# Chapter 2

# Materials and methods

This chapter contains generic methods including parasite maintenance, molecular procedures, and bioinformatic analyses of sequence data. Details of experimental designs and sample collection specific to each of the chapters are covered in each method section.

## 2.1 Parasite materials

Parasite materials for the *in vivo* timecourse dataset were provided by Prof. Michael Doenhoff from the University of Nottingham and the methods are provided in chapter 3. Parasite materials for the *in vitro* experiment (chapter 4 and 5) were obtained from the Wellcome Trust Sanger Institute (WTSI). All procedures involving mice were performed by authorised personnel according to the UK Animals (Scientific Procedures) Act 1986. The life cycle maintenance was a joint effort between multiple teams and is described below.

### 2.1.1 Maintenance of the *S. mansoni* life cycle

The life cycle at WTSI was initially set up by seeding *Biomphalaria glabrata* snails infected with *S. mansoni* (Puerto Rico stain) provided by Prof. Michael J. Doenhoff (University of Nottingham, UK) and naive *B. glabrata* snails from the Department of Pathology, Cambridge, UK. The parasites were propagated between snails and C57BL/6 or BALB/c mice. Uninfected snails were kept as a breeding colony in a light and temperature controlled room with a 12/12 hour light-dark cycle and an ambient temperature of 28 ºC. Younger snails with sizes between 3-5 mm were collected for infection with miracidia. To perform the infection, one snail was placed into each well of a 24-well plate and up to 30 miracidia, counted and collected under a dissecting microscope, were added to each well. The plates were left at 28 °C for an hour before the snails were placed in a new tank and kept in the same light-dark cycle as the breeding colony. Shedding of cercariae started after 3-4 weeks post-infection. The snails were checked for patent infection and transferred to a dark cupboard reserved

for *S. mansoni* infected snails to maximise shedding capacity and for health and safety reasons.

To collect cercariae, snails were pooled into beakers containing 1X conditioned aquarium water (Appendix A) and left under bright light between 40 minutes to 2 hours. Exposure of intact skin to the cercariae could lead to an infection; therefore, appropriate personal protective equipment including long-cuff gloves, a face shield, and a plastic apron was used while handling the parasites. Following shedding, approximately 250 cercariae per mouse were used for fortnightly infections of mice by intraperitoneal injection, or by subcutaneous infection through tails. After 7-8 weeks, mice were sacrificed and adult worms collected by portal perfusion. Livers from infected mice were processed for egg collection by mincing the livers or by digesting the livers overnight in a collagenase enzyme, and filtering for eggs. Eggs were hatched in sterile water and miracidia collected for infecting snails.

### 2.1.2 Snail husbandry

The infected snails were kept in an aquarium tank lined with a disposable plastic insert, filled with 1X condition aquarium water (Appendix A), aerated, fed three times a week with fish pellets, and kept in a dark cabinet. Uninfected *B. glabrata* were maintained as a breeding colony in the same light- and temperature-controlled room. Small snails (3-5 mm), suitable for infection with miracidia, were maintained separately from the large breeder snails in order to minimise loss during routine tank cleaning and for ease of collecting snails for miracidial infection. To separate breeders from their offspring, egg clutches were regularly removed and transferred to new tanks. I observed that the breeder snails preferentially laid eggs on silicone tubes compared to the surface of the tank. To simplify the removal of egg clutches, I therefore developed “The Octopus” apparatus, comprising a polymer platform holding a number of long silicone tubes hanging on top of the tank into the water (Figure 2.1).

