Transcriptome characterisation of cercariae and skin-stage schistosomula in the parasitic helminth *Schistosoma mansoni*.

Anna Victoria Protasio

Christ's College

University of Cambridge

A dissertation submitted for the degree of

Doctor in Philosophy

November 2011

Declaration

The work presented in this dissertation was carried out at the Wellcome Trust Sanger Institute (Hinxton) and at the Department of Pathology (Tennis Court Road, Cambridge) between October 2007 and September 2011. This dissertation is the result of my own work – contributions from collaborations are clearly referenced and have been approved by the collaborators. No part of this dissertation has been or is being submitted for any qualification in any other university.

This thesis does not exceed the word limit established by the Biology Degree Committee.

Acknowledgements

Many people contributed to my work during these 4 years at the Wellcome Trust Sanger Institute. I am grateful to my supervisor, Matt (Dr. Matthew Berriman) for his sound advice, support and for providing an excellent work environment. He has been the most approachable of supervisors and allowed me to work in a state of "guided independence", which strengthen my confidence and scientific thinking. I would also like to thank all the members of Team 133 - Pathogen Genomics at the Sanger Institute - for their help and tolerance; special thanks to Jason (Isheng) Tsai, Martin Hunt and Adam Reid for being remarkably helpful in all things bioinformatic and provided lively discussions. Magdalena Zarowiecki, Lia Chappell and again Adam Reid help proofreading these chapters. Many thanks to all members at the Pathogens Informatics (team led by Dr. Jacquline McQuillan) and the Library Production Team (led by Dr. Michael A. Quail) for their help and technical advice. I also thank the members of my thesis committee Prof. Gordon Dougan and Dr. Julian Parkhill for their useful comments and encouraging attitude. Part of the work presented in this thesis was performed at the Schistosomiasis Research Group in the Department of Pathology, University of Cambridge – led by Prof. David Dunne who provided good advice and background for this research. At "pathology", Frances Jones and Maureen Laidlaw provided invaluable help before during and after experiments; Colin Fitzsimmons placed the most interesting questions and shared long coffee break discussions. Thanks also to Prof. Mark Filed and Dr. Ka-Fai Leung (Dept. Pathology, University of Cambridge) for facilitating the use of their fluorescence microscope. I would also like to thank Prof. Karl Hoffmann from Aberystwyth University (Wales) for his (at the distance) support and ready advice in all technical, academic and career path matters; to him I owe my passion for schistosome biology. I would also like to acknowledge in these lines the support of my friends in Cambridge, especially Bronwyn, Myrto, and Greg; and the unconditional support of my family and friends in Uruguay.

This thesis is dedicated to my parents.

Summary

Schistosomiasis is an endemic parasitic disease affecting approximately a quarter of a billion people worlwide, mainly in developing and under developed countries of Africa, Southeast Asia and Central and South America. The causative agent is a plathyhelminth worm of the genus *Schistosoma.* Chemotherapy with praziquantel is possible but does not protect from re-infection; moreover, reduced susceptibility to this drug have raised the issue of potential outbursts of drug resistance. In this context, researchers have a strong interest in finding alternative routes of chemotherapy and have also established programs for vaccine development. It is thought that the most vulnerable point in the parasites' life cycle is the early stages of life in the mammalian host. The infectious larvae, the cercariae, infect the host by penetrating through the skin where parasites transform into schistosomula. *In vivo*, skin transformation occurs within hours and this process can be reproduced *in vitro* by inducing a mechanical transformation. The parasite profile of gene expression across this vulnerable transition is not well understood. What is more, a transcriptome comparison between naturally transformed parasites and those forced to transform *in vitro* has not been investigated.

This thesis aims at filling in these gaps in our knowledge of schistosome biology by investigating gene expression changes that the early schistosomula undergo upon infection. In order to address this question, RNA-seq transcriptome sampling of cercariae, 3-hours old and 24-hours old schistosomula and adult worms was used. Because RNA-seq differential expression analyses heavily relies on genome annotation, the transcriptome data was first used to improve the gene annotation of the *Schistosoma mansoni* genome. Second, RNA-seq data generated from 24-hours old skin- and mechanically transformed schistosomula were compared. Finally, the patterns of gene expression that accompany the transformation of the parasites from the cercariae stage to the schistosomula during its first 24 hours of infection were studied. This time course study allowed the identification of known biological processes with improved resolution while other newer developmental changes are also reported and examined. The resolution achieved in this study has no precedent in any other parasitic helminth and contributes to our understanding of schistosome biology.

