European Review, Vol. 10, No. 3, 389–408 (2002) © Academia Europaea, Printed in the United Kingdom

DNA microarrays: a bridge between genome sequence information and biological understanding

KEITH HARSHMAN and CARLOS MARTÍNEZ-A

Department of Immunology & Oncology, Centro Nacional de Biotecnologia/CSIC, UAM Campus de Cantoblanco, E-28049 Madrid Spain.
E-mail:harshman@cnb.uam.es

The development, refinement and increasingly widespread use of DNA microarrays have been important responses to the explosion of sequence information produced by genome science. The high sample densities possible with DNA microarrays, coupled with the complete or nearly complete genome sequences available for humans and model organisms, provide a powerful analytical method to measure both qualitative and quantitative variations in RNA and DNA. Principal among the applications of microarrays is the large-scale analysis of RNA expression, often referred to as *expression profiling*. The power of this application lies in its ability to determine the expression patterns of tens of thousands of genes in a single experiment. Additionally, the ability to detect DNA polymorphisms makes microarrays useful in studies designed to correlate DNA sequence variations with variations in phenotype. The unprecedented scale on which microarrays allow both experimentation and generation of results should make possible a more complete and comprehensive understanding of cells and cellular processes.

Introduction

Biology and biomedical research are experiencing a period of information growth that is extraordinary, even by contemporary standards. Driven primarily by the application of automated sequencing and computational methods to genome analysis, biologists now have access to the complete or nearly complete genome sequence of an increasingly long and diverse list of organisms, including Homo sapiens. As a result, longer and more complete lists of genes are available for each

https://doi.org/10.1017/S1062798702000303 Published online by Cambridge University Press

of the organisms being studied. Nonetheless, although the sequence information produced by the assorted genome sequencing projects aims to be comprehensive, in most cases the simple sequence of the genes contained within an organism's genome does not tell us very much about how genes function or how they cooperate in the genetic pathways that control the physiology of that organism. The challenge of the so-called post-genomic era is to make maximal use of the enormous amount of genome information in understanding the complex nature of living cells; in effect, to make its utility match or exceed its bulk. Functional genomics seeks to meet this challenge by building a bridge between the bulk of gene sequence and the utility of gene function.

The success of the genome sequencing project, itself made possible by technological improvement and innovation, necessitates further technological advances to exploit this enormous amount of sequence information efficiently. Functional genomics has responded with the development of a number of new tools and technologies, one of the most powerful of which is the DNA microarray. A DNA microarray is a very precise and high-density grid of single-stranded nucleic acid samples attached to a solid support. Typically, DNA microarrays are made with either synthetic oligonucleotides or PCR-generated fragments of cDNA clones (PCR is a technique for replicating specific pieces of DNA in quantity), such that each position in the grid contains many copies of the same DNA sequence. The utility of DNA in this format arises from the ability of single-stranded nucleic acids to hybridize with high specificity to a second strand containing the complementary sequence, thus forming double-stranded nucleic acid molecules. Because of this, DNA microarrays can be used to 'interrogate' complex mixtures of thousands of nucleic acids for both the presence and abundance of molecules of known sequence. The ability to interrogate complex mixtures of nucleic acids both qualitatively and quantitatively makes the DNA microarray an extremely useful technology with a broad range of applications. To date, the most widespread uses of this technology have been in the analysis of gene expression and, to a lesser degree, the analysis of DNA variation.

The identification of gene expression changes induced by a particular stimulus or associated with a change in cell state has long been a central strategy used for the elucidation and description of physiological pathways. Until recently, analysis of gene expression in any single experiment has been limited to, at most, a few tens of genes, but with DNA microarrays it is now possible to monitor simultaneously, in a rapid and straightforward manner, the expression levels of tens of thousands of different genes.[1,4] Additionally, the nature of microarray data sets allows genes to be grouped according to similarities in their expression patterns. Experience has shown that these groups normally contain genes of both known and unknown function, and that these groups of co-expressed genes often contribute, to a very large extent, to the same cellular function. This has important

consequences: first, functions can be ascribed to genes of previously unknown function; second, new functions can be ascribed to previously characterized genes; and third, association of known genes to physiological processes may uncover previously unknown aspects of these processes.

Variations in genome sequence and structure, that is, DNA variations, are responsible for a large part of the variation among individuals. Knowledge of the DNA variation that exists between individuals and correlation of this information with disease phenotype have played an important role in understanding the link between genes and human health. Thus far, the disease phenotypes studied have been relatively rare, affecting relatively few people. However, using the techniques and information being generated by the fields of genomics, human genetics, and functional genomics, DNA microarray technologies provide a high throughput format to identify the genes and describe the mechanisms that control the development of common forms of disease The ability afforded by DNA microarrays to measure both the differential expression of genes and DNA variation systematically on a genome-wide scale provides an exciting opportunity to investigate the gene functions and genetic pathways that control the physiology underlying both normal and diseased states. The versatility of the technology makes a review of all its applications beyond the scope of this article. Instead, we will focus our discussion primarily on expression profiling and cite some specific examples of how microarrays are being applied to tumour classification.

Microarray technology

DNA microarrays are the latest step in the evolution of a group of related bio-analytical techniques first developed in the 1960s and 1970s.[2] These techniques – new and old – all make use of the ability of single-stranded nucleic acid molecules to hybridize with molecules of complementary sequence to form a stable double-stranded duplex DNA molecule. Of particular relevance to the development of DNA microarrays are those techniques designed to analyse both DNA and RNA in which DNA clones were immobilized on solid support, initially nitrocellulose membranes.[3] The primary physical difference between today's microarrays and their forerunners of the 1970s and 1980s is the use of an impermeable solid support, usually glass, rather than nitrocellulose or nylon membranes on which to attach the nucleic acid. A rigid support has a number of practical advantages over a flexible membrane, since the resulting array is more uniform, can achieve much higher sample density, and can be used in assays that require less time and incorporate improved, usually fluorescence-based, detection technologies.

