Analysis of DNA microarray data. Part I: Technological background and experimental design

REVIEW

Jamilet Miranda, ✉ Ricardo Bringas

Centro de Ingeniería Genética y Biotecnología, CIGB
Ave. 31 e/ 158 y 190, Cubanacán, Playa, AP 6162, CP 10600, Ciudad de La Habana, Cuba
E-mail: ricardo.bringas@cigb.edu.cu

ABSTRACT

DNA microarrays have emerged as the most widely used technology for the massive quantification of gene expression and have been applied to a very diverge range of topics in molecular biology research over the last several years. One key element for a successful application of this technology is a thorough understanding of the steps to be followed in order to obtain and analyze expression data. In the present article we review the origins of the technology, its evolution and some of its more common applications, highlighting the importance of a clear definition of the objectives for the design of the experiment, the different sources of variability to be considered and the most common experimental setups.

Keywords: DNA microarrays, gene expression, experimental design

Biotecnología Aplicada 2008;25:90-96

RESUMEN

Análisis de datos de microarreglos de ADN. Parte I: Antecedentes de la tecnología y diseño experimental. Los microarreglos de ADN han emergido como la tecnología más utilizada para la cuantificación masiva de la expresión de genes y han sido aplicados a temas muy diversos entre las investigaciones biológicas en los últimos años. Un elemento fundamental para la aplicación exitosa de esta tecnología es el conocimiento de los pasos a seguir para la obtención y análisis de los datos de expresión. En el presente trabajo se hace un recuento del surgimiento de la tecnología, su evolución y algunas de sus aplicaciones más comunes, se subraya la necesidad de definir claramente en el diseño, los objetivos del experimento y se exponen las diferentes fuentes de variabilidad a tener en cuenta en el diseño y los tipos de diseño más comunes.

Palabras clave: microarreglos de ADN, expresión de genes, diseño experimental

*I*ntroduction

The advent of whole genome sequencing has provided access to the primary structure of the entire complement of genes of an organism, as well as their regulatory elements. The differences in this primary information between individuals of a population ultimately determine their differential interaction with the environment, and thus their study constitutes a major avenue of research for the solution of a number of agricultural and human health-related problems. This wealth of data, in turn, has stimulated the development in recent years of technologies for the genome-wide simultaneous analysis, in a single experiment, of all sequence elements related to a biological phenotype. One such technology is the DNA microarrays, which are simply collections of DNA fragments of known sequence bound to a solid support that can be used for the quantification of the levels of specific RNA or DNA molecules in biological samples. These fragments, or probes, are designed to be complementary to the target DNA or RNA species, and therefore allow for the quantitation of the target(s) in the sample by measuring its hybridization to the complementary probes printed onto the array.

DNA microarrays have turned not only into the most widely used tool for the genome-wide generation of gene expression profiles, but their use has been extended to the study of inter-individual genetic variation through the use of arrays specifically geared for the detection of SNPs (*Single Nucleotide Polymorphisms*) [1]. Additionally, their use in combination with methodologies such as chromatin immunoprecipitation has allowed the identification of regulatory sequences recognized by transcription factors, using arrays with probes complementary to the promoter regions of all the known genes of the organism under study [2]. Yet another problem successfully tackled with DNA microarrays has been the study of differential splicing, thanks to the design of arrays containing specific probes for each exon of the gene(s) under scrutiny [3].

The application of DNA microarrays for the generation of gene expression profiles from samples obtained under different experimental conditions has allowed the dissection of the molecular basis and mechanisms underlying a number of human diseases, such as viral infections [4, 5], schizophrenia [6-10], prostate cancer [11-15] and breast cancer [16, 17]. DNA microarrays have also been at the center of pharmacogenomic experiments addressing the changes induced at the molecular level by drugs specific for different disorders, such as those studying the emergence of tamoxifen resistance in breast cancer cells [18] and the effects of IL2 for cancer treatment [19-22].

Microarrays have also proven to be advantageous over more traditional techniques for the diagnosis of complex disorders. For instance, Alizadeh *et al.* [23] identified two B cell lymphoma subtypes that were hardly distinguishable by histological tests alone, characterized by a differential expression profile for a defined set of genes, which clearly correlated with survival.

✉ Corresponding author

1. Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, *et al.* Accessing genetic information with high-density DNA arrays. Science 1996;274:610-4.

2. Lee TI, Rinaldi MJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science 2002;298: 799-804.

3. Castle J, Garrett-Engele P, Armour CD, Duenwald SJ, Loerch PM, Meyer MR, *et al.* Optimization of oligonucleotide arrays and RNA amplification protocols for analysis of transcript structure and alternative splicing. Genome Biol 2003;4:R66.

4. Wang X, Yuan ZH, Zheng LJ, Yu F, Xiong W, Liu JX, *et al.* Gene expression profiles in an hepatitis B virus transfected hepatoblastoma cell line and differentially regulated gene expression by interferon-α. World J Gastroenterol 2004;10:1740-5.

5. Yang J, Bo XC, Yao J, Yang NM, Wang SQ. Differentially expressed cellular genes following HBV: potential targets of anti-HBV drugs?. J Viral Hepat 2005;12:357-63.

6. Hakak Y, Walker JR, Li C, Wong WH, Davis KL, Buxbaum JD, *et al.* Genome-wide expression analysis reveals dysre-gulation of myelination-related genes in chronic schizophrenia. Proc Natl Acad Sci USA 2001;98:4746-51.

