Expression profiling using cDNA microarrays

David J. Duggan, Michael Bittner, Yidong Chen, Paul Meltzer & Jeffrey M. Trent

Cancer Genetics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. e-mail: jtrent@nhgri.nih.gov

cDNA microarrays are capable of profiling gene expression patterns of tens of thousands of genes in a single experiment. DNA targets, in the form of 3´ expressed sequence tags (ESTs), are arrayed onto glass slides (or membranes) and probed with fluorescent- or radioactively-labelled cDNAs. Here, we review technical aspects of cDNA microarrays, including the general principles, fabrication of the arrays, target labelling, image analysis and data extraction, management and mining.

Ambitious projects aimed at cloning, mapping and sequencing the genomes of various organisms, including that of *Homo sapiens*, have been launched worldwide. In all cases, the fruits of these labours will provide a solid platform from which to attempt the larger goal of understanding how genomes result in the organisms they specify. The success of these international efforts is impressive. So far, complete genomic sequences of 17 organisms, including the eukaryote *Saccharomyces cerevisiae*, have been produced. The mapping (both genetic and physical) and sequencing phases of the Human Genome Project are ahead of schedule. Researchers have catalogued more than 1.1 million expressed sequence tagged sites (ESTs), corresponding with 52,907 unique human genes[1] (www.ncbi.nlm.nih.gov/UniGene). However, the function, expression and regulation of more than 80% of them has yet to be fathomed. The next phase of the human genome project will place strong emphasis on assigning function to these genes.

The ability to identify genes at the nucleic acid level rather than proceeding from a known protein to its chromosomal counterpart has prompted efforts to likewise extract functional information at the nucleic acid level. Two methods are currently in use. The 'sequence' approach has led to the discovery of a wide variety of sequence motifs encoding structural domains, such as DNA-binding and nucleotide-binding domains[2], thus providing clues to gene function. Another route for exploring the function of a gene is by determining its pattern of expression. The accumulation of expression data has yet to reach the point at which it is possible to speak of expression motifs, but it does suggest that this is a plausible outcome of the approach[3–5].

Various methods are available for detecting and quantitating gene expression levels, including northern blots[6], S1 nuclease protection[7], differential display[8], sequencing of cDNA libraries[9,10] and serial analysis of gene expression[11] (SAGE). Augmenting this coterie are two array-based technologies—cDNA and oligonucleotide arrays. These allow one to study expression levels in parallel[3,12,13], thus providing static information about gene expression (that is, in which tissue(s) the gene is expressed) and dynamic information (that is, how the expression pattern of one gene relates to those of others). The high degree of digital data extraction and processing of these techniques supports a variety of samples or experimental conditions.

Although both cDNA and oligonucleotide arrays are capable of analysing patterns of gene expression, fundamental differences exist between the methods. Here, we focus primarily on technical aspects of cDNA microarrays, although some comparison with the oligonucleotide array (see page 20 of this issue (ref. 14)) will be made where appropriate.

Principle of method

As reviewed by Ed Southern on page 5 of this issue, hybridization between nucleic acids (one of which is immobilized on a matrix) provides a core capability of molecular biology[15]. This method provides high sensitivity and specificity of detection as a consequence of exquisite, mutual selectivity between complementary strands of nucleic acids. Historically, most applications of this method have employed a single, pure, labelled oligonucleotide or polynucleotide species in the liquid phase and complex mixtures of polynucleotides attached to a solid support. Transcript abundance is assayed by immobilizing mRNA or total RNA (electrophoretically separated or in bulk) on membranes and then incubating with a radioactively labelled, gene-specific target. If multiple RNA samples are immobilized on the same matrix, one obtains information about the quantity of a particular message present in each RNA pool.

cDNA arrays alter this strategy in several ways (Fig. 1). In an array experiment, many gene-specific polynucleotides derived from the 3´ end of RNA transcripts are individually arrayed on a single matrix. This matrix is then simultaneously probed with fluorescently tagged cDNA representations of total RNA pools from test and reference cells, allowing one to determine the relative amount of transcript present in the pool by the type of fluorescent signal generated. Relative message abundance is inherently based on a direct comparison between a 'test' cell state and a 'reference' cell state; an internal control is thus provided for each measurement (Fig. 2). The scheme is similar when using radiolabelled probe, but it is not possible to carry out simultaneous hybridization of test and reference samples. In such cases, serial or parallel hybridization is required, introducing the possibility of higher variability in comparisons of expression level.

