GENE EXPRESSION STUDIES USING MICROARRAYS: PRINCIPLES, PROBLEMS, AND PROSPECTS

David Murphy

University of Bristol Research Centre for Neuroendocrinology, Bristol Royal Infirmary, Bristol BS2 8HW, England

A number of mammalian genomes having been sequenced, an important next step is to catalog the expression patterns of all transcription units in health and disease by use of microarrays. Such discovery programs are crucial to our understanding of the gene networks that control developmental, physiological, and pathological processes. However, despite the excitement, the full promise of microarray technology has yet to be realized, as the superficial simplicity of the concept belies considerable problems. Microarray technology is very new; methodologies are still evolving, common standards have yet to be established, and many problems with experimental design and variability have still to be fully understood and overcome. This review will describe the time course of a microarray experiment—RNA isolation from sample, target preparation, hybridization to the microarray probe, data capture, and bioinformatic analysis. For each stage, the advantages and disadvantages of competing techniques are compared, and inherent sources of error are identified and discussed.


Key words: functional genomics; microarray; cDNA; oligonucleotide; bioinformatics

The sequencing projects that have elucidated the human genome (http://www.nature.com/cgi-taf/dynapage.taf?file=/nature/journal/v409/n6822/index.html, http://www.ncbi.nlm.nih.gov/genome/guide/human/, http://www.sciencemag.org/content/vol291/issue5507/, and Refs. 8, 52) and will soon reveal the genetic complement of key model vertebrate organisms such as the mouse (http://www.ncbi.nlm.nih.gov/genome/guide/mouse/index.html), rat (http://www.ncbi.nlm.nih.gov/genome/guide/R_norvegicus.html), puffer fish (http://fugu.hgmp.mrc.ac.uk/), and zebra fish (http://www.ncbi.nlm.nih.gov/genome/guide/D_rerio.html) must rank among the greatest achievements of human civilization. This is all the more so as these efforts have been truly international—the Human Genome Sequencing Consortium includes scientists at 16 institutions in France, Germany, Japan, China, the United Kingdom, and the United States. Furthermore, public and charitable funding has ensured that genome data are freely available to all, enabling information to be put to immediate use to the maximal possible benefit of all humankind.

Despite the magnificent efforts, the exact number of genes needed to make a human being is still not known. The computing technologies that recognize transcription units in raw genomic information are still being developed. Current estimates range from 28,000 to 120,000 (9, 17, 28, 33, 42, 52). Irrespective of the exact number, the challenge remains to determine exactly what all of these genes do in terms of the development and physiological functioning of the organism. This is the task of a newly emerging discipline—functional genomics. A crucial aspect of func-
tional genomics is the description of global expression patterns. In this regard, we need to address two questions:

- Where in the organism are all of these genes expressed?
- How do the expression patterns of these genes change as a consequence of physiological cues and disease?

For many years, molecular biologists tackled the analysis of one or a few genes at a time. Techniques relied on the use of nucleic acid (in situ hybridization, Northern blotting, RNase protection, etc.) or antibody (immunocytochemistry, Western blotting, etc.) probes. Descriptions of gene expression were a haphazard process, largely dependent on the opportunistic availability of these probes and the intuition of researchers. The availability of genome information and the parallel development of microarray technology have provided the means to perform global analyses of the expression of thousands of genes in a single assay (15, 27). The results provide an assessment of the expression levels of the genes included on the microarray in a particular cell, tissue or organ. The basic concept of microarray analysis is simple (Fig. 1). RNA is harvested from a cell type or tissue of interest and labeled to generate the target—the free nucleic acid sample whose identity or abundance is being detected. This is hybridized to the tethered probe DNA sequences corresponding to specific genes that have been affixed, in a known configuration, onto a solid matrix. Hybridization, based on Watson and Crick base pairing, between probe and target provides a quantitative measure of the abundance of a particular sequence in the target population. This information is captured digitally and subjected to various analyses to extract biological information. Comparison of hybridization patterns enables the identification of mRNAs that differ in abundance in two or more target samples (Fig. 1). Thus microarrays provide a powerful tool with which to screen biological specimens for alterations in the expression of mRNAs that accompany, and may regulate physiological and pathological change.

Before we consider in detail the many reasons for not carrying out a microarray study, we should iterate the
reasons why this technique has elicited so much excitement in recent years. Microarrays are a novel way of getting fast answers to obvious questions. Although microarray “discovery research” is not in itself hypothesis driven, rigorous but unbiased data gathering can narrow the search for genes that represent the best targets for future hypothesis-driven studies. Thus microarray techniques are finding utility in

- gene discovery—the global description of genes potentially involved in developmental, physiological, and pathological processes (7, 22)

- gene regulation—the description of regulatory networks, based on the assumption that genes regulated in parallel share common control mechanisms (2, 3, 53, 56)

- diagnosis—identification of patterns of gene expression that define disease states and that may represent prognostic indicators (49)

- drug discovery and toxicology (8, 25).

