Options available—from start to finish—for obtaining expression data by microarray

David D.L. Bowtell

Peter MacCallum Cancer Institute, Locked Bag 1, A'Beckett St. Melbourne 3000, Victoria, Australia. e-mail: d.bowtell@pmci.unimelb.edu.au

The excitement surrounding microarray technology has been tempered by the limited ability of the general biomedical research community to gain access to it. Given that the hardware required for exploitation of the technology is becoming increasingly available, it is an appropriate moment to review options, be they commercially or publically available. Here, we provide a snapshot of the rapidly changing field of microarray-based RNA expression analysis and consider the components and procedures for putting together a complete system.

The use of DNA microarrays for comprehensive RNA expression analysis has caused a great deal of interest recently, although the concept is not new[1,2]. Technical developments that offer increased sensitivity, the prospect that all genes for a given organism could soon be scrutinized in this way and a general appreciation of the need to integrate information obtained from more traditional and reductionist approaches to biology make microarray-based expression analysis a powerful tool[3–13].

The components of a complete system can be divided into three parts (Fig. 1), involving sample preparation (which I designate 'the front end'), array generation and sample analysis (or 'middleware'), and data handling and interpretation (the 'back end'). Component emphasis depends heavily on the questions an investigator wishes to address. For example, there are many 'front end' considerations involving tissue acquisition and processing that are relevant to a researcher interested in prostate cancer but are largely irrelevant to one seeking to dissect signal transduction in established cell lines. Similarly, many 'middleware' issues concerning slide and filter generation will be irrelevant to users who opt for commercially available arrays. Despite this, many users may want to assemble all of the components involved—a difficult but not impossible undertaking.

The front end—from sample to RNA

RNA preparation from cell lines is simple and straightforward. Use of tissues and a need to microdissect adds a layer of complexity, and dealing with human tissues adds many more.

Ethical considerations. Collection of tissue removed at surgery usually requires approval from an institutional ethics committee and informed consent from individuals. In obtaining consent for work with tumour material, germline analysis is obviously more contentious than somatic analysis. This distinction is blurred, however, if RNA expression changes found in tumour material are characteristic of certain mutations. For example, if changes observed in breast cancer samples point to a *BRCA1* mutation, such an analysis could be considered to be a surrogate germline test, given the low incidence of somatic mutations in this gene. Researchers performing RNA microarray analysis with human material should therefore consider duty of care, and processes for follow-up of patients, possibly including genetic counselling. Another critical ethical consideration is whether the collection of tissue impacts on diagnostic procedures. For example, determining whether clear excision margins are obtained in cancer surgery is very important but may be compromised if material is removed for research purposes. Samples obtained from primary, early stage tumours hold great value for gaining an understanding of the molecular progression of disease. Early-stage malignancies, however, may be subject to more extensive pathological scrutiny for staging than end-stage disease in which the diagnosis has been made previously. In addition, early-stage degenerative or malignant disease is often less apparent clinically and less accessible than late-stage or bulky disease, where surgical intervention is much more likely. Issues such as these represent logistical challenges to collecting material of high quality. Finding ways in which samples can be obtained without compromising diagnostic imperatives requires close co-operation with both surgeons and pathologists. A discussion of some of the ethical issues associated with human tissues is available (http://209.143.140.244/napbc/tissue.htm, http://209.143.140.244/napbc/irb%20review.htm, http://bioethics.gov/cgi-bin/bioeth_counter.pl and http://www.health.gov.au/nhmrc/ethics/consult.htm), as well as model consent forms for collecting samples from cancer patients (http://www.pmci.unimelb.edu.au/tissforms/ and http://bioethics.gov/briefings/jan98/model.pdf).

Tissue banks. Although large numbers of archival samples are available in many clinical departments, often the samples are suboptimal with respect to RNA integrity, fixation, or critical patient information. The establishment of suitable tissue banks is a logical adjunct to any in-depth RNA analysis of human tissue; repositories must address issues of appropriate collection and storage and also ensure that the samples are accompanied by appropriate patient information, including treatment, outcome, epidemiological and family history data. The National Cancer Institute (NCI) coordinates a centralized tumour bank for North American researchers (http://www-chtn.ims.nci.nih.gov/). Commercial tissue banks, such as LifeSpan BioSciences, also provide access to a wide variety of human disease tissues (http://www.lsbio.com/).

Microdissection. Diseased tissue contains a mixture of normal tissue, inflammatory cells, necrotic tissue and, in cancer samples, areas of different grade. Similarly, healthy tissue also includes a range of cell types. All of these elements can combine to produce a complex RNA expression profile. Microdissection capability is thus critical for microarray studies involving tissues (see page 38 of this issue (ref. 14)) and is also useful for associated technologies such as comparative genomic hybridization (ref. 15). Current protocols for fluorescent labelling of RNA demand large quantities of RNA, which impedes the use of microdissected RNA on GeneChip® and glass slide arrays. Laser-based microdissection[16–18] offers a means of more rapidly obtaining pure material than conventional techniques. The commercially available laser capture microdissection microscope (http://www.arctur.com) is

