Brazilian Journal of Medical and Biological Research (2007) 40: 615-619
ISSN 0100-879X Short Communication

Pipeline for macro- and microarray analyses

R. Vicentini[1,2] and M. Menossi[1,2]

[1]Laboratório de Genoma Funcional, Centro de Biologia Molecular e Engenharia Genética, [2]Departamento de Genética e Evolução, Instituto de Biologia, Universidade Estadual de Campinas, Campinas, SP, Brasil

Abstract

The pipeline for macro- and microarray analyses (PMmA) is a set of scripts with a web interface developed to analyze DNA array data generated by array image quantification software. PMmA is designed for use with single- or double-color array data and to work as a pipeline in five classes (data format, normalization, data analysis, clustering, and array maps). It can also be used as a plugin in the BioArray Software Environment, an open-source database for array analysis, or used in a local version of the web service. All scripts in PMmA were developed in the PERL programming language and statistical analysis functions were implemented in the R statistical language. Consequently, our package is a platform-independent software. Our algorithms can correctly select almost 90% of the differentially expressed genes, showing a superior performance compared to other methods of analysis. The pipeline software has been applied to 1536 expressed sequence tags macroarray public data of sugarcane exposed to cold for 3 to 48 h. PMmA identified thirty cold-responsive genes previously unidentified in this public dataset. Fourteen genes were up-regulated, two had a variable expression and the other fourteen were down-regulated in the treatments. These new findings certainly were a consequence of using a superior statistical analysis approach, since the original study did not take into account the dependence of data variability on the average signal intensity of each gene. The web interface, supplementary information, and the package source code are available, free, to non-commercial users at http://ipe.cbmeg.unicamp.br/pub/PMmA.

Correspondence
M. Menossi
Laboratório de Genoma Funcional
Centro de Biologia Molecular e Engenharia Genética, UNICAMP
Caixa Postal 6010
13083-875 Campinas, SP
Brasil
Fax: +55-19-3521-1089
E-mail: menossi@unicamp.br

Research partially supported by FAPESP (Nos. 02/01167-1 and 03/07244-0). R. Vicentini was supported by a fellowship from Instituto UNIEMP and M. Menossi was the recipient of a research fellowship from CNPq.

Received February 21, 2006
Accepted February 2, 2007

Key words
- Intensity-dependent selection of expression ratios
- Pipeline
- BioArray Software Environment
- Normalization
- Macroarray
- Microarray

Expressed sequence tags (ESTs) have provided insights into transcribed genes in a variety of organisms and are widely used for gene discovery and expression analysis. DNA arrays have been used to study gene expression patterns on a genomic scale because they allow the simultaneous analysis of thousands of genes. DNA arrays consist of sets of probes, such as oligonucleotides or ESTs, that are arranged in an orderly manner on a surface and are hybridized with cDNA obtained from different individuals, tissues or physiological states. The signal obtained from each sample is a *bona fide* indicator of the amount of RNA in the cells, thus allowing the identification of genes with significant changes in their expression. Arrays can be constructed using either a hand-held or a

robotically controlled arraying pin tool. Each spot in the array has a specific address and represents an oligonucleotide or EST. Often the arrays may be constructed with technical replicates, where the same sequence is fixed in different spots. Surfaces such as glass (microarrays) (1) or nylon (macroarrays) (2) are ideal for arrays since nucleic acid can be immobilized, hybridized, and detected using the standard techniques of molecular biology.

However, because of the characteristic noise of array data and the usually limited number of experimental replicates, the selection of differentially expressed genes is not easily determined (3). This is especially the case in macroarray experiments (4), where the requirement for larger amounts of RNA can be a limiting factor in the number of replicates, and each labeled cDNA sample is hybridized to a different nylon membrane, thereby increasing the data variability (in contrast to microarrays, in which two samples labeled with different dyes are hybridized on the same slide).

The large amount of data obtained in these experiments makes the visualization and analysis of the results challenging. We have developed a suite of interactive scripts (pipeline for macro- and microarray analyses, PMmA) that can be used as a pipeline to manage and to analyze data from macro- and microarrays. PMmA works with data generated by several types of image analysis software, has a web interface, and can be used with a plug-in in the BioArray Software Environment (BASE), an open-source database (5). All modules in normalization, clustering and data analysis classes have plug-ins for installation in BASE. This make it easy for BASE users to run all analyses provided by PMmA.

An important aspect of PMmA is its flexibility in handling data from spotted single-color arrays, such as filter arrays (macroarrays). There are only a few free tools available for this technique (6), which is cheaper than microarrays and is easier to implement in molecular biology laboratories.

