CHAPTER 2

MATERIALS AND METHODS

2.1 Reagents

2.1.1 Aquarium water 10x (also known as Lepple water 10x)

- 3.78 mM calcium chloride dihydrate
- 5.00 mM magnesium sulphate heptahydrate
- 0.25 mM potassium sulphate
- 5.00 mM sodium carbonate anhydrous
- 0.002 mM ferric chloride

Aquarium water was diluted 1/10 in prior to use.

2.1.2 Supplemented DMEM

To obtain 500 ml of supplemented DMEM:

- 490 ml of high glucose Dulbecco's Modified Eagle's medium (DMEM) (Sigma, U.K.)
- 5 ml of penicillin–streptomycin solution (10,000 U, 10 mg streptomycin/ml; Sigma, U.K.)
- 5 ml of L-glutamine (200 mM, Sigma, U.K.)

2.1.3 Percoll solution

To obtain 10 ml of 70% Percoll solution:

- 7.0 ml Percoll (Sigma, UK)
- 0.6 ml sodium chloride 1.5 M
- 2.4 ml high glucose DMEM (Sigma, U.K.)

Percoll solution was kept in ice for at least 15 minutes prior to use.

2.1.4 Growth media

- Supplemented DMEM (see 2.1.2)
- 10% foetal calf serum (FCS) (PAA, UK)
- 1% Hepes buffer (PAA, UK)

2.2 Parasite material

In order to routinely obtain parasite material, part of the life cycle of *S. mansoni* is reproduced in the laboratory (Schistosomiasis Research Group, Dept. of Pathology – University of Cambridge, UK) using *B. glabrata* snails. *S. mansoni* (NMRI strain of Puerto

Rican origin) eggs were kindly provided by Prof. Michael J. Doenhoff (University of Nottingham, Nottingham, U.K.). Miracidia were allowed to hatch in aquarium water and were separated phototropically[1]. *B. glabrata* snails were infected with 2-6 miracidia each and kept in the dark for five weeks prior to shedding cercariae phototropically.

SAFETY NOTE: cercariae are the infected stage for the human host. Contact with skin may result in infection. Hence, protective clothes, gloves and goggles must be worn throughout all experimentation procedures that involved live parasitic material.

2.2.1 Collection of cercariae

Infected *B. glabrata* snails were kept in aquarium water in 40 cm x 15 cm tanks inside light-protected cupboards at constant room temperature of 28°C. To obtain freshly shed cercariae, snails (typically 30-40) were placed in small beakers in approximately 40 ml of aquarium water and exposed to the light for 1-2 hours. Cercariae are concentrates by pooling all water from beaker glasses and snails are returned to their tanks and placed in the dark. Approximated number of cercariae was estimated by counting individual in three aliquots of 10 ul each. Live cercariae were placed in 50 ml conical tubes and allowed to cool down on ice for 30 minutes. Then, cold cercarial suspensions were centrifuged at 1000g for 10 minutes and the supernatant was discarded. In order to preserve RNA, 1 - 2 ml of RNA*later* (Ambion, UK) were added to the cercariae pellet and stored at -80°C until the sample was used for RNA extraction. If cercariae were used to obtain skin-transformed schistosomula, they were kept at 28°C until used for not more than one hour. Optimal numbers of cercariae were obtained when snails were exposed to light with no less than 2 days between exposures.

2.2.2 Mechanically transformed schistosomula

Mechanically transformed schistosomula were obtained using a modified version of the protocol used by Brink *et al.,* (1977). Freshly shed cercariae, still in aquarium water, were cooled down on ice for 30 minutes, centrifuged at 1000g for 10 minutes at 4°C and then resuspended in 10 ml of supplemented DMEM. In order to induce tail detachment, cercariae were shaken vigorously for approximately 30 seconds using a vortex mixer and then subjected to 13-15 passages through a 21G syringe needle. Then, the parasite

[1] The beaker containing miracidia is placed under a source of light (lamp) and its walls covered with aluminium foil. Miracidia are phototropic and would swim towards a source of light and concentrating in the upper layers of water.

suspension was carefully placed on top of 10 ml of ice-cold Percoll solution (see 2.1.3) in 15 ml conical tubes. These were centrifuged at 4°C for 10 minutes at 1000g producing the separation of tails (top) and cercarial heads/schistosomula (bottom). Each fraction was placed in individual tubes and washed 3 times in supplemented DMEM. After the last wash step, tails' supernatant was discarded and 1 ml of TRIzol reagent (Invitrogen, UK) was added to the samples. These were stored at -80°C until RNA extraction. The schistosomula preparations were incubated at 37°C and 5% CO_2 for either 3 hours or 24 hours in growth media (see 2.1.4). After incubation period was completed, parasites were transferred to 15 ml conical tubes and centrifuged at 1000g for 5 minutes, supernatant was discarded and schistosomula were resuspended in 1 ml of TRIzol reagent (Invitrogen, UK) and stored at -80°C until RNA extraction.

2.2.3 Skin transformed schistosomula

Skin transformed schistosomula were obtained using a modified version of the protocol published by Clegg *et al.,* (1972). By allowing the cercariae to naturally penetrate through a layer of freshly excised mouse skin, the authors mimicked the transformation of cercariae into schistosomula.

2.2.3.1 Ethics statement

The procedures involving animals were carried out in accordance with the UK Animals (Scientific Procedures) Act 1986 and as authorised on personal and project licences issued by the UK Home Office.

2.2.3.2 Protocol:

For each experiment, a total of six mice were used. Mice were killed with an overdose of anaesthetics (followed by cervical dislocation) according to Home Office regulations. Hair was removed form the abdominal and dorsal skin areas using clippers and skin was later excided from the animal using dissecting scissors. Each animal provided an area of skin of approximately 7.5 cm^2; which was divided into two halves. Gel-like dermal tissue was removed by rubbing the skin gently (for approximately 5 minutes) with sterilized gauze soaked in supplemented DMEM. The organization of the transformation apparatus is presented in **Figure 2.1A.** Tube B of the assembly was filled with supplemented DMEM containing 2% FCS and one half of prepared skin was mounted covering the opening of tube B with the dermal side facing downwards. Tube A was placed above Tube B with a rubber O-ring in between. All pieces were kept in place by holding both tubes with metal clips (**Figure 2.1B**). The skin surface was washed three times with aquarium water and was then checked for leaks. All assemblies were placed in a water bath pre-warmed at 37°C; the lower part of the assembly (Tube B) was constantly kept at this temperature (**Figure 2.1C**). Five ml of aquarium water were placed in Tube A of each assembly to maintain the skin moist. The experiment was carried out in a room with controlled temperature of 28°C, resulting in Tube A being kept at this temperature. After 1 hour, the assemblies were taken out from the water bath and any air bubbles found in Tube B were removed by carefully raising the skin layer and replacing the air bubbles with 2% FCS supplemented DMEM. Then, the apparatus is assembled again and placed back in the 37°C water bath. Approximately 12,000-14,000 freshly shed cercariae kept in aquarium water were placed in each assembly. Schistosomula were harvested from Tube B after 3 hours of application of the cercariae and the schistosomula produced in each assembly were checked under the microscope. Samples with less than 4% tails/cercariae contamination were kept and only these were pooled and incubated at 37°C and 5% CO_2 for 21 hours. After incubation time was completed, schistosomula were centrifuged at 1000 g for 10 minutes and kept at -80°C in 1 ml TRIzol.

2.2.4 Schistosomula evaluation

Schistosomula preparations were evaluated using a Leica DM 1L inverted microscope (Leica, Milton Keynes, Bucks, UK) at 10x or 40x. The criterion used for evaluating "healthy" parasites is the one used in Mansour *et al.*, (2010). Briefly, parasites are catalogued as:

- Unaffected parasites are generally translucent in the inverted microscope and, depending on the life cycle stage, show typical worm-like movements
- Dead parasites are opaque and immobile
- Damaged parasites show a granular appearance and little movement; they may acquire a range of shapes.