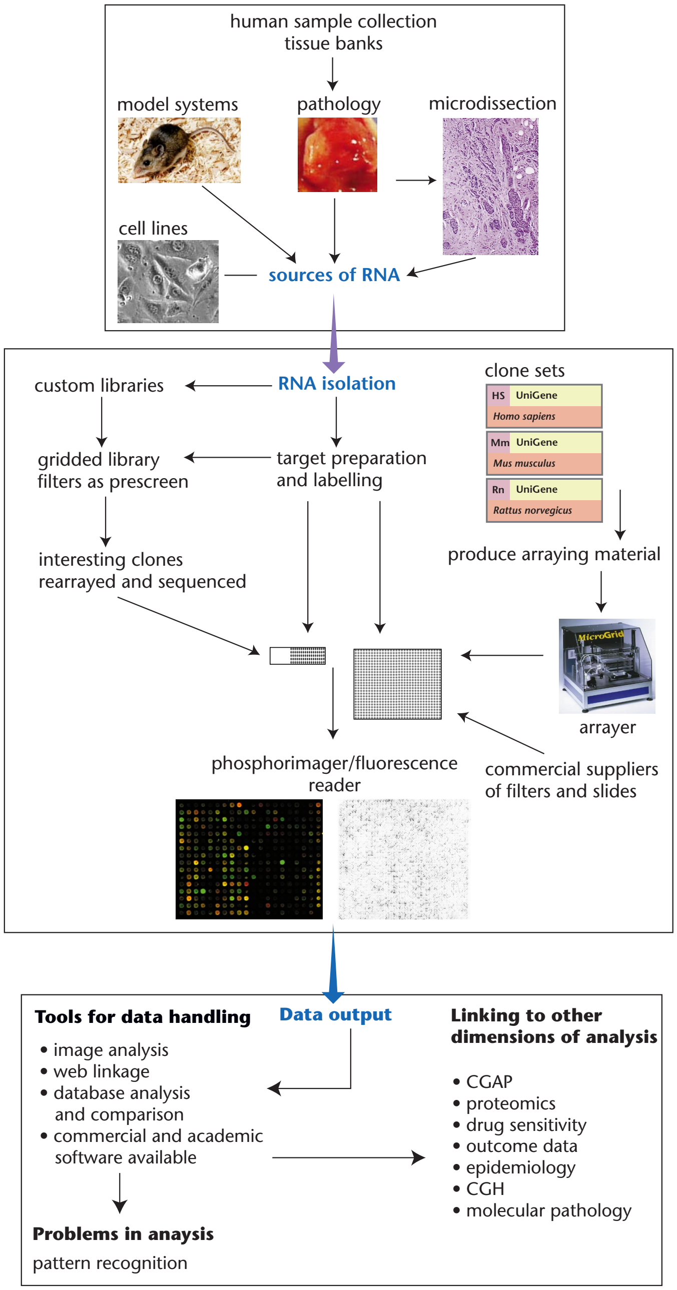

Fig. 1 Flow path schematic of the interacting components required for RNA-expression analysis using DNA-microarrays generated by robotic arraying of PCR products from cDNA clone sets.

thus a valuable adjunct to microarray studies. Strategies for using limited material include PCR-amplification of total cDNA before labelling, or the generation of P^{33}-labelled nucleic acids (or targets) for filters and glass slides, as these require relatively small amounts of total RNA. Xenografts provide an *in vivo* means of amplifying limited amounts of tumour cell material and may reduce levels of contaminating non-neoplastic host tissue, although they may fail to recapitulate the expression pattern of the primary tumour.

Cell lines. The process of obtaining large amounts of RNA from a homogeneous cell population is greatly simplified when using continuous cell lines. It is important to remember that most microarray analyses do not measure absolute levels of RNA but rather compare RNA levels between two samples. Attention to technical details such as density, pH and possible effects of inducers used in conditional systems is critical.

Middleware: making and using microarrays

With RNA in the tube, the researcher can proceed by either buying or making his/her own arrays. Issues of cost, speed, available human resources, flexibility and product range dictate the choice made.

Clone sets. Complexity of the available arrays remains a major issue and relates to the current state of identification of all the genes in a given organism and the clone sets that house these genes. Soon we will have the complete sequence of the human genome and that of many model organisms. For the moment, however, the sequence of most genomes—except those of *Saccaromyces cerevisiae*, *Caenorhabditis elegans* and some unicellular microbes—are incomplete. This necessarily limits what can be put down on arrays.

Traditional approaches to gene discovery such as the cloning of recessive or dominant mutations or the genes encoding specific proteins have identified approximately 7,000 human genes. In contrast, large-scale expressed sequence tag (EST) sequencing has greatly accelerated the rate of gene discovery. Initially promoted by Craig Venter and associates[19], EST projects have spread into both the commercial and academic arenas. In 1991, Merck and Washington University established a collaboration that ultimately fostered the deposition of 480,000 human EST sequences into GenBank. That number has now grown to over one million human ESTs, particularly through the efforts of Washington University and members of the IMAGE consortium[20] (Integrated Analysis of Genomes and their Expression; http://www-bio.llnl.gov/bbrp/image/image.html), and more recently, through those of the Cancer Genome Anatomy Project (CGAP; http://www.ncbi.nlm.nih.gov/UniGene/gene_discovery.html). EST sequences are deposited in dbEST (http://www.ncbi.nlm.nih.gov/dbEST/index.html), a division of GenBank, in which an automated process called UniGene compares ESTs and assembles overlapping sequences into clusters in a similar manner to shotgun sequencing projects (http://www.ncbi.nlm.nih.gov/UniGene/index.html). Some ESTs correspond with known genes, but the majority represent partially sequenced novel genes. Ideally each cluster would correspond with one gene, but as several non-overlapping clusters may exist for large or low abundance genes, the number of clusters is likely to exceed the number of separate genes from whose sequence they are derived. Additionally, errors in alignment programs can produce false clusters (over-clustering). Clone sets, comprising a single representative of each cluster (usually the most 5´ clone), are sold by licensed vendors (http://www-bio.llnl.gov/bbrp/image/idistributors.html; Table 1). The large number of ESTs identified by the Institute for Genomic Research (TIGR, http://www.tigr.org/) are now publically available. TIGR mouse, human, zebrafish, rat and plant clones can be viewed in their respective Gene Indices databases, where they have been assembled into tentative consensus sequences (and can be considered the equivalent of UniGene clusters) by comparison with TIGR and GenBank databases. The American Tissue Culture Collection provides single human and mouse TIGR clones, including a limited number of clones with the complete open reading frame of known genes, but, for the