All scripts in PMmA were developed in the PERL programming language (7) and statistical analysis functions were implemented as scripts in the R statistical language (8). PERL and R are freely available under the GNU General Public License and can be used on a wide variety of platforms. Consequently, our package is a platform-independent software because it will run in Linux, Windows and MacOS, and has low hardware requirements (the minimum computer free space required is 2 MB). The user can also build a local version of the web service, although he will need a web server with CGI allowed, and PERL and R language locally installed. For an optional interface with BASE, a local installation of this database is necessary. The basic steps for PMmA installation - the edition of a configuration file and the configuration of the web server to use the package - can be followed by any user with basic concepts in informatics.

The PMmA package consists of ten modules that work in a web interface to provide easier access to the main computational environment. The input data converter is flexible and can use data from any image quantification software that generates tab-delimited data and information on the volume (pixel intensity in the spot area) and background. The data normalization feature adjusts several effects of DNA array techniques and allows direct comparison of the data, thus compensating for variations between cDNA labeling, hybridization, and so on. The value used to normalize the data can be the median or average spot intensity, or a value selected by the user, such as the array global background. Flags with different meanings are generated by the module Arrayflags that can also combine the result with a map of spots in the array, thereby providing the user with a more complete output. These flags are calculated when spot

and background intensities are closer or when there is wide variation between intensity of replicate spots. The plot tool easily generates MA-plots, MS-plots and scatterplots using the normalized data. For experiments with three or more replicates, the Student *t*-test can be used to identify genes that are up- or down-regulated by a given treatment, whereas when little or no replication of array hybridizations is available, intensity-dependent selection of expression ratios (ISER) can be used. In fact, ISER is designed to deal with data from single-replicated experiments, but can also be used when more samples are compared or replicates are present (9). In this case, the algorithm is applied separately to the comparison of each pair of samples and to each replicate, thus selecting genes that, for any pair of samples, are identified as differentially expressed in all replicates or in a large (pre-defined) proportion of them. The results of ISER or the Student *t*-test can also be grouped into a more legible form that shows the percentage of genes that are up- or down-regulated in each experiment. Finally, the output data may be clustered using hierarchical clustering or self-organizing map algorithms. The pipeline kernel design is shown in Figure 1.

To show this scripts in action we used the PMmA to analyze 1536 ESTs of sugarcane exposed to cold for 3 to 48 h with control in 0 h (10). These data are taken from the Gene Expression Omnibus public database (11) under accession numbers GPL210 (platform), GSE83 and GSE84 (series), and GSM2431 to GSM2442 (samples). These data consist of a set of two high-density filters, each filter containing 768 ESTs in replicate, hybridized with a probe of cold treatment (0, 3, 6, 12, 24, and 48 h at 4ºC). Figure 2 shows the pipeline's graphic output for the comparison of 3-h (Data 1) versus 12-h (Data 2) cold treatment in the second filter.

We noted that PMmA identified 61 clones with low-quality spots, which were excluded from analysis. Thirty cold-responsive genes that were not previously identified by the authors were found in our analysis. Fourteen were up-regulated, two had a variable expression and the other fourteen were down-regulated in the treatments. These genes complemented the original results of Nogueira et al. (10), providing further characterization of the response of sugarcane to low temperatures, and their identification was certainly a consequence of using different statistical

