Chapter 3: Description, Characterisation and Verification of the Schistosome Microarray

3.1 Introduction

As emphasised previously (Chapter 1), the Chinese (SJC) and Philippine (SJP) strains of *S. japonicum* exhibit a number of morphological and other phenotypic differences, including pre-patent period (26 and 28 days, SJC and SJP respectively), tegument topography, adult worm size (SJC longer than the SJP), virulence (SJC is more pathogenic) and the sub-species of snail intermediate host infected [59-61]. Therefore in this study, it was hypothesised that there may be major differences at the gene expression level between the strains of *S. japonicum*. Extensive microarray analysis was undertaken in order to test this hypothesis and to compare gene expression levels in adult male and female worms. A microarray platform was constructed based on oligonucleotide probes, because they are less problematic than cDNAs, being more readily defined (see Chapter 2 [section 2.2.4]). Two extensive EST public datasets were used for *S. japonicum* [25] and *S. mansoni* [58] which provided a practical, reproducible base to design the microarray. This allowed construction of a 22,575 feature 60-mer microarray to investigate gene expression patterns to profile gender-regulated gene transcripts [112] in SJC and SJP.

3.2 Experimental platforms

Microarrays vary in platform size, probe length, platform type, probe construction and cost. Taking into consideration these factors, the 60-mer custom array chip (Agilent) was chosen for the following reasons:

As previously mentioned in Chapter 2 (section 2.6), 60-mers can tolerate more base mismatch in probe to target hybridization than shorter oligonucleotides [72]. In a study by Walker *et al*. [72] the use of long or short oligonucleotides in microarray probe design was investigated. They demonstrated that longer probes may be more suitable for cross-species gene expression [72]. Additionally, studies of species that show increasing divergence from the target organism could benefit from increasing the length of the microarray probe [72]. Hughes *et al*. [69] showed that the in-sequence placement of mismatch between target organism and sample is important and must be taken into consideration. Their study showed that 18 or more random mismatch bases on a 60-mer chip would reduce the signal to background levels [69]. In addition, an average of five or more randomly mismatched bases can reduce the signal to <50% [69]. Taking this into consideration as a comparison to this technology, the Affymetrix 25-mer probes are not as tolerant to base mismatch as the longer probes printed on the 60-mer microarray chip.

The combination of *in silico* design and *in situ* construction techniques employed by Agilent results in an ideal microarray. The *in silico* design has three main advantages in that it avoids duplex formation with non target molecules, it provides the most energetically favorable probes, and enables selection of non-overlapping probes for greater coverage of the transcriptome. Complimenting the *in silico* design the *in situ* bubble-jet system facilitates redesign of a probe in contrast to light directed technologies where a change to one feature would require the redesign of one or more masks [69]. Coupled together, this technology can provide simple future development of the microarray chip.

The Agilent platform is able to be used in many microarray scanners. The platform of the custom microarray is a standard 75 by 25 mm glass slide similar to a

microscope slide. This is important if the microarray platform is to be readily assessable to other research groups, since other platforms such as that produced by Affymetrix require specially produced DNA scanners.

The Agilent platform allows expression analysis using one or two colour microarray experiments. In the comparison of two total RNA samples, two-colour microarrays produce the most accurate results, particularly with small changes in expression levels, because two RNAs may be co-hybridized to the array. It is noteworthy that two colour microarrays may contain dye bias which makes gene expression sometimes misleading. This is because the fluorophore molecules used in most two colour arrays are different sizes [113]. This can be overcome by using a dye-swap replicate and by removing bias in the raw data by linear and 'locally weighted linear regression' (LOWESS) normalisation [114].

For these reasons, together with optimised wash stringency, the accessibility of product technical assistance, glass platform, manufacturing control and reagent quality, the 60-mer custom microarray was ideal for this study.

3.3 Design and construction of the 60-mer microarray

3.3.1 Target sequence probes

The oligonucleotide microarray was designed and constructed using the two available schistosome databases:

- *S. japonicum* EST database (http://schistosoma.chgc.sh.cn/) [25]
- *S. mansoni* Gene Index (http://www.tigr.org/tdb/e2k1/sma1/) [58]

The initial design strategy was to focus on the *S. japonicum* dataset and was later expanded to include sequences from *S. mansoni* providing a broader coverage of the

schistosome transcriptome [4]. The final microarray design consisted of 22,575 features (including 3,354 controls explained in section 3.4.2) (Figure 3.1), making this the largest microarray available for the analysis of the schistosome transcriptome [4]. The 19,221 target sequence probes on the microarray included 12,166 probes derived from *S. mansoni* contigs and 7,055 probes derived from *S. japonicum* contigs, effectively providing a wide coverage of the majority of the schistosome transcriptome (Table 3.1) [4]. The complete layout and details of the microarray probes are presented in a master file in supplementary Table 1, including the designated contiguous sequence (contig) numbers, positions on the microarray, assembled sequences used in design, known sequence identity, homology or annotation and any associated gene ontologies of probes [105]. The combination of the *S. mansoni* and *S. japonicum* probes on one microarray chip provides an extensive base that covers the majority of the schistosome transcriptome [105].