7. Vawter MP, Barrett T, Cheadle C, Sokolov BP, Wood WH, Donovan DM, *et al.* Application of cDNA microarrays to examine gene expression differences in schizophrenia. Brain Res Bull 2001;55: 641-50.

The technological developments associated with the arrival of DNA microarrays have inevitably stimulated the evolution, adaptation and creation of statistical and mathematical methodologies for the analysis of the arrays of gene expression values they generate, where the number of variables or genes (g) is typically much larger than the number of samples or tissues (n) under analysis (that is, $n << g$). The specific statistic tests to be used depend to a great extent on the objectives of the experiment, often requiring a combination of different statistical methods for data analysis due to the characteristically large complexity of biological systems. The comparison, prediction and discovery of experimental classes [24-26] constitute the most common goals for these experiments.

The present work reviews the current status of the application of DNA microarrays for the study of genome-wide gene expression profiles, placing a special emphasis on its technological antecedents and experimental design requirements.

Technological background and antecedents

Origins

The birth of DNA microarrays can be traced back in the literature to the publications in the mid-nineties by *Schena et al.* [27] and *Lockhart et al.* [28].

Schena et al. [27] were the first to describe the development of a microarray for monitoring in parallel the expression of multiple genes. Probes from a 96-well plate were printed on microscopy slides in an area of 3.5 x 5.5 mm. Once deposited on the glass, the probes were chemically and thermally treated in order to denature the DNA and fix it to the surface. The expression of a total of 45 *Arabidopsis thaliana* genes (plus 3 control genes from other organisms) was evaluated, duplicating each probe in adjacent wells in order to study the reproducibility of the printing and hybridization processes, and using fluorescently labeled cDNA reverse-transcribed from total *A. thaliana* RNA as a sample (Actually, two samples labeled with different fluorophores were analyzed simultaneously). The experiment yielded a total of 27 genes differentially expressed between samples from leaf or root tissues, and most importantly, pioneered some technological innovations that have later become staples of the methodology, such as the use of cDNA microarrays and the simultaneous analysis of two samples in a single experiment by means of double fluorescence labeling.

Lockhart *et al.* [28], on the other hand, developed techniques for the parallel measurement of the expression levels of thousands of genes. Their methodology was based on the quantification of the relative abundance of mRNA by hybridizing whole mRNA populations to high-density arrays of DNA probes. These arrays contained thousands of 20-mer oligonucleotides designed to be complementary to the transcribed 3' regions of known human genes, and were obtained by parallel *in situ* synthesis on the glass surface through a combination of methods borrowed from the arsenal of nucleic acid chemistry and the photolithographic procedures used in the microelectronics industry [29, 30]. Since an area of only 50 x 50 µm was required for the synthesis of each oligonucleotide, a total of more than 65 000 different probes could be squeezed into an area of 1.6 cm^2. The specificity of the hybridization was controlled by synthesizing the probes in pairs, where one probe was a perfectly matching oligonucleotide designated as PM (*Perfect Match*) and the other contained a centrally positioned single mismatch relative to the target (designated as MM, or *MisMatch*), thus affording an internal control for non-specific hybridization. Therefore, the PM-MM signal ratio, rather than the absolute intensity of the PM signal, was used for further processing. The simplest data analysis algorithms used for the estimation of the expression levels of the target gene in this experimental setup usually average the PM-MM differences after a background correction of each PM/MM pair, although other methods can be used for this estimation [31-33].

The microarray images were obtained, in the case of Lockhart *et al.* [28], with a confocal scanning microscope specially designed for this purpose. These images had a resolution of 7.5 µm, which is equivalent to an average of 45 luminance values in the 50 x 50 µm area corresponding to each printed probe that were further combined to yield a single value per probe. Obviously, the accuracy of this value increases with the number of luminance values per probe, which depends directly on the area occupied by the probe (the cell) and the resolution of the scanning device. Figure 1 illustrates the influence of the cell area/scanning resolution ratio on the number of luminance values or pixels obtained per probe: A 10:1 ratio (Figure 1a) generates an average of 100 pixels/probe, whereas a 5:1 ratio only generates 25 pixels/probe. The importance of this parameter is further reinforced by the fact that the pixels that most closely correlate with the intensity of the hybridization are those near the center of the square occupied by the fluorescent spot. As also illustrated by figure 1, the proportion of these pixels also increases with the cell area/scanning resolution ratio.

In their initial experiment, Lockhart *et al.* [28] designed hundreds of probe pairs for each gene to be evaluated, with the aim of estimating their sensitivity and specificity upon hybridization with a complex

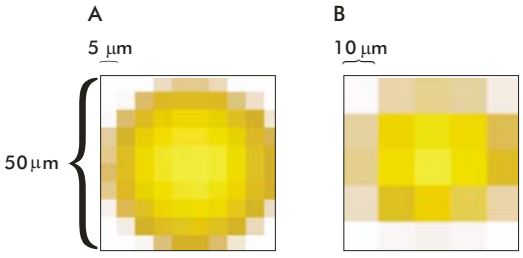

Figure 1. Illustration of signal reading from a microarray at different laser scanning resolutions. A signal with an approximate diameter of 50 μm is represented in both cases, with a scanning resolution of 5 μm in a) and 10 μm in b). For both examples the intensity readings from the peripheral areas of the spot depend mainly not on signal intensity *per se*, but on the area occupied by the signal within the section being scanned by the reading device. Consequently, the most accurate intensity readings correspond to the central pixels within the fluorescent spot, evidencing how a higher resolution results on much higher accuracies for intensity measurements.