2.2.5 Adult worms

Prof. Phil LoVerde (University of San Antonio, Texas, US) kindly provided seven-weeks old male and female adult worms from which RNA was extracted and libraries prepared as explain in 2.3.1 and 2.4 respectively.

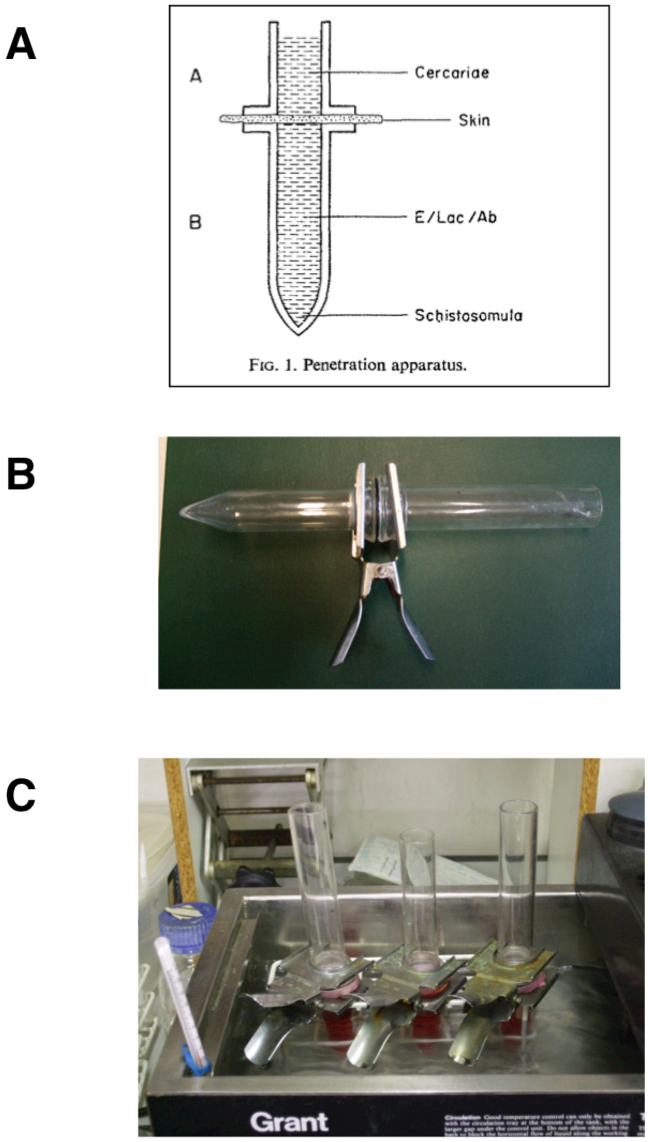

Figure 2.1 - Diagram and photographs of the skin transformation assemblies. A - Graphical representation of a transformation assembly (reproduced from Clegg *et al.,* (1972). B – Photograph of one of the transformation assemblies prior to use. C – Transformation assemblies in use during an experiment using only 3 assemblies. The lower part of the assembly is placed in a water bath with a constant temperature of 37°C while the upper part is left at a room temperature (28°C).

2.3 Molecular Biology and Biochemistry Techniques

2.3.1 RNA extraction

Total RNA from parasite material was extracted using TRIzol (Invitrogen, UK) according to manufacturer specifications with the exception of cercariae samples, where a modified TRIzol (Invitrogen) / RNeasy (Qiagen, UK) protocol (Hoffmann *et al.*, 2002) was used instead. After extraction, RNA quality was assessed using an Agilent RNA 6000 Nano - Bioanalyzer and quantified using a NanoDrop ND-1000 UV-Vis spectrophotometer.

2.3.2 Sodium acetate/isopropanol precipitation of RNA

This protocol was used mainly with two objectives: concentration of RNA samples and/or cleaning of RNA sample, usually from phenol/ethanol contaminants.

To the RNA solution, the following reagents were added:

- 1/10 volume of sodium acetate 3M pH 5.2-5.3 (Ambion, UK)
- 2.5 volume of 96-100% ethanol (Sigma, UK)
- 1ul glycogen (5mg/ml) (Ambion, UK).

The mixture was shaken vigorously and incubated for a minimum of 1 hour at -20°C (maximum 16 hours – overnight incubation). Then, precipitated RNA was recovered by centrifuging the mix at 16,000g for 15 minutes at 15°C. Typically, a white pellet placed at the bottom of the tube can be observed. After removing the supernatant, the pellet was washed twice with 1 ml of 70% ethanol (Sigma, UK) in DEPC water (Ambion, UK). Supernatant was discarded and the pellet left to air-dry in a covered box that allowed airflow. RNA was then resuspended in 20 ul of nuclease-free or DEPC water (Ambion, UK) and quantified using a Nanodrop ND-1000 UV-Vis spectrophotometer.

2.3.3 DNAse treatment – removal of genomic DNA in RNA samples

Prior to cDNA synthesis, traces of genomic DNA were removed from the RNA samples using DNaseI with the DNA-*free*™ Kit (Ambion, UK) following the manufacturer's instruction. DNAse treatment was not applied to samples dedicated to RNA-seq library preparation because this treatment can sometimes partially degrade RNA molecules.

2.3.4 First strand synthesis – cDNA synthesis

Up to 1 ug of DNAse-treated total RNA was used for reverse transcription reaction using SuperScript II and oligo-dT (Invitrogen, UK) and following manufacturers instructions. After first strand synthesis, cDNA was diluted by adding 183 ul of nuclease-free water (final volume 200 ul). If less than 1 ug of total RNA had been used as starting material, the amount of nuclease-free water added to the cDNA was scaled appropriately.

2.3.5 Oligonucleotides design

All oligonucleotides were designed using Primer3 (Untergasser *et al.*, 2007) with default parameters:

Primer Size (bases): minimum 18, optimal 20, maximum 27.

Primer Tm (°C): minimum 57, optimal 60, maximum 63.

Primer GC (%): minimum 20, maximum 80.

Product sizes: variable.

For qPCR, primers were also designed using default parameters except that the product sizes were limited to be between 100-150 bases. Where possible, oligonucleotides were designed in different exons as an extra control for DNA contamination.

Oligonucleotides were ordered from Sigma-Aldrich (UK) with the following specifications:

Purification method: desalt

Concentration: 100 uM in water

Stock primers were kept at -20°C.

2.3.6 Standard PCR

All standard PCR were performed using QIAGEN Fast Cycling PCR Kit (QIAgen, UK). Unless otherwise specified, reactions were performed in a total volume of 10 ul with 1 ul of template (cDNA obtained as described in 2.3.4), 1 ul of primer mix (10 uM each; final concentration 1uM), 3 ul nuclease-free water and 5 ul of QIAGEN Fast Cycling PCR master mix. Thermo cycler programme was as follows:

1. Initial denaturalization step and polymerase activation:	1 minute	94°C
2. Denaturalization step:	5 seconds	95°C
3. Annealing step:	5 seconds	58°C
4. Elongation step:	10 seconds	72°C
5. Repeat steps 2-4 for a total of 35 times.		
6. Final elongation step:	1 minutes	72°C.

All PCR were carried out in a MJ Research PTC-225 Peltier Thermal Cycler.

2.3.6.1 Validation of *trans*-spliced transcripts

Standard PCR was used to validate the presence of *trans*-spliced transcripts. In each reaction, the forward primer was SL1 while the reverse primer was gene specific (**Table 2.1**). Smp_024110, previously reported as a *trans*-spliced (Davis *et al.*, 1997), was used as a positive control. Smp_045200 was used as a negative control.

