Chapter 2: Microarrays and their Application in Parasitology

2.1 Introduction

Microarrays are specially produced slides which have thousands of individual DNA probes attached in an ordered array to the surface. They provide the user with the ability to view the expression level of thousands of genes simultaneously [63]. In 1995, Schena and co-workers reported the first cDNA microarray analytical procedure using 45 genes from the plant *Arabidopsis* which were printed onto a glass slide with the use of an arraying machine [64]. Since then, this technology has expanded, allowing for new applications in genomic research; for example light directed *in situ* synthesised DNA arrays may contain 135,000 or more probes on a single slide "chip" [65]. Moreover, experimental versions of commercial manufactured arrays now exceed one million individual probes per array [66]. This miniaturisation of the probes has allowed for greater sensitivity and more genes to be analysed per chip [65]. In addition, entry level array probe printing machines have made the production of chips less expensive in a general academic setting [67]. With the establishment of the discipline of microarray technology, a new generation of terminology and acronyms has evolved; examples of these are included in Table 2.1 [63].

Table 2.1 Key microarray terminology 'Adapted from Rosetta BioSoftWare: http://www.rosettabio.com/tech/geml/omg/lsr_ge_glossary.doc'.

Array	Refers to the physical substrate to which bio-sequence reporters are attached to create **features**.
Array Design	An **array design** is conceptual it is the layout or blueprint of one or more **arrays**.
Background/ Background noise	**Background** is the measured signal outside of a **feature** on an **array**. In many gene **expression** analysis methods, **background** subtraction is performed to correct measured signals for observed local and/or global **background**.
Channel	A **channel** is an intensity-based portion of an **expression** dataset that consists of the set of signal measurements across all **features** on an **array** for a particular **labelled** preparation used in a **hybridization**. In some cases, such as Cy3/Cy5 **array hybridizations**, multiple **channels** (one for each **label** used) may be combined in a single **expression** profile to create **ratios**.
Chip	The physical medium of many **arrays** used in gene **expression**.
contig	A **contig,** an abbreviation for "contiguous sequence" is a group of clones representing overlapping regions of a genome.
Control	The reference for comparison when determining the effect of some procedure or treatment. (Deletion, mismatch, positive, negative).
Error Model	An **error model** is an algorithm that computes quality statistics such as p-values and error bars for each gene **expression** measurement.
Expression	The conversion of the genetic instructions present in a DNA sequence into a unit of biological function in a living cell. Typically involves the process of transcription of a DNA sequence into an RNA sequence followed by translation of the mRNA into protein.
Feature	A **feature** refers to a specific instance of a position upon an **array**. Commonly referred to as a spot in a microarray experiment.
Feature Extraction	Quantitative analysis of an **array** image or scan to measure the **expression** values.
Filter/ed	A mathematical algorithm applied to image/array data for the purpose of enhancing image quality/defining expression analysis
Fluor/ Fluorophore/ Fluorescent label	A fluorescent tag bound to mRNA or cDNA extracted from a sample. When properly excited the fluor gives off measurable fluorescence which is the observable in an experiment.
Hybridization	Treating an **array** with one or more **labelled** preparations under a specified set of conditions.
Label	**Label** refers to fluorescent **labels**, for example, Cy3 and Cy5, commonly used to distinguish baseline and experimental preparations in gene **expression** microarray **hybridizations**.
Normalisation	**Normalisation** is the procedure by which signal intensities from two or more **expression** profiles (or **channels**) are made directly comparable through application of an appropriate algorithm.
Oligo / Oligonucleotide	Usually short strings of DNA or RNA to be used as **probes** (**features**) or spots. These short stretches of sequence are often chemically synthesised.
Probe	In some organisations, **probe** is used as a synonym for **feature**.
Ratio	Also referred to as "fold change". A **ratio** refers to a **normalised** signal intensity generated in a **feature** given **channel** divided by a normalised signal intensity generated by the same **feature** in another **channel**. The **channels** compared are typically baseline versus experimental, for example normal versus diseased or untreated vs. treated.
Target	Material that may **hybridize** to the **probe**, usually containing all of the mRNA (cDNA or cRNA) or gDNA of the subject organism.

