Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Experimental Design and Analysis of Microarray Data

Claire H. Wilson, Anna Tsykin, Christopher R. Wilkinson, Catherine A. Abbott

School of Biological Sciences, Flinders University, PO Box 2100, Adelaide, South Australia, 5001, Australia; Hanson Institute, Adelaide, SA, Australia; Child Health Research Institute, Adelaide, SA, Australia; School of Mathematical Sciences, University of Adelaide, Adelaide, SA, Australia (claire.wilson@flinders.edu.au, anna.tsykin@adelaide.edu.au, christopher.wilkinson@adelaide.edu.au; cathy.abbott@flinders.edu.au)

The advent of microarray technology has significantly changed the way we can quantitatively measure and observe gene expression at the mRNA level within a given biological sample of interest, allowing for the monitoring of tens to hundreds of thousands of genes within a single experiment. The two main array platforms are spotted two-colour arrays and one-colour in situ-synthesized arrays. Microarrays are used for a wide range of applications including gene annotation, investigation of gene-gene interactions, elucidation of gene regulatory networks and gene-expression profiling of Saccharomyces cerevisiae and other fungal organisms. Academic researchers and both the pharmaceutical and agricultural industries have an enormous interest in developing microarrays both as diagnostic tools and for use in basic research into how pathogens, such as fungi, interact with their host. Microarray experiments generate vast quantities of raw gene expression data, therefore good experimental design and statistical analysis is required for the extraction of accurate and useful information regarding the expression of genes. In this review we firstly provide an overview of the arrival and development of microarray technology. We then focus on the issues surrounding experimental design and the processing of microarray images, followed by a discussion on methods for cleaning and normalizing raw gene expression data and a final discussion of the importance statistical analysis plays in identifying differentially expressed genes.

1. INTRODUCTION

Since the late 1990's, following the successful sequencing of the Escherica coli genome, there has been a rapid advancement in genome-scale sequencing of both prokaryotic and eukaryotic organisms. At present the publicly accessible National
Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov)
Entrez Genome Project database contains 257 complete or in progress eukaryotic
genome projects. Seventy of these projects are fungal, however many more projects
are underway in both public and private laboratories that are not yet accessible. This
rapid increase in genomic knowledge has largely driven the emerging discipline of
functional genomics, fuelling the development of high-throughput technologies and
computational methods for the rapid interpretation and extrapolation of information
on a genome-wide scale. Functional genomics aims to functionally annotate every
gene within the genome, their interactions with other genes and their involvement in
gene regulatory networks, hence, allowing for the study of biological problems at
levels of complexity that have never before been possible. Functional genomics and
the need for genome-wide expression analysis has been a major driver in the
development of DNA, protein and combinatorial chemistry array technologies.
DNA microarrays, which are used for simultaneously measuring the level of mRNA
gene products from a given biological sample are currently the most advanced of
these technologies and will be the focus of this chapter. Although proteins are the
ultimate products of genes, measuring mRNA expression levels is a good starting
point for functional gene characterization and currently it is a considerably cheaper
technology then measuring direct protein levels which utilizes mass spectrometry
resources.

In general microarrays are used to measure the concentration of each mRNA from
a given sample, providing a snapshot of expression at a single time point or relative
to another sample. This is achieved by monitoring the combinatorial interaction of a
set of DNA or mRNA fragments with a predetermined library of polynucleotide
probes. Before the emergence of this technology, using techniques such as Northern
blots or quantitative real-time RT-PCR (qRT-PCR), researchers were only able to
measure expression at the mRNA level for a limited number of genes. With the
advent of microarrays it is now possible for researchers to uniquely and
quantitatively measure the expression of tens to hundreds of thousands of genes at
any given time within a given biological sample on a single platform. There is
enormous potential for expansion of our scientific knowledge and discovery through
the application of microarrays into the investigation of gene-gene interactions and in
pharmaceutical and clinical research to enable further understanding of disease and
the creation of future diagnostic tools to individualize molecular medicine.

Analysis and handling of the huge amounts of raw gene expression data
generated from microarrays has rapidly become one of the major bottlenecks from
the utilization of this technology with an increasing reliance on bioinformatic based
innovations. Research from the fields of biology, statistics, mathematics, computer
science and physics are drawn together to further our understanding of biological
processes. Like all experiments the value of the data generated can be greatly
affected by the choice of experimental design, and the implementation and analysis
phases. With the ultimate goal being to make inferences among biological samples,
their genes and associated levels of mRNA expression, many factors must be
considered and integrated during each phase. Microarray data must be integrated
with nucleotide sequence data, knowledge of protein structure and function, and
with phenotypic and clinical data. The overall goal of this chapter is to provide the
reader with an overview of current DNA microarray technologies and to introduce
issues regarding the experimental design and analysis of microarray data.

This chapter will firstly provide an overview of microarray technology, discussing
cDNA and oligonucleotide arrays and their development as well as the steps
involved in their assembly and applications within the fields of mycology and
biotechnology. Following this issues concerning experimental design and
microarray image processing will be discussed. The latter half of this chapter will
then talk about cleaning and normalization of raw gene expression data followed by
a discussion on methods for statistically analysing the data in order to identify
differentially expressed genes. Finally the chapter will conclude with comments
about future directions for the usage of microarrays.

For an overview of the computational methods used for the interpretation of
microarray data, with a particular focus on unsupervised and supervised methods
for clustering microarray data refer to the Tjaden and Cohen chapter in this volume.

2. MICROARRAY PLATFORMS
2.1 Introduction and Generic Features of Microarray Technology

Microarrays are a type of ligand assay based on the same principles as
immunoassays, and Northern and Southern Blots. Immunoassay technology first
started to appear back in the 1960's (Ekins and Chu 1999). DNA was first hybridized
and immobilized onto a solid-phase matrix consisting of plastic and agarose
supports back in the 1980's (Polsky-Cynkin et al. 1985). A research team led by Pat
Brown and Ron Davis at Stanford University are credited with engineering the first
DNA microarray, and in 1995 the same group produced the first modern microarray
analysis publication regarding the use of cDNA glass slide microarrays for obtaining
gene expression profiles (Schena et al. 1995). Stephen Fodor and his colleagues at
Affymetrix are credited with development of the first commercially available short
oligonucleotide microarray wafer chip (Lockhart et al. 1996), the GeneChip™,
making use of photolithography for the in situ-synthesis of nucleotides onto an array
(Fodor et al. 1991). Although the first reported use of microarrays was in Arabidopsis
thaliana (Schena et al. 1995) most of the early microarray research involved
utilization of the technology to identify differentially expressed genes in mammalian
and yeast fields with the first complete genome, S. cerevisiae, being spotted onto an
array in 1997 (DeRisi et al. 1997).

Development and automation of microarray technology, particularly within the
commercial sector, has been primarily allowed through the application and
emergence of advanced technologies such as specialized robotics, fluorescence
detection, photolithography, and image processing equipment and software
(McLachlan et al. 2004). Whether the microarrays, also known as gene expression
arrays, DNA chips, biochips or chips, are commercially- or home-made, cDNA or
oligonucleotide based, they all share a number of generic features in regards to their
underlying technology such as the probe or ‘spot’, the target or sample probe and
the solid-phase medium of the array platform. The probe or ‘spot’ typically refers to
the single stranded polynucleotide (DNA) that is fixed onto the array. Typically the
polynucleotide is of known sequence but may come from a custom unsequenced
cDNA library. The term ‘target’ or ‘sample probe’ is commonly used to refer to the
polynucleotide in the given sample solution which hybridizes to the fixed complementary probe sequence (Kohane et al. 2003). In general there are two basic sources for nucleotide probes on an array; each unique oligonucleotide probe is either individually synthesized base pair by base pair onto the solid array surface (Baldi and Hatfield 2002) or pre-synthesized DNA, cDNA, oligonucleotides or PCR products are directly attached to the array surface. If oligonucleotides are used the probes can be either short, 24-30 bases in length, or long, 60-70 bases in length, while the pre-synthesized PCR and cDNA products are typically hundreds to thousands of base pairs in length (Yauk et al. 2004). The DNA, PCR products or cDNA probes are generally amplified from a vector stored within a bacterial clone or are amplified from open-reading frames or nucleotide fragments of a chromosome (Kohane et al. 2003). Traditionally pre-synthesized probes are attached to the solid array surface using techniques such as robotic spotting or piezoelectricity. Typically the target or sample probe is RNA or cDNA synthesized from the total RNA or mRNA extracted from the biological sample of interest, which for detection purposes is synthesized using fluorescently or biotinylated labeled nucleotides.

Materials commonly used for the solid phase of the array platform include glass or plastic slides similar to a microscope slide, and silica chips. Less commonly used mediums include charged nylon filters, nylon meshes, silicon, nitrocellulose membranes, gels and micro-beads (Kricka and Forina 2001; Baldi and Hatfield 2002). Glass is a good choice of material for the solid phase and it is typically pre-coated with a product such as silicon hydride and poly-lysine or poly-amine to reduce background fluorescence and encourage electrostatic adsorption of the probes onto the slide (Baldi and Hatfield 2002; McLachlan et al. 2004). The solid phase of the array contains a grid with an ordered arrangement of tens to hundreds of thousands of “spots” or “wells” that are each capable of holding a droplet of the probe molecule (Kehoe et al. 1999). If a glass slide is used the spotted or in situ-synthesized probe is typically immobilized by air drying and ultraviolet (UV)-irradiation (Cheung et al. 1999). Each individual spot on the grid generally represents an individual gene, thus serving as an experimental assay for the relative levels of expression for that given gene.

Whether the platform is cDNA or oligonucleotide based, the underlying basis of microarray technology is complementary base pairing between probes and targets. This allows for the determination of the relative levels of mRNA expression of target sequences within the sample via measurement of the quantity of labeled target that binds to each immobilized spot of DNA (Colebatch et al. 2002). In a typical two-colour array experiment one sample (e.g. treatment sample) is labeled with a red Cy5 dye and the second sample (e.g. control sample) is labeled with a green Cy3 dye (Figure 1). After co-hybridization and scanning, the colour of the spot indicates the relative abundance of each sample. If a spot is significantly red it indicates that the gene in the treatment is in abundance, if it is green it indicates RNA from that gene in the control population is in abundance, if the spot is yellow it indicates equal levels of binding and hence no-change in gene expression, while a black spot indicates that binding has not occurred. With one-colour arrays, a single sample is hybridized to each chip, and the relative ratios of absolute spot intensities are compared to establish relative abundances.
2.2 Limitations of Microarray Technology

In regards to what microarrays are able to detect and measure there are a number of limitations that must be considered when planning experiments. DNA microarrays measure levels of mRNA only which do not always correlate well with protein levels. In addition cDNA microarrays are unable to detect and quantify the effects of translational or post-translational modifications on the activity of proteins.

Most probe sequences target the 3' end of mRNA transcripts, which means that they are generally insensitive to different mRNA isoforms, typically being unable to detect the impact of alternative splicing during precursor mRNA processing (Brewster et al. 2004). However this limitation can be overcome by designing probes to specific isoforms.

One must also be aware that whilst a given probe is supposedly on the array, it may not be able to bind its target sequence due to it being a poorly designed probe, or it may in fact cross-hybridize to a different sequence due to homology or contamination of probe source material. Such effects are likely to be platform specific, as different platforms utilize different probe sequences to the same target mRNA.