However, the superficial simplicity of the microarray concept belies considerable problems, and the full promise of the technology has yet to be realized. Microarray technology is very new; methodologies are still evolving, common standards have yet to be established, and many problems with experimental design and variability have still to be fully understood and overcome. Each stage in the course of a microarray experiment is thus a source of error (Fig. 1).

BIOLICAL SAMPLES

Microarray analysis has revealed a large degree of variability in gene expression patterns in particular tissues, even within supposedly genetically identical individuals. In addition, differences can be expected as a consequence of time of day, age, and physiological state (including sex and stage of the estrous cycle) and as a response to disease. Furthermore, variability can be introduced into a microarray experiment by way of sample heterogeneity and sampling error. The only way to overcome these difficulties is to

- standardize experimental conditions as far as possible

- pool multiple samples

- perform multiple replicates.

What Is Normal?

Before we can interpret microarray data related to physiological and pathological transitions, we need to answer the question, “What is normal?” For any system under study, we need to understand the magnitude and diversity of gene expression in the unperturbed state. Different strains of the same species have been shown to exhibit marked variability in the overall pattern of gene expression in the brain. Sandberg et al. (44) compared the expression profiles of more than 10,000 genes in six brain regions of two inbred mouse lines, C57Bl/6 and 129SvEv. Around 1% of genes were identified as being differentially expressed in at least one brain region. This is not unexpected, and indeed, these gene expression differences may represent a functional substrate for physiological and behavioral differences between strains. Similarly, it has been shown that gene expression patterns differ greatly between genetically identical individuals of the same age, sex, and physiological condition. Comparison of the expression profiles of 5,406 genes in a number of tissues of the C57Bl/6 inbred mouse strain revealed considerable variability, ranging from 0.8% in the liver to 3.3% in the kidney (40). It is accepted that, for any particular parameter, physiological “normalcy” is not a strict value but is rather a range of values presented by healthy individuals. It is perhaps no surprise that “normal” gene expression displays similar variability. Such variability between individuals emphasizes the need to pool samples, and to perform replicates.

Sample Collection

Ideally, a tissue sample from which target RNA is isolated must be pure. Although this might be possible for cultured cells, it is impossible for samples obtained from animals or from patient material due to the cellular heterogeneity of tissues. This is a problem compounded by human error and the variability inherent in tissue collection.

Whereas microarrays can be seen as a new and precise way to define cellular phenotype based on pat-
terns of gene expression, it follows that the ideal sampling methodology should enable the analysis of the cellular contents of single cells or of groups of selected cells (12, 14). The mRNA content of an individual cell can be isolated using a patch pipette to penetrate the cell membrane. The cellular contents are then aspirated into the pipette and transferred to a microcentrifuge tube for RNA isolation and target preparation. Samples can be obtained from randomly selected cells, from electrophysiologically characterized cells or from defined cell types in transgenic mice that specifically express reporters, such as enhanced green fluorescent protein (55). This level of analysis has now been taken to the subcellular level. Previous studies have revealed that a complex subset of mRNAs is present within the dendritic subdomain of neurons, where their local translation may contribute to synaptic plasticity (35). Microarray analysis has been used to screen for genes that are enriched in neuronal dendrites in response to various stimuli. A patch pipette was used to harvest individual dendrites and cell soma from primary rat hippocampal neuronal cultures treated with (RS)-3,5-dihydroxyphenylglycine, a metabotropic glutamate receptor agonist that modulates protein translation in dendrites. Targets derived from these samples were used to analyze microarrays, which revealed that a few mRNAs changed in abundance as a result of stimulation (12).

Another method for rapidly procuring pure, targeted, single or multiple cells from specific microscopic regions of tissue sections is laser capture microdissection (LCM) (6). A tissue sample is covered with a transparent plastic film and observed under the light microscope. The cells of interest having been identified, a focused infrared laser beam is activated. The heat of the beam melts the film, causing it to adhere to the targeted cells, which then can be lifted away, leaving the rest of the tissue section intact. Bonaventure et al. (i) have elegantly validated this approach by cataloguing gene expression profiles of seven rat brain nuclei or subnuclei, thus identifying putative specific markers of potential functional importance. Arcturus sells the industry standard LCM machine (http://www.arctur.com/about/technology/technology_lcm.htm).

