Chapter 13
Gene Expression Studies Using Microarrays

Camila Guindalini and Renata Pellegrino

Introduction

The sequencing of the human genome and other organisms has been accompanied by major methodological and scientific advances in biology and molecular genetics technologies. Currently, in the post-genomic era, it is expected that the data accumulated for over 15 years of projects are finally translated into practical applications. This has generated a growing interest in the scientific community and a series of expectations about future applications of genetics in the understanding and diagnosis of complex diseases like cancer, diabetes, psychiatric and neurological disorders in general.

Among the new emerged technologies, the development of microarrays or DNA chips should be highlighted. This technique allows the investigation of thousands of genes simultaneously and promises to revolutionize predictive medicine, diagnostic and pharmacology, by substantially increasing the analytical capacity of the molecular processes.

Today, the availability of this new research method has allowed scientists to examine global gene expression that occurs in different cell types or a specific tissue, when subjected or exposed to a certain pathological or experimental conditions. Moreover, it is also possible to examine structural variations in DNA sequence that may contribute to increased susceptibility to diseases, in a quick, economical and systematic approach.

Thus, the focus of studies on the pathophysiology of complex diseases tends in the short term, move from the characterization of individual processes and

C. Guindalini (✉) • R. Pellegrino
Department of Psychobiology, Universidade Federal de São Paulo (UNIFESP),
São Paulo, São Paulo, Brazil
e-mail: camila.guindalini@afip.com.br

M.L. Andersen, S. Tufik (eds.), *Rodent Model as Tools in Ethical Biomedical Research*, DOI 10.1007/978-3-319-11578-8_13

mechanisms to the investigation of biological systems as a whole, integrating and generating data that are more realistic and closer to the complexity of an organism. The basic concepts that underlie this technique, as well as the important points to consider in designing an experiment using microarrays, its advantages, prospects and future scientific directions will be discussed.

Gene Expression and Microarrays

The complete genome of a given organism is composed of thousands of genes. Genes are selected regions of the DNA molecules that serve as templates for synthesizing RNA, in a process called transcription (Fig. 13.1). In turn, RNA is, in the majority of the cases, used to guide the synthesis of polypeptides, which subsequently form proteins either directly or by supporting the different stages of gene expression. The RNA molecules which specify a particular polypeptide are known as messenger RNA (mRNA). In this sense, mRNA may be seen as an intermediate product and proteins as the major functional end-points of the DNA template. This, on the other hand, is not the case for non-coding RNA genes, which are genes that encode a functional RNA molecule that is not translated into a protein and include: transfer RNA (tRNA), ribosomal RNA (rRNA), as well as, microRNAs and short interfering RNA (siRNAs), molecules recently described to play a crucial role in gene expression. However, not all genes are active in every cells all of the time. Some are expressed in specific cell types, at particular stages of development, or even in a precise period of the day. In genetics, gene expression is the most basic level at which genotype influences the phenotype.

With the advent of the microarray technology, today scientists have the possibility to analyze the expression of thousands of genes in parallel and use this information to determine gene expression profiles. The analysis of all expressed genes in a target sample is also entitled transcriptome analysis and is increasing being conducted using microarray based approach. In this specific type of experiment, the aim

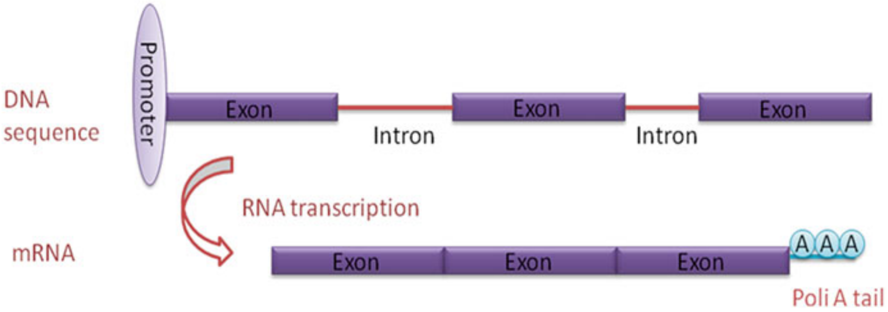

Fig. 13.1 Schematic representation of mRNA transcription process

is to quantify the types and amounts of mRNA molecules present in a particular sample. The number of mRNA molecules derived from a given gene may be seen as an approximate estimate of the level of its expression. The idea is to identify variations in the level of gene expression that may occur as natural biological responses due to the presence of particular disease, or some other experimental or pathological condition, with the assumption that the mRNA levels will reflect protein abundance and help explain the phenotype of interest.

