The Journal Of The International Federation Of Clinical Chemistry And Laboratory Medicine

eJIFCC

GENE EXPRESSION PROFILES – WHAT THE CLINICIAN NEEDSTO KNOW

Koraljka Gall-Troselj

Corresponding author's address:
Koraljka Gall-Troselj, MD, PhD
Laboratory of Molecular Pathology,
DMM Rudjer Boskovic Institute,
10 000 Zagreb, Croatia

Genetics and molecular medicine have an expanding need for rapid genotyping, mutational analyses and DNA re-sequencing technologies which have clear potential for automation and high-throughput screening. Expression technology is a tool for investigating expression patterns, identifying new genes (either for monogenic diseases or complex traits), identifying new pathways and possibly related drugs. Microarrays have recently gained widespread use. The aspiration of this technology is custom-tailored pharmacotherapy, with each patient treated effectively based on the gene expression signatures of the host/afflicting pathogen.
The term "microarray" originally stood for an array of a number of cloned DNA molecules affixed to a glass slide. A similar term "DNA chip" was used to describe an array of short DNA oligomeres directly synthesized on a slide. Recently, however, the microarray is also referred to by this name.

The basic biochemistry of reactions on chips includes labeling and hybridization of cDNA or cRNA targets derived from the mRNA to nucleic acid probes attached to the solid support. By monitoring the amount of label associated with each DNA location, it is possible to infer the abundance of each mRNA species represented. There are two basic approaches: hybridization of the test and control sample on the same chip by the use of fluorescent dyes; or, one sample-one chip hybridizations using non-fluorescent labeling. Although hybridization has been used for decades to detect and quantify nucleic acids, the combination of the miniaturization of the technology and the large and growing amounts of sequence information, have enormously expanded the scale at which gene expression can be studied.
Whole-genome analyses also benefit studies where the objective is to focus on small numbers of genes, by providing an efficient tool to sort through the activities of thousands of genes, and to recognize the key players. In addition, monitoring multiple genes in parallel allows the identification of robust classifiers, called "signatures", of disease. Often, these signatures are impossible to obtain from tracking changes in the expression of individual genes, which can be subtle or variable. Global analyses frequently provide insights into multiple facets of a project. A study designed to identify new disease classes, for example, may also reveal clues about the basic biology of disorders, and may suggest novel drug targets.

17.1 Types of DNA microarrays

During the past few years, quite a few types of DNA microarrays have been developed including: macroarray; cDNA arrays; high-density oligonucleotide microarrays (these two are most commonly used for transcriptome analysis); and microelectronic arrays.

Although both, cDNA arrays and high-density oligonucleotide microarrays work fairly well with expression analysis experiments, there are pros and cons for the use of each type of microarray. Longer cDNA probes offer lower cost, higher specificity and stronger signals, whereas shorter probes offer higher densities. Shorter oligos are able to distinguish transcripts with single mismatches making them attractive also for genotyping applications. Higher sensitivity is usually achieved by incorporating several different oligonucleotide sequences from the same gene allowing for multiple probing events in the same experiment. Nylon membranes robotically spotted with cDNA inserts or genomic fragments are typical macroarrays. The probe density is lower, with spacing between spots typically being 1–2 mm. The detection is usually based on chemiluminescent labeling. This type of arrays is suitable for the simultaneous analysis of tens to hundreds of genes.

High-density oligonucleotide arrays usually contain *in situ* synthesized 25-mer oligonucleotide probes utilizing the photolithography and solid-phase DNA synthesis techniques. The technique is still limited to commercial production. A standardised operating system, including hardware, software, a defined protocol, chemicals and analysis tools, has been constructed for GeneChip (Affymetrix Inc, USA) arrays which contain thousands of different oligonucleotides on a small glass surface. These have been designed and used for quantitative and highly parallel measurments of gene expression, to detect the presence of alternatively spliced transcripts and to discover polymorphic loci. Since 2004, the entire genome can be "viewed" on the GeneChip Humane Genome U133 Plus 2.0 array. It contains 1.3 million distinct oligonucleotides and is designed for expression analysis of 47,000 transcripts as well as variants. This includes over 30,000 well-characterized human genes. GeneChip microarrays utilize several probe sequences to interrogate a single gene's expression and also to include a mismatch sequence for each probe wherein the mismatch has a single base change. These mismatch probes permit comparing sequence specific hybridization signals and non-specific signals from background.

Contrary to macro- and micro-arrays, hybridization of NanoChipTM array can occur in minutes. Microelectronic arrays have been recently introduced to the market and are yet to be thoroughly tested by end-users. The NanoChipTM array

(http://www.nanogen.com) is a 99-site electronically-powered microarray. Each test site is electronically connected to a computer with platinum wires. The sequence-specific probes are electronically addressed to specific sites and the biotin-labeled RNA samples can be transported and hybridized to the complementary probes on the NanoChipTM array rapidly and precisely by electronically manipulating the charge at the test sites.