Three main formats of DNA microarray are currently in common use, each with advantages and disadvantages in terms of cost, flexibility, versatility and ease of set-up (see Figure 1). Two formats rely on the in situ synthesis of oligonucleotide probes whose sequence is taken from the sequence of known genes.[4] The first of these methods makes use of photolithographic methods adapted from the microelectronics industry to synthesize oligonucleotides up to 30 bases in length in situ on a glass surface. This method allows the fabrication of DNA microarrays of extremely high densities – hundreds of thousands of different oligonucleotides per cm^2. Because of the relatively short length of the oligonucleotides produced by this method, however, multiple probes are needed per gene investigated. Therefore, the density of interrogated genes on these arrays is currently in the order of 15 000 to 20 000 cm^{-2}. The second in situ method is conceptually similar to the first, in that oligonucleotides of predefined sequence are synthesized stepwise on a glass surface. This method uses inkjet-based technologies (similar to those used in many commercial colour printers) to construct the arrays. Densities on these microarrays currently fall in the range of 1700 to 2000 genes cm^{-2}. The third format uses high precision robotic devices to transfer prepared nucleic acid samples mechanically onto the solid support, usually a chemically treated glass microscope slide. The actual transfer to the slide is most often accomplished by the physical contact of an array of pins similar in shape and design to very small fountain pens that have been previously primed with the nucleic acid samples to be transferred. Less often, inkjet or piezoelectric spray nozzles, which do not physically contact the slide surface, are used to deliver the sample to the slide.[1] With mechanical transfer methodologies, the nucleic acid sample can be any of a range of types, but is usually a PCR product generated from a cDNA clone or, less often, an oligonucleotide whose design is based on known gene sequences. Sample densities on these microarrays are in the range of 1700 to 2000 genes cm^{-2}.

Regardless of the fabrication method, the result is the same: a very precise, high-density grid of single-stranded nucleic samples attached to a solid support. Key to their successful use is the fact that the identity of each nucleic acid sample at each position on the microarray grid is known to the investigator.

All of the microarray formats in widespread use today require that the nucleic acid sample to be 'interrogated' be labelled in some way to allow hybridization detection. By far the most common labelling method makes use of fluorescent dyes, either directly incorporated or attached to the interrogee sample. When a solution containing nucleic acids labelled in this way is placed onto a microarray, molecules that are complementary in sequence to probes present on the microarray will bind to those positions. The intensity of the resulting fluorescent signal at any site on the array reflects the abundance of the complementary sequence in the target sample. Because the detection is fluorescence based, the method is quite

sensitive; individual mRNA species can be detected at a threshold of roughly 1 : 100 000. In addition, the dynamic range of the detection system is large, and both strong and weak hybridization signals can therefore be detected in the same experiment. Diverse informatic tools for data acquisition, display and analysis have been developed and implemented for this type of microarrays data.

The use of DNA microarrays in the study of RNA variation

Like filter-based methods, DNA microarrays provide a convenient means to measure quantitative and qualitative variations in RNA and DNA;[5] as such, they provide a convenient format for a wide variety of different analytical techniques.

Gene expression profiling

The most widespread application of this technology is currently the comparative analysis of RNA expression. Nearly all physiological and pathophysiological functions and events – cell growth, homeostasis, differentiation and death – are in very large part, determined by the protein components of a cell. A key insight of modern molecular biology, which has been consistently supported and rigorously validated over 30 years of investigation, is that there is a normally very tight association between changes in RNA expression levels, changes in protein expression levels and changes in cell state. Still, if a cell's physiology is controlled principally by its proteins, why measure differential RNA expression? Would it not be better to measure protein levels directly? In fact, many methods have been developed for direct measurement of protein expression.[6] But compared with nucleic acid-based methods, current protein-based methods for the analysis of differential gene expression are laborious, typically have lower sensitivity, and generally have lower sample throughput capabilities. The analysis of changes in mRNA expression levels induced by a particular stimulus is therefore the central strategy used today for the elucidation and description of physiological pathways.

DNA microarrays with complete catalogues of genes make it possible to monitor gene expression changes in a comprehensive way. Microarrays have been made containing all the genes of a number of prokaryotes as well as the simple eukaryote *Saccharomyces cerevisiae*, and the time is quickly approaching when all human genes will be present on one or two microarrays. Analyses with complete gene sets generate broader and much less biased views of the expression changes associated with cellular responses. When technical limitations force one to limit the scope of an experiment, the focus usually centres on those genes that, based on previous knowledge, are most likely to be involved in the process under study. By removing the limit on the number of genes analysed, the need to limit

