KEYWORDS: DNA chip, microarray, probe hybridization

Special section on techniques:

You want Ketchup with your DNA Chips? An Overview of Expression Microarrays

Bradley Coe
Pathology, University of British Columbia

Submitted April 2003

Introduction

Since Mark Schena published the first manuscript on microarray technology in *Science* magazine during the fall of 1995 [1] the technique has been rapidly adopted by the scientific community for its ability to analyze the expression of genes at a scale difficult to accomplish with traditional methods. Traditionally, analyzing gene expression has been a time consuming process involving setting up a separate experiment for each gene. These traditional techniques include Reverse Transcriptase PCR, Real Time PCR and Northern Blots, the details of which are beyond the scope of this manuscript. Compared to these techniques, the analysis of thousands of genes at once via a single expression microarray experiment allows researchers to acquire quantitative data at a significantly reduced cost per gene [2].

The basis of microarray technology lies in the ability of complementary strands of DNA to bind to each other under appropriate conditions. The microarray itself consists of DNA elements representing a single gene present as microscopic (less than 1mm in diameter) spots on a 75mm by 25mm glass slide. The RNA sample to be examined is labelled with fluorescent molecules and hybridized to an array. After hybridizing the sample of interest to the array, a measurement of the fluorescence intensity for each DNA element is measured and can then be utilized to determine the genes relative expression [2].

Unfortunately expression microarray technology is not a simple technique in practice. There are multiple types of arrays to choose from, many methods to label the sample RNA, and no simple rules for the analysis of generated data. The aim of this manuscript is to provide a general introduction to microarray technology and cover some of the issues and advantages this new technique offers.

Array Types

There are three main types of expression microarrays which differ based on the DNA elements they are constructed with. These types of DNA elements include cDNAs, short oligonucleotides (e.g. 25bp) and long oligonucleotides (e.g. 70bp).

Expression microarrays constructed with cDNAs represent the first type of arrays produced. Each cDNA is selected from a library of clones and PCR amplified in order to generate enough material for spotting onto a glass slide (details on spotting technologies are discussed below). These arrays are preferable in some cases due to their affordability and customizability when compared to oligonucleotide based technologies. Additionally cDNA arrays can be more sensitive in detection due to the larger size (100 to thousands of base pairs) of the array target when compared to oligonucleotide approaches. This allows more accurate detection of low copy number transcripts. However the conserved nature of some genes can allow multiple members of a gene family to bind to a single spot thus preventing an accurate measurement of the expression of such genes. Although this type of noise is not very common due to the fact that redundancy in codons allows high divergence in the RNA sequence between two genes while preserving a similar amino acid sequence. Another disadvantage lies in the fact that different cDNAs will be of differing lengths and sequence compositions, and thus vary widely in their melting temperatures causing problems with picking hybridization and wash conditions which will result in optimal data for every gene [2].

Short oligonucleotide arrays were pioneered by Affymetrix and consist of small 25 base pair oligonucleotides synthesized directly onto the chip by photolithography. Each gene is represented by ten distinct pairs of 25mers, one member of each pair representing a perfect match to a section of the transcript sequence and the other containing a mismatch at the 13th base in order to assess levels of non-specific hybridization. During the design process Affymetrix balances the melting temperatures of each array target in order to reduce any non-specific hybridization that could be caused by the varying melting temperatures found across transcripts [2, 3].

Long oligonucleotide arrays represent a middle ground between cDNA and short oligonucleotide arrays. Similarly to cDNA arrays the targets are first created in a microplate format and then printed onto glass slides post synthesis. In the case of the Qiagen array ready oligonucleotide set a 70bp length was chosen to optimize the balance between the specificity of short oligonucletide arrays and the sensitivity of cDNA arrays. Similarly to the Affymetrix arrays these oligonucleotides were also chosen in such a way as to ensure similar melting temperatures for each target [4].

The choice of which array technology to use is not a simple one. Arrays constructed from cDNAs are the best choice for most labs which construct their own arrays due to their low cost and customizability [2]. However the slide to slide consistency of Affymetrix arrays has proven popular with many laboratories despite their proprietary nature and the high cost per experiment [2, 3]. Additionally the recent availability of long oligonucleotide arrays which attempt to optimize the advantages and disadvantages of the other array technologies add further confusion to the choice [4]. All of these microarray technologies can produce good quality data, so the choice of which to use must be made by taking into account the cost and availability of the arrays as well as the arrays utility in analysing the pathways of interest to the researcher. Since not every array will contain all of the genes a particular researcher is interested in this may become one of the major factors in choosing an array technology/provider.