Table 2.1 – Primer combinations used for the validation of *trans*-spliced transcripts. Primer sequences are presented in Appendix A.

Type of experiment	Forward primer	Reverse primer
Test	SL1	Smp_102510_R
Positive control	SL1	Smp_024110_R
Test	SL1	Smp_027360_R
Test	SL1	Smp_016410_R
Test	SL1	Smp_141320_R
Test	SL1	Smp_136960_R
Test	SL1	Smp_124050_R
Test	SL1	Smp_176420_R
Test	SL1	Smp_176590_R
Test	SL1	Smp_030020_R
Test	SL1	Smp_048880_R
Negative control	SL1	Smp_045200_R
Negative control	Smp_045200_F	Smp_045200_R

2.3.6.2 Validation of polycistronic transcripts

For validation of polycistronic transcripts, each putative polycistron was subjected to two PCR (**Table 2.2**); the first evaluates the presence of a transcript containing the intergenic region (using gene specific primers from both upstream and downstream genes[1]) while the second evaluates the presence of the *trans*-spliced gene (using the SL1 and a gene specific primers from the gene downstream of the *trans*-splice site). The polycistron enolase-UbCRBP (Davis *et al.*, 1997) was used as a positive control. In the case of the polycistron PCR these were verified by capillary sequencing of the PCR product.

[1] In the context of polycistronic transcripts, up stream and down stream refer to the position of the transcript with respect to the *trans*-splicing acceptor site.

Table 2.2 - Primer mixes used for detection of polycistronic transcripts. Primer sequences are presented in Appendix A.

Type of experiment	Forward primer	Reverse primer	Feature test
Positive control	enolase_poly_F	enolase_poly_R	polycistron
Test	Smp_006980-70_F	Smp_006980-70_R	polycistron
Test	SL1	Smp_006980-70_R	*trans*-splicing
Test	Smp_084900-890_F	Smp_084900-890_R	polycistron
Test	SL1	Smp_084890_R	*trans*-splicing
Test	Smp_079750-60_F	Smp_079750-60_R	polycistron
Test	SL1	Smp_079760_R	*trans*-splicing
Test	Smp_023160-70_F	Smp_023160-70_R	polycistron
Test	SL1	Smp_023170_R	*trans*-splicing

2.3.7 Nucleic acid separation

2.3.7.1 Agarose gel electrophoresis

PCR products were analysed in a 1.5% or 2% agarose (Sigma, UK) gel in 1x TBE (Tris Borate EDTA) using a molecular weight marker ranging from 100 bp to 1000 bp (Hyperladder IV, Bioline, UK). DNA staining was performed with ethidium bromide (Sigma, UK); gel image documentation was taken with a GelDoc-IT Imaging System.

2.3.7.2 Acrylamide gel electrophoresis

RNA samples from the fragmentation experiment (see 2.4.3) were analysed using the Novex® Gel System (Invitrogen, UK) using pre-cast Novex® 10% TBE-Urea (Invitrogen, UK). Samples were denatured by heating them in equal volume of loading buffer for 5 minutes at 65°C. Two molecular weight markers were used (25 bp ladder, Invitrogen, cat.no. 10597-011 and SRA Ladder catalogue number 1001665) and were treated in the same way prior loading them in the gel.

2.3.8 AlamarBlue® – metabolic activity of schistosomula

AlamarBlue® incorporates a colour indicator of metabolic activity of the mitochondrial function (Springer *et al.*, 1998). This indicator changes colour when the redox state of the growth media changes as a result of cell growth: the more growth/metabolic activity the cells have in culture, the more reduced the growth media will be and therefore the more colour the indicator will develop.

This indicator has previously been used to assess the viability of schistosomula (Mansour *et al.*, 2010) and a modified version of the protocol was used in this work.

In order to identify the minimum number of parasites required to detect a difference in metabolic activity after a 3 hours of incubation time, a titration using different number of mechanically transformed schistosomula was performed. Schistosomula were obtained as described in section 2.2.2. All experiments were carried out in flat-bottom, transparent 96-well plates. Aliquots of 250, 500 and 1000 parasites were placed in a total of 200 ul of supplemented DMEM in each well. Blank wells (negative control) consisted of only media. Ten ul of AlamarBlue® (Invitrogen, UK) were added to each well and the plates incubated for 3 hours or 24 hours at 37°C 5%CO_2. Absorbance was measured at 570 nm (with reference at 600 nm) using a microplate reader BioTek PowerWave HT (BioTek Instruments Inc., Winooski, VT, USA); data collection was performed using the software Gen5 (BioTek Instruments Inc. , Winooski, VT, USA).

In order to asses differences in metabolic activity between mechanically and skin transformed schistosomula (for details in the preparation of schistosomula see sections 2.2.2 and 2.2.3), 3-hour-old and 21-hour-old schistosomula (mechanically- and skin-transformed schistosomula) were placed in 200 ul of supplemented DMEM + 10 ul AlamarBlue (Invitrogen, U.K.) for 3 hours and absorbance measured. Several technical replicates were used in each experiment. Student's *t*-test was calculated to determine the significance of the mean's differences.

2.3.9 ConA-FITC staining

Concanavilin-A (ConA) is a lectin that binds to glycoproteins containing α−D-mannose or α−D-glucose. Samuelson *et al., (*1982) showed that ConA binds to the surface of schistosomula but not cercariae.

Mechanically transformed schistosomula or skin-transformed schistosomula were prepared as previously described (2.2.2 and 2.2.3 respectively) except that after transformation, no FCS was used in the incubation media as this interferes with the binding of ConA. Parasites were concentrated to 3000-5000 individual per ml in cold supplemented DMEM (see 2.1.2). ConA-FITC (Sigma, UK) solution was added to the parasites' suspensions to a final concentration of 50 ug/ml. Parasites were incubated in ConA-FITC solution for 30 minutes at 37°C and washed 3 times in supplemented DMEM. A Nikon Eclipse E600 epifluorescence microscope fitted with a Hamamatsu CCD digital camera was used for visualization and the Metamorph software (Molecular Devices) was used for image acquisition.

2.4 *RNA-seq library preparation for Illumina sequencing*

A total of 11 libraries were produced for this study. A list and details of each sample are presented in **Table 2.3.**

Table 2.3 – Summary of biological samples obtained for the generation of RNA-seq libraries.

Sample name	Parasite life cycle stage	Number of individuals	RNA yield	Starting amount of RNA
cerc10a (*)	cercariae	580,000§	52ug	10ug
cerc12 (**)	cercariae	NR	NR	10ug
cerc13 (**)	cercariae	NR	NR	10ug
somule1	3hr MT schistosomula	~250,000§	33ug	20ug
somule2 (**)	3hr MT schistosomula	100,000	23ug	11ug
somule3 (**)	24hr MT schistosomula	100,000	24ug	11ug
somule4 (**)	24hr MT schistosomula	114,000	15.2ug	11ug
somule5-6 (**)	24hr ST schistosomula	121,000§	30ug	11ug(2x)
adult2	7-week mixed-sex adult worms	NR	24ug	10ug

The preparation of some libraries were either assisted (*) or fully done (**) by the library production team led by Dr. Michael A. Quail at the Wellcome Trust Sanger Institute. NR – Not recorded; § indicates samples were RNA was pooled from two or more RNA extractions experiments in order to obtain sufficient starting material.

Detailed descriptions of the steps involved in library preparations are presented below.