8. 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;67:41-52.

9. Tsuang MT, Nossova N, Yager T, Tsuang MM, Guo SC, Shyu KG, *et al.* Assessing the validity of blood-based gene expression profiles for the classification of schizophrenia and bipolar disorder: A preliminary report. Am J Med Genet B Neuropsychiatr Genet 2005;133B:1-5.

10. Glatt SJ, Everall IP, Kremen WS, Corbeil J, Sasik R, Khanlou N, *et al.* Comparative gene expression analysis of blood and brain provides concurrent validation of SELENBP1 up-regulation in schizophrenia. Proc Natl Acad Sci USA 2005;102: 15533-8.

11. 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.

12. Welsh JB, Sapinoso LM, Su AI, Kern SG, Wang-Rodriguez J, Moskaluk CA, *et al.* Analysis of gene expression identifies candidate markers and pharmacological targets in prostate cancer. Cancer Res 2001;61:5974-8.

13. 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.

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

15. Zhao H, Lai F, Nonn L, Brooks JD, Peehl DM. Molecular targets of doxazosin in human prostatic stromal cells. Prostate 2005;62:400-10.

16. 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.

sample of cellular RNA. This experiment allowed the derivation of a set of rules for probe selection. A second experiment, which analyzed the expression of 118 genes, involved the design of an average of 300 PM-MM pairs per gene, which were selected from the 3' untranslated regions of the target mRNAs. Ten sets of 20 PM-MM pairs each were then randomly selected from the totality of the probes designed for each gene, and the changes in hybridization patterns for each set were compared with those for the full probe set of the target gene. The results revealed that a single set of 20 probes was sufficient for measuring changes in expression levels even for low-abundance mRNAs. These experiments set the foundation for the development of the technology used by Affymetrix [34], which remains the market leader in DNA microarrays. The technology has evolved through the years, with a steady increase in chip density (currently 5 microns) and scanning resolution.

The use of microarrays at a genomic scale

DeRisi *et al.* [35] used a microarray containing approximately 6 400 different DNA sequences corresponding to virtually every open reading frame identified in the genome of *Saccharomyces cerevisiae* to study gene expression profiles during the metabolic shift from fermentation to respiration. The experiment included an initial stage of anaerobic fermentation using glucose as carbon source, followed by a gradual switch to aerobic growth on ethanol as the glucose in the medium was depleted; taking samples for the study of gene expression every 2 hours along the process.

The results revealed a stable pattern of gene transcription during the anaerobic stage with only isolated cases of differential expression from sample to sample; however, the switch to aerobic growth as glucose concentration decreased resulted in a gradual increase in the number of differentially expressed genes. At the end, the anaerobic-aerobic switch resulted in a two-fold or higher increase in mRNA levels for 710 genes, and a two-fold or higher decrease in another 1 030 genes. Half of the differentially transcribed genes had no known function; on the other hand, there was a coordinated induction for cytochrome-C-related genes and those involved with the TCA and glyoxylate cycles as glucose was exhausted, together with a coordinated reduction in expression levels for genes involved in protein synthesis. This work showed that functionally related genes can be clustered based on similarities between their expression profiles alone, and that even within these clusters it is possible to infer a common regulatory pathway by the identification of regulatory sequences in their promoter regions. The results also evidenced the value of clustering methods for the analysis of microarray data, underlined the agreement between similar transcriptional profiles and the presence of upstream regulatory elements and confirmed the influence of regulatory genes on the expression levels of their target.

Experimental design

The sections below deal with topics concerning the definition of the biological hypothesis to be tested and experimental design. A diagram depicting the steps to be followed during the design and analysis of microarray experiments can be seen in figure 2.

Defining the biological hypothesis

As is usually the case in scientific research, microarray experiments require an *a priori* definition of the questions they are designed to answer. This question may be *e.g.* what genes are differentially expressed under 2 or more experimental conditions, or if it is possible to cluster different samples based on similarities or discrepancies between their gene expression profiles. In any case, a clearly defined experimental hypothesis is necessary in order to identify the type and number of samples per experimental condition to be used, as well as for establishing a strategy for data analysis. Next, an unprejudiced analysis on whether DNA microarrays are the right technology for obtaining the desired answer should be performed. The use of DNA microarrays is justified when performing genome-wide experiments without a deep knowledge of the behavior of individual genes, where it is desirable to identify a cellular or metabolic process linked to the scientific hypothesis under scrutiny. The results obtained from the experiment depend to a large extent on the available facilities for data analysis and interpretation, on which the use of Systems Biology approaches plays a major role.

Design

Choosing the optimal design during experimental planning for a microarray assay depends on the evaluation of different factors that have to be chosen depending on the scientific questions to be answered and on available resources. As a general rule, microarray experiments analyze a large number of variables (thousands of genes) under a small number of experi-

17. 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.