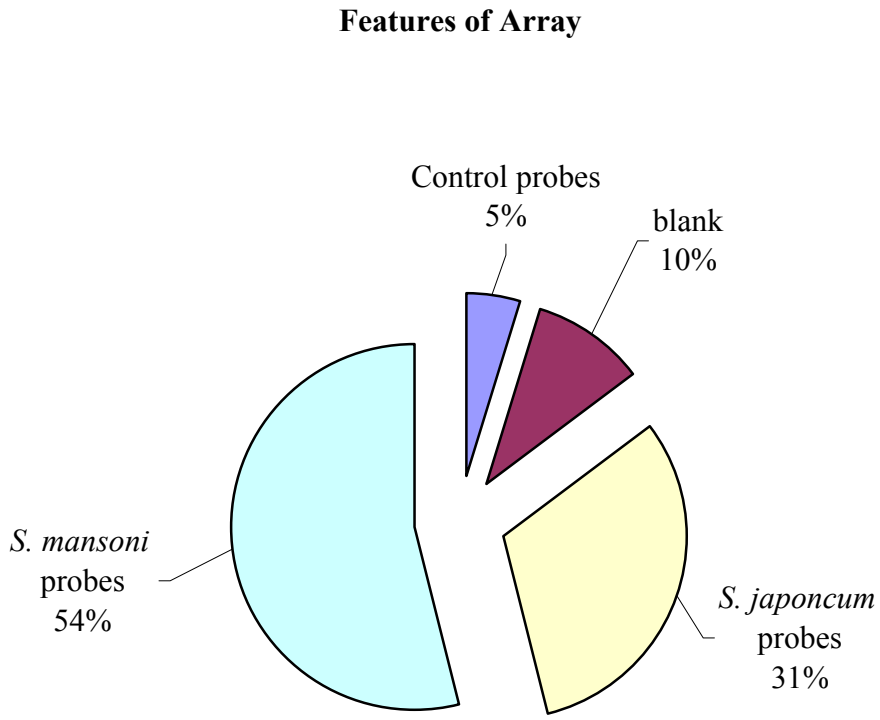

Figure 3.1. A summary of the features present on the schistosome microarray. The percentage of *S. japonicum* and *S. mansoni* probes present on the chip are shown; the probes were designed from the two extensive EST schistosome databases (http://schistosoma.chgc.sh.cn/) [25] and (http://www.tigr.org/tdb/e2k1/sma1/) [58], respectively.

Table 3.1 Summary of the 22,575 features on the 60-mer schistosome microarray.

22,575 features

- 1,080 Controls (not including blank features)
- 2,274 Blank features
- 7,055 *S. japonicum*
  - 5,663 homology annotation
  - 245 *S. japonicum* genes
  - 182 *S. mansoni* genes
  - 19 *S. haematobium* genes
  - 141 *C. elegans*
  - 1,719 *H. sapiens* clones
  - 1,386 *S. japonicum* clone information
  - 838 *M. musculus*
  - 127 *A. thaliana*
  - 88 *D. discoideum*
  - 229 *D. melanogaster*
  - 109 *P. falciparum*
- 12,166 *S. mansoni*
  - 6,581 homology annotation
  - 271 *S. mansoni* genes
  - 372 *S. japonicum* genes
  - 0 *S. haematobium* genes
  - 355 *C. elegans*
  - 1,651 *H. sapiens* clones
  - 666 *M. musculus*
  - 188 *A. thaliana*
  - 205 *D. discoideum*
  - 758 *D. melanogaster*
  - 261 *P. falciparum*
- 1,979 features with associated gene ontology
  - 1,011 Molecular function
  - 975 Biological process
  - 812 Cellular localisation

Further details may be found in the supplementary Table 1 master file (On disk) and in Figures 3.5 to 3.7.

A total of 42,437 target sequences from the two schistosome databases, were masked for vector contamination using an augmented version of the Univec database National Center for Biotechnology Information (NCBI), and simple repeats using 'Repeatmasker' [115]. Candidate probes were then selected from target sequences in unmasked regions with a 3′ bias, which in turn were scored and filtered using empirically derived base composition (BC) parameters. Probe base compositions were scored from one to four. Probes with a BC score of one have a much greater chance of forming a stable and consistent duplex with their intended targets, than probes with a BC score of four, which tend to exhibit hybridization intensities that are not indicative of target sample concentrations [4].

Candidate probes were screened for potential cross-hybridization based upon comparison with a similarity database. The similarity database serves to represent the entire transcriptome of the target species, making it possible to thermodynamically determine the hybridization efficiency of non-target duplex formation. For a given target transcript, candidate probes that showed cross-hybridization potential were removed from consideration if non-cross-hybridizing probes were available. The reconciliation for *S. japonicum* and *S. mansoni* involved conducting a similarity search between the two transcriptomes using a Basic Local Alignment Search Tool, (BLAST) [116], and removing all *S. mansoni* transcripts from the similarity database that showed a 95% similarity, and 80% overlap with *S. japonicum* [4]. Probes designed for the *S. mansoni* EST clusters were designated TC ####, while *S. japonicum* derived probes were designated Contig ####[i]. Annotation and Gene Ontology were provided by the *S. mansoni* [58] and *S. japonicum* [25] EST databases respectively [4].