The adaptable nature of the fabrication and hybridization methods allows the technique to be applied widely—the only limitations are the availability of clones for the solid phase and the quality of RNA samples derived from the cells (or tissues) to be compared. This is illustrated by diverse applications that include: investigating gene expression in the roots and leaves of *Arabidopsis thaliana*[3], human T cells exposed to phorbol ester[12], rheumatoid arthritis and inflammatory bowel disease[16], tumorigenic versus non-tumorigenic cell lines[4], the diauxic shift from anaerobic to aerobic metabolism in *S. cerevisiae*[5,17] (yeast),

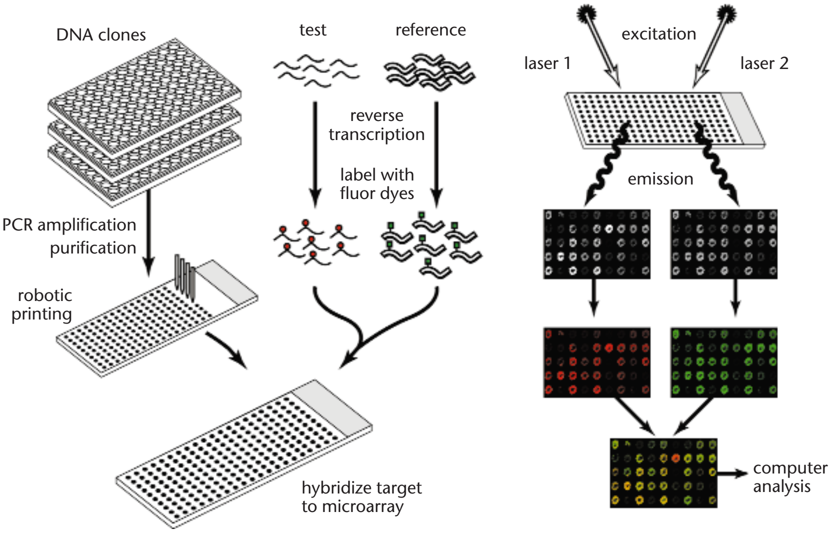

Fig. 1 cDNA microarray schema. Templates for genes of interest are obtained and amplified by PCR. Following purification and quality control, aliquots (~5 nl) are printed on coated glass microscope slides using a computer-controlled, high-speed robot. Total RNA from both the test and reference sample is fluorescently labelled with either Cye3- or Cye5-dUTP using a single round of reverse transcription. The fluorescent targets are pooled and allowed to hybridize under stringent conditions to the clones on the array. Laser excitation of the incorporated targets yields an emission with a characteristic spectra, which is measured using a scanning confocal laser microscope. Monochrome images from the scanner are imported into software in which the images are pseudo-coloured and merged. Information about the clones, including gene name, clone identifier, intensity values, intensity ratios, normalization constant and confidence intervals, is attached to each target. Data from a single hybridization experiment is viewed as a normalized ratio (that is, Cye3/Cye5) in which significant deviations from 1 (no change) are indicative of increased (>1) or decreased (<1) levels of gene expression relative to the reference sample. In addition, data from multiple experiments can be examined using any number of data mining tools.

murine T cells challenged with 4-phorbol-12-myristate-13-acetate[13] and in *Streptococcus pneumoniae*[18].

Fabrication

Production of arrays begins with the selection of the 'probes' to be printed on the array. In many cases, these are chosen directly from databases including GenBank (ref. 19), dbEST (ref. 20) and UniGene (ref. 1), the resource backbones of the array technologies (see page 25 of this issue (ref. 21)). Additionally, full-length cDNAs, collections of partially sequenced cDNAs (or ESTs), or randomly chosen cDNAs from any library of interest can be used. Arrays for higher eukaryotes are typically based on the EST portions of these projects, whereas for yeast and prokaryotes, probes are usually generated by amplifying genomic DNA with gene-specific primers. Given the expense of obtaining clones, producing DNA from them, and printing them, it is usually preferable to produce arrays with a low redundancy of representation, so as to survey the broadest possible set of genes.