18. Jansen MP, Foekens JA, van Staveren IL, Dirkzwager-Kiel MM, Ritstier K, Look MP, *et al.* Molecular classification of tamoxifen-resistant breast carcinomas by gene expression profiling. J Clin Oncol 2005;23:732-40.

19. Diehn M, Alizadeh AA, Rando OJ, Liu CL, Stankunas K, Botstein D, *et al.* Genomic expression programs and the integration of the CD28 costimulatory signal in T cell activation. Proc Natl Acad Sci USA 2002; 99:11796-801.

20. Mao M, Biery MC, Kobayashi SV, Ward T, Schimmack G, Burchard J, *et al.* T lymphocyte activation gene identification by coregulated expression on DNA microarrays. Genomics 2004;83:989-99.

21. Panelli MC, White R, Foster M, Martin B, Wang E, Smith K, *et al.* Forecasting the cytokine storm following systemic interleukin (IL)-2 administration. J Transl Med 2004; 2:17.

22. Kovanen PE, Young L, Al-Shami A, Rovella V, Pise-Masison CA, Radonovich MF, *et al.* Global analysis of IL-2 target genes: identification of chromosomal clusters of expressed genes. Int Immunol 2005;17:1009-21.

23. 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.

24. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006;7:55-65.

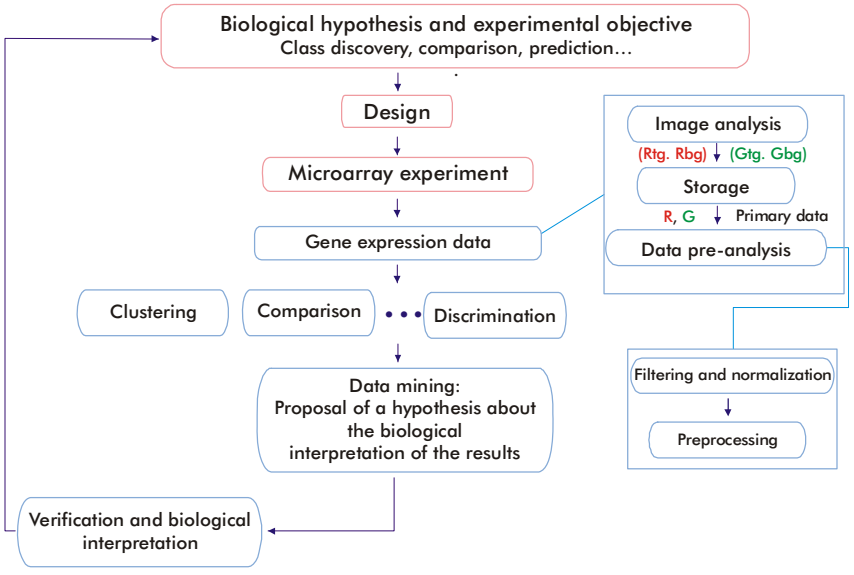

Figure 2. General workflow for a microarray experiment and the analysis of the resulting data. The initial steps requiring the definition of a biological question, the implementation of an experimental design and actually performing the experiment are represented in red; the data analysis phase, starting with image acquisition and concluding with the interpretation of the results and a proposal for a biological verification, is represented in blue.

mental conditions (dozens or hundreds of samples); therefore requiring a highly optimized design in order to maximize the chances of obtaining a valid result. To compound matters, microarray data are subjected to significant sources of variation, included those inherent to the system under analysis (*e.g.* inter-individual variation) and those resulting from the multiple steps necessary to use the technique. Choosing a specific microarray technology constitutes in itself another challenge, since different technological implementations of DNA microarrays usually lend themselves better to different experimental designs, and yet it is often necessary to factor a cost/benefit analysis into the equation. Therefore, proper planning and experimental design are essential for using this technology.

Defining the objective of the experiment based on the biological hypothesis

The biological hypothesis to be tested is the basis for the definition of the experimental objectives. In the case of microarray experimentation, most hypotheses can be tested by one of the following three experimental goals, presented below in association with common biological questions.

Question 1: Which genes are differentially expressed between two or more experimental conditions?

Objective: Class comparison

These experiments often compare affected *vs.* healthy tissue samples, or samples of cells treated or not treated with a specific drug, or mutant *vs.* wild-type organisms. Answering this type of questions requires a class comparison as the experimental objective, defined as a comparison of expression profiles among different samples. The classes to be compared must be defined beforehand, using no information about their expression profiles. An example of this type of assays is the experiment of Lapointe *et al.* [36], which studied healthy and tumor samples from a prostate cancer patient. Using classes defined according to known clinical parameters such as tumor grade, stage and recurrence of the disease, the authors obtained sets of genes with significant differences in expression levels between classes.

The present work shows, in figure 3, the results of an analysis of the data from this experiment, obtained from the Stanford Microarray Database (SMD; http://smd.stanford.edu/). Three classes were defined *a priori*: healthy, tumoral and metastatic. After defining the classes, the data from tumors *vs.* healthy tissue were compared, identifying a set of differentially expressed genes to which clustering methods were applied (Figure 3). Then, pairwise class comparisons were used to identify genes whose expression gradually decreases with the progression from healthy to tumor to metastatic tissue. This is exemplified by the SYNPO2 gene, which has been previously reported to be repressed in advanced prostate cancer [37] and has recently been used as a predictor for metastasis in prostate cancer [38].