Table 1 • Clone sets

	Genome Systems	Research Genetics
Currently available	**IMAGE clones regridded but not sequence verified** human known genes 1,939 EST Cluster reps 16,887 singletons 29,080 mouse known genes 1,037 EST Cluster reps 16,479 singletons 2,890 These clone sets were chosen using clustering methods developed by Incyte. A pre-screen against the TIGR human and mouse gene indices identified the genes of known function. The remainder were simply chosen as 5´ most representatives of clusters or as singletons. Singletons probably include some relatively rare unique clones and some cloning artefacts **Incyte/CGAP SV human clones** human known genes 5,524 EST cluster reps 4,320 GS have decided to reformat their sets to correspond more closely with the Unigene sets. The 9,844 human SV clones were obtained from Incyte and are thus free of some commercial inhibitions that apply to the IMAGE clones. This set is used on their current GEMArray V human microarray. A major priority is to include all the known human genes in this set. The ESTs (4,320) represent genes that have been identified within the public domain and in which Incyte does not have a proprietary interest. These clones were derived from a range of libraries, for which there is limited information available. The CGAP program will be an important process for increasing the number of ESTs in the GS SV human set	**Human regridded and sequence verified.** known genes ~4,000 EST cluster reps ~11,000 Currently 15,000, soon to increase to 20,000 IMAGE and CGAP clones. On paper, 4,800 of these genes have non-EST identifiers in UniGene, however, 4,015 should be considered true known genes; for example, entries such as 'matches genomic clone 111A1' are not included. RG uses Greg Schuler's UniGene builds for clustering. A clone is chosen from each cluster based on having a sequenced 3' end and is preferably a NCI-CGAP clone. Clones are robotically arrayed into 96-well plates, streak isolated, plasmid prepped, sequenced and BLAST performed. If the sequence matches its expected sequence, then it goes into the SV set. If the sequence is not what is expected, then a new clone for that cluster is selected and the process is repeated **Mouse regridded and sequence verified** known genes ~2,000 ESTs with similarity to other genes ~6,000 (released late 1998) ***Drosophila melanogaster*** Approximately 55,000 ESTs from the Berkeley *Drosophila* Genome Project, prepared from four different stage libraries, identified by 5´ end sequencing. See http://www.fruitfly/org/EST/ for details of the clones. A UniGene set has been defined and is currently under construction. They are currently being 3´ resequenced to verify and collapse further ***S. cerevisiae*** ORF-specific primers Six thousand primer pairs designed to amplify the complete coding region, including the start and stop codons, from genomic DNA. This is possible because very few yeast genes contain introns *Xenopus laevis*, rat and mouse libraries from Minoru Ko are also available
Cost	regridded, non-SV human and mouse set: $2 ea. SV human 9844 set: $6 ea. Individual SV $45 ea.	Individual sequence-validated clones are $45/clone. Entire set: $6/clone. Yeast ORF primer sets: $20 each for up to 25 pairs, $12 each for 25 or more. 200 μl of a 20 μM solution in TE
T1 contamination	All IMAGE clones are tested for T1 contamination on a sensitive strain. The Incyte/CGAP human revalidated set has no history of T1 contamination	low level T1 contamination present in human clone sets (<10%) and some *Drosophila* libraries. RG recommends that plasmid preps and PCR products are the safest way to work with these clones
Supply format	glycerol stocks in 96- or 384-well plates. Can also supply PCR products, amplified with universal primers, for some sets	96-well microtitre plates for bulk orders. Individual agar stabs for single orders. Also can supply PCR products, amplified with universal primers, for some sets
Information provided	a file with clone IDs, plate and well locations, annotation, *et cetera* For SV clones, supply a text file with the sequences.	clone ID, accession number, gene name, UG cluster ID, gene symbol, chromosome location; more is available at http://www.resgen.com (see the Human GeneFilter section)
Plans	aim to have a SV representative for every human and mouse Unigene cluster and to continually add clones to the set as more data are produced through public domain programs such as IMAGE and CGAP. Incyte/GS does not plan to release Incyte-proprietary clones in the near future the 9,844 SV human clones will be available on nylon filters soon A mouse SV set based on IMAGE clones should be available by the end of 1998. The current mouse ESTs were based on a 5´ sequencing program which has produced a large number of clusters The planned *Arabidopsis* (~6 months) and rat sets will be SV using clustering from TIGR *Drosophila* and soya bean sets planned	SV mouse (see above) *Drosophila* UniGene consisting of 3´/5´ SV clones rat UniGene consisting of 3´/5´ SV clones

SV, sequence-verified

moment, they do not sell complete sets (http://www.atcc.org/hilights/tasc2.html).

Genome Systems (GS; http://www.genomesystems.com/) and Research Genetics (RG; http://www.resgen.com/) are the two IMAGE clone vendors with the most developed clone sets. Both GS and RG have undertaken a process of clone validation through restreaking (to isolate single cells) and resequencing, as the original UniGene sets had a significant discrepancy between actual and designated clone sequence and many IMAGE clones were mixed. In addition to providing individual clones and clone sets, both companies sell filters on which clones or purified DNAs have been arrayed at high density to provide targets for reverse-transcribed probes and supply some clone sets both as bacterial colonies in microtitre plates and as PCR products. The latter have the advantages of avoiding both the risk of T1 phage contamination and the need to isolate plasmids for PCR, a step some labs feel is essential to obtain clean DNA for arraying purposes.