Microarrays are sensitive platforms, and this sensitivity means that they are often sensitive to nuisance variables such as confluence of cells or date of hybridization. Careful experimental design (what, when and how many samples are hybridized) can ensure one measures true biological signal and good analysis software will ensure the results are reliable and robust against the many noise sources.

Along with the overall cost of microarray experiments some of the more technical limitations include having sufficient quantities of the biological sample of interest to allow extraction of high quality DNA and/or RNA. Such issues will be discussed further in section 6.2.

2.3 Current Microarray Platforms

Microarray technology has, and is continuing to rapidly evolve, resulting in the availability of a number of array platforms with different strengths and weaknesses. The two main types in use are the one-colour systems, in which one hybridizes one sample per microarray, and two-colour systems, in which one competitively hybridizes two labeled samples to a single array.

Generally the spotted cDNA and long oligonucleotide arrays are produced and used with identical or similar methods (Barczak et al. 2003). Spotted array assembly uses a robotic spotter to mechanically pick up small volumes (in the nanoliter or picoliter range) of sequence specific probes onto its pins and then deposit them onto a specific grid address on the glass slide (Figure 1) (Baldi and Hatfield 2002). The size of these spots is dependent on the type of pin and robot that is used. Spot quality is variable and often print run specific. Typically space limitations mean that each clone or oligonucleotide in the library is only printed once on the array but some smaller libraries do permit printing duplicate spots on each slide. If choice is available, well separated duplicate spots are much better than those printed side by side. The main advantage of robotically spotted systems is that they are cheap and
highly customizable. Long oligonucleotide libraries tend to be more specific than cDNA libraries. The main disadvantages of cDNA spotted systems are the
difficulties associated in maintaining and replicating a large library with potential for contamination and unavoidable variance in cDNA concentrations during these processes. Designing sensitive and specific oligonucleotide probes is not easy and often involves a trial and error approach. However, cost considerations often mean that poorly designed oligonucleotides are retained and used for long periods.

This problem is largely overcome by Agilent Technologies (www.agilent.com, Palo Alto, CA) which have developed a modified ink-jet printer head to in situ-synthesize long oligonucleotides (60mers) producing a highly customizable two-colour system with excellent quality spots. The advantages of this system include: the slide quality tends to be higher and more reliable than robotically spotted arrays, and whilst standard layouts are available (e.g. the 44k human genome array), one has complete control over what is printed. The use of 60mer's also increases
specificity between the sample probe and target. The disadvantage is that these systems are more expensive, and there is a limit to the number of spots that may be printed on a single slide (less than 50,000).

The Affymetrix (www.affymetrix.com) GeneChip™ utilizes a sophisticated mask based photo-lithographic technique (developed for silicon chips) combined with solid-phase chemistry to in situ-synthesize 25mer probes at extremely high density (500,000 features per chip). For GeneChip™ arrays, each gene is typically represented by one probe set. Each probe set typically consists of 11 probe pairs which each cover a different 25 base pair section of the target mRNA (older chips used up to 20 probe pairs). Each probe pair consists of a complementary perfect match (PM) 25 mer oligonucleotide, and a mismatch (MM) 25mer oligonucleotide in which the 13th nucleotide in the sequence is changed to its complement, thereby functioning as a non-specific hybridization control (Brewster et al. 2004). However, the effectiveness of the MM oligonucleotide in this role is questionable as it has been shown that many MM's are also sensitive to the true target signal effectively invalidating such a role (Irizarry et al. 2003b). In contrast to the individually assembled spotted arrays, GeneChip™ arrays are produced as a wafer containing between 40 and 400 individual microarrays that are separated after probe synthesis (Kehoe et al. 1999). The main advantage of Affymetrix chips is they produce high quality and highly reproducible chips suitable for single colour hybridizations. Their main disadvantage is that they are very expensive, and not easily customizable.

Illumina’s (www.illumina.com, San Diego, CA) recently developed BeadChip™ technology is a single colour system that uses a fiber optic detection system in conjunction with micro-beads tagged with 50-mer long oligonucleotides. This results in very small feature (spot) size allowing each probe to be represented on average 30 times per chip (Jianbing et al. 2003; Kuhn et al. 2004). The advantages of this system are the high quality, relatively low cost of chips and reagents, and it is highly customizable. The disadvantages are that this is a new system, with poorly developed analysis software, and the high resolution scanner required is expensive.

Other recent and developing technologies include Amersham’s (www.amersham.com) CodeLink™ technology which uses a proprietary 3-D aqueous gel matrix slide surface with 30-base oligonucleotide probes. The 3-D gel matrix provides an aqueous environment that holds the probe away from the surface of the slide, allowing for maximum interaction between probe and target. This results in higher probe specificity and array sensitivity. CombiMatrix (www.combimatrix.com, Mukilteo, WA) and Nanogen (www.nanogen.com, San Diego, CA) utilize electrical addressing systems for the manufacture of DNA arrays onto semiconductor chips. It is believed by researchers at CombiMatrix that through using this technology it will be possible to produce a biological array processor with over 1 000 000 sites per square centimeter (Baldi and Hatfield 2002). CombiMatrix technology allows production of highly customized arrays even for very small orders and several fungal genomes are already available. Lynx (www.lynxgene.com, Hayward, CA) have developed a type of “fluid array” which employs a highly advanced tool developed by Sydney Brenner called massively parallel signature sequencing (MPSS) (Brenner et al. 2000). This platform allows for co-hybridization
of two sample probes, measuring the absolute mRNA levels for virtually every expressed gene within the given samples (Stolovitzky et al. 2005) but at present it is extremely expensive.

2.4 Further Applications of Microarrays

While gene expression profiling experiments are the most common applications of microarray technology, this technology has been developed for numerous other applications such as chromatin immunoprecipitation (ChIP), Tiling arrays, comparative genomic hybridization (CGH) and single nucleotide polymorphism (SNP) arrays, all of which will be discussed below.

ChIP-on-Chip experiments involve hybridizing two independent samples to the array: one being an immunoprecipitated sample containing all the transcription factor-bound DNA and the other a non-specific sample. This allows for rapid and precise mapping of binding sites for transcription factors and other DNA-binding proteins. ChIP-on-Chip experiments also allow for investigation of the activation state of chromatin, chromatin remodeling and functional studies of histone modification as performed by Bernstein et al. (2005) who coupled ChIP with tiling arrays to investigate histone modification patterns in human and mouse cells. Cawley et al. (2004) also combined ChIP with tiling arrays to map the binding sites for three DNA binding transcription factors for human chromosomes 21 and 22.

Tiling arrays contain evenly spaced probes (overlapping or separated) designed to exhaustively span all non-repetitive intronic and exonic (i.e. non-coding and coding regions) sequence from a given genome. Kapranov et al. 2002 reported use of the first human tiling array platform, designed to interrogate on average every 35 bases of the approximately 35 million base pairs of chromosomes 21 and 22. Application of tiling arrays has revealed that a great deal more genomic sequence is transcribed into RNA than can currently be accounted for using present gene annotation data. Such transcripts have been termed TUFs (transcripts of unknown function) and are referred to as the 'hidden transcriptome'. Tiling arrays also allow for identification of alternate spliced isoforms, trans-splicing (where one gene is spliced to the next gene down-stream providing evidence for multiple alternative splice forms), exon-skipping and evidence for non-coding and antisense RNAs as well. Bertone et al. 2005 constructed a series of high-density oligonucleotide tiling arrays representing the entire human genome to comprehensively identify novel transcribed regions.

Similar to the ChIP-on-Chip experiments, Tenenbaum et al. (2000) hybridized purified endogenous mRNA-binding proteins (mRNP) complexes to cDNA arrays to identify subsets of mRNA contained in endogenous messenger ribonucleoprotein complexes (mRNPs) such as ribosomes that are cell type specific. Arava et al. (2003) performed a similar study, being the first group to perform a complete genome-wide analysis of mRNA translation profiles using S. cerevisiae by separating free and ribosome bound mRNA on sucrose gradients then hybridizing these to cDNA arrays containing all known and predicted genes of S. cerevisiae. Interestingly they found that the number of ribosomes associated with mRNAs i.e. ribosome density, decreased with increasing open reading frame (ORF length). The purpose of their study was to carry out a comprehensive and detailed characterization of mRNA association with ribosomes in yeast cells growing in rich medium to probe the
general features of translational behaviour and identify mRNAs that behave distinctively.

CGH (comparative genomic hybridization) arrays are currently the most powerful method for simultaneously detecting and localizing loss or gain of genetic material by allowing for copy number changes within genomes to be assayed through direct hybridization of whole genomic DNA (Mantripragada et al. 2004). By hybridizing entire genomes or specific chromosomal regions of interest, CGH arrays can be used to detect genetic aberrations such as deletions, amplifications, unbalanced translocations and copy number polymorphisms that are often associated with cancer and other complex diseases, and thus mapping of associated breakpoints. CGH can be coupled with tiling arrays to obtain a more complete picture of genome-wide copy number changes. CGH can also be coupled with SNP (single nucleotide polymorphism) arrays.

SNPs are the most frequent form of genetic variation present in the human genome with the SNP Consortium having mapped the presence of over two million SNPs within the human genome (http://www.ncbi.nlm.nih.gov/SNP/). Studies of SNPs offer the possibility to identify disease loci. High-density SNP arrays such as the Affymetrix 10K, 100K or 500K high density SNP mapping arrays provide a high-throughput platform for such studies. SNP arrays have also become valuable tools for loss-of-heterozygosity (LOH) studies and can be coupled with CGH to analyze copy number abnormalities (Zhou et al. 2005). Other applications of SNP arrays include whole genome association, genotyping, genetic linkage analysis, linkage disequilibrium mapping, and genetic epidemiology.

3. APPLICATIONS OF MICROARRAYS WITHIN MYCOLOGY AND BIOTECHNOLOGY

Early microarray experiments focused on the identification of differentially expressed genes in studies involving mainly human and yeast samples. Some of the early yeast microarray experiments involved answering questions regarding the size and diversity of genomes from different yeast strains. Microarray experiments helped identify the discarding of gene encoding DNA fragments from different laboratory yeast strains (Lashkari et al. 1997) and how the gene expression of varying yeast strains changed in response to altered growth conditions (DeRisi et al. 1997). DeRisi et al. (1997) were the first authors to report the use of a microarray containing nearly the whole genome of \textit{S. cerevisiae}. \textit{S. cerevisiae} has been used in a large number of microarray experiments for multiple applications such as investigation of the consequence of loss of gene function (Giaever et al. 2002), the identification of cytoplasmic localized mRNAs (Shepard et al. 2003), functional genomic analysis of a commercial wine strain of \textit{S. cerevisiae} grown under varying nitrogen conditions with high-sugar (Backhus et al. 2001) and for “microarray karyotyping” of several \textit{S. cerevisiae} wine strains to determine the genomic differences between them which may account for some of the observed variations in their fermentation properties (Dunn et al. 2005).

