MICROARRAYS IN MEDICAL RESEARCH

Maria Kamarinos and Mark Waltham
Pharmacogenomics Unit, St Vincent's Institute of Medical Research, VIC 3065

Microarrays allow screening for expression changes in thousands of genes in a single experiment. Application and technical improvement of the technology is evolving at a rapid rate with the publication of hundreds of microarray based manuscripts in the last five years. Unfortunately, many researchers initiate projects with limited understanding of the technology and little scope for long-term follow up. Clearly a critical factor to ensure success of any microarray experiment involves thorough understanding of the technology and its limitations. In this review we will give an overview of the technology with respect to microarray expression profiling, including technical information on how to perform good microarray experiments, a discussion of the common pitfalls, where to purchase inexpensive Australian-made arrays, and what may be ahead for the researcher once the initial array data has been obtained. Other applications of microarray technologies, such as proteomics (antibody or lysate arrays) and array comparative genome hybridisation, will not be discussed here but excellent applications in those areas can be found in recent publications (1-3).

What is a Microarray?

In essence microarrays consist of individual deposits of DNA fragments organised at high density onto a solid support. Experiments revolve around the basic principle of competitive hybridisation between your sample and the DNA fragments deposited on the microarray. DNA fragments can be cDNA clones with one clone representing a single gene or oligonucleotides where one or multiple oligonucleotides represent each gene. Commonly, DNA fragments are deposited onto a glass slide surface but they can also be deposited on a nylon membrane. In addition to the standard cDNA and oligonucleotide microarrays, photolithographic technology can be applied to synthesise oligonucleotides directly on the glass slide surface. This technology has been used by the Affymetrix company to generate their product known as GeneChips.

The differences in physical composition and synthesis of the three microarray platforms have an impact on cost and availability, but these factors are largely out of the control of academic scientists. The important differences between the three platforms lie in how samples must be prepared, ease of data acquisition and options for data analysis. It is these latter differences that will guide the choice of platforms. Quantities of RNA required for array interrogation also differ slightly between platforms (**Table 1**) and this can also guide the choice.

The Basic Steps

Microarray experiments differ from many conventional experimental techniques in that the majority of the work lies in planning and data analysis with less time spent at the bench. The success of a microarray experiment is highly dependent on ensuring quality control measures are met at each stage of sample preparation and hybridisation. From hereon we will restrict our discussion to microarrays of spotted cDNAs or oligonucleotides on glass (currently the more frequently used platform) where two samples are compared by simultaneous hybridisation to one microarray. Other platforms such as nylon arrays and Affymetrix GeneChips where only one sample is hybridised to the array at a time involve a similar principle but are not discussed here in any detail. Recommended additional information on the use of microarrays can be found in (4) and (5).

The plan and design of a microarray experiment is a critical stage which must be considered carefully. The choice of experimental design is dependent on your aim with direct and indirect comparisons being the major types. Direct comparisons compare two samples on the same slide and indirect comparisons compare each sample to the same reference sample on two different slides (6). A well-designed microarray experiment must incorporate replicates with the major categories being technical replicates, dye-swaps and biological replicates. Technical replicates repeat a comparison on a different slide using identical samples to assess for variability in

Table 1. Quantities (μg) of total RNA required for analysis

	Standard Protocol	RNA Amplification*	
		One round	Two rounds
Isotope filter array	3 to 5	0.1	0.01
Spotted cDNA/oligo	15 to 30	1	0.1
Affymetrix	5 to 20	0.1	0.001

Quantities provided are approximate only and based on the experience of authors. *A number of RNA amplification strategies have been published and commercial kits available. Many are based on a procedure known as Erberwine amplification (10). Amplified RNA (aRNA) is, however, often unsuitable for interrogation on oligo-based arrays.