Array construction

The microarray technologies, which do not synthesize the target DNA directly on the chip demanded the development of a method for depositing the samples very accurately in order to create the high density arrays researchers demand. This is accomplished via microarray printing robots which come in many designs [2]. It is worth noting that high throughput spotting robots can easily cost a quarter million dollars or more and as such are outside the budgets of most labs which are forced to rely on commercially available chips.

The majority of spotters belong to the family of contact printers. These rely on a pin which is loaded with sample physically contacting the slide to deposit nanolitre scale volumes of printing solution [2].

The other popular method of printing, non-contact printing, works similarly to an ink-jet printer and benefits from the consistent features it creates. However this technology is limited by the viscosity of the product it can print and currently cannot match the small feature sizes and throughput produced by contact printing [2].

Another issue in array construction is the choice of an appropriate substrate on which to deposit the target DNA. Most microarrays are deposited on standard microscope sized glass slides which are coated with a substance that covalently binds to DNA. There are two main types coatings used on microarray slides which differ in how they form covalent interactions with the deposited DNA. Slides coated with aldehyde groups bind covalently to amino groups present in modified primers used to generate the target DNA while amine slides bind directly to the backbone of DNA through charge interaction [2].

Aside from the difference in how DNA is bound, aldehyde slides offer a slight advantage in decreased feature size due to their higher hydrophobicity but do involve a much longer post spotting processing time than amine slides. With amine slides the DNA is covalently attached to the slide by simply baking and UV crosslinking while aldehyde slides require a long (12 hour) dehydration step followed by washes with the reducing agent sodium borohydride [2].

Probe Generation

In order to perform microarray analysis on an RNA sample it must first be labelled in such a way that it can be quantitated by fluorescence after hybridization. The most popular dyes are cyanine 5 and cyanine 3, respectively, which are detectable with all commercial array scanners. All probe generation techniques can either be classified as direct labelling or indirect labelling. In direct labelling the fluorescent tag is covalently bound to the probe molecule via enzymatic or chemical means and in indirect labelling the fluorescent tag is attached indirectly through a bridge molecule such as an antibody to a biotin conjugated nucleotide analogue [2].

The most common method of direct labelling is reverse transcription with nucleotide analogues. This procedure is relatively simple and involves using a reverse transcriptase to generate single stranded cDNA from an oligo-dT primer. Labelling is accomplished by either including a nucleotide which is directly attached to a fluorescent tag into the transcription reaction or by incorporating an aminoallyl nucleotide which can later be covalently bonded to a reactive dye molecule. The aminoallyl technique benefits from more efficient incorporation due to the smaller size of the nucleotide analogue used resulting in brighter hybridizations and more uniform labelling [2].

Since mRNA makes up only < 5% of total RNA reverse transcription can require tens of micrograms of RNA (This number is highly variable) per microarray experiment. In order to improve the utility of microarrays for smaller samples researchers use labelling procedures which amplify the amount of starting mRNA. The Eberwine procedure is a T7 RNA polymerase based technique for amplifying the starting material. In this procedure an RNA sample is first reverse transcribed into single stranded cDNA using an olgo-dT primer containing a T7 promoter sequence. DNA polymerase is then used to convert the single stranded cDNA into a double stranded DNA. The T7 RNA polymerase is then used to produce many aRNA (amplified RNA) copies from each double stranded cDNA. Each round of Eberwine amplification produces a 100 fold amplification of the starting material and three rounds produces 1 million fold amplification. Although this technique greatly amplifies the amount of probe the fact that the probe is composed of RNA introduces the need for great care in avoiding its degradation. Additionally three rounds of the Eberwine procedure requires 2 to 3 days of steady work. The resultant RNA from this procedure can be labelled in several ways. Firstly by incorporating fluorescent nucleotide analogues in the synthesis direct labelling can be accomplished. Additionally incorporating biotin or streptavidin labelled nucleotides can allow indirect labelling with fluorescent antibodies in a post hybridization labelling step [2].

A popular indirect labelling technique which allows significant signal amplification from a small amount of starting RNA (2 -5 µg) is tyramide signal amplification (TSA). TSA relies on first reverse transcribing the sample RNA using biotin or streptavidin nucleotides, and then hybridizing the resultant cDNA to an array. Following the initial hybridization the arrays are incubated with an antibody against biotin or streptavidin which is conjugated to a horseradish peroxidase (HRP). In the presence of hydrogen peroxide the HRP oxidizes fluorescent tyramide molecules (cyanine 3 and cyanine 5 tyramides are available) which in turn rapidly bind to the microarray surface adjacent to the antibody. Although this procedure requires several hours of treatment after the initial hybridization the procedure is very robust (a kit is available from NEN Life Science Products) and produces high quality data using minimal starting material [2, 5].