2.2 Construction of microarrays

There are many ways to construct microarrays, but all share characteristics which may be described as follows [68]:

(1) Photolithography. This technique utilises photo-lithographic masks (a series of laser designed templates for an individual microarray chip) to control the exposure of light for each round of oligonucleotide synthesis, an example of which is the Affymetrix GeneChip® [69]. This technology has enabled analysis of nucleic acid expression from small samples and has recently allowed researchers to access arrays of over a 100,000 probes [66, 70, 71]. The disadvantages that are associated with this type of microarray, are the limited size of probes since the full length yield falls rapidly with synthesis [65], a sequence change within the array would require the manufacture of new masks and additionally, the small size of the probes may not be suitable for some experiments [72].

(**2) Ink-jet arrays.** These are non-contact printed chips, second only in density to photolithographic chips, examples of which are the Agilent 60-oligomer (mer) custom arrays [65]. This method utilises a robotic spotting *in situ* method to deposit complementary DNA (cDNA) onto a specially prepared surface [69], the details of which will be described in Chapter three. Non-contact printed microarrays are easier to produce and allow the production of longer probes increasing the specificity of hybridization [65].

(3) Simple oligonucleotide arrays. In this technique, the manufacture of oligonucleotides is performed separately and then chips are fabricated by simple array printing machines, which makes this method inexpensive compared to others [73]. The use of oligonucleotides as probes in these arrays enable specific hybridization, distinguishing single-nucleotide polymorphisms and splice variants [69].

(4) Complementary DNA (cDNA) array chips. These arrays are made from a selection of probes that are printed as full length, partially sequenced or randomly chosen cDNAs [74]. These cDNA probes are transferred to a glass slide by an array printing machine and stored until use [75]. The manufacture of cDNA array chips is readily available by using simple array printing machines, which also makes this an inexpensive method (Figure 2.1) [65, 73]. There are some limitations to cDNA arrays, in that they require a large amount of total RNA per hybridization [74], the PCR or cDNA products are not as specific as oligonucleotides [69] and often multiple experimental repeats are required to demonstrate gene expression measurement reproducibility [76, 77].

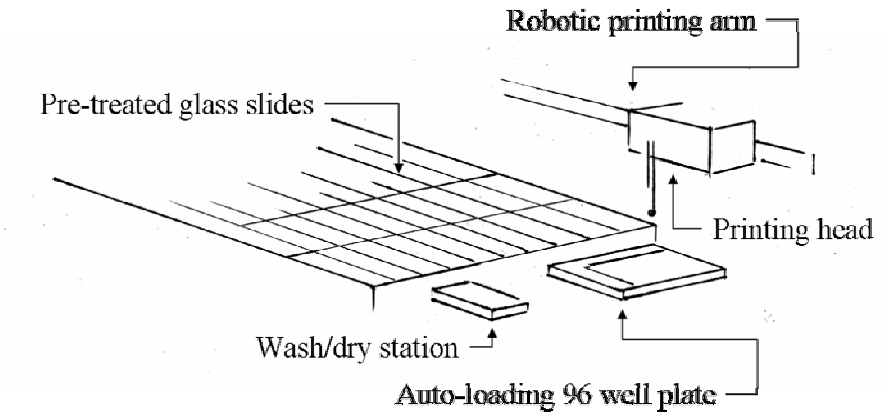

Figure 2.1 A typical cDNA microarray-printing machine 'Adapted from [67]'. The cDNA array may be produced by extracting messenger RNA (mRNA) from total RNA from the organism or tissue to be studied and creating cDNA by use of an oligonucleotide primer. The cDNA is inserted into a plasmid before being transferred into bacterial cells which are plated to grow into separate colonies. The now cloned plasmids containing the inserts are removed to have the cDNA amplified by oligonucleotide primers. This cDNA probe is transferred to a glass slide by an array printing machine and stored until use [75].