**ARRAY PLATFORMS**

There are two microarray platforms in common use—cDNA microarrays, which utilize cloned probe molecules corresponding to characterized expressed sequences, and oligonucleotide microarrays, made of synthetic probe sequences based on database information (20).

**cDNA Microarrays**

A cDNA microarray comprises a collection of gene sequences [usually pure PCR products ranging in size from 100–2,000 bp derived from cDNA and expressed sequence tag (EST) clones] that are applied individually to precise locations on a solid matrix, usually nylon or glass. Nylon membranes have a number of advantages:

- they are usually hybridized to more sensitive radioactive tags
- no specialized equipment is required
- they can be reused up to four or five times.

Glass, too, has its advantages:

- DNA samples can be covalently attached to a glass matrix
- glass is a durable material that sustains high temperatures and washes of high ionic strength
- because glass is nonporous, the hybridization volume can be kept to a minimum, enhancing the kinetics of base pairing between probe and target
- glass has a low inherent fluorescence, so does not contribute greatly to background “noise.”

Two different targets derived from two samples to be compared can be labeled with different fluors and simultaneously hybridized with a glass microarray in a single competitive reaction. In contrast, nylon microarrays are generally probed in serial or parallel hybridization reactions.
Fabrication of cDNA microarrays. For both glass and membrane matrixes, each microarray element is generated by the deposition of a few nanoliters of purified PCR product at a concentration of 100–500 μg/ml. Spots are typically 100 μm in diameter and can be deposited at a density of up to 20,000 features/cm². Spotting is achieved by contact (mechanical microspotted) or noncontact (ink jetting) methods.

MECHANICAL MICROSPOTTING. A DNA sample is loaded into a spotting pin by capillary action, and a small volume is transferred to a solid surface by physical contact between the pin and the solid substrate. After the first spotting cycle, the pin is washed, and a second sample is loaded and deposited to an adjacent address. Robotic control systems and multiplexed print heads allow automated microarray fabrication (29, 46).

INK JETTING. A DNA sample is loaded into a miniature nozzle equipped with a piezoelectric fitting (or other form of propulsion), which is used to expel a precise amount of liquid from the jet onto the substrate. After the first jetting step, the nozzle is washed and a second sample is loaded and deposited to an adjacent address. A repeat series of cycles with multiple jets enables rapid microarray production (47).


Target labeling and hybridization of cDNA microarrays. Sample RNA is converted to target by use of the enzyme reverse transcriptase, an oncoretroviral enzyme that uses RNA as a template for the synthesis of a single-stranded cDNA. Reverse transcriptase requires a short primer to initiate cDNA synthesis, and this is usually provided by oligo(dT), which anneals to the poly(A) tail found at the 3' end of the vast majority of mammalian mRNAs. The label incorporated into the cDNA can be either radioactive or fluorescent. Radioactive target is generated by incorporation of [33P]dCTP, a relatively weak emitter that reduces interference between the closely physically juxtaposed microarray elements. Clearly, the use of a radioactive target requires that comparison of different targets must be carried out using serial hybridizations to the same microarray or by parallel analyses using separate microarrays.

An advantage of fluorescence detection is that competitive hybridization to the same microarray (usually glass, see above) can be used to compare targets derived from different samples. The relative hybridization of the targets labeled with different fluoros to the same probe can be readily quantified. The fluorescent labels Cy3-dUTP and Cy5-dUTP are frequently paired, as they have high incorporation efficiencies with reverse transcriptase and good photostability and yield and are widely separated in their excitation and emission spectra, allowing highly discriminating optical filtration. However, it should be noted that the different fluoros produce targets with different characteristics. Thus microarray experiments must either be repeated with the fluoros swapped around or be performed with the same fluor in parallel on different probes.

RNA purity is a critical factor in hybridization performance, particularly when fluorescence is used, as cellular protein, lipid, and carbohydrate can mediate significant nonspecific binding of labeled cDNAs to matrix surfaces.