Microarrays can be applied to a wide range of studies including the comparison of disease versus non-diseased tissue, determining the effect of specific gene mutations or gene knockouts within a given cell line or whole organism and also in
evolutionary studies. Microarrays are also rapidly becoming a valuable tool for cancer and viral diagnosis and treatment with the American Food and Drug Administration's making the first approval for use of a microarray as a genetic test in 2004. This microarray, called the AmpliChip® Cytochrome P450 Genotyping Test, was manufactured by Roche Molecular Systems, Inc., of Pleasanton California U.S.A. and was cleared for use with the Affymetrix GeneChip® Microarray Instrumentation System (Affymetrix, Santa Clara, CA), allowing for physicians to individualize drug administration and dosage. In March 2003, a microarray designed to detect a wide range of known viruses and novel viral families was used during a severe acute respiratory syndrome (SARS) outbreak to reveal the presence of a previously uncharacterized coronavirus from a patient sample (Wang et al. 2003). With such applications in mind it is of no surprise that the pharmaceutical companies are increasingly utilizing microarrays throughout the many stages of drug development (Gerhold et al. 2002).

There is also huge potential for using microarrays as diagnostic tools within the field of agriculture. Microarrays have been used for the identification of pathogen's such as individual Fusarium fungi species on cereal grain (Nicolaisen et al. 2005), and for studying the complex signaling that exists between plants and their host-pathogen/symbiont relationships with other organisms. Once a genetically modified organism (GMO) has been generated, microarrays can be used to characterize the effect of the gene modification to ensure that it doesn't result in any undesirable phenotypic effects (Brewster et al. 2004). For investigation of fungal-plant interactions, DNA microarrays have been specifically constructed for examination of symbiotic interactions between Arbuscular mycorrhizal (AM) fungi and legumes (Franken and Requena 2001), between Ectomycorrhizal fungi and eucalyptus trees (Voiblet et al. 2000) and numerous studies examining fungal pathogenic interactions between Alternaria brassicola and A. thaliana (Schenk et al. 2000) and Cochliobolus carbonum and maize (Baldwin et al. 1999). Microarrays have also been used to investigate the virulence of Aspergillus fumigatus, a fungal pathogen of humans (Rementeria et al. 2005) and to investigate how hypoviruses affect fungal development including asexual and sexual sporulation (Allen et al. 2003).

4. EXPERIMENTAL DESIGN OF MICROARRAY EXPERIMENTS
4.1 General Experimental Design of Microarrays

The standard statistical experimental design concepts of control, randomization, and replication apply equally well to microarray experiments. Before embarking on an experiment one must decide what biological questions one seeks to answer. On this basis one can chose a suitable probe library (this will define the chip type), and what samples or pairs of samples, will be hybridized against this probe library. One must also consider the feasibility of randomizing sample collection and date of hybridization to avoid confounding results. It is also important to consider how much replication is possible. Finally, thought should be given to how the data will be analysed. Typically many of these choices are driven by cost, but some forethought can help refine and prioritize the biological aims of the experiment thus maximizing the information gained from an experiment.
Microarray experimental design is largely concerned with choosing the hybridization strategy. However, before looking at this in detail we should first review the basic steps in a microarray experiment. A typical microarray experiment begins with the extraction of mRNA or total RNA from a specific biological sample of interest followed by synthesis of fluorescently labeled cDNA target (Figure 1). It is important that during the extraction of RNA, all traces of genomic DNA are removed in order to keep background levels low as it is a common source of contamination. The isolated mRNA or total RNA is typically reverse transcribed by first-strand cDNA synthesis and labeled for detection during the scanning process. For two-coloured microarrays the most commonly used fluorescent probes are the cyanine dyes Cy3 (green) and Cy5 (red) whilst one-colour arrays typically use biotin/streptavidin conjugated probes for labeling. The labeled target is then denatured by heating to obtain single stranded polynucleotides from the sample, which upon cooling will hybridize to its complementary probe fixed on the array. In order to promote the specific and complementary binding of the labeled sample to its probe while reducing the level of background noise, it is of critical importance that the hybridization conditions are optimized (Brewster et al. 2004). Following hybridization, the array is washed to remove any unhybridized sample probes, and the amount of target hybridized to each spot is then quantified by scanning and image processing (see section 5). Microarray experiments thus include a large number of technical steps, and the resultant level of background noise can be influenced at many points and can depend upon the skills of the technicians performing extraction, labeling and hybridization. It is important to be aware of how the data is generated, so that it may be appropriately analysed, but we will not discuss this further instead concentrating on issues related to choosing what to hybridize.

As mentioned earlier the most important design step is deciding the aims of the experiment, and prioritizing what comparisons are of most interest. This will guide subsequent choices on platform type, RNA sources, and hybridization strategy. Many of these design choices are inter-related. Two-colour spotted arrays are cheap, but typically more noisy than the one- or two-colour commercial arrays systems (Irizarry et al. 2005). One also desires to perform sufficient biological replication to build confidence in the results. Using a cheaper system allows for the use of more replication or time points in a time course study, however the gain may not be huge due to increased noise of such systems. Sometimes the availability of a specific probe library will make a single platform the obvious choice, and similarly the desire for a specific comparison may make one design standout.

When deciding between whether to use a one- or two-colour system the choice is generally made on the basis of the aims and complexity of the experiment, and the desire to link the experiment to other experiments to be performed at a later date, or in other laboratories. Two-colour microarrays are analogous to matched pairs experiments - by co-hybridizing samples we control for many array and hybridization specific variables. To achieve similar power, one-colour arrays must be technically more stable and reproducible. With Affymetrix, the production methods and robotic control of hybridization ensures that this is generally the case (Irizarry et al. 2005). In general two-colour competitive hybridization is good for small-scale
experiments, but as the scale increases, one faces problems in choosing which pairs of samples to co-hybridize, shifting the balance towards one-colour systems. Also if hybridizations are to be conducted over long time scales then one-colour systems may be more appropriate than using a two-colour system with a common reference design (discussed later).

4.2 Replicate Experiments, Reproducibility and Randomization

A typical question concerning researchers contemplating a microarray experiment is "How many replicates are required?". A typical response is "As much as one can afford". Confidence in results is based on their reliability which is derived from performing replicated experiments using different biological samples thus increasing the number of degrees of freedom. Biological replicates must be obtained from different biological samples involving separate RNA extractions for each to ensure an adequate measure of biological variability. The amount of variability between biological replicates will depend upon whether the material is derived from cell-lines, in-bred or out-bred species or strains (which are most variable).

One also has the option of performing technical replication, where technical replicates involve RNA that has been extracted from the same biological sample but has been independently labeled for hybridization. The need for technical replication is related to the level of variability inherent in the microarray platform being used. Dye swap replicates are technical replicates in which the dye labels are reversed e.g. Sample A is labeled with Cy3 in the first slide, and Cy5 in the second, whilst Sample B is labeled with Cy5 in the first slide, and Cy3 in the second. Dye swap replicates can be used in most two colour designs, to reduce any possible effects due to differential dye responses which are not completely removed by normalization procedures (discussed in section 6). One related approach is to perform dye swapping on biological replicates. In the opinion of Glonek and Solomon (2004), if hybridizations are to be replicated then they should be performed as dye-swapped replicates.

It is important for researchers to realize that under no circumstance does technical replication account for biological variability, it purely provides an estimated measure of the level of experimental variability (Kerr 2003) while increasing the likelihood of detecting differentially expressed genes.

Instead of providing adequate biological replicates, some researchers will pool samples with the belief that pooling will provide a means of reducing the biological variability and the number of arrays required for the experiment. However pooling generally doesn’t provide a valid basis for adequate statistical analysis of the resulting data set. While pooling does lead to a reduction in observed biological variance it also results in the elimination of all independent biological replication making it impossible to compare individual samples from which the pools were derived (Simon and Dobbin 2003). However, pooling is sometimes necessary when there are insufficient quantities of RNA from an individual sample.

Another important consideration is randomization. For any experiment in which there is a treatment the biological samples should be randomly assigned to the treatment groups. In the opinion of Kerr (2003), if the microarray data is particularly susceptible to technical variation then arrays should also be chosen randomly from
the batch of arrays for each planned hybridization to remove any possible systematic variation that may be related to the order in which the arrays were printed.

There have been several comparative cross-platform studies performed which indicate that in general biological variability tends to be higher than technical variability, and that commercial platforms tend to have less technical variability than in-house printed arrays (Yauk et al. 2004; Irizarry et al. 2005).

After determining the overall aim of the microarray experiment, i.e. what biological question is to be answered, one of the first steps involved with the design is selection of the functionally relevant biological sample, whether it is a cell type, tissue or whole organism such as fungi. Treatments and conditions relating to growth and isolation of the samples need to be identified, performed and kept tightly constant across all specimens in order to minimize biological variation arising from environment. Kazan et al. (2001) suggest that in order to gain the best possible comparisons, a separate control accounting for differences imposed by treatments must be used for each treatment in the experiment. Once the biological sample is selected and control and treatment groups obtained then the next stage in the process is generation of the sample which requires the isolation of either total RNA or mRNA. If RNA becomes degraded during the experimental process then it will be unsuitable for labeling by most standard techniques. Some manufacturers do offer alternative protocols when RNA degradation cannot be avoided. To help determine the integrity and quality of the sample the RNA is typically examined by “denaturing” gel electrophoresis. The presence of RNA degradation in Affymetrix arrays can also be detected in-silico using RNA degradation plots and NUSE boxplots which are part of the affy and affyPLM packages of the freely available Bioconductor analysis package (Gentleman et al. 2004).

4.3 Models and Assumptions of Experimental Design

There are several models currently used for the design of microarray experiments. These models cover issues concerning the labeling and allocation of arrays and the order of sample probe hybridization to the arrays. The choice of experimental design is dictated by the biological questions being asked, availability of microarray platforms, suitability of analysis software and constraints related to amounts of RNA and financial considerations. At present the most commonly used models are the reference design, the balanced block design, the loop design, the dual-label or dye-swap design (only applicable to two-colour arrays) and the time course and factorial designs (Figure 2). Each of these will be briefly discussed below.

The reference design (Figure 2B) involves a common reference being used for comparisons with treatment effects across a number of different experiments. To be useful, the reference must contain detectable levels of all genes expressed in samples co-hybridized with it, therefore the reference is often a pooled sample of all experimental conditions. This design often makes the assumption that the effect of dye bias is equal across all comparisons with the reference (Maindonald et al. 2003) as the reference is generally labeled with the same dye on each array. Therefore any gene-specific dye bias not removed by normalization will affect all arrays in a similar fashion (Simon and Dobbin 2003). One reason for using this design is when there is limited availability of RNA from one or all of the samples. The orientation of dye
labeling is applied in the same direction so that samples for comparison are always labeled with the same dye and the reference is always labeled with the same dye (Churchill 2002). For spotted arrays, this design uses an aliquot of common reference RNA as one of the samples co-hybridized to each array so that comparisons between the reference and sample of interest can occur on the same spot (Simon and Dobbin 2003). In the opinion of Maindonald et al. (2003) a direct pair-wise comparison for two treatments should be more precise than indirect comparison of all treatments through a reference. Similarly, Churchill (2002) argues that use of a reference sample is unnecessary and results in inefficient experiments as half of the gene-expression measurements are made on the reference sample which is generally of no interest to the researcher. Reference designs can still be appropriate in complex experiments where each sample is involved in several comparisons.

The direct pair-wise comparison (Figure 2A) preferred by Maindonald et al. (2003) is referred to as the balanced block design. The major advantage of this design is that it can limit the number of arrays per experiment thus reducing the overall cost. A disadvantage of this design is that it generally has a higher level of signal-to-noise ratio then the reference design due to the variation of spots between arrays and within arrays. Simon and Dobbin (2003) report that the efficiency of the block design is reduced when there is increased biological sample variation.