About 20 years ago, the microarray technology was known as *macroarrays* with experiments performed on large membrane sheets made of nitrocellulose spotted with complementary DNA (cDNA), representing around 1–10,000 genes, and were used for comparative hybridization of RNA samples. This technology, although an advancement in comparison to classic methods such as Northern and Southern blotting has moved through to the chip technology, which is available today. Microarrays are small, solid supports onto which the sequences of cDNA or oligonucleotides derived from thousands of different gene sequences, hereafter called *probes*, are immobilized at specific locations in an orderly and fixed manner. The solid supports are typically glass microscope slides, silicon chips or nylon membranes, where the probes are attached to a chemical matrix via surface engineering by a covalent bond. There are a number of different variations on the microarray technology and there are different names for the commercial microarrays, such as DNA/RNA Chips, BioChips or GeneChips. The protocol basically starts with the extraction of total RNA from the specimen and the isolation of the mRNA. The mRNA transcripts are then converted to a form of fluorescent dye labeled nucleotides, normally referred as *targets*, and subsequently, hybridized to the microarray (Fig. 13.2). During the hybridization, the target will bind to the probes on the array by sequence complementarily and the excess sample will be submitted to a washing off procedure. At this point, each probe on the microarray should be bound to a quantity of labeled target that is proportional to the level of expression of the gene represented by that probe. The amount of fluorescent emission on each probe will be used to generate a signal intensity, which will afterwards be processed by bioinformatics tools and provide information on the level of expression of all the corresponding genes.

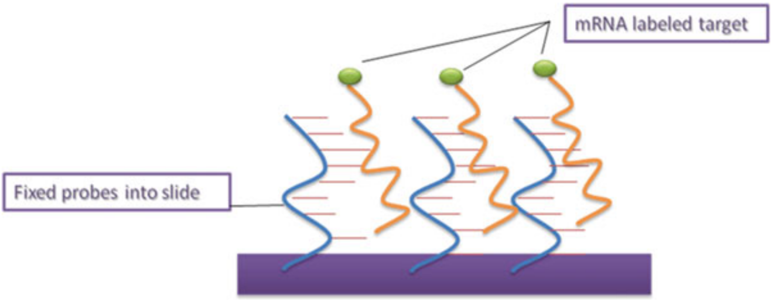

Fig. 13.2 Representation of hybridization of the fluorescently labeled target RNA sample to the synthesized probes immobilized at specific locations on a solid support of the microarray

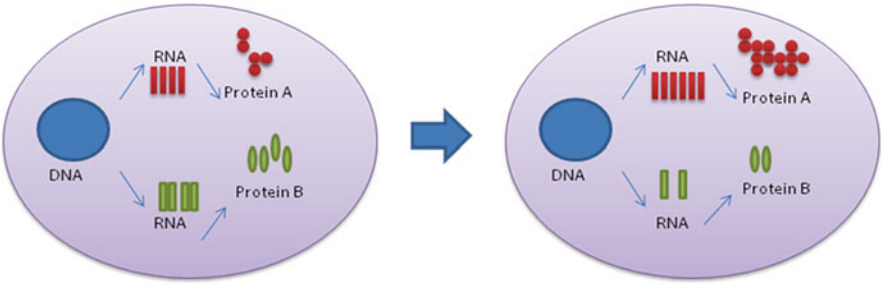

Fig. 13.3 An example of the microarray technology application. The baseline gene expression of two hypothetical genes and their expression levels modified, as a consequence of an experimental manipulation, or altered physiological condition

Microarrays have been applied to many types of biological approaches such as, responses to environmental changes, classifications of tumors, characterization of therapeutic drugs, among others (Fig. 13.3). At the present time, the main large-scale application of microarrays is comparative gene expression analysis. Because of the greater facility in acquiring samples and the nature of the disease itself, the most successful application of microarray technology has been to the study of tumor tissues. In recent years, the technology has been applied to the identification of specific patterns of gene expression that characterize different types of cancer, predict prognosis and responses to specific therapies. However, the efficiency and robustness of microarray analysis have been presented in areas as diverse as: neurological diseases, asthma, psychiatry diseases and cardiovascular diseases, with very interesting and promising results.

