Options available—from start to finish—for obtaining data from DNA microarrays II

Andrew J. Holloway, Ryan K. van Laar, Richard W. Tothill & David D. L. Bowtell

doi:10.1038/ng1030

Microarray technology has undergone a rapid evolution. With widespread interest in large-scale genomic research, an abundance of equipment and reagents have now become available and affordable to a large cross section of the scientific community. As protocols become more refined, careful investigators are able to obtain good quality microarray data quickly. In most recent times, however, perhaps one of the biggest obstacles researchers face is not the manufacture and use of microarrays at the bench, but storage and analysis of the array data. This review discusses the most recent equipment, reagents and protocols available to the researcher, as well as describing data analysis and storage options available from the evolving field of microarray informatics.

We reviewed the options for the manufacture and use of DNA microarrays in the original *Chipping Forecast*[1] in 1999. Since then, the number of companies that produce microarrays equipment and reagents has increased, and so has the choice of protocols available to researchers. Perhaps the biggest change, however, has been the explosion in options for data storage and analysis, partly because effective array laboratories are now producing vast quantities of data. Here we discuss the significant advances in the field of microarray research made in the three years since our original review.

Making or accessing DNA microarrays

The basic concept behind all microarrays is the precise positioning of DNA fragments (probes) at high density on a solid support so that they can act as molecular detectors. In practice, microarrays vary according to the solid support used (such as glass or filters), the surface modifications with various substrates, the type of DNA fragments on the array (such as cDNA, oligonucleotides or genomic fragments), whether the gene fragments are presynthesized and deposited or synthesized *in situ*, and the machinery used to place the fragments on the array (such as ink-jet printing, spotting, mask or micromirror-based *in situ* synthesis).

Currently, combinations of these variables are used to generate three main types of microarray: filter arrays, spotted glass slide arrays, and *in situ* synthesized oligonucleotide arrays. Both filter and spotted arrays are produced readily in academic facilities; they can also be purchased from commercial vendors for those not inclined toward ‘do it yourself’. By contrast, arrays of oligonucleotides that are synthesized *in situ*, such as the Affymetrix GeneChip, require complex equipment and are only produced in commercial settings[2–4]. Below we first discuss issues that affect the manufacture of filter and spotted arrays and then consider the developments made in producing *in situ* synthesized oligonucleotide arrays.

Probes for filter and spotted arrays. The first step in the production of spotted DNA microarrays or filters is the generation of ‘array-ready’ material, which serves as the feedstock for printing. In gene expression microarrays, either synthetic oligonucleotides or cDNA fragments are used as probes. For most researchers, the ideal microarray for expression profiling would be a complex array of sequence-validated probes, in which each sequence is unique, shows minimal cross-hybridization to related sequences and provides, collectively, a comprehensive representation of the expressed fraction of the genome including splice variants. It would also be richly annotated in terms of the functions of the genes that correspond to the probe sequences. In a similar way, a nonredundant set of fragments that provide a comprehensive representation of a genome would be ideal for carrying out comparative genomic hybridization[5–8].

So far, the principal source of probe fragments used for arraying have been bacterial cDNA and bacterial artificial chromosome (BAC) clone sets, although sets of long oligonucleotides are increasingly providing a viable alternative. Considerable progress has been made in the past few years in improving the complexity and reliability of the cDNA and BAC clone sets. For complex organisms such as mice and humans, however, there are still some shortcomings in the libraries available. Frequently these libraries contain a certain amount of redundancy, misannotation and contamination.

Sets of cDNA clones, comprising a single representative of each cluster, are distributed by licensed vendors to researchers (see the IMAGE Consortium: http://image.llnl.gov/image/html/idistributors.shtml) as bacterial cultures in a multiwell plate format. The most comprehensive sets are currently distributed by the ResGen Invitrogen Corporation (http://www.resgen.com/) and the Resource Center of the German Human Genome Project (RZPD: http://www.rzpd.de/; see Web Table A online). Specifically for microarray analysis, Lion Bioscience (http://www.lionbioscience.com/) has developed sets of mouse, rat and dog cDNA clones that have been selected for size, 3′ bias and the removal of poly(A) tails, and which therefore show limited cross-hybridization (Web Table A online). Incyte Genomics (http://www. incyte.com), a main provider of validated sets in the past, stopped making their clone sets available in May 2001

The Ian Potter Foundation Centre for Cancer Genomics and Predictive Medicine and The Trescowthick Research Laboratories, Peter MacCallum Cancer Institute, Locked Bag 1, A’Beckett Street, Melbourne 8006, Victoria, Australia. Correspondence should be addressed to D.D.L.B. (e-mail: d.bowtell@pmci.unimelb.edu.au).

but supplies array-ready material from these sets for spotting (http://www.incyte.com/expression/easy_to_spot/catalog.jsp?page=index).

In addition to the IMAGE Consortium, other main contributors of cDNA clones sets include RIKEN (http://genome.rtc.riken.go.jp/home.html), The Institute for Genomic Research (http://www.tigr.org/) and other individuals (see, for example, ATCC Bioproducts: http://www.atcc.org/SearchCatalogs/tasc2.cfm#who). In many of these cases, individual clones or sets of clones are available publicly, although access to some clones involves the licensing or sharing of any intellectual property generated from them. For species for which no libraries are available, some researchers have had success with arrays made from random collections of clones[9–11].

For some applications of microarrays, such as the sensitive detection of genomic losses in tumor cells[8], large insert clones such as BACs are more appropriate than cDNAs. Some of the companies that supply cDNA clone sets also supply curated BAC libraries (ResGen: http://www.resgen.com/; Incyte Genomics: http://www.incyte.com/). For species for which a pre-existing BAC library is not available, some companies (for example, Genomex: http://www.genomex.com/) will prepare custom libraries. Many of the BAC clone sets, such as the RPCI-11 series of human BAC clones[12], were generated as part of genome sequencing efforts. Because most of these libraries contain up to a tenfold redundancy across genomes, they require individual clones to be selected to reduce clonal overlap for microarraying purposes.

An academic supplier of BAC libraries curated specifically for microarraying purposes is BACPAC Resources at the Children's Hospital Oakland Research Institute (http://www.chori.org/bacpac/). In addition to 53 BAC, 7 phage artificial chromosome (PAC) and some fosmid libraries, representing numerous vertebrate and invertebrate genomes, BACPAC Resources also offers a human BAC library containing 3,500 clones with a spacing of 1 clone per 1,000,000 base pairs. These clones have been verified for use as fluorescent *in situ* hybridization probes. A more complex human BAC library containing 30,000 clones with minimal spacing, and a similar mouse library are currently under construction. These new libraries are expected to become available by late 2002.

