

COMMENTARY

Analysis of gene expression by microarrays: cell biologist's gold mine or minefield?

Almut Schulze and Julian Downward

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

Published on WWW 7 November 2000

SUMMARY

The development of DNA microarrays to study simultaneously the level of mRNA expressed from thousands of genes offers great promise to cell biologists. Microarrays can be used to gain detailed information about transcriptional changes involved in a specific pathway, potentially leading to the identification of novel components of the signalling system. They can also be used to obtain a fingerprint of the transcriptional status of the cell under a given condition, which may be useful for characterising the pathways used in response to novel stimuli. The use of microarrays will generate huge amounts

of expression data, contributing to the transformation of biology from a data-poor to a data-rich science. Whether this leads to real advances in the understanding of cell biological problems will depend on the development of methodologies, both in experimental biology and in bioinformatics, that allow meaningful knowledge to be extracted from all this information.

Key words: DNA microarrays, Gene expression, Cell biology, Cell signalling, Cancer

INTRODUCTION

During the five years or so since their introduction for use in gene expression monitoring, DNA microarrays have come a long way. From a few dozen cDNAs spotted on glass or nylon membranes to the completion of the 'full genome chip' for *S. cerevisiae*, we have reached a point at which genome-wide monitoring of expression levels of all human and mouse genes seems close at hand. With a working draft of the human genome already completed and the recently reported lower estimates for the total number of human genes (Aparicio, 2000) perhaps less than 40,000, this might well be achieved within the next couple of years. Nevertheless, with the technology available today, cell biologists can already generate large amounts of information about expression profiles of cells from different tissues or changes in gene expression as a result of a particular stimulus or treatment. At the same time, the increasing commercialisation of microarrays has brought their use within the reach of ordinary cell biology labs in academia. Here, we address the promises and pitfalls of this technology, in particular whether the huge amount of data that it can produce will help cell biologists or end up being a source of further confusion.

TOOLS OF THE TRADE

Although many different systems have been described by

various academic groups and commercial suppliers, DNA microarrays can mainly be divided into two groups, according to the arrayed material. Oligonucleotide-based arrays can either be produced by spotting pre-synthesised short oligonucleotides onto glass or synthesised in situ on the surface of silicon wafers by photo-lithography (often referred to as oligonucleotide chips). In contrast, cDNA microarrays consist of longer DNA fragments, either inserts from cDNA libraries or PCR products generated from gene-specific primers, which are printed onto glass slides or nylon membranes (Granjeaud et al., 1999). Oligonucleotide chips do not require the sometimes problematic access to large collections of unambiguous cDNA clones, since they can be generated from sequence information alone, and are potentially able to distinguish between genes that have high sequence similarity and splice variants. However, oligonucleotide chips require more specialised equipment than cDNA microarrays to manufacture.

In general, the arrayed material has been termed the probe, because it is equivalent to a probe used for northern analysis, whereas the material that is hybridised to the microarray has been termed the target. To generate the target, mRNA from cells or tissue is labelled either during first strand cDNA synthesis or by amplification protocols, which can involve several rounds of in vitro transcription. The target is then hybridised to the probes on the array surface and detected by fluorescence scanning or phosphorimaging. The high

