

Showcase on Research

Showcase on Microarray Technology

Sean Grimmond

Institute of Molecular Bioscience, University of Queensland, St Lucia, QLD 4072

Introduction

The last decade has seen a quantum shift in our approach to molecular genetics. Reverse genetics and positional cloning has been superseded by the completion of large-scale genome sequencing projects. In the case of the mammalian genomes, a combination of EST sequencing and exon-finding from complete genomic sequences has led to the first prediction of entire gene complement or transcriptome. The challenge (and the race) for the entire biological research community is now to integrate this enormous but finite set of genes into understanding of mammalian biology.

Determining gene function over the last two decades has followed a well-worn path. First, functional predictions were made based on homology to known function domains. Second, RNA expression patterns in key tissues or samples were used to infer possible function. This type of approach is limited by the rate at which genes can be interrogated and the requirement of large amounts of RNA required to assay expression. Microarray expression stands at the forefront of new methodologies being used to monitor expression on a genome-wide scale.

Microarray expression profiling: the concept

Hybridisation-based methodologies have been standard molecular biology tools for quantifying and analysing nucleic acids for the last three decades. Microarrays rely on the similar methodologies, namely detecting the hybridisation of gene-specific probes to RNA molecules. Instead of a single DNA probe being labelled and hybridised to RNA tethered to a semiporous substrate (the nylon filter), a complex population of labelled RNA molecules are hybridised in a very small volume to a battery of DNA probes tethered to discrete locations on a non-porous substrate. In most cases, more than one RNA target sample is applied to array by labelling samples with different fluorophores. After high stringency washing the amount of target an-

nealed to each probe is recorded by fluorescence imaging (whether by confocal laser scanning or CCD camera imaging). Finally, image analysis software converts captured images from fluorescence images to numerical output. These data are then compiled and screened to find genes that display differential expression.

Microarray manufacture

Microarray production can be divided into two groups. One is where the probes are synthesised *in situ* on a substrate, such as the Affymetrix GeneChip. The other includes those where nucleic acid material is printed onto discrete locations on a substrate. The Affymetrix GeneChip consists of many thousands of short oligonucleotides synthesised on the surface of a silicon wafer substrate using a process known as photolithography. A panel of perfectly matching and partially matching oligonucleotide probes are included for each gene to improve discrimination of specific versus non-specific signals. This type of microarray preparation requires industrial scale manufacturing facilities.

Printed microarrays are generated using an accurate robotic system that collects small volumes of probe material in a printing system and then deposits minute volumes of material at discrete locations on the array. The size of arrayed spots is typically in the range of 80-250 microns and elements are printed at densities of 100-6000 spots per cm². At these extreme densities, at least 80,000 probes can be printed on a 20 mm x 60 mm slide.

The delivery system used to deposit probe material can be divided into those requiring contact with a substrate using carefully tooled pins and non-contact printing tools such as piezoelectric or ink jet printing technologies. The engineering of the infrastructure needed to create spotted arrays has been refined to such a point that printed microarrays can be rapidly manufactured in academic laboratories, rather than large industrial establishments and this has led to their popularity.

The material printed onto microarray

substrates is often dictated by the task proposed. RNA expression studies have generally involved the printing of PCR-purified inserts of commercially available sequence verified cDNA clone sets (e.g. http://www.resgen.com/products/IMAGE_Clones.php3). DNA dosage studies have required arrayed bacterial artificial chromosome (BAC) clones from key genomic regions and mutation screening has required sequence-specific oligonucleotides to be deposited.

Recently, advances in the high throughput synthesis of long oligonucleotides has led to 50-70mers being used as a source of probes for expression arrays. These elements have several advantages including exon specificity, greater discrimination of homologous sequences and controlled molarity and are likely to become a more popular source of array probes as cost of production falls.

Applications of DNA microarrays will now be considered.

Gene expression studies

Much of the pioneering work leading to the success of microarrays came from into expression profiling of yeast. The transcriptional consequences of well-characterised biochemical events such as cell cycle progression and diauxic shock were recorded for the entire yeast transcriptome shortly after completion of the entire yeast genomic sequence (1). These studies involved performing a series of expression profiling experiments on RNA samples taken from yeast cells over time.