2.3 M.I.A.M.E.

The need for a standard of Minimum Information About a Microarray Experiment (MIAME) was first highlighted during a meeting organized by the European Bioinformatics Institute in 1999. After development and discussion, MIAME was proposed as standard practice and reported in the journal *Nature Genetics* in 2001 [78]. MIAME is a detailed list of information that describes the experimental process from construction of the chip to data analysis [63]. The MIAME standard [79] is made up of two major sections: (a) Array design and (b) Gene expression description.

The array design description contains two further sub-sections:

(1) Array related information, including design name, platform type, and number of features.

(2) Information about the probes, sequence, type, attachment, location on array and controls.

The gene expression experiment description contains three further sub-sections

(1) Experimental design, including authors, type of experiment, experimental factors (time dose), and quality controls.

(2) Samples used, extract preparation and labeling, sex, developmental stage, type, biomaterial manipulations, protocol, conditions, treatments, hybridization extract preparation protocol, external controls.

(3) Hybridization procedures and parameters, batch serial numbers, blocking agent, wash procedure, quantity of target.

Most of these experimental conditions are good research practice and, in general, they provide a basis that allows the research community access to microarray generated data [78]. It is recognised that this standard still has some failings in that it only focuses on documentation of experimental details [80], the majority of which can

be addressed by increasing the number of independent target validations and following good experimental practice. Taking the failings of MIAME into consideration the 'minimal information standard' presented will add to the design and utilisation of larger databases that will correlate individual microarray datasets [63].

2.4 Genomics, expressed sequence tags and microarray construction

The advances in sequencing methods and associated bioinformatics [81] have enabled the establishment of many large scale sequencing projects for a range of organisms including a number of parasites. These include several of medical and veterinary importance such as *Plasmodium*, *Brugia malayi* and *Schistosoma sp.* [25, 58, 82, 83]. The abundance and richness of this new sequence data provide the basis for the design and construction of parasite microarrays [63].

The availability of a very large number of ESTs and genomic sequences [84, 85] and then the complete genomic sequencing of *Plasmodium* [36, 86], has provided huge insights and information about the malaria parasite. Subsequent microarray analysis has provided the basis for a better appreciation of the biology and pathogenesis of malaria [86]. The filarial nematodes are another prominent group of parasites subject to large scale EST and genomic sequencing. These studies have helped in our understanding of the complexity of the genome of *Brugia malayi*, for which over 25,000 partially sequenced cDNA clones have been submitted to the EST databases [87, 88]. Additionally, these sequencing efforts have led to major advances particularly in the area of chromosome mapping and functional genomics [63, 82, 89].

Taenia solium, arguably the third world's major parasite responsible for brain disorders, is also a target of a genome project [90]. The project consists of two major stages, the first being to determine basic parameters of the parasite which will include

characterising several thousand adult worm and cysticerci ESTs, and genomic clones [90]. This will be followed secondly by the production of synthetic oligonucleotides from identified ESTs [90]. These oligonucleotides will enable the study of gene expression or transcriptional analysis through microarrays [90].

The publication of three draft kinetoplastid's parasite genomes in July 2005 has provided new insights into the biology of *Trypanosoma brucei*, *T. cruzi* and *Leishmania major* [91-94]. The approximate 29,000 genes in total will enable new insights into the evolution of the parasites [91]. Furthermore, new therapeutic and vaccine targets will be identified through projects that will follow this research such as microarrays.

2.5 Applications

2.5.1 Comparison of gene expression during the parasite life cycle

A good example of the use of microarrays for comparison of gene expression in life cycle studies is the transformation expression of *Trypanosoma cruzi* from trypomastigote to amastigote in an axenic system [95]. This study was based on the use of approximately 4,400 probes, including 3,014 genomic sequences and 1,248 open reading frame library probes, to investigate the expression of genes in trypomastigotes and developing amastigotes. Mining *et al*. [95] used green fluorescent protein (GFP) gene expression based on a bacterial system for library selection, with the aim to create a microarray to identify vaccine targets and amastigote-specific genes. The GFP gene encodes a spontaneously fluorescent protein isolated from coelenterates. In such gene expression studies, if the open reading frame lacking a start codon is inserted upstream of the GFP sequence, after transformation, all colonies that contain an insert will

fluoresce under an ultraviolet light. Mining *et al*. [95] showed most differential expression was due to the up-regulation of 60 genes in the developing amastigote, including 25 novel and 14 previously characterised *T. cruzi* genes. In order to validate these results, they used real time PCR as an independent measure of gene expression [95]. The real time PCR results confirmed the microarray findings with 12 of the 13 genes showing similar expression profiles [95].