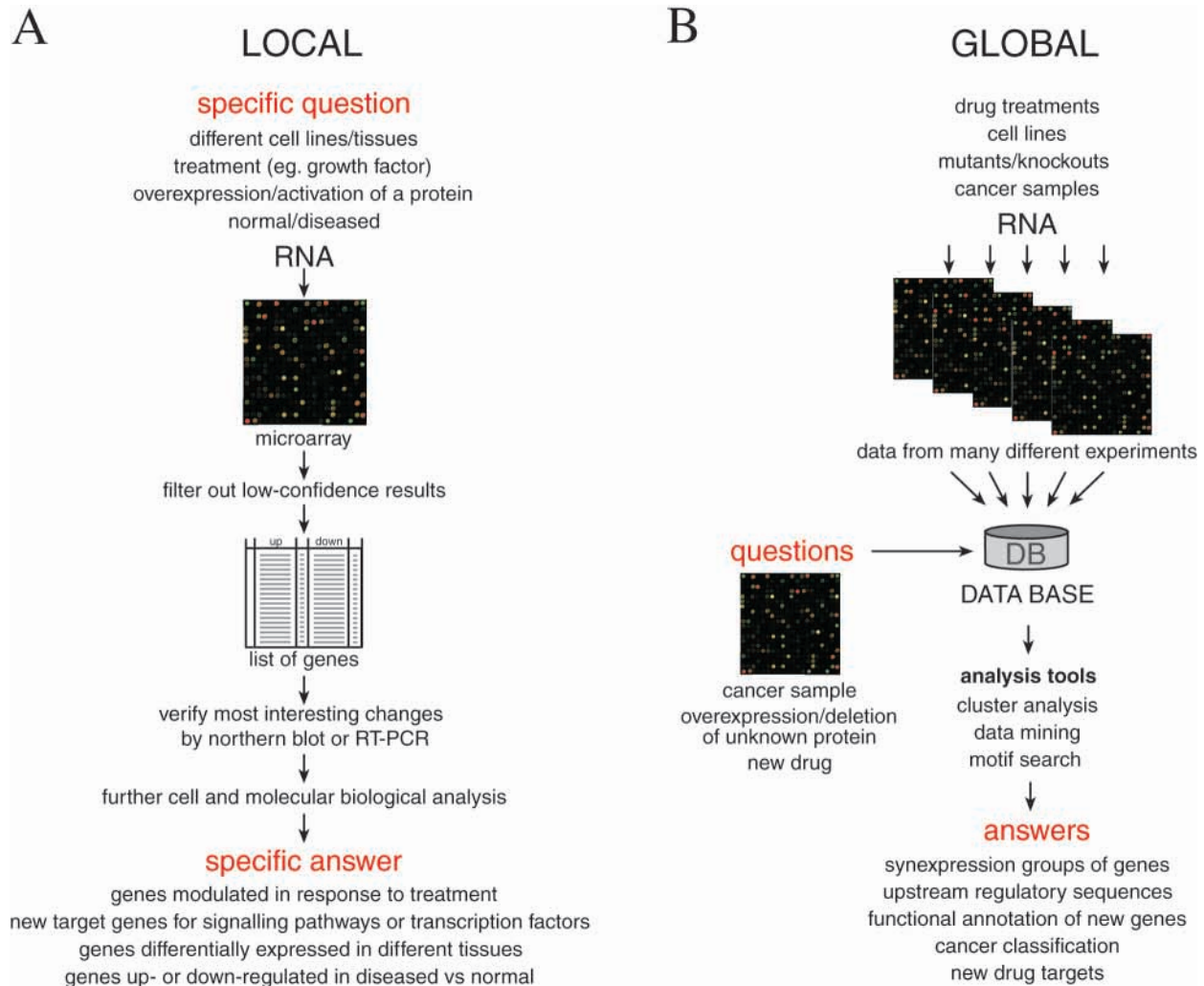


Fig. 1. Cell biologists perspectives of microarray experiments. (A) Local view: To answer a specific biological question, DNA microarrays can be employed to generate lists of genes that are modulated in cells in response to a particular treatment, between different cell lines, tissues or tumour samples or as a result of overexpression or activation of a protein. Conceptually these types of experiments resemble other comparative techniques, such as differential display or subtraction cloning. (B) Global view: Initially, large amounts of data generated in potentially unrelated sets of experiments by different experimenters are accumulated and stored in central repositories. Once data sets are standardised to allow cross comparison, researchers can query the database, using analysis tools to answer different biological questions. Genes that are modulated in the same way in a variety of experiments can be identified as being co-regulated, which can allow functional classification as well as potential identification of common regulatory elements in promoter sequences. In addition, different samples can be grouped according to similarities in their gene expression profiles – for example, for the molecular classification of cancer types. Data from new experiments (e.g. an unknown cancer sample or a mutation in an unknown gene) can be compared with the database, allowing identification of matching profiles that allow conclusions about the nature of a sample or function of a gene.

reproducibility of in situ synthesis of oligonucleotide chips allows comparison of signals generated by samples hybridised to separate arrays. In the case of spotted arrays, this is in general not possible. The use of different fluorescent dyes (Cy3 and Cy5) allows mRNAs from two different cell populations or tissues to be labelled in different colours, mixed and hybridised to the same array, which results in competitive binding to the probes. Calculation of the difference in signal intensities of the two dyes generates the relative change in mRNA abundance between the two samples for each gene present on the array.