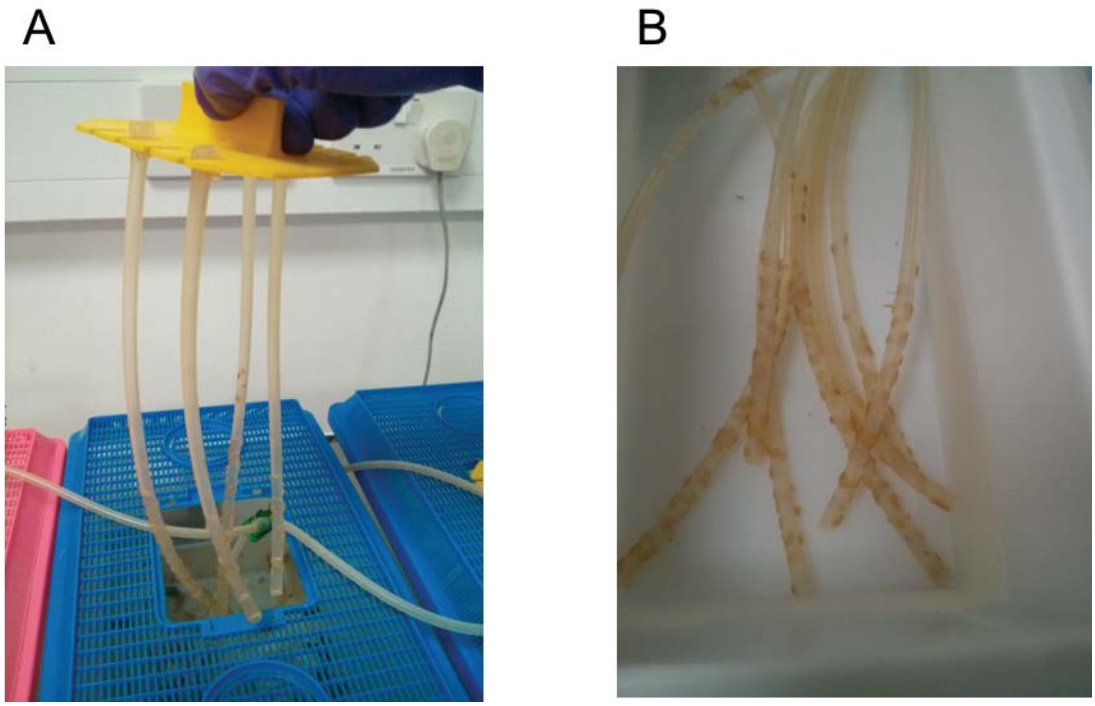


***Figure 2.1 Transfer of snail eggs using "The Octopus"***

A) "The Octopus" holds up to eight silicone tubes which are submerged in water and provide a substrate for snail egg-laying. B) The silicone tubes with snail eggs are transferred to a new tank for hatching and cultivating juvenile snails.

### 2.1.3 Mouse maintenance

The Research Service Facility staff of the WTSI maintained mouse colonies. Mice infected with *S. mansoni* were provided with food and water *ab libitum* and were checked daily for general health.

## 2.2 Molecular methods

### 2.2.1 RNA extraction from parasites

RNA was extracted from parasite materials with a modified phenol-chloroform method followed by column purification. Frozen samples in TRIzol® reagent (15596-026, Invitrogen) were thawed on ice. Next, the samples were resuspended by gently pipetting and transferred to 2 ml tubes containing ceramic beads (MagNA Lyser Green Beads, Roche). The parasite materials were homogenised in a MagNA Lyser Instrument (FastPrep-24) at maximum speed for 20 seconds twice, with 1 minute rest on ice in between. After this, 200 µl of chloroform-isoamyl alcohol 24:1 was added to each tube, followed by vigorous shaking by hand for 5 seconds. The tubes were

centrifuged at 13,000 x g for 15 minutes at 4ºC to separate the aqueous and organic solvent layers. The top aqueous layer containing RNA was carefully transferred into a new RNase-free 1.5 ml tube. To these tubes, an equal volume of 100% ethanol was added and mixed by pipetting. The mixture was then transferred to Zymo RNA Clean & Concentrator-5 column (R1015, Zymo Research) and processed according to the manufacturer's protocol. Finally, 15 µl of RNase-free water was added to the column and centrifuged for 30 seconds to elute the RNA. The elution was repeated once using an additional 15 µl of the RNase–free water.