Table 2. Free software for microarray analysis

ScanAlyze, image analysis, http://rana.lbl.gov/EisenSoftware.htm
TIGR Spotfinder, image analysis, http://www.tigr.org/software/
SNOMAD, normalisation, http://pevsnerlab.kennedykrieger.org/snomadinput.html
BioConductor, normalisation, statistical analysis, http://www.bioconductor.org/
TIGR Mev, clustering, http://www.tigr.org/software/
Cluster and TreeView, clustering, http://rana.lbl.gov/EisenSoftware.htm
BRB Tools, clustering, statistical analysis, http://linus.nci.nih.gov/BRB-ArrayTools.html
BASE, data management, http://base.thep.lu.se/
Onto-Express, helps interpret genetic profiles, http://vortex.cs.wayne.edu:8080

experimental technique. Biological replicates use a different biological sample on the second slide, for example an organ from a second animal with the same biological condition. Dye-swaps, where the labels are flipped, are useful to help identify dye biases that can occur when one dye preferentially binds to certain genes. The overall number of replicates depends on your resources but it is important to include enough replicates to enable estimation of variance.

The quality of RNA samples used must be high and thorough quality control is an integral part of the planning process. It is important that isolated RNA is free of contamination from carbohydrates, lipids or other substances used in the purification procedure and that all samples are purified using the same method. DNA contamination must also be removed as most labeling procedures label both DNA and RNA with equal efficiency. Labeled DNA will result in higher background readings and a masking of true expression values. RNA should be intact as degraded RNA can lead to the loss or reduction of signals from certain genes and biased expression patterns (7).

Once RNA samples have passed quality control and experimental comparisons have been planned the labelling process can begin. RNA samples are converted into cDNA that either has the fluorophore incorporated (direct labeling) or has aminoallyl-modified nucleotides in it to which the flurophores are coupled after the reverse transcriptase reaction (indirect labeling). All unincorporated flurophores must be removed by washing. The two samples to be compared are labeled in separated reaction with different flurophores, commonly Cy3 and Cy5. After the cleanup step samples are combined and applied to a single microarray slide and hybridisations performed overnight. The next day slides are washed and scanned to extract images. Data analysis then begins.

Scan Analysis and Statistics

Data analysis has three major steps including image analysis, normalisation and statistical analysis. Image analysis allows extraction of foreground and background intensity readings for both fluorophores for each gene. The background intensity is subtracted from total intensity for each gene in both channels. A ratio of these readings is formulated that defines relative gene expression between the samples. Data is then normalised to reduce experimental bias and statistical analysis is applied to identify differentially expressed genes. Statistical methods such as modified t-tests and ANOVA are commonly used to allow assignment of confidence levels (p-values) to the ratio measurements (8). Further downstream analysis such as clustering can also be applied to look for global expression patterns and is particularly useful for analysis and data visualisation of multisample analyses. Quite a number of commercial software packages are available (many are sold along with scanning instruments), although some good-quality, free packages are available that can assist in all steps (**Table 2**). Image analysis of Affymetrix chips requires the dedicated software package GeneChip Operating Software (GCOS) that is available with the complete Affymetrix system.

Differentially expressed genes identified from a microarray experiment should ideally undergo validation. Even the best designed and executed microarray experiment can yield false positives and data that has not undergone validation must be viewed with caution. Expression changes are ideally confirmed on independent replicate samples (where possible) using techniques such as quantitative PCR, Northern or Western analysis. Such 'downstream validation' is usually an essential component of project to refine lists of differentially expressed genes.

Made in Australia

The price and availability of the major species of expression microarrays (human, mouse, rat and yeast) has dramatically improved for the Australian research community over the last three years. This can be attributed to three major factors: (1) the establishment of spotting facilities within a number of academic centres as well as the Australian Genome Research Facility (AGRF, Melbourne); (2) a grant from the Australian Cancer Research Fund (ACRF) to purchase cDNA libraries and oligos; (3) a substantial reduction in price for Affymetrix chips (www.affymetrix.com); (4) some reductions in price for consumables associated with microarray hybridisations (Amersham, http://www.amershambiosciences.com). **Table 3** lists the sites and information of the major array spotting facilities within Australia (the stocks of arrays available for purchase do however fluctuate throughout the year). The purpose of the grant from ACRF ($1.3 million) has been to facilitate the coordinated distribution of cDNAs and oligos to the centres and this has been extremely beneficial in assisting mass production of arrays. This grant also supports an annual Australian microarray conference and helped seed the Australasian Microarray and Associated Technologies Association (AMATA) whose objectives and activities can be found at www.microarrays.com.