Technical Considerations

Experimental Design

A proper experimental design is crucial for obtaining useful conclusions from a project. The choice of the design ideally includes an assessment of the biological variation, the technical variation, the cost and duration of the experiment, as well as the availability of biological material (Fig. 13.4). The experimental plan can also depend on the methods that will be used to analyze the data afterwards. In certain cases, the parameters needed to find the optimal design must be obtained by a pilot experiment. Microarray experiments have multiple sources of variation, including variation from measurement errors associated with the array assays, laboratory process, and biological variation, representing the variability among the subjects under study. Therefore, experimental designs should ensure that effects of interest are not confounded with ancillary effects. For example, it is well-known that even when genetically identical, variability between animals in the same group may be observed. Therefore, it is very important to have a maximum control of experimental

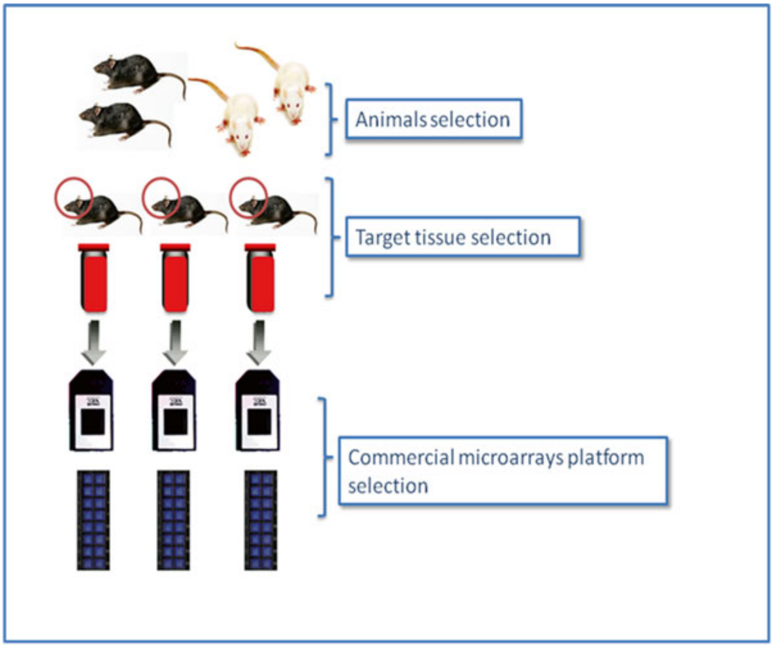

Fig. 13.4 The experimental design overview of a microarray experiment

conditions, establishing uniform procedures for the handling and treatment of the animals. Moreover, the number of animals per cage, diet, gender, age, length of fasting, circadian patterns, stress conditions and the random assignment of the animals to the different treatment groups are important factors that should be carefully established, in an attempt to eliminate potential source of variability. Notably, when separating the tissues or cell lines for the microarray experiment, gloves should be used at all times during the extraction procedure and while handling materials and equipment to prevent contamination. All equipment should be as free as possible from contaminating RNases and should, as often as possible, be treated with diethylpyrocarbonate (DEPC) and autoclaved with baking. The collection of the sample is also crucial and should be performed in a minimum period of time to prevent RNA degradation, since RNA integrity is critical for successful quantitation. In addition, the sample should be immediately snap-frozen in liquid nitrogen or dried ice and kept at –80 °C until the RNA extraction procedure takes place.

The Importance of Replicates

In microarrays studies, there are common strategies to control for technical assays and biological variations. Performing technical and/or biological replicates of the experiment being conducted is one of the classical approaches used by researchers to increase the power of the study. The technical replicate relates to the multiple labeling of the same RNA sample, with the motivation to reduce the variability related to assays and laboratory conditions (array to array difference, reagent lots, dye incorporation, apparatus, and operator, among others). Biological replicates involve the isolation of RNA from different samples independently (multiple cell lines, multiple biopsies, multiples animals, and multiple patients). The main principle of a biological replicate is to control for the biological diversity between samples.