### 2.2.2 RNA extraction from human cells

In addition to cells which were part of the co-culturing experiments, spare cells of each cell type were cultured and collected to use as test samples for RNA extraction. This was not required for parasite materials as standard protocols were available in our laboratory. The yield of RNA from the test extractions ranged between 2.7 to 34 µg which exceeded the recommended amount for the Zymo Clean & Concentrator kit (5 µg). Therefore, QIAGEN RNAeasy Plus Universal mini kit (QIAGEN RNAeasy Plus Universal mini kit , QIAGEN), recommended for the extraction of up to 100 µg of RNA was used. The extraction followed the manufacturer's protocol with minor modifications. Two hundred µl of chloroform-isoamyl alcohol 24:1 was used instead of 180 µl, and 100% ethanol was used instead of 70%. The elution was done twice, each time in 30 µl nuclease-free water. Batches of 11-12 samples were processed

### 2.2.3 Measurements of RNA concentration and purity

Extracted RNA was measured for its concentration and integrity using a Agilent RNA 6000 Nano kit (5067-1511, Agilent Technologies), and assessed for its purity using a NanoDrop spectrophotometer. The integrity of the RNA was inferred from the presence of the distinctive 18S rRNA peak for parasite samples, and two rRNA peaks (18S and 28S) for human samples; broad peaks indicate that the RNA is degraded. For parasite samples, an extra peak of 28S rRNA (approximately 4700 nucleotides (nt)) infer contamination with mouse tissue (Figure 2.2). RNA contaminated with mouse tissue was commonly observed for *S. mansoni* schistosomule samples collected at day 6 post-infection. The source of the mouse tissue contamination was an artefact of the collection method. Minced lung tissues were incubated and filtered to obtain lung-stage schistosomules. Some of the minced lung tissues passed through

the filter and was collected with the schistosomules. Detailed methods for the collection of lung schistosomules are presented in chapter 3. The purity of RNA was determined from the 260/230 and 260/280 ratios measured by the Nanodrop spectrophotometer. Ratios of ~2.0 indicated pure RNA and lower ratios suggested contamination with chemical or biological molecules.