2.4.1 RNA extraction and RNA quality control prior to library preparation

The first step in the preparation of RNA-seq libraries is RNA extraction (see section 2.3.1) and quality control of extracted RNA. Checking the integrity of RNA is a crucial quality control in the generation of RNA-seq samples. In general, all RNA extractions yielded good quality RNA. As an example, capillary electrophoreses results (Bioanalyzer®) obtained from intact and partially degraded samples are presented in **Figure 2.3A** and **2.3B**. In other systems, such as mammalian samples, the Bioanalyzer® software provides a figure representing RNA's integrity called RIN (RNA integrity number), which is based in the ratio of the concentration of 18S and 28S ribosomal subunits. *S. mansoni*'s 28S rRNA subunit is nicked *in vivo* (Tenniswood *et al.*, 1982) and during electrophoresis it migrates together with the 18S rRNA subunit. These appear as only one band, or a single peak in the Bioanalyzer® electropherogram preventing the RIN from being calculated. Therefore quality assessment of *S. mansoni* RNA is done based on the presence of partially degraded molecules. If the RNA is degraded these molecules will elute earlier than the 18S ribosomal subunit (**Figure 2.3B**). In terms of quantities, most of the samples yielded enough RNA for the production of one and sometimes two libraries. However, in some cases it was necessary to pool RNA from different extractions in order to obtain enough starting material. In all cases, the integrity of each individual RNA extraction was checked.

2.4.2 mRNA purification from Total RNA

Polyadenylated molecules are extracted from the total RNA sample using magnetic beads covalently bound to poly-dT oligomers. Ten ug of total RNA aliquots (sample) were diluted with nuclease-free H_2O to a final volume of 50 uL in a 1.5 ml RNase free non-sticky tube (Ambion, UK). Samples were heated at 65°C for 5 minutes to disrupt secondary structures, then placed on ice. One hundred ul of Dynal oligo(dT) beads (Invitrogen, UK) were aliquoted into a 1.5 mL RNase free non-sticky tube and washed twice[1] with 100 ul of Binding Buffer (20 mM Tris-HCl pH 7.5, 1.0 M LiCl and 2 mM EDTA). After removing the supernatant, beads were resuspended in 50 uL of Binding Buffer, and 50 uL of total RNA in solution was added to the beads. Tubes were incubated at room temperature (18°C - 25°C) for 5 minutes in constant gentle rotation. After removing the supernatant, the beads were washed twice with 100 ul of Washing Buffer B (10 mM Tris-HCl, pH 7.5, 0.15 M LiCl, 1 mM EDTA). After the last wash, supernatant was removed and 20 ul of 10 mM Tris-HCl was added and the tubes placed in a dry heat block at 80°C for 2 minutes to elute polyadenylated molecules.

[1] Separation of the magnetic beads from the solution (for example, for washing beads or removing supernatant) was achieved using a 6-tube Magnetic Stand (Ambion, UK)

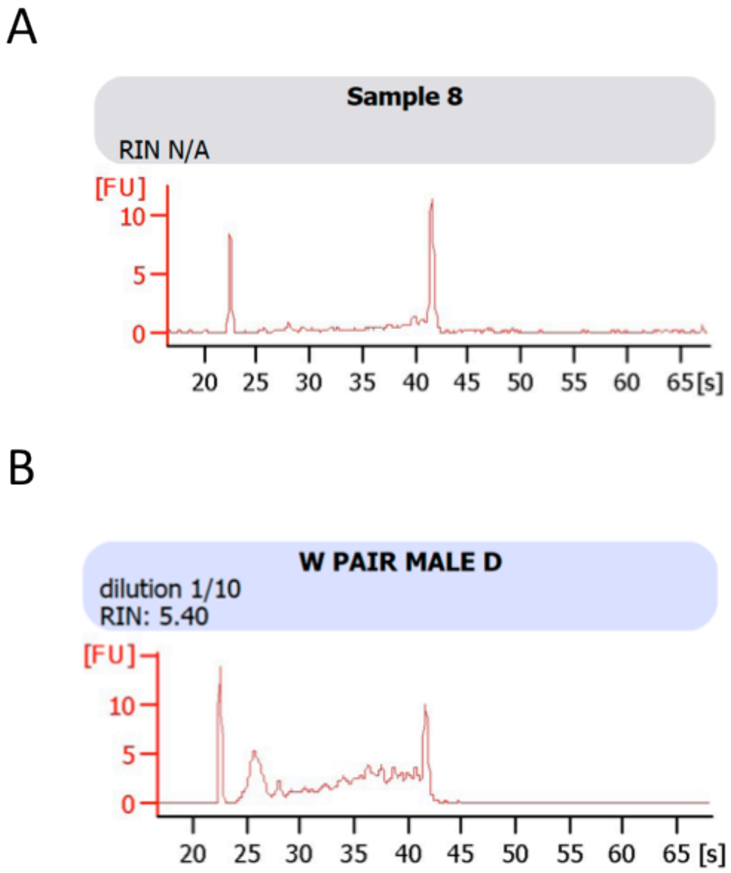

Figure 2.3 – The library preparation procedure requires quality control or RNA samples. A – Electropherogram of total RNA used for the library preparation of somule1 isolated using a QIAgen column. From left to right, the first peak corresponds to a size marker and the second peak corresponds to the 18S rRNA. The 28S rRNA is missing in *S. mansoni*. B – Electropherogram of partially degraded total RNA isolated using TRIzol reagent. Smaller RNA molecules resulting from the degradation of 18S rRNA appear between retention times of 25 and the 40 seconds. [S], retention time in the column in seconds; FU, relative fluorescent units.

After 2 minutes, the tube was placed again in the magnetic stand; the supernatant was removed and added to a tube containing 80 ul of Binding Buffer[1]. The remaining beads were washed twice in 100 ul of Washing Buffer B. The mix of Binding Buffer and RNA sample was heated at 65°C for 5 minutes to disrupt the secondary structures, then cooled on ice and added to the beads suspension. Tubes were incubated at room temperature (18°C - 25°C) for 5 minutes in constant gentle rotation. After removing the supernatant, the beads were washed twice with 100 ul of Washing Buffer B. After removing the supernatant from the last wash step, 10 ul of 10 mM Tris-HCl was added to the beads and the mix was placed in a dry heat block at 80°C for 2 minutes to elute polyadenylated RNA. Immediately, beads were placed on the magnetic stand and supernatant containing the purified polyadenylated RNA was transferred to a fresh 200 ul thin wall PCR tube. Typically, 9 ul of RNA in solution was recovered.

Two rounds of extraction yielded approximately 150-300 ng of polyadenylated RNA (typically from 10 ug of total RNA as starting material). Given that mRNA is present in 1-5% of the total RNA, the obtained quantities are consistent with the expected amount.

2.4.3 Fragmentation of mRNA

The length of the DNA molecules destined for sequencing is an important factor of the sequencing process. There are two important considerations. First, it is recommended that the length of the molecules are at least two times greater than the planned number of cycles at which the DNA molecules will be sequenced, which will determine the length of the sequencing read. This is important because when the molecules are too small reads sequenced from the forward and reverse strands will overlap in the middle. Although this does not necessarily represent a disadvantage, longer DNA molecules will provide more information by generating non-overlapping reads. Better results are obtained when larger molecules are sequenced even though a gap is introduced between the forward and reverse reads. Larger molecules will span a larger stretch of RNA and therefore having higher chances of connecting exons found far apart. Given that millions of reads will be generated, the unsequenced part of an individual template will be covered by other sequenced reads after alignment to the genome. Second, the molecules cannot be too large otherwise the bridging reaction that produces a "polony" (see Chapter 1 section 1.3 for an

[1] This second round or extraction of polyadenylated RNA molecules is suggested in the Dynal oligo(dT) beads instructions manual to reduce contamination with other RNA species in the polyadenylated RNA extraction.

explanation of the sequencing protocol) would be spread across a larger surface therefore compromising the process of reading the signal generated from the process of sequencing.

