Navigating gene expression using microarrays — a technology review

Almut Schulze and Julian Downward

Signal Transduction Laboratory, Imperial Cancer Research Fund, 44Lincoln's Inn Fields, London WC2A 3PX, UK
e-mail: schulze@icrf.icnet.uk, downward@icrf.icnet.uk

Parallel quantification of large numbers of messenger RNA transcripts using microarray technology promises to provide detailed insight into cellular processes involved in the regulation of gene expression. This should allow new understanding of signalling networks that operate in the cell and of the molecular basis and classification of disease. But can the technology deliver such far-reaching promises?

The use of microarrays to analyse gene expression on a global level has recently received a great deal of attention. There has been much speculation about the power of this approach in the analysis of disease and the unravelling of cellular signalling pathways. As the finishing touches are put on the human genome sequence and it becomes clear that it contains less than 40,000 genes, one might hope that we are advancing towards being able to describe fully the regulation of gene-expression networks and how they malfunction in disease. But does the reality live up to the hype? In this review we describe the technology underpinning the microarray revolution, and attempt to address the ways in which it can be used and the type of solutions it can be expected to provide. We also consider the pitfalls of the technology, whether it will provide new insights into cell biology, and how to avoid drowning in a glut of uninterpretable data.

Upstream considerations: microarray technology

The principle of a microarray experiment, as opposed to the classical northern-blotting analysis, is that mRNA from a given cell line or tissue is used to generate a labelled sample, sometimes termed the 'target', which is hybridized in parallel to a large number of DNA sequences, immobilized on a solid surface in an ordered array[1]. Tens of thousands of transcript species can be detected and quantified simultaneously. During recent years, DNA microarray technology has been advancing rapidly. The development of more powerful robots for arraying, new surface technology for glass slides, and new labelling protocols and dyes, together with increasing genome-sequence information for different organisms, including humans, will enable us to extend the quality and complexity of microarray experiments.

Array platforms. Although many different microarray systems have been developed by academic groups and commercial suppliers, the most commonly used systems today can be divided into two groups, according to the arrayed material: complementary DNA (cDNA) and oligonucleotide microarrays (Fig. 1). The arrayed material has generally been termed the probe since it is equivalent to the probe used in a northern blot analysis. Probes for cDNA arrays are usually products of the polymerase chain reaction (PCR) generated from cDNA libraries or clone collections, using either vector-specific or gene-specific primers, and are printed onto glass slides or nylon membranes as spots at defined locations. Spots are typically 100–300 μm in size and are spaced about the same distance apart. Using this technique, arrays consisting of more than 30,000 cDNAs can be fitted onto the surface of a conventional microscope slide. For oligonucleotide arrays, short 20–25mers are synthesized *in situ*, either by photolithography onto silicon wafers (high-density-oligonucleotide arrays from Affymetrix[2], http://www.affymetrix.com) or by ink-jet technology (developed by Rosetta Inpharmatics, http://www.rii.com, and licensed to Agilent Technologies). Alternatively, presynthesized oligonucleotides can be printed onto glass slides. Methods based on synthetic oligonucleotides offer the advantage that because sequence information alone is sufficient to generate the DNA to be arrayed, no time-consuming handling of cDNA resources is required. Also, probes can be designed to represent the most unique part of a given transcript, making the detection of closely related genes or splice variants possible. Although short oligonucleotides may result in less specific hybridization and reduced sensitivity, the arraying of presynthesized longer oligonucleotides (50–100mers) has recently been developed to counteract these disadvantages[3]. However, the high cost of commercially available, *in situ*-synthesised oligonucleotide arrays can make them inaccessible for academic laboratories, and purchase of large numbers of long oligonucleotides also incurs significant cost.

Spotted arrays allow a greater degree of flexibility in the choice of arrayed elements, particularly for the preparation of smaller, customized arrays for specific investigations. As a result, cDNA gridded arrays have so far been the technique most frequently used in academic labs (see http://cmgm.stanford.edu/pbrown/mguide/index.html and http://www.nhgri.nih.gov/DIR/LCG/15K/HTML for information about cDNA array technology). In addition, arraying of unsequenced clones from cDNA libraries can be useful for gene discovery. However, with prices for oligonucleotide synthesis falling all the time, spotted long-oligonucleotide arrays could be a viable alternative for the future.