There is no precise rule to define the number of replicates needed per microarrays experiments. However, for statistical instance, the maximum number of samples that can be handled within the biological experiment is important for the accuracy of the experiment. At the present time, it is advocated by experts in the field that, when possible, one should always substitute technical replicates with biological replicates, since individual variability are suppose to be higher then the variability derived from the technical process. A recent published guideline suggest three biological replicates for cell line work, six for animal tissues and at least ten for human samples. Furthermore, when considering two or more groups for analysis, more samples per conditions are required. Conversely, when running a time course experiment, fewer replicates per time point should be sufficient.

Another point to consider when designing an experiment is that the choice of tissue to be analyzed by the microarray technique should be based on its relevance to the physiology of the pathology of interest and/or to the location where a specific process is taking place. In addition, it is important to note that gene expression may not be only tissue-specific, but also cell-specific. Thus, the expression profile of certain population of cells may be modified, if analyzed together with different cell populations. Accordingly, new technologies such as microdissection and laser capture, which allow the extraction of specific individual cells, are already being used by several groups. As a result, the microarray technology should constantly adapt to enable the achievement of highly specific and accurate results from ever smaller amounts of RNA.

The Impact of Pooling

In microarray experiments, sometimes pooling RNA samples before labeling and hybridization may be considered, in cases where there is insufficient RNA from each individual sample, or to reduce the number of arrays for the purpose of saving cost or of simplifying the laboratory procedures. The basic assumption of pooling is that the expression of a particular mRNA molecule in the pool is close to the average expression from individuals that comprise the pool. However, it has been exhaustively discussed that pooling individual samples has a number of disadvantages: (1) the potential risk for pooling bias, e.g. significant differences between the gene expression indentified from the pooled sample and the average signal that would be derived from the individual measurements; (2) the impossibility of detecting and eliminating technical or biological outliers, which would have an effect on the data obtained from the pool; (3) the loss of information about the individual variability, which would eliminate the feasibility of indentifying specific characteristic of a given individual or clustering samples in clinical or pharmacologic subgroups; (4) difficulty in estimating variance between samples, and relying only on the observed fold-change to select genes, since it would not be possible to incorporate any statistical assessment regarding the reliability of the findings.

Nevertheless, if pooling is chosen as the research strategy, one may consider using as many independent pools as possible, so that the sets of pooled samples in each array will represent a biological replication. It has been demonstrated that in certain situations, pooling an increased number of specimens allows the researcher to reduce the number of arrays without losing precision. In addition, being more specific in the biological question under study and considering the results of a pooling experiment only as a screening exercise for future in depth analysis, while recognizing the possibility of detecting false negative and positive findings, may also help researchers to extract reliable information from a pooling experiment.

The Extraction and Quality Control Checking

The RNA quality is the most important factor that will establish the success or failure of any microarray assay. The artifacts caused by nuclease activities, potential cold shock reactions and contaminations can be avoided if the experiment process is strictly controlled and well planned. In this sense, the collection and preprocessing stage are crucial for high-quality RNA isolation. When it is not practical to extract RNA from tissue samples immediately, the samples should be snap frozen in liquid nitrogen or dried ice within 30 min after dissection. As an alternative, RNA stabilizing solutions can be used in an attempt to maintain the integrity of RNA during longer periods of time. Several methods are available to adequately isolate RNA from tissue and cell lines samples. The most common of these is the guanidinium thiocyanate-phenol-chloroform extraction. The method is very useful in providing high-quality concentration of RNA, however technical guidelines suggest that this method should not be used alone. The microarrays assays are very sensitive and since phenol may remain in the RNA solution after extraction, lowering the efficiency of the experimental reactions, the subsequent purification of the sample using a column-based method to remove the phenol residues and keep the purity of RNA is highly recommended. The procedure of RNA extraction is further complicated by the ubiquitous presence of ribonuclease enzymes in cells and tissues, which can rapidly degrade RNA. Therefore, maximum care should be applied during the entire process from tissue collection to RNA purification, in an attempt to maintain the integrity of the samples.