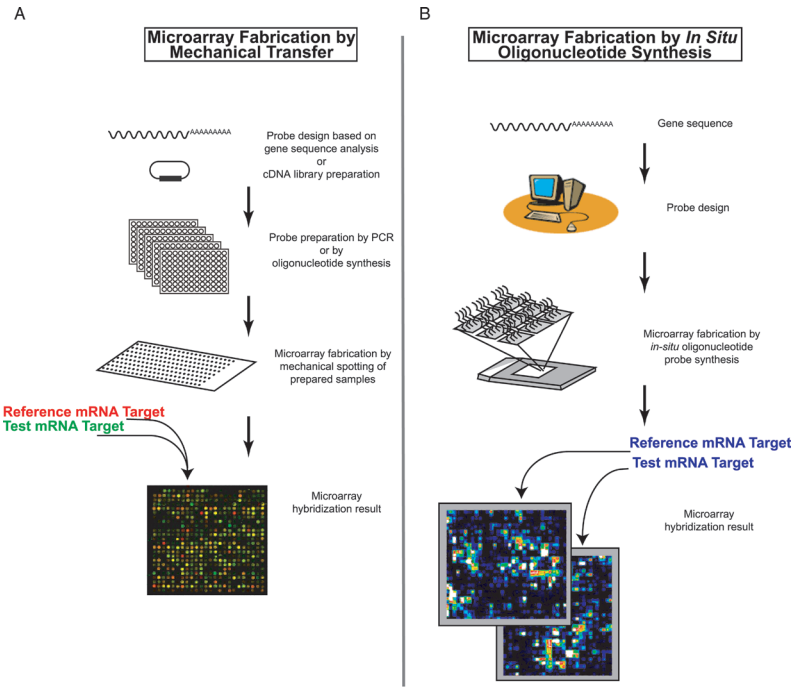

Figure 1

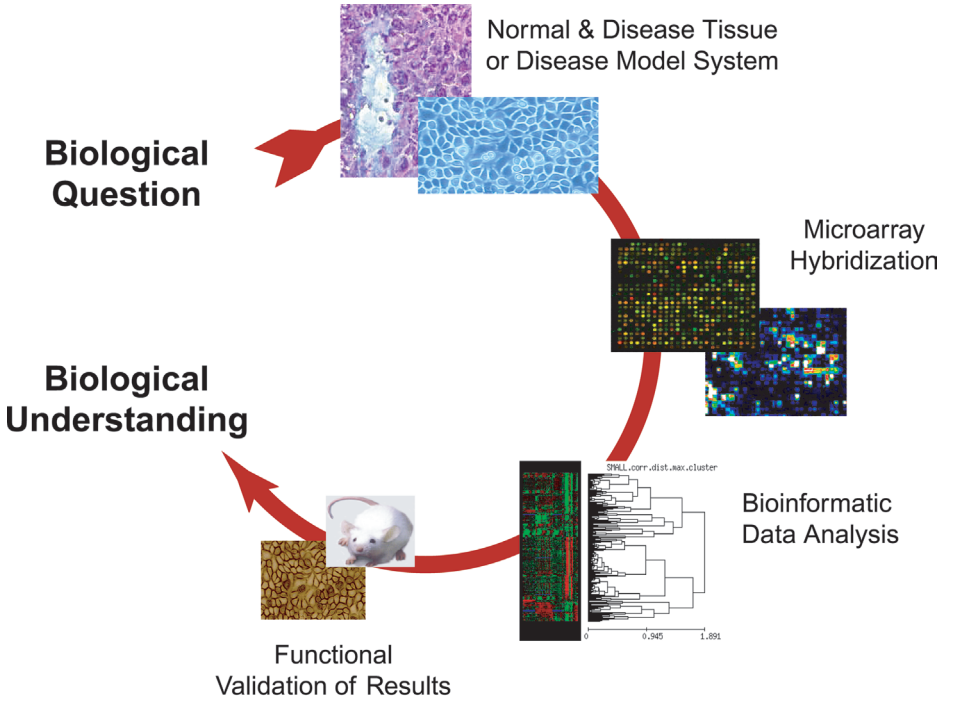

Figure 2

https://doi.org/10.1017/S1062798702000303 Published online by Cambridge University Press

Figure 1. Microarray fabrication and use workflow. (A) Fabrication by mechanical transfer. The nucleic acid to be spotted onto the microarray is prepared by either (1) synthesizing oligonucleotides whose design is based on known gene sequences, or (2) PCR amplifying inserts from cDNA clone gene libraries. The samples are then transferred to chemically treated glass microscope slides using high-precision arraying robots[24]. The use of chemically treated glass, post-transfer processing steps and, in some cases, probe chemical linkers result in the covalent attachment of the probe to the microarray. Target mRNA is isolated from the Test and Reference tissues or cell populations, labelled with different fluorescent dyes, and co-hybridized to the microarray surface. The hybridization reaction allows competitive binding between individual mRNAs in the differentially labelled target samples at positions on the microarray with the corresponding gene sequence. High-resolution fluorescence scanning of the microarray at two different wavelengths corresponding to the labelling dyes yields relative fluorescence signal intensities. The ratio of the two signals at a given gene position reflects the relative abundances of the corresponding mRNAs in the Test and Reference samples. (B) Fabrication by in situ synthesis. Oligonucleotide probe sequences are designed based upon known gene sequences. Photolithographic or inkjet technologies are used for the stepwise in situ oligo synthesis directly on the glass surface. Reference and Test mRNA are isolated, labelled and hybridized separately to different microarrays. Following fluorescent scanning of the two arrays, the signal intensities are used to calculate the relative mRNA abundances in the Test and Reference samples for the genes represented on the microarray.

Figure 2. mRNA expression profiling workflow – answering biological questions. Careful definition of the biological question is fundamental to designing effective expression profiling experiment. Following definition of the question, RNA is extracted from cell populations or tissues that constitute the experimental system. The RNAs are fluorescently labelled and either hybridized individually to two arrays or co-hybridized to a single microarray. Following scanning and quantification of data, the results are analysed using statistical methods that cluster genes into groups with similar expression profiles. From these results, predictions are made regarding the involvement of one or more genes in the biological process under investigation. Formal proof of the involvement normally requires the result be validated using an alternative experimental technique.

https://doi.org/10.1017/S1062798702000303 Published online by Cambridge University Press

the investigation to the 'usual suspects' is thus also removed. As a consequence, surprising results, often critical to gaining new insight and understanding, are more likely to be obtained.

