# *Analysis of DNA microarray data. Part II: Quantification and analysis of gene expression*

REVIEW

✍ Jamilet Miranda, Ricardo Bringas

Center for Genetic Engineering and Biotechnology, Havana, Cuba
Ave 31 e/ 158 and 190, Playa, PO Box 6162, Havana 10 600,
E-mail: jamilet.miranda@cigb.edu.cu

***ABSTRACT***

The first step in a DNA microarray experiment is to define the biological question to be addressed and the selection of an appropriate experimental design. In a second step samples are processed and hybridized to the chips. Next, the fluorescent images are obtained and a processing step begins in which the images are analyzed, the expression values assigned and mathematical-statistical methods suited to the goals of the study are applied. Here a particular emphasis is made on the methodology used to quantify and analyze gene expression data. The most frequently used programs are briefly described and working plans are proposed for some of the most common experimental objectives. Additionally, we discuss and comment the applications of this technology in the field of Oncology, where it has enabled the discovery and classification of new cancer subtypes, and has helped to identify new therapeutic targets, as well as improving the prediction of disease stages.

*Keywords*: DNA Microarrays, gene expression, quantification, statistical analysis

*Biotecnología Aplicada 2008;25:301-311*

***RESUMEN***

**Análisis de datos de microarreglos de ADN. Parte II: Cuantificación y análisis de la expresión génica.** Los experimentos de microarreglos de ADN constan de una primera etapa, en la que se define la pregunta biológica objeto de investigación y se selecciona el diseño experimental que mejor responda a los objetivos. La segunda etapa del experimento comienza una vez que las muestras se procesan e hibridan en los chips. Cuando se obtienen las imágenes fluorescentes, se inicia la etapa de procesamiento, en la que se analizan las imágenes para asignar los valores de expresión, y se aplican métodos estadístico-matemáticos que permitan cumplir los objetivos de la investigación. En este artículo se enfatiza en la metodología para la cuantificación y el análisis de los datos de expresión. Se describen los programas más utilizados para estos análisis y se proponen esquemas de trabajo para acometer algunos de los objetivos más frecuentes. Además, se comentan aplicaciones de esta tecnología en Oncología, donde ha habido avances en cuanto a: la clasificación de nuevos subtipos de cáncer, la identificación de nuevos blancos terapéuticos y la predicción de estadios de la enfermedad.

*Palabras clave*: Microarreglos de ADN, expresión de genes, cuantificación, análisis estadístico

## *I*ntroduction

The availability of complete genome sequences has marked the advent of genome wide high throughput technologies that provide information on the different levels of physiological regulation taking place in living beings. These technologies, of which DNA microarrays are a prominent example, typically generate large volumes of biological data that require the development of customized information systems to cope with tasks such as data collection, management and analysis. In the specific case of DNA microarrays, the development of statistical methods for the analysis of datasets with a high number of variables but a limited number of measurements has steadily gained importance as the technology has matured and become widely used.

During the interpretation of the results from microarray experiments, and especially in studies focused on the molecular basis of disease and other biological phenomena, the statistical analysis of expression matrices must be complemented with the study of other available sources of information and biological ontology, such as databases for protein-protein interactions, transcriptional regulators and, in general, functional annotation databases obtained through either experimentation or predictive algorithms. Such an integrated approach is already a fundamental part of the arsenal of Systems Biology [1]. Although some applications are already available for the integration and statistical analysis of this information [2], it is still a challenge the development of statistical-mathematical algorithms that can lead to the formulation of biological hypotheses based on gene/protein interaction networks linked to expression data.

In the first part of this paper we discuss matters such as defining the biological hypothesis, experimental goals and design, which are essential for the data obtained to be able to answer the scientific questions posed by the researcher. The present part deals with relevant methodological topics for the analysis of microarray data, which are of importance whether the scientist is dealing with the results of their own experiments or analyzing data from public repositories.

## *Q*uantification of gene expression

In DNA microarray experiments, once the samples are hybridized, the chips are scanned and the corresponding images are generated. The gene expression data from each sample is obtained through image analysis, which generates what is commonly known as an expression matrix, with each row representing a gene and each column corresponding to a sample.

1. Hwang D, Smith JJ, Leslie DM, Weston AD, Rust AG, Ramsey S, *et al*. A data integration methodology for systems biology: experimental verification. Proc Natl Acad Sci USA 2005:102:17302-7.

2. Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004; 20: 578-80.

✍ Corresponding author

The initial analysis usually consists of transforming the raw data contained in this matrix into normalized and pre-processed data, which is then processed in later stages using statistical methods. One of the challenges of microarray experiments is the collection, management and analysis of these data.

## Image analysis for the quantification of gene expression

On reading each chip images are generated consisting of matrices of 16-bit pixels, which have individual luminescence values ranging from 0 to 65 535 ($2^{16}$-1). The intensity value for each gene is not calculated from an individual pixel, but rather from a set of pixels. The steps taken to assign an intensity value to a gene are:

### *Localization of the signal*

This process is usually automatic. It locates the pixel rectangle where the signal should be contained and assigns the coordinates of this rectangle to the corresponding gene.

### *Segmentation*

This process classifies the pixels forming the image of the chip as either signal or background. This is a very important step, since the intensity value assigned to each gene depends to a large extent on the differences in luminescence between the signal and background pixels. Although there are segmentation algorithms that assume a circular shape for the signal (classified as either fixed circle segmentation methods -assigning a circle of the same diameter for all signals on the chip- or adjustable circle segmentation methods, which use individual pixel intensities to estimate a separate diameter for each probe), they are considered to be unsatisfactory due to the often irregular shape of the signals. Therefore, the most widely used algorithms use the individual luminescence values of the pixels forming the signal to determine its contour [3, 4], resulting in better and more accurate estimates of the level of gene expression.

### *Calculation of the intensity (signal)*

After locating the rectangle corresponding to each gene on the chip and identifying the area corresponding to the signal within this rectangle based on the segmentation algorithms discussed above, the intensity of the signal is calculated from the luminescence of the pixels above that of the background.

### *Background correction*

After the processes of localization, segmentation and calculation of the intensity, the luminescence of background pixels from adjacent areas of the signal are also analyzed and quantified in order to estimate the contribution of non-specific, background interactions to the calculated intensities and therefore, to subtract this contribution from the gene expression values obtained. There are different methods for this purpose, differentiated mainly by whether they estimate a common background correction for the complete surface of the chip or a local correction coefficient for each signal or for clusters of adjacent signals. Since background levels can and often do vary within a microarray, local correction methods have been largely favored for this purpose [5, 6].

### *Signal exclusion criteria*

During the process of calculating the intensity values, the signals must be examined to detect inconsistencies among their constituent pixels and to evaluate their possible elimination from the dataset. For instance, any signal with a high variability in the luminescence of its pixels must be eliminated; this is commonly the case for low-intensity signals which are close to background intensity.

Yang et al. demonstrated how in some cases the use of background correction algorithms can actually significantly worsen the accuracy, given that they usually increase the variability of low-intensity signals while the different segmentation procedures introduce low levels of variability into the resulting precision [7]. Additionally, Wang *et al*. showed, by using a quality scoring function, that the values from high quality signals were less variable than those from low-quality ones, thus demonstrating how the inherent variability in measurements of intensity ratios is inversely related to signal quality [8].

### *Storage of expression data*

Several Laboratory Information Management Systems (LIMS) custom-tailored for the storage and analysis of microarray gene expression data have been developed (Table 1). These systems comply with the guidelines of an international standard established by the Microarray Gene Expression Data (MGED) Society, known as MIAME (Minimum Information About a Microarray Experiment) [9], based on structured tables for the storage of data from the samples, the experimental conditions studied, and the expression values themselves. Additionally, there is an increasing trend towards the incorporation of data analysis as an integral part of these systems. BASE (BioArray Software Environment, http://base.thep.lu.se) [10], which is one of the most popular LIMS, is a database server that contains the information on the biomaterials, the primary expression data and images,

3. Beucher S, Meyer F. The morphological approach to segmentation: the watershed Transformation. In: Dougherty E, editor. Mathematical morphology in image processing. New York: Marcel Dekker; 1993. p. 433-81.