Target preparation. Another important difference between *in situ*-synthesized, high-density oligonucleotide arrays (Affymetrix) and spotted arrays lies in target preparation (Fig. 1). In both cases, mRNA from cells or tissue is extracted, converted to DNA and labelled, hybridized to the DNA elements on the array surface of the array, and detected by phospho-imaging or fluorescence scanning. The high reproducibility of *in situ* synthesis of oligonucleotide chips allows accurate comparison of signals generated by samples hybridized to separate arrays. In the case of spotted arrays, the process of gridding is not accurate enough to allow comparison between different arrays. The use of different fluorescent dyes (such as Cy3 and Cy5) allows mRNAs from two different cell populations or tissues to be labelled in different colours, mixed and hybridized to the same array, which results in competitive binding of the target to the arrayed sequences. After hybridization and washing, the slide is scanned using two different wavelengths, corresponding to the dyes used, and the intensity of the same spot in both channels is compared. This results in a measurement of the ratio of transcript levels for each gene represented on the array. To be able to compare a large number of samples, the same reference RNA — sometimes a mixture of all the samples of one experiment or a commercially available standard — can be used.

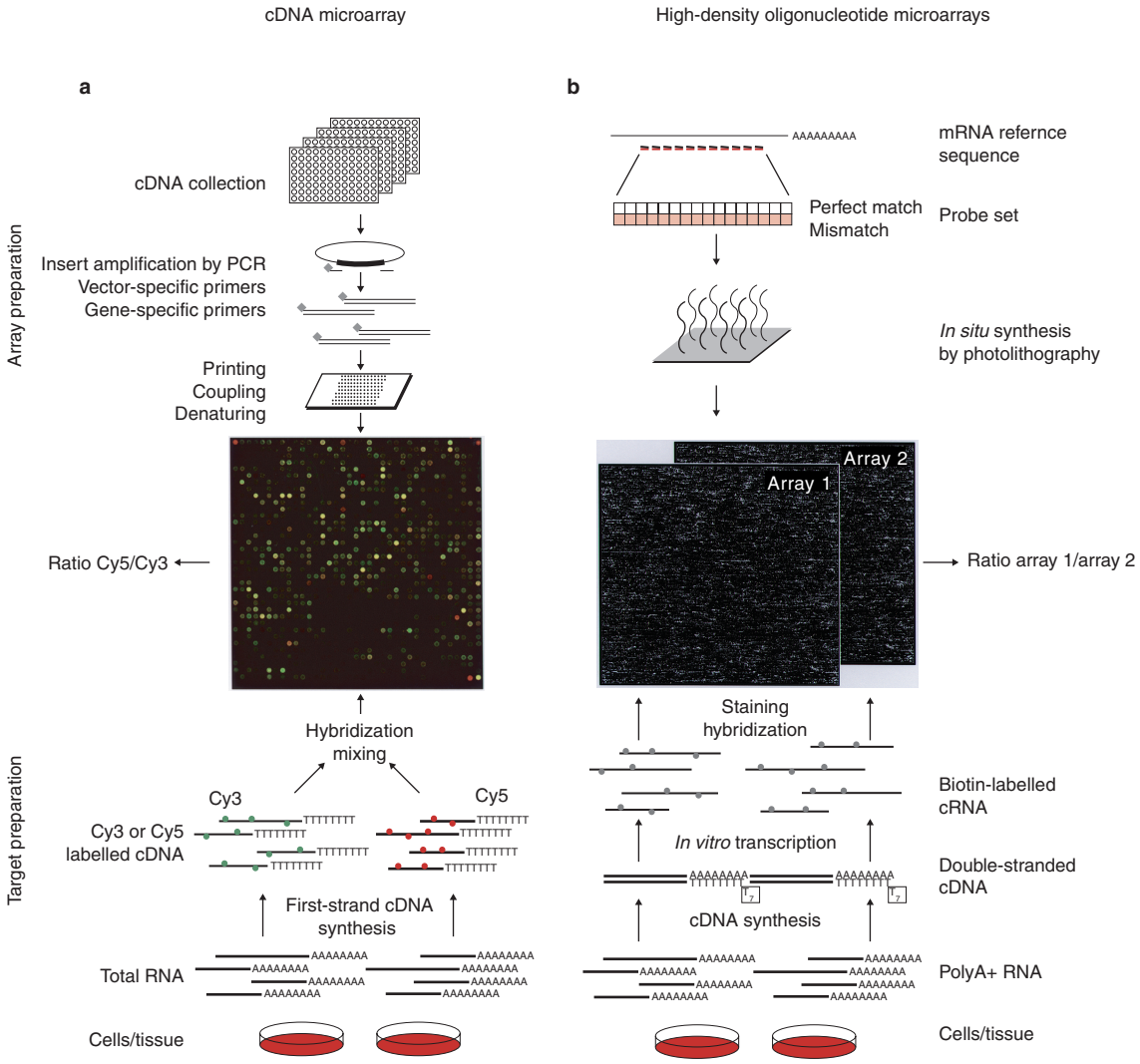

Figure 1 Schematic overview of probe array and target preparation for spotted cDNA microarrays and high-density oligonucleotide microarrays. **a**, cDNA microarrays. Array preparation: inserts from cDNA collections or libraries (such as IMAGE libraries) are amplified using either vector-specific or gene-specific primers. PCR products are printed at specified sites on glass slides using high-precision arraying robots. Through the use of chemical linkers, selective covalent attachment of the coding strand to the glass surface can be achieved. Target preparation: RNA from two different tissues or cell populations is used to synthesize single-stranded cDNA in the presence of nucleotides labelled with two different fluorescent dyes (for example, Cy3 and Cy5). Both samples are mixed in a small volume of hybridization buffer and hybridized to the array surface, usually by stationary hybridization under a coverslip, resulting in competitive binding of differentially labelled cDNAs to the corresponding array elements. High-resolution confocal fluorescence scanning of the array with two different wavelengths corresponding to the dyes used provides relative signal intensities and ratios of mRNA abundance for the genes represented on the array. **b**, High-density oligonucleotide microarrays. Array preparation: sequences of 16–20 short oligonucleotides (typically 25mers) are chosen from the mRNA reference sequence of each gene, often representing the most unique part of the transcript in the 5′-untranslated region. Light-directed, *in situ* oligonucleotide synthesis is used to generate high-density probe arrays containing over 300,000 individual elements. Target preparation: polyA$^+$ RNA from different tissues or cell populations is used to generate double-stranded cDNA carrying a transcriptional start site for T7 DNA polymerase. During *in vitro* transcription, biotin-labelled nucleotides are incorporated into the synthesized cRNA molecules. Each target sample is hybridized to a separate probe array and target binding is detected by staining with a fluorescent dye coupled to streptavidin. Signal intensities of probe array element sets on different arrays are used to calculate relative mRNA abundance for the genes represented on the array.