Prior to running a microarray experiment, RNA quality must be adequately checked. There are three characteristics of the isolated RNA that may be measured: quantity, quality and integrity. The most commonly used method to perform the inspection is spectrophotometer analysis though UV absorption measurements. This will allow the determination of the sample concentration and the presence of contaminants, such as proteins and phenol residues. In brief, the absorbance is measured at 260 and 280 nm and the ratio of absorbance at 260 and 280 nm is used to assess the purity of DNA and RNA. A ratio of ~2.0 generally indicates pure RNA. Since RNA has its maximum absorption at 260 nm, if the ratio is appreciably lower, it may indicate the presence of protein, phenol or other contaminants that absorb strongly at or near 280 nm. RNA quality is also usually assessed by electrophoresis on an agarose gel, followed by staining with ethidium bromide (Fig. 13.5). The presence of clear 28S and 18S ribosomal RNA bands are indicative of non-degraded RNA. However, it is important to remember that a number of technical conditions such as saturation of ethidium bromide fluorescence, the amount of sample loaded, agarose quality and concentration may influence the visual evaluation and should always be taken into consideration and standardized as accurately as possible. Moreover, it is also not clear if clear 28S and 18S bands do reflect the characteristics of the underlying mRNA population, which are know to present a more rapid degradation. One excellent alternative to improve the assessment of RNA quality and to standardize the process of RNA integrity interpretation is pro-

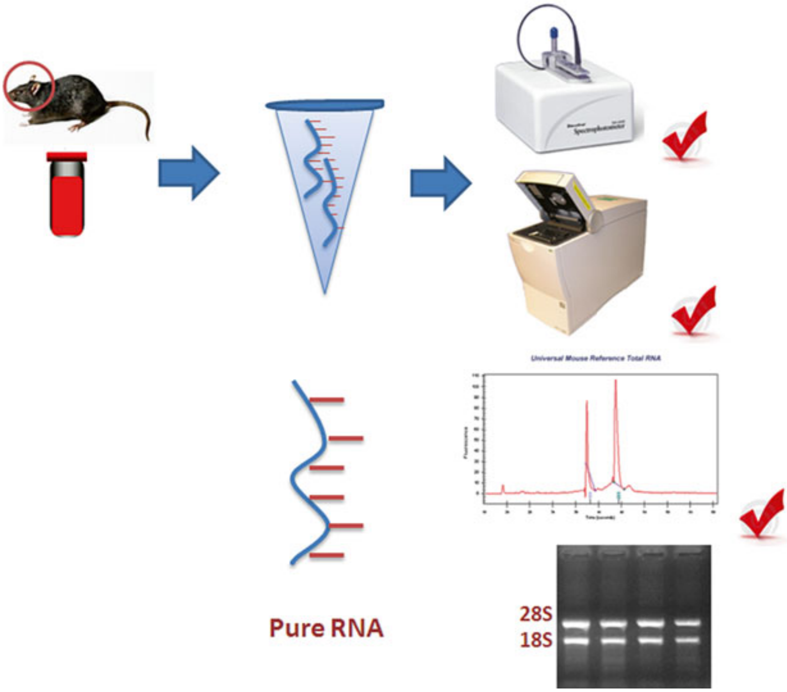

Fig. 13.5 The RNA extraction and quality control checking: sample selection, RNA extraction, spectrophotometer inspection, capillary electrophoresis and ribosomal band visualization on the agarosis gel

vided by the equipment Agilent 2100 bioanalyzer (Agilent Technologies Inc., Palo Alto, CA), a commercially available system that employs chip-based nucleic acid separation technology. RNA samples are separated by capillary electrophoresis on a microchip device (LabChip 7500; Caliper Technologies, Mountain View, CA) and subsequently detected via laser induced fluorescence detection. An electropherogram and gel-like image will be generated by the software, providing the sample concentration, the ratio of the 18S to 28S ribosomal subunits and a more accurate and standardized visualization of the RNA quality and integrity. This new technology introduces an interesting tool for RNA quality assessment, which is called RNA Integrity Number (RIN) and was developed to reduce the subjective interpretation and potentially incorrect determination of RNA quality. The software classifies eukaryotic RNA according to a numbering system that ranges from 1 to 10, with 1 indicating important levels of degradation and 10 representing highly intact and pure RNA. The acceptable number of RIN for microarrays experiments is 6 or higher. The entire process from RNA extraction to quality control samples is represented in Fig. 13.5.