4. Adams R, Bischof L. Seeded region growing. IEEE Trans. Pattern Anal. Mach. Intell 1994;16:641-47.

5. Yang YH, Buckley MJ, Dudoit S, Speed TP. Analysis of Cdna microarray images. Brief. Bioinf 2001;2:341-9.

6. Jain AN, Tokuyasu TA, Snijders AM, Segraves R, Albertson DG, Pinkel D. Fully automatic quantification of microarray image data. Genome Res 2002;12(2): 325-32.

7. Yang YH, Buckley MJ, Dudoit S, Speed TP. Comparison of methods for image analysis on cDNA microarray data. J Comput Graph Stat 2002;11:108-36.

8. Wang X, Ghosh S, Guo S. Quantitative quality control in microarray image processing and data adquisition. Nucleic Acids Res 2001;29:E75-5.

9. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, *et al.* Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001;29: 365-71.

Table 1. The most popular systems for the storage and management of microarray data

| Available LIMS* | Institute | Database management system | URL |
|---|---|---|---|
| BASE (BioArray Software Enviroment) | Oncology Department, Lunds University | MySQL, PostgreSQL | http://www.lu.se/ |
| MaxdSQL (Manchester Array Express Database) | Microarray Bioinformatics Group, Manchester University | Oracle, MySQL, PostgreSQL | http://www.bioinf.man.ac.uk/microarray/maxd/ |
| MADAM (MicroArray Data Manager) | The Institute of Genomic Research - TIGR | MySQL | http://www.tigr.org/software/tm4/madam.html |
| SMD (Stanford Microarray Database) | Stanford University | Oracle, LAD (The Longhorn Array Database) es una implementación de SMD en PostgreSQL | http://genome-www5.stanford.edu/MicroArray/SMD/download/ |
| ArrayExpress | EMBL-EBI | Oracle | http://www.ebi.ac.uk/arrayexpress/ |

*Laboratory Information Management Systems.

and also has facilities for the normalization, visualization and analysis of gene expression data.

There are several LIMS that also double as public microarray data repositories. Among the most important ones due to the volume of data and their constant update are the SMD (Stanford Microarray Database, http://genome-www5.stanford.edu/) [11, 12], providing access to more than 300 experiments, ArrayExpress (http://www.ebi.ac.uk/ arrayexpress/) [13, 14] from EMBL-EBI (European Molecular Biology Laboratory-European Bioinformatics Institute), with more than 1800 experiments, and GEO (Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/) [15, 16] from NCBI (National Center for Biotechnology Information), with more than 4900 public expression data series. These databases attract the interest of researchers not only because of the chance to analyze data from individual experiments, but because of the opportunity they provide for studying global and specific molecular mechanisms for diseases and other biological phenomena through meta-analyses of microarray data obtained from different experiments [17-19].

## Pre-analysis: filtering, normalization and pre-processing of expression data

After image analysis and intensity assignment it is always necessary to transform and process the primary data, since these intensities not only reflect the levels of mRNA expression but also contain biases associated with variations during chip printing, sample labeling and other sources of variability. The processes of filtering, normalizing and pre-processing the data intend to remove these biases.

### *Data filtering*

A first step in processing primary intensity data is to eliminate the values that most probably arise from experimental errors. One of the criteria used for this is the calculation of the coefficient of variation (CV) for each gene (defined as the ratio of the standard deviation (SD) and the mean intensity for multiple signals from the same gene), filtering out any results above a certain threshold. Another criterion for filtering out invalid data is to eliminate signals that are above the threshold corresponding to over saturated values. It is also recommended to visually inspect the images to detect defects of the array such as scratches, fogs, edge effects and bubbles, and eliminating the corresponding intensities before the data normalization stage [20].

### *Normalization*

The process of normalization must be the first transformation applied to gene expression values and it is an essential step before data analysis. This transformation attempts to minimize systematic errors arising during the quantification of the hybridization of mRNA samples in order to more easily identify biological differences [21] and to be able to compare expression levels between chips. Normalization is generally applied either within each chip or between multiple chips, and therefore demands a careful selection of the most appropriate method and variables or spatial regions of the chip (gene set) to be used for data standardization.

### *Set of genes to be used for normalization*

In general, normalization methods assume that:

1) The expression levels of most genes changes and the signals are normalized based on a restricted set of genes with invariable expression between the experimental conditions (reference genes), or:

2) There will not be changes in the expression for most genes of the array between the different experimental conditions; therefore the signals can be normalized based on the intensities of all signals from the chip [22, 23].

Methods based on the first assumption are used generally when the array contains only a selection of genes known to be associated to the biological problem under study, *e.g.* genes related to a specific disease. In this case it is possible to use a set of reference genes whose expression is previously known to be constant across the different experimental conditions, as it is the case of genes for essential functions which must always be expressed at similar levels, also known as housekeeping genes. Another approach is to include control probes from genes not expressed in the sample, as is the case of genes from evolutionarily distant organisms, although this variant is not widely used. Methods based on the second assumption, on the other hand, are widespread when using chips with genomic coverage. Additionally, Yang *et al.* [21] have proposed other methods for normalization which separate the area of the chip per printing groups and apply the chosen method on a per-group basis, thus avoiding edge effects between different printing groups in the same chip.

### *Normalization methods*

There are several normalization methods:

- Global or linear (applicable to cDNA and Affymetrix-type chips): The normalization factor is the same for every gene in the chip.

- Intensity-dependent (applicable to cDNA and Affymetrix-type chips): The normalization factor depends on the intensity of each signal.

- Location-dependent (applicable to cDNA chips): The normalization factor depends on the position of each signal on the surface of the chip.

These methods have been extensively used for the normalization of gene expression data [21, 24, 25] and can be applied either within an array or between pairs of arrays.

The selection of a normalization method depends on the type of chip used for the microarray experiment. As an example, let define R as the set of intensities for the red signals and G as the set of intensities for the green signals of a cDNA chip.

Here, the global, or linear method, assumes that there is the same amount of RNA in the samples compared, and they therefore contain the same number of molecules. It also assumes that the probes printed on the chip represent a random sample of the genes from an organism. If these two assumptions hold true it follows that the same number of labeled molecules from each sample must hybridize to the array and, therefore, the total sum of intensities from all probes in the array must be the same for each sample. Based on

10. 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.

11. Sherlock G, Hernández-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, *et al.* The Stanford Microarray Database. Nucleic Acids Res 2001;29:152-5.

12. Ball CA, Awad IA, Demeter J, Gollub J, Hebert JM, Hernández-Boussard T, *et al.* The Stanford Microarray Database accommodates additional microarray platforms and data formats. Nucleic Acids Res 2005;33:D580-2.

13. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, *et al.* ArrayExpress- A public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2003;31:68-71.

14. Parkinson H, Sarkans U, Shojatalab M, Abeygunawardena N, Contrino S, Coulson R, *et al.* ArrayExpress- A public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2005; 33:D553-5.

15. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207-10.

16. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, *et al.* NCBI GEO: mining tens of millions of expression profiles-database and tools update. Nucleic Acids Res 2007;35:D760-5.

17. Segal E, Friedman N, Koller D, Regev A. A module map showing conditional activity of expression modules in cancer. Nat Genet 2004; 36:1090-8.

18. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. Large scale-meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. Proc Natl Acad Sci USA 2004; 101:9309-14.

19. Kim RD, Park PJ. Improving identification of differentially expressed genes in microarray studies using information of public databases. Genome Biol 2004; 5:R70

20. Troyanskaya Missing value estimation methods for DNA microarrays. Bioinformatics 2001;17(6):520-5.

21. Yang IV, Chen E, Hasseman JP, Liang W, Frank BC, Wang S, *et al.* Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol 2002;3(11).