Many of the initial cDNA clone sets were compromised by contamination with T1 phage, by multiple clones in individual wells and by incorrect sequence assignment[13], and substantial efforts have been taken to reduce the error rate in clone sets. Between 1 and 5% of clones in well-maintained clone sets are said to be misassigned—that is, they do not contain the specified sequence—although this figure has been disputed[14].

Operationally, the clone error rate depends on the degree of care taken both by the facilities that produce clone sets and by the individual laboratories that use them. DNA for arraying is typically prepared from clone sets by high-throughput polymerase chain reaction (PCR), rather than by the purification of recombinant constructs such as plasmids[15]. Once the early stages of microarray production are achieved, investigators often seek to increase the complexity of their arrays, for example, stepping up from arrays of 5,000 clones to 20,000–40,000 clones. Many have found, however, that the ability to maintain high-quality, error-free clone preparation and printing is not easily scalable. In short, it is one thing for a person to spend a month doing PCRs and running validation gels with the promise of carrying out long-awaited microarray experiments; it is another to repeat this process using the same equipment on 5–10 times as many clones and to remain focused.

Given the logistical difficulties associated with handling large numbers of bacterial clones, it has become very attractive to obtain large sets of oligonucleotide probes that obviate much of the work involved in producing array-ready material and take advantage of the rapid growth in genome sequence information. The production of complex, spotted oligonucleotide microarrays has become progressively more accessible as the cost of oligonucleotide synthesis has fallen and the yield of full-length long oligonucleotides or 'longmers' has improved. Because oligonucleotide sets, unlike cDNA sets, are not limited by the availability of physical clones, in principle, sets could be generated in-house from sequence information. But the design of oligonucleotides is a complicated procedure[16,17], and in practice, most investigators use commercially available oligonucleotide sets that encompass large numbers of genes ascertained from the latest draft of the relevant genomic sequence (Web Table B online).

A forerunner of this approach was the generation of short oligonucleotide primers to amplify fragments corresponding to expressed regions of the genome. The use of gene-specific PCR to generate array material was first applied to yeast[18], and other genomes have been subsequently amplified in this way, including bacterial genomes[19] and plants[20]. Although this is an appealingly systematic approach, it is not without its challenges in terms of ensuring effective primer design for efficient high-throughput PCR amplification. A more direct approach is to use oligonucleotides of 50–70 bases that correspond to known or predicted genes and to print these directly onto spotted arrays (Web Table B online). Because the oligonucleotides are much shorter than cDNAs, the base composition is likely to influence their performance strongly, and an effective oligonucleotide design is required. This is usually accomplished by using available cDNA sequence information and sequence prediction programs such as ArrayOligoSelector (http://sourceforge.net/projects/arrayoligosel).

There is still a lack of good comparative data for cDNA arrays versus long oligonucleotide arrays. One of the best analyses of the utility of oligonucleotide arrays has been provided by Hughes *et al.*[4], who used an ink-jet printer to synthesize large sets of human and yeast oligonucleotides. Hughes *et al.*[4] also evaluated several key features of oligonucleotide design for sensitivity and specificity. Their systematic study showed that oligonucleotides of 60 bases can provide excellent results in terms of specificity and sensitivity. We have found that, depending on the origin of the oligonucleotide set, there can be surprising lack of concordance between results obtained with cDNA arrays and those obtained with oligonucleotide arrays. We therefore urge that care should be taken when changing from cDNA to oligonucleotide probes and suggest that investigators use test samples to compare sets from different providers.

Printing substrates for spotted glass arrays. Spotted arrays are typically printed on glass to allow visualization of the bound, fluorescently labeled targets. Glass slides have continued to be the favored solid support for immobilizing probes for reasons of availability, low fluorescence, transparency, resistance to high temperature, physical rigidity and the variety of surface chemical modifications possible. The nonporous nature of glass means that targets have direct access to probes without the limitations of internal diffusion. The use of a nonporous substrate reduces the background problems that typically arise with the high concentrations of targets and the agents designed to deplete repeat sequences.

Initial studies used standard glass microscope slides that had been thoroughly washed and then coated with poly-L-lysine[21]. Poly-L-lysine is still popular as a substrate owing to its low cost, its ease of manufacture and its generally good results. Slides for printing have undergone considerable commercial development, and alternative substrates on slides with highly uniform surface

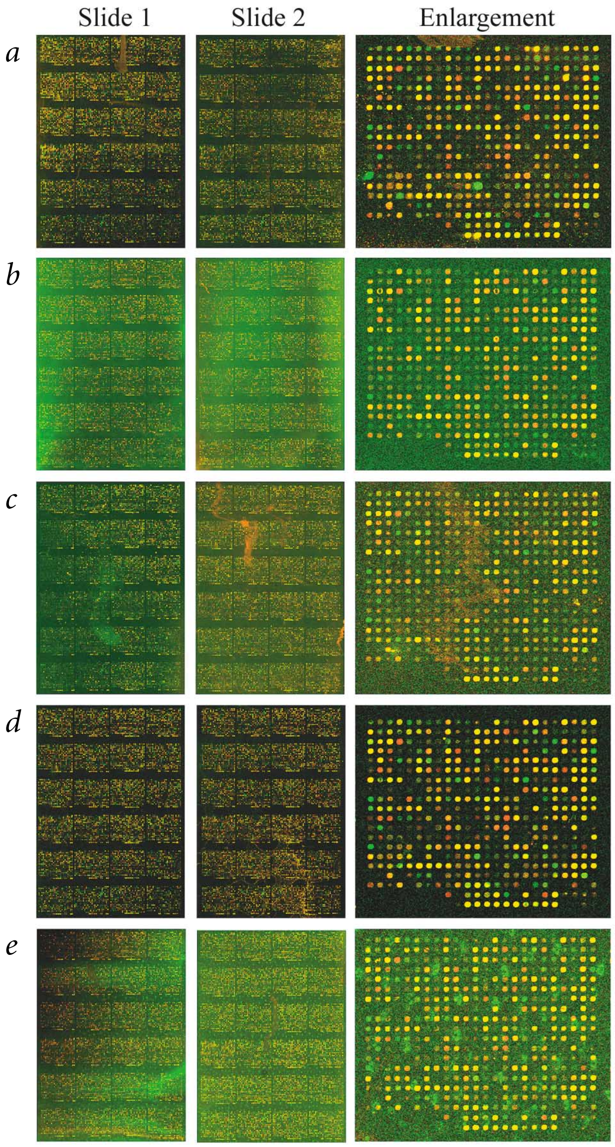

Fig. 1 Comparison of commercial slides for printing of cDNA material. Five types of slide (***a*–*e***) were used for printing 10,500 cDNA arrays in a single print run. cDNA was printed at roughly 100 µg/ml in 150 mM sodium phosphate buffer (pH 8). After printing, each slide was blocked according to the respective manufacturer's protocol. Duplicates of each slide type (slide 1 and slide 2) were hybridized with an aliquot of a large pool of labeled cDNA (amino terminally labeled total RNA from Jurkat (Cy3-labeled) and MCF7 (Cy5-labeled) cells) and processed using the same stringency washes. Each slide was scanned using identical settings on a Packard Bioscience Scanarray 5000. The images show that there is considerable variation in spot morphology (see the enlargements) and background among the types of slide, with specific effects consistent between duplicate slides.

properties have been produced (Web Table C online). There is probably no 'best' substrate at present, because substrate choice depends on the type of material that is printed (for example, cDNAs or oligonucleotides) and its purity, and particularly on the protocols used to subsequently label and hybridize targets to the array. Clearly, the type of substrate can markedly affect the signal intensity, degree of background and durability of the slide. The striking variation in signal and background that can occur when different substrates are used for spotted cDNA arrays is shown in Fig. 1.