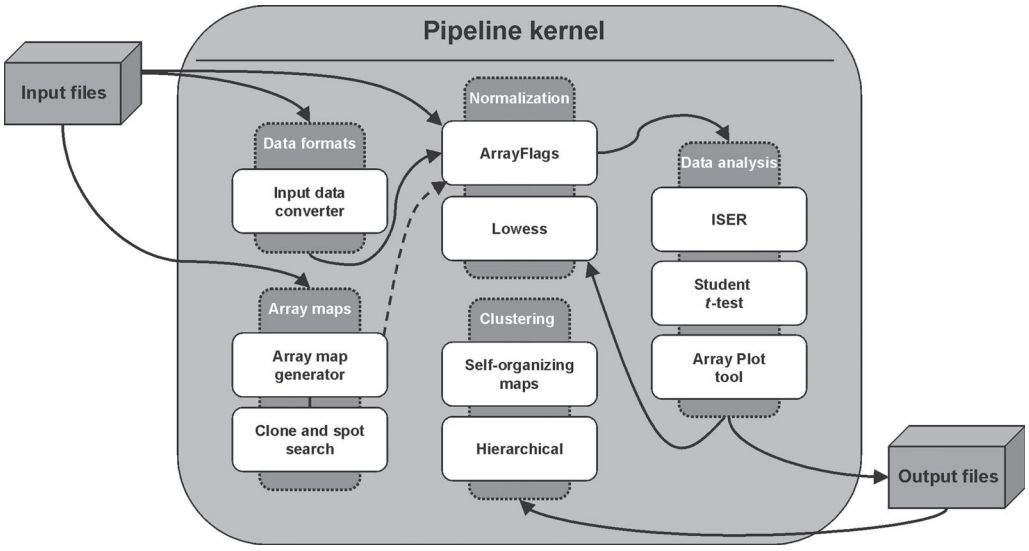

Figure 1. The diagram of the pipeline for macro- and microarray analyses. A pipeline kernel diagram of pipeline for macro- and microarray analyses shows the ten modules in five classes (dotted boxes). Solid arrows indicate the flow of information and the dashed arrow indicates an optional flow. Note that the output of the analysis may be edited prior to appropriate clustering. ISER = intensity-dependent selection of expression ratios.

analysis approaches. In the original study, data analysis consisted of calculating the expression ratios (treatment/control) and their logarithms for each time point and each replicate. Then, since these log ratios were assumed to be normally distributed, their mean and standard deviation were calculated and the genes that showed a log ratio more than 1.65 standard deviations distant from the mean and a ratio ≥2 (respectively ≤0.5 for repressed genes) in both replicates were considered to be differentially expressed. This approach does not consider the dependence of data variability on the average signal intensity of each gene and therefore it may select more false-positive genes than ISER.

PMmA is available free and is a fully integrated tool that will facilitate the analysis of data from DNA arrays. Besides being integrated with BASE, an open-source database for DNA array analysis, PMmA normalizes the data, generates information about variations in expression, identifies up- or down-regulated genes using the Student *t*-test and a new intensity-dependent algo-