A Typical Experimental Protocol

After sample quality control checking, a typical microarrays protocol may be performed using either total RNA or mRNA. The experiment starts with the target RNA being first reverse transcribed using a T7-Oligo(dT) Promoter primer in the first-strand cDNA synthesis reaction. Subsequently to the second-strand cDNA synthesis mediated by RNase H, the resulting double-stranded cDNA is purified and serves as a template in the following *in vitro* transcription (IVT) reaction. In this step, the complementary RNA (cRNA) is synthesized in the presence of T7 RNA Polymerase and a biotinylated nucleotide analog/ribonucleotide mix. The labelled cRNA products are then cleaned up and submitted to a fragmentation reaction to finally be hybridized to the microarray slide. Immediately following the hybridization, the microarray is submitted to a washing off procedure for the removal of non-specific bonding sequences. The amount of remaining hybridized target molecules is proportional to the quantity of the originally isolated mRNA. Finally, the microarray slide is scanned while connected to specific software that processes the data and quantifies the intensity of fluorescence at each point. This information will then be used for the relative quantification of differently expressed genes. The detailed assay is shown in Fig. 13.6.

Data Analysis

Microarray data sets are commonly very large, and analytical precision is influenced by a number of variables. Statistical challenges include taking into account effects of background noise and appropriate normalization of the data using

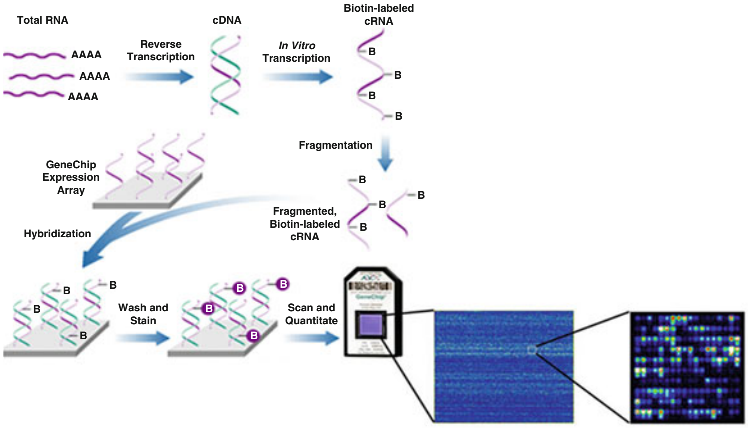

Fig. 13.6 Example of typical microarray experimental protocol. Reproduced with authorization of Affymetrix Inc

algorithms methods. Nowadays, a number of powerful freely available as well as commercial software's packages that incorporate different microarray analysis algorithms have been developed to allow researchers to capture, manage, and analyze effectively data from DNA microarray experiments. After the capture and imaging, the data obtained by a microarray experiment are subjected to a series of analytical processes, which involves the standardization and elimination of experimental noise so that obtained expression estimates reflect the true changes in mRNA abundance, as precisely as possible. After background correction and normalization, the transformation of intensity values into adequate data for statistical analysis is the subsequent step for indentifying differently expressed genes between groups. The most straightforward approach to select potentially regulated genes is ranking the results with respect to fold change. For example, genes demonstrating a twofold or greater change may be regarded as differently expressed and further selected for in-depth examination. However, this approach will have an obvious drawback, in the way that such a selection does not provide the investigator with any measure of the reliability of the observed change, since it does not take inter-experimental variability into account. Moreover, it is possible that a given gene with high fold change may as well as be greatly variable, and therefore its selection will end up providing poor information on its regulation and will be fairly imprecise. In this sense, statistical tests, such as t-tests, Analysis of Variance, Maximum-likelihood analyses, F-statistics, and the non-parametric equivalents, as well as the new generation of modified t-statistics, are alternative methods to identify significant changes between group means. Nevertheless, the choice of method used to identify differently expressed genes can have an important influence on the selected gene list and ultimately, this decision should be based on biological, rather than on statistical considerations. If the research question relies on the identification of absolute changes in gene expression, and not in the variation within the groups, the use of the fold change is recommended. On the other hand, statistical tests or the combination of both methods are more appropriate if one is interested in changes in gene expression relative to the underlying noise for a given gene.