Acquiring enough mRNA to perform the analysis can be a problem in some circumstances. Experiments involving tissue

culture cells usually do not present a problem, but when material is to be obtained from a particular cell population within a tissue, low-yield methods such as laser-capture-assisted microdissection can be needed to assure purity of the cells analysed. In order to generate enough target to profile, extensive amplification is required; protocols for this have been successfully developed (Luo et al., 1999), but it remains unclear whether the amplification introduces significant experimental bias into the target sample (Lockhart and Winzeler, 2000).

Pairwise comparison of mRNA populations, either of two samples or of several samples to a common control, can be very useful for many biological questions: for example, in analysis

of the response of cells to a given stimulus or the consequences of ectopic expression or activation of one particular protein in a given cellular background (see Fig. 1A). It is very likely that microarrays will replace other comparative techniques such as differential display or subtraction cloning, once full genome microarrays are available. Studies using this 'local' approach (see below) have provided considerable insight into particular cellular pathways: for example, previously unrecognised potential target genes for the transcription factors MYC, WT1 and p53 were identified (Lee et al., 1999; Collier et al., 2000; Zhao et al., 2000). This type of experiment typically yields a list of genes that are found to be up- or down-regulated in the sample. The challenge is then to filter out false positives. Many investigators have found variability of microarray data to be high, especially among genes expressed at low levels. Increasing the number of replicate experiments is an advantage but might be problematic owing to high cost or limited starting material. Verification of hits by different techniques, such as northern analysis or RT-PCR is very important, but can become unfeasible for large numbers of genes. Also data about genes that are found to be unchanged have to be treated with particular care because reliable detection of differential expression of each gene must be established first: some probes might simply not respond – for example, owing to adverse secondary structure. Although variability resulting from experimental noise can be a significant problem, and unambiguous detection of expression levels of all genes represented on the microarray might not be proven, further testing and improvement of microarray technology, in particular in the choice of probes, will improve confidence in the data.

A more conceptual problem with the interpretation of expression data for identification of new target genes for signalling pathways or known transcription factors is the discrimination between primary events and secondary events, which result from the accumulation of the gene products of the direct target genes. Drugs that inhibit protein synthesis, although useful for the study of individual genes, can be misleading because they can have significant effects on gene expression by themselves. It is also important to realise that the contribution of post-transcriptional mechanisms to gene expression is significant. Investigators have attempted to monitor only mRNAs that are translated, by using purification and polysome-bound mRNA (Johannes et al., 1999). This technique offers an interesting addition to the tools available for analysing the proteome in addition to the transcriptome of cells or tissues – in particular because other proteomics techniques do not yet offer the resolution and sensitivity of microarray-based mRNA detection.

DNA microarrays also offer the opportunity to compare large numbers of samples to generate a more global view of a biological system. In this type of experiment, many samples are compared with each other or, in the case of comparative two-colour hybridisation, compared with a common control, which can be a mixture of all samples used. Owing to the large amount of information, sophisticated data analysis and data mining become indispensable (see Fig. 1B). During the past few years much effort has been made to develop and improve mathematical models that can be used to sieve through those mounds of data. In particular, different cluster-analysis tools that assemble genes into groups that have similar expression patterns are available

(Sherlock, 2000; see also rana.stanford.edu/clustering/ and www.genome.wi.mit.edu/MPR/). These algorithms have been used to find groups of genes that exhibit similar temporal expression patterns during the yeast cell cycle (Cho et al., 1998; Spellman et al., 1998) or after growth factor treatment in fibroblasts (Iyer et al., 1999). The same approach can reveal relationships between different RNA samples according to similarities in the expression levels of a large number of genes. Samples from different cell types of the lymphatic system or different cancer types have been successfully classified by mRNA profiling (Golub et al., 1999; Alizadeh et al., 2000). Although these methods have the potential to generate new insight into complex biological systems, they are also potentially inaccessible to a wider community of experimental biologists. Therefore close collaboration between experimenters, mathematicians and informaticists becomes a key issue in the development and meaningful use of powerful tools for the analysis of gene expression data.