As array technology has advanced, more sensitive and quantitative methods for target preparation have had to be developed. In cases in which the quantity of RNA is not limited, incorporation of nucleotides coupled to fluorescent dyes during synthesis of the first strand of cDNA is the method of choice, as it provides the most linear relationship between starting material and labelled product. However, most protocols require between 25–100 μg total RNA, which is often not readily available in studies using primary cells or tissues. Various procedures have been developed to increase sensitivity and reduce the amount of RNA required. One strategy is target amplification by *in vitro* transcription, whereby up to 50 μg of labelled cRNA can be produced from 1 μg of mRNA. In addition, several rounds of *in vitro* transcription can be combined with cDNA synthesis to enhance the amplification even further[4]. Using these protocols, it is even possible to profile the transcripts of a single cell[5]. Another strategy is post-hybridization amplification using labelled antibodies or molecules carrying large numbers of fluorophors[6]. Several studies have used target-amplification tech-

niques to compare the expression profiles of defined cell populations extracted from tissue sections by laser-capture microdissection. However, suitable controls are required to ensure that amplification has not introduced significant experimental bias into the target preparation[7]. This problem has been particularly evident in the expression profiling of tumour samples. In the case of solid tumours, obtaining pure populations of tumour cells for microarray analysis would require microdissection. However, a recent study[8] using grossly dissected breast-cancer specimens has demonstrated a way to circumvent the problem of sample heterogeneity. Expression profiles from whole solid tumours can be compared to profiles from potential untransformed infiltrating cell types, such as lymphocytes or endothelial cells, to identify a subset of genes with expression patterns that are specific to the tumour cells. Subsequent data analysis and sample clustering can then be carried out only on this 'intrinsic gene subset', which in the case of the recent study was sufficient for tumour classification[8].

Data analysis, reproducibility and validation. The data of a microarray experiment typically constitute a long list of measurements of spot intensities and intensity ratios, generated either by pairwise comparison of two samples or by comparing several samples to a common control. The challenge is then to sieve through this mound of data to find meaningful results. Replication has been shown to reduce markedly the number of potential false positive results, but may be difficult because of high cost or limitation of the amount of sample. However, as the efficiency of incorporation of nucleotides labelled with different fluorescent dyes during target-sample preparation may not be equal, reciprocal labelling with swapped colours is recommended. Our own experiences have led us to analyse independent duplicate samples with reciprocal labelling — using four microarrays as standard for each experimental point. Several researchers have found that variability of microarray results can be significant, especially for genes with low expression levels, and replication is needed to establish a high degree of confidence in the data (see also Mills *et al.*, pp. 175–178, this issue). Another problem arises from the fact that data about genes that are found to be unchanged have to be treated with particular care. It is very possible that certain DNA elements on the array simply fail to detect the right transcript species, as a result of cross-hybridization or adverse secondary structure. Verification of a subset of results by alternative techniques such as northern hybridisation, RNase protection or PCR with reverse transcription (RT-PCR), in particular 'quantitative' or 'real-time' RT-PCR[9], can help to establish an estimate of the variability of a given experimental system. More experience in using microarray technology, particularly concerning the choice of DNA sequences to be arrayed, will improve confidence in the reliability of the data.