Investigators are advised to test a range of substrates systematically to find those that best meet their needs and budgets, because slides can constitute a substantial fraction of the cost associated with construction of microarrays. It is also worth checking that batches of slides have uniform geometry, as a minor variation in edge length can accumulate over a platen of slides with disastrous consequences for printing. Similarly, there is no standard slide dimension across the industry at present, and some microarray readers may not be able to focus on slides whose width or thickness falls outside a predetermined range.

Arrayers for printing glass slide and filter arrays. Glass slide microarrays were first produced in Patrick Brown's laboratory at Stanford University[22]. The microarrays were produced by an *xyz*-axis gantry robot that used banks of pins to ferry small volumes of DNA solutions from the wells of 96-well plates to the prepared surfaces of a series of glass slides. Initially, the availability of commercially produced microarray robots was very limited, and to build their own arrayers, the early pioneers of the technology made use of the detailed specifications provided by the Stanford group (http://cmgm.stanford.edu/pbrown/mguide/index.html).

An increasing number of commercial robots can position a print head precisely over a field of glass slides, and many investigators have opted for the purchase of these devices as a fast, relatively painless (but not inexpensive) way of entering into the arraying field (Web Table D online). The development of pin-based gantry microarray robots over the past few years has been evolutionary rather than revolutionary. The machines work on the same principles as the earlier arrayers, but they are more automated and more commonly have features such as climate control and plate stackers.

The development of high-precision printing pins that can deliver smaller, more uniform spots over many slides probably has been more significant than the development of the arrayers themselves (Web Table E online). A practical consequence of more efficient probe delivery by newer pin designs is the need for arrayers with slide platens that can accommodate more than 100–200 slides, so that printable material is not wasted at the end of the print cycle.

Although cDNAs can be renewed for use at moderate expense, the use of expensive and non-renewable oligonucleotide sets has made the issue of waste reduction even more important. The recent development of noncontact ink-jet and piezo printing machines may potentially reduce the wastage of print material while offering increased precision and speed. Many companies now manufacture noncontact arrayers, which use a variety of configurations (Web Table D online). Although these provide a solution to wastage and have other benefits over standard contact printing, they are an expensive alternative, and individual laboratories need to weigh the benefit of the investment.

Perhaps the most interesting recent development in laboratory-based array manufacture is the promise of bench-top machines that allow the *in situ* synthesis of oligonucleotide arrays. At least two companies, febit and NimbleGen (Web Table F online), are developing devices that use maskless, micromirror technology to accomplish base addition during the *in situ* synthesis of oligonucleotides[2]. These machines could potentially revolutionize the use of genome information by investigators, by allowing the flexible and rapid design of new microarray devices.

Filter arrays are sometimes called macroarrays because of their generally lower probe density and to distinguish them from their more glamorous cousins. Glass microarrays and filter macroarrays are to some extent seen as alternatives, but in reality both formats have their strengths and weaknesses, and they probably should be seen as complementary rather than competing technologies. Filter arrays can be produced rapidly using minimally purified PCR material, require tiny amounts of RNA for radioactive target labeling, use widely available phosphoimager instrumentation to read, and are relatively cheap to produce and use[15]. Their chief disadvantage is the need to carry out sequential hybridizations of targets to the same filter or parallel hybridizations to duplicate filters to compare gene expression between samples. This key difference distinguishes fluorescence-based microarrays and filter arrays, and probably contributes to the greater ability of fluorescent microarrays to detect differences in expression, especially those of low-abundance genes[23].

The printing of filter arrays using quill pins is constrained by the fragility of the membrane and the tendency of quill pins to wick unacceptably onto the porous surface of the filter. The Affymetrix 417 and 427 pin and ring printers (Web Table D online) are used widely for producing filter arrays in academic facilities because this method combines a solid pin with the speed of a traveling reservoir (ring) of printing material.

Obtaining microarrays from commercial or core facilities. A key consideration in using microarray technology is whether to adopt a do-it-yourself approach and spot arrays in the laboratory or to purchase arrays from a commercial supplier. Given the complexity of manufacturing microarrays, many investigators have chosen to enter the field by obtaining microarrays for their experiments from someone else. Initially, the options for doing this were very limited, because commercially produced arrays, such as Affymetrix GeneChip, cost several thousands of dollars. In addition, very few academic facilities were in a position to part with their hard won batches of glass slide arrays.

The situation regarding purchase of arrays has changed markedly in the past few years as the price of commercial arrays has tumbled. Affymetrix GeneChip arrays have increased in complexity and in the number of species represented, and the unit cost per probe has decreased several-fold; thus, GeneChip arrays are now within the reach of academic users. Affymetrix has also introduced new arrays for single-nucleotide polymorphism analysis. Agilent Technologies (http://www.agilent.com/) has entered the field with spotted oligonucleotide and Incyte cDNA collection arrays, and oligonucleotide arrays synthesized *in situ* by ink-jet. A comprehensive summary of the commercial arrays is available on the web (http://ihome.cuhk.edu.hk/~b400559/array.html).

In parallel with increasingly affordable commercial options, academic faculties have pooled their resources to create core facilities to produce microarrays for several laboratories or institutes and to guide investigators through standardized protocols. The development of core microarray facilities makes sense, given the cost of the hardware and the complexity of the process. Anyone who has participated in the development of a microarray core facility will know, however, that such a development is not without difficulties as impatient investigators wait for the new microarray facility to deliver.

We have used a structure of two independent core facilities located in research institutes to establish and develop the technology, which is then ported to a service-orientated genomics facility to make slides on a large scale for a wide community of users (http://www.vicmicroarray.org/). The developer nodes also provide investigator training in the use of microarrays. Development and service roles are equally important, but not necessarily compatible; this is particularly true in the early days of a core facility, when many new techniques are to be established. Although it is probably not necessary to separate these roles physically, it is essential to have enough staff to cater for both activities. An excellent survey of the configuration of core facilities, including the average number of staff, is available on the web (http://abrf.org/ResearchGroups/Microarray/EPosters/MARG_Survey_2000_Poster.pdf).