Microarray technologies are an important development in molecular diagnostics as well as in the development of personalized medicine in several areas. Currently the most important of these currently is personalised treatment for cancer.

17.2 Cancer (re)classification and prognosis

The biggest challenge in cancer diagnosis cannot be the identification of cancer lesions per se. Cancers occuring in the same organ may be morphologically indistinguishable but may belong to biologically and clinically distinct classes. The ability to classify a group of unknown samples into subcategories holds much promise for cancer diagnosis and patient-tailored medicine. Although advanced cytogenetics and molecular analysis tools have been used in subclassification of some cancers (e.g., acute leukaemia), molecular classification solely based on gene expression is a viable alternative and may provide further information regarding the disease state. In addition, expression profiles unique to each subtype of cancer can be used for developing therapy and monitoring therapeutic efficacy.
The proof of principle was first illustrated in a study involving 72 acute leukaemia samples in which supervised learning methods based on known identities of acute myelogenous leukaemia (AML) and acute lymphocytic leukaemia (ALL) generated a profile (classifier). The profile was successfully used to classify a group of unkown samples into the correct category.

This study established the feasibility of expression-based tumour classification.

Although not an exclusive approach to diagnostic cancer classification, expression-based outcome analysis sets a primary goal to identify previously unrecognized prognostic subtypes. It deals primarily with the gene expression correlates of the treatment outcomes and predicting of outcomes using molecular "predictors". Many tumour classification studies mentioned above have correlated tumour subtypes with clinical outcomes. Nevertheless, studies performed primarily in lymphoma, leukaemia, and breast cancer have established the feasibility of outcome prediction solely based on gene expression.

One should know....

To obtain reliable and reproducible data, one should carefully plan the experiment. The high cost of microarray experiments dictates the necessity to optimize all steps involved including selecting the biological system, isolation of RNA and the choice of microarray. Contrary to homogenous cell populations such as cell lines or purified cell populations, studies involving tissues or organs add more complexity. They contain several diverse cell populations and the gene expression profiles obtained from them may not truly represent the "real" conditions. This is of special concern in microarray experiments using tumour samples. RNA isolated from the biopsy sample may contain both normal and cancerous cells and the expression profile in the cancer cell may be diluted by the normal background. The current trend of isolating a homogenous population using fractionation techniques prior to RNA isolation when possible, attempts to resolve the uncertainty associated with signal origin.

The high throughput nature of the data acquisition process makes data mining the bottle-neck in the process. Some criteria that need to be determined when using any of available softwares include: normalization routines (allow for accounting the variability across multiple chips in a single study or between studies); filtering strategies; statistical testing routines; and data representation. The choice of microarray platform and image analysis software dictates the detection algorithm used with little input from the end-user.

17.3 Experimental models

The major goal is to compare mRNA expressed in one subset of cells/tissue with a control set. As already stated, the RNA or mRNA must be isolated from the target cells or the tissue sample of interest. This is usually not problematic when dealing with isolated cells in culture but not at all easy when working with organ samples. Since cells in culture can be exposed to a particular stimulus all at once, the reaction can be assumed to be uniform and differences in RNA expression reflect the specific answer to the stimulus.

In complex tissue samples, like organs, it is always difficult to answer specific questions since different cell types/areas of an organ may react differently. For that reason, only the isolation of the specific areas/cells permits asking precise questions. One good example where the discrimination of gene expression is important is differentiation between cancerous and normal or inflamed and control tissue. In all these cases, accurate separation (by use of laser-capture dissection), between the samples of interest and proper controls is an essential. The problem here is that the amount of RNA, which can be isolated from such small samples, is limited and very often further analysis needs amplification steps.

Direct labeling of the cDNA using reverse transcriptase and properly labelled nucleotides is the method of choice when enough RNA is available. Protocols vary from 100ng to 2mg RNA as the amount of starting material required. Labelled nucleotides include radioactive, fluorescent, biotin, digoxigenin or aminoallyl-tagged dNTP. The choice of label depends on preferences and the equipment of the laboratory for detection after hybridization.

Proper controls are necessary. The presence of several housekeeping genes as positive controls as well as bacterial or yeast genes as negative controls are very important steps in the development of reliable standards. In addition the representation of several stretches of the same gene and the inclusion of hybridization controls by adding mismatch controls adds further certainty in the specificity and information derived by this technology. The inclusion of a control by spiking the sample with, for example, phage RNA and having the corresponding sequences on the array, allows for normalization and quantification of this particular sequence as well as detecting differences in the overall hybridization between different sets of arrays.