Because of the complexity of the data sets generated by microarray experiments, the use of data-analysis software is essential. Several data-analysis tools have been developed by commercial suppliers (such as GeneSpring from SiliconGenetics, http://www.sigenetics.com/), and others are available from public sources (see http://genome-www4.stanford.edu/MicroArray/SMD/restech.html for an overview).

Downstream considerations: what can a cell biologist learn using microarrays?

The use of microarrays to analyse gene expression is becoming increasingly familiar, but what it can and cannot achieve is still the subject of much debate. Microarrays can be used to investigate problems in cell biology in various ways; the different experimental approaches fall between two extremes. At one end, the investigator is interested only in finding the single change in gene expression that might be the key to a given alteration in phenotype; in some ways this is equivalent to looking for a needle in a haystack, and could be thought of as an entirely local approach to analysis of gene-expression changes (Fig. 2). At the other extreme, the aim is to look at overall patterns of gene expression in order to understand the architecture of genetic regulatory networks, a global approach that could ultimately lead to complete description of the transcription-control mechanisms in a cell.

Following individual pathways. So far, much of the interest of cell biologists in microarrays has been directed towards identifying individual genes, the regulated expression of which can explain particular biological phenomena. However, it is probably fair to say that this has not been the most successful use of microarrays to date. The problem is that microarray experiments, whether based on oligonucleotides or cDNA, are highly capable of generating long lists of genes with altered expression, but they provide few clues as to which of these changes are important in establishing a given phenotype — this deduction is left to the ingenuity of the experimenter. A given stimulus could potentially lead to changes in the mRNA levels of hundreds of genes, particularly in mammalian systems. Faced with such a mass of data, the temptation is to look for genes that conform to existing ideas about how the system might work. The benefits of an unbiased approach will be lost if exploration is limited to our current framework of understanding.

How can this problem be minimized? Careful experimental design is critical, and several key issues should be considered. It is important that the samples for comparison are as closely matched as possible: a comparison of apples and oranges is unlikely to be the best way to investigate the regulation of skin colouration! Although this may seem obvious, it is very easy, when analysing growth-signalling pathways, for example, to end up comparing proliferating and quiescent cells, in which the vast majority of differences in gene expression are the consequences, rather than the causes, of the induction of proliferation. These problems can be reduced by using inducible systems — either by addition of extracellular factors or by the use of inducible expression vectors — and looking at earlier rather than later time points. In some cases it may be possible to block secondary changes in gene expression (those that are dependent on the initial changes in protein levels) using inhibitors of protein synthesis, although these can have significant effects in their own right and should be treated with caution. The comparison of cells constitutively expressing activated signalling proteins with parental cells does not allow primary and secondary changes in gene expression to be distinguished, and runs the risk of identifying clonal differences in expression that are not connected to the gene of interest but are generated by selection of the engineered cell line.

Through the use of well-designed time courses and titrations, it may be possible to reduce greatly the number of gene-expression changes that could account for a given biological response. It is possible that a secondary functional screen may be available to narrow this down further, but in most cases careful analysis of the published literature will be required before direct experimental investigation of gene function is undertaken. This can often take one into entirely unfamiliar areas, requiring much time and patience and is also complicated by the lack of unified gene ontology. The use of advanced sophisticated literature-searching tools, such as Medminer (http://discover.nci.nih.gov/textmining/filters.html) or high-density array-pattern interpreter (HAPI, http://array.ucsd.edu/hapi), which correlates gene accession numbers with MeSH (medical subject heading) keywords, can be very useful for known genes. Functional annotation of new sequences and expressed sequence tags remains a challenge, but is crucial for the interpretation of genome-wide expression data. For further resources, see Mills *et al.* on pages 175–178 of this issue.