Consequently, in order to obtain a population of smaller RNA molecules where the majority of RNA species would be represented, the library preparation protocol introduces a fragmentation step followed by size selection (Mortazavi *et al.*, 2008). In the fragmentation step polyadenylated RNA is subjected to controlled degradation by heating the RNA sample to high temperatures (>65°C) in the presence of Zn^{2+} or Mg^{2+} salts. In order to investigate the range of fragments created, total RNA was subjected to fragmentation at 70°C for 5, 10 and 15 minutes (**Figure 2.4**). These results suggested that fragmentation is very efficient as the ribosomal bands present in the control sample are no longer evident in the fragmented samples. What is more, the RNA molecules with higher molecular weight (~500 nt) tend to disappear as the incubation time is increased. In subsequent experiments, the fragmentation time was therefore tightly controlled to avoid excessive fragmentation of RNA molecules, which would lead to loss of sample.

For library preparation and according to the circulating protocol, fragmentation was carried out at 70°C for 5 minutes. A detailed protocol is presented here.

One ul of 10x Fragmentation Buffer (Ambion, UK) was added to the 9 ul of RNA solution obtained from the previous step (mRNA extraction using magnetic beads) and the mix was heated at 70°C for exactly 5 minutes. The fragmentation reaction was stopped immediately by adding 1 ul of Stop Buffer (Ambion, UK) and was then placed on ice. In order to clean the mix from the fragmentation salts and stop buffer, a sodium acetate/isopropanol precipitation of RNA (see 2.3.2) was performed. The resulting RNA pellet was resuspended in 10.5 ul of nuclease-free water.

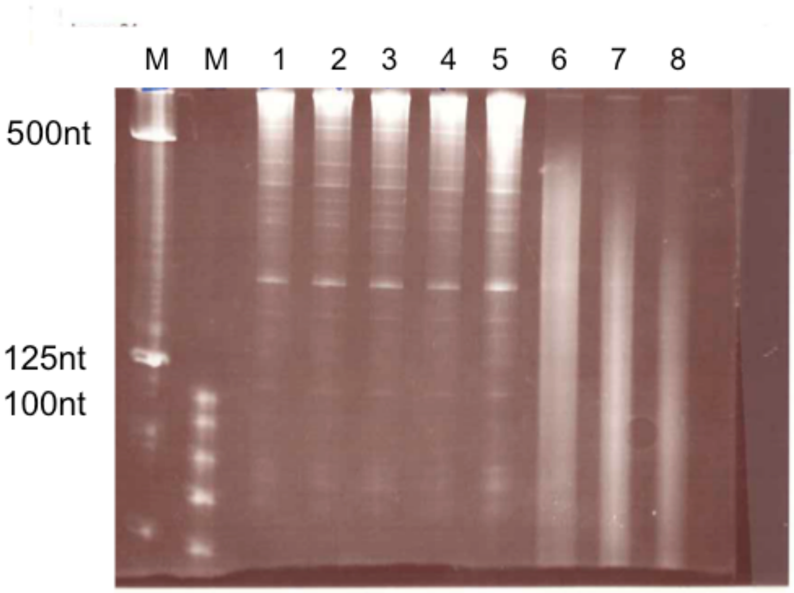

Figure 2.4 – Test on fragmentation of RNA for RNA-seq library preparation. Chemical fragmentation of total RNA is performed in the presence of salts at high temperature (70°C). Lanes 1 to 5 are controls of non-fragmented RNA; lane 6, 7 and 8 are RNA samples incubated at 70°C for 5, 10 and 15 minutes respectively. Molecular markers are indicated as "M". (nt = nucleotides)

2.4.4 First strand cDNA synthesis

Although the principle is the same as previously described (Section 2.3.4), the protocol includes the use of random hexamers instead of oligo-dT for priming the RNA.

Procedure:

The fragmented RNA sample from the previous step was placed in a 200 ul thin wall PCR tube and 1 ul of random hexamer primers 3ug/ul (Invitrogen, UK) was added. The mix was incubated at 65°C for 5 minutes and then placed on ice. The following mix was prepared separately:

- 4 ul of 5x first strand buffer (Invitrogen, UK)
- 2 ul of 100 mM DTT (Invitrogen, UK)
- 1 ul of dNTP mix (10 mM each – ThermoScientific, UK)
- 0.5 ul of RNaseOUT (40U/μL) (Invitrogen, UK)

And was then added to the RNA samples. The mix was incubated at 25°C for 2 minutes prior to the addition of 1ul of SuperScript II (200U/ µL, Invitrogen, UK). The reaction mix was incubated in a thermal cycler (MJ Research PTC-225) with the following program:

• Step 1	25°C	10 min
• Step 2	42°C	50 min
• Step 3	70°C	15 min
• Step 4	4°C	Hold

2.4.5 Second strand cDNA synthesis

This step generated fragmented double stranded cDNA.

First strand cDNA from the previous step was diluted with 61 ul of nuclease-free water and the following reagents were added:

- 10 ul of 5 x second strand buffer (100 mM Tris-HCl pH 6.9, 23 mM $MgCl_2$, 450 mM KCl, 0.75 mM beta-NAD+, 50 mM $(NH_4)_2SO_4$ – Invitrogen, UK)
- 3 ul of dNTP mix (10 mM each, ThermoScientific, UK).

The solution was mixed gently and placed on ice for 5 minutes. Then, the following reagents were added:

- 1 ul of RNaseH (2U/µL, Invitrogen, UK)
- 5 ul of DNA Pol I (10U/µL, Invitrogen, UK)

The solution was carefully mixed and incubated at 16°C in a thermal cycler (MJ Research PTC-225) for 2.5 hours. Afterwards, DNA was purified using a QIAquick PCR spin column (Qiagen, UK) and eluted in 30µL of EB solution (Qiagen, UK)

2.4.6 End Repair

Unless otherwise stated, reagents in this step are part of the Pair-end DNA sample prep (Illumina, UK) kit[1].

DNA from previous step was diluted with 45 ul of nuclease-free water and the following reagents were added:

- 10 ul of T4 DNA ligase buffer with 10mM ATP
- 4 ul of dNTP mix (10mM each)
- 5 ul of T4 DNA polymerase (3U/µL)

[1] This kit is no longer available from this provider. Similar kits might be found from Illumina, UK or other providers.

- 1 ul of Klenow DNA polymerase (5U/μL)
- 5 ul of T4 PNK (10U/μL)

The mix was incubated at 20°C for 30 minutes and the resulting DNA was purified using the QIAquick PCR spin column (Qiagen, UK) and eluted in 32 ul of EB solution (Qiagen, UK).

2.4.7 Addition of a single "A" base

Unless otherwise stated, reagents in this step are part of the Pair-end DNA sample prep (Illumina, UK) kit.

The following reagents were added to the DNA solution from the previous step:

- 5 ul of Klenow buffer
- 10 ul of dATP (1 mM)
- 3 ul of Klenow 3' to 5' exo- (5U/μL)

The mix was incubated at 37°C in for 30 minutes and the DNA was then purified using the QIAquick MinElute column (Qiagen, UK) and eluted in 19 ul of EB solution (Qiagen, UK).

2.4.8 Adaptor ligation

Unless otherwise stated, reagents in this step are part of the Pair-end DNA sample prep (Illumina, UK) kit.

The following reagents were added to the DNA sample obtained from the previous step:

- 25 ul of DNA Ligase buffer
- 1 ul of Adaptor oligo mix
- 5 ul of DNA Ligase (1U/μL)

The mix was incubated at room temperature (18-25°C) for 15 minutes and the DNA was purified using Agencourt AMpure SPRI beads (Beckman Coulter Genomics, UK).