Because of the large number of genes being tested in one experiment, the probability of identifying false positives is substantially enhanced when increasing number of tests are being performed. Therefore corrections for multiple testing methods, such as False Discovery Rate (FDR) and Family-Wise Error Rate (FWER), should be performed before the differentially expressed genes are selected and further analyses are conducted. Once a definitive list of potentially regulated genes is produced, the next step is to biologically interpret the data, using clustering and functional analyses methods for a more detailed understanding of the gene expression profile observed. Clustering is an exploratory data analysis tool that aims to group similar objects to respective categories, according to some measure of similarity. Typically, clustering is used as a strategy to present and summarize the microarray data in the format of dendrogram or heatmaps. Both samples and genes can be clustered, therefore highlighting the overall similarity of samples within a given group, providing discriminative information based on certain selection of genes, indentifying groups of possible related genes, or even providing an illustration of existing gene patterns within the set of microarrays. Functional analysis, also known as functional enrichment, is a method that integrates the gene list indentified by the experiment with the available literature, normally public databases, extracting information on potential biological pathways altered by the experiment (Fig. 13.7). This approach is especially important since analyzing genes as independent entities disregard the fact that genes do not work in isolation but in pathways. Available online programs such as Gominer™, The Database for Annotation, Visualization and Integrated Discovery (DAVID) and Ingenuity Pathway Analysis® provide information on enriched biological themes, gene ontology terms, enriched functional-related gene groups, other functionally related genes not in the list, gene-disease associations, among others.

Validation of Differentially Expressed Genes

Finally, the verification of positive results using a second independent technique is a well-established strategy performed among microarray users to validate their findings. This replication of data is extremely important, since small inconsistencies in protocols may cause subtle changes in expression levels of genes, increasing the chance of indentifying false-positive and false-negative signals. The most common method to confirm microarrays findings is the quantitative real-time polymerase chain reaction (qRT-PCR). The method is a rapid, sensitive and less complex technique for gene expression analyses and offers the opportunity for the investigation of multiple targets in a relative small standardization time. The selection of the gene set for validation analysis depends on many factors such as the original experimental design, relative difference in expression among the samples, biological function and availability of appropriate reagents (primers and antibodies). One important

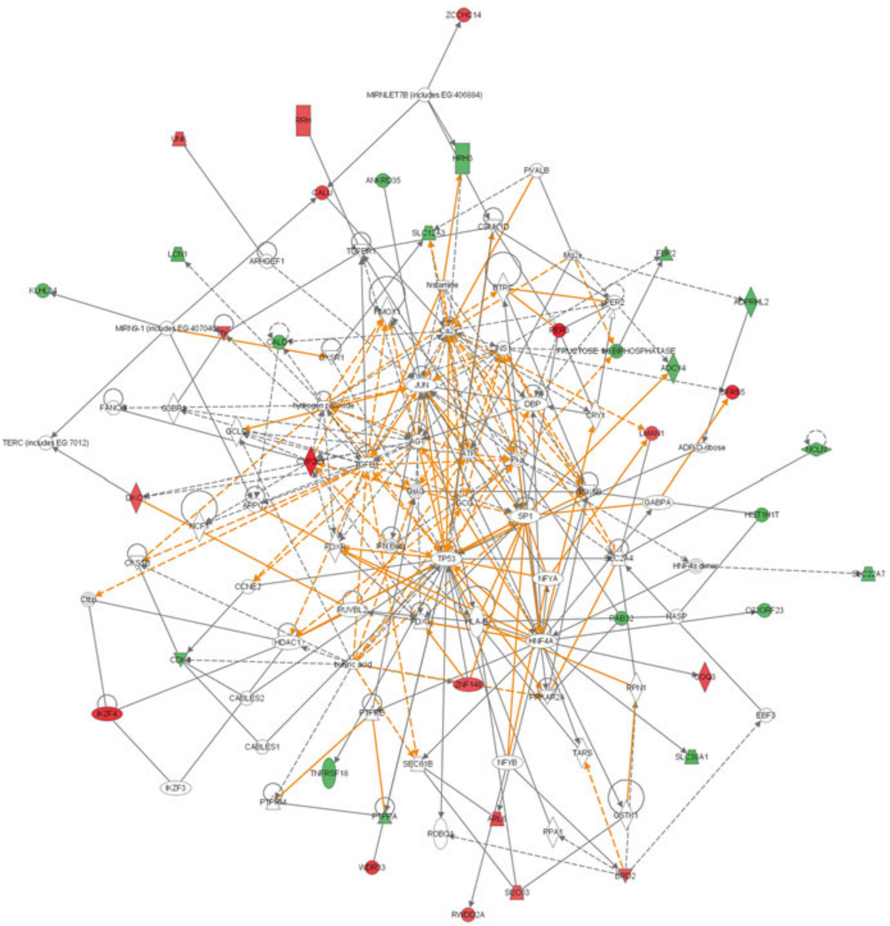

Fig. 13.7 Example of a functional pathway analysis result integrating the list of down (*green*) and up (*red*) regulated genes indentified by the experiment with the available literature and with other functionally related genes (*black*) not indentified