When cluster analysis of the resulting data set is the main objective of the experiment then a particular useful experimental model is the loop design (Figure 2C). This model typically involves the co-hybridization of two differing samples onto a single array with the aliquot for each sample being split between two arrays allowing for the arrays to ultimately be used for linking each of the samples together in a loop pattern therefore allowing for all pair-wise comparisons to be made between samples while controlling the size and variability of spots (Simon and Dobbin 2003). In comparison to the reference design, the loop design is more balanced with respect to the dyes as each sample is labeled at least once with each of the dyes used (Kerr and Churchill 2001). A major down-side to this design however is the increased variance due to the requirement for modeling indirect effects relating to the arrays that link two samples of interest (Simon and Dobbin 2003). The loop design is also generally less robust in respect to the occurrence of bad arrays as one or more bad arrays result in breaking of the loop whereas in the reference and block designs they can simple be removed from subsequent data analysis.

Large time-course experiments (Figure 2E-F) are those where samples are collected and measured at many different time points usually in response to a treatment and comparisons are then made between the time points (i.e. between 0, 6 and 12 hrs). A crucial factor for the validity of time-course designs is the actual times used and the overall number of time points. If poorly designed these experiments will become costly in terms of equipment and consumables and it is generally impossible to perform pair-wise comparisons on all samples.

The previous designs are examples of single factor designs in which we study different levels of a specific factor. Often we will be interested in investigating several factors at once, such as comparing several different treatments over time. Such designs are known as factorial designs (Figure 2D), and allow investigating
each effect plus the presence of interactions between factors (e.g. which genes would show changes in slope if different treatments were plotted over time). In conjunction with analysis utilizing linear models, such as methods implemented in Limma (see section 6.6) factorial designs allow highly effective identification of genes that for example respond to stimulation differently in test and control groups. Further

Fig. 2. Graphical representation of experimental designs. Boxes represent hybridized arrays and lines or arrows represent comparisons between samples. For two-coloured arrays, by convention arrow tails represent the Cy3 (green) labeled sample while arrow heads represent the Cy5 (red) labeled sample. (A) allows for direct pair-wise comparison of all genes between two biological samples. (B) allows for indirect comparisons to be made between samples and a common reference. (C) represents a loop design. (D) displays a factorial design where three samples taken at 0 hours are being compared with three samples taken at 1 hour. Such designs can become much more complex with increasing contrasts of interest. When this occurs it is more useful to use a reference design. (E) and (F) represent varying types of time-course designs. All designs can involve the use of dye-swaps.
considerations on how to optimally choose hybridizations when comparing several factors such as different cell lines and treatments are discussed by Glonek and Solomon (2004).

5. SCANNING OF MICROARRAYS AND IMAGE PROCESSING

After hybridization the gene expression data must be extracted from the microarray and analysed. This requires scanning of the array to measure the fluorescent signal for each spot on the array followed by analysis of the resulting image to extract the foreground and background intensity values that are used in subsequent analysis. One typically has little control over scanning equipment hence microarray scanners will not be discussed here. An area where choice is more readily available is the image processing software. Image analysis and the resulting acquisition of data is an important aspect of microarray experiments and can potentially have a large impact on subsequent data analysis. The commercial platforms (e.g. Affymetrix and Agilent) tend to have their own customized image analysis software packages, which leave little choice to the end user. However, in-house spotted arrays are typically more variable, requiring more care in the choice of image analysis software to ensure unnecessary variation is not introduced. Due to the wide variety of microarray platforms and various forms of labeling, no single microarray scanning device or image analysis software is suitable for all purposes. In this section we will only address the scanning and processing of images from two- and one-colour microarrays that emit a fluorescent signal.

Currently, there is a wide range of both commercial and freeware image analysis software available. Some of these are designed specifically for glass slide arrays while others can be used in conjunction with a variety of array platforms such as the nylon filter arrays. Regardless of platform, all image analysis software is generally designed to perform three fundamental processes: Addressing or gridding, segmentation, and intensity extraction or data acquisition. Addressing or gridding is the process of locating each spot on the slide and assigning it to a coordinate by taking advantage of the rigid layout of the spots. Segmentation is the method used to differentiate and classify the foreground pixels for each spot and background pixels. Information extraction or data acquisition is the process of calculating the foreground and background intensities for each spot based on the assignment of pixels during segmentation.

During the image processing stage a number of problems can arise which can result from insufficient labeling and concentration of the sample probes or too little of the probe on the solid-phase for binding of the sample probe as well as insufficient exposure time (Cheung et al. 1999). Other problems specifically relate to the presence of poor-quality spots which can have a drastic impact on the data set if not reduced (McLachlan et al. 2004). Spots are classified as poor quality when they have variable diameters and contours, a background signal that is higher than the foreground signal or the presence of spatial artifacts (McLachlan et al. 2004). A good image analysis program should therefore have the capability of collecting quality measures for each spot that can be used to flag unreliable spots or arrays. However, in general flagged spots should be down-weighted in analysis, rather than completely eliminated (Ritchie 2004).
5.1 Scanning Microarray Images

All microarrays that have been hybridized with a fluorescently labeled target use an optical system to scan slides to produce a digital record. This record contains the fluorescence intensity for every pixel at each grid location on the array. The intensity is proportional to the number of sample probes hybridized to the spotted probe (Cheung et al. 1999). Commonly used scanners are typically based on a confocal laser microscopy system where a separate laser is used as a source of excitation light for each fluorescent dye, and a photomultiplier (PMT) tube is used as the detector. In brief the fluorescent dyes become ‘excited’ by the laser light, absorbing its energy, and resulting in the emission of photons (Cy3 dyes produce a band from 510-550nm, whilst Cy5 dyes produce a band in the 630-660nm range). A detector such as the PMT is scanned across the surface and measures the intensity at each point (Baldi and Hatfield 2002; Yang et al. 2002a). Depending on the scanner used a number of settings such as the power of the excitation laser and the voltage of the PMT can be varied to improve the sensitivity of image acquisition so that low signals can be detected.

Images generated from microarrays contain the raw data of the experiment. Typically a 16-bit grey scale image is produced for each fluorescence frequency. For two-colour systems these can be combined into a single falsely coloured red-yellow-green image.

5.2 Addressing or Gridding

The scanned microarray image is imported into an image analysis program, where the first stage of analysis is to locate each spot on the slide by addressing or gridding of the array. The majority of image analysis software systems now provide reasonable automatic or semi-automatic gridding procedures with slight variations occurring between each. Aside from the comment that the results of automated gridding should be visually scanned to ensure the accuracy of the process, addressing and gridding will not be discussed here in detail; however this has been reviewed recently by Smyth et al. (2003).

5.3 Segmentation (Identifying Fore and Background pixels)

Segmentation of a microarray image is the process of dividing the image into different regions based on certain properties. For spotted arrays it involves the classification of pixels as being foreground or background (Yang et al. 2001). In regards to two-colour arrays, there is presently a number of differing segmentation methods used for production of the spot mask. It has been shown that the choice of segmentation method can introduce variability into the resulting microarray data, hence care must be taken when selecting the method for use (Ahmed et al. 2004; Ritchie 2004).

The four main groups of segmentation methods used for two-colour microarray images are: fixed circle segmentation; adaptive circle segmentation; adaptive shape segmentation; and histogram segmentation (Yang et al. 2001). Given that spots on many in house printed arrays are often irregularly shaped, fixed or adaptive circle systems should be avoided (Yang et al. 2001). Histogram segmentation uses a
thresholding system for classifying pixels as either foreground or background (Chen et al. 1997), but often results in the over- and under-estimation of foreground and background intensities (Smyth et al. 2003). Adaptive shape segmentation methods such as the watershed (Beucher and Meyer 1993) and seeded region growing (Adams and Bischof 1994) implemented within the Spot software make no assumptions in regards to the spots circularity and size and hence are suitable for use with both commercial and non-commercially produced arrays.

5.4 Intensity Extraction or Data Acquisition

Determining spot intensity requires computing the average pixel value of the foreground pixels of a spot. Background intensity needs to then be approximated using a suitable method and subtracted from the spot intensity to provide a foreground fluorescent intensity. For two-colour arrays foreground and background intensity needs to be calculated for both the Cy3 (green) and Cy5 (red) channels. Correction for background intensity is necessary as it is likely that not all of the measured spot intensity comes from the fluorescent label of the hybridized sample. Signals can also result from factors such as non-specific hybridization, from fabrication artifacts on the glass caused by chemicals, dust and spatial variation. If the method for background estimation is poorly chosen, correction of foreground can result in negative intensities and hence missing values when log intensities are computed, typically resulting in the loss of low-intensity data (Smyth et al. 2003). Poor choice of background correctors can also introduce extra noise (variability) making detection of true differential expression more difficult (Yang et al. 2002a; Ritchie 2004). For both one-colour and two-colour arrays the reported intensity level (spot, background, PM or MM) is a summary of fluorescence measurements detected in a series of pixels.

As for segmentation, there a number of differing methods implemented in the software for calculation of background intensity. Generally however, these methods can be classified on the basis of whether they use a constant or global value, a local background estimate or a morphological background estimated by applying a non-linear filter to a local window around the spot (Yang et al. 2002a). Local background methods only consider the intensities of small regions surrounding the spot mask while global background generally considers the intensity of background for the whole array. In the experience of Smyth et al. (2003) local background methods result in over estimation of the background while global methods can result in an under-estimation of the background. Morphological opening tends to give a less variable background estimate which is not upwardly biased by the presence of bright pixels (Yang et al. 2002a; Ritchie 2004). Ritchie (2004) has performed an in-depth study of different background estimators and advises that if the purpose of the experiment is to select differentially expressed genes then the use of a morphological opening background estimator is recommended. Morphological opening has been available in the Spot (http://www.cmis.csiro.au/index.htm) software package for several years, and more recently has been included in GenePix (http://www.axon.com) and InaGene’s (http://www.biodiscovery.com) offerings. If one only has the choice of a local background corrector (e.g. an older version of GenePix implementing median scale normalization) or no background correction,
Ritchie (2004) advises that more reliable (less variable) results are obtained with no background correction.

After background correction, data analysis software for the two colour-arrays calculates the log-differential expression ratio being \( M = \log_2 \frac{R}{G} \) for each spot and the log-intensity of the spot will being \( A = \frac{1}{2} \log_2 RG \), which is a measure of the overall brightness of the spot.

6. ANALYSIS OF MICROARRAY DATA

Microarray data analysis techniques have rapidly evolved, currently there are a wide variety of methods available to identify differentially expressed genes and infer functional information. These include analysing single genes to investigate how the behaviour of each gene changes between a control and a treatment or multiple gene analysis where clusters of genes are analyzed to determine common functionality, pattern-identification, gene-gene interactions and gene regulatory networks. Cluster analysis of microarray data is covered in the proceeding Tjaden and Cohen chapter, and therefore won't be discussed further. Overall success in identifying differentially expressed genes during the microarray data analysis stage is heavily dependent on the suitability of the chosen experimental design, which also governs what type of analysis to use and this was discussed previously in section 4.3.