Once a set of genes has been identified as candidates for explaining the phenotype of interest, their function has to be modulated in such a way as to directly determine their degree of involvement. To validate the involvement of the identified gene products in the studied phenotype, functions of genes on the shortlist need to be modulated artificially to see whether the generation of the biological end point can be influenced. Several technical and conceptual problems can arise here. One is that in mammalian systems it is

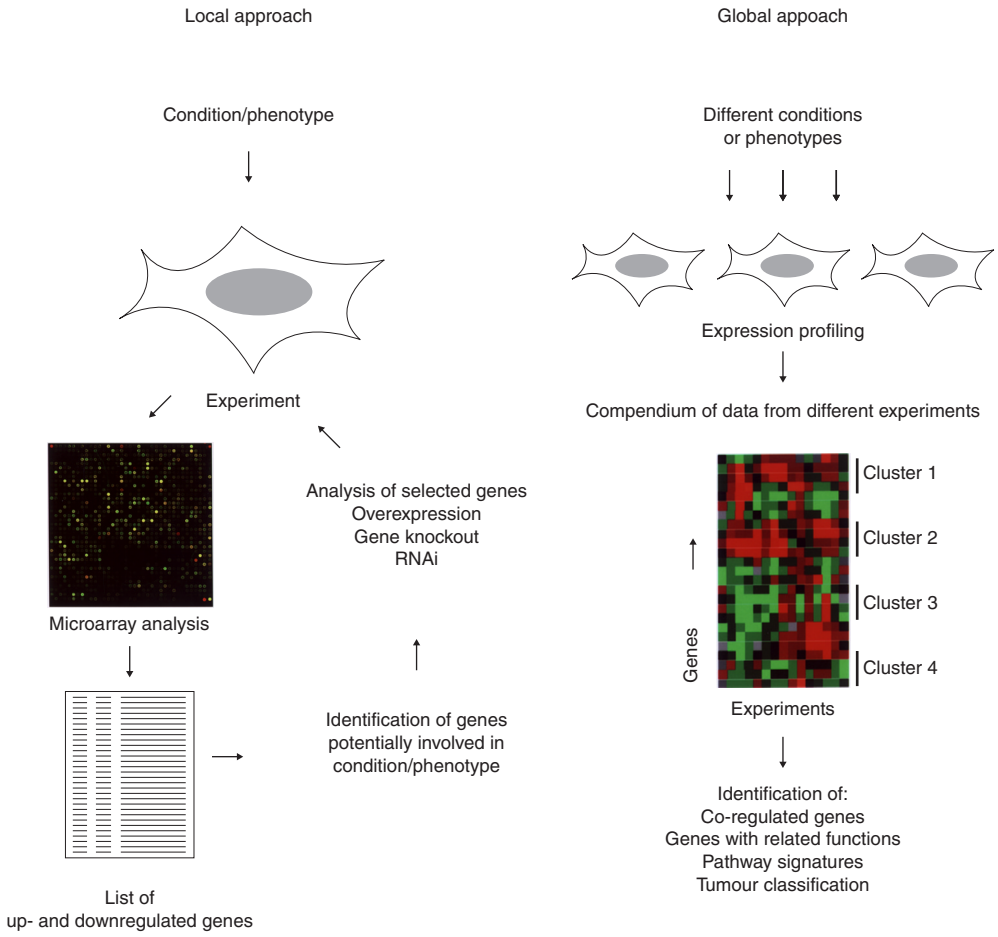

Figure 2 Different approaches towards microarray experiments in cell biology. For a local approach, results from a microarray experiment can be used to identify a shortlist of genes that could be modulated in response to a given stimulus, condition or experimental manipulation. The causal involvement of modulation of gene expression in the establishment of a given phenotype, such as transformation or protection from apoptosis, must be established individually. In the global approach, expression profiles from many different conditions, stimuli or phenotypes are compared. By using cluster-analysis tools it is possible to identify co-regulated genes, genes with related functions, and signatures of individual signalling pathways within the data set. Different samples can be grouped according to their expression profiles, which are used for molecular classification of cancer types.

currently very difficult to modulate gene expression or protein function on a broad scale. By using conventional techniques of molecular and cellular biology, such as dominant negative or active mutants, antisense or knockout technology, a researcher may be able to influence the activity of a handful of gene products, but not dozens or hundreds. However, to modulate the activity of scores of mammalian genes in a reasonable time frame is beyond the current ability of most laboratories. Future technical advances may improve this, perhaps following the situation in lower eukaryotic model systems where the advent of RNA interference and the establishment of systematic gene knockouts has revolutionized the ease of large-scale gene function.

Even if the functions of relatively large numbers of mammalian gene products, can be modulated by the experimenter, there is still the problem that such manipulation may affect signalling networks in unexpected ways. For example, the concern that some proteins may provide permissive functions, such that ablation of their activities leads to failure of a signalling pathways even when the proteins in question are not directly involved. Another problem is that functional redundancy could mean that two related gene products have overlapping functions and would therefore both need to be ablated to block a signalling pathway. Furthermore, in complex pathways, activation of a single gene may itself be insufficient to induce a phenotype, even if that component is a necessary part of the pathway. Although gene ablation or activation can be effective in characterising simple linear pathways, in reality many signalling systems are part of combinatorial networks that might not be so amenable to this kind of analysis.