22. Quackenbush J. Microarray data normalization and transformation. Nature Genetics Supp 2002;2:496-501.

23. Kroll TC, Wolfl S. Ranking: a closer look on globalisation methods for normalisation of gene expression arrays. Nucleic Acids Res 2002;30:e50.

this simple assumptions a normalization factor k is calculated by adding the intensities from both channels:

$$k = \frac{\sum_{i=1}^{N} R_i}{\sum_{i=1}^{N} G_i}$$

and the normalized expression ratio $T'_i$ for each probe on the chip would be:

$$T'_i = \frac{R'_i}{G'_i} = \frac{1}{k} \frac{R_i}{G_i}$$

This transformation is equivalent to subtracting a constant from the logarithm of the expression ratio:

$$\log_2 (T'_i) = \log_2 (T_i) - \log_2 (k)$$

There are variants of this method depending on whether the mean or the median of the intensities are considered to be the same within each array or across all arrays; it can also be applied to only a subset of genes instead of all the genes in the chip. Kroll *et al*. performed a comparative analysis of the methods of normalization that are based on an intensity ratio factor [23], studying different variants for the normalization factor such as the mean of the expression of the set of reference genes; the sum, the mean, the median, the quartile or percentile of the logarithms of all expression values and the mean excluding the highest intensity values; and concluded using the mean of the central values of intensity as the factor after excluding 5 and 10% of the highest values is a simple and robust method for the normalization of this type of data. One problem in these methods, however, is that they do not take into account intensity and block effects which have been described in other studies [26-28].

Dudoit *et al*. [24] suggested the use of a graph constructed with the logarithm of the primary data:

$$M = \log_2 (R / G) \quad \text{v.s} \quad A = \frac{1}{2} \log_2 (R*G)$$

which is useful for the identification of signal noise arising from differences in labeling efficiency. This is known as an MA graph, because of the name of the variables being charted.

If the same sample, labeled with both fluorophores, is hybridized to the same chip, it is expected that the value of $\log_2 (R/G)$ equals 0; however, in most cases what is actually observed is a deviation from 0 both for high and low intensities. Lowess (LOcally Weighted rEgression and Smoothing Scatterplots) [29] is a normalization method that can eliminate these labeling-specific biases relying on the intensity values. Lowess can be used to estimate a function to adjust the values of the MA-graph; this function will only be affected by differentially expressed genes, which will therefore behave as outliers in this case. By performing this normalization, the following transformation is applied:

$$\log_2 R / G \longrightarrow \log_2 R / G - c (A) = \log_2 R / k (A)*G$$

where $A=\log_2 \sqrt{(R*G)}$ and the values $R'$, $G'$ normalized by *lowess* are transformed as follows:

$$G' = 2^{c (A)} *G \quad , \quad R' = R$$

This method can be applied to the whole chip or a function can be estimated for each partition of the chip.

Yang *et al*. [21] studied this intensity-dependent effect by performing hybridizations in 30 cell lines, constructing MA-graphs for each one and calculating the standard deviation before and after a lowess correction. This demonstrated the convenience of the lowess method for intensity-dependent alterations.

On the other hand, Yang *et al*. [7] performed a wider comparison by using cDNA microarray data and including different normalization methods (global, lowess based on all chip probes, *lowess* per chip per printing block, *lowess* across every pair of chips with reverse labeling -which is only used to eliminate fluorophore biases in cDNA arrays- and a method of variance regularization applied per chip, per printing blocks and between chips), which showed that using lowess on each chip per printing block yielded the best results. This work also evidenced the need for this type of normalization method, which -unlike global methods- can eliminate biases that depend on signal intensity and on the spatial location of the probe over the surface of the chip.

The current general consensus for normalization is to normalize each chip globally and with *lowess* per printing group; adding a reversed labeling normalization step when using two fluorophores. However, some groups still search for improved normalization algorithms that not only yield better results, but are based on different biological hypotheses. For instance, Fan *et al*. [30] presented a method based on a semi linear model, applied within the same chip, which estimates intensity and block effects by using 100 replicates of the same gene. In their method, the estimated values of the effects are eliminated from all intensity values, followed then by a global normalization to correct for other effects. Methods such as this are a useful alternative when the biological premises of the lowess method (that the expression of most genes remains constant or that the number of up- or down -regulated genes is similar within each printing block) are not valid.

### *Pre-processing of expression data*

Once normalized, the expression values must undergo a treatment called pre-processing, which consists of a series of additional data transformations that attempts to partially correct a number of problems that may remain in the experimental results. Some of these transformations are:

#### Treatment of the replicates within each array

It is recommended to analyze and filter out inconsistent replicates, followed by the calculation of the mean or median expression values for each gene based on the intensities of all its valid replicates in the array.

#### Filling missing or blank data

Missing values can appear in an expression matrix for a number of reasons, including insufficient resolution of the scanner, or problems with the image or the chip itself. The most frequent solutions to deal with this problem are to fill-in the missing data with the median or the mean of the intensities for the corresponding

24. Dudoit S, Yang YH, Callow MJ, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica 2002;12:111-39.

25. Steinhoff C, Vingron M. Normalization and quantification of diferential expression in gene expression microarrays. Brief Bioinform 2006;7:166-77.

26. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations, and assessment of gene effects. Nucleic Acids Res 2001;29:2549-57.

27. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Stat Sci 2003;18:71-103.

28. Ma S, Kosorok MR, Huang J, Xie H, Manzella L, Soares MB. Robust semiparametric cDNA microarray normalization and significance analysis. Biometrics 2006;62:555-61.

29. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. J Am Stat Assoc 1979;74:829-36.

30. Fan J, Tam P, Woude GV, Ren Y. Normalization and analysis of cDNA microarrays using within array replications applied to neuroblastoma cell response to a cytoquine. Proc Natl Acad Sci USA 2004; 101:1135-40.

gene or with weighted values for the k nearest neighbors (kNN-Imputation); the latter variant is more robust [20]. Although the estimation of missing values is a necessary step for the application of clustering me-thods to identify groups of co-expressed genes or samples, it is dispensable in other cases (*e.g.*, when using a t-test) and can in fact give misleading results under those circumstances.

### Filtering out flat patterns

It is common to filter out genes that maintain a constant level of expression across all experimental conditions, since keeping them within the dataset can bias the results of later analyses, particularly when using clustering algorithms [31]. This filtering takes advantage of the fact that those genes usually have an expression profile with a very low standard deviation, and is implemented by visualizing all genes in the chip by standard deviation ranks, using the result to define an standard deviation threshold, and then eliminating the genes below this threshold.

### Exploratory data analysis

The techniques of exploratory data analyses are used to gather more detailed information about the available data, identifying relationships between variables without previous information on them and, occasionally, reducing or selecting a subset of the variables best suited for explaining and predicting the behavior of the system. Many of the techniques for exploratory data analysis fall within the field of descriptive statistics, and are based on the use of univariate analyses and the visualization of the distribution of the variables together with the estimation of their mean, median and standard deviation. During microarray studies it can be useful to visualize histograms of the number of genes per rank for each of these statistics. There are also multivariate exploratory techniques such as Principal Component Analysis [32, 33] that are even more useful due to the high number of variables that can be examined, which enable the construction of components as linear orthogonal combinations of the original variables. Each principal component accounts for a defined percentage of the variability observed in the system under study (preferably, the first two components should explain 80% or more of this variability) and the process leads to a coefficient that represents the weighting of each variable on each component. The most important variables, or genes, will therefore be those with the highest absolute values for the coefficients of the first and second component; if most of the variability is explained by the first component, later analysis can then be confined to the variables with the highest first-component coefficients.