GS has been able to add to the human IMAGE clones via its access to additional human cDNA from Incyte (http://www.incyte.com), an organization that has sequenced more human cDNA (3 million) than any other. Access to the Incyte LifeSeq database is currently limited to approximately 25 pharmaceutical partners (no academic institutions have subscribed). Of the human clones in LifeSeq, 2.3 million are Incyte-proprietary. By using the Incyte clone set, GS has recently produced a sequence-verified set of 9,844 human clones that includes many known genes present in the UniGene set (~5,000 of ~7,400). The GS

Table 2 • Filters

	Genome Systems	Research Genetics	Clontech
Currently available	human and mouse	sequence validated human known and unknown genes *S. cerevisiae* OFR primer set	human, mouse, and rat: ten additional application-targeted arrays including human cancer, neurobiology, apoptosis and haematology/immunology and mouse stress/toxicology
Size and density	22 cm^2, 18,342 clones double spotted	5×7 cm containing 5,184 insert sequences plus housekeeping and genomic DNA controls, the latter for orientation and to monitor evenness of hybridization	80×120 mm. 588 double-spotted genes plus controls
Complexity and source	*human IMAGE clones not restreaked or resequenced* GDA 1.2 known genes 1,939 EST cluster reps 16,887 Singletons 2,908 *human IMAGE clones restreaked but not resequenced (filter 1 only, filter 2&3 non-regridded or resequenced)* GDA 1.3 cluster rep. known singletons filter 1 14,938 1,939 — filter 2 — — 9,215 filter 3 — — 18,394 *mouse IMAGE not restreaked* GDA 1.0 known genes 1,037 EST cluster reps 16,479 singletons 2,890 Clones sets are selected following a pre-screen against the TIGR human and mouse gene indices to identify known genes. Remainder chosen as 5´ most representatives of clusters or as singletons (more detail: http://www genomesystems.com/GDA/index.html)	*human IMAGE/CGAP* releases I–III are a mixture of known and unknown genes, all of which are sequence verified *human known gene array v1.0* release I contains 4,015 SV known genes *yeast* complete set of 6,144 yeast ORFs is spotted onto two filters	known genes only; arrays contain 200–500-bp gene-specific fragments chosen with proprietary software for optimized hybridization characteristics. All fragments are SV and spotted in duplicate to ensure reproducibility and reliability. High sensitivity and low background due to low complexity probe synthesized from gene-specific primer mix; delivers linear dynamic range of three orders of magnitude. Broad-coverage arrays: 588 cDNAs. Application-targeted arrays: 100–300 cDNAs from specific research areas
Purified DNA arrayed or lysed bacterial colonies	lysed colonies	purified DNA	PCR-amplified, SV fragments
Cost of filters	GDA 1.2 $1,395 for 2. GDA 1.3 $999/filter Filter 1 of GDA 1.3 corresponds closely with GDA 1.2	human filters $960/filter yeast ORF $1,325/set	$550–950 for kit with 2 membranes, reagents for making complex cDNA probe, hybridization buffer and complete protocol
Cost of individual clones for follow up	$21.95 each for unverified (current clone set). $45 for verified-clone sets	$45/individual clone $6/clone for the entire set $24/yeast ORF	primer pairs ($85–115) and sequences ($10–20) available for confirmation
Recommended number of times for reuse	four	five	three
Level of T1 contamination (if any) in the clones used for arraying	none in the Incyte human set T1 contamination of the mouse IMAGE derived set	low level T1 contamination present in human clone sets (<10%).	not applicable
Analysis options	distribute software for retrieving the raw data from the phosphorimager files; also have an analysis service with processed and sorted data delivered over the web $199/filter	Pathways software is available for analysing arrays for $2,850; it imports images from the Packard Cyclone or MD Storm phosophorimagers and analyses them. Filters can be stored and compared quickly using different normalization methods (control points, housekeeping genes, or all points on the filter). Images and comparisions can be viewed in false colour (blue/red or red/green). Full reports are generated from a pre-loaded database. No array reading service available	AtlasImage software program available Jan 1999. Isotopic detection method allows straightforward visual analysis or quantitation using phosphorimager. AtlasInfo database provides gene information and links to public databases
Plans	• filters generated with 9,844 SV human available Oct/Nov 1998 • filters generated with PCR products available late 1998 • yeast, *Arabidopsis*, rat, *Drosophila* and other model organisms in planning stages	• a set of filters containing only known genes, released in November 1998 • mouse revalidated set in 1999 • rat and *Drosophila* filters in 1999 • tissue-specific arrays planned (prostate, breast and ovary with non-sequence validated clones)	• 25 different broad-coverage and application-targeted arrays by June 1999 • glass microarrays available spring 1999

GDA, gene discovery array; SV, sequence-verified

Table 3 • High-density microarrays

	Incyte/Synteni: UniGEM V	**Affymetrix** GeneChip®
Currently available	human: 10,000 SV Incyte clones, 5,524 known genes and 4,320 EST cluster reps mouse: 10,000 SV IMAGE clones. 60% of the elements representing known or homologous genes are ESTs *S. cerevisiae*: complete open reading frame set custom arrays (below)	human: Hu6800 set, Hu35K set = 42K genes/ESTs total from UniGene mouse: Mu6500 set, Mu11K set, Mu19K set = 30K genes/ESTs from UniGene, TIGR (some sequences redundant between 6500 set and Mu11K) *S. cerevisiae*: Ye6100 set
Specifications	'1×3' slides. Approximately 10,000 array elements per cm^2. detection limits: genes expressed once per cell (<1 in 100,000) dynamic range: 2–100-fold changes in differential expression	1,700–7,000 sequences analysed per array. 40 specific probes used to analyse each sequence. Sensitivity range 1:100,000. Detect changes greater than twofold between experiments over a 3 log range. Minimal false changes
Availability and cost	do not sell slides, only the data from them. Incyte sells access to a series of standard GEM microarrays based on public domain accessible clone collections, including the UniGEM, which are based on the Unigene collection. Custom microarrays are fabricated in minimum lots of 25 GEM microarrays with 1,000–10,000 cDNAs from customer provided cDNA. Slides are $4,000 per two samples competitively hybridized	$0.29 to $0.50 per human gene or EST transcript profiled; for mouse, $0.26–0.38. Discounts are available off of the list price on the basis of different access programs
Reader compatibility	not applicable	In addition to GeneChip® arrays, Affymetrix sells a fluidics station and reader through the Amersham network; this system is currently available in North America and Europe; available in Japan soon
Cost of individual clones for follow up	public domain clone stabs are available at $24; public domain resequenced clones are available at $66 each; private clones available only in a resequenced state at $350 each	not available—must obtain cDNA from other sources
Usage	single use	single use
Availability of software for analysing the data and cost	GEMTools and LifeArray for collection, analysis and management of data sets. GEMTools is available in a Windows NT-SQL Server operating environment from $1,895 (single use) to $300,000 (client server) and process data from Incyte databases and MD platforms. LifeArray is available in a subscription format for data management of	GeneChip® software suite contains software that automates control of instrumentation, captures hybridization data and converts it into relative expression level information GeneChip® EDMT, priced from $2,500–$3,250 per seat depending on geography, facilitates the mining of expression data contained in the GATC™ database created by GeneChip LIMS GeneChip® LIMS is a client/server software that facilitates the tracking of experimental and expression information in a relational database architecture. The server software is priced from $70,000–$98,000 and access licenses are priced from $2,500–$3,250 per seat depending on geography
T1 contamination of source clones	none in human clones T1 contamination of the IMAGE clones but tested before shipment	not applicable
Plans	*Pseudomonas aureginosa*, *Mycobacterium tuberculosis*, *E. coli*, Rat UniGEM, second human UniGEM and second mouse UniGEM	rat, *E. coli* and other prokaryotes, *C. elegans*, *Arabidopsis*, *Drosophila*