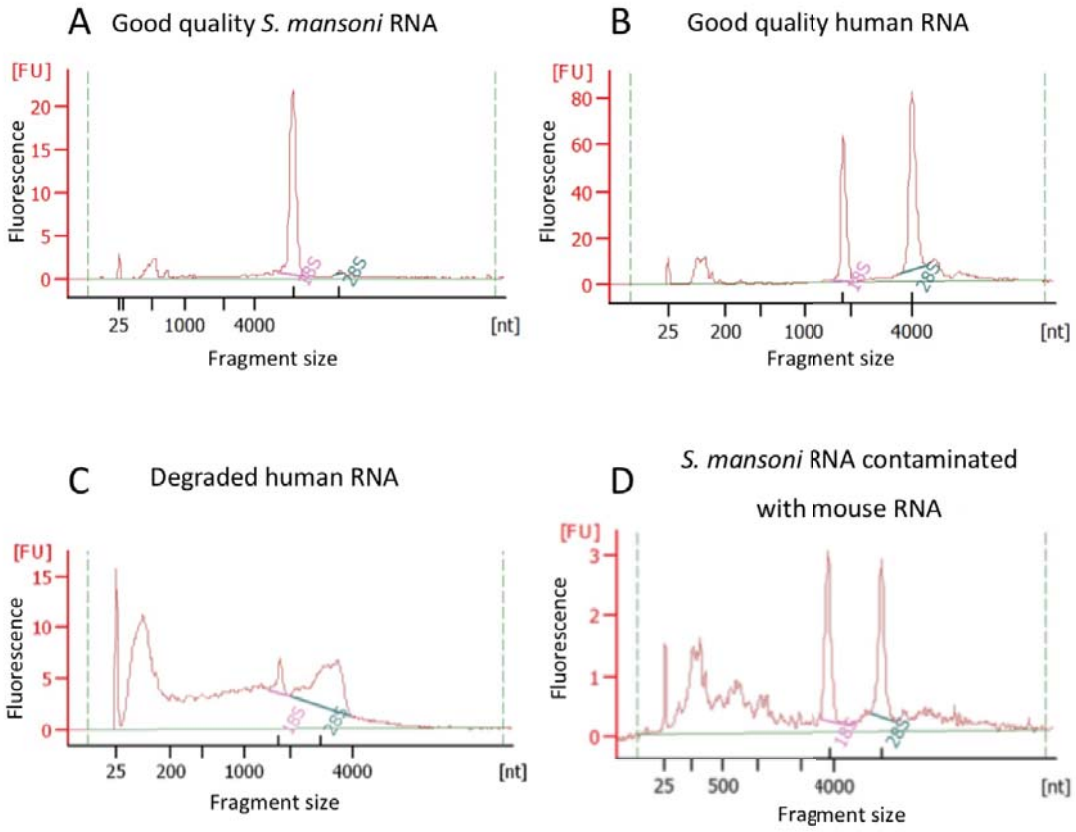


***Figure 2.2 RNA qualities assessed from Agilent Bioanalyzer electropherograms***

A) Electropherogram from Agilent Bioanalyzer output showing good quality RNA from *S. mansoni* with little or no degradation inferred from a shape peak at approximately 4000 nt representing schistosome 18S and nicked 28S ribosomal RNA (Tenniswood and Simpson, 1982). B) Good quality RNA from a human sample. C) An example of degraded RNA inferred from the absence of a sharp rRNA peak and the shift of RNA abundance towards shorter length. This example is from human RNA in a separate experiment. D) Good quality RNA from *S. mansoni* but with contamination of mouse RNA.

### 2.2.4 Library production and sequencing

All RNA-seq libraries were produced by the library production team at WTSI. Following quantity and quality checking, variable amounts of total RNA were sent for

library preparation. For chapter 3, up to 1 ug of RNA was used; for chapter 4, the entire RNA sample was used (range between 319 to 4,350 ng); and for chapter 5, 500 ng of RNA was required.

The libraries were produced using the Illumina ®TruSeq™ Stranded RNA Sample Preparation v2 Kits (RS-122-2101 and RS-122-2102, Illumina). Briefly, mRNA was pulled down from total RNA using oligo-dT beads. The mRNA was then fragmented into 200-300 base pairs fragments, and reverse-transcribed into cDNA using random hexamer primers and free nucleotides. The second strand synthesis was performed similarly but replacing dTTP with dUTP, which is essential to mark the second strand for quenching during the amplification. To improve the binding efficiency of an adapter with a T-overhang, an additional dATP was incorporated to the 3' end of both strands. Different adapter index sequences were incorporated into samples to allow multiplexing of samples. After adapter ligation, a PCR reaction was performed with primers specific to the adapter sequences that also contain the Illumina adapter fork region. The PCR reaction was completed for 10-14 cycles followed by library clean up using Agencourt AMPure XP Beads (A63881, Beckman Coulter). The libraries were then quantified by qPCR and a suitable amount was loaded onto a sequencing lane.

The sequencing was performed at the WTSI by the sequencing facilities on an Illumina HiSeq 2500 platform using either rapid run (chapter 5) or normal run (chapter 3 and 4) modes. All sequencing data was produced as 75 bp paired-end reads.

### 2.2.5 Overall QC of sequencing outcome

Read outputs from the sequencing step were quality assessed using various parameters in an automatic standard pipeline managed by core sequencing informatics (NPG team), WTSI. Of relevance to this thesis is the mapping of reads to a set of reference genomes. A small randomly selected subset of reads were mapped to reference genomes from model organisms, pathogens, and other microbes. List of the genomes is available in Appendix B. These data were used to indicate possible contamination in the samples.