IMPLICATIONS FOR THE CELL BIOLOGIST

As outlined above, analysis of gene expression using microarrays can be used to investigate cell behaviour at two different levels, either 'local' or 'global'. In the former case, the aim is to discover the specific gene(s) whose change in expression level is responsible for the establishment of a given phenotype in response to a particular stimulus. By contrast, the latter attempts to establish the overall architecture of genetic regulatory networks, ultimately trying to elucidate the complete wiring diagram of the cell.

Microarrays in the study of local signalling events

The use of gene expression microarrays by cell biologists is most likely to focus on the 'local', at least initially (see Fig. 1A). The connection between an input to the cell – for example, treatment with a growth factor or adhesion to the extracellular matrix – and a change in cell behaviour, such as induction of proliferation or protection from apoptosis, involves both direct signalling events and changes in gene expression. The direct events may involve, for example, activation of kinase cascades or ion fluxes, and these can in part mediate their effects through changes in gene expression, as well as through post-translational modulation of protein function. We know much about early direct signalling events, but much less about changes in gene expression. It will be attractive to many to catalogue the effects of stimuli or conditions of interest on the gene expression programme of the cell. Although this is becoming technically feasible, serious thought needs to be given to how we can take this beyond simply creation of lists to a greater understanding of signalling pathways and their functions. One study addressing regulation of gene expression by early signalling events analysed changes in transcript levels of nearly 6000 mouse genes in response to activation of mutant forms of the PDGF receptor (Fambrough et al., 1999). The authors found that different mutants of this receptor do not in fact activate different signalling pathways, and consequently modulate expression of distinct sets of genes, but instead actually induce broadly overlapping expression patterns. This result shows that gene expression analysis by microarrays could change our view of signalling processes as linear

pathways and provide more insight into the cross-connections within signalling networks (Pawson and Saxton, 1999).

In common with other 'unbiased' approaches, such as yeast two-hybrid screening, microarrays generate large quantities of data very rapidly, but often require a great deal of subsequent analysis for investigators to understand the significance, if any, of the hits. As microarrays become more comprehensive in terms of the proportion of the expressed genome represented, so will the lists of genes whose expression levels change significantly grow in length. Faced with several hundreds or even thousands of genes showing major changes in mRNA level, the cell biologist must try to pinpoint those changes that are significant to the biological response under investigation. Even after artefacts are weeded out, the list of genes is likely to be too long for each hit to be analysed individually. The obvious reaction of most will be to scan the list looking for genes that fit preconceived notions of how the system might work; immediately the benefits of an unbiased approach will be jeopardised, because connections beyond our current understanding may well be ignored! Improved experimental design is likely to be essential in this regard, with particular attention being paid to performing time courses and dose responses, and using the simplest possible experimental models.

What can be done next with the initial list of regulated genes? A shortlist of genes most likely to be important needs to be drawn up, ideally through the use of a secondary screen, if available; alternatively, this can be done through judicious analysis of published literature. This step can often take one into unfamiliar areas and be extremely time consuming; the use of advanced literature-database-searching tools such as Medminer (discover.nci.nih.gov/textmining/filters.html) can be very useful at this stage. Finally, the functions of genes on the shortlist need to be modulated artificially to reveal whether the generation of the biological end point can be influenced. A number of technical and conceptual problems can arise here. One is that it is currently difficult to modulate gene expression or protein function on a broad scale, at least in mammals: using dominant negative mutants, antisense or knockout technology, one can address the functions of a handful of genes, but not dozens or hundreds. Future technical advances may improve this; the situation in lower eukaryotic model systems has been revolutionised by the advent of RNAi and systematic gene-knockout resources. At present, just getting hold of more than a handful of full-length cDNAs in convenient vectors is a major undertaking, although this seems set to change with the establishment of the Harvard Institute of Proteomics cDNA archive (www.hip.harvard.edu/). In addition, Invitrogen sells many full-length cDNAs (www.invitrogen.com/genestorm/), although at prices that would rule out bulk purchase for most laboratories. A number of smaller genes are also present as full-length EST clones.