A limitation of cDNA microarray technology is the large amount of RNA required to produce an adequate signal over noise (11). This is a particular issue with low-abundance transcripts. Fluorescence detection requires ≥10 μg of total RNA (equivalent to a million cells), whereas radioactive detection enables detection with as little as 0.1 μg of starting total RNA (10,000 cells). However, as described above, the ultimate aim is to carry out expression profiling with as few cells as possible, preferably single cells. For targets to be derived from such samples, some form of amplification process needs to be incorporated into the procedure. PCR (43) is a highly efficient method for exponentially amplifying a population of single-stranded cDNA. However, the nonlinear amplification results in a target in which sequence representation is skewed compared with the original mRNA pool. In
contrast, the amplified antisense RNA (aRNA) procedure (13, 50) is a linear procedure that produces a target more representative of the initial mRNA population. An mRNA sample is converted into cDNA using an oligo(dT) primer that contains a bacteriophage T7 RNA polymerase promoter site. After the cDNA is rendered double stranded, T7 RNA polymerase is used to transcribe antisense RNA copies. The procedure can produce up to $10^6$-fold amplification. Both Ambion (http://www.ambion.com/catalog/CatNum.php?1750) and Arcturus (http://www.arctur.com/products/riboamp_main.htm) sell linear amplification kits.

Data capture. Once targets have been hybridized to probes and the microarray has been washed to remove as much unbound and nonspecifically bound target as possible, the array must be scanned to determine how much target is bound to each probe spot. Data are captured from microarrays hybridized with $^{33}$P-labeled target by means of a phosphorimager system (e.g., the Molecular Dynamics Storm and Typhoon machines; http://www.mdyn.com). Microarrays hybridized with fluorescent targets are stimulated with a laser. The emitted light is then captured by either a charge-coupled device or a confocal scanner. A number of companies produce machines for scanning fluorescently labeled microarrays (http://ihome.cuhk.edu.hk/~b400559/array.html#Scanner and http://www.lab-on-a-chip.com/files/mascanner.pdf).

Advantages of cDNA microarrays. cDNA microarrays are a relatively accessible and cost-effective technology. Hybridization does not need specialized equipment, and data capture can be carried out using equipment that is very often already available in the laboratory. Prefabricated microarrays are relatively cheap, and custom chip manufacture is within the reach of many researchers, affording flexibility of design as necessitated by the scientific goals of the experiment.

The long target sequences (≥2 kbp) increases detection sensitivity.

Disadvantages of cDNA microarrays. Sequence homologies between clones representing different closely related members of the same gene family may result in a failure to specifically detect individual genes. However, closely related genes can often be distinguished by using probes corresponding to the 3'-untranslated region of an mRNA, as these regions often display gene-specific sequence diversity.

The state of the double-stranded DNA on the microarray is ill defined and may well have constraining contacts with the matrix and inter- and intrastrand cross-links that will affect hybridization.

Each sample must be synthesized, purified, and stored before microarray fabrication.

Microarray fabrication is dependent on the curation of extensive clone sets. Even the best maintained sets are prone to mix-ups, with clones not containing the sequence that they are supposed to. Halgren et al. (24) sequenced 1,189 IMAGE consortium cDNAs (http://image.llnl.gov/) obtained commercially from Research Genetics (http://www.resgen.com). Only 62.2% were uncontaminated and contained cDNA inserts that had significant sequence identity to published data for the ordered clones; 7.1% contained both a correct and an incorrect plasmid; and 5.9% contained multiple, distinct, incorrect plasmids, indicating the likelihood of multiple contaminating events. Through this kind of analysis will emerge systems that will enable the better curation of clone stocks.

Oligonucleotide Microarrays

Oligonucleotide microarrays are made by synthesizing single-stranded probes on the basis of sequence information in databases. A number of technologies are available (http://www.lab-on-a-chip.com/suppliers/inform.html). For example, oligonucleotide synthesis has been combined with ink jet spotting. Motorola Life Sciences (http://www.motorola.com/lifesciences) synthesize 30-mer oligonucleotides “offline” and spots them onto slides coated with a three-dimensional, branched polymeric substrate gel surface (Motorola Life Sciences recently sold their microarray business to Amersham Biosciences; http://www.amersham.com). Agilent Technologies (http://www.chem.agilent.com/Scripts/IDS.asp?Page=1624), in partnership with Rosetta Inpharmatics (http://www.rii.com), has described oligonucleotide synthesis
in situ using an ink-jet printing method employing standard phosphoramidite chemistry (26).

**Affymetrix GeneChips.** The industry leader in the field of oligonucleotide microarrays is undoubtedly Affymetrix Corporation (http://www.affymetrix.com), which uses photolithography-directed combinatorial chemical synthesis to manufacture so-called GeneChips, microarrays bearing hundreds of thousands of different oligonucleotides on a derivatized glass surface (Fig. 2 and Refs. 19, 34, 40). By use of the Affymetrix Fluidics system, GeneChips are hybridized with fragments (35–200 residues long) of biotinylated target RNA derived from 5 μg of total cell RNA or 0.2 μg of poly(A)-selected mRNA. Hybridized probe is recognized by a streptavidin-phycoerythrin conjugate, and then the fluorescent image is captured using Affymetrix Microarray Reader.