[i] # = a numerical sequence

3.3.2 Control probes

The custom made microarray contained 3,354 control probes (Figure 3.2) as follows:

- BrightCorner positive control probes were included, located at the array corners, in a pattern that allows unambiguous orientation of the array image. Currently the external probe name used is different for 44 k and 22 k layouts; future layouts will use "BrightCorner". The corner features should always light up brightly to indicate a successful hybridization independently of other performance characteristics such as labelling. Corner probes hybridize to the 10x control targets that were spiked into the hybridization solution.

- The control probes r60 are a set of different probe sequences that correspond to different synthetically generated mRNAs. They may also be used with RNA spike-in controls (not present in the current microarray) to monitor the quality of the experiment (accuracy, sensitivity, dynamic range) [69].

- QC Fusion probes were used for internal 'quality control' purposes to monitor microarray processing, image analysis and experimental quality, an example being to monitor layers in oligo synthesis.

- Pro25G probes, visualised as a zigzag pattern across the array, were used for internal manufacture quality control purposes only. The intensity of these does not reflect either the success or failure of the experimental process.

- COR, CAB and C1 probes (*Arabidopsis. sp.* genes *cor* and *cab* and the *Escherichia coli* gene *C1*) may be used as probes for spike-in controls, or as

negative controls for animal-derived samples. In the current microarray these were used for internal quality control purposes.

- Negative control 3xSLv1 probes designed not to hybridize to targets because of secondary structure. They were used in conjunction with "BrightCorner" probes to identify array corners

- Open features with the probe name: "na" may be used to customise Agilent's commercially available microarrays, or be used for future product development.

Some of the above mentioned control probes were used in conjunction with the program Feature Extraction (Agilent). These may be examined in further detail in the supplementary Table 1 and "Feature Extracted files for microarray experiments" available on disk.

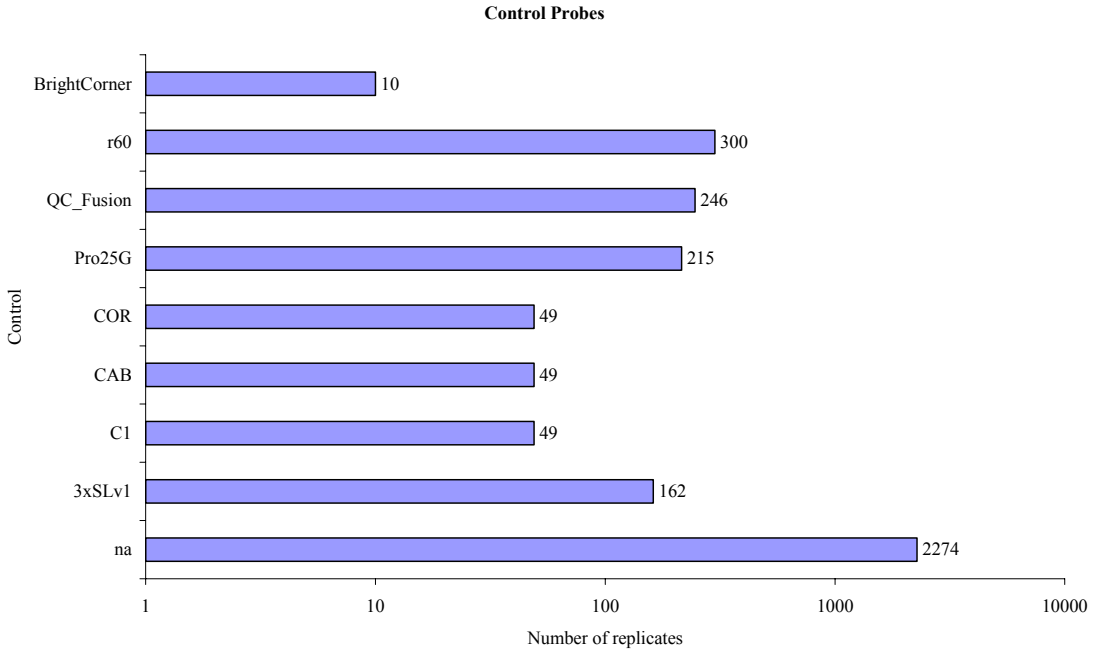

Figure 3.2 Number of control probes present on the schistosome microarray. Control probes include "BrightCorner" probes used to indicate successful hybridization; r60 probes for inclusion of "spike-in controls" (not included in the experiments); QC_Fusion and Pro25G probes for internal manufacture quality control; COR, CAB *Arabidopsis sp.* and C1 *E. coli* genes which can used for "spike-in" or negative controls; 3xSLv1 negative control probes used in conjunction with BrightCorner probes; and "na" which are open features for future development of the microarray chip.