## 2.3 Data analysis

### 2.3.1 Mapping and quantifying read counts

#### *2.3.1.1 Schistosome reads*

Schistosome reads were mapped to the *S. mansoni* genome version 5 (GeneDB, Logan-Klumpler *et al*., 2012) using TopHat (Kim *et al*., 2013, version 2.0.8) with default parameters except the following: -g 1 (only report 1 alignment for each read); --library-type fr-firststrand (for dUTP Illumina library); -a 6 (minimum anchor length); -i 10 and --min-segment-intron 10 (minimum intron length); -I 40000 and --max-segment-intron 40000 (maximum intron length); --microexon-search (find alignment to micro-exons). The resulting BAM files of accepted hits were sorted using SAMtools (by read names; -n option), indexed, and used to obtain read counts per gene. A GFF file containing gene annotations and their genomic coordinates was downloaded from GeneDB.org and filtered to keep only the longest transcript for each gene. The number of read counts per gene was calculated using HT-Seq (Anders *et al*., 2015, version 0.7.1) based on the GFF file. Next, the GFF file and sorted BAM files of mapped sequencing reads were used as inputs for HTSeq to obtain read counts per gene and used for analysis in R. HT-seq was run with default parameters except with strand option set to suit dUTP libraries (-s reverse), and alignment score cut-off increased (-a 30).

#### *2.3.1.2 Human reads*

In order to reduce computing power required for the mapping step, human reads were mapped to a reference transcriptome instead of to a reference genome (transcriptome version GRCh38, release 88 (Ensembl, Aken *et al*., 2016)). The mapping and quantification of reads per transcripts was performed using Kallisto (Bray *et al*., 2016, version 0.42.3). Kallisto uses pseudoalignment which match k-mers from each sequence read to reference transcripts, bypassing alignment by individual bases and by doing so, reducing computational power and time required. The output is a table of reads per transcript. Read counts per transcript were converted into read counts per gene by matching reference identifiers from Ensembl (Aken *et al*., 2016). The read counts per gene were used as an input for gene expression analysis in R software environment (R Core Team, 2016).

### 2.3.2 Read count and differential expression analysis

Analyses were performed using RStudio version 0.99.489 (RStudio Team, 2016), with R version 3.3.1. Versions of R packages used for analyses and data visualisation are listed in Appendix C. DESeq2 (Love *et al*., 2014) was used to import read counts into the analysis environment (function: 'DESeqDataSetFromHTSeqCount'), to investigate overall transcriptomic differences between samples using principal component analysis (PCA) (function: 'plotPCA'), and to identify differentially expressed genes in the timecourse and in pair-wise comparison. PCA used regularized log–transformed read count data as input. Read counts were normalised based on negative binomial distribution and scale factors. Differential expression analyses were performed with either likelihood-ratio tests (when the whole timecourse was considered) or with the Wald test (when used with pairwise comparisons) and returned adjusted p-values according to the Benjamini–Hochberg procedure to control false discovery rate (Benjamini *et al*., 2001). Differentially expressed genes were defined as those with adjusted p-value < 0.01 and $\log_2$ fold change ($\log_2$FC) in expression > 1 or < -1 (i.e. 2-fold change) for chapter 3, and 5; or adjusted p-value < 0.01 and $\log_2$FC in expression > 0.5 or < -0.5 (i.e. 1.4-fold change) for chapter 4.