SV, sequence-verified

resequenced human clones are also free from low-level T1 contamination present in IMAGE clones, which can be a serious problem (http://www-bio.llnl.gov/bbrp/image/phage.html). Both RG and GS clone sets also contain a large number of ESTs. Whether those ESTs present are likely to be expressed in your favourite tissue or cell line is hard to predict. Little or no information is provided concerning the basis for selection of ESTs; they appear to represent clones from a range of libraries with no preselection on the basis of biological interest. That situation is changing as the focus of the Washington University human EST project shifts to CGAP clones and companies provide clones that have interesting expression patterns.

EST sequences of other organisms, such as mouse, rat, *Drosophila melanogaster* and *Arabidopsis Thaliana*, have accumulated at different rates. A limitation of the mouse EST project (http://genome.wustl.edu/est/mouse_esthmpg.html) is that sequencing has been carried out from the 5′ end of cDNA. As the length of the 5′ ends of cDNA is variable, the number of clusters obtained is greater than if oligodT-primed cDNA had been sequenced from the 3′ end. As a result of this, and because fewer mouse cDNAs have been sequenced and clusters are smaller, the proportion of novel genes in the current mouse clone sets is substantially fewer than in the human sets. Although no sequence-validated sets are currently available for the mouse, both RG and GS are performing 3′ resequencing of a collection of mouse clones. These clones have been selected because they either correspond to known mouse genes or because they appear to be related to other genes of interest, thus effectively collapsing some of the earlier clustering. The sequence-validated, reclustered mouse set should be available at the end of 1998.

Celera (http://www.celera.com/), a commercial offshoot of TIGR, is sequencing the *Drosophila* genome as a prelude to their human genome sequencing project; the *Drosophila* sequence is expected in 1999. Obtaining the entire genomic sequence of *S. cerevisiae* allowed a near-complete set of genes to be generated by PCR, which have been arrayed and analysed[7]. Although it will be more difficult to identify coding sequences in more complex organisms, the convergence of genome and EST sequencing projects over the next couple of years will ensure the identification of non-redundant clone sets that encompass all genes for a variety of other species. For the time being, it is important to understand the limitations of the clone sets available.

Commercially available filter arrays, GeneChip® and slide arrays. Filter arrays have the advantage of being relatively affordable and needing no special equipment to use, although potential users should be aware that large format phosphorimager screens may be required with larger (22 cm^2) filters. Filters are also useful for scarce RNA (for example from microdissected tissue), as only approximately 50 ng of total RNA is required for a single experiment (100 µg of RNA is typically required for a fluorescent probe). The major disadvantage of filters is that comparison of expression between two samples

Table 4 • Arrayers

	Genetic Microsystems GMS 417 arrayer	**Cartesian Technologies** PixSys NQ series	**Beecher Instruments**	**Genomic Solutions** Flexys	**BioRobotics** MicroGrid and Total Array system
Specifications:					
• capacity	42 slides	50 slides	48 or 96 slides	14 trays of 24 slides	108 slides: 4–24 filters
• speed	1 spot/s/pin	0.5 s per print,75 s per transfer	8–16 spots/48 slides/1 min	print 24 slides in 3–4h	23 spots/s
• dimensions	80 cmL×53 cmD×49 cmH	90 cmD×63 cmW×41 cmH	1.5 m×0.8 m×1.9 m high	1.425 m×0.900 m×0.780 m	56×54×60 cm
• pen design	solid pen	inkjet dispenser	quill, carbide	solid pen, titanium	solid pen and quill available
• pen number	4 (8–32 in development)	1,4,8 channels	1,2,4,8 or 16	32,48,64 (384 plate)	4,12,16 (96) 16,48,64 (384)
• print density	200 μm pitch	400 μm pitch	250 μm pitch	variable pitch in 100 μm increments	200 μm–1 mm pitch
Features[a]	PS (3: 96, 384), TC[b]	PS (100×96 or 384)[b], TC[b], HC[b]	PS (20:96), HC, FA	PS[b] (15: 96, 384), LR[b], TC[b], FA[b]. Other: replicates, picks and re-arrays bacterial libraries and grids/arrays on membranes	PS (24: 48, 96, 384), LR, BC, TC[b], FA[b]. Other: replicates, picks and re-arrays bacterial libraries and grids/arrays on membranes
Technical support	US & ROW: GMS toll free & info@geneticmicro.com UK & Ireland. Japan, China, Taiwan, Korea	factory-trained field support; distributors in UK, Japan, China, Taiwan, Korea, Australia, India	available from Beecher	Europe: Genomic Solutions Ltd. Japan: Genomic Solutions KK. US & ROW: Genomic Solutions. help@genomicsolutions.com	Europe: *Bio*Robotics Ltd. US: *Bio*Robotics, Boston and West coast. Factory trained distributors in Japan, Australia, New Zealand, Taiwan, Korea
Cost	US$45,000	US$65,000	contact sales@beecherinstruments.com	US$80,000–$190,000 depending on options	£45,000–£70,000 depending on options