2.4.9 Gel purification of double stranded cDNA – size selection

DNA sample was size-separated in a 2% low-melting point agarose gel (Sigma, UK) prepared in ice-cold 1x TBE buffer leaving at least two wells separation between samples or sample and molecular weight marker. Electrophoresis was carried out at 120V for 45-60 minutes or until good separation of the molecular weight marker bands was achieved. Using the molecular weight marker as a guide, a gel slice corresponding to where the samples had been loaded was cut between 300-400 bp. DNA was recovered from the agarose gel using the QIAquick gel extraction kit (Qiagen, UK) and eluted in 30µL of EB solution (Qiagen, UK).

2.4.10 PCR enrichment of purified double stranded cDNA templates

Unless otherwise stated, reagents in this step are part of the Pair-end DNA sample prep (Illumina, UK) kit.

A PCR master mix was set up as follows:

- 10 ul of 5x Phusion Buffer
- 1 ul of PCR primer PE[1] 1.0
- 1 ul of PCR primer PE 2.0
- 0.5 ul of 25 mM dNTP mix
- 0.5 ul of Phusion DNA polymerase
- 7 ul of nuclease-free water

And was then added to the DNA solution obtained from the previous step. PCR was performed under the following programme:

Thermo cycler programme was set up as follows:

1. Initial denaturalization step and polymerase activation:	30 seconds	98°C
2. Denaturalization step:	10 seconds	98°C
3. Annealing step:	30 seconds	65°C
4. Elongation step:	30 seconds	72°C
5. Repeat steps 2-4 for a total of 15 times.		
6. Final elongation step:	5 minutes	72°C.

Amplified cDNA was then purified using Agencourt AMpure SPRI beads (Beckman Coulter Genomics, UK).

2.4.11 Verification of library sizes and adaptor contamination

A final quality control step was performed to verify that the DNA fragment sizes were within the expected range. To this end, 1 ul of the amplified DNA was analyzed in the Agilent Bioanalyzer DNA 1000 chip (Agilent, UK) according to manufacturer specifications. The size of the molecules present in the RNA-seq libraries ranged from 300-500 bp. At this point, it is common to find contaminating adaptors that had been carried over from the

[1] PCR primer sequences (PE primers) are propriety of Illumina®.

adaptor ligation step. Contaminating adaptors can be easily removed from the sample by another size selection step prior to sequencing. This step is routinely done and does not compromise the quality of the library.

2.4.12 Quantification of libraries.

It is important that an accurate measurement of the DNA concentration is done prior to sequencing. Precise quantification of DNA is done using real-time PCR. The library team (led by Dr. Michael A. Quail, WTSI) performed this quality control step for all the samples submitted in this study.

2.5 Sequencing

Libraries listed in 2.4 were sequenced as 76 base pair reads using the Illumina® Genome Analyzer IIx platform. Sequencing data were submitted to ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) under the accession number E-MTAB-451.

Biological replicates *vs.* technical replicates.

It is desirable that each sample has biological and a technical replicate. Biological replicates control for the variability arsing from the biological diversity of the subject of study. In the present study and due to the nature of the samples, it was necessary to pool samples from different experiments (e.g. schistosomula resulting from different transformation experiments carried out in different occasions) in order to obtain enough RNA for library preparation. Therefore, biological replicates could not be provided.

Technical replicates are also desirable. In this study, two types of technical replicates were assessed. One of them is a control of the library preparation protocol. In this case, a RNA sample is divided in two and these are subjected to parallel library preparation protocols. The second type of technical replicate is a control of the sequencing. In this case, the same library is subjected to two round of sequencing. The analysis of biological and technical replicates is presented in detail in Chapter 3 section 3.2.2.1.

2.6 Bioinformatic procedures

2.6.1 Alignment of RNA-seq reads to genome.

The alignment tool TopHat (Trapnell *et al.*, 2009) was chosen to map RNA-seq data to the genome. Contrary to reads generated from genomic DNA, reads generated from eukaryotes RNA samples will sample exon-exon boundaries and therefore aligners that do not take this into consideration will fail to find a suitable location for the complete length of that read in the genome.

Tophat provides a complete *de novo* splice site junction finder; it does not need to have *a priori* information about known splice site junctions. The first step in TopHat is the alignment of reads using Bowtie (Langmead *et al.*, 2009). With default parameters, Bowtie will report to TopHat reads that have up to 2 mismatches within the first 28 bases and up to 10 alignments might be reported for each read. Only low complexity reads are discarded at this stage. Then, TopHat infers "islands" from the regions of the reference where contiguous coverage is detected. These islands are regarded as putative exons and are generated as mini-assemblies of reads. Where there is a discrepancy in the sequence of the island and the reference the reference is used to make a base call; it will also extend the island in both 3' and 5' direction (by a default of 45 bases), based on the reference sequence. This is done in order to address the issue that coverage will naturally be lower in these regions because Bowtie would have aligned hardly any reads to them. In order to map reads to splice junctions, TopHat lists all the possible canonical splice acceptor and donor sites (GT-AG, GC-AG and AT-AC when reads are longer > 75bp) within each island and then generates putative introns based on the distance between the splice sites (minimum of 70 bases maximum of 20,000 with default parameters) found in nearby, yet not necessarily adjacent islands. Then, the reads not initially aligned are searched for reads that would span the junctions using a seed-and-extend approach. Pair-end data is also used. The software reports all spliced alignments.

2.6.1.1 RNA-seq reads alignment for gene expression studies

RNA-seq reads were aligned to the reference genome using TopHat (Trapnell *et al.*, 2009) (version 1.3.1) with default parameters except for minimum and maximum intron sizes which were set to 10 and 30,000bp respectively. Other parameters that were specified included the type of library sequenced (set to standard cDNA Illumina library; --library-type fr-unstranded) and the mate pair distance (or insert size; -r option), which was calculated individually for each library based on the mapping alignment of the reads to the transcriptome. The output filtered to show reads that map uniquely to the reference (reads that map to several locations in the genome are excluded).

The number of reads aligned to each exon was calculated using BEDTools (Quinlan *et al.*, 2010). The final count of reads per exon was parsed into reads per transcript and used to calculate RPKM values (Mortazavi *et al.*, 2008) for each transcript (reads per kilobase per million of mapped reads). This value provides the means to rank transcripts based on their expression levels within each library but is not suitable for comparing levels of expression for a given transcript across libraries.

2.6.2 Finding the RPKM threshold for discriminating background RPKM

Procedure developed by Dr. Adam Reid (Pathogen Genomics group - WTSI).

Most of the reads generated from RNA-seq will map specifically to locations in the genome where a gene is found. However, a proportion of reads will be mapped non-specifically generating noise in the signal. This could be because of artefacts in the sequencing (e.g. the read is of bad quality) or to the presence of contaminants in the sample (e.g. DNA contamination). In order to define a minimum level of expression that would discriminate between signal and noise, a threshold RPKM value must be calculated. For RNA-seq data representing libraries of mature mRNA transcripts, it is usually assumed that introns and intergenic regions are not expressed. Therefore, the reads mapping to these regions can be used as a measure of noise. Some introns and intergenic regions will be expressed, probably due to intron retention or expression of non-coding RNAs, making this approach a rather conservative one: it is likely that the established cut-off produces more under calling of expressed features rather than over calling them leading to a reduction in the false positives.

The procedure calculates RPKM values for different feature types across a given scaffold, in this case Schisto_mansoni.Chr_1.unplaced.SC_010. The evaluated feature types are: exon, intron, UTR (100 bp up- and down-stream of a gene) and intergenic regions, which are all defined by the existing genome annotation. Background noise was calculated using 500 bp non-overlapping windows across all the feature types, with each window considered as a potentially expressed unit. BAM files resulting from the TopHat alignments of all libraries (with different number of sequenced reads) were used producing similar results. **Figure 2.5** presents the results from one of the libraries tested.