In this regard, the human UniGene database represents an excellent model of the kind of informational base one needs both to choose clones and to evaluate expression profiles. It includes a summary of information about the function of a particular gene, its genomic location, clones that contain the gene and connections to other relevant databases and literature sources. On the other hand, no other organisms have such a well-developed EST database, a limitation, given that cDNA microarrays also permit the 'assay' of uncharacterized cDNAs (which may represent genes with informative expression patterns).

cDNA arrays are produced by spotting PCR products (of approximately 0.6–2.4 kb) representing specific genes onto a matrix. These are usually generated from purified templates, so that cellular contaminants do not find their way onto the array. Typically, the PCR product is partially purified by precipitation, gel-filtration, or both —to remove unwanted salts, detergents, PCR primers and proteins present in the PCR cocktail. For both glass and membrane matrices, each array element is generated by the deposition of a few nanoliters of purified PCR product, typically of 100–500 µg/ml (see page 18 of this issue (ref. 22)) Printing is carried out by a robot that spots a sample of each gene product onto a number of matrices in a serial operation. The first spotting robots relied on contact printing with a device not unlike a fountain pen. Many variations on this design are now available (see page 31 of this issue (ref. 21)), in addition to a 'spotter' that is essentially a capillary tube, to which a low but constant pressure is applied. Non-contact printing modes, using either piezo or ink-jet devices, are also being evaluated.

The types of membranes commonly used are nitrocellulose and charged nylon commercial varieties that are used for various blotting assays. Glass-based arrays are most often made on microscope slides, which have low inherent fluorescence. These are coated with poly-lysine, amino silanes or amino-reactive silanes[12], which enhance both the hydrophobicity of the slide and the adherence of the deposited DNA. They also limit the spread of the spotted DNA droplet on the slide.

In most cases, DNA is cross-linked to the matrix by ultraviolet irradiation. After fixation, residual amines on the slide surface are reacted with succinic anhydride to reduce the positive charge at the surface. As a final step, some percentage of the DNA deposited is rendered single-stranded by heat or alkali (see page 19 of this issue for a detailed description of procedures[22]). The state of bound DNA is ill-defined. It is deposited in double-stranded form, intra-strand cross-linked to some extent, and may well have multiple constraining contacts with the matrix along its length (induced by drying the DNA onto the matrix; Fig. 3). It is therefore probably not the best hybridization probe. One can imagine that oligonucleotide matrices, with their short chains and single points of constraint at each chain end, may well be a far more accessible probe for hybridization. Against this advantage, however, must be weighed the disadvantages of using short-chain detectors. Chief among these are the variations in melting temperature due to AT–GC composition, and the reduction in specificity due to truncating the number of nucleotides from hundreds to as few as twenty. A format in which the accessibility of a simply tethered, single-stranded probe could be combined with the specificity of a long probe would provide a considerable improvement for the field.

Target labelling and hybridization

The targets for arrays are labelled representations of cellular mRNA pools. Typically, reverse transcription from an oligo-dT primer is used. This has the virtue of producing a labelled product from the 3′ end of the gene, directly complementary to immobilized targets synthesized from ESTs. Frequently, total RNA pools (rather than mRNA selected on oligo-dT) are labelled, to maximize the amount of message that can be obtained from a given