Abbreviations

ATP	Adenosine triphosphate
CDS	Coding Sequence
ConA	Concanavilin A
DEPC	Diethyl pyrocarbonate
DMEM	Dulbecco's Modified Eagle's Medium
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
EST	Expressed Sequence Tag
FCS	Foetal calf serum
GO	Gene Ontology
MT	Mechanically-transformed schistosomula
NADH	Nicotinamide adenine dinucleotide
Npp	Neuropeptide precursors
nt	Nucleotides
ORF	Open Reading Frame
PCR	Polymerase Chain Reaction
qPCR	Quantitative Polymerase Chain Reaction
RPKM	Reads per Kilobase per million of reads mapped
SL	Spliced leader
ST	Skin-transformed schistosomula
TCA	Tricarboxylic acid cycle
TBE	Tris Borate EDTA
UbCRBP	Ubiquinol-cytochrome C reductase binding protein
UTR	Untranlated Region
WTSI	Wellcome Trust Sanger Institute

Table of Contents

CHAPTER 1

INTRODUCTION

1.1 Introduction

Schistosomiasis is a parasitic disease caused by platyhelminths of the genus *Schistosoma*. It has been estimated that ~780 million people live at risk of infection and ~200 million are infected (Steinmann *et al.*, 2006). The global distribution and infection rates of schistosomiasis (**Figure 1.1**) has varied little in the last 20 years (Davis, 2002; WHO, 2011). Only in Africa, where the greatest prevalence of infection occurs (WHO, 2011), it has been estimated that ~150,000 people die each year due to schistosomiasis-related causes (van der Werf *et al.*, 2003).

Symptoms of infection can be quite mild, which leads to long lasting infections often left undiagnosed. Continuous accumulation of parasite eggs in the liver causes hepatomegalia and liver failure. Infected patients are treated with praziquantel, which kills adult parasites and stops egg lying. This is orally administered, highly tolerated and cheap drug. However, this treatment does not prevent re-infection – very common in endemic areas and principally among young children - leading to long-term schistosomiasis infections with the concomitant establishment of chronic inflammation (Pearce *et al.*, 2002). This has a direct effect on morbidity, which contributes to the further impoverishment of the affected populations (King, 2010).

Without a vaccine, mechanisms of prophylaxis rely on tackling transmission through the reduction in the number of infected individuals, distribution of information regarding water contact habits and improvement of sanitary conditions in endemic areas (Davis, 2002). Although efforts have been implemented in these areas, the number of infected people has changed little over the last decades (Steinmann *et al.*, 2006).

In order to develop mechanisms of intervention against schistosomiasis infections, it is important that the process of infection is well characterised. The infectious agent for the human host is the cercariae. These microscopic free-living larvae are released by snail hosts and can infect humans by penetrating the skin and transforming into schistosomula. This transformation is characterised by phenotypic, metabolic and physiological changes that make the schistosomula a vulnerable stage in the parasite's life cycle.

The work presented in this thesis focuses on the characterisation of changes in gene expression occurring to the parasites during the transformation from the free-living cercariae to the parasitic schistosomula after 24 hours of infection.

The first step in the characterisation of such changes is the generation of reliable gene annotation (Chapter 3). Previous annotation has relied on *in silico* predictions and a limited dataset of experimental data. In this work, four time points of the parasite life cycle transcriptome were sampled and sequenced using high-throughput technology, which allowed an unprecedented level of accuracy to be obtained in the gene annotation of this organism (Protasio *et al.,* 2012). Additionally, it was possible to map trans-splicing events in a genome-wide scale and the existence of polycistronic transcripts in *S. mansoni* demonstrated for the first time.

Schistosomula from *in vivo* infections are usually recovered in very small numbers often not sufficient for high-throughput studies. The vast majority of available data generated from high-throughput gene expression studies have been gathered using a mechanical transformation approach of the schistosomula (which does not include invasion of host skin). Available comparisons of skin-transformed and mechanically transformed schistosomula are limited to changes occurring to the parasites surface (Brink *et al.*, 1977) and changes observed through electron microscopy (Cousin *et al.*, 1981). The only available gene expression study featuring a comparison from the gene expression perspective was done in *S. japonicum* and was not targeted to the skin stage of the parasites (Chai *et al.*, 2006). Because the skin stage is regarded as one of the vulnerable stages, it is important to understand what are the effects of the mechanical transformation in the parasites transcriptome. To that end, skin-transformed and mechanically transformed schistosomula were study of the bases of their transcriptome differences and similarities (Chapter 4). These results would help validate previous and future studies.