point to consider is that usually, commercial arrays contain a number of different isoforms of the same gene. In this sense, if an inadequate probe selection is performed, the researcher may design primers and perform qRT-PCR of transcripts that were originally not altered on the original experiment. This is one of the main reasons of inconsistencies between arrays and qRT-PCR results. As an additional recommendation, when possible, microarrays findings should be replicated using the original samples (technical replication), as well as using independent new samples (biological replication). Other methods, such as Northern and Western blot analyses, which measures RNA and protein levels, respectively, are also frequently used to validate the microarrays findings. The limitations of those techniques include time for conducting the experiment and the small number of genes that can be interrogated.

Of note, in the last years, the advances in research and the development of more robust platforms have increased the confidence in gene expression data derived from microarray experiments and it is probable that, in the future, the validations procedures may be an optional step to be performed by the researcher.

Future Perspectives

The number of studies involving the use of microarrays for the identification of new genes and molecular mechanisms has grown exponentially. We are moving to a new scientific level, in which the complex pathophysiology mechanisms of a number of diseases are now closer to be understood. The promise for the future is that biomarkers identified by this new technology will help the understanding of a number of conditions, and will eventually be directly incorporated into the diagnosis and treatment of diseases. However, despite robust and conceptually simple, the use of microarrays is still rather limited due its cost, which is considered an important bottleneck for a number of research groups. In addition to a careful experimental design, which involves the acquirement of high quality samples to the choice of the correct and most appropriate platform of analysis, another point to consider is the need for the implementation of bioinformatics tools and statistical analysis capable of managing and interpreting the massive amount of data that is generated, after each experiment. The latter seems to be a crucial factor for the success and reproducibility of the assays. Nevertheless, the next few years await further advances in the development of the technique, making it more accessible to the scientific community, both in financial and analytical terms. If used in an appropriate context and accompanied by appropriate biostatistics methods of analysis, microarray technology can be an important screening tool, which is capable of revealing valuable clues related to the pathophysiology of complex diseases, ultimately offering conditions for the development of new research strategies.

References

Chuaqui RF, et al. Post-analysis follow-up and validation of microarray experiments. Nat Genet. 2002;32:509–14.

Fodor SP, et al. Multiplexed biochemical assays with biological chips. Nature. 1993;364:555–6.

Göhlmann H, Talloen W. Gene expression studies using affymetrix microarrays. 1st ed. Chapman & Hall/CRC: Boca Raton; 2009.

Guindalini CSC, Tufik S. Use of microarrays in the search of gene expression patterns—application to the study of complex phenotypes. Rev Bras Psiquiatr. 2007;29:370–4.

Jafari P, Azuaje F. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. BMC Med Inform Decis Mak. 2006;6:27.

Kendziorski CM, et al. The efficiency of pooling mRNA in microarray experiments. Biostatistics. 2003;4:465–77.

Koremberg MJ. Microarrays data analysis: methods and applications, Series methods in molecular biology, vol. 377. New York: Humana Press; 2007.

Morey JS, et al. Microarray validation: factors influencing correlation between oligonucleotide microarrays and real-time PCR. Biol Proced Online. 2006;8:175–93.

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

Shendure J. The beginning of the end for microarrays? Nat Methods. 2008;5:585–7.

Shi L, et al. Reproducible and reliable microarray results through quality control: good laboratory proficiency and appropriate data analysis practices are essential. Curr Opin Biotechnol. 2008a;19:10–8.

Shi L, et al. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. BMC Bioinformatics. 2008b;9:S10.

Shih JH, et al. Effects of pooling mRNA in microarray class comparisons. Bioinformatics. 2004;20:3318–25.

Slonim DK, Yanai I. Getting started in gene expression microarray analysis. PLoS Comput Biol. 2009;5:1–4.

Stanislav M, et al. Sources of variation in Affymetrix microarray experiments. BMC Bioinformatics. 2005;6:214.

Zhang SD, Gant TW. A statistical framework for the design of microarray experiments and effective detection of differential gene expression. Bioinformatics. 2004;20:2821–8.