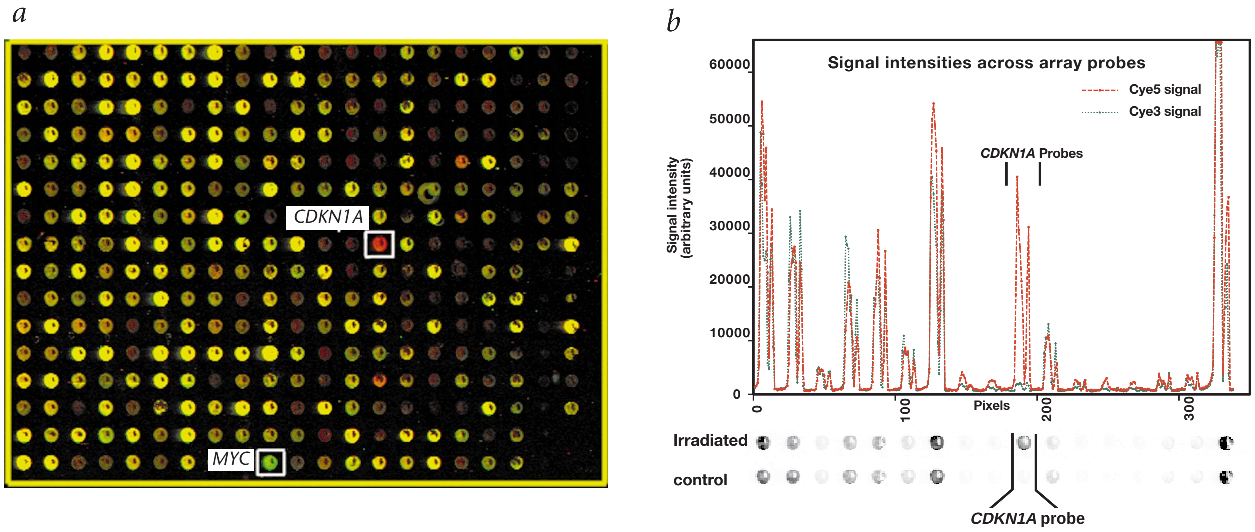

Fig. 2 Quantitations from two-colour hybridization. ***a***, A segment of an array to which targets from γ irradiated ML1 cells (red) and untreated ML1 cells (green) are hybridized. Highly differential hybridization is visible at the detectors for *CDKN1A* and *MYC* (boxed). ***b***, Intensity along a horizontal axis running through *CDKN1A* and several detectors on either side. The intensity profiles are nearly coincident at each probe except *CDKN1A*. At *CDKN1A*, the signal from the uninduced cells is near the threshold of detection, whereas the signal from the induced cells is considerably greater.

amount of tissue. The purity of RNA is a critical factor in hybridization performance, particularly when using fluorescence, as cellular protein, lipid and carbohydrate can mediate significant non-specific binding of fluorescently labelled cDNAs to slide surfaces. For radioactive detection, ^{33}P dCTP is preferred to more energetic emitters, as array elements are physically close to each other and strong hybridization with a radioactive target can easily interfere with detection of weak hybridization in surrounding targets. As fluorescent labels, Cye3-dUTP and Cye5-dUTP are frequently paired, as they have relatively high incorporation efficiencies with reverse transcriptase, good photostability and yield, and are widely separated in their excitation and emission spectra, allowing highly discriminating optical filtration.

A clear limitation to the application of this technology is the large amount of RNA required per hybridization. For adequate fluorescence, the total RNA required per target, per array, is 50–200 µg (2–5 µg are required when using poly(A) mRNA). For mRNA present as a single transcript per cell (approximately 1 transcript per 100,000), application of target derived from 100 µg of total RNA over an 800 mm^2 hybridization area containing 200-µm diameter probes will result in approximately 300 transcripts being sufficiently close to the target to have a chance to hybridize. Thus, if the fluorescently tagged transcripts are, on average, 600 bp, have an average of 2 fluor tags per 100 bp and hybridize—all of them—to their probe, approximately 12 fluors will be present in a 100-µm^2 scanned pixel from that probe. Such low levels of signal are at the lower limit of fluorescence detection, and could easily be rendered undetectable by assay noise. Although radioactive targets may have a higher intrinsic detectability, they too reach a level of dilution that prohibits effective detection, thus precluding experimentation on very small numbers of cells (Fig. 4).

A variety of means by which to improve signal from limited RNA has been proposed. These are being evaluated by our laboratory and many others. Efficient mixing of the hybridization fluid should bring more molecules into contact with their cognate probe, increasing the number of productive events. This entails, however, a larger 'mixing' volume, which might offset the potential gain. Methods that produce multiple copies of mRNA using highly efficient phage RNA polymerases have been developed[23]. A version of this approach, in which labelled target (cRNA) is made directly from a cDNA pool, having a T7 RNA polymerase promoter site at one end via *in vitro* transcription, has been applied to arrays[13]. Post-hybridization amplification methods have also been reported in which detectable molecules are precipitated at the target by the action of enzymes 'sandwiched' to the cDNA target[24]. Detection of hybridized species using mass spectroscopy or local changes in electronic properties can also be imagined[25,26].