For large-scale users, it is currently cheaper to make your own arrays than to buy them, and there are additional advantages in terms of flexibility in array design. But it is likely that at some time in the next few years the cost of making arrays will equal the cost of purchasing them from a commercial vendor. Perhaps the only way that this trend may be challenged is if affordable devices become available that allow laboratory-based *in situ* synthesis of oligonucleotide arrays, which will reduce further the price of in-house prepared arrays (Web Table F online).

A consideration common to all array users, regardless of whether the arrays are produced in an academic core facility or commercially, is the ability to gauge the quality of the arrays and the data produced. Numerous useful strategies for carrying out such evaluation have been proposed[24], including protocols for determining a measure of spot quality[25].

Using microarrays

The dominant application of microarrays has been in measuring gene expression in different situations, including analysis of diseased versus normal tissues[26], profiling tumors and predicting outcomes[27–35], studying gene regulation during development[36], and following the stimulation of cells *in vitro*[37]. Other array applications include comparative genomic hybridization[5], chromatin immunoprecipitation[38,39], mutation detection[40], genotyping[41,42] and microarray-mediated localized cell transfection[43].

Expression analysis. Expression analysis using glass slide microarrays is typically done by the competitive hybridization of two targets (typically known as test and reference), each labeled with a specific fluorescent dye such as Cy3 or Cy5 (ref. 21). Because levels of gene expression are relative, the nature of the RNA pairs is an important consideration when designing experiments for spotted arrays. Researchers may carry out individual pairwise comparisons or compare each sample against all others[44]. As the number of samples increases, the latter option rapidly becomes impractical—in terms of both the number of pairwise combinations needed and the amount of RNA required for each sample.

The comparative nature of gene expression measurements with spotted arrays creates many problems in terms of archiving data and comparing data among different experiments and laboratories. In principle, some of these problems could be alleviated by adoption of universal references (such as for mouse, human and *Arabidopsis*) by the microarray community.

Reference RNAs and oligonucleotides. Although there is no current consensus on references, Brown's research group at Stanford University has described a pool of RNAs derived from 11 diverse human tumor cell lines (Web Table G online) that has become a kind of *de facto* universal human reference RNA. Because the reference pool is derived from immortalized cell lines, it is possible to generate more reference material, although inevitably there is some batch variation. To avoid creating 'islands' of data, a large quantity of reference RNA must be made at the outset. Where several batches are required, growth conditions of the cells should be controlled tightly to reduce batch-to-batch variation.

Some researchers have elected not to generate their own reference RNA, but to purchase similar, pooled RNA from several commercial suppliers, including Stratagene (http://www.stratagene.com/displayProduct.asp?productId=439) and Clontech (http://www.clontech.com/archive/APR02UPD/ControlRNA.shtml). The degree of batch-to-batch variation in commercially supplied reference is not clear (although the manufacturers state that it is low). Usually only the type of cells is indicated, and not the exact cell lines, which can make it difficult both to decide whether a particular pool is appropriate and to interpret results.

Recently there have been developments in using a reference that is not derived from mRNA. A labeled oligonucleotide that is complementary to every feature on an array has been shown to be an effective reference, without the complications associated with references derived from mRNA[45]. In contrast to glass slide arrays, each labeled target is hybridized to a separate Affymetrix GeneChip array. This avoids some of the problems associated with relative measurements of gene expression that are fundamental to two-color competitive hybridization. Some users of Affymetrix arrays incorporate a reference-like pool of spiked RNAs and have developed algorithms to facilitate both comparisons across groups of arrays, in a manner akin to spotted array reference comparisons, and the determination of absolute concentrations of cellular mRNA species[46]. Ultimately, the development of microarrays or other processes to allow high-throughput, parallel measures of absolute RNA abundance are needed to provide a robust description of the transcriptome of specific cellular lineages, developmental stages and disease states.

Protocols and hardware. In early microarray technologies, the processes involved were often more of an art than a science: few array reagents were commercially available or, if they were, they often lacked reliability. The microarray field has now reached a stage where there is some consensus about the best approaches for producing reliable array results, and some of these ideas are being formulated as comprehensive manuals (http://ihome.cuhk.edu.hk/~b400559/book_mray.html). Accompanying this has been the inevitable near saturation of the market with reagents, often sold as kits, that are designed to purify and label RNA, and to hybridize labeled probes.

RNA labeling protocols. Expression analysis labeling protocols are based on the reverse transcription of mRNA, either from highly purified poly(A) mRNA or total RNA extracts. Extensive purification of RNA is essential to remove all contaminating protein, polysaccharide and other organic material, especially RNases. Many protocols have been developed for the extraction of high-quality RNA using various in-house and commercial kits and reagents (Web Table H online). Initial protocols for target labeling were based on direct labeling, whereby reverse transcription of mRNA is primed using a poly(dT) primer in the presence of fluorescently labeled nucleotides (typically Cy3- or Cy5-conjugated dCTP or dUTP). Cy3- or Cy5-conjugated nucleotides are bulky, however, which makes their incorporation using standard enzymes very inefficient. In addition, rates of incorporation can differ between dyes, potentially resulting in dye biases[47]. In an attempt to alleviate this problem, reverse transcriptases are becoming available that may allow a more efficient incorporation of fluorescently labeled nucleotides, for example Fluoroscript reverse transcriptase (Invitrogen).

An alternative method to direct labeling, called indirect or amino allyl labeling, circumvents the need to incorporate bulky fluorescent dyes during reverse transcription (http://www.microarrays.org/pdfs/amino-allyl-protocol.pdf). In this method, an amino allyl modified dUTP is used instead of a prelabeled nucleotide. After reverse transcription, the free amine group on the amino allyl dUTP can be coupled to a reactive *N*-hydroxysuccinimydl ester fluorescent dye. Although this technique is longer than direct labeling, its benefits—including better sensitivity, absence of dye biases and decreased cost—seem to be worth the extra effort. There are now several *N*-hydroxysuccinimidyl dyes available, including the standard Cy3 and Cy5 dyes (Amersham: http://www1.amershambiosciences.com/; Molecular Probes: http://www.probes.com/; see Web Table I online). As the reactivity of the dyes can vary markedly depending on the product, the batch and the supplier, we recommend that a range of dyes be tested.

RNA amplification and detection. Direct and indirect labeling requires a substantial amount of total RNA, typically between 20 and 75 µg. RNA amplification and high-sensitivity techniques have been developed to overcome this obvious limitation (Web Table H online), and many of these methods are based on 'Eberwine' amplification[48]. There are numerous commercially available kits for carrying out this kind of amplification; alternatively, protocols and individual reagents are easily accessible to the researcher (Web Table I online). Tyramide signal amplification[49] and 3DNA dendrimer technology (Genisphere: http://www.genisphere.com/) are directed toward increasing signal strength without using an RNA amplification step.