By choosing an RPKM cut-off of 2, 90-95% of the introns and intergenic regions are removed compared to only 23% of the exons and UTR regions.

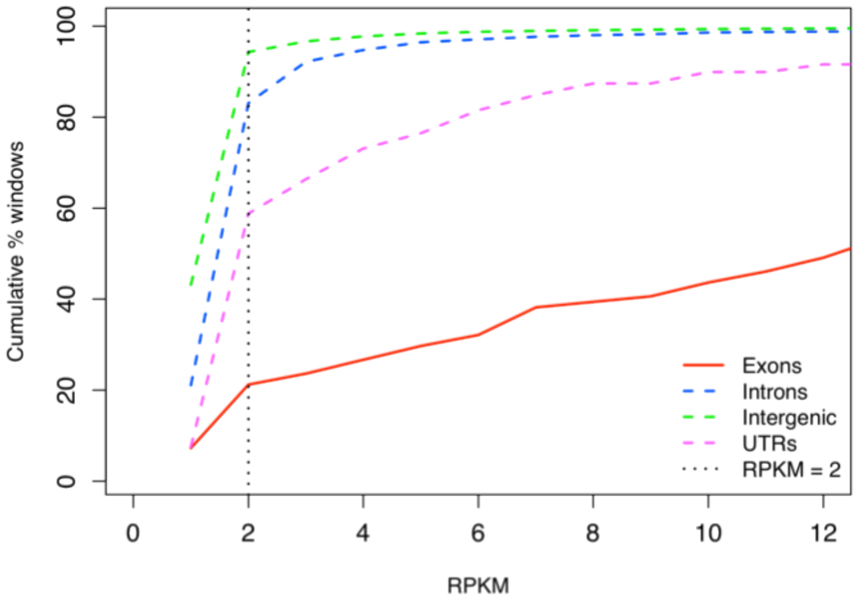

Figure 2.5 – Empirical calculation of minimum RPKM value. Method developed by Dr. Adam Reid (Pathogens Genomics group, WTSI).

2.6.3 Identification of *trans*-spliced and polycistronic genes

The procedures described in this section were designed by myself and implemented in collaboration with Dr. Martin Hunt and Dr. Isheng J. Tsai (Team133 – Pathogen genomes - WTSI).

RNA-seq data was screened for reads containing the 36-nucleotide sequence corresponding to *S. mansoni* splice-leader (SL) (Rajkovic *et al.*, 1990). Although this seems a rather strict criterion and a shorter minimum number of bases matched could have been used, this approach guarantees specificity (see below). Subsequently, the SL sequence was clipped off the reads and the remainder of the read (and its mate pair) were mapped to the genome using SSAHA2 v2.5 (Ning *et al.*, 2001) allowing putative *trans*-spliced acceptor sites to be identified. A *trans*-spliced acceptor site is defined as the first base in the genome that corresponds to the *trans*-spliced transcript that can be identified by mapping the trimmed SL-containing read. *Trans*-splicing acceptor sites can fall in one of four different places: up stream of the start of a gene (up to a maximum of 500 bp), within an exon, within an intron, or down stream of a gene. Only the first three categories are considered putative *trans*-spliced transcripts. **Figure 2.6** shows the number of trans-splicing events found with increasing number of supporting reads. *Trans*-splicing events with a minimum of four reads were considered for down stream analysis.

In order to identify polycistronic units, we looked for genes found within 200 bp and up to 2000 bp upstream of a putative *trans*-spliced transcript. Where a gene was found within the specified distance, the gene pair was catalogued as a putative polycistronic unit.

Further studies using a shorter minimum requirement for reads containing the splice leader sequence may result in a higher number of identified trans-spliced events.

2.6.4 Correlation of RNA-seq and microarray data

Microarray studies covering the same life cycle time points surveyed in this thesis have been previously published (Fitzpatrick *et al.*, 2009; Parker-Manuel *et al.*, 2011). These data were used to study the correlation between RNA-seq and microarrays expression data.

Normalized intensity values from the work of Fitzpatrick *et al.,* (2009) were obtained from the supplementary materials and methods (Fitzpatrick *et al.*, 2009); normalized $\log_2$ intensity values from Parker-Manuel *et al.,* (2011) were obtained from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) accession numbers GSE22037 and GPL10466.

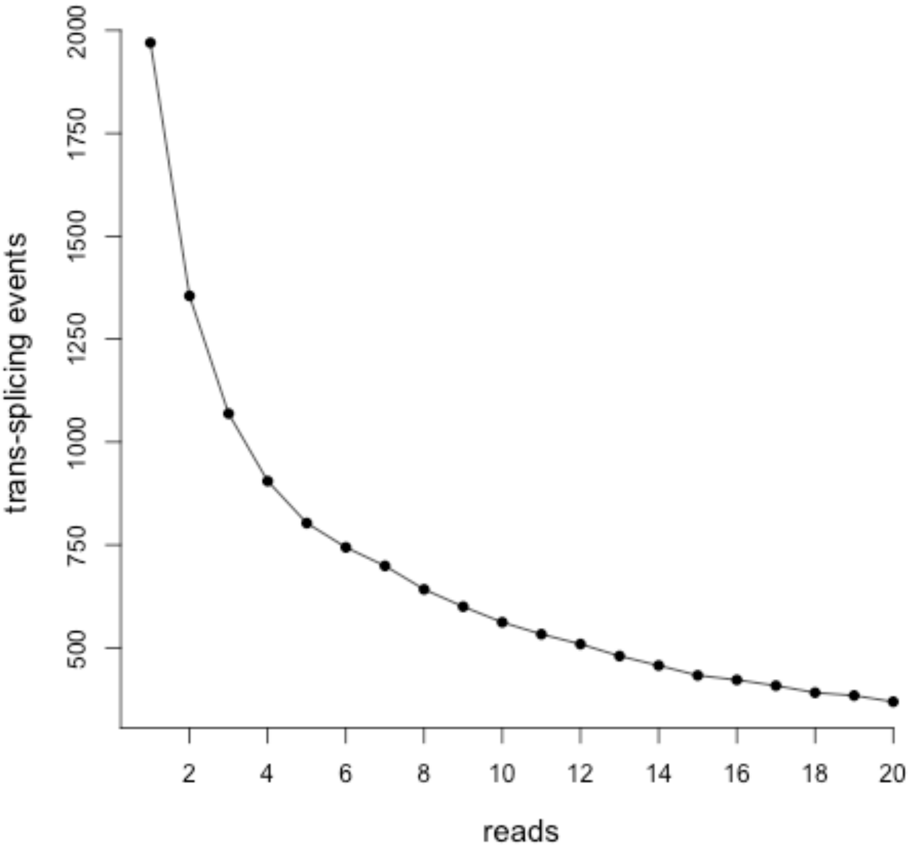

Figure 2.6 – Total number of *trans*-splicing events detected depends on the number of reads supporting such event. Almost half of the *trans*-splicing events are supported by fours reads or less [plot suggested by Mr Ferenc Kiss – University of Wurzburg, Germany].