Image analysis and data extraction

The highly regular arrangement of detector elements and crisply delineated signals that result from robotic printing and confocal imaging of fluor-detected arrays renders image data amenable to extraction by highly developed, digital image processing procedures. Grids specifying target locations can be readily overlaid on the images. Local sampling of background can be used to specify a threshold which true signal must exceed. Mathematical morphology methods can be used to predict the likely shape and

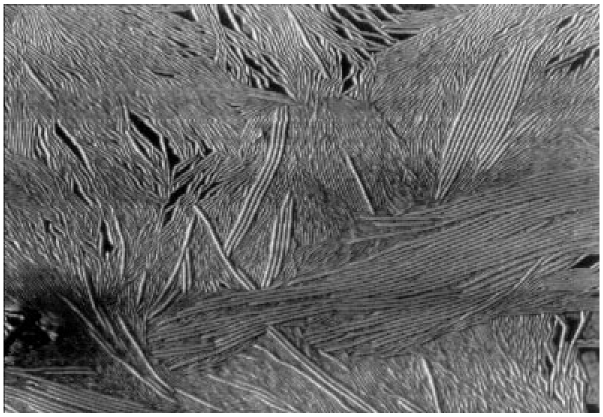

Fig. 3 Atomic force microscopy of DNA on a microarray. This is a micrograph of a portion of a hybridization probe from a yeast microarray, taken after the array was subjected to hybridization. The DNA is clearly deposited at a sufficient density to allow many kinds of strand-to-strand interactions. The width of the picture represents a scanned distance of 2 µm. Image kindly provided by J. DeRisi (Stanford) and E. Carr (Hewlett-Packard).

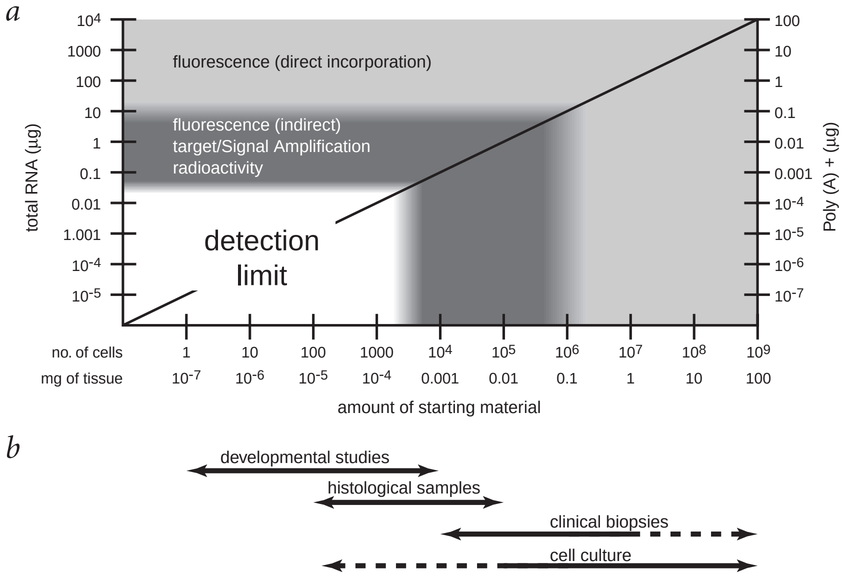

Fig. 4 Detection schemes and applications of cDNA microarrays. Quantitative changes in gene expression can be detected using several schemes for which the limits of detection vary (***a***). Direct incorporation of fluorescent nucleotides into the cDNA target can be used to examine expression profiles from 10 µg or more of total RNA. Indirect fluorescence, as well as target and signal amplification and radioactivity, on the other hand, can be used to detect expression profiles from as little as 50 ng of total RNA. This detection limit allows for the investigation of expression profiles from numerous biological sources including cell culture, clinical biopsies (including autopsy material) and histological samples (***b***). Improvements in technology will permit the detection of expression profiles from less than 50 ng of total RNA, increasing the utility of the technology with respect to studies in development. The limits of the various techniques are constantly changing, and this chart is meant only to illustrate of current performance levels.