One final problems that limits the usefulness of microarrays in the analysis of individual pathways is the fact that regulation of mRNA levels is only one aspect of biological control. Protein levels are also controlled at several post-transcriptional steps, and protein activity is controlled by post-translational modification. Ultimately, the global study of cellular proteins by proteomics may be able to provide the complete picture, although the fundamental differences in nucleic acid and protein chemistry indicate that proteomics will always be considerably more labour-intensive than the study of gene transcription. It should be noted, however, that analysis of gene-expression patterns is no less powerful a concept than proteomics when it comes to identification of the characteristics of signalling pathways or disease states.

Analysing gene-expression networks. The true power of microarrays in analysing cell function has been more obvious when used to provide global pictures of expression patterns, rather than to

identify a single critical gene (see Fig. 2). Large-scale analyses of changes in gene expression during the yeast and mammalian cell cycles[10–12] or in response to stimulation of fibroblasts with serum or platelet-derived growth factor[13,14] have been undertaken to identify genes with similar expression patterns. To achieve this, sophisticated software tools that identify common patterns of expression have had to be developed. Genes with similar patterns are clustered, and phylogenetic trees of related clusters can be drawn. During the past two years several new strategies for cluster analysis of microarray data have been described (reviewed in refs 15, 16). Several software tools for cluster analysis have been developed and are available from public sources (such as GeneCluster from Whitehead/MIT Center for Genome Research, http://www.genome.wi.mit.edu/MPR/; Expression Profiler from European Bioinformatics Institute, http://expsrv.ebi.ac.uk/; and Cluster from Lawrence Berkeley National Laboratory, http://rana.lbl.gov/EisenSoftware.htm).

In general, one can discriminate between unsupervised methods, such as *k*-means clustering, principal-component analysis and self-organizing maps[17,18], and supervised methods, in which a subset of the data is used to train the system to discriminate within the remaining set[19]. Such studies are based on the assumption that genes with closely related expression patterns may be controlled by the same regulatory mechanisms. Great advances have been made in yeast, in which it has been possible to identify common DNA motifs in the promoter regions of co-ordinately regulated genes[10,12,20,21]. This can lead to the identification of new transcription factors that are implicated in the transcription of a co-regulated set of genes. Groups of co-ordinately regulated genes (so-called 'synexpression groups') have been identified in higher eukaryotes[22], but as the organization of regulatory elements is far more complex than in yeast, the usefulness of using such pattern-recognition methods tools to identify common motifs in mammalian promoters remains unclear. However, detailed analysis of binding sites for known transcription factors within the promoters of co-regulated genes can already be useful in understanding the coupling of signalling processes with gene expression within the cell. The aim is to identify similar patterns or combinations of binding sites for different transcription factors on the promoters of co-regulated genes, but this requires a more comprehensive knowledge of transcription-factor binding sites and promoter structures. A recent technical development that could help to identify genes that are regulated by particular transcription factors involves using chromatin immunoprecipitation to generate target sample DNA for a microarray containing all yeast intergenic regions. This allows the identification of all sites within the genome at which specific transcription factors bind[23,24]. Overall, these approaches may eventually lead to the generation of accurate and comprehensive maps of transcription-control mechanisms.

Apart from conclusions about common mechanisms of transcriptional control, identification of co-regulated genes may also be used to assign potential functions to new genes. Although ambiguous, the idea behind this strategy is that genes with similar expression patterns in many different growth conditions are functionally related ('guilt by association'). A good example of the power of combining many data sets has been provided recently by Hughes *et al.*[25]. They placed hundreds of data sets, representing the gene-expression profiles of yeast in response to various drugs or mutations of known genes, into a compendium database. They then compared the expression profiles of cells with mutations in genes of unknown function with the compendium; in many cases the function of the gene of interest could be predicted with considerable accuracy by matching its profile to those corresponding to known mutants in the compendium. This method could also be used to determine the gene products or pathways that are targeted by a drug of unknown function, possibly providing valuable short-cuts to the characterization of new drugs. Deletion of the gene that encodes the desired target of the drug should prevent all the changes in expression caused by the drug, leaving only "off-target" side-effects resulting from unwanted interactions with other proteins[26].

It is to be hoped that analysis of these huge quantities of data will ultimately yield fundamental insights into the wiring of the cell. Many large microarray data sets — cataloguing changes in gene expression in response to mitogens, drug treatments, activation of specific pathways, and so on — have been placed in the public domain. In part, the logic behind this has been the hope that others will be able to make more sense of the expression patterns than the original investigators, particularly by comparing several data sets. This obviously requires a degree of standardization of databases, which is being addressed but has yet to be fully achieved[27] (see http://www.ebi.ac.uk/arrayexpress and http://www.ncbi.nlm.nih.gov/geo). For maximum comparability of data sets, standards may also need to be agreed for experimental protocols across the entire community.