One important factor to consider in all of the labelling techniques is that differences in the sizes of the tagged nucleotides can lead to differences in their incorporation efficiencies. As such it is common to perform a flip flour experiment where the dyes are switched and results are compared to verify outliers [2, 6].

Hybridization

Microarray technology is based on the fact that complementary DNA sequences bind to each other under the appropriate conditions.

The temperature at which DNA strands best associate is a function of the GC contents and strand length. The temperature at which 50% of the base pairs are specifically associating is referred to as the Tm. Since G and C base pairing involves 3 hydrogen bonds while A and T base pairing involves 2 hydrogen bonds the strength of association between two strands is proportional to the GC content. The length of the complementary regions also determines the association strength. In addition the optimal binding temperature can be modified by the addition of denaturing agents such as formamide which reduce the Tm and salts which stabilize the negative charges of the DNA backbone further promoting association [2].

Despite the complexities, optimal temperatures and hybridization buffer conditions have already been worked out for all of the commercially available arrays and sticking to the recommended conditions will normally produce the best data [2].

An additional factor to consider is the difference between single and dual/multi channel hybridizations. In a single channel hybridization only one sample is hybridized to the array. This protocol is only suitable for Affymetrix style arrays which demonstrate high slide to slide consistency due to the photolithography process. Spotted arrays demonstrate higher degrees of variation in spot intensity due to variable DNA deposition in each array element and as such are only suitable for competitive multi channel hybridizations. An additional concern in single channel hybridizations is that saturation of any single array element will result in non-representative expression level data. In multi channel hybridization two or more samples are hybridized to the array at once and the ratio between the reference and samples signals on each spot represents the expression difference between them [2].

Scanners

Post hybridization fluorescence images are generated for each array experiment in order to allow downstream analysis of the microarray experiment. In the imaging of fluorescently labelled microarrays two main technologies are used. These include CCD (charge coupled device) and confocal laser based systems.

CCD based systems such as the Arrayworx scanner from API use a sensor similar to the one found in consumer digital cameras to detect the fluorescence intensity across an array. These systems incorporate a light source behind an excitation filter which specifically excites only one of the dyes used at a time. The

CCD itself sits behind an additional filter which allows only the light generated by the fluorescence activity of the dye to be detected. Measurements of light intensity by the CCD is accomplished by converting the photons that hit a discrete picture element (pixel) into electrons and then converting the electron count into a 16 bit number (0-65535) for storage in a tiff format image file. Since a CCD is rather limited in resolution multiple pictures are taken of each slide using paired excitation and emission filters for each dye present in the experiment. These pictures are then "stitched" together electronically to create a single high resolution image of the array. In order to adjust for varying hybridization intensities one simply adjusts the exposure time to linearly increase the signal. Additionally pixels can be binned together to produce a lower resolution image with increased signal intensities [2,7].

Confocal laser systems such as those from Axon use a laser to specifically excite the fluorescent dye on a slide and then a photomultiplier tube (PMT) to convert the resulting fluorescence into electrons which can be counted and converted into a 16 bit value for storage in a tiff image. By taking multiple readings over the surface of the slide an image of the array is constructed for each channel for downstream analysis. The confocal imaging systems have advantages in their high specificity due to the gating of PMTs and the exact specifications of the lasers which allow precise measurements with minimal crosstalk between fluorescent channels compared to the filtering system used by CCD scanners. However the confocal systems suffer from complicated set up where changes must be made to both PMT sensitivity and laser power in order to optimize the dynamic range of the resultant image for each array. Additional complication arises from the fact that these settings are not linear and small changes can cause large differences in the image intensities [2,8].

Although both scanning technologies are very reliable, the high cost of microarray scanners (they can easily cost in excess of 50 thousand dollars) can lead many labs choosing to have third party companies perform their hybridization experiments for them. Although this is typically a costly solution (around one thousand dollars per experiment in the case of Affymetrix arrays) it is much more economical than purchasing a scanner for labs with low throughput array needs.

Analysis

After acquiring images of an array experiments one must extract their gene expression data. Multiple programs are available to perform this step and are typically included with the purchase of a microarray scanner. Typically, images are first segmented into features corresponding to the spots on the microarray. The average pixel intensity is then calculated for each feature and a value corresponding to the local background is subtracted. The local background is usually obtained as a median of the pixel intensities in an area surrounding the feature. The purpose of this subtraction is to account for non-specific binding of probe to the array surface, fluorescence generated by the substrate itself and other imaging related noise [2].