A more informative series of experiments using cDNA probes was performed on *T. cruzi* by Baptista *et al*. [96]. This group used 710 ESTs, representing 665 unique genes, to create a microarray for examining gene expression and genomic organisation in different isolates of *T. cruzi;* they identified 68 probes differentially expressed between two strains using genomic DNA. Independent verification of hybridization variation was shown by Southern blots. The analysis of the Southern blots confirmed some of the microarray expression results, but the comparative genome analysis was unidirectional, effectively representing only a small portion of genomic differences between isolates. Additionally any variation shown by Southern blots may correspond to repetitive elements within isolates [96]. Gene expression analysis between the two strains revealed that 84 of 730 probes were differentially expressed, and of these, 9 were validated by Northern blotting. Gene copy number between strains showed only 7/35 and 11/49 probes with higher hybridization with the Silvio and CL Brener strains, respectively [96]. This demonstrated that hybridization visualised by the microarray results is mainly due to gene expression with only a small proportion due to gene copy number [63].

2.5.2 *Plasmodium falciparum* microarrays

Oligonucleotide microarrays have been used to explore expression profiling, gene function and the transcriptome of *P. falciparum* [97-99].

Bozdech *et al*. [97] used a first generation 70-mer oligonucleotide microarray representing approximately 6,000 open reading frames (ORFs) to analyse the gene expression of the trophozoite and schizont stages of *P. falciparum* [97]. This group developed a software package, OligoSelector, in order to design their microarray probes, ORF-specific DNA, which were derived from public databases of *P. falciparum*. Bozdech *et al*. [97] noted extensive differential expression between the two malarial stages, demonstrating the significant advantage of *in silico* design of probes compared to using cDNA clones. The microarray probes had high hybridization efficiencies due to the selection of ORF candidates demonstrating unique 70-mer sequences [97]. This study is a good example where public data bases can be accessed for probe sequence design to create a target-specific microarray [63].

More extensive expression profiling of *P. falciparum* was shown in a study by Le Roch *et al*. [98]. This group used 367,226 probes on multiple chips to discover potential gene function by expression profiling of different life stages of the malaria parasite. As this study encompassed the nine different lifecycle stages of *P. falciparum* (human and mosquito life stages) there was an advantage of using probes designed from genomic sequences as opposed to life stage specific ESTs [63]. However, it is noteworthy that this would be a disadvantage if the study was based upon species differential expression, as genomic DNA probes can hybridize to non coding targets. This study was able to show the shift in transcriptional energy from protein synthesis to cell surface structures through expression levels which varied by five orders of magnitude [98]. The authors also concluded that a particular gene expression profile can elaborate on its cellular profile by clustering the expression of known and unknown targets. Additionally, uncharacterised genes may have their cellular processes

represented by characterised contigs within an expressed cluster demonstrating that arrays may be used to identify potential gene function [98].

Carret *et al*. [99] used a custom made 25-mer Affymetrix malaria microarray to access the amplification suitability for analysis of the *P. falciparum* genome. With no more than 80 ng of starting material they were able to show that the non-PCR based multiple displacement method demonstrated clear advantages in amplification of limited target quantity [99]. This study demonstrates that microarrays may also be used as a pilot method as a basis for further studies, additionally, amplification methods can also be used for expression studies, especially where there is limited starting material.

Problems in microarray construction, resulting from the very limited amounts of mRNA that can be isolated from different parasite life cycle stages [63] can reduce project goals. This issue can now be overcome due to the introduction of new cDNA library construction kits, which can be used in microarray construction. These new kits, an example of which is the Amino Allyl Message Amp II antisense RNA Kit (Ambion) [63] are able to amplify fluorophore-labelled targets which will significantly help in microarray construction and analysis where available material for analysis is limited.