As well as aiding the understanding of networks that regulate gene transcription, the use of microarray technology is also contributing directly to the annotation of the human genome sequence. The best example of this has been the use of the draft human genome sequence to design long oligonucleotides that correspond to predicted exons[28]. These were used on a genome-wide scale on microarrays to investigate whether the exons predicted by gene-finding programs are actually expressed. It seems likely that further work of this type will be required for all of the genes in the human genome to be identified. The consideration of gene expression at the level of individual exons will also be important in understanding the extent to which alternative splicing occurs in the human genome.

Cancer and other diseases. Microarrays hold much promise for the analysis of diseases, and their use has been especially intensive in the case of cancer. The identification of single gene products that are expressed in tumour cells but not in normal tissue is of great pharmacological interest, particularly in the establishment of tumour markers for diagnostic purposes. In addition, proteins that are expressed on the surface of tumour cells can be used to design targeting strategies for chemo- or immunotherapy. Transcriptional profiling of solid tumours is complicated by the fact that they may contain variable amounts of infiltrating tissue, such as stroma, endothelial or lymphoid cells. Isolation of purified tumour cells by laser-capture microdissection results in extremely low RNA yields, which require substantial target amplification and can introduce significant bias into the results. A recent technical advance, which should prove to be useful in verifying potential tumour markers identified using whole solid tumours, is the development of tissue microarrays. Small sections from many different tumours are combined in a paraffin block, sectioned and applied to microscope slides. Using this technique, up to 1,000 tumour specimens can be stained in parallel for immunohistochemistry or fluorescence *in situ* hybridization[29] to correlate gene expression with chromosomal rearrangements.

Apart from identification of tumour markers, microarray analysis can distinguish between clinically distinct subtypes of leukemia[30], lymphoma[31], breast cancer[8] and melanoma[32]. Gene-expression profiles have also been used to discriminate between normal and colon-tumour-derived tissue[33]. Microarrays have also been used extensively on tumour-derived cell lines and to classify the NCI-60 cancer cell-line panels[34,35] according to their tumour origin. The effects of more than 70,000 chemical compounds on these cell lines have been correlated with alterations in their gene-expression profiles[35] in an attempt to identify genes that are involved in drug sensitivity or resistance. Although these studies have offered new prospects for refined tumour diagnosis and treatment by identifying different subtypes of tumours that can be targetted with tailored therapies, the real question is whether microarray analysis can also help us to understand tumour biology. By comparing the expression profiles of tumour samples using many genes, it is possible to identify those genes whose expression characterizes a particular tumour

type, providing a signature of tumour gene expression. It will also be interesting to compare the expression signature of a particular tumour type to a compendium of data generated *in vitro* by measuring the responses of closely related cell lines in culture to many different stimuli, such as hormones, growth factors, or targeted activation of signalling intermediates[36]. Using this strategy, it should be possible to draw conclusions about which signalling pathways are activated in a particular tumour type, leading to the identification of pathways that might provide therapeutic targets. One example of this approach was the identification of the signalling pathways involved in the metastatic potential of melanoma cells[37].

Finally, gene-expression data can be combined with analysis of genomic alterations, for example by microarray analysis of comparative genomic hybridization, whereby genomic DNA is hybridized to immobilized bacterial artificial chromosome (BAC) or cDNA clones[38], to identify gene amplifications or deletions that are involved in tumour development. In addition to the analysis of cancer, microarrays have been used to analyse changes in gene expression in several other diseases, including muscular dystrophy[39], Alzheimer's disease[40], schizophrenia[41] and HIV infection[42].

Conclusions

The use of microarrays to explore gene expression on a global level is a rapidly evolving technology that seems set to become more powerful with the completion of the sequencing of the human genome. The biochemistry of the microarrays is proving very useful, and it seems likely that the significant advances in the next few years will come in the interpretation of the data sets generated. At present, it seems that we may only be scratching the surface when it comes to extracting useful information from these large quantities of data. This will be improved by advances in bioinformatics, but will also require more thoughtful experimental design. Whether or not microarray analysis is hypothesis-driven is open to debate, but if it is to fulfil its true potential it needs to progress beyond mere 'fishing expeditions'. Recent work in the field indicates that this is already beginning to occur, and that microarray technology is set to contribute much to the post-genomic future of biology. □