After one has extracted the data for each feature on an array the next step lies in normalizing the data. Normalization is required in order to account for differences in the amount of sample labelled, dye fluorescent intensities, and labelling efficiencies between reactions. In other methods of analysing gene expression such as northern blots, RTPCR and real time quantitative PCR people typically use the signal from a housekeeping gene to compare expression of a gene of interest across samples. Unfortunately, this approach has proven overly simplistic in microarray analysis. As such many labs now rely on techniques that centre the mean signal across the entire array at a ratio of one to one between the two channels. Another problem that is typically corrected for in microarray analysis is that, at low intensities, ratio values can decrease below their actual values. This is corrected by sorting the data into a plot of mean intensity versus ratio for each feature. A regression line is then fitted to the data and a residual is calculated to straighten the regression line and centre it at a ratio of representing equivalent expression. This correction is referred to as LOWESS normalization and several software packages are available to perform this normalization. One particularly useful tool is the SNOMAD (Standardization and NOrmalization of MicroArray Data) website (http://pevsnerlab.kennedy krieger.org/snomadinput.html) [6] which can normalize data by a LOWESS transformation and apply various other data correction algorithms, the details of which are well beyond the scope of this paper.

After the data has been successfully extracted from a microarray experiment, a researcher is typically left with the "what next" problem. There are several commercial and freeware packages available to help interpret data. Some commonly used tools include cluster analysis which groups genes together based on similar expression patterns and more comprehensive pathway based tools which can help assign some meaning to your data. One example of such a comprehensive pathway based tool is The Dragon (Database Referencing of Array Genes ONline) Database which is a free web based tool (http://pevsnerlab.kennedykrieger.org/dragon.htm) to help map expression data onto pathways from the KEGG (Kyoto Encyclopedia of Genes and Genomes) database (http://bioinfo.weizmann.ac.il:3456/kegg/kegg.html).

It is important to note that data from microarray experiments must be taken with a grain of salt. Despite the vast research into methods to analyze microarray data no-one is yet sure how many replicates will be required to generate highly statistically significant data as even the best performed experiments can yield 1% or more false positives. Additionally there is no current method to determine what a minimal significant expression change is. Currently many people set an arbitrary 2 fold or higher cut-off for every gene in their analysis [9] however the actual situation is probably more complex due to differences in hybridization kinetics and expression levels between genes which may require a unique cut-off for each gene analyzed [2].

Issues in Experimental design

One of the biggest experimental issues plaguing microarray researchers is the choice of an appropriate reference for hybridization. Due to the large degree of heterogeneity between different tissue types, the choice is not simple. Other confounding issues include deciding whether the normal sample needs to be from the same patient as the sample of interest and from where the normal sample should be acquired. Essentially, the complexity lies in answering the question of what is normal. In the case of time course experiments or experiments comparing a treatment group to a no treatment group running a standard normal can be acceptable as the differences of interest are those found between experiments rather than those within an experiment [2].

Additional concerns lie in determining how many replicate experiments are needed in order to confirm data as statistically significant. This will prove necessary as currently each potential gene needs to be confirmed via another technology to gain widespread acceptance in publication [2,6].

Conclusion

Overall, despite its many complexities, microarray analysis is a very powerful technique for analysing gene expression. The ability to scan the expression of thousands of genes simultaneously more than makes up for the high cost and time involved in a single microarray experiment by providing data at a fraction of the cost per gene of other methods.

References

1) Schena, M., Shalon, D., Davis, R. W., and Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467-470, 1995

2) Schena, M. Microarray Analysis *Wiley-Liss* 2003

3) http://www.affymetrix.com/technology/index.affx

4) http://qiagen.com/catalog/auto/cget.asp?p=microarray_products

5) http://lifesciences.perkinelmer.com/downloads/H78414.pdf

6) Colantuoni, C., Henry, G., Zeger, S., Pevsner J. SNOMAD (Standardization and NOrmalization of MicroArray): web-accessible gene expression data analysis. *Bioinformatics* 11:1540-1, 2002

7) http://www.api.com

8) http://www.axon.com

9) Nautiyal, S., DeRisi, J.L., and Blackburn, E.H. The genome-wide expression response to telomerase deletion in *Saccharomyces cervisae*. *PNAS* 99: 9316-9321, 2002

Local Contacts

Microarray analysis is utilized by many Labs including: The BC Cancer Research Centre (www.bccrc.ca), The Prostate Centre (http://prostatelab.org/) and many Labs on UBC Campus

(www.ubc.ca).