Even if the functions of relatively large numbers of genes can be modulated by the experimenter, there is the concern that ablation of the activities of many gene products causes the failure of a signalling pathway even when the protein in question is not directly on the pathway. On the other hand, in complex pathways, activation of a single gene may not be sufficient by itself 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. In addition, gene expression levels need to be controlled by the experimenter within the normal physiological range if overexpression artefacts are to be avoided. Of course, one must also remember that microarrays will not directly address changes in protein function, especially those caused by post-translational modification. These can be analysed systematically through proteomic approaches, such as coupling two-dimensional gel electrophoresis with mass spectrometry. At present, proteomic analysis, especially that focusing on post-translational modification, is much less sensitive than analysis of mRNA levels using microarrays; the fundamental differences between nucleic acid and protein chemistry suggest this is likely to remain the case for some time.

Microarrays in the analysis of global cell behaviour

Microarrays will also be used from an increasingly global perspective in cell biology. Generation of databases of changes in gene expression in response to many varied stimuli will allow comparison of patterns of regulatory control. For example, genes that are coordinately regulated can be identified by cluster analysis: these have been termed synexpression groups (Niehrs and Pollet, 1999). Coordinate regulation is likely to imply that genes share common transcriptional control pathways and might have similar cis-regulatory elements upstream of the coding region of the gene. Again, studies in higher eukaryotes are lagging behind the yeast field. Several groups have generated databases containing genome-wide expression profiles for mutants, different growth conditions, stress and drug treatments, which can be queried by different investigators in different ways (Fig. 1B). These studies have shown that it is possible to identify common DNA motifs in the promoter regions of co-regulated genes (Cho et al., 1998; Roth et al., 1998; Spellman et al., 1998; Tavazoie et al., 1999). This will eventually lead to the identification of binding sites for known or novel transcription factors that are implicated in the transcription of a co-regulated set of genes. Recently, an example of this type of analysis has seen the combination of hundreds of data sets representing the gene expression profiles of yeast following various mutations or drug treatments. This compendium is compared with expression profiles of cells that have mutations in genes of unknown function; in many cases the function of the gene of interest can be predicted with considerable accuracy by identification of which known expression profile they most closely match (Hughes et al., 2000). Although synexpression groups have been identified in higher eukaryotes, the organisation of regulatory elements is significantly more complex than in yeast, and application of similar pattern-recognition methods to identify common motifs in mammalian promoters remains a challenge.

Genes in coordinately regulated clusters might not only have similar control mechanisms but also participate in similar cellular processes or pathways. In completely sequenced organisms such as yeast, this possibility is providing useful hints to the function of uncharacterised genes. However, the reliability of any correlation between expression pattern and function has yet to be established (Lockhart and Winzler, 2000). Nevertheless, this approach will become more prevalent in mammalian systems following completion of the sequencing and annotation of the human genome.

Ultimately global analysis of gene expression might generate detailed maps of genetic regulatory networks. It will allow genome databases to be annotated with detailed information about the regulatory pathways controlling the expression of each gene, even when those genes have not been studied individually, and could provide insight into gene function beyond that provided by coding sequence homology. These advances will all benefit cell biologists, but it may take some time before this information is available in sufficient amounts and well organised enough to be analysed usefully. The establishment of publicly available databases of microarray data in standardised format will be essential (Brazma et al., 2000; see www.ebi.ac.uk/arrayexpress/ and www.ncbi.nlm.nih.gov/geo/).