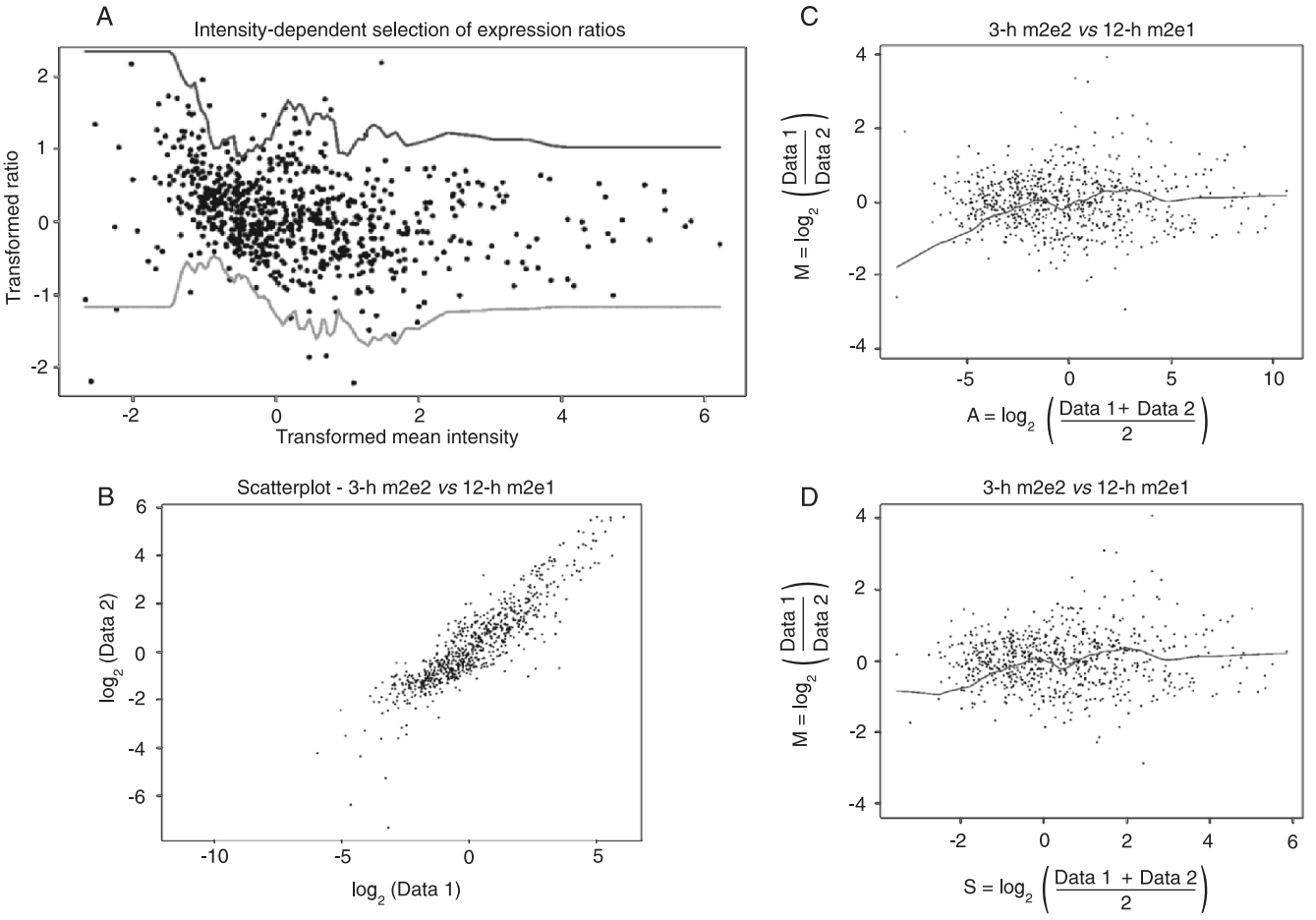

Figure 2. Data analysis with the pipeline for macro- and microarray analyses. Hybridization with RNA from sugarcane plants exposed to cold stress for 3 and 12 h. *A*, Intensity-dependent selection of expression ratios graphic output displaying the selection of differentially expressed genes, between 3 (Data 1) and 12 h (Data 2), based on the data variability dependence of intensities. *B*, Scatterplot displaying the $\log_2$ normalized data. *C* and *D*, MA- and MS-plots with lowess correction. m2e1 and m2e2 codes indicate the membrane (m) and experiment (e) numbers defined by the user.

rithm developed by our group (ISER) (9), and generates scatter plots, MA-plots, MS-plots, hierarchical clustering (12), and self-organizing maps (13). The use of the ISER algorithm can correctly select almost 90% of the differentially expressed genes. It is more sensitive to intensity-dependent bias in the data, thus showing a superior performance compared to the other methods of analysis, and was unaffected by the normalization adopted. Additional new methods of analysis will be included in future versions of PMmA, mainly the SAM method (14). This method is clearly better than the *t*-test. The fundamental problem with the *t*-test for array experiments is that the repetition numbers are often small. In this case the problem is to obtain accurate estimates of the standard deviation of individual gene measurements based on only a few measurements. PMmA is particularly useful for the community that uses cDNA arrays. This package will help small labs with no access to a bioinformatic center to perform modern computational analyses of genetic data.

Acknowledgments

The authors thank Cristiane de Souza Rocha (Laboratório de Genoma Funcional, Centro de Biologia Molecular e Engenharia Genética, Universidade Estadual de Campinas, Campinas, SP, Brazil) for technical assistance.

References

1. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 1995; 270: 467-470.
2. Desprez T, Amselem J, Caboche M, Hofte H. Differential gene expression in *Arabidopsis* monitored using cDNA arrays. *Plant J* 1998; 14: 643-652.
3. Lee ML, Kuo FC, Whitmore GA, Sklar J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc Natl Acad Sci U S A* 2000; 97: 9834-9839.
4. Freeman WM, Robertson DJ, Vrana KE. Fundamentals of DNA hybridization arrays for gene expression analysis. *Biotechniques* 2000; 29: 1042-1055.
5. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol* 2002; 3: SOFTWARE0003.
6. Dudoit S, Gentleman RC, Quackenbush J. Open source software for the analysis of microarray data. *Biotechniques* 2003; Suppl: 45-51.
7. The Source for Perl. http://www.perl.com. Accessed July 11, 2006.
8. The R Project for Statistical Computing. http://www.r-project.org. Accessed July 11, 2006.
9. Drummond RD, Pinheiro A, Rocha CS, Menossi M. ISER: selection of differentially expressed genes from DNA array data by non-linear data transformations and local fitting. *Bioinformatics* 2005; 21: 4427-4429.
10. Nogueira FT, de Rosa VE Jr, Menossi M, Ulian EC, Arruda P. RNA expression profiles and data mining of sugarcane response to low temperature. *Plant Physiol* 2003; 132: 1811-1824.
11. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002; 30: 207-210.
12. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 1998; 95: 14863-14868.
13. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* 1999; 96: 2907-2912.
14. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 2001; 98: 5116-5121.