### 2.3.3 Genes clustering by timecourse expression profile

Genes were clustered based on their expression profiles over the timecourse using self-organising maps constructed in the R package Kohonen (Wehrens and Buydens, 2007). The recommended inputs of mean-normalised, regularized log-transformed counts (rlog-transformed) were used for the self-organising maps. Rlog transformation is a robust transformation for stabilizing variances within genes especially when library sizes (size factor) vary between samples (Love *et al*., 2014). The library sizes between some samples varied up to 10-fold. To mean-normalise the regularized log-transformed counts, the mean value of a row was subtracted from each value within that row so that changes in expression fluctuated around zero and clustered based on expression pattern rather than absolute expression level. This normalised value was used to calculate means of replicates for each gene at each time point and used as input for clustering. Genes were grouped based on their expression pattern into 96 clusters (user defined). To reduce noise, only genes that were differentially expressed in at least one time point (likelihood ratio test, adjusted p-

value < 0.01) were used as inputs for clustering. Self-organising map outputs from the package Kohonen provide representative values (codebook vector) for each cluster which represent changes over time of genes in the cluster. This information was used to produce a dendrogram so that similarity between clusters could be used to find genes with similar expression patterns over the time points. Hierachical clustering was used to group clusters based on their representative values.

### 2.3.4 GO term enrichment

Gene Ontology (GO) term enrichment (biological process terms) was performed using the TopGO R package (Alexa and Rahnenfuhrer, 2016). The test used a *weight* algorithm and Fisher's exact test statistic to identify enrichment of GO terms among the input genes. To determine enrichment, all *S. mansoni* genes were used as a reference background. For the human dataset, expressed genes were used as a reference background. This was because each of the human cell types showed very distinct transcriptomic profiles. Separate lists of expressed genes were produced for each of the three cell types used. Expressed genes were genes that have FPKM > 0 in at least one replicate in a particular cell type. FPKM is Fragments Per Kilobase of transcript per Million mapped reads, and was calculated as

$$\frac{number\ of\ reads\ of\ a\ gene}{\left(\frac{transcript\ length}{1000}\right) \times \left(\frac{total\ number\ of\ reads\ in\ the\ library}{1{,}000{,}000}\right)}$$

GO annotations for *S. mansoni* were downloaded from GeneDB (January 2016). For human, GOslim annotations were downloaded from Ensembl GRCh38, release 88 (Ensembl, Aken *et al*., 2016) and supplemented with genes annotated for innate immune function from InnateDB (Breuer *et al*., 2013). To do this, GO:0002376 (immune system process) was added to the genes annotated with innate immune functions by InnateDB.

### 2.3.5 Pathway enrichment and pathway network

Pathway enrichment was used to provide another layer of specificity for deducing functional changes. Human transcriptome analysis benefits from richer pathway information and supporting evidence that is available due to the large size of the human-research community. Therefore, I used pathway databases for human

transcriptome analysis but not for schistosome transcriptome analysis because pathway descriptions in schistosomes primarily rely on homology based mapping to infer pathways from other species (Fabregat et al., 2016; Kanehisa et al., 2017a; Mi et al., 2017). It cannot necessarily be assumed that pathways in parasites follow those of better-characterised model species.

Pathway enrichment analysis were generated through the used of three databases - Reactome, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Ingenuity Pathways Analysis (IPA) (Fabregat et al., 2016; Ingenuity® Systems; Kanehisa et al., 2017b) Knowledge of human functional genomics, although growing rapidly, is still limited in scope. Hence combining information from various updated sources can provide further affirmation. The InnateDB online server was used for pathway analysis with the KEGG database. Reactome and IPA pathway analyses used their respective interfaces. Input to the analysis was a list of differentially expressed genes from the human dataset, given as Ensembl gene identifiers.

Many pathways from within the same database and between databases often shared similar genes, and pathways are sometimes named differently between databases. To incorporate this information, the enrichment outputs from the three databases were integrated by drawing a network of enriched pathways, linked by shared genes. To do this, each enriched pathway was compared to every other enriched pathway, and if there was at least one gene (differentially expressed input gene) in common between them, then the two pathways were linked on the network. The network was visualised on Cytoscape (Cline *et al*., 2007). Each network node represents a pathway, node colours represent sources of the enrichment analysis. Having edges between nodes indicated that the nodes shared at least one gene. A small distance between nodes indicated a large number of shared genes.