Question 2: Based on the previous knowledge of the expression profile of a set of genes for different types of samples, can a new sample be classified as belonging to a specific type?

Objective: Class prediction

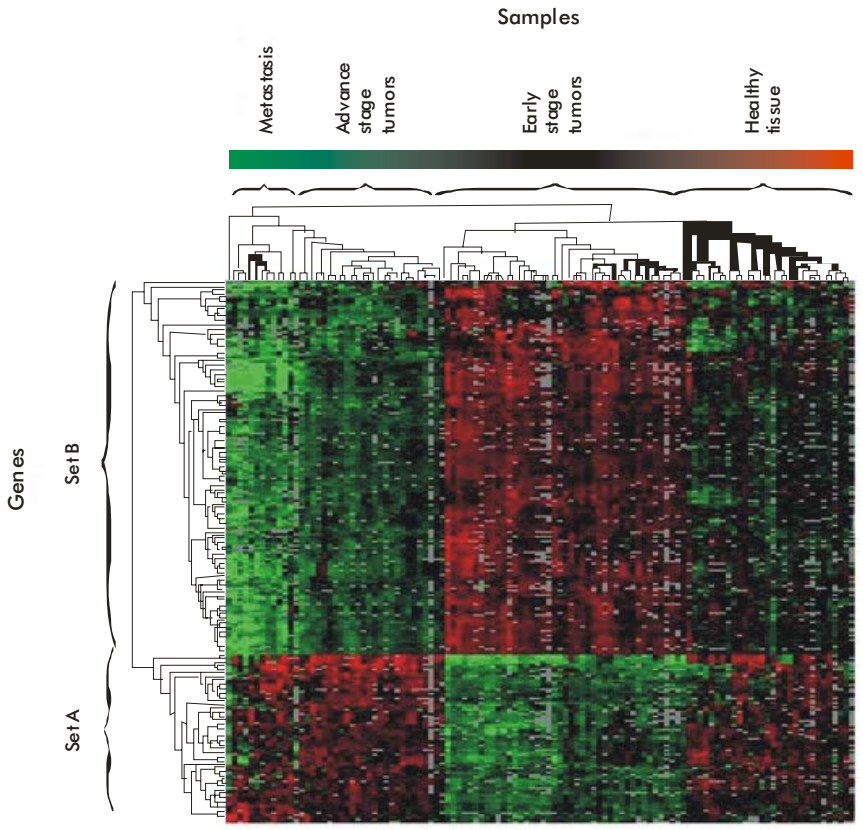

Figure 3. Two-dimensional clustering of 242 genes with differential expression when comparing healthy and prostate cancer tissues. The data were obtained from the study of Lapointe *et al* [35]. This exploratory analysis shows how the genes from set A are repressed in early tumor stages and then increase their expression as the disease progresses. The genes from set B display the opposite behavior.

The question associated to a class prediction objective usually tries to find a multivariate function based on gene expression that allows the classification, with a specified level of accuracy, of a new sample or tissue as belonging to one of several predefined groups, depending on the expression levels of a number of key genes. In other words, these experiments as a general rule try to identify a molecular signature or predictor represented by a set of genes whose expression profiles allow discriminating, with a high certainty, whether a sample belongs to a certain group. Often, the ulterior motive of this type of research is the development of cheaper, single-purpose DNA microarrays including only probes for the genes belonging to the predictor, which can be used as powerful tools for the diagnosis and prognosis of complex diseases. One of the most eloquent examples is their utilization for predicting metastasis from samples of primary breast tumors [39]. A predictor can also be applied for clinical decision making, such as the selection of a treatment or the definition of risk groups.

Question 3: Can the study of expression profiles in the samples define new subtypes that can be associated to other sample characteristics?

Objective: Class discovery

The type of question associated to a class discovery objective consists on the identification of new subtypes within the population under study. The main

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

26. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-7.

27. Schena M, Shalon, D, Davis RW, Brown PO. Qunatitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270: 467-70.

28. Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, *et al*. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675-80.

29. Fodor SP, Rava RP, Huang XC, Pease AC, Holmes CP, Adams CL. Multiplexed biochemical assays with biological chips. Nature 1993;364:555-6.

30. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SP. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA 1994;91:5022-6.

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

difference between this and the other two objectives described above is the absence of predefined classes. A good example of this type of experiment is provided by Alizadeh *et al.* [23], who studied different samples of diffuse large B-cell lymphomas (DLBCL) and managed to identify 2 DLBCL groups or subclasses based on the differential expression of hundreds of different genes. These subclasses, in turn, were associated with very different clinical manifestations, which suggested the classification of these subgroups of DLBCL as different clinical entities.

Selection of the best experimental platform

Before defining the type of experimental design to be used, it is advisable to choose a specific technological platform or microarray, since this decision will determine the maximum number of variables (genes) to be studied. Obviously, this number directly influences the experimental design and, particularly, the definition of the number of samples to be included.

The use of genomic coverage microarrays -that is, those comprising all or most genes from a genome- is warranted only if the hypothesis to be tested requires the massive analysis of all genes from the organism under study. If, on the contrary, testing the hypothesis requires the study of a limited number of genes, other techniques such as RT-PCR (Reverse Transcription/ Polymerase Chain Reaction) may constitute cheaper and more accurate alternatives.