Hybridizing arrays. Until recently, the options available for hybridizing spotted arrays were limited. Many researchers found that the conventional methods (under coverslips and in chambers[15]) gave variable results. Probe distribution was often problematic, resulting in variations in gene expression dependent on the spatial position[50]. But many commercial instruments now have the potential to allow automated, highly reproducible hybridization (Web Table J online).

The instruments available vary from ones based on simple approaches, such as a vibrating temperature-controlled platform (Thermo Hybaid: http://www.thermohybaid.com/), to ones based on complex systems of probe application and mixing (see, for example, Ventana Discovery: http://www.ventanadiscovery.com). Although the advantages and disadvantages of the systems vary, the volume of probe required for hybridization is a consideration that is applicable to all. Frequently, the instruments require a substantial dilution of the probe with a consequent loss of signal. Other factors include the reliability of the sealing of the hybridization chamber, and the possibility that the sealing mechanisms may not be compatible with all array layouts. The instruments currently available can be regarded as first generation, and significant advances in design are likely.

Scanning arrays. The binding of the target to the probe is detected by scanning the array, typically using either a scanning confocal laser or a charge coupled device (CCD) camera–based reader (Web Table K online). Like arrayers, scanners have gradually improved in sensitivity, reliability and their available features, such as autofeeders. Although in principle the latter are attractive to the high-end user, batch scanning may not be possible if adjustments of scanner settings are required for each slide. The ability of the scanner to be upgraded is potentially important as new dyes are developed.

With large numbers of experiments it is prudent to scan all arrays using, at the very least, the same model of scanner, if not the same unit. We have observed that results from two scanners reading the same slide can be subtly different, and this can have consequences in data analysis. At an extreme, samples could be potentially clustered on the basis of the scanner used. The Minimum Information About a Microarray Experiment (MIAME) protocols[51] (see below) include provision for recording the scanner used in an experiment.

Informatics

Along with the rapid development of microarray technologies, there has been an unprecedented amassing of data. Storage and analysis of these data can be a headache for microarray researchers. Although at present there is no clear standard solution for microarray data storage and analysis software, there are many open-source, public domain and commercial solutions vying for a share of this evolving market. Most of the available products are still in the early phases of the software development process; consequently, new and improved versions of these are being released frequently to keep up with consumer expectations and to fix programing 'bugs'. An exhaustive record of microarray software can be found on Y.F. Leung's website (http://ihome.cuhk.edu.hk/~b400559/array.html).

For some laboratories, combinations of public-domain or noncommerical software (see Web Table L online for examples) are capable of fulfilling data storage and analysis needs. A downside of this approach is the limited support or training available for noncommercial software applications. Laboratories producing large amounts of data may find that they require the support and the usually greater programming stability that comes with commercial solutions.

Several products have been released that integrate data acquisition, pre-processing and analysis. For example, the commercial GeneTraffic (http://www.iobion.com/), the academic TM4 (http://www.tigr.org/software) and the open-source BASE (http://base.thep.lu.se/) created at Lund University[52] aim to provide all the tools needed for data storage, quality-control metrics, normalization and statistical analysis in a web-based application. Comprehensive solutions are also likely to offer image analysis and data extraction in the near future. Some of the available software for the various steps in microarray analysis, and considerations for their use, are discussed below.

Data extraction software. Most commercial microarray scanners are supplied with data extraction software, such as QuantArray (http://www.packardbioscience.com/products/521.asp) and GenePix[53], that is designed to accommodate the usually unique parameters of the scanned images generated. Research is continuing to define more precise and automated approaches to spot detection and, particularly, the vexed issue of background measurement.

Improvement in printing processes has simplified grid measurements and the detection of spot boundaries. But there is no consensus on the best approaches to background subtraction. Options include fixed (a user- or software-specified value), local (the intensity of regions immediately surrounding individual spots measured) and global (the intensity of the area outside the array grid measured) background measurement. Further variation exists in each of these techniques on the exact formula used to produce the final background value for each feature.

The storage of primary scanned data (usually in the form of tiff images) is potentially important if investigators want to take advantage of future developments in image extraction and/or analysis software[54].

Storage of microarray data. When a microarray laboratory begins to produce reliable large-scale microarray data sets, two pressing questions arise: what data should be stored, and how should those data be stored to facilitate efficient analysis? When considering what to store, it is important to ensure that investigators can retrace their steps or reanalyze their data with new analysis tools. In addition, sufficient information is needed to interpret the quality of data submitted for publication and to allow others to repeat published studies.

The MIAME[51] set of protocols, developed by the Microarray Gene Expression Database Group (MGED: http://www.mged.org/), is currently the leading proposal for data submission and database standards at present. MIAME 1.0 was approved at the MGED 3 meeting at Stanford University in May 2001 (ref. 55). The document seeks to capture information regarding (i) experimental design; (ii) array design; (iii) the extraction, preparation and labeling of samples used for hybridization; (iv) hybridization conditions; (v) measurements such as images, quantification and specifications; and (vi) normalization controls (http://www.mged.org/Annotations-wg/index.html)[51]. Although not yet formalized, complying with MIAME (or something similar) has become essential for publishing array findings in several journals[56,57].

As reviewed recently[54] and summarized in Web Table M online, there are many options for storing the output from a microarray experiment. Storing the raw image files retains maximum information, allowing the use of different normalization, image extraction and quality metrics to be used subsequently. A useful list of available databases and their adherence to the MIAME structure has been prepared and is available online (http://www.wehi.edu.au/bioweb/Suzanne/databases.html). Gardiner-Garden and Littlejohn[58] have weighed up the pros and cons of the leading products for data storage, although—reflecting the rapid evolution of the field—some of the applications described in their review are no longer available.

Normalization of microarray data. Owing to the complicated process of producing and hybridizing spotted microarrays, it is not uncommon for a certain amount of systematic variation to exist in the data produced. Normalization is a routine, but important, step in the analysis of almost all microarray data[47,59]. The selection of a normalization algorithm needs to be made with a view to the type and degree of systematic bias present and is an important step between obtaining raw data and analyzing the biological issue under investigation (see also review by J. Quackenbush, pages 496–501, this issue)[60]. Caution must be taken in applying global transformations to a data set to avoid to overenhancing (or diminishing) the information contained in the arrays and to ensure that any results obtained represent biological, not systematic, variation.

Factors such as differing intensities of dye incorporation, minor irregularities in probe distribution during hybridization, topographical slide variation or scanner introduced bias, are examples of factors that can cause considerable systematic variation in a microarray experiment[47]. Most of these factors can be corrected for by using one or more statistical procedures, some of which are built into specialized microarray databases and analysis packages (Web Table L online). But even when a range of normalization options is available in a user-friendly format, it can still take some experience to select the most appropriate method confidently.