3.3.3 Bubble jet system

The bubble jet *in situ* system used in the construction of the schistosome microarray, unlike standard ink jet printing, does not deposit oligonucleotides on the platform. In phosphoramidite oligonucleotide synthesis an organic linker is covalently delivered to a surface by a ink jet printer head [69, 117]. The coupling step adds a nucleoside to the already present organic linker. After deblocking the previous layer, a new nucleoside is attached and deblocked. This is repeated, attaching and deblocking one nucleoside at a time [117] (Figure 3.3); in this way an oligonucleotide may be built *in silico*. The

advantage with this *in situ* system coupled with the *in silico* design, is that future modification is cheap and efficient, unlike photolithography where 100 masks are required to create a 25-mer oligonucleotide [117].

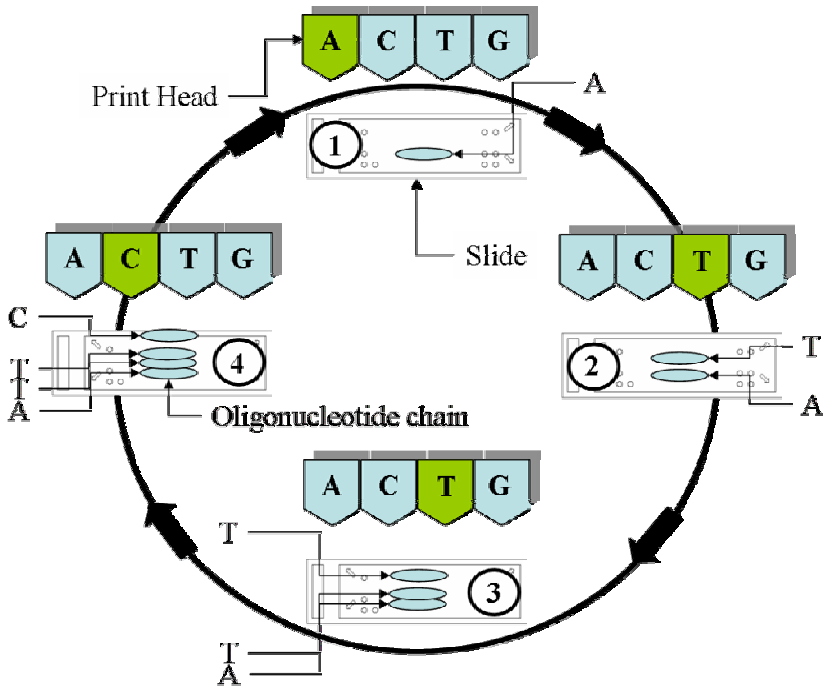

Figure 3.3 Ink-jet printing. The figure shows *in situ* phosphoramidite oligonucleotide chains been synthesised directly on the slides surface. (1) A primary layer of nucleosides, either Guanine (G), Cytosine (C), Adenosine (A) or Thymine (T) are deposited on the surface of the slide. In this example A is deposited following the activation of the respective (highlighted) print head unit. The nucleosides are then deposited on top of each other to form oligonucleotide chains across the chip, an example is (2) T, (3) T and (4) C. This mechanism of oligonucleotide synthesis will continue until the chain is an optimal 60 bases in length [69].

3.3.4 Probe design gene ontology

Annotation and 'gene ontology' (GO) for the microarray were provided by the *S. japonicum* and *S. mansoni* EST databases (Supplementary Table 2). Uncharacterised genomes such as those of both schistosome species will display some redundancy of contigs. This can be demonstrated in Figure 3.4 where probes with the same gene

identification annotations have been grouped to show redundancy within the schistosome microarray. This redundancy by annotation does not always reflect repeated contiguous sequence. All 60-mer probes from the two data sets were unique being selected using a Smith-Waterman algorithm to identify repeats [118]. Since the original annotations were derived by sequence search, the results present similar sequences which may only reflect a part of each contig. This partial identification may be the same for several contigs although the sequence may vary. This redundancy may also be visualised in the ontology based functional categories of the contigs.