Guilt by association

We are ignorant of the function of large numbers of genes and for many others our knowledge is incomplete. Expression profiling offers a straightforward way to make predictions about gene functions that later can be confirmed experimentally (Figure 2). This method relies upon the observation that genes with similar expression patterns are very often functionally related or participate in the same process or pathway. This approach, sometimes referred to as 'guilt by association', relies heavily on statistical methods to detect similarities in expression profiles and using those similarities to group genes into clusters with similar expression behaviour. As the predicted function is often based only on a statistical evaluation of the expression data, validation by some other experimental method is usually needed to prove the functional association.

Candidate gene expression and 'molecular phenotypes'

A related expression profiling application of DNA microarrays reverses the emphasis of the gene and pathway discovery approach; instead of measuring gene expression changes and correlating them with a change in cell phenotype or state, the phenotype change is correlated with the intentional and controlled overexpression or suppressed expression of a specific gene. Again, this is a method of gene functional classification long used in molecular biology: overexpress or suppress a specific candidate gene and observe the change in cellular phenotype. The difference in the microarray approach is that the phenotype is not revealed, for example, as a morphological change. Instead it consists of expression changes in specific genes or sets of genes. These expression changes constitute a 'molecular phenotype' that can implicate the candidate gene in specific cellular pathways or processes. This approach has been especially useful in identifying the downstream targets of transcription factors.

Tumour profiling

A third application of expression profiling, and one that has generated some of the most intense interest and excitement, is in the analysis and classification of human disease. Within this application, the most active area of investigation has been the study of cancer. The traditional histopathological approach to tumour classification has used a mixture of morphological, immuno-histochemical, and clinical criteria to classify malignancies. Despite significant progress, these methods often fail to predict accurately the clinical course of many tumours as

well as the response to, and effectiveness of, treatment. It was recognized early on that gene expression patterns determined using DNA microarrays could provide a potential means for classifying tumours into more biologically meaningful and clinically useful categories. Although the clinical utility of expression profile-based tumour classification remains controversial and is nowhere near universally accepted, we present here some early results that give cause for optimism.

One of the first major attempts at tumour classification using microarray expression profile data was with acute leukaemias.[7] These leukaemias, which arise from either lymphoid precursors (acute lymphoid leukaemias; ALL) or myeloid precursors (acute myeloid leukaemias; AML) are typically classified based on a combination of morphological, histochemical and cytogenetic criteria. A procedure was developed in which expression profiles for the different leukaemia samples were sorted by their degree of correlation with known AML-ALL clinical and histopathological differences. The investigators then selected the 50 genes whose expression most closely correlated with these ALL-AML distinctions. Using these genes, a class prediction algorithm was developed and applied to the original leukaemia samples; this algorithm correctly identified more than 90% of samples as either AML or ALL. Furthermore, when applied to a new group of 34 leukaemia samples, the class prediction algorithm correctly classified 85% of the samples. This study clearly demonstrated the ability of expression profiling with microarrays to differentiate between cancer subtypes and to define groups of genes that accurately and reproducibly classify tumours into those subtypes.

In an attempt to identify and diagnosis haematologic cancer subtypes for which there are no morphological means of classification, microarrays have been used to classify the most common subtype of non-Hodgkin's lymphoma, diffuse large B-cell lymphomas (DLBCL).[8] Subtype classification would be particularly useful for DLBCL, as patients with this lymphoma have highly variable clinical courses more than half fail to achieve long-term survival after receiving standard therapeutic regimes. In these studies, the DLBCL cases were subdivided into two groups based upon differences in gene expression patterns. Interestingly, the expression profiles grouped tumours of patients with similar clinical courses of disease, indicating that expression profiling could be used to predict clinical outcome. This example not only illustrates the potential power of expression profiling to supplement and extend the histopathological methods in standard use for tumour classification, but also its potential as a clinically relevant tool.

Expression profiling has also been used to classify a number of solid tumour types. One of the first and most thorough examples in which the technique was tested was in a study of breast tumours.[9] Breast tumour biopsies tend to be heterogeneous in nature, made up of a mixture of cell types including normal and malignant cells, stromal cells, infiltrating inflammatory cells, as well as cells at

various stages of necrosis, creating an even greater challenge for expression-profiling based classification (see Figure 3 for a discussion of sample preparations methods). Nevertheless, although the tumours in this study showed great variability in their expression patterns, they could be classified into four different subtypes. The differences in expression characteristics of these subtypes appear to reflect differences in known aspects of mammary epithelial cell biology. Another notable feature of this study was the inclusion of tumour sample pairs in the analysis; biopsies taken before and after a four-month chemotherapy regime, which showed that 75% of the 'before and after' pairs clustered together; that is, the two members of each pair were most often more similar to each other than to any other tumour sample. This indicates that both the reproducibility of the microarray analysis technique as well as the stability of each tumour's unique expression profile are quite high.