Regardless of platform, the raw microarray data acquired from an image needs to be processed in order to remove poor quality spots and then normalized to correct for systematic variation before further downstream analysis. Downstream analysis involves identifying genes that are differentially expressed. This requires the selection and calculation of a suitable statistic to be used for ranking of the genes, followed by selection of an appropriate cut-off point for differential expression whereby genes having a rank value above the cut-off are considered to be differentially expressed and those having a value below are considered not to be. Within this process it is often common to calculate the false-discovery rate, in order to determine the number of expected false-positives and false-negatives that are to be included or excluded from the final list of differentially expressed genes. More precisely false-positives are genes having no differential expression that appear within the final list of differentially expressed genes while false-negatives are genes having true differential expression that are excluded from the final list.

Before normalization and further analysis the raw microarray data usually undergoes a logarithmic transformation to the base 2. Transformation of the raw data can help minimize some of the systematic variation by eliminating measurements for poor quality spots, and may facilitate in the identification of differentially expressed genes. In the case of two-colour arrays, the logarithmic transformation converts the intensity ratios into differences between the two channels at each spot, making up-regulated and down-regulated values of the same scale comparable as the non-transformed ratios tend to treat up- and down-regulated genes differently. There are a number of variations to the standard logarithmic transformation such as shift transformations (Newton et al. 2001; Kerr et al. 2002); curve fitting transformations (Yang et al. 2002b) and variance stabilizing transformation methods (Rocke and Durbin 2001; Cui et al. 2003).
The final stage of microarray data analysis is biological interpretation of the results to determine their functional relevance, and confirmation of observed differential expression by other experimental means. Biological interpretation of the data requires the utilization of additional bioinformatics techniques for the correlation of expression data with other types of data such as genomic, proteomic or metabolomic data.

6.1 Sources of Experimental and Biological Variation

Poor quality or noise within the microarray data arises from the many sources of variation throughout the experimental process. If not removed, the noise will ultimately affect the observed levels of differential expression. Throughout the preceding sections of this chapter some of the sources of experimental variation have already been mentioned however they will be re-iterated here.

Variation arising during the early stages of the experimental process may occur due to the use of poor quality RNA, RNA degradation, and the presence of genomic DNA within the RNA sample. A number of variations can also arise during both the labeling and hybridization stages. If the conditions of hybridization, such as temperature and duration, are not optimized and kept constant, further variations will arise during the stages of scanning and image processing and there will be a higher occurrence of non-specific hybridization of the samples to the probes. Non-specific hybridization and the presence of foreign artifacts on the arrays such as dust, clothing fibers, skin and scratching of the slides is also a significant problem for both of the common array platforms. During the labeling process non-controllable sequence bias can occur, resulting from some of the fluorescent dye labels showing preferential binding to some nucleotides over others. Another variation specific to spotted arrays relates to the lengths of the probes. As the probe length can vary from a hundred to a thousand bases in length, there is a higher likelihood of the probes cross-linking which results in a decreased number of available probe molecules for the binding of sample probes. Also, the actual process of robotically spotting the probes onto the array can introduce a great deal of systematic variation due to inconsistencies occurring with location, size and shape of spots on the individual arrays and between the arrays. These variations typically result from there being slight differences between the print-tips on the robotic arrayer, or from using arrays that were generated at different times. One of the most common forms of bias affecting spotted microarrays is that of dye-bias also known as red-green bias. Dye-bias arises due to there being differences between the labeling efficiencies and scanning properties for the two fluorescent dyes, Cy3 and Cy5. For any array platform, there is often a major problem with saturation of the probes, occurring when the probe intensities reach the maximum level of intensity acquired by the scanner. Saturation can result in loss of information regarding differential expression by masking highly expressed genes. Saturation effects can be minimized during the scanning process by adjusting the settings of the scanner, however, this may result in the exclusion of low expressed genes hence, normalization of the data may be more desirable and will be discussed below.
6.2 Normalization of Microarray Data

One would like to remove or minimize any non-biological variation present. This process is generically referred to as normalization. Normalization is platform specific, with different approaches required for one- and two-colour systems.

A wide range of normalization algorithms are available such as local versus global and linear versus nonlinear normalization. Selection of a suitable algorithm requires some assessment to be made in regards to what type and degree of systematic variation is present and whether normalization is required within-arrays, between-arrays or both. It is important that the normalization process does not remove or reduce any variation arising from biological differences between RNA samples or the printed probes. While normalization is generally considered necessary, over normalizing the data can introduce biases that are more detrimental to the identification of differentially expressed genes than a small difference in scale. Selection of a suitable normalization method can be aided by viewing exploratory plots such as M vs. A plots (Figure 3) to investigate if there is any obvious curvature deviating from the horizontal line at zero, or boxplots of each array to determine the difference in spread of log-ratios for each array, or of print-tip groups for each individual array in the case of spotted arrays. The majority of normalization methods aim to scale individual intensities so that the mean or median intensities are balanced within and between arrays, allowing for meaningful comparisons to be made.

Many of the normalization methods make the assumption that the majority of genes are not differentially expressed, hence the average, or geometric mean, of the ratio is one and that the average, or arithmetic mean, of the log ratio is zero i.e. it is assumed that the expression level for the average gene does not change during the experimental conditions. Normalizing all or the majority of genes present on the array generally provides the most reliable and stable estimation of spatial and intensity dependent trends present within the data (Smyth et al. 2003). However, at times it is more useful to normalize a subset of genes back to a set of control genes or a set of housekeeping genes present on the array.

6.2.1 Normalization of gene chip (single colour) arrays

One-colour arrays such as the Affymetrix GeneChip allow for hybridization of only a single sample to each individual array. Appropriate normalization is therefore vital, as different arrays need to be compared against each other in order to determine a meaningful estimate of the level of differential expression in a given gene. In the following discussion we will also consider probe set summarization methods, as some normalization methods are applied to probeset summaries, whilst others are applied at the level of individual probes.

For several years Affymetrix have supplied their Microarray Analysis Suite version 5 (MAS5) for probeset summarization and array normalization. Probe set summarization is performed by using a Tukey biweight estimator to robustly obtain the average difference between the PM and modified MM signals for each probe pair in the probe set. If PM > MM, then the modified value is just the MM, but if PM < MM, the MM is modified to ensure the difference is always positive. This approach is used as the MM is supposed to measure non-specific signal binding to the PM (i.e.
background signal), and thus should always be less than or equal to MM, however in practice this is often not the case.

Fig. 3. Exploratory M vs. A plots are used to view the effect of normalization on two-colour microarray data. MA plots are a rotated plot of the deviation from red to green allowing easy visual identification of intensity dependent effects. A) displays a raw MA plot, demonstrating the need for normalization. B) displays the plot after global normalization as performed by the GenePix software. This normalization is generally not recommended as it simply shifts the median M values to zero and doesn't remove intensity dependent effects which are extremely common. C) displays the plot after performing the recommended print-tip intensity dependent loess normalization. This normalization approach involves fitting individual loess lines through each of the print tip groups on the array to bring the mean M in all print tip groups to zero.
Li and Wong (2001) noted that variation between probes in a probe set was often substantially (5 times or more) than the variation in values for a given probe across arrays. This massive within probe set variance thus advocates treatment of probe specific effects when trying to summarize probesets and Li and Wong (2001) advocated a multiplicative model in their dChip software. Irizarry et al. (2003a,b) used a series of experiments spiking in RNA of known concentrations to study probe sets effects and further noticed that many MM's showed concentration dependent effects (that is they were often sensitive to true signal in addition to any non specific background signal) and concluded that a more appropriate approach was a additive model, which they used in their highly successful RMA approach. RMA performs a background correction and normalization step on individual probes, before obtaining a probeset summary. The probeset summary comes from an additive model where the observed intensity in a probe is modeled as the sum of the true probeset expression value, a probe affinity term, and an error term. RMA then uses the estimate probeset expression value as its measure. Before considering the RMA approach we will examine the normalization system used in MASS, which is based upon the summarized average difference values for each probe set.

The normalization approach used by Affymetrix in MASS is to scale intensities so that each array has the same average value. A reference array is defined (typically a control sample), and ratio of intensity in the reference to a given array for each probeset is obtained. The scaling factor which is applied to this array is the trimmed mean of these ratios. This method is obviously highly dependent upon the choice of a good reference array and does not perform well if there are non-linear relationships between arrays. To rectify the problem of non-linearity Schadt et al. (2001) and Li and Wong (2001) both propose normalization methods that make use of non-linear smooth curves, fitting a non-linear regression of the baseline array values onto the experimental array values. However these approaches also depend upon the choice of a suitable reference array.

Whilst developing the RMA method, Bolstad et al. (2003) considered both existing, and several new approaches to normalization. In particular they considered complete data methods, in which all available data (rather than pairs of arrays) are used to determine the normalization. Firstly they utilize a background correction method in which they estimate the observed PM signal as being due to true signal plus a noise signal (due to non-specific hybridization and optical noise). The observed distribution of all PM values on the log2 scale has a log normal form (normal + exponential decay) from which appropriate background values may be estimated.

Once data is background corrected the quantile normalization method analyses intensity distributions by performing pairwise comparisons of quantile-quantile plots for multiple arrays. Assuming that there is an underlying common distribution of intensities across arrays, the method then aims to give each array the same intensity distribution by taking the mean quantile and substituting it as the value of the data item in the original dataset (Bolstad et al. 2003). Bolstad et al. (2003) found that quantile normalization was able to reduce the variation of a probe set measure across multiple arrays to a greater degree than the Affymetrix scaling method and the non-linear method by Schadt et al. (2001). Specifically they found that
The performance of the quantile method was most favorable in terms of speed as well as bias and variability measures and thus it is the recommended normalization method for high-density oligonucleotide arrays.

The RMA method for normalization and probe set summary provides an approximate 5 fold reduction in variance compared to the MAS5 approach giving a massive increase in sensitivity allowing the detection of true differential expression. However, a slight bias tradeoff is made, in that RMA compresses fold change estimates by 10-20% compared to MAS5. This fold change compression has more recently been addressed in a updated version of RMA known as GCRMA (Wu and Irizarry 2004). This approach uses sequence specific models for background estimation, resulting in similar estimates of true signal with MAS5, whilst retaining the low variance. Thus for probe set summarization and array normalization we would highly advise using RMA or GCRMA.

Finally we should note that Affymetrix have recently updated their analysis algorithms, dropping their MAS5 approach and utilizing the PLIER algorithm. Whilst few details are available, the PLIER algorithm is a model based approach broadly similar to RMA, and thus represents a substantial improvement over MAS5.

6.2.2 Normalization of two-colour spotted arrays

A major source of variation affecting the analysis of two-colour microarray data is that arising from dye bias. Normalization methods therefore need to minimize this bias by balancing the fluorescence intensities of the green (Cy3) and red (Cy5) dyes, as dye bias and other variations can result in a shift in the average ratio of the Cy3 and Cy5 channels, thus the intensities may also need to be rescaled. To reveal the extent of dye bias it is useful to view M vs. A plots for each array (Figure 3). Intensity-dependent normalization may not be the only type of normalization required. Yang et al. (2002b) address three main forms of normalization being: within slide normalization; paired-slide normalization for dye-swap experiments and between-slide normalization.