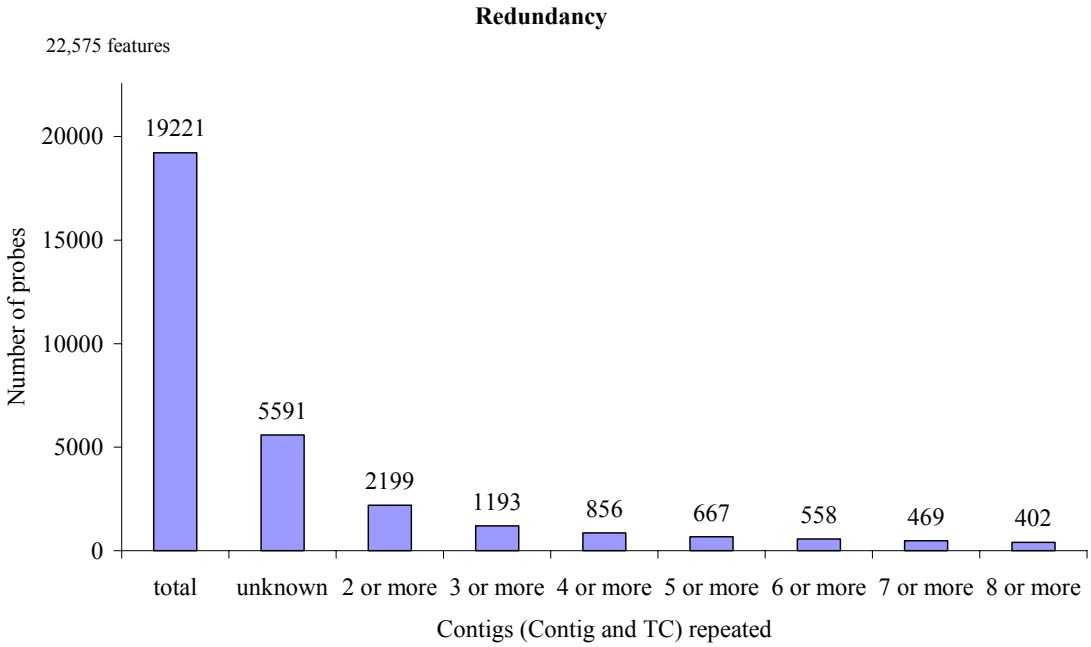

Figure 3.4 Number of probes with repeated annotation on the schistosome microarray. Probes were sorted by gene identification annotation. Probes with the same gene identification were counted and grouped either as 2, 3, 4, 5, 6, 7 or 8 'or more' according to the number of times they were repeated. The redundancy figure does not include control probes or features that contain no putative description.

3.3.5 Gene ontology

The Gene Ontology Consortium (GOC) first announced the construction of three independent ontologies accessible on the world-wide web [119] in 2000 in *Nature Genetics* [120]. This was a joint project of three model databases, FlyBase, Mouse Genome Informatics and *Saccharomyces* Genome Database and it was anticipated that other organism databases would be included in the future [119]. The GOC website has been accessed over 6,000,000 times since 1999 including the incorporation of many more data bases [119]. The incorporated GOs on the schistosome microarray provided putative identification of 1,979 contigs. These putative identifications are a guide to the

possible function of individual probes, the functions of which can highlight probes that are important in the transcriptomes of the two schistosome species. The GO is based on three major categories:

- Molecular function is defined as the biological activity of a gene product and describes such activities as "catalytic" or "binding", at the molecular level [119, 120]. Examples include enzymes, transporters or ligands [120].
- Biological process refers to a biological objective to which the gene or gene product contributes, and describes a series of events accomplished by one or more ordered assemblies of molecular functions [119, 120]. Examples include translation, pyrimidine metabolism or alpha-glucoside transport [119, 120].
- Cellular component (cellular localisation) refers to the place in the cell where the gene product is active, either externally or internally [120]. The cellular component has the proviso that the gene product is part of some larger entity, which may be an anatomical structure (nucleus) or gene product group (protein dimer), and includes such terms as ribosome or proteasome [119, 120].

An overview of the three major categories of gene ontology and prominent sub-categories represented on the schistosome microarray are summarised in Figures 3.5-3.7 and Table 3.1; these include lower level processes as defined by GO (Definitions available at AmiGO http://www.godatabase.org/cgi-bin/amigo/go.cgi).