In order to study the correlation between oligonucleotide probes and gene models, the 389,211 60-mer probes found in the array of (Parker-Manuel *et al.*, 2011) and the 35,078 unique 50-mer probes present in the array of (Fitzpatrick *et al.*, 2009) were mapped to the *S. mansoni* genome (v5.0) using SSAHA2 (Ning *et al.*, 2001) (with default parameters except for "-solexa" and "-identity 100") and only perfect matches (100% identity) that unambiguously matched one location in the genome were selected for subsequent analysis. The coordinates where the microarray probes were mapped to the genome were recorded and the number of "reads per probe" location was calculated using the CoverageBed programme from BEDtools (Quinlan *et al.*, 2010). In this particular case, reads per probe can be used instead of RPKM values because all probes are of the same length and therefore normalization by this parameter is not necessary. Log$_2$ values of both normalized microarray intensities and RNA-seq "reads per probe" were used to calculate the Spearman's rank correlation for each comparison: RNA-seq *vs.* Fitzpatrick *et al.,* (2009) and RNA-seq *vs.* Parker-Manuel *et al.,* (2011). For the microarray data from Fitzpatrick *et al.,* (2009), correlations were calculated for cercariae, 3 hours old schistosomula, 24 hours old schistosomula and adult samples while for the Parker-Manuel *et al.,* (2011) dataset only the cercariae sample was correlated.

For the analysis of constitutively expressed probes 5.2.1, these were obtained from supplementary materials and methods of Fitzpatrick *et al.,* (2009). To identify which transcript correlated to each probe, these data were extracted from the probe-transcript correlation generated as explained before in this section.

2.6.5 Differential gene expression analysis

Two differential expression experiments are presented in this thesis. Chapter 4 features the differential expression analysis between 24-hour old mechanically transformed and skin transformed schistosomula while Chapter 5 presents pair wise differential expression studies in a time course experiment (cercariae -> 3hr schisotosmula -> 24hr schistosomula). In both cases, two different figures are used: RPKMs and reads per transcript. RPKMs are used to evaluate whether a given transcript is expressed or not, as described in section 2.6.2, or to rank genes according to their expression within one sample. After calculating the mean RPKM for each transcript across replicate samples, transcripts with RPKM < 2 in all stages were removed from the "reads per transcript" dataset. This filtered dataset was used as input for the edgeR package (Robinson *et al.*, 2010) implemented in Bioconductor (Gentleman *et al.*, 2004) and written in R programming language (R Development Core Team, 2011). The output of this analysis is a table of transcript names with their respective $\log_2$ fold changes and associated p-values for a given comparison. P-values were adjusted (adjusted p-value) using a method for multiple testing (Benjamini *et al.*, 2001) and the different cut-offs are specified in each results chapter.

The statistical model behind edgeR is fully described Robinson *et al.,* (2010); a short outline is presented here.

With the aim of normalizing data across sequencing reads, other statistical models [such as RPKMs (Mortazavi *et al.*, 2008)] use a standardization approach based on the scaling of libraries. This is valid if a given RNA species is represented in the same proportion across all samples; which is a very unlikely situation in real samples. If a group of genes are exclusively expressed in sample A but not in sample B, the rest of the genes in sample A have less "representation" within sample A and may appear as under represented when compared to sample B even though they might be expressed at the same level. EdgeR proposes a statistical model where two assumptions are made:

- biological replicates follow the Negative Binomial distribution[1]
- the majority of genes are not differentially expressed; therefore the gene's overall expression between samples can be equated.

These principles are implemented as a TMM normalization (Trim Mean of M values), where the "trimmed mean is the average after removing the upper and lower x% of the data" (Robinson *et al.*, 2010). This normalization factor is calculated across all available samples by choosing one as a reference and calculating the TMM factor for all the rest (non-reference) samples. When the case is a two-sample comparison, one "relative scaling factor" is calculated and applied to each sample. The values actually subjected to being trimmed are the log-fold changes (M-value) and absolute intensities (A-value).

2.6.6 Gene Ontology (GO) term enrichment

Gene Ontology (GO) is a controlled vocabulary system developed and maintained by the Gene Ontology Consortium (Ashburner *et al.*, 2000). Genes and gene products are

[1] Negative binomial distribution: or over dispersed Poisson distribution. This is particularly useful to model data where the sample variance exceeds the sample mean. This usually happens in sets of data where each event can be virtually infinite count; such is the case of RNA-seq data. The formulation is very similar to the Poisson distribution but a second parameter is introduced in the negative binomial, which can be used to adjust the variance independently of the sample mean.

assigned controlled vocabulary terms based on sequence similarity. These terms are assigned in three categories: Biological Process, Cellular Compartment and Molecular Functions. GO terms are interconnected by parent-child relationships: for example the terms "DNA methylation" and "DNA replication" are children of "DNA metabolism".

The TopGO package (Alexa *et al.*, 2006) is a Gene Ontology (GO) term enrichment analysis tool implemented in Bioconductor (Gentleman *et al.*, 2004) and written in R programming language (R Development Core Team, 2011).

A summary of the principles used by TopGO is presented here.

TopGO analyses GO term enrichment in a global approach considering the whole hierarchy of the GO topology tree. Briefly, it groups all genes based on the relationship of their assigned GO terms and then maps the individual genes from a list of genes of interest (for example, up regulated genes in the cercariae to schistosomula comparison) to the GO topology/tree. The higher the number of members of a particular gene group mapped to a given GO term (or "node") and its neighbour nodes, the more important the GO term is. A test statistic (for example, Fisher's test) can be used to estimate the significance of this occurrence and based on this a scoring process is performed from the bottom to the top of the tree (from child to parent). The above description is common to many GO term enrichment packages. The novel contributions of TopGO to the downstream analysis are:

1. The *elim* algorithm. TopGO *elim*inates or removes genes mapped to significant GO terms in ancestors (parent) or already significant nodes. When a given node is found to be significant (with a p-value lower than the threshold) all genes found in this node are removed from ancestor nodes. This approach guarantees that provided its p-value, the most specific node (higher level GO term) is reported instead of a less informative one. This could cause the removal of significant nodes (because p-values of parent and child nodes are not compared at this stage) – which leads to the second implementation.
2. The *weight* algorithm. If a given node A is more significant than one of its children, the genes common to both get down *weight*ed in the children, producing less significant children nodes. If at least one child of a node A is more significant than the node A itself, genes common to the children and the node A are down-weighted in the node A and all its ancestors, making node A less significant.

The reported p-values are a combination of the *elim* and the *weight* method.

2.6.7 InterProScan – looking for conserved protein domains and signatures.

InterProScan (Hunter *et al.*, 2009) is a search engine that looks for conserved protein domains (sometimes called signatures) occurring in the amino acid sequences. It uses a combination of protein domain databases including the manually curated protein domain database Pfam (Finn *et al.*, 2010) and others that rely on automatic annotation such as SMART (Letunic *et al.*, 2009) and PROSITE (Sigrist *et al.*, 2010) among others. The significance cut-off chosen to assign a conserved domain to an amino acid sequence was an e-value of $1e^{-5}$.

2.6.8 SignalP, TargetP and TMHMM – prediction of signal peptides and *trans*-membrane domains

The software SignalP [version 4.0 - (Emanuelsson *et al.*, 2007)] was used to predict the presence of signal peptide or anchor signatures in amino acid sequences. The software TargetP [version 1.1 - (Emanuelsson *et al.*, 2007)] was used to predict the subcellular localisation of amino acid sequences. The software TMHMM [version 2.0 - (Emanuelsson *et al.*, 2007)] was used to predict the occurrence of *trans*-membrane domains within a given amino acid sequence. All these programmes were used with default parameters for eukaryotes non-plant organims.

2.6.9 Finding *S. mansoni* neuropeptide receptors using tBLASTn

Neuropeptide precursor sequences of many different platyhelminths including *S. mansoni* and *S. japonicum* were obtained from the supplementary materials presented in McVeigh *et al.,* (2009). These amino acid sequences were used as queries against the full set of gene models (spliced sequences) of *S. mansoni* using tBLASTn (Altschul *et al.*, 1990). Best hits were chosen based on the highest sequences identity.