This initial 'proof-of-principle' has been followed by studies designed to test the predictive capabilities of breast tumour expression profiling more rigorously. In the first of these studies, classification of estrogen receptor (ER) status as well as lymph node status, two clinical parameters that bear significantly on the choice of treatment, were determined using expression profile data.[10] This is important, as the standard diagnostic assays and procedures for ER and lymph node status are subject to error and, in the case of lymph node status, require a significant degree of surgical intervention. Clearly, any technique that could improve diagnostic and prognostic accuracy and decrease the need for invasive intervention would improve cancer treatment. The goal of the second study was to develop an expression-profile based method to classify breast cancer patients into groups likely or unlikely to need chemo or hormonal therapy following tumour removal.[11] Even though these therapies reduce the risk of developing metastases, the majority of women who receive them would not develop metastases in the absence of the treatment, and the patients who do not benefit from the treatments nonetheless suffer their toxic side effects. It would be highly beneficial to be able to identify those patients whose tumours are unlikely to metastasise and who are therefore unlikely to benefit from chemotherapy. This study first analysed tumours from a group of women for whom the clinical outcome progression to metastasis or not was known. Expression profile data were used to select a group of genes whose behaviour 'predicted' the clinical outcome. When this 'prognosis-classifier' was applied to a similar but independent set of breast cancer patients and tumours, it correctly predicted nearly 90% of the clinical outcomes.

These and other successful applications of expression profiling to tumour classification have led some to suggest that this technique may eventually replace more conventional tumour diagnostic techniques.[12] Others have cautioned restraint in such predictions, as the studies to date have been relatively few in

number and small in scale.[13] Still, these early studies indicate that expression profiling of a relatively small number of genes can provide a molecular means of identifying clinically important tumor subtypes not identified using standard methods, and that these subtypes may identify specific subgroups of patients that will benefit from distinct treatment regimes. Carefully controlled, large-scale expression profiling studies on large numbers of clinically well-described tumour samples are needed before the true utility of this technique can be judged accurately. However, regardless of potential clinical applications, it is nonetheless clear that the results of expression profiling studies with tumours will add to our understanding of the genes and genetic pathways that underlie mechanisms controlling normal and disease pathology.

Bioinformatics

Biology has not traditionally been considered a science that generates, manipulates and analyses large data sets. This began to change with the rise of genomics and the associated large-scale sequencing projects and has continued and accelerated with the birth and growth of functional genomics. Each of the expression profiling examples described above produced massive amounts of data that needed to be stored, explored, analysed and interpreted in the context of other sources of biological information. It has become clear that bioinformatics and computer science are critical components in the design and successful implementation of any DNA microarray study.

A number of different statistical methods have been used to detect patterns in expression profiling data and to cluster genes accordingly.[14] Although effective, no one would argue that these methods cannot be further refined and modified nor that there are not useful approaches still untried. There is also an obvious need for methods to integrate microarray data and results with other collections or databases of biological and medical information. With this integration will certainly come more complete and more profound insights into the relationships and mechanisms that control physiological processes. This drive to extract and integrate data and information is not unique to DNA microarray-based investigation. Indeed, it is one of the defining characteristics of functional genomics and is forcing molecular biologists to consider issues formerly thought the concern only of physicists, mathematicians and computer scientists.

The use of DNA microarrays in the study of DNA variation

Identifying the genes that control phenotypic traits

Variations in genomic DNA sequence and structure are frequent and contribute in important ways to many biological processes, including disease development.

https://doi.org/10.1017/S1062798702000303 Published online by Cambridge University Press

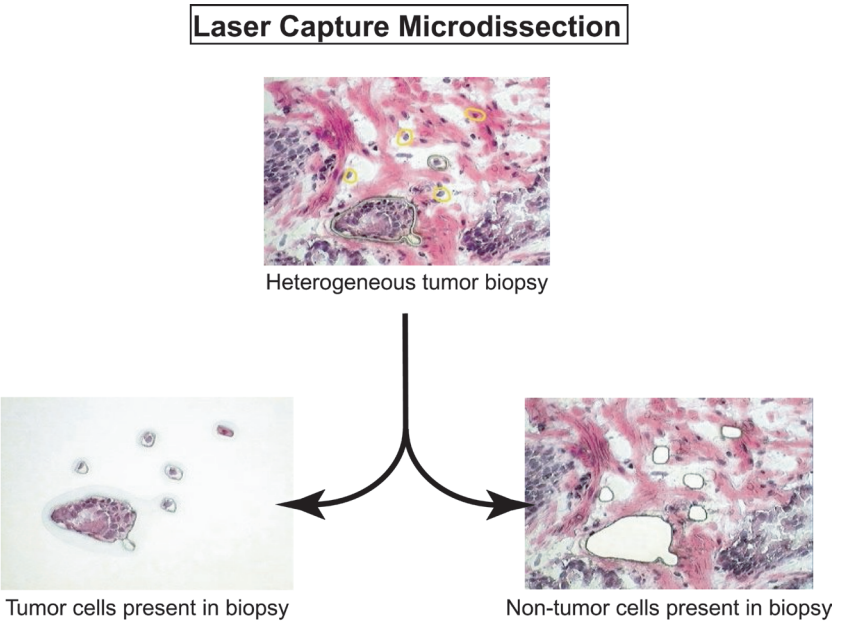

Figure 3

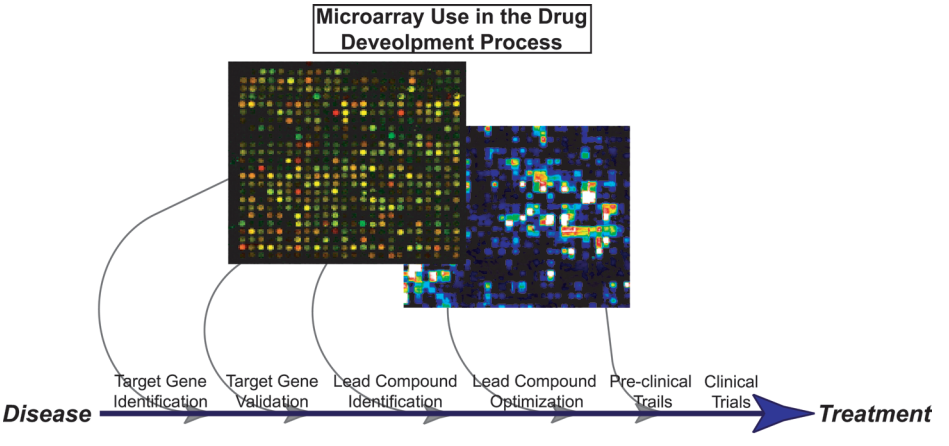

Figure 4

https://doi.org/10.1017/S1062798702000303 Published online by Cambridge University Press