**Advantages of Affymetrix GeneChips.** Because probes are based entirely on sequence information, there is no need to prepare and verify physical intermediates such as bacterial clones, PCR products, or cDNAs. As such, there is less possibility of probe mix-ups. However, Affymetrix has experienced some problems in this regard. Up to one-quarter of the sequences on one set of mouse microarrays were wrong. The company had used sequences from the public sequence databases that were known to be ambiguous and that actually corresponded to the wrong strand from the DNA double helix. As a result, the oligonucleotides could not detect their target mRNAs (http://www.affymetrix.com/support/technical/product_updates/mgu74_product_bulletin.affx).

The use of multiple, short-sequence detectors enables splice variants and closely related members of a gene family to be distinguished (Fig. 2B). By use of probes representing regions of genes that significantly diverge or “are significantly unique” between family members, microarrays can distinguish transcripts that are up to 90% identical.

For each probe designed to be perfectly complementary to a target sequence, a partner probe is generated that is identical except for a single base mismatch in its center (Fig. 2B). This probe mismatch strategy, along with the use of multiple probes for each transcript, helps identify and minimize the effects of non-specific hybridization and background signal and allows the direct subtraction of cross-hybridization signals and discrimination between real and non-specific signals.

Short-chain oligonucleotides with single points of constraint are probably more accessible for hybridization to target than cDNA probes.

**Disadvantages of Affymetrix GeneChips.** There are several disadvantages to the Affymetrix GeneChips. First is a need for access to expensive specialized equipment. Second, oligonucleotide chips are only available from commercial manufacturers. Custom oligonucleotide microarrays can be commissioned, but at great expense. Third, prefabricated GeneChips are themselves very expensive, although the price, particularly for academic users, is falling. Fourth, although short-sequence probes confer high specificity, they may have decreased sensitivity/binding compared with cDNA microarrays. Low sensitivity is compensated for by employing multiple probes.

**Comparability of Different Microarray Platforms**

Technical problems inherent in probe manufacture and use still confound the extraction of meaningful data from comparative microarray experiments. Such sources of variability include

- slide heterogeneity
- spotting variation
- printing irregularities
- background noise.

Furthermore, little work has been carried out on the comparability of different microarray platforms. In the absence of common standards, different platforms are likely to give different results. Difficulties with intraplatform comparability will be compounded by different gene and probe choices.

**ANALYSIS OF MICROARRAY DATA**

Microarray experiments produce a huge amount of data. A single microarray run can produce between
100,000 and a million data points, and a typical experiment may require tens or hundreds of runs (21). Quite simply, for the first time in the history of the biomedical sciences, our ability to generate data in vast quantities is running ahead of our ability to make sense of them. Moving from data to knowledge is a
considerable challenge. Although procedures for the assessment, curation, and presentation of microarray data are rapidly evolving, statistical approaches are neither routine nor standardized (10, 76). The Microarray Gene Expression Database (MGED; 53) consortium has the goal of facilitating the adoption of common standards for microarray experiment annotation and data representation, as well as the introduction of standard experimental controls, and data normalization methods. The projects being pursued by MGED are:

- **MIAME**—formulation of the minimum information about a microarray experiment required to interpret and verify the results (5)
- **MAGE (Microarray and Gene Experiment)**—establishment of a data exchange format, or mark-up language (MAGE-ML), and object model (MAGE-OM) for microarray experiments
- **ontologies**—development of ontologies for microarray experiment description and biological material (biomaterial) annotation.

The hope is that researchers will be able to share MAGE-compatible data seamlessly. Some central internet sites have been established for depositing microarray data (e.g., http://www.ebi.ac.uk/arrayexpress).

A variety of microarray analysis software packages are available from commercial and academic sources (http://ihome.cuhk.edu.hk/~b100559/array.html#Software and http://www.lab-on-a-chip.com/suppliers/inform.html).

**Low-Level Analysis**

Primary image data having been collected from a microarray experiment, the aims of the first level of analysis, so-called low-level analysis, are background elimination, filtration, and normalization, all of which should contribute to the removal of systematic variation between chips, enabling group comparisons. Background noise is removed from cDNA microarrays by subtracting nonspecific signal from spot signal. In contrast, preprocessing of Affymetrix data is intrinsic to the perfect match and mismatch strategy (Fig. 2B). Normalization in both cases involves comparing different microarrays relative to some standard intensity value. This could be the overall intensity of the microarray, the overall intensity of all of the genes on the microarray, the intensity of so-called housekeeping genes (the expression of which are supposedly constant), or spiked targets, containing a known and constant amount of a labeled control. Negative normalization controls might be represented by target sequences from a different organism. Data are often then subjected to log transformation to improve the characteristics of the distribution of the expression values.