Microarrays will also have several uses that are more applied, but still likely to be important to cell biologists. One such area is in diagnosis of disease. This has received particular attention in the case of cancer, where microarray analysis has proven capable of distinguishing subtypes of leukemia and lymphoma (Golub et al., 1999; Alizadeh et al., 2000). Gene expression profiles have also been used to classify 60 cancer cell lines according to their tumour origin (Ross et al., 2000) and to discriminate between normal and colon-tumour-derived tissue (Alon et al., 1999). In the case of solid tumours, homogeneity of the tissue sample is crucial and laser capture micro-dissection has been used to analyse selectively gene expression in cells from invasive and metastatic breast cancer tissue sections (Sgroi et al., 1999). In addition, detection of changes in gene copy number due to amplification or deletion in tumour cells by comparative genomic hybridisation (CGH) is another promising application in the analysis of tumorigenesis (Pollack et al., 1999). Another approach to dealing with possible sample heterogeneity has been developed (Perou et al., 2000); this uses downstream bioinformatics analysis to factor out the contribution of contaminating cell types – for example, stromal or lymphoid cells in a tumour. Profiling in these cases can concentrate on only those genes known to be expressed principally by the tumour cells and not the contaminating cell types, the ‘intrinsic gene subset’. Although this may greatly reduce the number of genes that can be analysed, it still leaves plenty to provide a useful profile for comparison against other tumours. These studies show that microarray technology coupled with sophisticated sample preparation and data analysis tools will probably yield novel information about the nature of genes involved in tumour development, invasiveness and metastasis and in sensitivity to therapy.

Another use for microarrays in cancer research is the investigation of the mechanism of action of drugs. Treatment of cells with a drug will give a profile of changes in expression that will be similar for other drugs with related mechanisms of action. Genetic removal of the target of the drug should prevent all the changes in expression caused by the drug's action on this target, leaving only ‘off-target’ effects (Marton et al., 1998). This type of analysis will be beneficial for pharmaceutical companies in the development of improved drugs, but would also probably be very sobering if applied to the typical pharmacological tool box used by cell biologists in studying signal transduction. In a recent study, Scherf et al. correlated the growth inhibitory action of over 70,000 chemical compounds on 60 different cancer cell lines with their

expression profiles to identify genes that are implicated in drug sensitivity or resistance (Scherf et al., 2000). This led to the generation of a database containing information about drug sensitivity as well as gene expression profiles for all the cell lines used, which can be correlated with additional data about the cells, such as genomic alterations or activation of known oncogenes. Comparing the expression profile of a new cell line or tumour sample with the expression profiles in the NCI-60 database can allow predictions to be made about drug sensitivity or resistance, without one having to screen the entire compound collection.

This emphasises the usefulness of large databases of gene expression profiles from different cells, tumour samples or mutants (see Fig. 1B). The database compendium approach (Hughes et al., 2000) mentioned earlier has amply demonstrated the power of this methodology in yeast. Similar approaches will be very useful in establishing functions for mammalian genes, although techniques for the deletion of large numbers of genes are significantly more challenging, and caveats about functional redundancy have to be considered.

CONCLUSIONS

Gene expression profiling by the use of microarrays is a rapidly evolving technology that will have a major impact on all areas of biology. Following the completion of the sequence of the human genome, its influence will become more profound. Developments in yeast, and to some extent in mammals, show that microarrays will have enormous value in analysing the function of uncharacterised genes and environmental influences. However, the danger that microarrays will lead to a glut of data that cannot be interpreted meaningfully is very real. Careful design of experiments to address specific questions, rather than blind fishing for ‘interesting’ genes, will be of utmost importance. In addition, standardisation of protocols and databases with public access will be essential if the most is to be made from this technology.

REFERENCES

- Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X. et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511.
- Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Nat. Acad. Sci. USA* **96**, 6745–6750.
- Aparicio, S. A. (2000). How to count human genes. *Nature Genet.* **25**, 129–130.
- Brazma, A., Robinson, A., Cameron, G. and Ashburner, M. (2000). One-stop shop for microarray data. *Nature* **403**, 699–700.
- Cho, R. J., Campbell, M. J., Winzler, E. A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T. G., Gabrielian, A. E., Landsman, D., Lockhart, D. J. and Davis, R. W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73.
- Coller, H. A., Grandori, C., Tamayo, P., Colbert, T., Lander, E. S., Eisenman, R. N. and Golub, T. R. (2000). Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc. Nat. Acad. Sci. USA* **97**, 3260–3265.
- Fambrough, D., McClure, K., Kazlauskas, A. and Lander, E. S. (1999). Diverse signaling pathways activated by growth factor receptors induce