Figure 3. Sample preparation. Solid tumours are rarely homogeneous; they are usually a mixture of cancer and normal cells, infiltrating inflammatory cells, and cells undergoing necrosis. This cellular heterogeneity presents problems and special challenges for expression profiling experiments that are not encountered with blood-based cancer samples. Because of this heterogeneity, the expression profiles of most solid tumours are a composite of the various cell types present in the biopsy sample. A number of strategies have been developed to circumvent this problem. The most straightforward of these is the isolation of pure or near-pure tumour cell samples from the biopsy sample, physically separating them from neighbouring non-tumour cells by gross dissection. A more sophisticated, thorough and precise method is Laser Capture Microdissection (LCM).[25] In LCM, a specially equipped microscope is used to identify visually, using morphological criteria, the cancer cells of interest in tissue sections obtained from the heterogeneous tumour biopsy. Once identified, tumour cells are dissected from the surrounding cells using a laser integrated into microscope optics. Following dissection, the selected cells can be procured by any of a number of mechanical transfer methods or using the laser itself to catapult the sample into a collection tube. Although very effective at isolating homogeneous cell populations from heterogeneous biopsy samples, LCM and related techniques suffer from being able to process only very limited numbers of cells – very often fewer than are required to perform a standard expression profiling experiment. In response to this limitation, methods are being developed to amplify the expression-profiling signal generated from small quantities of cells in an unbiased way.

Figure 4. Applications of DNA microarray technology in the drug development process. Listed above the bold arrow are the major steps in the drug development process. There are important applications of DNA microarrays in nearly every one of these steps. Expression profiling methods are useful in the target gene identification step as well as in the target gene validation, lead compound identification and lead compound optimization steps where they can serve to assay molecular phenotypes; microarrays can be used for SNP determination in human genetic studies designed to identify disease causing genes (target gene identification) or in pharmacogenetic studies designed to stratify patients into groups that are most likely to respond well to the drug treatment (pre-clinical trials).

Indeed, much of what is referred to as the 'genetic basis of disease' is the result of quantitative variations in genome sequence and structure. The size of the variations ranges from small changes, such as insertions, deletions or sequence changes of as little as a single base, to huge ones, including amplifications and deletions of entire chromosomes or chromosomal regions. A wide variety of DNA hybridization-based methods have been developed to identify and quantify these variations. However, assay throughput is often a severely limiting factor. In response, a number of different assays designed to measure DNA variation have been adapted to the microarray format, they differ in many aspects but have in common a greatly increased throughput capability compared to their non-array-based equivalent.

The principal application of microarray technology in this area has been the detection of single nucleotide polymorphism (SNP). SNPs are single-base differences in a DNA sequence that can be observed between individuals in the population and which are found in the human genome with an average frequency of 1 per 1000 base pairs. Because of this relatively high frequency, they are useful in genetic studies designed to identify the genes that underlie and influence common phenotypic traits, including the development and susceptibility to many forms of human disease.[15] Identification and characterization of these genes and their disease-causing variants are likely to facilitate the development of more effective therapeutics, as well as genetic tests designed to identify individuals with a high risk of developing disease. In a related application, SNPs are used to study the variable response individuals often show to drugs. The field of pharmacogenetics aims to explain and understand the variability in drug responses attributed to human genetic variability.[16] It is hoped that this information can be used to classify individuals into groups that are likely to respond positively or negatively to treatment with a particular drug. With this information, therapeutic regimes can be individually chosen that maximize treatment efficacy and minimize the adverse side effects associated with the drugs.

Mutation and polymorphism detection

A significant portion of the mutations that change or inactivate the normal function of many genes are nothing more than small variations – insertions, deletions and base changes – in the wild type sequence of those genes. The same microarray-based methods used in detecting SNPs for genetic analysis can thus also be applied to screening for mutation causing gene changes. As the understanding of human disease genetics improves and more disease-causing gene mutations are identified, this application of microarray technology is likely to increase in importance.

Comparative genome hybridization

Genome-wide scans of amplified and deleted chromosomal regions by comparative genome hybridization (CGH) are useful in identifying amplified and deleted regions in tumour DNA. These regions are of interest since they often contain oncogenes or tumour suppresser genes whose amplification or deletion can be significant steps in the process of tumourigenesis.[17] Furthermore, characterization of these regions can be clinically useful in tumour diagnosis, as specific chromosomal amplifications and/or deletions are often associated with specific tumour subtypes. The DNA microarray-based CGH procedure is more sensitive and gives higher resolution results than the standard CGH procedures.