Databases and Data Standards

A useful strategy to further investigate genes identified from an array study comes from examining other microarray data stored in web-accessible databases. For example, in an experiment performed to identify differentially expressed genes following growth factor stimulation of a cell line, we can take this gene set and look at the relative expression level across a broad range of cell or tumour types, including clinical specimens. In this approach we can explore if individual or the entire set of genes are more abundant in specific tumour types or if they associate with poorer patient prognosis. Given that the same probe sets (cDNAs, oligos) will often be represented on an expression array irrespective of where the array was manufactured, the chances that a particular gene or genes of interest are represented in the same species database can be quite high. The Affymetrix technology platform has particular advantages in this respect.

Understandably these databases are growing extremely fast. The increasing numbers of samples and experimental conditions represented in these databases provides for an increasingly valuable research tool that can often complement simpler array profiling experiments from any one lab. Examples of popular and freely assessable microarray databases include the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) and the National Human Genome Research Institute Microarray Project (http://www.nhgri.nih.gov/DIR/Microarray/main.html). GeneLogic's Affymetrix subscription database (ASCENTA system; www.genelogic.com) is particularly powerful boasting expression profiles from approximately 9,000 human, mouse and rat samples.

In order to ensure that a certain quality level of deposited data is maintained and that exact details of samples and array information can be traced, a number of organisations have set about to standardise array data and reporting. One such organisation is the Microarray Gene Expression Data society (MGED, www.mged.org) which has developed guidelines for reporting the microarray results known as Minimal Information About a Microarray Experiment (MIAME) (9). The primary objectives of these standards are to help support continuing expansion of the databases, aid in the development of cross-platform data mining tools, and facilitate the sharing of high quality, appropriately annotated data within the life sciences community. It is important to note that these guidelines have also been widely adopted by major scientific journals as a requirement for publication. Details of MIAME reporting requirements are usually available from the various academic and commercial centres where arrays are produced.

Table 3. Major Australian microarray spotting facilities for eukaryote expression analysis

Queensland

(1) SRC Microarray Facility (IMB) www.microarray.imb.uq.edu.au
Human (oligo), Mouse (cDNA & oligo), Drosophila (cDNA), custom.

New South Wales

(1) The Ramaciotti Centre for Gene Function Analysis (UNSW) www.genomics.unsw.edu.au
Human (cDNA & oligo), Mouse (cDNA & oligo), Rat (oligo), Yeast (oligo), custom.
(2) Garvan Institute (Core Research Facility) www.garvan.org.au
Affymetrix workstation.

Australian Capital Territory

(1) John Curtain School of Medical Research, http://jcsmr.anu.edu.au
Affymetrix workstation.

Victoria

(1) Australian Genome Research Facility (Melbourne) www.agrf.org.au
Human (cDNA & oligo), Mouse (cDNA & oligo), *Affymetrix workstation*, custom.
(2) Peter MacCallum Cancer Institute Microarray Facility (PMCI) www.ccgpm.org
Human (cDNA & oligo), Mouse (cDNA & oligo), custom.

South Australia

(1) Adelaide Microarray Facility (Uni. of Adelaide) www.microarray.adelaide.edu.au
Human (cDNA & oligo), Mouse (oligo), Rat (oligo), custom.

Western Australia

(1) Lotterywest State Microarray Facility www.lsmaf.org.au
Human (oligo), Mouse (oligo), *Affymetrix workstation*, custom.

Listed are the major facilities where microarrays can be purchased (eukaryote species) or have facilities available for academic collaboration. Availability of arrays will vary considerably at each centre, though web sites will usually indicate species and numbers that are currently available. Protocols, useful links and other information is also available from each facility or via the AMATA web site (Australasian Microarray and Associated Technologies Association, www.microarrays.org.au). At present, nine complete Affymetrix workstations are in operation within Australia. Several (as indicated) are available for contract or academic collaborations.