- broadly overlapping, rather than independent, sets of genes. *Cell* **97**, 727-741.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S.** (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531-537.
- Granjeaud, S., Bertucci, F. and Jordan, B. R.** (1999). Expression profiling: DNA arrays in many guises. *BioEssays* **21**, 781-790.
- Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D. et al.** (2000). Functional discovery via a compendium of expression profiles. *Cell* **102**, 109-126.
- Iyer, V. R., Eisen, M. B., Ross, D. T., Schuler, G., Moore, T., Lee, J. C. F., Trent, J. M., Staudt, L. M., Hudson, J. Jr, Boguski, M. S., Lashkari, D., Shalon, D., Botstein, D. and Brown, P. O.** (1999). The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83-87.
- Johannes G., Carter, M. S., Eisen, M. B., Brown, P. O., Sarnow, P. and Botstein, D.** (1999). Identification of eukaryotic mRNAs that are translated at reduced cap binding complex eIF4F concentrations using a cDNA microarray. *Proc. Nat. Acad. Sci. USA* **96**, 13118-13123.
- Lee, S. B., Huang, K., Palmer, R., Truong, V. B., Herzlinger, D., Kolquist, K. A., Wong, J., Paulding, C., Yoon, S. K., Gerald, W., Oliner, J. D. and Haber, D. A.** (1999). The Wilms tumor suppressor WT1 encodes a transcriptional activator of amphiregulin. *Cell* **98**, 663-673.
- Lockhart, D. J. and Winzeler, E. A.** (2000). Genomics, gene expression and DNA arrays. *Nature* **405**, 827-836.
- Luo, L., Salunga, R. C., Guo, H., Bittner, A., Joy, K. C., Galindo, J. E., Xiao, H., Rogers, K. E., Wan, J. S., Jackson, M. R. and Erlander, M. G.** (1999). Gene expression profiles of laser-captured adjacent neuronal subtypes. *Nature Med.* **5**, 117-122.
- Marton, M. J., DeRisi, J. L., Bennett, H. A., Iyer, V. R., Meyer, M. R., Roberts, C. J., Stoughton, R., Burchard, J., Slade, D., Dai, H., Bassett, D. E. Jr, Hartwell, L. H., Brown, P. O. and Friend, S. H.** (1998). Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* **4**, 1293-1301.
- Niehrs, C. and Pollet, N.** (1999). Synexpression groups in eukaryotes. *Nature* **402**, 483-487.
- Pawson, T. and Saxton, T. M.** (1999). Signaling networks – do all roads lead to the same genes? *Cell* **97**, 675-678.
- Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A. et al.** (2000). Molecular portraits of human breast tumours. *Nature* **406**, 747-752.
- Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D. and Brown, P. O.** (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genet.* **23**, 41-46.
- Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J. C., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D. and Brown, P. O.** (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.* **24**, 227-235.
- Roth, F. P., Hughes, J. D., Estep, P. W. and Church, G. M.** (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature Biotechnol.* **16**, 939-945.
- Scherf, U., Ross, D. T., Waltham, M., Smith, L. H., Lee, J. K., Tanabe, L., Kohn, K. W., Reinhold, W. C., Myers, T. G., Andrews, D. T. et al.** (2000). A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **24**, 236-244.
- Sgroi, D. C., Teng, S., Robinson, G., LeVangie, R., Hudson, J. R. Jr and Elkhouloun, A. G.** (1999). In vivo gene expression profile analysis of human breast cancer progression. *Cancer Res.* **59**, 5656-5661.
- Sherlock, G.** (2000). Analysis of large-scale gene expression data. *Curr. Opin. Immunol.* **12**, 201-205.
- Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B.** (1998). Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273-3297.
- Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. and Church, G. M.** (1999). Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281-285.
- Zhao, R., Gish, K., Murphy, M., Yin, Y., Notterman, D., Hoffman, W. H., Tom, E., Mack, D. H. and Levine, A. J.** (2000). Analysis of p53-regulated gene expression patterns using oligonucleotide arrays. *Genes Dev.* **14**, 981-993.