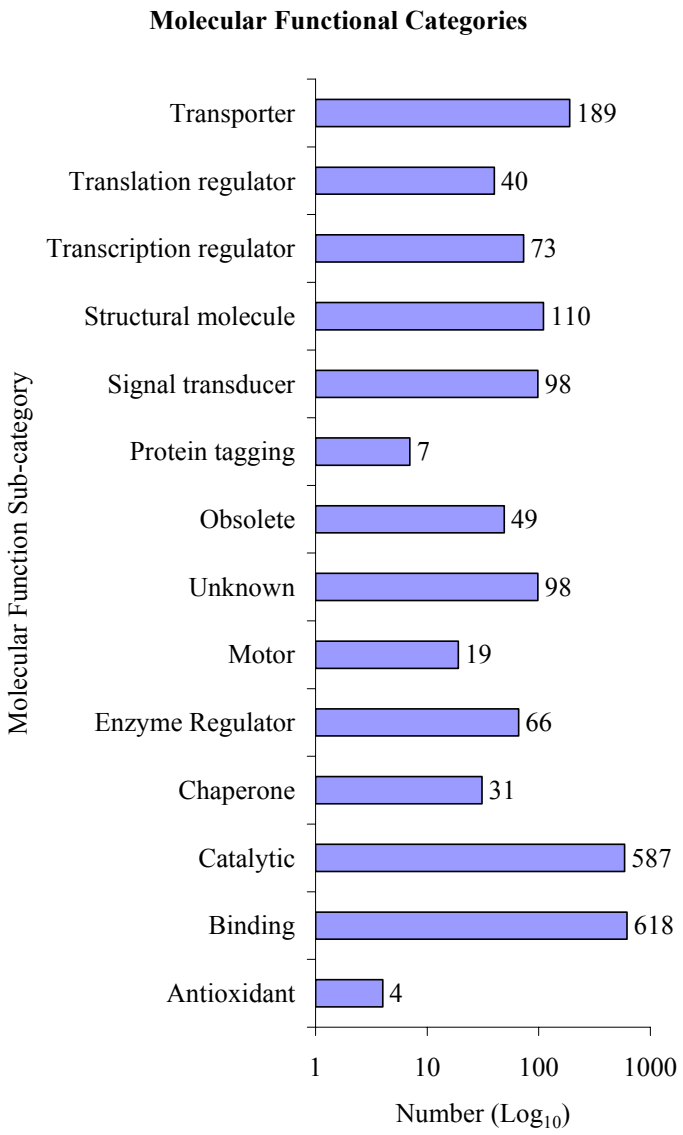

Figure 3.5 Number of genes within the molecular functional sub-categories of the schistosome microarray. Probes that had GO identification were sorted by sub-category of molecular function. The figure shows the number of probes in each sub-category. The probes may have GO identification of one or more sub-categories.

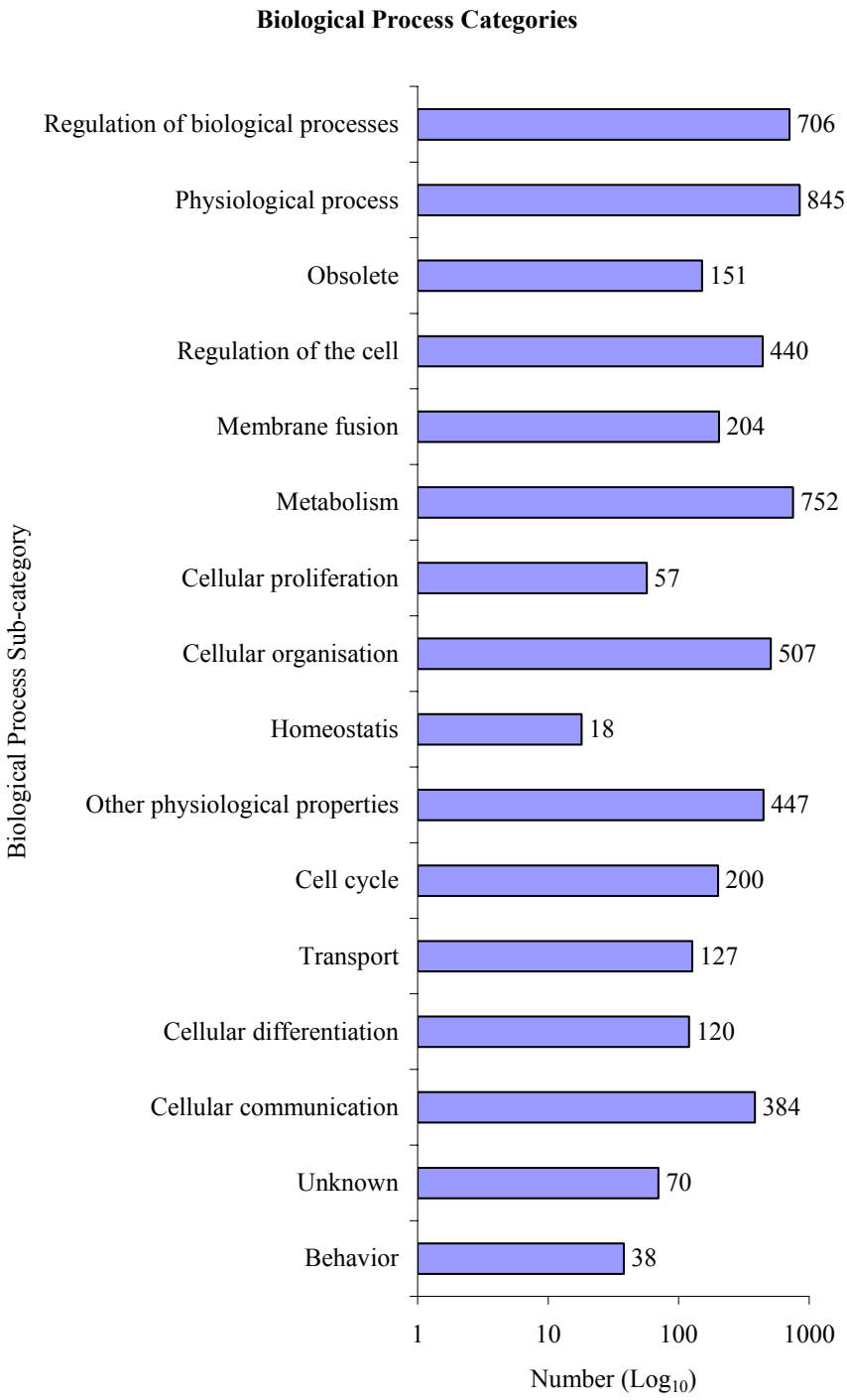

Figure 3.6 Number of genes within the biological process sub-categories of the schistosome microarray. Probes that had GO identification were sorted by sub-category of biological process. The figure shows the number of probes in each sub-category. The probes may have GO identification of one or more sub-categories.