At present, there are few normalization algorithms that address the issue of irregular spatial distribution across the surface of an array. When a global variation in mean expression, or a bias towards one particular channel, has been corrected for, there is often evidence of a nonrandom distribution of the differentially expressed genes present. Options for addressing this anomaly include pin-group normalization with SMA (http://stat-www.berkeley.edu/users/terry/zarray/Software/smacode.html) and the more sophisticated SNOMAD method[50], which attempts to address both intensity and spatial bias. Figure 2 illustrates the ability of SNOMAD to detect variations in local mean signal intensities for a given two-color microarray.

Pin-group normalization, implemented in the open-source R statistical language[61], uses information provided about the grid and subgrid layout of the array to carry out a LOWESS-based transformation of the data. The SNOMAD technique (also R-based) identifies and corrects specific regions of an array where artifacts show a systematic spatial pattern. The LOWESS method is again used in this technique to calculate the local mean signal intensity across the surface of an array.

More development and improvement of normalization algorithms are required to produce accurate and reliable means of detecting and correcting for systematic variation. The MGED Normalization Working Group website (http://www.dnachip.org/mged/normalization.html) contains extensive information on some of the available options for normalizing microarray data.

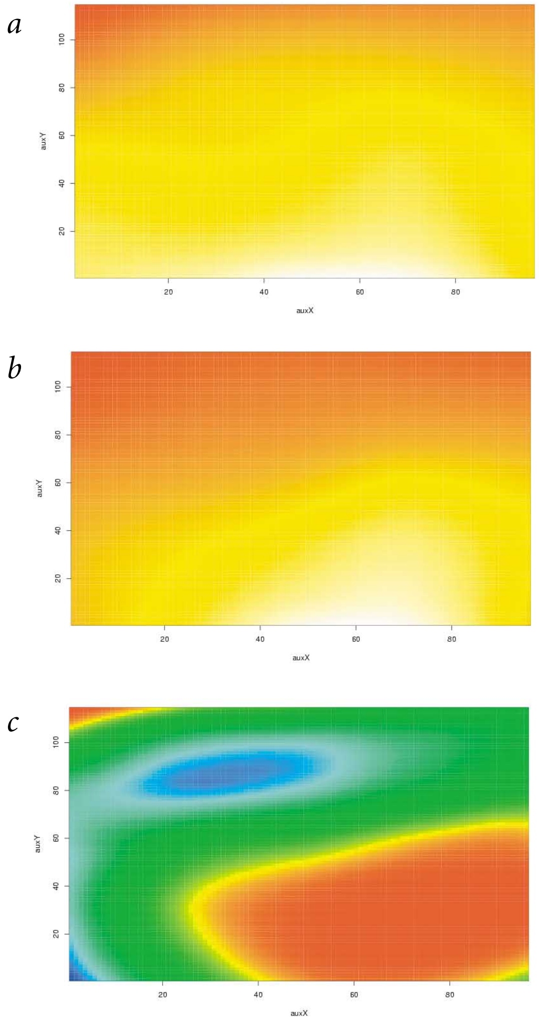

Fig. 2 Output from the local mean normalization step of SNOMAD normalization. ***a***,***b***, Locally calculated mean element intensity for each channel of a particular two-color microarray slide. There is clear variation in the distribution of the mean intensity. Values are colored from white to yellow to red in order of increasing intensity. ***c***, Differences between the locally calculated mean spot intensities. Blue or red regions of the image reflect sections of the array that may be suffering from a hybridization (or other spatial) artifact.

Analysis of microarray data. Initial approaches to analyzing microarray data focused heavily on the use of unsupervised hierarchical clustering techniques (see also the review by D. Slonim, pages 502–508, this issue)[62]. The twin programs Cluster, which organizes related gene expression data, and Tree View, which allows clustered microarray data to be visualized easily[63], are commonly used for this approach (Web Table L online). Although not suited to all analyses, hierarchical clustering techniques are a simple, powerful method of organizing such data. If an experiment is designed to be an exploratory endeavor, rather than to answer a particular question, then clustering is generally an excellent choice when beginning data analysis. By contrast, where there are data to guide the initial analysis, then computational methods that 'train' an algorithm to recognize patterns in data—the so-called 'supervised learning programs'[64–67]—are proving to be more effective than unsupervised approaches.

Several public domain and commercial solutions are now available for scientists who want to carry out a statistical analysis of microarray data but are not familiar with specialized statistical software (Web Table L online). Although these programs are not a substitute for understanding the statistical issues relevant to one's experimental design, they do reduce manual data manipulation and present a user-friendly interface. Using programs such as Silicon Genetics' GeneSpring (http://www.silicongenetics.com/), the Whitehead Institute's GeneCluster (http://www.genome.wi.mit.edu/cancer/software/software.html) or Biodiscovery's GeneSight (http://www.biodiscovery.com/genesight.asp), one can view, normalize and extract a biologically annotated list of differentially expressed genes at a specified confidence level on an array with only a few clicks of the mouse.

Many of the supervised and unsupervised data analysis techniques frequently published can now be done in graphical user interface (GUI)-style programs. User-friendly packages such as GeneSpring and GeneCluster (free to academic users; see Web Table L online) make it possible for nonprogramers to carry out a range of techniques including normalization, hierarchical clustering, *k*-means clustering, principal component analysis, self-organizing maps, profile similarity searches, gene filtering and simple machine-learning analysis. Quality journals are now publishing papers in which the total analysis of data has been carried out using one or more commercially available software products (for example, Silicon Genetics maintains a list of publications citing use of their products: http://www.silicongenetics.com/cgi/SiG.cgi/Products/GeneSpring/citations.smf), supporting the idea that these products are gaining the necessary sophistication needed to analyze the large data sets associated with microarray experiments.

For researchers who wish to explore other analysis protocols, microarray data sets can also be analyzed using methods that have been applied to data from other disciplines for many years. Statistical approaches that have been used to investigate data from epidemiological, environmental or social studies, for example, frequently can be adapted to microarray studies. Consequently, industry-standard statistical packages such as Matlab (http://www.mathworks.com), R (http://www.r-project.org), S-Plus (http://www.insightful.com/) or Minitab (http://www.minitab.com/) can be used in place of specialized microarray software (Web Table L online).

The recent adaptation of advanced machine-learning techniques, such as neural networks, support vector machines and decision trees[64,67,68], has demonstrated the potential that exists for powerful microarray analysis by methods that have been used traditionally in disparate fields such as finance, computing and engineering. Many machine-learning tools and algorithms are available from industry, such as Microsoft's Bayesian Network Tool (http://research.microsoft.com/adapt/MSBNx/), and academic sources, such as the University of Waikato's WEKA system (http://www.cs.waikato.ac.nz/ml/weka/)[69] and GeneCluster. At present, only a few of the specialized microarray analysis packages offer any form of machine-learning tool. One of these is GeneSpring (http://www.silicongenetics.com/), which contains a simple yet effective *k*-nearest neighbor 'class prediction' tool.