**High-Level Analysis**

High-level microarray analysis is often called “data mining,” the uncovering of relevant patterns of interest in data from a particular problem domain. Typically this will involve data processing using various statistical techniques to identify the patterns. In addition, data needs to be packaged, presented, archived, and compared with other types of information.

**Statistical analysis**. The statistical analysis of microarray data is probably the most difficult problem associated with the use of these techniques. The aim is to apply standard statistical approaches to determine gene expression and gene expression alteration significance, thus enabling the extraction of significant biological information from a morass of noise and variability. However, present methodologies do not deal well with the number of possible combinations. Statisticians are experienced with handling data involving a limited number of variables, but a large number of samples (e.g., the average weight of persons in England is a problem of a single variable and 49 million samples). Microarrays turn this problem on its head, producing thousands of variables from a small number of samples. A number of different methods have been explored.

**FOLD CHANGE**. Simple and intuitive, this method, involves the calculation of a ratio relating the expression level of a gene under control and experimental conditions. An arbitrary ratio (usually 2-fold) is then selected as being “significant.” Because this ratio has no biological merit, this approach amounts to nothing more than a blind guess. The selection of an arbitrary threshold results in both low specificity (false positives, particularly with low-abundance transcripts or
when a data set is derived from a divergent comparison) and low sensitivity (false negatives, particularly with high-abundance transcripts or when a data set is derived from a closely linked comparison). It is now accepted that the use of the fold change method should be discontinued.

**UNUSUAL RATIO.** This method selects genes for which the ratio of control and experimental values is an arbitrarily selected distance from the mean control-to-experimental ratio. This is usually taken to be ±2 standard deviations. This can be calculated by applying z-transformation, subtraction of the mean, and division by the standard deviation to the log ratio values. As a fixed proportion threshold is used, the unusual-ratio method will always identify the most affected genes. However, genes will be reported, even if there are no differentially expressed genes. Although flawed, the unusual-ratio method is commonly used for the analysis of cDNA microarrays.

**UNIVARIATE STATISTICS.** If log ratios follow a normal distribution, a probability (P value) that the gene is erroneously reported as being differentially regulated above a given threshold can be assigned using a univariate statistical test (e.g., the t-test). However, such tests require correction. From a statistical point of view, interrogating R genes on a microarray is the same as running R parallel tests. The Bonferroni correction takes this into account, and adjusts P to P/R. However, because a microarray experiment involves an R of thousands, no differentially expressed genes would ever be reported as reaching significance. Less conservative correction methods have been reported (10).

**ANALYSIS OF VARIANCE.** Ultimately, the analysis of microarray data, and the selection of differentially expressed genes, will be achieved by analysis of variance (ANOVA) based on explicit experimental models.

**Identifying patterns in microarray data.** The output from the analysis of a microarray experiment is usually a large data spreadsheet filled with numbers related to the signal intensity for each gene on the chip. Further analysis is required to identify groups of genes that are similarly regulated across the biological samples under study. A variety of mathematical procedures have been developed that partition genes or samples into groups, or clusters, with maximum similarity, thus enabling the identification of gene signatures or informative gene subsets. Methods for classification are either unsupervised or supervised. Supervised methods use existing biological information about specific genes that are functionally related to “guide” or “test” the cluster algorithm. With unsupervised methods, no prior test set is required.

The most commonly employed unsupervised classification methods are the clustering techniques (16). They fall within the categories of hierarchical and nonhierarchical (partitional) clustering. Most cluster analysis techniques are hierarchical; the resultant classification has an increasing number of nested classes, and the result resembles a phylogenetic classification. Hierarchical clustering has the advantage that it is simple and the result can be easily visualized. Non-hierarchical clustering techniques, such as k-means clustering (51), partition objects into different clusters without trying to specify the relationship between/among individual elements. A self-organizing map (SOM (48)) is a neural-network-based divisive clustering approach. A SOM assigns genes to a series of partitions on the basis of the similarity of their expression vectors to reference vectors that are defined for each partition. It is the process of defining these reference vectors that distinguishes SOMs from k-means clustering.

Principal component analysis [PCA (1, 41)] is a mathematical decomposition technique that picks out the most abundant themes to reoccur in an experiment. A set of expression patterns, called principal components, is identified, and linear combinations of these are assembled to represent the behavior of genes in a data set. PCA can be applied to both genes and experiments as a means of classification. In most implementations of PCA, it is difficult to define accurately the precise boundaries of distinct clusters in the data or to define genes (or experiments) belonging to each cluster. However, PCA is a powerful technique for the analysis of gene expression data when used with another classification technique, such as k-means clustering or SOMs, that requires the user to specify the number of clusters.