A useful middle ground is often the use of arrays including a smaller number of genes, all related to the biological problem being researched. Such arrays are already marketed by several manufacturers, and include gene sets which have been previously determined to be involved in a specific disease, metabolic pathway or other biological function. Although these arrays measure a smaller number of variables and are therefore much cheaper, they have the important disadvantage of containing a set of probes already constrained to be relevant for the problem at hand, thus limiting the possibilities for the obtaining of original results. Another possibility is the design, in house, of a purpose-made array that only includes relevant genes, as defined by other research tools available to the researcher. However, this topic will not be reviewed on this work.

Brief description of the most used technologies

In spite of the availability of several technological alternatives for the manufacture of microarrays, those based on cDNA arrays [27] and the Affymetrix platform [28, 34] have remained as the most popular within the research community.

cDNA. The cDNA probes for each gene are robotically printed as a two-dimensional array on the surface of a solid support, with glass remaining the most common choice for this purpose. Two RNA samples are simultaneously tested, each labeled with a different fluorophore (usually Cy5 and Cy3). Once the signals are read, their intensities are interpreted as a direct indication of the expression level of the hybridizing mRNA. The 2 samples may correspond to different experimental conditions to be compared for relative expression, or many samples to be analyzed can be paired each to the same control sample, in order to later estimate their relative expression levels by determining their differences with the control. There are also more conventional cDNA microarray technologies in which the probes are fixed to nylon-based membranes and the sample is radioactively labeled.

Affymetrix. As described above, a representative set of probes is designed for each gene, printing a PM-MM pair per probe. In this case, the printing process uses a silicone chip [40]. Each sample is labeled and individually hybridized to the array. After washing, the expression level of each gene is estimated with an algorithm that analyzes the intensities of each set of gene-specific PM-MM pairs. The reading wavelengths used for this technology are smaller than those used for cDNA arrays. One of the most recent products from Affymetrix, the *Human Gene 1.0 ST Array*, can detect approximately 29 000 human genes, each represented by 26 different probes distributed along the complete sequence of the target gene; therefore corresponding to a total of more than 750 000 different probes. Each probe is printed into a 5 x 5 μm spot (http://www.affymetrix.com/).

Choosing a specific platform must take into account their advantages and disadvantages. The *Affymetrix* platform is more reliable, but also more expensive than cDNA-based arrays, which also have the advantage of further cost reductions if two samples are simultaneously analyzed via double labeling, reducing the number of arrays according to the design being pursued. The results obtained with *Affymetrix* technology are usually more accurate and reproducible. The probes in these arrays are more homogeneous than those of cDNA microarrays, and inter-array variability is decreased by minimizing the effects produced by an uneven distribution of the sample over the array during hybridization. These facts have led many laboratories to perform preliminary screening experiments with *Affymetrix* technology using a genomic coverage chip, followed by the selection of a few hundred candidate genes based on these results, which are later analyzed with less accuracy, but cheaper technologies such as cDNA arrays, that ultimately afford the researcher the possibility of using larger numbers of biological replicates.

Selecting a specific experimental design

Once the biological hypothesis, the experimental objective, and the technological platform have been defined, it is possible to choose the best experimental design for the assay. As detailed above, there is a significant difference between using *Affymetrix* microarrays or cDNA arrays with double labeling for the samples. In the first case, each sample is labeled and hybridized independently, whereas in the latter the use of dual labeling techniques allows the simultaneous hybridization of two samples that may correspond to different experimental treatments or to a sample and a control.

It should be noted that there are differences in labeling efficiency for both fluorophores in cDNA array-based experiments (that is, the same sample evaluated with two different labels yields different fluorescence intensities) [41, 42]. Eliminating this "dye bias" requires the measurement of experimental replicates with reversed labeling. However, Dobbin *et al.* [41, 43] argue that a reversed labeling experimental replicate is

32. Li C, Wong WH. Model based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proc Natl Acad Sci USA 2001;98:31-6.

33. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4:249-64.

34. Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. High density synthetic oligonucleotide arrays. Nat Genet 1999;21 Suppl 1:20-4.

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

36. Lapointe J, Li C, Higgins JP, van de Rijn M, Bair E, Montgomery K, *et al.* Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci USA 2004;101:811-6.

37. Lin F, Yu YP, Woods J, Cieply K, Gooding B, Finkelstein P, *et al.* Myopodin, a synaptopodin homologue, is frequently deleted in invasive prostate cancers. Am J Pathol 2001;159:1603-12.

38. Yu YP, Tseng GC, Luo JH. Inactivation of myopodin expression associated with prostate can cer relapse. Urology 2006; 68:578-82.

39. 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.

40. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251:767-73.

41. Dobbin K, Shih JH, Simon R. Statistical design of reverse dye microarrays. Bioinformatics 2003;19:803-10.

42. Rosenzweig BA, Pine PS, Domon OE, Morris SM, Chen JJ, Sistare FD. Dye bias correction in dual-labeled cDNA microarray gene expression measurements. Environ Health Perspect 2004;112:480-7.

not necessary for each pair of samples A and B, but rather can be performed using biological replicates of A and B, thus increasing the throughput of the assay. The reverse labeling results can be used for normalization, eliminating the average labeling bias (although some dye bias may remain for specific genes).