For clarity, let us illustrate the use of this technique for the analysis of the 50 genes with the highest difference in expression levels between the healthy and tumorous tissues from prostate cancer, taken from the study of Lapointe *et al.* [34]. Figure 1 shows the values of each sample in a scatter plot where the axes represent the values of the first and second principal components. Both components account for 80% of the variability of the system under study (76.7 and 3.3%, respectively), and it is easy to notice how they separate the sample into two groups. This high percentage makes it possible to link these two principal components to the most evident phenomena in the analyzed data while at the same time selecting the genes with the highest coefficients in both components for later analyses. In this particular example, the first component is associated to a phenomenon of gene repression (the 30 genes with the highest coefficients) and the second component is associated with overexpression (two genes with the highest coefficients). One of the overexpressed genes is *AMACR*, known to exhibit a high activity in prostate tumors [35].

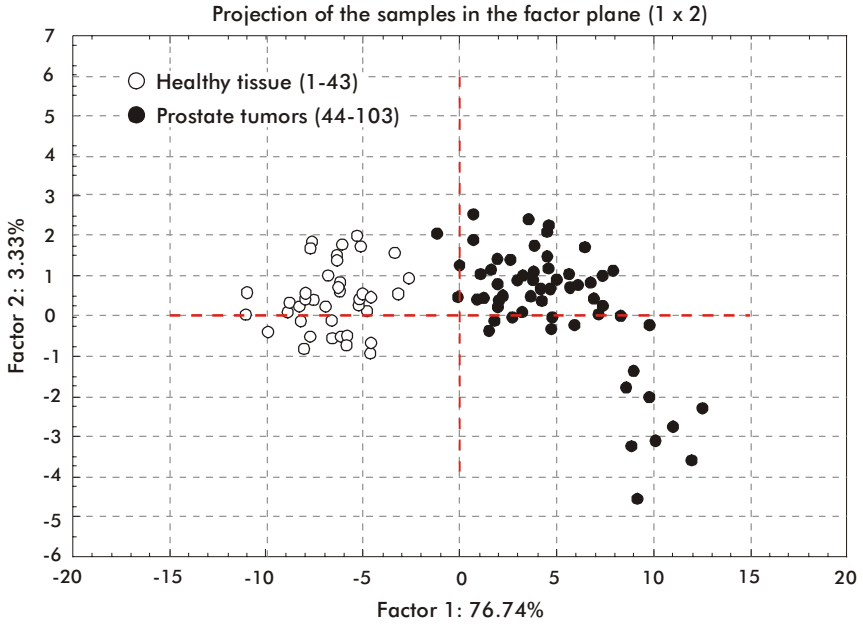


Figure 1. Projection of the values from the 103 samples onto the plane formed by the axes of the first and second principal components or factors. These components are obtained by the application of a Principal Component Analysis on the 50 genes with the most pronounced changes in expression between the classes formed by healthy tissue samples and prostate tumor samples. These two components separate the two experimental groups under study. The chart shown above is the output from program Stat 6.1, showing in each axis the percentage of the total variability accounted for by each factor (76.74% for the first component and 3.33% for the second). The expression data were taken from the public study of Lapointe *et al.* [34].

## Statistical-mathematical analyses according to the goals of the experiment

There is a close relationship between the goals of the experiment and the most suitable statistical method to be used [36] (Table 2). Statistical methods for the analysis of microarray data can be classified as supervised or unsupervised [37]. Supervised methods require the definition of classes or experimental groups. These include methods oriented towards the identification of genes with differential expression patterns between defined classes (comparison methods) and methods geared for the prediction of class membership (prediction methods). The later case requires a previous step of selection of the variables. Unsupervised methods (clustering methods) are mainly used for the identification of genes with similar expression patterns without the knowledge of their classification in a particular class (Figure 2).

31. Herrero J and Dopazo J. Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. J Proteome Res 2002;1:467-70.

32. Mardia KV, Kent JT, Bibby JM. Multivariate Analysis. London: Academic Press; 1979.

33. Venables WN, Ripley BD. Modern Applied Statistics with S. Springer-Verlag; 2002.

34. Lapointe J, Li C, Higgins JP, Van de Rijn M, Bair E, Montgomery K. Gene expression profiling identifies clinically relevant sub-types of prostate cancer. Proc Natl Acad Sci USA. 2004;101:811-6.

35. Kumar-Sinha C, Shah RB, Laxman B, Tomlins SA, Harwood J, Schmitz W, Conzelmann E, Sanda MG, Wei JT, Rubin MA, Chinnaiyan AM, et al. Elevated alpha-methy-lacyl-CoA racemase enzymatic activity in prostate cancer. Am J Pathol 2004 Mar; 164(3):787-93.

36. Simon R, Radmacher MD. Dobbin K. Design of studies using DNA Microarrays. Genet Epidemiol 2002;23:21-36.

37. Smyth GK, Yang YH, Speed T. Statistical issues in cDNA microarray data analysis. Methods Mol Biol 2003;224:111-36.

## Selection of significant genes: supervised methods

Class comparison methods are required to identify the genes whose expression profile changes significantly across the different experimental conditions. These are supervised learning methods that require the input of the group or experimental condition to which each sample belongs and as results identify differentially expressed genes between these defined experimental groups or conditions.

In general, the problem can be reduced to the selection of an adequate statistical test and the computation of p values for the samples. The choice of a statistical test depends on the number of conditions to be compared. The use of the fold change FC:

$$FC = \log_2 \left( \frac{\overline{x}_1}{\overline{x}_2} \right), \quad FC \geq 2 \quad \text{or} \quad FC \leq -2$$

is not recommended as a measure of differential expression, since in this case the differences in variance dominate the analysis.

The most accepted method for comparing two conditions is a modified *t*-test [24, 38] such as the one described by Tusher *et al.* [39], implemented in the SAM software (Significance Analysis of Microarrays, http://wwwstat.stanford.edu/~tibs/SAM/); in addition, there are SAM versions currently available for the comparison of multiple experimental conditions. In general, the modifications to the *t*-test are applied to the denominator. Tusher *et al.* propose a *d*-test where the modification to the *t*-test [40] consists of adding an $s_0$ value, such that for each gene it is possible to calculate:

$$d = \frac{\overline{x}_1 - \overline{x}_2}{\sigma^2 + s_0}$$

where $s_0 > 0$, $\sigma$ is the variability of the expression among the classes, and $\overline{x}_1 - \overline{x}_2$ represents the difference between the expression means of the gene between both classes.

There are different methods available for testing the hypothesis of differential expression [37, 38]: *1)* Those assuming that the data follows a normal distribution, such as the *t*-test for two groups or the *F* test for multiple conditions, *2)* Non-parametric tests such as the Wilcoxon's for two conditions, the Kruskal-Wallis's for multiple comparisons, which do not require a normal distribution of the experimental data since they are based on the use of sum ranks, or others less known non-parametric statistics[41-43], and *3)* Procedures based on Bayesian statistics [44-46]. The diagram in figure 3 shows a proposal for the use of these statistic tests, where the assays are grouped according to their requirements for a normal distribution of the data and the number of classes to which they can be applied. In any case, the results obtained after the application of the relevant method is a list of genes, ranked according to the value of the selected statistical test for differential expression.

It should also be stressed that this type of experiment usually makes statistic inferences on thousands of variables (genes) and, therefore, the results demand very small p values (in the order of $10^{-3}$ to $10^{-4}$). A very popular alternative is the calculation of un-adjusted *p* values, for which the use of resampling algorithms is recommended [47] since microarray data often do not follow a normal distribution. If the researcher wishes to account for any dependencies among the observed variables -which is a frequent event in functionally related genes- it is also necessary to compute *p* values which have been adjusted for multiple hypothesis tests with permutations [27].