placement of the hybridization signal. By applying these methods it is possible to accurately detect even weak signals[27] and extract a mean intensity above background for the target. In contrast, extraction of data from film or phosphor-image representations of radioactive hybridizations presents many difficulties for image analysis. If the array is on a membrane, there is frequently nonlinear warping of the matrix, which means that the observed array will not have the strict geometric regularity of an array printed to a stiff matrix, such as glass. This introduces difficulty in developing highly accurate grids to specify target locations. The spread of detectable particles from a disintegrating nuclide to the detector is highly sensitive to variations in distance between source and detector, and produces a smooth transition from the highest levels of intensity to background. This ensures that the image produced by radioactive exposure is composed of sections at many focal planes, and renders impossible the application of single, simple, point-spread functions to reconstitute a 'focused' representation of the data. The smoothness of the transition from maximum signal intensity to background signal intensity makes consideration of local background for each signal a difficult proposition as one does not observe an abrupt, readily discerned transition between signal and background, but a smooth curve without a sharp derivative.

In carrying out comparisons of expression data using measurements from a single array or multiple arrays, the question of normalizing data arises. All experiments are carried out under conditions of a large excess of immobilized probe relative to labelled target. The kinetics of hybridization are therefore pseudo-first order, and inter-probe competition is not a factor. Under these conditions, the linear differences arising from exact amount of applied target, extent of target labelling, efficiencies of fluor excitation and emission, and detector efficiency can be compounded into a single variable and the information from each detection channel normalized. It is best to achieve normalization by adjusting the sensitivity of detection (photomultiplier voltage with fluorescence or exposure time with radioactivity) so that the measurements occupy the same dynamic range in the detector. There are essentially two strategies that can be followed in carrying out the normalization. One is based on a consideration of all of the genes in the sample, and the other, on a designated subset expected to be unchanging over most circumstances. In either case, variance of the normalizing set can be used to generate estimates of expected variance, leading to predicted confidence intervals. In instances of closely related samples, the transcript level of many genes will remain unchanged, making global normalization a useful tool. As samples become more divergent, the fraction of genes showing altered transcript levels increases, and global normalization yields a poorer estimate of normalization than would be achieved using a subset of constantly expressed genes. Explicit methods have been developed which make use of a subset of genes for normalization, and extract from the variance of this subset statistics for evaluating the significance of observed changes in the complete dataset[27].

An aspect common to all array techniques is the extent of reliability and variance in measurements. So far, most array methods have been validated by probing northern blots of the biological samples. As with sequencing, the best comparisons and measures of reliability can be made only when large data sets containing significant repetitions and overlapping data are freely available. One can, however, clearly envisage strengths and weaknesses. The simple and highly determined nature of immobilized hybridization probes in oligonucleotide arrays make them likely to yield the highest level of reproducibility of absolute measurement for a given element. The ability of cDNA arrays to achieve element-by-element normalization with two-colour fluorescence detection and to use a single, highly specific immobilized probe could provide the most accurate measurements of relative expression levels. All methods should readily disclose large changes in transcript levels among those genes readily detected.

Data management and mining

All array methods require the construction of databases for the management of information on the genes represented on the array, the primary results of hybridization and the construction of algorithms to make it possible to examine the outputs from single and multiple array experiments (ref. 27; see also, page 51 of this issue (ref. 28)). Methods applied to microarray data analysis have essentially been correlation-based approaches that apply methods developed for the analysis of data which are more highly constrained (such a protein or amino acid sequence comparisons) than at the transcript level. This level of analysis on large data sets could provide new perspectives of the operation of genetic networks. Comparison of expression profiles will undoubtedly provide useful insights into the molecular pathogenesis of a variety of diseases (ref. 29; see also, page 48 of this issue (ref. 30)). It will not, however, deliver the kind of intimate understanding of the highly inter-related control circuitry that is necessary to achieve true understanding of genome function. A number of recent publications suggest that to achieve this objective, we should reconsider our perception of transcriptional control as a simple on-off switch to a model whereby control is analogous to a highly gated logic circuit, where numerous, often contradictory, inputs are summed to produce a response[31–33]. To reach these goals, biologists must expand the arsenal of tools they use to analyse expression data—recruiting statisticians and mathematicians to consider multivariant problems of a size never before attempted.