2.5.3 Analysis of infected host tissues using microarrays

Parasitic infections may be monitored by microarray investigation of the gene expression response of infected host tissue. Sexton *et al*. [100] showed there was transcriptional changes in more than 1,000 genes which occur in both the brain and spleen tissue of malaria-infected mice. This group utilised a commercially produced oligonucleotide mouse array to monitor the effects of the malaria infection which they showed promoted a modification in specific gene expression profiles of the host tissues. These modifications included an early infection suppression of erythropoiesis as well as

an up-regulation of genes that control glycolysis. This study was elegant, showing definite expression changes in tissues using high quality microarray chips together with an excellent independent validation system. Despite this, MIAME standards were not adhered to, including no information of sampling numbers and whether tissue samples were pooled or not [63]. It is known that significant variation in gene expression can occur between individuals [63]. In the design of microarray experiments care must be taken to ensure that the numbers of samples are relevant to the hypothesis being tested. For example, in pooling samples, subtle fluctuations in gene expression occurring in individual cells or tissues can be lost [63]. Reflectively, the differential gene expression in one individual or cell would not represent a species or cell line [101].

A similar study by Hoffmann *et al*. [102] examined gene expression in the livers of cytokine-deficient mice using cDNA microarrays [63]. This group hypothesised that severe schistosome-induced liver disease can develop via two detrimental genetic programs [102]. They identified gene expression profiles that were associated with type 1 and 2 cytokine immune responses and expanded the knowledge of the disease mechanisms attributed to granuloma formation caused by *S. mansoni* infection [102]. Although this study had no independent validation it provides a benchmark in the field of schistosome microarray research.

2.5.4 Comparative gene expression between related species

There have been relatively few microarray investigations of inter-specific variation between species, even fewer in parasite studies [63]. A major limitation when investigating expression differences between species is the variable degree of homology of the target total RNA that is probed [63]. Base differences between target samples will result in a limited hybridization, which may be misinterpreted as a low gene

expression level; yet biological meaningful data can still be obtained from such experiments demonstrating hybridization efficiencies of probes [103].

Three good examples of such successful interspecies analysis include cDNA-based microarrays from fish, *Astatotilapia* [103] and *Salmo* [104], and schistosomes namely *S. japonicum* and *S. mansoni* [105]. The gene expression studies between different fish species clearly demonstrated that the more closely related species produced better hybridization data compared to other, more divergent species [103, 104]. The two studies demonstrated that it is feasible to use a microarray platform to examine a wide range of species to generate evolutionary and ecologically relevant data [63]. Taxonomic classification can be demonstrated by large scale hybridization investigation of genes. Additionally, such large scale studies can be used to identify small variations that occur between closely related species or strains. The hybridization variation demonstrated by the species and inter-species microarray studies can also be adapted to the study of parasites, a principle utilised in a comparative study of the transcriptomes *S. japonicum* and *S. mansoni* by Gobert G N, McInnes R, **Moertel L P**, Nelson C, Jones M K, Hu W and McManus D P. [105], as described in the following paragraphs.

2.6 Microarray tools for schistosome research

One of the eight selected diseases targeted for study and control by the World Health Organization (WHO) is human schistosomiasis [8]. One initiative of WHO was the formation of the *Schistosoma* Genome Network [106], which consists of a number of laboratories who have exploited the use of ESTs, and genomic sequencing strategies to identify novel genes [106]. Sequencing efforts have culminated with the release of considerable numbers of ESTs for *S. japonicum* [25] and *S. mansoni* [58], which are

readily accessible in the public databases. Together with *in silico* design, automated oligonucleotide synthesis and spotting, these data sets have enabled researchers to access raw sequences to generate several schistosome microarrays as laboratory tools [63].