Lastly, the genes which are differentially expressed are selected from the gene list, based on whether or not the relevant statistic test has a value above the chosen threshold or, for adjusted *p* values, under the chosen threshold. This stage also makes use of controls, such as the estimation of the false discovery rate (FDR) and/or the family-wise error rate (FWER), with different modifications for the calculation of the expected proportion of false positives or false negatives for microarray data [27, 48-50].

Table 2. Summary of the most frequently used statistical methodologies, according to the objective of the experiment

| Objective of the experiment | Most frequently used statistical methods | Reasons for the selection |
|---|---|---|
| Class comparison | t -test, **F-test**, Wilcoxon, Kruskal Wallis, SAM | The F-test can compare two or more experimental conditions, with better accuracy than non-parametric tests (Wilcoxon, Kruskal Wallis) in microarray data |
| Class prediction | **kNN**, D LDA, Naive Bayes, QDA, LDA, LOCLDA, SVM | Although kNN is a simple method, comparison studies of discriminant methods have consistently singled out kNN as the best performer for microarray data according to Dudoit *et al.* [52] |
| Class discovery | k-means, SOM, HCL, **SOTA** | SOTA is a clustering method combining SOM and HCL. It contains an implementation of a stopping criterion for the division of clusters based on the estimation of variability within the cluster. A specific design for microarray data developed by Herrero *et al.* [82] is also known as SOTArray |

*The methods recommended by the authors are shown in bold typeface, with the reasons behind their choice in the third column.

## Searching for a molecular signature: supervised methods

Class prediction methods are needed to find a molecular signature (*i.e.*, a reduced set of genes whose expression profiles allow the classification of the sample), which are also supervised learning methodologies. In this case the goal is to find a multivariate predictor that can assign an unknown sample or individual to a specific class. The comparative study published by Dudoit *et al.* [51] revealed that the most accurate method for this purpose is that known as kNN (k Nearest Neighbors) [52], which yielded the smallest number of classification errors.

38. Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 2002;18:546-54.

39. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 2001;98:5116-21.

40. Welch BL. The generalization of 'students' problem when several different population variances are involved. Biometrika 1947;34:28-35.

41. Troyanskaya OG, Garber ME, Brown PO, Botstein D, Altman RB. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 2002;18:1454-61.

42. Zhao Y, Pan W. Modified nonparametric approaches to detecting differentially expressed genes in replicated microarray experiments.Bioinformatics 2003;19:1046-54.

43. Yan X, Deng M, Fung WK, Qian M. Detecting differentially expressed genes by relative entropy.J Theor Biol 2005; 234:395-402.

44. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 2001;8:37-52.

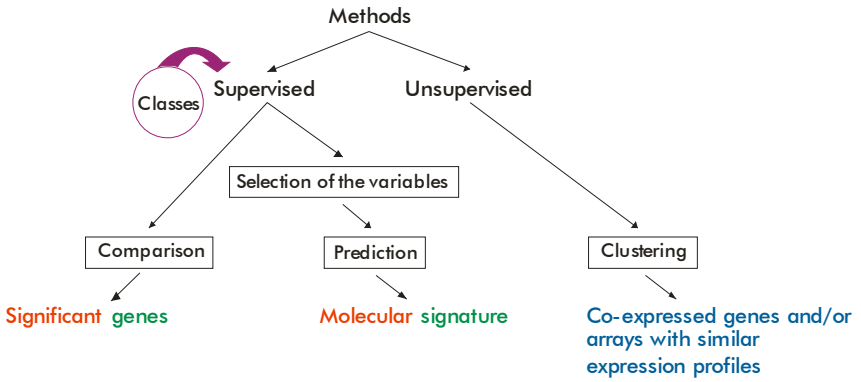


Figure 2. Diagram depicting the general workflow followed during the statistical analysis of microarray data. Unless un-supervised methods are used, experiments of class comparison or prediction require the a priori definition of the experimental group (class) to which the sample belongs.

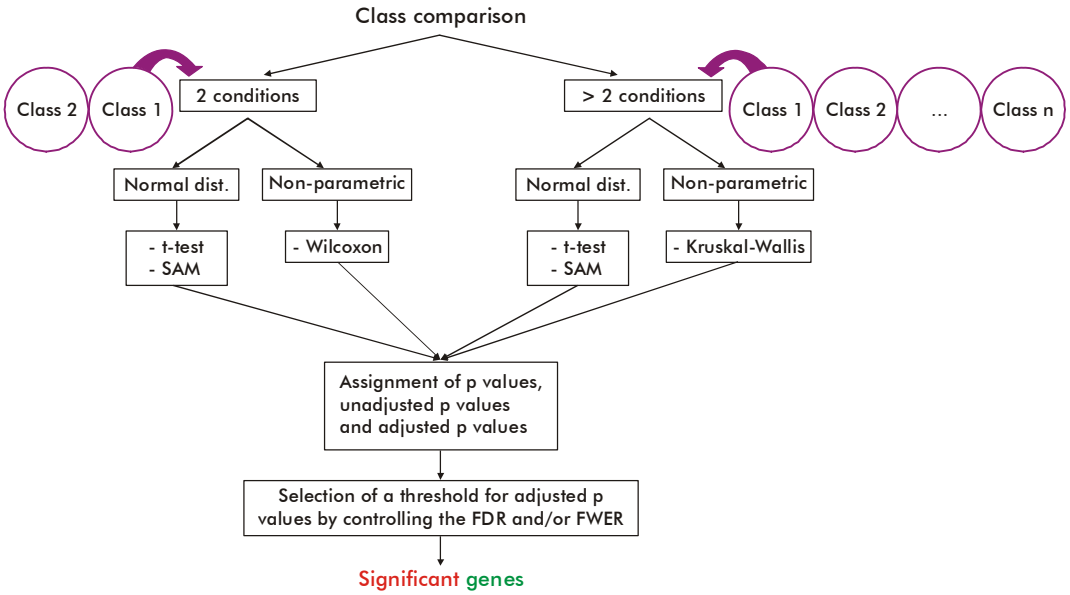


Figure 3. Proposed workflow for statistical analysis focused on the discovery of genes with statistically significant expression changes during the comparison of two or more experimental conditions. Some statistical tests assume a normal distribution of the data; non-parametric tests make no such assumptions. Regardless of the specific test, a p value is computed and the samples are permuted between the groups under study for obtaining an adjusted p values. Finally, the adjusted p values are computed using multiple hypotheses testing with permutations. The genes with a adjusted p values below a specific threshold will be considered to have statistically significant changes in expression between the compared classes. The selection of a specific threshold value takes into account the FDR/FWER ratios.

Before developing the predictor, it is necessary to select its constituent variables (genes). This selection is necessary because it can be reasonably assumed that only a subset of the genes under evaluation will be useful for distinguishing among classes. A method very frequently used for this selection is to choose the relevant genes according to the statistical significance of univariate tests (*t* test, *F* test or Wilcoxon's rank test) for differences among classes. Those genes with statistically significant differences are selected for their inclusion in a multivariate model. The threshold employed for testing statistical significance is important, since a more stringent criterion results in a simpler model with fewer variables, but runs the risk of omitting important genes; additionally, the complete procedure generally requires large sample sizes in order to identify enough relevant genes for the construction of an accurate predictor. One common strategy is to perform a screening with a low-stringency significance threshold, estimating the rate of erroneous classification for the resulting models by crossing-over validation. Other alternatives that have been used are multiple hypothesis testing and Principal Component Analysis during the variable selection stage [53, 54].

## Clustering into gene expression profiles: unsupervised methods

Clustering methods, in the context of microarray data, are used for constructing groups of genes or samples with similar expression profiles, using a measurement of distance [51]. The most frequently used distance metrics are the Euclidean distance and Pearson's correlation coefficient. In the case of hierarchical clustering methods, it is also necessary to define a method for estimating the distances between gene clusters [55].