Reference design. This type of design is based on using a -preferably universal- reference sample that is hybridized to the array under the same conditions as the experimental sample (Figure 4a). In order to facilitate inter-experimental comparisons, it is recommended that all laboratories use the same reference sample for all the assays [44, 45]. Although a reference design facilitates the analysis and comparison between samples tested at very different times or in different laboratories, it has the disadvantage of reducing the throughput (since half the hybridizations are performed with the reference sample), thus increasing the costs.

Balanced block design. This design alternative, proposed by Dobbin and Simon [46], consists of the hybridization of a different sample from each group to each array, alternating the sample dye assignment order according to their group (Figure 4b). This design is suited for simple experimental settings where only two types of samples are under comparison and is characterized by a very efficient use of the available microarrays, as each sample pair consumes only a single array. It is not without drawbacks, however, since it does not lend itself to the use of clustering methods or to comparisons of expression profiles between different arrays and experimental groups.

Loop design. In this design, proposed by Kerr and Churchill [47], the sample pairs to be compared are distributed in such a way that each sample is hybridized to two different arrays, using on each case a different fluorophore (Figure 4c). This design is not used often, since it requires twice the number of microarrays as the balanced block design, is not well suited for clustering algorithms, and has more complex demands for its analysis than the reference design.

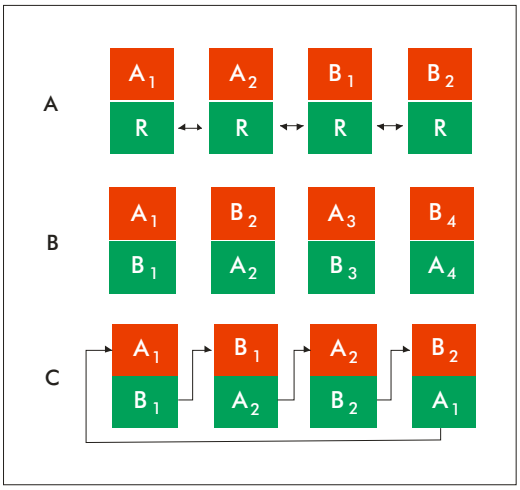

Figure 4. Main types of microarray experimental design. a) Reference design; b) Balanced block design; c) Loop design. Ax, Bx represent two different sets of experimental samples and R represents a reference sample. For case a), every chip always combines an experimental with a reference sample, whereas cases b) and c) represent alternative design proposals where the experimental samples can simultaneously be used as a reference.

Sources of variability to be accounted by the design

Some of the sources of variability to be taken into account during the design of microarray experiments have been mentioned above. The following listing contains the most important causes of experimental variability:

- The biological heterogeneity of the population and samples under study
- The process for obtaining and manipulating the samples
- The extraction of RNA and its enzymatic amplification (if performed)
- Sample labeling (Labeling efficiency, physical properties of the fluorophore)
- Hybridization and reading, depending on the PMT voltage and laser power

Small variations caused by any of the factors listed above may lead to significant changes in the measured expression levels and, therefore, to erroneous experimental conclusions. However, the influence of these biases can be minimized with the selection of proper controls, a number of replicates adequate to the expected levels of variability, and through statistical normalization [48].

Another potential source of variability is the possible contamination of the tissue sample from which the RNA samples are purified. This problem is not restricted to DNA microarray experiments, and can be addressed through techniques such as LMM (Laser Microbeam Microdissection) [13], which strive for a higher accuracy during tissue selection.

In any case, there is a very tight interdependence between the type of design, the experimental objective and the statistical method for analyzing the generated experimental data. The table shows a proposal for a design type depending on the goals of the experiment.

Sample selection

The principle of sample homogeneity remains a cornerstone of sample selection procedures for microarray experiments [49]. The requirements of homogeneity can be fulfilled by using a population of controls obtained through random sampling from the same

43. Dobbin K, Shih JH, Simon R. Questions and answers on design of dual-label microarrays for identifying differentially expressed genes. J Natl Cancer Inst 2003; 95:1362-9.

44. Novoradovskaya N, Whitfield ML, Basehore LS, Novoradovsky A, Pesich R, Usary J, *et al*. Universal Reference RNA as a standard for microarray experiments. BMC Genomics 2004;5:20.

45. Khan RL, Gonye GE, Gao G, Schwaber JS. A universal reference sample derived from clone vector for improved detection of differential gene expression. BMC Genomics 2006;7:109.

46. Dobbin K, Simon R. Comparison of microarrays designs for class comparison and class discovery. Bioinformatics 2002; 18:1438-45.

47. Kerr MK, Churchill GA. Statistical design and analysis of gene expression microarray data. Genet Res 2001;77:123-8.

48. Fan J, Ren Y. Statistical analysis of DNA microarray data in cancer research. Clin Cancer Res 2006;12:4469-73.

Table. Types of experimental design recommended according to the experimental objective

Objective/ Type of design	Reference*	Balanced block	Loop	Objective/ Type of design
Class comparison	Recommended for the comparison of more than 2 classes	Recommended for the comparison of 2 classes	A loop design is less efficient than a balanced block and requires more sophisticated analysis methods. Not recommended	Class comparison
Class prediction	Recommended	Not recommended	Not recommended	Class prediction
Class discovery	Recommended for implementing clustering methods	Not recommended	Not recommended	Class discovery

*As shown in the table the reference design is not only appropriate for every objective, but also eases future inter-experimental comparisons

population of the experimental samples. Care should also be taken during the selection of experimental samples, ensuring that they constitute a faithful representation of the features under study. Additionally, these gene profiling studies must be conducted in such a way that the accuracy of the measurements for experimental cases and their controls remains comparable; *e.g.* when studying tumor *vs.* healthy tissue, so that the potential for the introduction of measurement error biases is minimized.