[a]PS (Plate stacker, number: format of plates); LR, lid remover; TC, temperature control; HC, humidity control; FA, filtered air; BC, bar code reader; ROW, rest of world. [b]Optional extra.

requires hybridization of each sample to separate duplicate filters, or to a single filter that must be stripped and hybridized sequentially. The sensitivity of lysed colony filter arrays is reported to be limited to high- and medium-abundance genes[21]. In contrast, hybridization of fluorescently labelled nucleic acids to slide arrays or gene chips can detect low abundance genes[10]—an important point as most genes fall within this category[22–24]. Direct comparison of GeneChip®, slide and filter arrays is required, however, to settle the considerable debate concerning the relative sensitivity of filters hybridized with P^{33}-labelled targets versus GeneChip® or slides hybridized with fluorescent targets.

Commercial filter arrays of clone sets are available from Clontech, GS and RG. (Table 2; arrayed libraries are also available from these and other companies). For those considering an investment, a detailed comparison of cost, species range and complexity is warranted. Important considerations also include whether the clones used to produce the arrays are restreaked and sequence-verified, whether DNA or lysed colonies are arrayed, and the number of known genes and ESTs. Clontech filters only include known genes, pre-selected and grouped for their involvement in specific processes such as apoptosis. Current GS filters use lysed bacterial colonies, whereas purified DNA is arrayed on both Clontech and RG filters. The lower complexity and higher purity of arrayed DNA is thought to increase the sensitivity of these filters. GS will soon release sequence-verified human arrays that include a large proportion of known genes present in the UniGene set (approximately 5,000). When considering the cost of commercial filters, it is important to bear in mind the additional cost of obtaining individual clones for follow-up experiments, especially as the expression of a substantial number of the genes may vary in a typical experiment.

GeneChip® arrays (see page 25 of this issue (ref. 25)) and commercially available glass slides and compete at the more sophisticated end of microarray analysis (Table 3). At the moment, both of these are relatively expensive and fairly limited (this may change pending legal challenges; http://www.affymetrix.com/press/pr980901.html). At present, options include Affymetrix GeneChip® arrays (http://www.affymetrix.com) and slide arrays from Incyte (which has recently acquired Synteni (http://www.synteni.com/)). Incyte does not sell slide arrays as such but provides a service whereby samples applied to slides and the data returned. Molecular Dynamics and Clontech have recently announced that they will provide slide arrays early next year (http://www.mdyn.com/). Genometrix (www.genometrix.com) provides custom synthesis of large numbers of low complexity arrays (up to several hundred probes). Using a proprietary method for arraying oligonucleotides, they mass-produce slides at low cost (approximately $10/array for orders of 1,000–10,000 individual arrays) and are developing devices for high-throughput analysis of these arrays. Hyseq (http://www.hyseq.com/) have developed a novel method where hybridization of DNA targets with all possible pentamer or heptamer oligonucleotides allows inference of sequence from the pattern of oligonucleotide hybridization[26–28]. This strategy has been applied to measuring the abundance of individual cDNA in libraries from tissues of interest, thereby providing an estimate of individual gene expression. Hyseq do not sell chips or filters but offer this type of analysis in house.

Options for producing slide and filter arrays. Few researchers are interested in producing consumables for their experiments. The cost and range of GeneChip® and glass slide arrays, however, means that many will favour producing slides and filters for use in academic settings in the short- to medium-term at least (see pages 10 (ref. 29) and 15 (ref. 30)). The first glass slide arrays were produced in Pat Brown's laboratory at Stanford (http://cmgm.stanford.edu/pbrown/index.html) and from there the technology spread to those of a few others, including that of Jeff Trent (National Human Genome Research Institute (NHGRI, http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/), Vivian Cheung (University of Pennsylvania; http://w95vcl.neuro.chop.edu/vcheung) and Geoff Childs (Albert Einstein College of Medicine; http://sequence.aecom.yu.edu/bioinf/funcgenomic.html). Each has made important refinements to microarray technology and been remarkably helpful in disseminating it, including making available detailed protocols. Brown's web site also contains detailed specifications for building an arrayer and associated software. (Additional protocols and some hardware details are available at http://chroma.mbt.washington.edu/mod_www/).

Recently, several companies have produced arrayers (Table 4). Each is a relatively simple XYZ axis robot that can position the print head with a similar degree of precision. Critical determinants when choosing among them are whether the machine has a proven track record in the field, the technical support network available,