A considerable quantity of machine-learning resources can be found online (see, for example, http://www.ai.univie.ac.at/oefai/ml/ml-resources.html and http://directory.google.com/Top/Computers/Artificial_Intelligence/Machine_Learning/). As more sophisticated learning algorithms are shown to produce unparalleled pattern or class discovery and prediction with gene expression data, one can expect these methods to trickle down into the dedicated, biologist-friendly, microarray analysis software. In the meantime, a more advanced knowledge of data manipulation and statistical testing may be required to use software and algorithms not designed specifically for the microarray field.

Gene annotation. Efficiently obtaining appropriate and readily understandable information about genes in microarray experiments is a crucial part of meaningful interpretation of the experiment. In a perfect world, every gene would have a unique identifier, and complete molecular information would be available, including an annotated sequence that described promoter elements, intron–exon structure, splice variants and related genes. Biological information associated with that gene would be available readily through links to other databases, ranging from PubMed to protein structure databases. Although there are active steps towards achieving this molecular nirvana, we are still a long way from it.

An essential part of synthesizing information from diverse sources is the use of structured, controlled vocabularies. Such vocabularies will allow automated searches to filter information and draw meaningful inferences about possible associations between the data obtained in a microarray experiment and information that already exists. An attempt to address some of these issues has been made by the Gene Ontology Consortium[70,71] (see also review by C. Stoeckert, pages 469–473, this issue)[51]. The goal of the Gene Ontology project is to produce a comprehensive controlled set of terms that can be used to describe genes in all organisms. The use of medical subject headings (MeSH) for describing scientific literature through MEDLINE is a familiar example of this concept[72,73]. The Gene Ontology project began with the development of shared vocabularies for the model organism databases—FlyBase, Mouse Genome Informatics Database and *Saccharomyces* Genome Database—and uses terms that describe molecular function, cellular location and biological processes. Some software such as GeneSpring allows the user to sort genes contained in both commercial and custom arrays into categories proposed by the Gene Ontology Consortium.

It is particularly important that cDNA clones are richly annotated, but there is a great challenge in reconciling information from several sources. Most laboratories identify cDNA clones by the GenBank accession number or the IMAGE clone ID, but this practice is not standardized. There are several files available at the National Center for Biotechnology Information that show how IDs from different projects relate to each other (see, for example, the LocusLink ftp site: ftp://ncbi.nlm.nih.gov/refseq/LocusLink/). It is important to work with annotation systems that provide the most current and up-to-date information possible. Some sources of data that are kept reasonably current are listed in Web Table N online. The SOURCE database created at Stanford University conveniently combines the data from a range of other databases and allows batch annotation of clone sets.

Conclusions

Although not a mature technology, microarray devices have come a very long way in a short period of time and are now an established industry in their own right. The future would seem to lie in our ability to use full genome information for comprehensive array manufacture and to reduce the aspects of the technology that are labor-intensive and that introduce systematic variations in the data. The development of methods that allow absolute, rather than relative measures of gene expression is a principal goal if durable descriptions of gene expression patterns, analogous to DNA sequence database information, are to be achieved. As this technology develops and the number of users expands, one can expect the continued conception and development of new and potentially revolutionary microarray-based solutions for the whole spectrum of biological issues.

Note: Web Tables A to N are available on the Nature Genetics website.

Acknowledgments

We are grateful to M. Murphy, K. Belfrage, S. Katsabanis and D. Diyagama for their assistance with all aspects of microarray production and use and to Cold Spring Harbour Laboratory Press, in particular K. Jaansen, J. Sambrook and the contributors to Microarrays, A Molecular Cloning Manual, for granting permission to reproduce tables included in this manuscript. The PMCI microarray facility was established with the generous support of the Ian Potter Foundation and the Wellcome Trust. The ongoing operation of the Facility is supported by a Science Technology Innovation Initiative Grant from the State Government of Victoria, Australia.

1. Bowtell, D.D. Options available—from start to finish—for obtaining expression data by microarray. *Nature Genet.* **21**, 25–32 (1999).
2. Singh-Gasson, S. *et al.* Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nature Biotechnol.* **17**, 974–978 (1999).
3. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. High density synthetic oligonucleotide arrays. *Nature Genet.* **21**, 20–24 (1999).
4. Hughes, T.R. *et al.* Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnol.* **19**, 342–347 (2001).
5. Pollack, J.R. *et al.* Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nature Genet.* **23**, 41–46 (1999).
6. Albertson, D.G. *et al.* Quantitative mapping of amplicon structure by array CGH identifies CYP24 as a candidate oncogene. *Nature Genet.* **25**, 144–146 (2000).
7. Snijders, A.M. *et al.* Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genet.* **29**, 263–264 (2001).
8. Pinkel, D. *et al.* High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nature Genet.* **20**, 207–211 (1998).
9. Hayward, R.E. *et al.* Shotgun DNA microarrays and stage-specific gene expression in *Plasmodium falciparum* malaria. *Mol. Microbiol.* **35**, 6–14 (2000).
10. El-Sayed, N.M., Hegde, P., Quackenbush, J., Melville, S.E. & Donelson, J.E. The African trypanosome genome. *Int. J. Parasitol.* **30**, 329–345 (2000).
11. Lee, J.M., Williams, M.E., Tingey, S.V. & Rafalski, J.A. DNA array profiling of gene expression changes during maize embryo development. *Funct. Integr. Genomics* **2**, 13–27 (2002).
12. Osoegawa, K. *et al.* A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* **11**, 483–496 (2001).
13. Halgren, R.G., Fielden, M.R., Fong, C.J. & Zacharewski, T.R. Assessment of clone identity and sequence fidelity for 1189 IMAGE cDNA clones. *Nucleic Acids Res.* **29**, 582–588 (2001).
14. Knight, J. When the chips are down. *Nature* **410**, 860–861 (2001).
15. Bowtell, D.D. & Sambrook, J.F. *DNA Microarrays: A Molecular Cloning Manual* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2002).
16. Relogio, A., Schwager, C., Richter, A., Ansorge, W. & Valcarcel, J. Optimization of oligonucleotide-based DNA microarrays. *Nucleic Acids Res.* **30**, e51 (2002).
17. Rouillard, J.M., Herbert, C.J. & Zuker, M. OligoArray: genome-scale oligonucleotide design for microarrays. *Bioinformatics* **18**, 486–487 (2002).
18. 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).
19. Wei, Y. *et al.* High-density microarray-mediated gene expression profiling of *Escherichia coli*. *J. Bacteriol.* **183**, 545–556 (2001).
20. Hegde, P. *et al.* A concise guide to cDNA microarray analysis. *Biotechniques* **29**, 548–556 (2000).
21. Schena, M. *et al.* Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl Acad. Sci. USA* **93**, 10614–10619 (1996).
22. 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).
23. Bertucci, F. *et al.* Sensitivity issues in DNA array-based expression measurements and performance of nylon microarrays for small samples. *Hum. Mol. Genet.* **8**, 1715–1722 (1999).
24. Yue, H. *et al.* An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression. *Nucleic Acids Res.* **29**, E41 (2001).
25. Wang, X., Ghosh, S. & Guo, S.W. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res* **29**, E75 (2001).
26. Lock, C. *et al.* Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. *Nature Med.* **8**, 500–508 (2002).
27. Alizadeh, A.A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
28. Bittner, M. *et al.* Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature* **406**, 536–540 (2000).
29. Dhanasekaran, S.M. *et al.* Delineation of prognostic biomarkers in prostate cancer. *Nature* **412**, 822–826 (2001).
30. Golub, T.R. *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).
31. Hedenfalk, I. *et al.* Gene-expression profiles in hereditary breast cancer. *N. Engl. J. Med.* **344**, 539–548 (2001).
32. Shipp, M.A. *et al.* Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Med.* **8**, 68–74 (2002).
33. Sorlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).
34. van't Veer, L.J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
35. Volm, M., Koomagi, R., Mattern, J. & Efferth, T. Expression profile of genes in non-small cell lung carcinomas from long-term surviving patients. *Clin. Cancer Res.* **8**, 1843–1848 (2002).
36. Miki, R. *et al.* Delineating developmental and metabolic pathways *in vivo* by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc. Natl Acad. Sci. USA* **98**, 2199–2204 (2001).