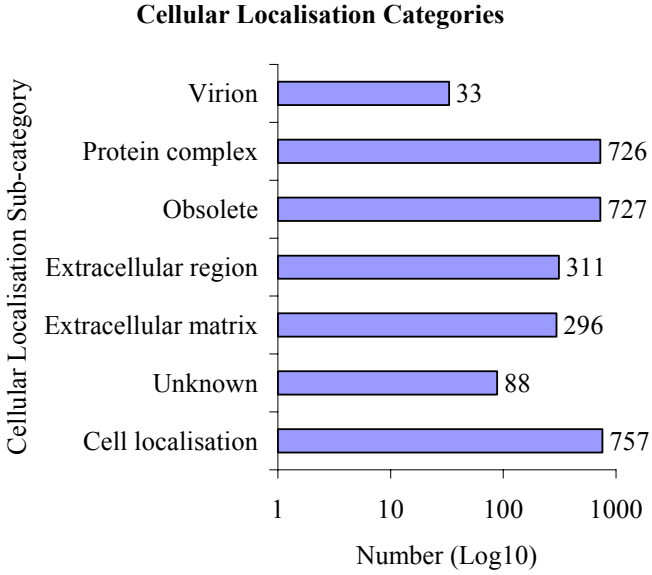

Figure 3.7 Number of genes within the cellular localisation sub-categories of the schistosome microarray. Probes that had GO identification were sorted by sub-category of cellular function. The figure shows the number of probes in each sub-category. The probes may have GO identification of one or more sub-categories.

3.4 Clustering

Gobert G N, McInnes R, **Moertel L P**, Nelson C, Jones M K, Hu W and McManus D P. [105] used the above mentioned microarray to identify important biological functions and hybridization differences between adult worms of *S. mansoni* and *S. japonicum*. This study reported 3,103 and 3,422 probes hybridized to *S. japonicum* and *S. mansoni* targets, respectively, at a p-value of $\leq$0.001 [105]. As shown in Table 3.1, many of the schistosome probes on the schistosome microarray contained no GO description. Through a series of four microarray replicates, probes were clustered by gene-ontology based on hybridization differences between the species. Using this method putative gene ontology was shown for many uncharacterised probes on the array providing a valuable insight into the transcriptome of the schistosome worms [105].

The Feature Extracted 'tab deleted' text files generated from the data from chips 51-54 (Supplementary Feature Extracted files) were transferred along with known GO annotated gene lists (Derived from supplementary Table 2) into GeneSpring 6.1 (Agilent, Santa Clara, USA). The files 51-54 which included 2 dye swaps and a replicate, were combined to generate a gene tree by clustering based on a distance metric (Pearson correlation). Branches of the gene tree included individual clustering of molecular and biological functions (Supplementary Tables 3 and 4).

Based on the assumption that probes that contain similar GOs are co-regulated, the categories produced putative identification of contigs with molecular and biological functions, as outlined in Figure 3.8 and a complete listing in supplementary Table 5. Although it was shown [105] that probes will favorably hybridize to the species of design, there was hybridization from both *S. japonicum* and *S. mansoni*. These results indicate that probes designed on *S. mansoni* may be used for analysis of same species experimentation due to equal sequence mismatch. This analysis may be improved by

the addition of more replicates or different experimental design. If a complete schistosome life cycle comparison was used as the basis of this experiment a gene tree could be generated using quality threshold (QT) clustering or principle component analysis (PCA) [121, 122]. Through QT clustering, complexity of the gene tree will reduce through minimum size and maximum correlation coefficient of each cluster [121]. Similarly, PCA will allow reduction of the complexity of data by discovery of a number of principal components that define most of the data variability, both of which are available in the GeneSpring program [121]. The clustering will additionally be more specific if the analysis is based on differences generated as a result of changing experimental conditions such as drug exposure. For example, gene expression of adult worms of both *S. mansoni* and *S. japonicum* exposed to increasing amounts of praziquantel or artemether (Chapter 1) would provide another variable that would generate a more specific tree. This is because similar genes may be inhibited by the exposure of these or other drugs. The data shown here indicate that gene function may to some extent, be predicted by hierarchical clustering [123] of four replicates, yet can only be confirmed by protein localisation and analysis.