Acknowledgements

A host of talented investigators have contributed to the NIH Microarray Project, including: Y. Jiang, A. Glatfelter, G. Gooden, J. Kahn, M. Boguski, G. Schuler, O. Ermolaeva, E. Dougherty, T. Pohida, P. Smith, S. Leighton, J. Hudson, A. Fornace, S. Amundson, S. Zeichner, C. Xiang, R. Simon, J. DeRisi & P. Brown.

1. Schuler, G.D. *et al.* A gene map of the human genome. *Science* **274**, 540–546 (1996).
2. Henikoff, S. *et al.* Gene families: the taxonomy of protein paralogs and chimeras. *Science* **278**, 609–614 (1997).
3. 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).
4. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* **14**, 457–460 (1996).
5. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
6. Alwine, J.C., Kemp, D.J. & Stark, G.R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. *Proc. Natl Acad. Sci. USA* **74**, 5350–5354 (1977).
7. Berk, A.J. & Sharp, P.A. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell* **12**, 721–732 (1977).
8. Liang, P. & Pardee, A.B. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. *Science* **257**, 967–971 (1992).
9. Adams, M.D. *et al.* Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* **252**, 1651–1656 (1991).
10. Okubo, K. *et al.* Large-scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression. *Nature Genet.* **2**, 173–179 (1992).
11. Velculescu, V.E., Zhang, L., Vogelstein, B. & Kinzler, K.W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
12. Schena, M. *et al.* Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA* **93**, 10614–10619 (1996).
13. Lockhart, D.J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* **14**, 1675–1680 (1996).
14. Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nature Genet.* **21**, 20–24 (1999).
15. Southern, E. Mir, K. & Shchepinov, M. Molecular interactions on microarrays. *Nature Genet*. **21**, 5–9 (1999).
16. Heller, R.A. *et al.* Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl Acad. Sci. USA* **94**, 2150–2155 (1997).
17. 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).
18. de Saizieu, A. *et al.* Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. *Nature Biotechnol.* **16**, 45–48 (1998).
19. Benson, D.A., Boguski, M.S., Lipman, D.J. & Ostell, J. GenBank. *Nucleic Acids Res* **25**, 1–6 (1997).
20. Boguski, M.S., Lowe, T.M. & Tolstoshev, C.M. dbEST—database for “expressed sequence tags”. *Nature Genet.* **4**, 332–333 (1993).
21. Bowtell, D.L. Options available—from start to finish—for obtaining expression data by microarray. *Nature Genet.* **21**, 25–32 (1999).
22. Cheung, V.G. *et al.* Making and reading microarrays. *Nature Genet.* **21**, 15–19 (1999).
23. Phillips, J. & Eberwine, J.H. Antisense RNA amplification: a linear amplification method for analyzing the mRNA population from single living cells. *Methods* **10**, 283–288 (1996).
24. Chen, J.J. *et al.* Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics* **51**, 313–324 (1998).
25. Thorp, H.H. Cutting out the middleman: DNA biosensors based on electrochemical oxidation. *Trends Biotechnol.* **16**, 117–121 (1998).
26. Marshall, A. & Hodgson, J. DNA chips: an array of possibilities. *Nature Biotechnol.* **16**, 27–31 (1998).
27. Chen, Y., Dougherty, E.R. & Bittner, M.L. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J. Biomed. Optics* **2**, 364–374 (1997).
28. Ermolaeva, O. *et al.* Data management and analysis for gene expression arrays. *Nature Genet.* **20**, 19–23 (1998).
29. Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
30. Debouck, C. & Goodfellow, P. DNA microarrays in drug discovery and development. *Nature Genet.* **21**, 48–50 (1999).
31. McAdams, H.H. & Shapiro, L. Circuit simulation of genetic networks. *Science* **269**, 650–656 (1995).
32. Yuh, C.H., Bolouri, H. & Davidson, E.H. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. *Science* **279**, 1896–1902 (1998).
33. Evan, G. & Littlewood, T. A matter of life and cell death. *Science* **281**, 1317–1322 (1998).