Two important findings were made from these experiments. The first was that changes in the physiological state of organism results in the modulation of expression of specific sets of genes or "gene pathways". The second was that genes displaying similar expression profiles often played roles in the same "gene pathway". The implications of these observations meant that expression profiles could be used to predict physiological events taking place within cells and that function can be predicted, even for previously uncharacterised novel genes. Similar ob-

Showcase on Research

Showcase on Microarray Technology (contin.)

servations have been confirmed now in higher eukaryotic cell culture and tissues.

Expression profiling of large panels of patient samples can also be used to provide transcriptional insight into complex states and heterogenous diseases.

Studies of the expression from 65 cancer cell lines has led to the defining of classes of lineage-specific or tissue-specific genes (2). Also, new expression-based taxonomies have been defined by performing expression profiling on melanomas, leukemias and breast carcinomas (reviewed in ref. 3). In the case of diffuse large B cell lymphoma, expression profiling defined subtypes with significant prognostic implications.

Genomic analyses

Microarray-based assays of DNA dosage and sequence content have also recently been developed. Mapping variation of gene dosage above and below $2n$ has been studied in cancer genetics, as regions that are frequently amplified or deleted often contain oncogenes and tumour suppressor genes, respectively. High-density microarrays made up of probes that are either large genomic BAC clones or even conventional cDNA micro-arrays have been used for this purpose (4).

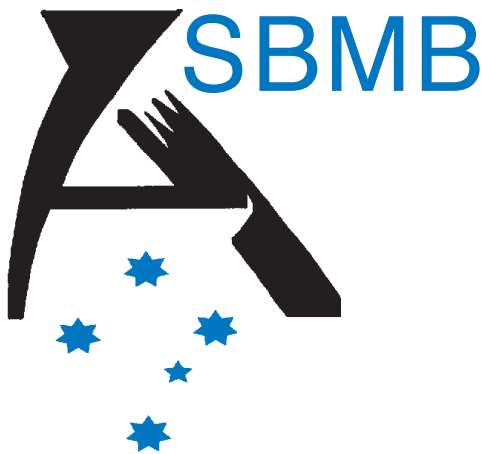
One of the other fruits of the human genome project has been the recording of vast numbers of small nucleotide polymorphisms (SNPs) between individuals. Many efforts are under way to derive methods that allow researchers to exploit these polymorphisms to rapidly map loci. Traditionally the detection of SNPs has required the hybridisation of perfect and mismatched oligonucleotide probes against DNA targets or primer extension studies. Both these methods have been modified to work on a microarray substrate allowing for massive increases in processivity. Affymetrix have generated re-sequencing chips for specific genes including TP53 and critical genes from HIV genome (5).

Informatics and data mining

The third essential component to microarray studies is bioinformatics. While, the greatest strength of microarrays is their ability to generate measurements in a massively parallel fashion, this amounts to nothing but confusion unless appropriate data management and storage and data mining tools are available.

Any given microarray experiment requires the identity of every probe to be integrated with all related information for that gene (chromosomal position, biochemical pathway data, clinical involvement, all PubMed information, orthologs, etc). The system that handle integrates this data must be accessible and amenable to rapid and regular updating as much of this data is constantly evolving.

Once each probe has been identified, the amount of test and reference signal detected for that spot must be accurately determined. Ideally this includes both filtering and flagging procedures to prevent artifactual measurements arising from dust, local background or other experimental



Showcase on Research

Showcase on Microarray Technology (contin.)

variables affecting the data. There is a plethora of publicly available and commercial software solutions for this task. As the size and density of microarrays increase it is necessary to employ better and image analysis software to automate addressing of probes and making measurements.

Expression data from single experiments is relatively "shallow" in that only the small fraction of differentially expressed genes are of any use. The real power of this sort of expression profiling requires the combination of expression data from a series of experiments. Before data can be mined from a panel of experiments, the data need to be normalised so that the values from each experiment are comparable. Median "centring" has been used as a standard method for many profiling experiments, an expanding number of statistical strategies and aids are being developed to perform this task.

Once a large data set has been compiled and normalised, the data are ready to be mined. The two most commonly asked questions of these data are as follows. Which genes are being expressed in a similar fashion? Which genes possess expression profiles that correlate with a particular phenotype? There is currently a near-exponential increase in the number of bioinformatic tools designed to perform these tasks. In the case of pattern recognition, a mixture of hierarchical clustering, partitioning methods and machine learning tools are being used on array data (reviewed in ref. 6).