Microarrays have been used to show differential expression within the schistosome life cycle. Dillon *et al*. [107] used a 6,000 feature microarray to identify genes preferentially expressed in the lung schistosomulum of *S. mansoni*. This group used an array comprising of ESTs exclusive to the lung schistosomula to visualise gene expression in seven life stages; early liver and adult worms, eggs, germ balls from developing daughter sporocysts, cercariae and day two and seven schistosomula [107]. They showed that there were many genes preferentially expressed in the lung stage of schistosomes. However, this paper did not report on the actual number of these genes; additionally only a small number of ESTs were used in independent validation. A far better example of life stage gene expression profiling by arrays was the study of Vermeire *et al*. [108]. This group used a cDNA microarray consisting of 7,335 unique elements from *S. mansoni* to compare gene expression in the miracidium and mother sporocyst stage [108]. They found that 361/273 probes showed stage-associated expression in the miracidium and sporocyst, respectively. Additionally, they used 22 oligonucleotides to independently verify the microarray data by real time PCR. The major limitations of this study were that cDNA probes were spotted on the chip, and without obtaining these cDNAs or chips, other laboratories will not be able to verify these findings. Such limitations are common to 'in house' printed arrays, containing sequence mismatches in reported ESTs. This would not be a problem if the microarray chip probes are generated *in situ* where probes are created "on chip" as per reported sequence. This is further explored in Chapter 3.

In this current study a 60-mer microarray was constructed from two extensive EST public datasets for *S. japonicum* [25] and *S. mansoni* [58] to explore the differential expression differences of the two species including three major schistosome studies:

- Transcriptomics of *S. mansoni* and *S. japonicum* adult worms (Chapter 3) [105].
- Transcriptome profiling of lung schistosomula, *in vitro* cultured schistosomula and adult *S. japonicum* (Chapter 3) [109].
- Analysis of strain- and gender-associated gene expression in the human blood fluke, *S. japonicum* (Chapters 4 and 5) [4].

The characteristics of the microarray are fully described in Chapter 3; the array contains the largest number of schistosome features compared with previous studies (Table 2.2). This has clearly provided a powerful resource for characterising the schistosome transcriptome [105], which will be expanded upon in Chapters 3-6 of this thesis.

Table 2.2. Summary of microarray elements in previously reported schistosome studies.

(A) *Schistosoma mansoni* cDNA microarray (After Hoffmann *et al.* [110]).
576 features (printed in duplicate)

- 7 blank (spotting buffer only)
- 521 *S. mansoni* PCR amplified ESTs
- 48 Controls
  - 24 Positive controls
    - 4 *S. mansoni* genomic DNA
    - 4 mucin-like protein
    - 4 p48
    - 12 chorion
  - 24 Negative controls
    - 8 yeast tRNA
    - 8 pBluescript DNA
    - 8 lambda DNA

(B) *Schistosoma japonicum* cDNA microarray (After Fitzpatrick *et al.* [15]).
743 features

- 459 probes
  - 233 unknown
  - 1 not in data base
- 6 Positive controls (genomic DNA)
- 278 Negative controls
  - 206 blank (spotting buffer only)
  - 24 yeast tRNA
  - 24 pBluescript DNA
  - 24 lambda DNA

(C) *Schistosoma mansoni* oligonucleotide microarray (After Fitzpatrick *et al.* [111]).
8,160 features

- 7,335 *S. mansoni* oligonucleotides
- 825 Controls
  - 120 *A. thalina*
  - 84 *B. subtillus*
  - 621 buffer/negative controls
- 3,605 gene ontology terms assigned
- 1,242 sequences assigned one or more gene ontology terms
- 476 unique gene ontology terms
  - 249 Molecular function
  - 161 Biological process
  - 66 Cell component (Cellular localisation)

Previous schistosome microarray analysis includes cDNA microarrays described in: (A) Hoffmann *et al.* [110]; (B) Fitzpatrick *et al.* [15]; (C) Fitzpatrick *et al.* [111]. The study conducted by Vermeire *et al*. [108] used the same microarray platform as (C) Fitzpatrick *et al*. [111]. In addition the microarray described by Dillon *et al*. [107] was comprised of ESTs exclusively from lung schistosomula from *S. mansoni* with no detailed description of microarray content.