Within-slide normalization methods, i.e. those that are applied to a single array, can be carried out by performing a form of location and intensity dependent normalization for each individual slide and one of the several forms of global normalization. As with the one-colour arrays, global normalization aims to correct the log-ratio values by subtracting a constant value, typically estimated from the mean or median M-values of a subset of genes whose expression is expected to remain constant. For two-colour arrays, global normalization makes the assumption that the red and green intensities can be related by a constant factor, with the aim being to shift the log-ratios to zero (Figure 3B). Sadly, despite there being evidence of spatial or intensity dependent dye biases in most experiments, global normalization methods are still generally the most widely used despite their inability to correct such types of variation (Yang et al. 2002b).

In general a more appropriate technique is the robust intensity-dependent Loess normalization method (Figure 3C) (Yang et al. 2002b). This method assumes the deviation from an M value of 0 varies in an intensity dependent way (i.e. over the range of A values observed. This is most obvious in MA plots where one observes curvature in the raw data (Figure 3A). At each value of A, a robust locally weighted
regression line is obtained to robustly locate the central M value of the points. This value is then subtracted from all points at this value of A, thus shifting the central cluster of points back to the zero line (Figure 3C). Outliers (which in this case are differentially expressed genes) do not influence this calculation. Intensity-dependent Loess normalization may be applied globally over an array, or individually to each print tip group on an array.

This latter case may be necessary due to slight variations within the print-tips, the robotic spotter tip length or opening may vary during the array assembly process leading to spatial variation across the slide. Print tip intensity dependent loess also performs a de-facto spatial normalization as well, although this may occasionally perform poorly if there is strong intensity gradient within the print tip group (perhaps due to a local hybridization artifact). It is also possible to use spot quality weights in these methods, so poor quality spots do not influence the normalization procedure. In general it is recommended that print-tip intensity dependent loess is used as the default normalization method for two-colour microarrays. Occasionally more specialized forms of normalization are required such as 2D spatial normalization (Cui et al. 2003).

When dealing with replicate experiments, the relative gene expression levels may have different spreads in their log-ratios due to differences in experimental conditions. If significant, an adjustment of scale in which M-values from a series of arrays are scaled so that each array has the same median absolute deviation will be required to balance out the relative expression levels between experiments and hence between arrays (Yang et al. 2002b). After within-slide normalization, all normalized log-ratios will be centered around zero, regardless of the normalization method i.e. whether lowess or non-linear, however it is useful to examine boxplots displaying the spread of log-ratios for individual arrays to determine if scaling is required between arrays (Smyth et al. 2003). Failing to perform scale-normalization could lead to one or more slides having undue weight when averaging log-ratios across experiments to an average of log-ratios across slides. One common method of scale normalization is to divide each intensity by the total of the intensities on the slide, so that all slides then have the same total intensity. Another lowess normalization performed in a similar manner to that for GeneChip involves fitting a robust regression line through the M vs. A plot instead of a lowess curve.

An alternative form of normalization for two-colour arrays is the single-channel normalization method proposed by Yang and Thorne (2003) which allows for meaningful information to be individually obtained from the Cy3 and Cy5 channels of two-colour microarrays. This method removes systematic intensity bias that is not due to real gene expression separately from each channel both within and between arrays. Ultimately, single channel analysis allows for comparisons of absolute intensities between separate arrays for which no direct comparisons have been made. The cost is that single channel data from two-colour systems is considerably more noisy, requiring roughly four times the number of arrays that would be needed had a direct comparison been made.
6.3 Identifying Differentially Expressed Genes

A common aim of microarrays is to reliably identify differentially expressed genes between two conditions. This can be quantified by a t-test, which is simply a measure of the mean difference between conditions (mean M value), divided by the standard error in this difference (standard error in M = standard deviation/square root of n). One obtains significant t values if the absolute value of the ratio is large, as this implies that the observed mean difference is much larger than the variance in its measurement.

Genes identified as being differentially expressed are those that display a significant change in their expression between two samples of interest. Identification of differentially expressed genes within the normalized data set requires two steps; firstly selection and calculation of a suitable statistic for ranking the genes in order of evidence for differential expression from strongest to weakest and secondly selection of a suitable cut-off value for the ranking statistic where any gene having a value falling above the cut-off is considered to be differentially expressed. Although relatively simple in principle, in reality identifying differentially expressed genes is actually quite a complex problem due to the measured intensity values being affected by numerous sources of fluctuation and noise (Draghici et al., 2003).

A common, however flawed method, for ranking the genes is to simply consider the average fold change, or M values for each gene. Use of M values as the ranking statistic is a poor choice as it ignores any variability between replicates and there is no means by which to calculate the level of confidence you can have in regards to the supposed differential expression. Using simple fold change cut-offs can lead to an increased number of false-positives and false-negatives (Cui and Churchill, 2003).

Commonly used statistical methods that can be used to rank genes from replicated data are the Student's t-test and its many variations, one-way analysis of variance (ANOVA), empirical Bayes analysis and the Wilcoxon (or Mann-Whitney) test. A simulated comparative study by Troyanskaya et al. (2002) found that both the t-test and the Wilcoxon test resulted in a low number of false-positives while successfully identifying a large number of the differentially expressed genes. The Student's t-test, or simply t-test, is one of the simplest and most common methods that can be used to compare two conditions provided that true biological replication has been used in the experiment. In general, methods based on calculating the t-statistic are able to identify differentially expressed genes by examining the difference between the means, relative to the spread, or variance of the data by determining the ratio of the difference between two means and measuring the variability, or error variance, between the two data sets for each gene (Cui and Churchill, 2003). The ordinary t-statistic however is still not ideal in the context of microarrays as it is sensitive to genes with unusually low variance, resulting in an excessive number of false-positives in the list of differentially expressed genes.

Genes identified as having a small estimated sample, or error variance, may still have a good chance of giving a large t-statistic even when they are not differentially expressed (Smyth et al. 2003). Smyth (2004) has developed an empirical Bayes based moderated t-test which produces reduced false-positive rates compared to the standard t, and more ad-hoc moderated t-tests such as that used in SAM. This moderated t-test was based on the B statistic developed by Lonnstedt and Speed
The B statistic is essentially a calculated log posterior odds ratio of differential expression versus non-differential expression that takes into account gene-specific variances while combining the information across many genes. Smyth (2004) extended the hierarchical model of Lönnstedt and Speed (2002), resetting the statistic in the context of general linear models with arbitrary coefficients and contrasts of interest. The hybrid classical/Bayes approach is proposed by Smyth (2004) in terms of moderated t-statistics, where the posterior odds of differential expression are shown to depend on the data through the moderated t-statistics. This approach can be further generalized to a moderated F-statistic, allowing for tests to be conducting that simultaneously involve two or more contrasts (Smyth 2004). Motivated by both of these preceding methods, Tai and Speed (2004) propose a one-sample multivariate empirical Bayes statistic (the MB-statistic) to rank genes from replicated microarray time course experiments.

ANOVA models, such as the classical F test, are basically generalization of the t-test that are more suitable for use when two or more conditions are to be compared and can be roughly divided into fixed, random and mixed effects models. The fixed-effects and mixed ANOVA models are generally a more powerful method to use when there are several sources of variation in the data and when consideration needs to be given to multiple factors (Cui and Churchill, 2003). Basically, ANOVA models make multiple estimations of variance in order to determine the overall level of variability within multi-factorial experiments by comparing the variation among replicated samples within and between conditions to determine differential expression. A novel method proposed by Draghici et al. (2003) makes use of a log-linear statistical model and an ANOVA approach to model the noise characteristic of multi-channel arrays. This model is then used to identify differentially expressed genes for a given confidence level (Draghici et al. 2003).

6.4 Determining Significance, False Positive and False Negatives

Once the genes have been suitably ranked the next step in the process is selection of a suitable cut-off value for the differential expression. At the same time the significance or confidence that can be given to the observed differential expression needs to be determined while giving someone allowance and control in regards to the amount of multiple testing needed to conduct a test for each gene such as controlling the family-wise error rate (FWER) or the false discovery rate (FDR) (Smyth, 2004). A simple, however informal, graphical method for assigning significance is to display the genes by their ranking statistic in a normal or t-distribution plot then selecting the genes whose points deviate markedly from the grouped bulk of genes. Depending on the user, manual selection of a cut-off value can typically result in either an over- or under-estimation in regards to differential expression.

False-positives and negatives can be respectively classed as being either a Type I or Type II error (Cui and Churchill, 2003). Calculation of both help determine the confidence that one can have in the results of their data, and in general both types of errors need to be balanced when selecting a cut-off value for differential expression. The problem of multiple testing is that it can increase the number of false-positives and false-negatives within the final list of differentially expressed genes. One of the
most stringent approaches to the multiple testing problem is to control the FWER which determines the probability of accumulating one or more false-positives errors within the final list of differentially expressed genes, thus increasing confidence that the final list is free from such errors. The simplest procedure for controlling the FWER is the Bonferroni correction (Cui and Churchill, 2003) while Dudoit et al. (2000) propose a more rigorous method for controlling the FWER making use of a resampling method which computes a step-down adjusted p-value for each gene. A less stringent and possibly more powerful method for addressing the multiple testing problem is to control the FDR, which determines the expected proportion of false-positives within the list of differentially expressed genes. In contrast to methods that determine significance levels, the FDR is typically computed after the list of differentially expressed genes has been generated therefore providing a post-data method of confidence. Due to its low stringency, the FDR provides an increased number of genes identified as being truly differentially expressed than that of the FWER.

Finally we should note that p-values from microarray experiments are at best approximate. P-values should be seen as an evidence based ranking system. Genes with small p-values have strong evidence, whilst those with large values have weak evidence. The range of p-values gives us a measure of the relative confidence between those at the top of the list, and those further down. Experience has shown that approaches such as moderated t-tests, and FDR adjustments are on the right track, but that exact p-values should be treated with some caution. For this reason, the setting of p-value cut-offs is an arbitrary process, and one should perform exploratory plots before deciding on appropriate cut-off values.

6.5 Verification of Differential Expression

The final stage in the analysis of microarray data requires bioinformatics analysis of the final list of differentially expressed genes, to either characterize the non-annotated gene or to determine the functions and pathways that each gene is involved with. Results of the bioinformatics analysis will aid selecting the genes of most interest from within this final list. Selected genes will then need to have their differential expression confirmed by some of the more traditional or alternative methods for measuring gene expression such as northern blots, qRT-PCR, in situ-hybridization, ribonuclease protection assays, and serial analysis of gene expression (SAGE). Further biological studies may involve altering gene function with targeted mutations, antisense technology or protein inhibition. Ultimately the goal is to come to a conclusion to the biological question that was trying to be answered by the microarray experiment.