The use of DNA microarrays in drug discovery and development

The development of safe, effective drugs is typically a prolonged and expensive process consisting of many steps. The process begins with the identification of proteins or 'targets' likely to play an important role in disease development. Once identified, the relevance of the protein in the development of the disease is established in a process called *target validation*. Once the protein's role in disease development has been well established, chemical compounds and/or proteins can be screened to find drugs that interact with the target protein and thereby halt or delay disease progression. Following its identification and prior to entering clinical trials, a lead compound invariably passes through an extensive process of modification to optimize and improve a range of characteristics that increase its efficacy and decrease its possible side effects. Finally, compounds that pass through this optimization process can successfully be considered candidates for pre-clinical trials and, ultimately, clinical trials. Considering their power and versatility, it is not surprising that DNA microarrays have found a number of important applications in drug development processes (see Figure 4 and Ref. 18). Their use in the identification of the genes and pathways associated with the normal cellular processes, as well as the development and progression of disease, is providing many new drug targets. Genes identified in these studies will include not only those that control normal and disease physiology, but also genes controlling disease symptoms. Different classes of drugs are thus likely to originate from these studies. As described above, specific expression profiles associated with well-defined disease phenotypes can be considered as molecular phenotypes that can be used to replace the physiological endpoint in many of the screens and assays associated with the drug development process. The use of such surrogate phenotypes can greatly simplify and speed the steps of target validation, lead identification and lead optimization, particularly when the phenotype screen being replaced requires the use of animal model systems. Finally, in pre-clinical and clinical trials, SNP detection with microarrays can be combined with

https://doi.org/10.1017/S1062798702000303 Published online by Cambridge University Press

pharmacogenetic data to segment or stratify patient populations into likely responders, non-responders and adverse reactants to any particular drug. Identifying and removing from clinical trials individuals likely to not respond or to respond adversely will improve the safety and cost-effectiveness of these final stages of the drug development process.

An expanding range of applications

A sign of the versatility of the DNA microarray format is the varied list of applications to which it has been applied in studying nucleic acid variations. The list is long and, although these applications are less widely used than RNA expression profiling or SNP analysis, they have still contributed significantly to our increased understanding of biology and physiology. A very abbreviated list includes: the use of microarrays to show that variability in the effectiveness of various tuberculosis vaccines is likely the result of variable genomic deletions that arise in the *Mycobacterium* strain used to generate the vaccine;[19] using microarrays to select effective antisense oligonucleotides to inhibit gene expression;[20] determining the DNA binding sites of protein whose functions involve physical interaction with the genome;[21] characterizing the patterns and dynamics of chromosomal replication in eukaryotes;[22] and using microarrays to detect directly (i.e. without the use of PCR) the presence of specific environmentally important microbes in soil and sediment samples.[23] The list of applications is sure to increase and diversify as the technology is evolving and improving at a rapid rate. Additionally, it is now becoming accessible to an increasing large and diverse group of researchers. Therefore, the number and breadth of DNA microarray applications will certainly increase, as will its importance as a research and diagnostic tool.

Conclusions and perspectives

Over the past decade, genome sequencing has produced an enormous amount of sequence information about humans, agriculturally important plants, animal and plant model systems, as well as a number of important micro-organisms. In some ways, this information can be considered as somewhat crude or limited in usefulness – a genetic 'parts list' that provides relatively little insight into the functions of genes or how they cooperate in the gene networks that control cellular physiology. Functional genomics is providing both the tools and the intellectual infrastructure to refine genome sequence information into an understanding of gene and gene network function, organization and coordination.

The DNA microarray is one of the most powerful of the new functional genomic tools. The examples we discussed here provide an idea of the potential this technology has to influence dramatically biomedical science and clinical practice.

Yet, even as its importance in understanding gene function is increasing, the technique is quickly evolving and improving. Rapid technical advances are expected in the following areas:

(1) Fabrication techniques to increase sample density. This will include improvements in technology already in use as well as the use of new formats and technologies for microarray fabrication. Single DNA microarrays containing every human gene should soon be available.
(2) Increased sensitivity of detection. The desire to do expression profiling experiments using a small number of cells (e.g. clinical biopsies) or even a single cell has resulted in the RNA amplification and nucleic acid labelling techniques being in a state of constant refinement and improvement. While these improvements are being made, alternative detection technologies are being evaluated and, when appropriate, adapted to the microarray format. Included among these detection technologies are atomic force microscopy, scanning optical microscopy, electrochemical detection and mass spectrometry.
(3) Bioinformatic methods to extract information from microarray data sets and to integrate that information with biological and medical knowledge databases. In particular, robust automated information extraction methods capable of retrieving from the published literature information about genes that have been clustered according to expression profiles should greatly assist in the functional classification of uncharacterized genes.

The DNA microarray is not the only technology to result from the rise of functional genomics and its goal to exploit and interpret genome sequence data. Proteomics, the large-scale analysis of proteins, is poised to overcome the technical problems that have so far limited its usefulness for gene expression analysis in comparison to nucleic based-methods.[6] Existing proteomics technologies, such as mass spectrometric analysis of proteins separated by two-dimensional electrophoresis, are being improved to fit functional genomic applications. Additionally, new proteomic technologies are being developed, some of which use the microarrays format to produce arrays of peptides, proteins and antibodies. Finally, other standard methods of biological investigation are being adapted to fit the functional genomics approach of classification based on high throughput, comprehensive, whole genome-scale analysis. Some of these include: classification based on computational methods that determine sequence conservation across, and similarities between, species; large-scale structural studies that aim to classify unknown protein function based upon structural similarities with proteins of known function; and systematic mutational

https://doi.org/10.1017/S1062798702000303 Published online by Cambridge University Press

inactivation of single genes coupled with characterization of those mutants to determine gene function.