Table 5 • Scanners

	Genomic Solutions *GeneTAC^{TM} 1000*	Packard *BioChip Imager*	General Scanning *ScanArray3000*	Beecher Instruments	Molecular Dynamics *Avalanche*	Genetic Microsystems *GMS 418 Array Scanner*
Basic design	1.36 million pixel CCD camera with high energy Xenon light source. Filter wheel allows for up to 4 fluors per slide. Image area 28×36 mm	Epi-fluorescence confocal scanning laser system. 543 nm (CY3) and 633 nm (CY5) HeNe laser	Scanning confocal laser GHeNe 543nm and RHeNe 632nm	Scanning laser confocal, 2 simultaneous PMT channels. 3 lasers: 488 nm @ 75 mw 532 nm @ 100 mw 633 nm @ 35 mw	Confocal optics, 9 element lens. HeNe and NdYag lasers. Four emission filters	Scanning laser digital imaging epifluorescence microscope. 532 nm (35 mW) and 635 nm (35 mW)
Resolution	27 μm	50 μm, 20 μm, or 10 μm	10 μm	10–100 μ pixels	10 μm or 20 μm	10 μm
Speed	automatic processing up to 24 slides per run. 20 s–2 min per fluor	not available	4 min 20×20mm	80 mm/s max scan speed	5 min single colour/slide 11 min two colour	~3 min/entire slide
Sensitivity	not available	<34.5 molecules/μm^2 CY3 dye, less than 11.5 molecules/μm^2 of CY5 dye. Dynamic range, four orders of magnitude	<0.5 molecule fluor/μm^2 or 0.15 amol end-labelled nucleotide. 16 bit dynamic range		<10 Cy3 molecules/μm^2 @ SNR = 3. 16-bit TIFF gives 5 orders of magnitude dynamic range	dynamic range linearity over 6 orders of magnitude. 16 bit TIFF file output
Software	Visage HDG	OptiQuant	ImaGene (BioDiscovery), Optional autoload		ImageQuant 5.0 and ArrayVision for microarray analysis	
Track record	now available	release early 1999	released late 1997	released late 1997	fourth generation MTAP, public release Sept 1998	available February, 1999
Tech support	Europe: Genomic Solutions Ltd. Japan: Genomic Solutions KK. US & ROW: Genomic Solutions. Worldwide: help@genomicsolutions.com	worldwide	worldwide	available from Beecher Instruments	worldwide	US & ROW: GMS toll free & info@geneticmicro.com. UK & Ireland. Japan, China, Taiwan, Korea: Takara
Cost	US$60,000	TBA	US$40,000–$80,000	contact: sales@beecherinstruments.com	US$110,000	US$50,000

ROW, rest of world

cost, ease of use, capacity, features such as bar code reader, temperature control, plate stacker, microtitre plate lid remover and pen design. Beecher Instruments was started by one of the engineers who developed the robot used by the NHGRI; it now sells an equivalent arrayer and reader, which has the advantage of having been successfully 'field tested' for more than two years. The BioRobotics microGrid (http://www.biorobotics.co.uk/) combines many features into a compact machine. Initially designed for robotic gridding of clones from 96- or 384-well microtitre plates onto filters, it can also replica-plate libraries and be upgraded to print glass slides and filters and re-array bacterial libraries (also known as cherry picking). Genomic Solutions (www.genomicsolutions.com) produces a complete system of arrayer, hybridization station, reader and analysis software. Genetic MicroSystems (http://www.geneticmicro.com/) also produce a relatively affordable machine. Molecular Dynamics conduct a Microarray Technology Access Program (MTAP), where participants gain (for a substantial fee) early access to microarray technology developed by Molecular Dynamics and Amersham. Although Molecular Dynamics produce arrayers for MTAP participants, it does not plan to make its arrayer available to those outside of the program in the near future.

One of the most important factors affecting the performance of the arrayer are the shape, reproduciblity and durability of the pens (also referred to as tips, pins and quills). Uneven pens deliver unequally during a print run and tax the abilities of image analysis programs. Precision tips are available from several suppliers, including Beecher Instruments, Majer Precision Engineering (http://www.majerprecision.com), who custom-machine high-precision pens from a range of materials and Telechem International (http://www.wenet.net/~telechem/), who also offer related microarray equipment and consumables.

Readers. Filters are hybridized with P^{33}-labelled probes and signal is detected using phosphorimager screens. Phosphorimager systems are produced by Molecular Dynamics, Packard Instrument company and Fuji. The Packard Cyclone instrument is relatively low in cost but offers a high degree of resolution for array work. Analyses can be carried out by eye, by sending a GIFF data file via the web to a company to be read (for a charge), or using commercial or public domain software (see below).

The Affymetrix fluorescence reader, produced by Hewlett Packard, is currently customized for GeneChip® arrays. Hewlett Packard plan to build readers capable of reading both GeneChip® and glass slides; estimates suggest that it will be available at the end of 1999. General Scanning (http://www.genscan.com/) released the ScanArray 3000, a compact scanning confocal laser, late in 1997 (Table 5), and Beecher Instruments sell a reader based on the machines used at NHGRI and the National Cancer Institute (NCI). Like the arrayer, the Beecher reader is not supported by a service network but its high degree of sensitivity has provided a benchmark for other commercial readers. Molecular Dynamics have recently released the Avalanche reader, which is based on one developed during the MTAP program. Genetic Microsystems and Packard Instruments are developing readers due for release early in 1999. The above readers are laser confocal scanning devices, except for the Genomic Solutions reader that uses a CCD camera and filter blocks, facilitating upgrade to reading different fluorophors. Although there are obviously a range of readers available, it is very difficult to assess their relative performance at this early stage; direct comparison would be very useful.

Back end: moving and handling data

Informatics. A typical array experiment generates thousands of data points and creates serious challenges for storing and processing data. Informatics can be categorized as either 'tools' or 'analysers'. Tools include software that operate arraying devices and perform image analysis of data from readers, databases to hold and link information, and software that link data from individual clones to web databases. Some involve fairly straightforward software but are nevertheless quite extensive. The Brown laboratory has made available software for operating custom built arrayers (http://cmgm.stanford.edu/pbrown/mguide/software.html).

The quality of image analysis programs is crucial for accurate interpretation of signals for slide and filters. Yidong Chen (NHGRI) has developed a sophisticated image analysis program for slides and filters, deArray, that is available but not supported (www.nhgri.nih.gov/DIR/LCG/15K/HTML/). Mark Boguski and colleagues have developed software that is capable of both analysing microarray data and linking to databases such as Entrez and UniGene (ref. 31; see also, page 51 of this issue (ref. 32)), and this can be downloaded from the web (www.nhgri.nih.gov/DIR/LCG/15K/HTML/). Commercial readers and arrayers provide software for data analysis (Tables 2,3): Synteni have developed a sophisticated program for analysing microarray data (GemTools); RG sells the Pathways package to analyse their filters; and the Visage suite can be purchased from Genomic Solutions, separate from their hardware. Silicon Genetics (http://www.sigenetics.com) provides the GeneSpring package for analysing data from Affymetrix GeneChip® and other microarray experiments.