There is no need for an *a priori* definition of the group, class or experimental condition of each sample included in the analysis when using clustering methods; in fact, clustering algorithms can suggest a new clustering plan for the samples based on the similarity between the expression profiles of the genes under study. These methods, when applied to expression data, are useful for identifying clusters of co-expressed genes and distinct patterns of gene expression in the samples without the need for predefined classes that supervise the analysis [56, 57].

The clustering method most frequently used for microarray data is known as hierarchical clustering. This unsupervised methodology derives a series of partitions for the data; in this case, each data point is formed by the expression profile of a sample or gene. There are in turn different variants of hierarchical clustering, such as agglomerative and divisive clustering; the latter is better if the data are to be divided into a few groups of a few elements each. In any case, the end result of these methods is a tree-like structure known as a dendrogram.

There are alternatives to hierarchical clustering methods. One of the most widespread alternatives is the k-Means technique, which has the disadvantage of having to know beforehand the number of groups into which the data will be classified. However, the estimation of k is a known problem that arises whenever it is necessary to map a data structure to a group structure, and it has been intensively studied in the context of gene expression studies [58, 59]. A widely used criterion proposes the selection of k as the lo-

45. LönnStedt I, Speed TP. Replicated microarray data. Statistica Sinica 2002; 12:31-46.

46. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments.Stat Appl Genet Mol Biol 2004;3:Article 3.

47. Westfall PH, Young SS. Re-Sampling Based MultipleTesting New York: Wiley; 1993.

48. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical bayes analysis of a microarray experiment. J Am Stat Assoc 2001;96: 1151-60.

49. Storey J, Tibshirani R. Estimating False Discovery Rates Under Dependence, with Applications to DNA Microarrays Data. Stanford University, Technical Report; 2001:28.

50. Tsai CA, Hsueh HM, Chen JJ. Estimation of false discovery rates in multiple testing: application to gene microarray data. Biometrics 2003;59:1071-81.

51. Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. American Statistical Association 2002;97(457):77-87.

52. Hastie TJ, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2001.

53. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673-9.

west number of groups yielding only small variations in the y-axis of a FOM (Figure of Merit) chart [60], and there are other techniques based on the evaluation of group stability [61].

One of the most common mistakes during the analysis of microarray data is to use clustering analyses to solve problems of class prediction and comparison [62]. These analyses do not yield valid, statistically significant quantitative information on what genes are differentially expressed between classes, and should only be used in this case as an exploratory technique. Class prediction and comparison are better approached with supervised methods, as described in the preceding sections.

## Available software for microarray data analysis

**Bioconductor**: This is an international free software project for the analysis and interpretation of genomic data, written in the R statistical language (http://www.r-project.org) and including a large number of computational algorithms for the analysis of gene expression [63]. The Bioconductor project (http://www.bioconductor.org) has a number of packages that together offer a wide range of statistical applications for several types of genomic analyses: ctc [64] for clustering expression values, multtest [27, 65] and maanova [66] for multiple comparisons of experimental conditions, marray [67], containing *loess/lowess* functions for local regression and samr [68] (an implementation of the SAM method for detecting differential expression) among others.

**MeV** (*The Institute for Genomic Research*, USA): This is one of the most popular programs for microarray data analyses because of the large variety of mathematical methods it includes [69]. This is a free software application that is easy to install, written in Java and showing a user-friendly interface. It has options for performing basic data transformations, filtering, normalization and clustering of genes or experimental conditions using different measurements of distance and methods such as k-Means [70], HCL (Hierarchical Clustering) [55], SOM (Self Organizing Maps) [71, 72] and SOTA (Self Organizing Tree Algorithm) [73, 74]. It also contains the t-test, ANOVA [75] and SAM with FDR control for the discovery of genes differentially expressed between experimental conditions (The latter is one of the discovery tests most frequently used in the literature). MeV also has other features, such as RN (Relevance Networks) [76], which uses a gene entry with its expression profile to draw out its most related gene networks, using a minimal coefficient for node correlation defined by the user. The software is available at http://www.tigr.org/software/tm4/.

**GEPAS** (Centro de Investigación Príncipe Felipe, Valencia): This is a web application for the analysis of gene expression profiles [77-79] implemented as a series of interconnected modules, which performs preprocessing, normalization (using the DNMAD application for cDNA chips or Expresso for Affymetrix systems) [80], the determination of differential expression with T-Rex [78, 81] and clustering algorithms (SOM, Som Tree [82], SOTArray). GEPAS can also be used to construct a class predictor with methods such as SVM (*Support Vector Machine*) [83], DLDA (*Diagonal Lineal Discriminant Analysis*), k NN [51] and others as implemented in the Prophet tool [84]. In order to facilitate the interpretation and extraction of biological information from clusters of related genes, GEPAS also contains FatiGO [85] and Fatigo+ [86, 87], which can be used to associate these genes to GO (Gene Ontology) [88] terms and to the KEGG database [89] of metabolic pathways. The FatiScan application [90] allows for the detection of blocks of functionally related genes (GO, KEGG) in gene lists sorted according to the results of an analysis of differential expression or any other theoretical or experimental criterion (http://gepas.bioinfo.cipf.es/).

**BRB-ArrayTools** (*National Cancer Institute-NCI*, USA): This is a professional, integrated package for the visualization and statistical analysis of gene expression data which is installed as a Microsoft Excel plug-in [91]. It contains almost all the features mentioned above, although it emphasizes topics related to experimental design (http://linus.nci.nih.gov/BRB-ArrayTools.html).

## Functional annotation of the results through data mining techniques

Once the analyses above have been performed (a stage that constitutes only the initial part of the analysis of a microarray experiment) the researcher obtains clusters of related genes, which have to be linked to other sources of public information. By approaching the data as a whole, using a comprehensive analysis for their interpretation that takes into account their statistical significance, the researcher can devise the hypotheses that must be verified experimentally. The most accepted procedure for establishing this link is through gene/protein networks in systems that integrate different sources of biological data, thus being able to observe the behavior of clusters of functionally related genes and their relationships instead of focusing on individual genetic units.

Dopazo J [92] considers that there are currently two separate generations of methods for the analysis and interpretation of the data obtained from high-throughput technologies. The more traditional is known as "threshold-based functional analysis" and is usually implemented as a two-stage process, where first the genes of interest are selected according to a threshold of statistical significance, and then their frequency is determined in biologically relevant terms. The other generation, "threshold-free functional analysis", is currently still under active development and analyzes the behavior of clusters of functionally related genes without previously filtering the results. In order to show the capabilities of this second technological generation, Dopazo used the data from a gene expression study on diabetes mellitus performed by Mootha *et al*. [93], where the application of a *t*-test to compare two groups (17 controls with normal glucose tolerance *vs.* 26 cases, 8 with reduced glucose tolerance and 18 with type 2 diabetes mellitus) did not yield differentially expressed genes using a significance threshold of 0.05. The author used the T-rex program from the GEPAS package and did not find any case of differential expression. However, when the segmentation test implemented in the FatiScan application, as well as other

54. West M, Blanchette C, Dressman H, Huang E, Ishida S, *et al*. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA 2001;98:11462-7.

55. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863-8.

56. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680-6.

57. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, *et al*. Distinct type of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503-11.

58. Belitskaya-Levy I. A generalized clustering problem, with application to DNA microarrays. Stat Appl Genet Mol Biol 2006;5:Article2.

59. Bolshakova N, Azuaje F. Estimating the number of clusters in DNA microarray data. Methods Inf Med 2006;45:153-7.

60. Yeung KY, Haynor DR, Ruzzo WL. Validating clustering for gene expression data. Bioinformatics 2001;17:309-18

61. Smolkin M, Ghosh D. Cluster stability scores for microarray data in cancer studies. BMC Bioinformatics 2003;6:4:36.

62. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. J Natl Cancer Inst 2003;95:14-8.