### 2.3.6 Pathway comparison between cell types

Changes of gene expression in two pathways were compared between cell types. The two pathways were chosen because of their occurrance in multiple pathway enrichment analyses, and their relevance to *S.* mansoni infection biology. These were the *extracellular matrix organisation* pathway and *coagulation and complement* pathway. The list of genes in the *extracellular matrix organisation* pathway was

obtained from the Reactome database (pathway identifier: R-HSA-1474244). Genes in coagulation and complement pathways were obtained from KEGG (pathway identifier: hsa04610). Gene names from the database were converted to ENSG identifiers. Changes in gene expression for each cell type were compared based on $log_2$FC of co-cultured*vs.*worm-free pairwise comparison. Only genes that were expressed (FPKM > 0) in at least one replicate for each cell type were included in the comparison. $Log_2$FC values were displayed only for genes with an adjusted p value of less than 0.01.

### 2.3.7 Protein structural prediction

I-TASSER online server (v5.0) (Roy *et al*., 2010; Yang *et al*., 2015; Zhang, 2008) was used to predict protein 3D structure from amino acid sequences, using default parameters. Amino acid sequences were obtained from GeneDB in November 2015. TM-scores indicate similarity between two structures. The value ranges between 0-1 with a higher value inferring better match.

### 2.3.8 Protein domains and motif search

InterProScan online server (v60 and v61) (Finn *et al*., 2017) was used to identify protein domains from amino acid sequences. Amino acid sequences were obtained from GeneDB in December 2015. SignalP 4.1 (Petersen *et al*., 2011) was used to resolve conflicts between transmembrane and signal peptide predictions from InterProScan (using Phobius predictor; Käll *et al*., 2004). MyHits (Pagni *et al*., 2004) and ScanProsite (de Castro *et al*., 2006) were used to identify functional motifs in amino acid sequences. CathDB (Sillitoe *et al*., 2015) was used to explore protein structural domains and to search by structural match.

### 2.3.9 Gene phylogenetic tree

Information on homologues, orthologues and paralogues of genes was obtained from WormBase ParaSite release 8 and 9 (Howe et al., 2016a), which uses the Compara pipeline to create gene trees (Vilella *et al*., 2009).

### 2.3.10 Artemis and BamView

Mapping of RNA-seq reads to genomic regions was visualised using BamView (Carver *et al*., 2010) within Artemis (Carver *et al*., 2008) to assess the accuracy of

gene models. Short read mappings came from *S. mansoni* data produced in this thesis. Long-read mappings (ISO-seq, or isoform sequencing) were obtained from a separate project of the Parasite Genomics team. In the project, *S. mansoni* RNA were processed into full-length RNA libraries and sequenced on a PacBio platform (Rhoads and Au, 2015) to obtain long reads covering full-length isoforms. Screenshots of long-read mapping were provided by Alan Tracey, Parasite Genomics team.

### 2.3.11 Seaview sequence alignment

Multiple sequence alignment was performed using SeaView version 4 (Gouy *et al*., 2010) with ClustalW alignment program (Sievers *et al*., 2014). Inputs were amino acid sequences from the reference *S. mansoni* genome (obtained December 2015) or from GenBank accession identifiers resulting from Basic Local Alignment Search Tool (BLAST).

## 2.4 Data presentation

R package ggplot2 (Wickham, 2009) was used to produce most plots and graphs in this chapter including PCA plots, volcano plots, gene expression dot plots, and graphical representation of GO term enrichment. PCA plots and gene expression dot plots used output from DESeq2 'plotPCA' and normalised counts respectively. Volcano plots used pairwise differential expression as input. Gene clustering expression graphs used generic functions in R. Clustering dendrograms were created using R package APE (Paradis *et al*., 2004) using hierarchical clustering of cluster representative values (codebook vector). Heatmaps were generated with R package pheatmap (Kolde, 2015) using rlog transformed values of genes as input.