A difference in the hybridization signal of more than twofold is expected to represent a distinct difference in expression of the specific gene. The question arises, how to choose which array is useful for getting maximal/optimal and also meaningful results in a given experiment? It primarly depends on whether the goal of the

experiment is to get an overview on as many as possible known genes or the detection of new genes.

Different appraches have proven their usefulness. The total analysis of 65,000 genes with regard to leukaemia has changed the perspective on disease classification with consequences even for therapy. On the other hand after these pioneering activities, it is now clear that only a handful of genes allow the differentiation of leukaemia. This has reduced the number of genes to be analysed in clinical practice. Microarray technology (including oligonucleotide arrays and cDNA arrays) offers great promise for functional genomics research, and potentially, can transform the diagnosis and treatment of diseases. Limitations include the paucity of RNA from small samples, the quality of RNA from medical samples, the detection limit and the price for arrays.

17.4 An overview

Microarray techniques introduce a new challenge for clinicians and pathologists. Being in charge of the first assessment of the disease, they need to know how to use these new methods optimally. The study described in the American Journal of Pathology – 2002, was the first one designed to evaluate the microarray results on protein levels (in addition to RealTime PCR that was also used for checking the microarray results on the RNA level) of a breast tumour series (55 samples). It was designed to evaluate the interest and limitations of immunohistochemistry performed on a large scale (TMA – tissue microarray) that was constructed with three cores per tumour sample. It was very interesting that, in a majority of cases, cDNA array and TMA data obtained on the same breast tumour samples gave different results. One of the "problematic" genes was gene p53. This was not surprising, as p53 protein detection is not dependent on mRNA overexpression, but is the result of the increased half-life of a mutated protein. In normal cells, p53 protein half-life is short and expression levels are low and undetectable by IHC. In cancer cells, most p53 mutations lead to products that are not ubiquitinated and accumulate in the nuclei where they can then be detected.
It has to be stated here that, for many genes, there is little correlation between the abundance of the mRNA transcript and the steady-state levels of the encoded protein. Post-transcriptional and post-translational mechanisms are likely to influence protein expression, thus blurring the correlation between mRNA and protein levels. Proteins encoded by very low levels of RNA, below the detection level of cDNA arrays, can be detected by IHC because of increased protein stability (the case of p53), or the high sensitivity of the antibody. Reciprocally, elevated levels of RNA may produce only small amounts of detectable proteins. It is also possible that the chosen antibody may detect only certain forms of a protein that do not correspond to the cDNA spotted on the DNA array, because of the alternative splicings of mRNA for example. Finally, distinct areas of a heterogeneous tumour may be submitted to RNA and protein analyses.

The brighter side of results described in this paper is an excellent correlation between RNA and protein levels in one-third of the tested molecules; among them: ERB-B2, BCL-2 and ER.

To summarize, the correlation between these techniques was shown in one-third of the selected markers and the absence of correlation in the other two-thirds. If protein levels of a target molecule, or a group of molecules, correlate with its selection by cDNA array, IHC on TMA offers a powerful tool quickly to evaluate the clinical relevance of differentially expressed genes. Thus, it is critical to determine to which extent changes in mRNA expression are accompanied by similar changes at the protein level. One very interesting observation was that the level of mRNA, but not protein expression levels of the THSB1 gene had prognostic value. The explanation for this phenomenon is beyond the scope of this paper. Even if the intrinsic prognostic power of the cDNA array data and clustering analyses derives from the combined expression of several genes, and not from an individual gene, it may be interesting for routine clinical application to test each of these genes as a candidate marker and to determine how its expression may alone distinguish the tumour classes.

For validation studies that are under way, it is particularly important to bear in mind that differences between mRNA and protein expression levels are possible with respect to intensities and to prognostic relevance. These differences underline the complementarity or synergy between expression measurements from cDNA arrays and IHC on TMA. Also, the need for other high-throughput technologies such as cDNA arrays containing alternatively spliced transcripts, protein arrays, and *in situ* hybridizations on TMAs. The combination of these complementary approaches will accelerate even more the identification of new diagnostic and prognostic markers as well as new therapeutic targets. These will improve the diagnosis and management of patients.

Literature

1. Sridar V. Chittur. DNA Microarrays: Tools for the 21st Century Combinatorial Chemistry & High Throughput Screening, 2004;7,531-537.

2. Lee C-H, MacGregour P. Using microarrays to predict resistance to chemotherapy in cancer patients. Pharmacogenomics 2004;5:611-625.

3. Ginestier C, Charafe-Jauffret E, Bertucci F, Eisinger F, Geneix J, Bechlian D et al. Distinct and complementary information provided by use of tissue and DNA microarrays in the study of breast tumour markers. Am J Pathol 2002;161:1223-1233.

4. 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;286:531–537.