The comprehensive sequences produced by genome science have produced a unique set of opportunities and challenges to which the response is functional genomics. A range of powerful analytical and experimental tools are being used to ask and answer questions about the complex nature of living cells. The drive now is to extract the maximum amount of information using these tools and to integrate that information as completely as possible with as many other types and sources of information as possible. The result should be a profound understanding of the function of genes and gene networks that regulate cellular processes and how they give rise to the collective properties of whole organisms. Improved understanding of normal and aberrant cellular processes will, in turn, lead to a much enhanced capacity to prevent, treat and cure disease.

Acknowledgements

We are grateful to Dr Luis López-Fernández and Catherine Mark for critical reading of the manuscript and to Drs Jose Alberto Garcia-Sanz and Inmaculada Segura for help preparing the illustrations. The Department of Immunology and Oncology was founded and is supported by the Spanish National Research Council (CSIC) and by the Pharmacia Corporation.

References

1. D. J. Duggan, M. Bittner, Y. Chen, P. Meltzer and J. M. Trent (1999) Expression profiling using cDNA microarrays. *Nature Genetics*, **21**, 10–14.
2. E. M. Southern (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. *Journal of Molecular Biology*, **98**, 503–517.
3. G. A. Beltz, K. A. Jacobs, T. H. Eickbush, P. T. Cherbas and F. C. Kafatos (1983) Isolation of multigene families and determinations of homologies by filter hybridization methods. *Methods in Enzymology*, **100**, 266–285.
4. R. J. Lipshutz, S. P. Fodor, T. R. Gingeras and D. J. Lockhart (1999) High density synthetic oligonucleotide arrays. *Nature Genetics*, **21**, 20–24.
5. E. S. Lander (1999) Array of hope. *Nature Genetics*, **21**, 3–4.
6. A. Pandey and M. Mann (2000) Proteomics to study genes and genomes. *Nature*, **405**, 837–846.
7. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander (1999) Molecular classification of cancer: class discovery by gene expression monitoring. *Science*, **286**, 531–537.

8. A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, Jr, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown and L. M. Staudt (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.
9. C. M. Perou, T. Sorlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, O. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lonning, A. L. Borresen-Dale, P. O. Brown and D. Botstein (2000) Molecular portraits of human breast tumors. *Nature*, **406**, 747–752.
10. M. West, C. Blanchette, H. Dressman, E. Haung, S. Ishida, R. Spang, H. Zuzan, J. A. Olson, Jr, J. R. Marks and J. R. Nevins (2001) Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences (USA)*, **98**, 11462–11467.
11. L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. M. Hart, M. Mao, L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards and S. H. Friend (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
12. S. A. J. R. Apricio, C. Caldas and B. Ponder (2000) Does massively parallel transcriptome analysis signify the end of cancer histopathology as we know it? *Genome Biology*, **1**, 1021.1–1021.3
13. J. R. Master and S. R. Lakhani (2000) How diagnosis with microarrays can help cancer patients. *Nature*, **404**, 921.
14. Brazma and J. Vilo (2000) Gene expression data analysis. *FEBS Letters*, **480**, 17–24.
15. N. J. Risch (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
16. A. D. Roses (2001) Pharmacogenetics. *Human molecular genetics*, **10**, 2261–2267.
17. D. G. Albertson, B. Ylstra, R. Segraves, C. Collins, S. H. Dairkee, D. Kowbel, W. L. Kuo, J. W. Gray and D. Pinkel (2000) Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nature Genetics*, **25**, 144–146.
18. C. Debouck and P. N. Goodfellow (1999) DNA microarrays in drug discovery and development. *Nature Genetics*, **21**, 48–50.
19. M. A. Behr, M. A. Wilson, W. P. Gill, H. Salamon, G. K. Schoolnik, S. Rane and P. M. Small (1999) Comparative genomics of BGH vaccines by whole-genome DNA microarray. *Science*, **284**, 1520–1523.
20. M. Sohail, H. Hochegger, A. Klotzbucher, R. L. Guellec, T. Hunt and E. M. Southern (2001) Antisense oligonucleotides selected by hybridization to scanning arrays are effective reagents *in vivo*. *Nucleic Acids Research*, **29**, 2041–2051.

21. V. R. Iyer, C. E. Horak, C. S. Scafe, D. Botstein, M. Snyder and P. O. Brown (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature*, **409**, 533–538.
22. M. K. Raghuraman, E. A. Winzeler, D. Collingwood, S. Hunt, L. Wodicka, A. Conway, D. J. Lockhart, R. W. Davis, B. J. Brewer and W. L. Fangman (2001) Replication dynamics of the yeast genome. *Science*, **294**, 115–121.
23. J. Small, D. R. Call, F. J. Brockman, T. M. Straub and D. P. Chandler (2001) Direct detection of 16S rRNA in Soil Extracts by Using Oligonucleotide Microarrays. *Applied Environmental Microbiology*, **67**, 4708–4716.
24. D. J. Lockhart and E. A. Winzler (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.
25. K. Schutze and G. Lahr (1998) Identification of expressed genes by laser-mediated manipulation of single cells. *Nature Biotechnology*, **16**, 737–742.

About the Authors

Carlos Martínez-A is the Director of the Department of Immunology and Oncology at the Centro Nacional de Biotecnologia. He was previously at the Basel Institute for Immunology, the University of Umeå, the Clínica Puerta de Hierro and the Institut Pasteur in Paris.

Keith Harshman is a Scientist in the Department of Immunology and Oncology at the Centro Nacional de Biotecnologia. He was previously Director of Central Nervous System Disease Research at Myriad Genetics, Inc in Salt Lake City.