6.6 Introduction to Software

Currently there is a diverse range of public and commercially available software for the analysis of microarray data (Table 1). The most commonly used commercial analysis software systems are GeneSpring and GeneSite, neither will be discussed further. Many of these, particularly the commercial software, are represented by a Graphical User Interface (GUI), making analysis simpler and more accessible to a wide range of people by providing a predefined set of operations for the analysis of
Table 1. List of commercially and freely available analysis software for microarrays

| Tool Name       | Description                                                                 | URL                        |
|-----------------|-----------------------------------------------------------------------------|----------------------------|
| GeneSpring      | t-test, various clustering                                                 | www.sigenetics.com         |
| GeneSight       | t-test, various clustering, non-linear normalization                        | www.biodiscovery.com       |
| R               | environment for statistical computing                                        | www.r-project.org          |
| Bioconductor    | set of microarray analysis tools                                            | www.bioconductor.org       |
| affy,           | background correction, normalization, and probeset summarization for Affymetrix GeneChip™ arrays using robust multichip analysis (RMA) | www.bioconductor.org/packages |
| affyImGUI       | Graphical User Interface for limma and affy packages                        | www.bioconductor.org/packages |
| affyPLM         | Quality control (NUSE, RLE) for affymetrix GeneChip™.                       | www.bioconductor.org/packages |
| GCRMA           | background correction, normalization, and probeset summarization for Affymetrix GeneChip™ arrays using Gene Chip Robust Multichip Analysis (GCRMA) | www.bioconductor.org/packages |
| limma           | Empirical Bayes analysis, normalization, analysis of two-colour and single colour arrays | www.bioconductor.org/packages |
| limmaGUI        | Graphical User Interface for limma package                                  | www.bioconductor.org/packages |
| marray          | Diagnostic plots, reading data, normalization of two-colour arrays          | www.bioconductor.org/packages |
| TM4             | various normalization, t-test, Mann-Whitney test, clustering, dissimilarity measures and graphical options | www.tigr.org/software |
| BASE            | LIMS and data analysis system for microarray experiments                     | base.thep.lu.se            |
| Microarray      | t-test, Mann-Whitney test, various normalization                             | www.affymetrix.com         |
| Analysis Suite (MAS) version 5.0 |                                                                             |                            |
| MAANOVA         | ANOVA programs for microarray data. Diagnostics, normalization, ANOVA, Clustering (Matlab, R and Java) | www.jax.org/research/churchill/software/anova |
| Cyber-T         | Differential gene expression, t-test or t-test with Bayesian framework      | http://visitor.ics.uci.edu/genex/cy bert/index.shtml |
| dChip           | Differential gene expression, t-test, model-based analysis for GeneChips™    | www.dchip.org              |
| MAExplorer      | Differential gene expression, scatter plots, k-means clustering dendrograms  | www-lecb.ncifcrf.gov/MAExplorer |

Microarray data. Other programs are command-line driven, such as the set of microarray analysis tools provided by the Bioconductor project (Gentleman et al. 2004) that are designed for use with the R (Ihaka and Gentleman 1996) statistical computing environment. The Bioconductor project is an international initiative for the collaborative creation of extensible open development software for computational biology and bioinformatics. It contains many peer reviewed and award winning algorithms such as RMA which is part of the Affy package (Gautier et al. 2004), robust normalization for two colour microarrays as advocated by Yang.
and Speed (2002), and in the Marray package (Yang and Dudoit 2005) and linear modeling and empirical Bayes adjusted t and F-statistics (Smyth 2003) in the Limma package (Smyth et al. 2005). While some of the analysis software is platform specific, others can readily accept microarray data generated from one- and two-colour microarray platforms. R provides an extensive environment for detailed bioinformatics data mining of microarray datasets such as clustering, principal components analysis, chromosomal clustering, Gene Ontology clustering and over-representation analysis.

Affymetrix provides its own analytical software the Affymetrix Microarray Suite (Table 1), for the analysis of its GeneChip™ data; however a number of publicly available tools have been developed for the storage, management and analysis of Affymetrix probe level data, such as the Affy package of Bioconductor. Affy provides a number of algorithms for background correction such as the robust multichip analysis (RMA) method of Irizarry et al. (2003a,b), which performs background correction, normalization and probe set summarization as well as providing an implementation of Affymetrix's MAS 5.0 algorithm (Gautier et al. 2004). A number of the implemented methods are designed to produce a range of diagnostic plots for the data e.g., 2-D spatial images, boxplots and histograms.

Both the Marray and Limma R packages provide functions for the analysis of two-colour spotted microarray data, providing functions to produce diagnostic plots of spot statistics such as boxplots, scatter-plots, and spatial colour images. Specifically, Limma allows for the use of the empirical Bayes linear modeling approach described by Smyth (2004) for the analysis of designed experiments and identification of differentially expressed genes. Limma also permits appropriate treatment of two levels of replication such as duplicate spots printed on slides and replicated slides or alternatively, a mix of technical and biological replication.

Quality control of Affymetrix chips is provided in the Affy and affyPLM packages. These include RNA degradation plots, NUSE (normalized unsealed standard error) and RLE (Relative Log Error) box plots that are particularly useful for identifying poor quality arrays (Gautier et al. 2004; Bolstad 2005). The linear model functions of the Limma package and those for identifying differentially expressed genes are applicable to all microarrays platforms including Affymetrix GeneChips™ and other single-channel microarray experiments. The Marray package provides some alternative functions for reading and normalizing spotted microarray data, providing flexible location and scale normalization routines for log-ratios. These two packages have a reasonable level of overlap however the Limma package is based on a more general separation between within-array and between-array normalization (Smyth et al. 2005). Wettenhall and Smyth (2004) have also generated a graphical user interface model of the Limma package, limmaGUI (Wettenhall and Smyth 2004). A sister package to limmaGUI has been developed, affylmGUI (http://bioinf.wehi.edu.au/affylmGUI/R/library/affylmGUI), providing a GUI for analysis of Affymetrix microarray data. These GUI's provide a simple point-and-click interface to many of the commonly-used Limma and Affy functions, and provide automated construction of appropriate design and contrast matrices for analysis with Limma so users have to only specify the contrasts they wish to compare.
Some other publicly available alternatives to the open-source software of Bioconductor and R are the TM4 (Saeed et al. 2003) microarray analysis suite of Java-based tools and the BioArray Software Environment (BASE) (Saal et al. 2002) which provides a Web-based approach, using standard browsers to interact with a central microarray database and appropriate data analysis tools. The major advantage of the BASE approach is that it removes the need for users of the system to have to ensure that their software is current and the calculations underlying any analysis are passed on to more powerful central servers, keeping the user's desktop computers free.

When selecting the most suitable software to use for analysis of microarray data, the end choice will ultimately depend on the bioinformatics savvy and statistical knowledge of the user as well as the availability of money if considering the use of commercial software.

7. CONCLUSION

Since the advent of high-throughput microarray technology there has been a focus on improving this technology and developing suitable experimental designs and statistical algorithms for image processing, data cleaning, and identifying differentially expressed genes. In parallel, there has also been a focus on the development of user friendly software for the implementation of these algorithms and techniques for image processing and analysis phases of microarray experiments. This review has summarized the current status of microarray technology and issues concerning the experimental design, image processing and statistical analysis of microarray experiments while the proceeding chapter by Tjaden and Cohen will go into more detail regarding the statistical algorithms used for clustering microarrays.

Future directions of microarrays will follow in a similar manner with their being a strong focus on increasing the high-throughput capability of microarrays, improving protocols, improving the analysis of the vast quantities of raw gene expression data and the generation of user-friendly analysis software while reducing the overall cost of performing a microarray experiment. Although microarrays have been used for fungal research, virtually from as soon as the technology emerged, there is still huge potential for increasing our mycology knowledge base by utilizing microarray technology and then applying this knowledge within the pharmaceutical and agricultural industries via biotechnology based research.

Out of the 70 fungal genome projects currently accessible from NCBI (http://www.ncbi.nlm.nih.gov), 9 of these contain completed sequences of fungal genomes, 37 of the projects are in the process of assembling fungal genome sequence fragments and 29 of them are still in the process of sequencing. On top of this there are still many more fungal genome projects that are underway in both public and private laboratories that have not yet been made accessible. The sequence information from all of these projects can be used to develop fungal species and strain specific whole genome microarrays allowing for high-throughput gene expression studies on a genome-wide scale. Generation of whole-genome arrays will allow for a rapid elucidation of the cellular mechanisms that allow fungi to exist either alone or in a host-pathogen or host-symbiont relationship. Such studies may also lead to the identification of novel pathways by which fungi affect organisms
such as plants or humans that can then be targeted by the pharmaceutical or agricultural industries for the development of drugs or fungicides for eradication of invasive, disease causing fungal pathogens. In regards to the symbiotic relationship of fungi with plants, whole-genome expression analysis studies will help in the elucidation, and possibly identification of cellular and novel mechanisms by which fungi are needed for growth of the host plant. Overall, utilization of microarrays within mycology will provide insight into the function of varying cellular functions of fungi at the gene level.

REFERENCES
Adams R and Bischof I (1994) Seeded region growing. IEEE Transactions on Information Technology in Biomedicine 16: 1651-1656.
Ahmed AA, Vias M, Iyer NG, Calsda C and Brenton JD (2004) Microarray segmentation methods significantly influence data precision. Nucleic Acids Research 32: e50.
Allen TD, Dawe AL and Nuss DL (2003) Use of cDNA microarrays to monitor transcriptional responses of the chestnut blight fungus Cryptonectria parasitica to infection by virulence-attenuating hypoviruses. Eukaryotic Cell 2: 1253-1265.
Arava Y, Wang Y, Storey JD, Liu CL, Brown PO and Herschlag D (2003) Genome-wide analysis of mRNA translation profiles in Saccharomyces cerevisiae. Proceedings of the National Academy of Sciences of the United States of America 100: 3889-3894.
Backhus LE, DeRisi J and Bisson LF (2001) Functional genomic analysis of a commercial wine strain of Saccharomyces cerevisiae under differing nitrogen conditions. FEMS Yeast Research 1: 111-125.
Baldi P and Hatfield W (2002) DNA Microarrays and Gene Expression. From Experiments to Data Analysis and Modeling, Cambridge University Press, Cambridge, United Kingdom.
Baldwin D, Crane V and Rice D (1999) A comparison of gel-based, nylon filter and microarray techniques to detect differential RNA expression in plants. Current opinion in Plant Biology 2: 96-103.
Barczak A, Rodrigues MW, Hanspers K, Koth LL, Tai YC, Bolstad BM, Speed TP and Erle DJ (2003) Spotted long oligonucleotide arrays for human gene expression arrays. Genome Research 13: 1775-1785.
Bernstein BE, Kamal M, Lindblad-Toh K, Bekiranov S, Bailey DK, Huebert DJ, McMahon S, Karbsson EK, Kulbokas EJ, 3rd, Gingeras TR, Schreiber SL and Lander ES (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. Cell 120: 169-181.
Bertone P, Gerstein M and Snyder M (2005) Applications of DNA tiling arrays to experimental genome annotation and regulatory pathway discovery. Chromosome Research 13: 259-274.
Beucher S and Meyer F (1993) The morphological approach to segmentation: the watershed transformation. Mathematical morphology in image processing. Optical Engineering 34: 433-481.
Bolstad B (2005) affyPLM: Fitting Probe Level Models. www.bioconductor.org/repository/devel/vignette/AffyExtensions.pdf.
Bolstad BM, Irizarry RA, Astrand M and Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.
Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J and Corcoran K (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nature Biotechnology 18: 630-634.
Brewster JL, Beason KB, Eckdahl TT and Evans IM (2004) The Microarray Revolution. Biochemistry and Molecular Biology Education 32: 217-227.
Cawley S, Bekiranov S, Ng HH, Kapranov P, Sekinger EA, Kampa D, Piccolboni A, Sementchenko V, Cheng J, Williams AJ, Wheeler R, Wong B, Drenkow J, Yamanaka M, Patel S, Brubaker S, Tammana H, Helt G, Struhl K and Gingeras TR (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell 116: 499-509.
Chen Y, Dougherty ER and Bittner ML (1997) Ratio based decisions and the quantitative analysis of cDNA microarray images. Journal of Biomedical Optics 2: 364-374.

Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R and Childs G (1999) Making and reading microarrays. Nature Genetics 21: 15-19.

Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. Nature Genetics 32 Suppl: 490-495.

Colebatch G, Trevaskis B and Udvardi M (2002) Functional genomics: tools of the trade. New Phytologist 153: 27-36.

Cui X, Kerr MK and Churchill GA (2003) Transformations for cDNA Microarray Data. Statistical Applications in Genetics and Molecular Biology 2: 1-20.

DeRisi JL, Iyer VR and Brown PO (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278: 680-686.

Draghici S, Kulaeva O, Hoff B, Petrov A, Shams S and Tainsky MA (2003) Noise sampling method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. Bioinformatics 19: 1348-1359.

Dudoit S, Yang YH, Callow MJ and Speed TP (2000) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Department of Statistics, UC Berkeley, CA, pp. Technical Report 578.

Dunn B, Levine RP and Sherlock G (2005) Microarray karyotyping of commercial wine yeast strains reveals shared, as well as unique, genomic signatures. BMC Genomics 6.

Ekins R and Chu FW (1999) Microarrays: their origins and applications. comment. Trends in Biotechnology 17: 217-218.

Fodor SP, Read JL, Pirrming MC, Stryer L, Lu AT and Solas D (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251: 767-773.

Franken P and Requena N (2001) Analysis of gene expression in arbuscular mycorrhizas: new approaches and challenges. New Phytologist 150: 517-523.

Gautier L, Cope L, Bolstad BM and Irizarry RA (2004) affy - analysis of Affymetrix GeneChip data at the probe level. Bioinformatics 20: 307-315.

Gerhold DL, Jensen RV and Gullans SR (2002) Better therapeutics through microarrays. Nature Genetics 32 Suppl: 547-551.

Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegeman JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Ross M, Ross-Macdonald P, Scherens B, Schimack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW and Johnston M (2002) Functional profiling of the Saccharomyces cerevisiae genome. Nature 418: 387-391.

Glonek GF and Solomon PJ (2004) Factorial and time course designs for cDNA microarray experiments. Biostatistics 5: 89-111.

Ihaka R and Gentleman R (1996) R: Language for data analysis and graphics. Journal of Computational Graphics 5: 299-314.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B and Speed TP (2003a) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Research 31: e15.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U and Speed TP (2003b) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249-264.

Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, Gabrielson E, Garcia JG, Geoghegan J, Germino G, Griffin C, Hilmer SC, Hoffman E, Jedlicka AE, Kawasaki E, Martinez-Murillo F, Morsberger L, Lee H, Petersen D, Quackenbush J, Scott A, Wilson M, Yang Y, Ye SQ and Yu W
(2005) Multiple-laboratory comparison of microarray platforms.[see comment]. Nature Methods 2: 345-350.
Jianbing F, Diping C, Chanfeng Z, Lixin Z and Wenyi F (2003) High-density fiber optic array technology and its applications in functional genomic studies. Chinese Science Bulletin 48: 1903-1905.
Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP and Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. Science 296: 916-919.
Kazan K, Schenk PM, Wilson I and Manners JM (2001) DNA microarrays: new tools in the analysis of plant defence responses. Molecular Plant Pathology 2: 177-185.
Kehoe DM, Villand P and Somerville S (1999) DNA microarrays for studies of higher plants and other photosynthetic organisms. Trends in Plant Science 4: 38-41.
Kerr MK (2003) Design considerations for efficient and effective microarray studies. Biometrics 59: 822-828.
Kerr MK, Afshari CA, Bennett L, Bushel B, Martinez J, Walker NJ and Churchill GA (2002) Statistical analysis of a gene expression microarray experiment with replication. Statistica Sinica 12: 203-217.
Kerr MK and Churchill GA (2001) Experimental design for gene expression microarrays. Biostatistics 2: 183-201.
Kohane IS, Kho AT and Butte AJ (2003) Microarrays for an Integrative Genomics, The MIT Press Massachusetts Institute of Technology Cambridge, Massachusetts London, England.
Kricka LJ and Forina P (2001) Microarray technology and applications. Clinical Chemistry 47: 1479-1482.
Kuhn K, Baker SC, Chudin E, Lieu M, Oeser S, Bennett H, Rigault P, Barker D, McDaniel TK and Chee MS (2004) A novel, high-performance random array platform for quantitative gene expression profiling. Genome Research 14: 2347-2356.
Lashkari DA, DeRisi JL, McCusker JH, Namath AF, Gentile NC, Hwang SY, Brown PO and Davis RW (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proceedings of the National Academy of Sciences of the United States of America 94: 13057-13062.
Li C and Wong WH (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proceedings of the National Academy of Sciences of the United States of America 98: 31-36.
Lockhart DJ, Dong H, Byrne MC, Folletti MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H and Brown EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. see comment. Nature Biotechnology 14: 1675-1680.
Lonnstedt I and Speed T (2002) Replicated microarray data. Statistica Sinica 12: 31-46.
Maindonald JH, Pittelkow YE and Wilson SR (2003) Some considerations for the design of microarray experiments. Science and Statistics 40: 367-390.
Mantripragada KK, Buckley PG, Stahl TD and Dumanisk JP (2004) Genomic microarrays in the spotlight. Trends in Genetics 20: 87-94.
McLachlan GJ, Do K and Ambroise C (2004) Analyzing Microarray Gene Expression Data, John Wiley & Sons, Hoboken, New Jersey.
Newton M, Kendzierski C, Richmond C and Blattner F (2001) On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. Journal of Computational Biology 8: 37-52.
Nicolaesen M, Justesen AF, Thrane U, Skouboe P and Holmstrom K (2005) An oligonucleotide microarray for the identification and differentiation of trichothecene producing and non-producing Fusarium species occurring on cereal grain. Journal of Microbiological Methods 62: 57-69.
Polsky-Cynkin R, Parsons GH, Allerd L, Landes G, Davis G and Rashchitchian A (1985) Use of DNA immobilized on plastic and agarose supports to detect DNA by sandwich hybridization. Clinical Chemistry 31: 1438-1443.
Rementeria A, Lopez-Molina N, Ludwig A, Vivanco AB, Bikandi J, Ponton J and Garaitz J (2005) Genes and molecules involved in Aspergillus fumigatus virulence. Revista Iberoamericana de Micologia 22: 1-23.
Ritchie ME (2004) Quantitative quality control and background correction for two-colour microarray data. Department of Medical Biology, The Walter and Eliza Hall Institute of Medical Research, University of Melbourne.
Rocke DM and Durbin B (2001) A model for measurement error for gene expression arrays. Journal of Computational Biology 8: 557-569.

Saal LH, Troein C, Vallon-Christersson J, Grubberger S, Borg A and Petersen C (2002) BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. Genome Biology 3: software0003.0001-0003.0006.

Saeed AI, Sharov J, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V and Quackenbush J (2003) TM4: A Free, Open-Source System for Microarray Data Management and Analysis. Biotechniques 34: 374-378.

Schadt EE, Li C, Ellis B and Wong WH (2001) Feature Extraction and Normalization Algorithms for High-Density Oligonucleotide Gene Expression Array Data. Journal of Cellular Biochemistry Supplement 37: 120-125.

Schenck PM, Kazan K, Wilson I, Anderson JP, Richmond T, Somerville SC and Manners JM (2000) Coordinated plant defense responses in Arabidopsis revealed by microarray analysis. Proceedings of the National Academy of Sciences of the United States of America 97: 11655-11660.

Shepard KA, Gerber AP, Jambhekar A, Takizawa PA, Brown PO, Herschlag D, DeRisi JL and Vale RD (2003) Widespread cytoplasmic mRNA transport in yeast: identification of 22 bud-localized transcripts using DNA microarray analysis. Proceedings of the National Academy of Sciences of the United States of America 100: 11429-11434.

Simon RM and Dobbin K (2003) Experimental design of DNA microarray experiments. Biotechniques Supplement: 16-21.

Smyth GK (2004) Linear models and empirical Bayes for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology 3: Article 1.

Smyth G, Thorne N and Wettenthal J (2003) limma: Linear Models for Microarray Data User's Guide, http://bioinf.wehi.edu.au/lmma/usersguide.pdf;

Smyth GK, Yang YH and Speed T (2003) Statistical issues in cDNA microarray data analysis. Methods in Molecular Biology 224: 111-136.

Stolovitzky GA, Kundaje A, Held GA, Duggar KH, Haudenschild CD, Zhou D, Vasicek TJ, Smith KD, Aderem A and Roach JC (2005) Statistical analysis of MPSS measurements: Application to the study of LPS-activated macrophage gene expression. Proceedings of the National Academy of Sciences of the United States of America 102: 1402-1407.

Tenenbaum SA, Carlson CC, Lager PJ and Keene JD (2000) Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. Proceedings of the National Academy of Sciences of the United States of America 97: 14085-14090.

Tai YC and Speed T (2004) A multivariate empirical Bayes statistic for replicated microarray time course data. Department of Statistics, University of California, Berkeley, (submitted), pp. Technical Report #667.

Troyanskaya OG, Garber ME, Brown PO, Botstein D and Altmann RB (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 18: 1454-1461.

Voigt M, Duplessis S, Encelot N and Martin F (2000) Identification of symbiosis-regulated genes in Eucalyptus globulus-Psilotus tinctorius ectomycorrhiza by differential hybridization of arrayed cDNAs. The Plant Journal 25: 181-191.

Wang D, Urisman A, Liu Y, Springer M, Ksiezak TG, Erdman DD, Mardis ER, Hickenbotham M, Magrini V, Eldred J, Latreille JP, Wilson RK, Ganem D and DeRisi JL (2003) Viral Discovery and Sequence Recovery Using DNA Microarrays. PLoS Biology 1: 257-260.

Wettenhall JM and Smyth GK (2004) limmaGUI: A graphical user interface for linear modeling of microarray data. Bioinformatics 20: 3705-3706.

Wu Z and Irizarry RA (2004) Preprocessing of oligonucleotide array data. Nature Biotechnology 22: 656-658.

Yang YH, Buckley MJ, Dudoit S and Speed TP (2002a) Comparison of methods for image analysis on cDNA Microarray Data. Journal of Computational & Graphical Statistics 11: 108-136.

Yang YH, Buckley MJ and Speed TP (2001) Analysis of cDNA microarray images. Briefings in Bioinformatics 2: 341-349.
Yang YH and Dudoit S (2005) Bioconductor’s marray package: Plotting component, www.bioconductor.org/repository/devel/vignette/marrayPlots.pdf,
Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J and Speed TP (2002b) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research 30: e15.
Yang YH and Speed T (2002) Design Issues for cDNA Microarray Experiments. Nature Reviews 3: 579-588.
Yang YH and Thorne NP (2003) Normalization for Two-color cDNA Microarray Data.
Yauk CL, Berndt ML, Williams A and Douglas GR (2004) Comprehensive comparison of six microarray technologies. Nucleic Acids Research 32: 1-7.
Zhou X, Rao NP, Cole SW, Mok SC, Chen Z and Wong DT (2005) Progress in concurrent analysis of loss of heterozygosity and comparative genomic hybridization utilizing high density single nucleotide polymorphism arrays. Cancer Genetics and Cytogenetics 159: 53-57.