A proper number of replicate measurements is another important design consideration. Replicates can be experimental or biological; experimental replicates are those designed to estimate experimental variability and are often implemented by placing multiple copies of the same probe on the array, or evaluating the same sample in different arrays, but are never directly related to the biological problem under investigation. On the other hand, biological replicates address the inherent variability between the individuals of the study population. Obviously, although experimental replicates improve the accuracy of the measurements they are unable to provide information about the intrinsic variability of the biological system being studied, which has led some authors to propose discarding experimental replicates altogether in favor of biological replicates [50, 51]. Still, there are specific cases in which experimental replicates are essential, like, for instance, during the evaluation of a predictor for clinical diagnosis.

Although falling within the topic of experimental design, the subject of sample size determination will not be treated on the present work, since it is widely discussed in the available literature [52-56].

Performing microarray experiments

Main steps

Independently from the experimental platform, there are common steps to all microarray experiments: RNA extraction, hybridization of each sample to one or several arrays, scanning of the array and digitization of the image, identification of the area on the image corresponding to each spot and assignment of signal intensities to each gene represented on the array as a measure of their expression levels on the sample.

Figure 5 shows the main steps required for the obtention of primary data from a microarray experiment. Assuming the availability of previously printed chips, the experiment consists on continuous processes of labeling, hybridization and reading of the samples under similar experimental conditions, followed by the stage of image analysis, which allows the measurement of the intensities for each probe contained in the chip for each experimental sample.

Further recommendations

- Every experiment should be performed by a single researcher

- The array to be used for each specific hybridization should be randomly assigned

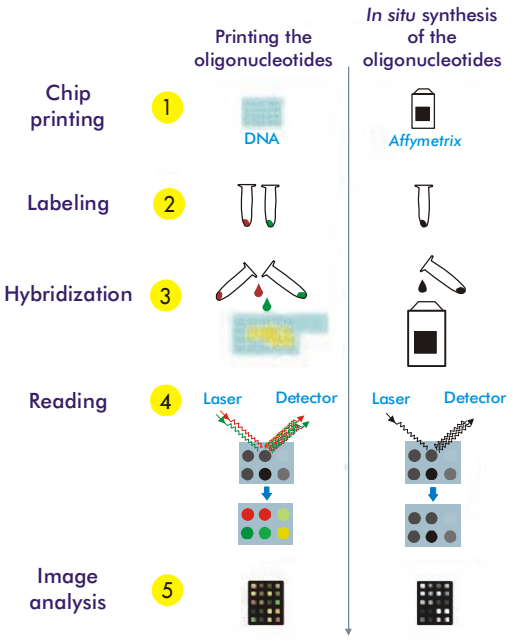

Figure 5. Main steps leading to the acquisition of primary data from a microarray experiment. The two most common technological platforms are represented: cDNA arrays in the left panel and *Affymetrix* chips on the right panel. Both cases require printing the chip, labeling the samples, hybridizing them to the chips, and using a scanner for reading and analyzing the images. The main differences between both methodologies reside in the processes of chip design and printing.

- Before applying a treatment, it is necessary to know the basal differences between the samples to be compared

- When the variability of the population is high, the best choice is to use a reference design

Conclusions

The summary presented here of the different stages for the design and implementation of a microarray experiment has described briefly the major technological platforms and the main objectives to be followed during biomedical research with this methodology, stressing the importance of the experimental design stage and the particular relevance, within the latter, of a properly stated biological hypothesis in order to choose the best experimental objective and design. Additionally, some ideas and recommendations about this stage have been discussed, which may help the researcher obtaining more accurate and reliable results. A case has also been made for the use of public microarray data to integrate this knowledge to the search for the molecular mechanisms underlying complex disorders such as cancer, exemplifying this last point through an analysis performed with available data from prostate cancer studies. The foreseeable developments of the technology include further increases in accuracy and throughput, which no doubt will modify the current practices for experimental design and will result in a wider spectrum of potential applications

Received in October, 2007. Accepted for publication in May, 2008.

49. Repsilber D, Fink L, Jacobsen M, Blasing O, Ziegler A. Sample selection for microarray gene expression studies. Methods Inf Med 2005;44(3):461-7.

50. Kerr, MK. Design considerations for efficient and effective microarray studies. Biometrics 2003;59:822-8.

51. Landgrebe J, Bretz F, Brunner E. Efficient two-sample designs for microarray experiments with biological replications. In Silico Biol 2004;4:461-70.

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

53. Pavlidis P, Li Q, Noble WS. The effect of replication on gene expression microarray experiments. Bioinformatics 2003;19: 1620-27.

54. Zien A, Fluck J, Zimmer R, Lengauer T. Microarrays: how many do you need?. J Comput Biol 2003; 10:653-67.

55. Dobbin K, Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 2005;6:27-38.

56. Pawitan Y, Michiels S, Koscielny S, Gusnanto A, Ploner A. False discovery rate and sample size for microarray studies. Bioinformatics 2005;21:3017-24.