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270,** 467–470 (1995).
2. Wodicka, L., Dong, H., Mittmann, M., Ho, M. H. & Lockhart, D. J. Genome-wide expression monitoring in *Saccharomyces cerevisiae*. *Nature Biotechnol.* **15,** 1359–1367 (1997).
3. Kane, M. D. *et al.* Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res.* **28,** 4552–4557 (2000).
4. Luo, L. *et al.* Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nature Med.* **5,** 117–122 (1999).
5. Brady, G. Expression profiling of single mammalian cells — small is beautiful. *Yeast* **17,** 211–217 (2000).
6. Stears, R. L., Getts, R. C. & Gullans, S. R. A novel, sensitive detection system for high-density microarrays using dendrimer technology. *Physiol. Genomics* **3,** 93–99 (2000).
7. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405,** 827–836 (2000).
8. Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).
9. Bustin, S. A. Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J. Mol. Endocrinol.* **25,** 169–193 (2000).
10. Cho, R. J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2,** 65–73 (1998).
11. Cho, R. J. *et al.* Transcriptional regulation and function during the human cell cycle. *Nature Genet.* **27,** 48–54 (2001).
12. Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9,** 3273–3297 (1998).
13. Iyer, V. R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283,** 83–87 (1999).
14. Fambrough, D., McClure, K., Kazlauskas, A. & Lander, E. S. Diverse signaling pathways activated by growth factor receptors induce broadly overlapping, rather than independent, sets of genes. *Cell* **97,** 727–741 (1999).
15. Brazma, A. & Vilo, J. Gene expression data analysis. *FEBS Lett.* **480,** 17–24 (2000).
16. Gaasterland, T. & Bekiranov, S. Making the most of microarray data. *Nature Genet.* **24,** 204–206 (2000).
17. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95,** 14863–14868 (1998).
18. Toronen, P., Kolehmainen, M., Wong, G. & Castren, E. Analysis of gene expression data using self-organizing maps. *FEBS Lett.* **451,** 142–146 (1999).
19. Brown, M. P. *et al.* Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA* **97,** 262–267 (2000).
20. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.* **16,** 939–945 (1998).
21. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nature Genet.* **22,** 281–285 (1999).
22. Niehrs, C. & Pollet, N. Synexpression groups in eukaryotes. *Nature* **402,** 483–487 (1999).
23. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290,** 2306–2309 (2000).
24. Iyer, V. R. *et al.* Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409,** 533–538 (2001).
25. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126 (2000).
26. Marton, M. J. *et al.* Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* **4,** 1293–1301 (1998).
27. Brazma, A., Robinson, A., Cameron, G. & Ashburner, M. One-stop shop for microarray data. *Nature* **403,** 699–700 (2000).
28. Shoemaker, D. D. *et al.* Experimental annotation of the human genome using microarray technology. *Nature* **409,** 922–927 (2001).
29. Sallinen, S. L. *et al.* Identification of differentially expressed genes in human gliomas by DNA microarray and tissue chip techniques. *Cancer Res.* **60,** 6617–6622 (2000).
30. Golub, T. R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286,** 531–537 (1999).
31. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403,** 503–511 (2000).
32. Bittner, M. *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406,** 536–540 (2000).
33. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96,** 6745–6750 (1999).
34. Ross, D. T. *et al.* Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* **24,** 227–235 (2000).
35. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **24,** 236–244 (2000).
36. Schulze, A., Lehmann, K., Jefferies, H. B. J., McMahon, M. & Downward, J. Analysis of the transcriptional program induced by Raf in epithelial cells. *Genes Dev.* **15,** 981–994 (2001).
37. Clark, E. A., Golub, T. R., Lander, E. S. & Hynes, R. O. Genomic analysis of metastasis reveals an essential role for RhoC. *Nature* **406,** 532–535 (2000).
38. Pollack, J. R. *et al.* Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genet.* **23,** 41–46 (1999).
39. Chen, Y. W., Zhao, P., Borup, R. & Hoffman, E. P. Expression profiling in the muscular dystrophies. Identification of novel aspects of molecular pathophysiology. *J. Cell Biol.* **151,** 1321–1336 (2000).
40. Ginsberg, S. D., Hemby, S. E., Lee, V. M., Eberwine, J. H. & Trojanowski, J. Q. Expression profile of transcripts in Alzheimer's disease tangle-bearing CA1 neurons. *Ann. Neurol.* **48,** 77–87 (2000).
41. Mirnics, K., Middleton, F. A., Marquez, A., Lewis, D. A. & Levitt, P. Molecular characterization of schizophrenia viewed by microarray analysis of gene expression in prefrontal cortex. *Neuron* **28,** 53–67 (2000).
42. Geiss, G. K. *et al.* Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. *Virology* **266,** 8–16 (2000).