63. Gentleman RC, Carey VJ, Bates DJ, Bolstad BM, Dettling M, Dudoit S, *et al*. Bioconductor: Open software development for computational biology and bioinformatics. Bioconductor Project Working Papers. Working Paper 1. 2004. Available from: URL: http://www.bepress.com/bioconductor/paper1.

64. Lucas A, Gautier L. Cluster and Tree Conversion. Bioconductor 2006. Available from: URL: http://bioconductor.org/packages/2.2/bioc/html/ctc.html.

65. Pollard KS, Ge Y, Taylor S, Dudoit S. Resampling-based multiple hypothesis testing. Bioconductor's multtest Package. Bioconductor 2005. Available from: http://bioconductor.org/packages/2.2/bioc/html/multtest.html.

66. Wu H, Yang H, Churchill GA. R/MAANOVA: An extensive R environment for the Analysis of Microarray Experiments. Bioconductor 2008. Available from: http://bioconductor.org/packages/2.2/bioc/html/maanova.html.

67. Yang YH. Exploratory analysis for two-color spotted microarray data. Bioconductor 2008. Available from: URL: http://bioconductor.org/packages/2.2/bioc/html/marray.html.

68. Tibshirani R, Chu G, Hastie T. The samr Package. Bioconductor, 2005. Available from: http://www-stat.stanford.edu/~tibs/SAM/Rdist/index.html.

69. Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, *et al*. TM4: a free, open-source system for microarray data management and analysis. Biotechniques 2003;34(2):374-8.

second generation programs such as GSEA [93, 94], PAGE [95] and SAFE [96] were applied, they all identified a common set of genes which, although lacking statistical significance by more traditional criteria, were known to be linked to diabetes. These results evidence the superiority of the approach of Systems Biology over preceding approaches, although the development of, and research on these new methods are far from finished.

### *Verification of the results*

The first part of this review lists the main sources of variability and error for microarray experiments [97]. The simultaneous examination of thousands of genes in each experiment inevitably leads to a higher than usual number of false positives during statistical analysis. Therefore, it is necessary to use other experimental techniques to validate and confirm the presence of a differential gene expression profile among the genes identified with microarrays.

One of the most sensitive experimental techniques to detect and quantify mRNA in tissue samples is the use of quantitative reverse transcription-real time polymerase chain reaction (Q-RT-PCR). This makes it one of the most robust [98, 99] and frequently used methods [100-103] to verify the expression of genes derived from statistical analysis of microarray data. Other analytic techniques, such as Northern blotting and Ribonuclease Protection Assay, have also been used for this purpose [104].

## Applications of the technology in the field of Oncology

### *Examples of class discovery*

- Tamayo *et al*. used microarrays to study gene expression in HL-60 cells and, with the help of SOM [72], obtained biologically relevant gene clusters which were involved in the processes of cellular differentiation.

- Alizadeh *et al*. discovered new lymphoma subtypes with clustering methods [57].

- Bittner *et al*. found a subclassification within melanomas which had not been identified morphologically by other techniques [105]. The subset was obtained by mathematical analysis of gene expression data from a series of samples. Most importantly, the genes whose expression profile was distinctive to the subset were differentially expressed in invasive stage melanomas.

### *Examples of class comparison*

- Prostate cancer has been intensively studied with DNA microarrays. These studies fall into four fundamental groups: those comparing normal to tumoral tissue samples [34, 106-108], those comparing samples from benign prostatic hyperplasia to tumor cells [109], those examining the effects of clinical treatments, *e.g.* samples before and after using Doxazosin [110], and those studying the molecular evolution of prostate cancers refractory to treatment [111, 112], although all these experiments share the common goal of finding genes with differential expression profiles within different experimental conditions and discovering co-expressed genes. If the experiments that have analyzed normal *vs.* tumoral tissue are screened for overlapping results, it can be seen that the expression profile of genes such as CAMKK2, FASN, SIM2, CAV2, LIM and AMACR has behaved similarly in many of them, showing statistically significant differences to the compared groups.

### *Examples of class prediction*

- Breast cancer: Van't Veer *et al*. reported the identification of a set of 70 genes that can be used as predictors for metastasis, obtained from the examination of the gene expression profiles of primary breast cancer tumors from 117 young patients [113]. An initial analysis of the expression of 25 000 human genes showed that there were 5000 genes with statistically significant differences when comparing the tumor samples to a reference sample. These genes were then processed with an unsupervised bidimensional hierarchical clustering algorithm, which clustered the tumors according to the similarity of their expression profiles for these 5000 genes and clustered in turn the genes according to the similarity of their expression profiles in the set of analyzed tumors. As a result, the analysis clustered the tumors in two main groups, one dominated by patients with a negative prognosis for the next 5 years and another composed predominantly of patients with a positive prognosis in the same period, evidencing the predictive power of the analysis of gene expression profiles. Additionally, when the gene clusters were cross-analyzed against the histopathological data, the results matched the results published in the literature. Next, 78 patients with negative lymph nodes were selected to search for prognostic signature in their expression profiles; after five years 44 of these patients were disease-free and the remaining 34 patients had developed metastases. With the aim of classifying patients with either good or bad prognosis a three-stage supervised method was implemented. In a first step, 231 out of the 5000 candidate genes were selected, based on the correlation of their expression profiles with disease progress (correlation coefficient $< -0.3$ or $> 0.3$), then these 231 genes were sorted according to their correlation coefficient and lastly, the predictor was constructed by the sequential addition of sets of the first 5 genes from the top of the sorted list, estimating the classification error by cross-validation for each iteration. The best accuracy (83%) was obtained with a set of 70 genes which were then proposed as the predictor. Upon applying this predictor to the original sample, only 13 patients were classified erroneously, 5 of them from the good prognosis group and 8 from the bad prognosis group. Later studies comparing the survival of a group of 295 patients with the results of the predictor [114] confirmed these results.

- Lung cancer: Chen *et al*. [115] found a molecular signature composed of 5 genes (ERBB3, LCK, DUSP6, STAT1, MMD) associated to survival from NSCLC (Non-Small Cell Lung Cancer). Samples from 125 patients afflicted with this cancer were studied in arrays containing probes for 672 genes which had been shown to be associated to invasiveness in a previous experiment that compared normal *vs*. NSCLC tissue [116]. The genes with a coefficient of variation below 3% were excluded from the analysis, thus selecting 485

70. Soukas A, Cohen P, Socci ND, Friedman JM. Leptin-specific patterns of gene expression in white adipose tissue. Genes Dev 2000;14:963-80.

71. Kohonen T. Self-organizing formation of topologically correct feature maps. Biological Cybernetics 1982;43(1):59-69.

72. 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 USA 1999;96:2907-12.

73. Dopazo J, Carazo JM. Phylogenetic reconstruction using an unsupervised growing neural network that adopts the topology of a phylogenetic tree. J Mol Evol 1997;44:226-33.

74. Herrero J, Valencia A, Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics 2001;17(2): 126-36.

75. Zar JH. Biostatistical analysis. 4th ed., New Jersey: Prentice Hall; 1999, p. 663.

76. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. Proc Natl Acad Sci USA 2000;97:12182-6.

77. Herrero J, Al-Shahrour F, Díaz-Uriarte R, Mateos A, Vaquerizas JM, Santoyo, *et al*. GEPAS: A web-based resource for microarray gene expression data analysis. Nucleic Acids Res 2003;31:3461-7

78. Montaner D, Tárraga J, Huerta-Cepas J, Burguet J, Vaquerizas JM, Conde L, *et al*. Next station in microarray data analysis: GEPAS. Nucleic Acids Res 2006;34: 486-91.

79. Vaquerizas JM, Conde L, Yankilevich P, Cabezon A, Minguez P, Díaz-Uriarte R, *et al*. Gepas an experiment-oriented pipeline for the analysis of microarray gene expression data. Nucleic Acids Res 2005; 33:616-20.

80. Vaquerizas JM, Dopazo J, Díaz-Uriarte R. DNMAD: web-based Diagnosis and Normalization for MicroArray Data. Bioinformatics 2004;20:3656-8.