One approach to supervised modeling is linear discriminant analysis (LDA), which uses a training set
consisting of all classes of interest and then tries to set up a model that classifies an unknown sample unambiguously into one of the already established classes [23] (http://www.stat.berkeley.edu/users/terry/zarray/Html/discr.html).

**Relational and functional databases.** Microarray data need to be interpreted within the context of gene function and the functional relationships between genes. This demands relating microarray data with existing biological knowledge. However, this project has had to face up to the linguistic ambiguities of the existing scientific literature; supposedly rigid, solid scientific concepts are often couched in imprecise terms. What is needed is a common gene language. Thus the aim of the Gene Ontology Consortium (http://www.geneontology.org) is to “produce a dynamic controlled vocabulary that can be applied to all organisms even as knowledge of gene and protein roles in cells is accumulating and changing.”

**dChip.** One of the best Affymetrix analysis packages is dChip (31, 32, 45), available online at no cost (http://www.dchip.org). dChip is based on a statistical model for Affymetrix expression data at the probe level. This approach facilitates automatic probe selection in the analysis stage to reduce errors caused by outliers, cross-hybridizing probes, and image contamination. Data are then normalized using a rank-selection method. The program selects a set of genes with the property that the rank of a gene in this set according to its expression measurement in one microarray is similar to its rank using values for the second microarray. Genes thus selected tend to be nondifferentially expressed, and this forms a valid basis for the computation of a normalization relation. By pooling information across multiple microarrays, it is possible to assess standard errors for the model-based expression indexes (MBEI) calculated for each gene. After obtaining MBEIs, dChip can perform some high-level analysis, such as hierarchical and functional clustering, involving ANOVA-based gene filtering, comparative analysis, PCA, and LDA. Some of these functions require R (http://www.r-project.org), a statistical package and language, as the engine for computational and graphic tasks.

**CONFIRMATIONAL STUDIES**

Because of the statistical issues raised by microarray technology, it is necessary that findings be confirmed using independent methodological criteria, preferably with separate samples rather than with the tissue or RNA used to derive the original targets.

A rapid, high through-put, but expensive, method for confirmation of microarray data is quantitative (real time) RT-PCR using the TaqMan (Applied Biosystems; http://www.appliedbiosystems.com/products/productdetail.cfm?prod_id=42), iCycler (Bio-Rad Laboratories; http://www.bio-rad.com/iCycler/), LightCycler (Roche Diagnostics; http://www.lightcycler-online.com/) machines. TaqMan PCR (http://www.appliedbiosystems.com) exploits the 5’-nuclease activity of Taq DNA polymerase in conjunction with DNA probes labeled with quencher and reporter dyes. A positive PCR reaction results in the removal of the reporter dye from the influence of the quencher dye, leading to an increase in measurable fluorescence. The real-time reaction information allows quantification of target nucleic acid. Advantages of TaqMan are that it is a closed-tube assay, reducing the risk of contamination, and no post-PCR processing (such as gel electrophoresis) is required. Multiple reactions, detecting more than one sequence per reaction, are possible using different quencher dyes.

Alternatively, Northern blots or ribonuclease protection assays provide the benefit of direct quantification. Finally, in situ hybridization can be used as a sensitive measure of gene expression changes in specific cell types within a mixed tissue. This is important, as significant gene expression changes detected on a microarray may be related to a small fraction of the cells in a tissue.

As steady-state levels of RNA are not necessarily reflective of the final steady-state level of the functional protein translation product of an mRNA, further studies might involve the use of specific antibody probes in Western blot or immunocytochemical studies.

Because a microarray experiment may reveal putative changes in the expression of tens or hundreds of genes, it is practically impossible to confirm all of the
data. However, it is incumbent upon investigators to evaluate a reasonable number of genes. That said, confirmational studies may raise other issues. Although a microarray experiment might indicate an increase or decrease in the expression of a gene, an independent method might reveal a greater or a lesser change. Does such a result represent sufficient confirmation of the microarray findings, or does a quantitative difference raise new questions about the validity of the microarray data?