Multidimensional analysis. RNA-expression analysis represents only one parameter by which cells or tissues may be characterized. Depending on the experiment, epidemiological or molecular pathological data, genomic changes (gains or losses) or sensitivity to drugs[33] may be additional parameters that will influence the interpretation of microarray data. The ability to combine RNA and protein expression data to comprehensively profile both transcriptional and post-transcriptional changes in cells and tissues is particularly appealing, although the number of proteins that can be profiled at this stage is substantially less than the number of genes. Although it is more difficult to identify proteins that are differentially expressed, techniques for rapid and reproducible two-dimensional gel protein separation and mass spectrometry-based protein identification make high-throughput proteomics a highly desirable adjunct to microarray RNA expression analysis[34,35].

The means for carrying out RNA expression analysis are rapidly increasing; as the technology infiltrates the larger research community, we shall be better to gauge its impact on our understanding of biological systems.

Acknowledgements

I thank R. Dickins and E. Sloan for many useful comments during the preparation of the manuscript.

1. Gress, T.M. *et al.* Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm. Genome* **3**, 609–619 (1992).
2. Friemert, C., Erfle, V. & Strauss, G. Preparation of radiolabeled cDNA probes with high specific activity for rapid screening of gene expression. *Methods Mol. Cell Biol.* **1**, 143–153 (1989).
3. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray [see comments]. *Science* **270**, 467–470 (1995).
4. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645 (1996).
5. Schena, M. *et al*. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA* **93**, 10614–10619 (1996).
6. Derisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* **14**, 457–460 (1996).
7. 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).
8. Lashkari, D.A. *et al*. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl Acad. Sci. USA* **94**, 13057–13062 (1997).
9. Welford, S.M. *et al*. Detection of differentially expressed genes in primary tumor tissues using representational differences analysis coupled to microarray hybridization. *Nucleic Acids Res.* **26**, 3059–3065 (1998).
10. 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).
11. 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).
12. Lockhart, D.J. *et al*. Expression monitoring by hybridization to high-density oligonucleotide arrays [see comments]. *Nature Biotechnol.* **14**, 1675–1680 (1996).
13. Cho, R.J. *et al*. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.* **2**, 65–73 (1998).
14. Cole, K.A., Krizman, D.B. & Emmert-Buck, M.R. The genetics of cancer—a 3D model. *Nature Genet.* **21**, 38–41 (1999).
15. Kallioniemi, A. *et al*. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**, 818–821 (1992).
16. Schutze, K. & Lahr, G. Identification of expressed genes by laser-mediated manipulation of single cells [In Process Citation]. *Nature Biotechnol.* **16**, 737–742 (1998).
17. Emmert-Buck, M.R. *et al*. Laser capture microdissection [see comments]. *Science* **274**, 998–1001 (1996).
18. Simone, N.L., Bonner, R.F., Gillespie, J.W., Emmert-Buck, M.R. & Liotta, L.A. Laser-capture microdissection: opening the microscopic frontier to molecular analysis. *Trends Genet.* **14**, 272–276 (1998).
19. Adams, M. *et al*. Complementary DNA sequencing: expressed sequence tags and Human Genome Project. *Science* **252**, 1651–1656 (1991).
20. Lennon, G., Auffray, C., Polymeropoulos, M. & Soares, M.B. The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* **33**, 151–152 (1996).
21. Pietu, G. *et al*. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. *Genome Res.* **6**, 492–503 (1996).
22. Bishop, J.O., Morton, J.G., Rosbash, M. & Richardson, M. Three abundance classes in HeLa cell messenger RNA. *Nature* **250**, 199–204 (1974).
23. Hastie, N. & Bishop, J. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **9**, 761–774 (1976).
24. Zhang, L. *et al*. Gene expression profiles in normal and cancer cells. *Science* **276**, 1268–1272 (1997).
25. Lipshutz, R.J., Fodor, S.P.A., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nature Genet.* **21**, 20–24 (1999).
26. Drmanac, S. *et al*. Accurate sequencing by hybridization for DNA diagnostics and individual genomics. *Nature Biotechnol.* **16**, 54–58 (1998).
27. Strezoska, Z. *et al*. DNA sequencing by hybridization: 100 bases read by a non-gel-based method. *Proc. Natl Acad. Sci. USA* **88**, 10089–10093 (1991).
28. Drmanac, R. *et al*. DNA sequence determination by hybridization: a strategy for efficient large-scale sequencing. *Science* **260**, 1649–1652 (1993) [published erratum appears in *Science* **163**, 596 (1994)].
29. Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P. & Trent, J. Expression profiling using cDNA microarrays. *Nature Genet.* **21**, 10–14 (1999).
30. Cheung, V.G. *et al.* Making and reading microarrays. *Nature Genet.* **21**, 15–19 (1999).
31. Ermolaeva, O. *et al*. Data management and analysis for gene expression arrays. *Nature Genet.* **20**, 19–23 (1998).
32. Bassett, D.E. Jr, Eisen, M.B. & Boguski, M.S. Gene expression informatics—it's all in your mine. *Nature Genet.* **21**, 51–55 (1999).
33. Weinstein, J.N. *et al*. An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
34. Kahn, P. From genome to proteome: looking at a cell's proteins [news]. *Science* **270**, 369–370 (1995).
35. Shevchenko, A. *et al*. Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. *Proc. Natl Acad. Sci. USA* **93**, 14440–14445 (1996).