Getting it Finished

Perhaps the most unanticipated aspect for those starting in the microarray area is the time taken to complete an entire project. This is particularly true for academics where the main end objective is to publish with a focussed conclusion pertaining to the original biological question. This is sometimes because confirmation of the gene list and performing follow-up studies takes considerably longer than expected. Often, however, it can be the result of the technology generating 'too much information'. This of course is not the case for every microarray experiment and will vary depending on the particular samples and approach - but often what seems to be a simple system to study can turn out to be far more complicated. In other words, occasions when you may expect very few differentially expressed genes (for say a simple binary comparison) can be quite the opposite. For many in this situation, an attempt to rationalise the biological significance of a list of many genes (perhaps hundreds) identified from a single microarray experiment can be overly daunting. Many at this juncture will: (a) give up and not bother to do anything with the dataset; (b) select a subset of the genes that 'make sense' and pursue those in further validation or functional studies; or (c) publish a compendium of identified genes offering scant explanation as to why they are differentially expressed and leaving it for others to potentially make sense of it at a later date. Unfortunately option (a) is probably the most frequent course of action. Researchers will either redesign their experiments or decide that microarrays are not for them. As for option (b) and (c), it is our opinion that neither of the two are necessarily the more appropriate action, and it seems that the course taken will usually depend on the personality of the individual researchers.

To counter these situations are the many instances when microarray experiments give a simple and relatively straight forward answers to the system under investigation. These occasions are not always from the simple binary comparisons using samples that are already well characterised by other methods, nor are they necessarily datasets where very few genes are identified as differentially expressed. Due to this, perhaps the best advice to provide regarding microarray technology is to start out with realistic expectations on what the technology can deliver, plan it well, and be aware and prepare for what may be ahead after the hybridisation. While the experiences of people who have used microarrays may be varied, there is little debate that over the last five years the technology has delivered remarkable insight into biological processes and has enormous future potential in many fields within the molecular life sciences. From the authors' point of view, the most appealing aspect of the technology is that discovery of biological processes can often be derived without any preconceived ideas or expectations. Discovery in these instances can be the most gratifying.

References

1. MacBeath, G. (2002) *Nature Genet.* **32**, 526-532
2. Zhu, H., Bilgin, M., and Snyder, M. (2003) *Ann. Rev. Biochem.* **72**, 783-812
3. Ishkanian, A.S., Malloff, C.A., Watson, S.K., deLeeuw, R.J., Chi, B., Coe, B.P., Snijders, A., Albertson, D.G., Pinkel, D., Marra, M.A., Ling, V., MacAulay, C., and Lam, W.L. (2004) *Nature Genet.* **36**, 299-303
4. Holloway, A.J., van Laar, R.K., Tothill, R.W., and Bowtell, D.D.L. (2002) *Nature Genet.* **32**, 481-489
5. Churchill, G.A. (2002) *Nature Genet.* **32**, 490-495
6. Yang, Y.H., and Speed, T. (2002) *Nature Rev. Genet.* **3**, 579-588
7. Aurer, H., Lyianarachchi, S., Newsom, D., Klisovic, M.I., Marcucci, U., and Kornacker, K. (2003) *Nature Genet.* **35**, 292-293
8. Draghici, S. (2002) *Drug Discovery Today* **7**, S55-S63
9. Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., Gaasterland, T., Glenisson, P., Holstege, F.C.P., Kim, I.F., Markowitz, V., Matese, J.C., Parkinson, H., Robinson, A., Sarkans, U., Schulze-Kremer, S., Stewart, J., Taylor, R., Vilo, J., and Vingron, M. (2001) *Nature Genet.* **29**, 365-371
10. Van Gelder, R.N., von Zastrow, M.E., Yool, A., Dement, W.C., Barchas, J.D., and Erberwine, J.H. (1990) *Proc. Nat. Acad. Sci. USA* **87**, 1663-1667