81. Gautier L, Cope L, Bolstad BM, Irizarry RA. Affy-analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 2004;20:307-15.

82. Herrero J, Dopazo J. Combining hierarchical clustering and self-organizing maps for exploratory analysis of gene expression patterns. J Proteome Res 2002; 1:467-70.

83. Vapnik V. Statistical learning theory. John Wiley and Sons 1999. New York.

84. Medina I, Montaner D, Tarraga J, Dopazo J. Prophet, a web-based tool for class prediction using microarray data. Bioinformatics 2007;23:390-1.

85. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics 2004;20: 578-80.

out of the 672 original genes, and the 125 samples were randomly assigned to a training set or to a test set. In order to find genes associated with death or the recurrence of the disease, the expression values were transformed, assigning codes according to the intensity levels with the purpose of performing a regression analysis. The hazard ratios of a univariate regression showed the associations of the expression level of each gene with survival, with hazard ratios below one associated to 'protective' genes and hazard ratios above one associated to 'risk' genes. A signature of 16 genes significantly correlated with survival was selected; from these, 5 genes predicted the survival of the patients with 96% accuracy. The mean survival time for the 101 patients evaluated during the search for the predictor was 20 months, and the patients classified as high-risk according to the molecular signature had a mean survival time below those with a low-risk signature (20 *vs.* 40 months, $p < 0.001$). This molecular signature was validated with another 60 patients of Chinese origin and 86 western patients from a public repository of microarray data for NSCLC. The presence of a high-risk signature on the tumors was associated with an increased risk of recurrence and lower survival.

## Conclusions

This paper describes and reviews the stages for the quantification and statistical analysis of data from microarray experiments. The coming years will probably witness the development of new algorithms and statistical methods allowing for a better interpretation of the information provided by microarray experiments, in addition to a smoother integration and complementation with other high throughput technologies operating at genomic scales. It is also expected that solutions to some of the known problems of microarray experimentation, such as the large size of the expression matrices (commonly containing thousands of rows -genes- versus hundreds of columns -observations, samples), and the integration with data from microarray experiments performed with differing technologies and controls, will be solved as well. The development of statistical-mathematical methods will no doubt parallel the technological development of the methodology and the future appearance of new high-throughput techniques. A thorough understanding of the technology, therefore, will be a basic requirement in obtaining results with a high scientific impact.


86. Al-Shahrour, F, Minguez P, Vaquerizas, JM, Conde, L, Dopazo, J. Babelomics: a suite of web-tools for functional annotation and analysis of group of genes in high-throughput experiments. Nucleic Acids Res 2005;33:460-4.

87. Al-Shahrour F, Minguez P, Tárraga J, Montaner D, Alloza E, Vaquerizas, *et al*. BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. Nucleic Acids Res 2006;34: 472-6.

88. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. Gene Ontology: tool for the unification of biology. Nat Genet 2000;25:25-9.

89. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res 2004; 32:D277-80.

90. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information. Bioinformatics 2005;21:2988-93.

91. Simon RM, *et al*. Analysis of gene expression data using BRB-Array Tools. Cancer Inform 2006;2:1-7.

92. Dopazo J. Functional interpretation of microarray experiments. OMICS. 2006;10: 398-410.

93. Mootha VK, Lindgren CM, Eriksson KF, Subramanian A, Sihag S, Lehar J, *et al*. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet 2003; 34:267-73.

94. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA 2005; 102:15545-50.

95. Kim SY, Volsky DJ. PAGE: Parametric Analysis of Gene Set Enrichment. BMC Bioinformatics 2005;6:144.

96. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics 2005;21: 1943-9.

97. Miranda J, Bringas R. Análisis de datos de microarreglos de ADN. Parte I: Antecedentes de la tecnología y diseño experimental. Biotecnol Apl (en prensa).

98. Su LJ, Chang CW, Wu YC, Chen KC, Lin CJ, Liang SC, *et al*. Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap resampling scheme. BMC Genomics 2007; 8:140.

99. Dallas PB, Gottardo NG, Firth MJ, Beesley AH, Hoffmann K, Terry PA, *et al*. Gene expression levels assessed by oligonucleotide microarray analysis and quantitative real-time RT-PCR-how well do they correlate? BMC Genomics 2005;6(1):59.

100. Bektic J, Wrulich OA, Dobler G, Kofler K, Ueberall F, Culig Z, *et al*. Identification of genes involved in estrogenic action in the human prostate using microarray analysis. Genomics. 2004;83(1):34-44.

101. Vawter MP, Ferran E, Galke B, Cooper K, Bunney WE, Byerley W. Microarray screening of lymphocyte gene expression differences in a multiplex schizophrenia pedigree. Schizophr Res 2004;67(1):41-52.

102. Wiese AH, Auer J, Lassmann S, Nährig J, Rosenberg R, Höfler H, *et al*. Identification of gene signatures for invasive colorectal tumor cells. Cancer Detect Prev 2007;31(4): 282-95.

103. Jura J, Wêgrzyn P, Korostyñski M, Guzik K, Oczko-Wojciechowska M, Jarzab M, *et al*. Identification of interleukin-1 and interleukin-6-responsive genes in human monocyte-derived macrophages using microarrays. 2008 (en prensa).

104. Kothapalli R, Yoder SJ, Mane S, Loughran TP. Microarray results: how accurate are they?. BMC Bioinformatics 2002;3:22.

105. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, *et al*. Molecular classification of cutaneous malignant melanoma by gene ex-pression profiling. Nature 2000; 406:536-40.

106. Ashida S, Nakagawa H, Katagiri T, Furihata M, Iiizumi M, Anazawa Y, *et al*. Molecular Features of the Transition from Prostatic Intraepithelial Neoplasia (PIN) to Prostate Cancer: Genome-wide Gene-expression Profiles of Prostate Cancers and PINs. Cancer Res 2004;64:5963-72.

107. Zhao H, Ramos CF, Brooks JD, Peehl DM. Distinctive gene expression of prostatic stromal cells cultured from diseased versus normal tissues. J Cell Physiol 2007;210:111-21.

108. Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodríguez J, Moskaluk CA, *et al*. Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer. Cancer Res 2001;61:5974-8.

109. Luo J, Duggan DJ, Chen Y, Sauvageot J, Ewing CM, Bittner ML, *et al*. Human Prostate Cancer and Benign Prostatic Hyperplasia: Molecular Dissection by Gene Expression Profiling. Cancer Res 2001;61:4683-8.

110. Zhao H, Lai F, Nonn L, Brooks JD, Peehl DM. Molecular Targets of Doxazosin in Human Prostatic Stromal Cells. Prostate 2005; 62:400-10.

111. Stanbrough M, Bubley GJ, Ross K, Golub TR, Rubin MA, Penning TM, *et al*. Increased expression of genes converting adrenal androgens to testosterone in androgen-independent prostate cancer. Cancer Res 2006; 66:2815-25.

112. Tamura K, Furihata M, Tsunoda T, Ashida S, Takata R, Obara W, *et al*. Molecular features of hormone-refractory prostate cancer cells by genome-wide gene expression profiles. Cancer Res 2007;67:5117-25.

113. Van't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AA, Mao M, *et al*. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6.

114. Van de Vijver MJ, He YD, Van't Veer LJ, Dai H, Hart AA, Voskuil DW, *et al*. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002; 347:1999-2009.

115. Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, *et al*. A Five-Gene Signature and Clinical Outcome in Non-Small-Cell Lung Cancer. N Engl J Med 2007;356: 11-20.

116. Chen JJ, Peck K, Hong TM, Yang SC, Sher YP, Shih JY, *et al*. Global Analysis of Gene Expression in invasion by a lung cancer model. Cancer Res 2001;61:5223-30.

*Received in June, 2008. Accepted for publication in December, 2008.*