37. Iyer, V.R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87 (1999).
38. Lo, A.W. *et al.* A novel chromatin immunoprecipitation and array (CIA) analysis identifies a 460-kb CENP-A-binding neocentromere DNA. *Genome Res.* **11**, 448–457 (2001).
39. Shannon, M.F. & Rao, S. Transcription. Of chips and ChIPs. *Science* **296**, 666–669 (2002).
40. Ahrendt, S.A. *et al.* Rapid p53 sequence analysis in primary lung cancer using an oligonucleotide probe array. *Proc. Natl Acad. Sci. USA* **96**, 7382–7387 (1999).
41. Lindblad-Toh, K. *et al.* Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *Nature Biotechnol.* **18**, 1001–1005 (2000).
42. Lindblad-Toh, K. *et al.* Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nature Genet.* **24**, 381–386 (2000).
43. Ziauddin, J. & Sabatini, D.M. Microarrays of cells expressing defined cDNAs. *Nature* **411**, 107–110 (2001).
44. Kerr, M.K. & Churchill, G.A. Statistical design and the analysis of gene expression microarray data. *Genet Res.* **77**, 123–128 (2001).
45. Dudley, A.M., Aach, J., Steffen, M.A. & Church, G.M. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. *Proc. Natl Acad. Sci. USA* **99**, 7554–7559 (2002).
46. Hill, A.A. *et al.* Evaluation of normalization procedures for oligonucleotide array data based on spiked cRNA controls. *Genome Biol.* **2**, research0055 (2001).
47. Yang, Y.H. *et al.* Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**, e15 (2002).
48. Van Gelder, R.N. *et al.* Amplified RNA synthesized from limited quantities of heterogeneous cDNA. *Proc. Natl Acad. Sci. USA* **87**, 1663–1667 (1990).
49. Karsten, S.L., Van Deerlin, V.M., Sabatti, C., Gill, L.H. & Geschwind, D.H. An evaluation of tyramide signal amplification and archived fixed and frozen tissue in microarray gene expression analysis. *Nucleic Acids Res.* **30**, E4 (2002).
50. Colantuoni, C., Henry, G., Zeger, S. & Pevsner, J. Local mean normalization of microarray element signal intensities across an array surface: quality control and correction of spatially systematic artifacts. *Biotechniques* **32**, 1316–1320 (2002).
51. Stoeckert, C.J., Causton, H.C. & Ball, C.A. Microarray databases: standards and ontologies. *Nature Genet.* **32**, 469–473 (2002).
52. Saal, L.H. *et al.* BioArray software environment: a platform for comprehensive management and analysis of microarray data. *Genome Biol.* **3**, software0003.1–0003.6 (2002).
53. Fielden, M.R., Halgren, R.G., Dere, E. & Zacharewski, T.R. GP3: GenePix post-processing program for automated analysis of raw microarray data. *Bioinformatics* **18**, 771–773 (2002).
54. Geschwind, D.H. Sharing gene expression data: an array of options. *Nat. Rev. Neurosci.* **2**, 435–438 (2001).
55. Kellam, P. Microarray gene expression database: progress towards an international repository of gene expression data. *Genome Biol.* **2**, reports4011 (2001).
56. Microarrays standards at last. *Nature* **419**, 323 (2002).
57. Coming to terms with microarrays. *Nature Genet.* **32**, 333–334 (2002).
58. Gardiner-Garden, M. & Littlejohn, T.G. A comparison of microarray databases. *Brief. Bioinform.* **2**, 143–158 (2001).
59. Bilban, M., Buehler, L.K., Head, S., Desoye, G. & Quaranta, V. Normalizing DNA microarray data. *Curr. Issues Mol. Biol.* **4**, 57–64 (2002).
60. Quackenbush, J. Microarray data normalization and transformation. *Nature Genet.* **32**, 496–501 (2002).
61. Ripley, B.D. The {R} project in statistical computing. *MSOR Connections. Newsletter of the LTSN Maths, Stats & OR Network* (The University of Birmingham, Edgbaston, U.K.) **1**, 23–25 (2001).
62. Slonim, D.K. From patterns to pathways: gene expression data analysis comes of age. *Nature Genet.* **32**, 502–508 (2002).
63. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
64. Brown, M.P. *et al.* Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA* **97**, 262–267 (2000).
65. Pomeroy, S.L. *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442 (2002).
66. Khan, J. *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.* **7**, 673–679 (2001).
67. Xu, Y. *et al.* Artificial neural networks and gene filtering distinguish between global gene expression profiles of Barrett's esophagus and esophageal cancer. *Cancer Res.* **62**, 3493–3497 (2002).
68. Ramaswamy, S. *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA* **98**, 15149–15154 (2001).
69. Holmes, G. & Hall, M.A. A development environment for predictive modelling in foods. *Int. J. Food Microbiol.* **73**, 351–362 (2002).
70. The Gene Ontology Consortium. Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425–1433 (2001).
71. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. *Nature Genet.* **25**, 25–29 (2000).
72. Delozier, E.P. & Lingle, V.A. MEDLINE and MeSH: challenges for end users. *Med. Ref. Serv. Q* **11**, 29–46 (1992).
73. Lowe, H.J. & Barnett, G.O. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *J. Am. Med. Assoc.* **271**, 1103–1108 (1994).