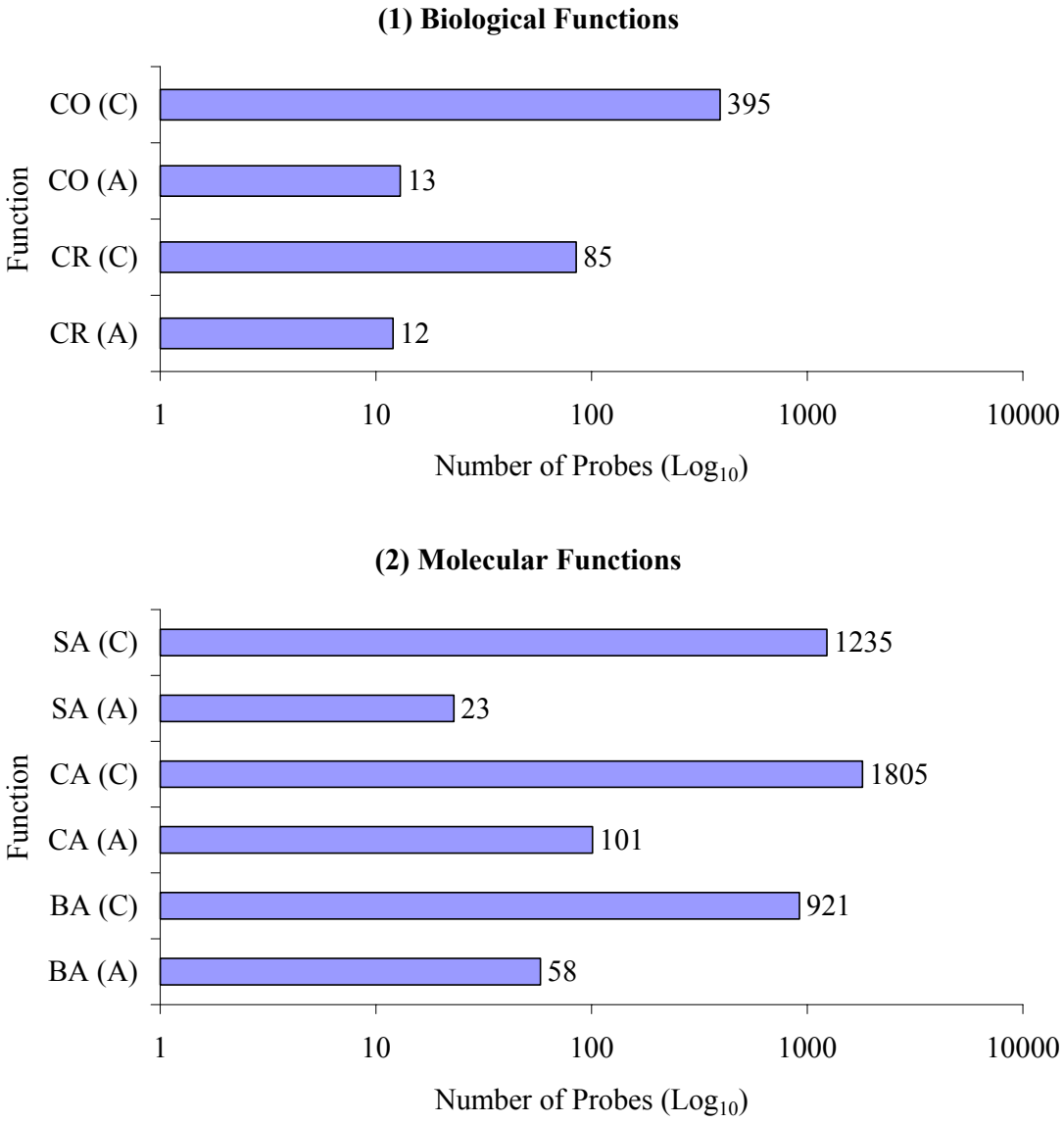

Figure 3.8 Broad gene clustering. Putative assignment of broad gene ontology category was performed by GeneSpring (Version 6.1) hierarchical clustering. (1) Biological functions: Cellular Organisation (CO) and Cellular Regulation (CR); and (2) Molecular functions: Structural Activity (SA), Catalytic Activity (CA) and Binding Activity (BA), clustered (C) from already assigned (A).

3.5 Summary

Chai M, McManus D P, McInnes R, **Moertel L**, Tran M, Loukas A, Jones M K and Gobert G N. [109] used the custom made microarray to investigate the transcriptomic profile of lung schistosomula in comparison to *in vitro* obtained schistosomula and adult *S. japonicum*. There were 3,777 (p-value $\leq$0.001) differentially expressed probes between adult and *in vivo* lung schistosomula. Additionally they showed 6,662 (p-value $\leq$0.001) differentially expressed probes between mechanically-transformed/cultured schistosomula and *in vivo* lung schistosomula [109]. Real time PCR was used as an independent verification of expression of a selected group of over-expressed probes, demonstrating directional regulation of the selected contigs. It was noted that real time PCR results often show greater expression of selected probes, which is common to microarrays probably due to probe saturation and/or non-specific hybridization effects [124]. Although this study would have benefited from using a 'house-keeping gene' for normalisation of the real time results, it was able to show an impressive number of differentially expressed genes between the samples at a high 99.9% confidence interval. The oligonucleotide microarray described here provides an ideal platform for analysis of genomic data from both *S. mansoni and S. japonicum*. Through the use of this microarray the current study was able to analyse reproducible gene expression in both species [4, 105, 109]. Additionally, studies using this platform have greatly expanded on previous published data (Chapter 2) of species, strain and gender-associated differential expression in *S. japonicum* [4, 105, 109], highlighting genes that are important for further understanding its functional biology (Chapters 4-6).