PRESENT AND FUTURE CHALLENGES

Hardware

Scientists in academia and industry are diligently addressing the technical problems of microarrays. The quality, reproducibility, comparability, sensitivity, and dynamic range of microarrays will improve. In 1965, Gordon Moore, the founder of Intel, observed that the number of transistors per semiconductor chip doubles every 18–24 months (36). Microarrays are on a similar trajectory. In 1998, an Affymetrix microarray contained fewer than 1,000 genes; by 2000, it boasted of 12,000. The ultimate aim is to represent all of the expressed sequences of the genome on a single chip. Toward this end, Affymetrix has recently released the Human Genome U133 GeneChip Set, comprised of two microarrays containing almost 45,000 probe sets corresponding to more than 39,000 transcript variants representing greater than 33,000 of the best characterized human genes (http://www.affymetrix.com/products/arrays/specific/hgu133.affx). However, until the day when all transcription units have been identified, microarrays will remain incomplete. Although this is acceptable, microarrays should be unbiased in their selection of genes.

Another factor driving the development of bigger chips is cost; as size and volume increase, prices will surely drop.

Software

We continue to “search for a body of mathematics that will serve as a natural language for gene expression information” (54). The role of this mathematics will be to

• combine expression data with other sources of information to improve the range and quality of biological conclusions
• develop techniques that enable the modeling of gene networks in the context of the functioning of integrated biological systems.

It is to be anticipated that, as these aims are realized, human judgement and expertise will be replaced by artificial intelligence.

Experimental Design

Careful contemplation of the ultimate research objective for a study will ensure that appropriate type and number of treatment groups are incorporated into an experimental design. Every microarray study should include a sufficient number of independent experiments to allow statistical evaluation of claims of an increase or decrease in gene expression (30, 38). The number of microarrays and replicates needed to achieve statistical significance is dependent on the coefficient of variation. Reproducibility must be demonstrated, including rigorous evaluation of the run-to-run variability for each gene. This will permit appropriate adjustments to be made that will reduce the false discovery rate.

Another challenge that can be overcome by good experimental design is the need to distinguish among primary and secondary effects and subsequent events. An initial perturbation of a biological system will induce gene expression changes that will be followed by more alterations related to secondary, cellular changes, and subsequent modulations at the organismal level. All of these levels of cause-and-effect plasticity are of interest and could be dissected by incorporating a broad range of time points. Similarly, effects not directly related to an experimental perturbation can be eliminated by inclusion of appropriate control groups. For example, if studying gene expression changes as a consequence of drug interactions with a specific receptor, controls might include comparisons with groups using a receptor pathway inhibitor or a nonactive analog.
TOWARD AN UNDERSTANDING OF GENE FUNCTION

The microarray-based approach to the problem of gene function clusters genes according to their expression behavior under defined conditions and to assign function. The hypothesis of this “guilt-by-association” approach is that clustered genes may be co-regulated and therefore may be involved in similar functions. However, sequence and expression analysis alone is insufficient to fully inform us about gene function. To make sense of these data, the hypotheses that emerge from analysis of system expression information must be tested empirically. This will involve the integration of genomic knowledge with biochemistry, cell biology, genetics, structural biology, and proteomics. Ultimately, hypotheses must be tested within the physiological integrity of the whole organism. This will demand the development of a new, high-throughput systems biology coupled with rapid and efficient gene transfer techniques.

The Wellcome Trust is thanked for support. Dr. Mohamed Ghorbel and Greig Sharman (University of Bristol) are thanked for critical expression behavior under defined conditions and to assign function. The hypothesis of this “guilt-by-association” approach is that clustered genes may be co-regulated and therefore may be involved in similar functions. However, sequence and expression analysis alone is insufficient to fully inform us about gene function. To make sense of these data, the hypotheses that emerge from analysis of system expression information must be tested empirically. This will involve the integration of genomic knowledge with biochemistry, cell biology, genetics, structural biology, and proteomics. Ultimately, hypotheses must be tested within the physiological integrity of the whole organism. This will demand the development of a new, high-throughput systems biology coupled with rapid and efficient gene transfer techniques.

The Wellcome Trust is thanked for support. Dr. Mohamed Ghorbel and Greig Sharman (University of Bristol) are thanked for critical expression behavior under defined conditions and to assign function. The hypothesis of this “guilt-by-association” approach is that clustered genes may be co-regulated and therefore may be involved in similar functions. However, sequence and expression analysis alone is insufficient to fully inform us about gene function. To make sense of these data, the hypotheses that emerge from analysis of system expression information must be tested empirically. This will involve the integration of genomic knowledge with biochemistry, cell biology, genetics, structural biology, and proteomics. Ultimately, hypotheses must be tested within the physiological integrity of the whole organism. This will demand the development of a new, high-throughput systems biology coupled with rapid and efficient gene transfer techniques.

REFERENCES