Once a pattern has been identified within an expression data set, a subsequent task is to try and associate that cluster with a possible biological pathway. In many cases this is not an easy task as it is unlikely that the microarray experimenter is a master of all aspects of biochemistry and cell biology. For this reason, the development of visualisation tools that integrate gene expression data with the vast amounts of biochemical, genetic, metabolic and functional data we have accrued with expression clusters are essential. This area of bioinformatics is in its infancy but will be a key component in accelerating our ability to interpret these data.

The final and possibly most exciting challenge at this time is the need to develop systems that allow us to look at expression data in an additive fashion,

rather than experiment by experiment. The careful collation of each microarray experiment into an appropriately designed relational database can allow the researcher to analyse global expression changes or to determine how a given gene behaves in hundreds of differential environmental, pathological, genetic and pharmacological scenarios.

The power of this sort of analysis can be seen at the Stanford Microarray Database where the expression behaviour of a given gene can be reviewed for hundreds or even thousands of microarray experiments (www.DNAchip.org). This sort of analysis can rapidly define in what context the expression of a given gene is modulated.

Microarraying in Australia

Despite the development of microarray technologies having only occurred very recently, Australia has rapidly developed a national network with the goal of providing high quality microarray resources to its research community. As there is a substantial outlay required for implementing manufacture of microarrays (including the purchasing of robotic arrayers, array readers, computing infrastructure and software) most major cities have formed nodes or centres to generate arrays.

The Australian Genome Research Facility (AGRF) has also committed to the provision of microarray reagents to the research community. A recent initiative by the AGRF has to provide access to yeast, mouse and human clone sets as well as microarray resources throughout Australia. Large initiatives into plant expression profiling have also been undertaken by the CSIRO. Recent national meetings have been held to disseminate this technology to interested parties and will continue to be expanded upon, as more resources become available.

The future

It is difficult to predict where this technology is likely to be within 2-3 years, due to the rapid pace at which further developments are occurring. It is clear that many standard techniques are being adapted to fit the microarray model so that genes or their products can be as-

sayed in a massively serial fashion. Recent advances have intragenic region microarrays to study transcription-factor binding sites binding regions in promoters (7), antibody-based arrays to quantify specific proteins in complex samples (8) and even high throughput immunohistochemistry or *in situ* hybridisation on tissue microarrays (9). Even more exciting is the recent assaying of gene function by microarray where full-length cDNAs arrayed onto a substrate are transiently transfected into mammalian cells layered upon the surface of a microarray and phenotypic changes are recorded (10).

Conclusion

Microarray technologies are rapidly providing biologists with the ability to analyse the genome and transcriptome in an unprecedented fashion. Bioinformatic tools to interpret these data are also being developed at a rapid pace. Microarray experiments are being used as hypothesis-generating machines, predicting key genes in important biological processes and inferring function for novel genes. Our ability to define more subtle patterns in large expression data sets are set to improve with the evolution of data-mining tools and are likely to make enormous advances in biomedical research.

References

1. DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997) *Science* **278**, 680-686
2. Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, V.C., Myers, T. G., Andrews, D. T., Scudiero, D. A., Eisen, M. B., Sausville, E.A., Pommier, Y., Botstein, D., Brown, P. O., and Weinstein, J. N. (2000) *Nat. Gen.* **24**, 236-244
3. Liotta L., and Petricoin, E. (2001) *Nat. Rev. Gen.* **1**, 48-56
4. Pollack, J. R., Perou, C. M., Alizadeh, A.A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P.O. (1999) *Nat. Gen.* **23**, 41-46
5. Fan, J. B., Chen, X., Halushka, M. K., Berno, A., Huang, X., Ryder, T., Lipshutz, R. J., Lockhart, D. J., and Chakravarti, A. (2000) *Genome Res.* **10**, 853-860
6. Quackenbush, J. (2001) *Nat. Rev. Gen.* **2**, 418-427
7. Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. (2001) *Nature* **409**, 533-538
8. Haab, B. B., Dunham, M. J., and Brown, P. O. (2001) *Genome Biol.* **2**, 1-13
9. Kononen, J., Bubendorf, L., Kallioniemi, A., Barlund, M., Schraml, P., Leighton, S., Torhorst, J., Mihatsch, M. J., Sauter, G., and Kallioniemi, O. P. (1998) *Nat. Med.* **4**, 844-847
10. Ziauddin, J., and Sabatini, D. M. (2001) *Nature* **411**, 107-110