Finally, a time course analysis of the gene expression during the first 24 hours of infection is presented (Chapter 5). Close time points were selected for this study guaranteeing the maximum possible resolution. The combination of the powerful RNA-seq approach as the cutting edge technique for gene expression measurement and the newly improved genome assembly and annotation, led to the identification of developmentally regulated processes and genes that would help improve the knowledge of this critical stage in the parasites' development.

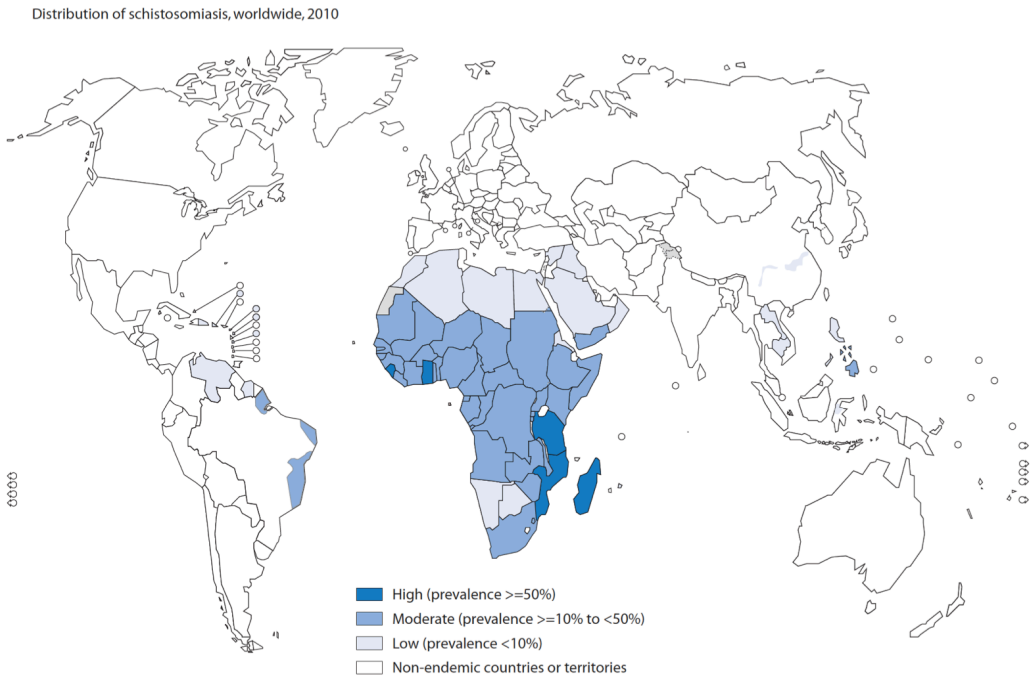


Figure 1.1 – Global distribution of schistosomiasis estimated as for the year 2009 [reproduced from (WHO, 2011)]. Shades of blue represent prevalence[1] of infection among individuals.

[1] Prevalence is the percentage of infected individuals within the population

1.2 Biology of schistosomes

Schistosome parasites are platyhelminths, blood dwelling trematodes belonging to the subclass digenea. Unusually for platyhelminths, they have separate male and female individuals (dioecious). They infect vertebrates (i.e., mammals, birds) and use them as their definitive host while aquatic or amphibious snails serve as intermediate hosts. Human schistosomiasis is caused mainly by five *Schistosoma* spp.: *S. mansoni, S. japonicum, S. haematobium, S. intercalatum* and *S. mekongi*, which are found in sub-Saharan Africa, Southeast Asia, regions of South America and the Caribbean (**Figure 1.1**). *S. haematobium* and *S. mansoni* are distributed across all sub-Saharan Africa and are the main agents for schistosomiasis in this region with some areas showing infections with *S. intercalatum* parasites. Only *S. mansoni* is found in South America and the Caribbean as a result of the slave trade between the 15th and 19th century (Morgan *et al.*, 2005). *S. japonicum* and *S. mekongi* are found in Southeast Asia with the former being the subject of a massive public health programme in China (Davis, 2002).

The following sections introduce the life cycle of schistosomes, principal aspects of the pathology and the current mechanisms of intervention.

1.2.1 Life cycle of schistosomes

The life cycle of schistosomes comprises two parasitic and two free-living phases. During the parasitic phases schistosomes colonize an intermediate snail host or a definitive vertebrate host. The free-living stages occur in fresh water environments and provide a link between the two parasitic stages (**Figure 1.2A**) breaching the physical gap between the two hosts. Each schistosome species shows a preference for a particular genus of snail host: *S. mansoni* infects snails of the genus *Biomphalaria*, while *S. japonicum* and *S. haematobium* infect snails of the genus *Oncomelania* and *Bulinus* respectively. However, not all the *Biomphalaria* species are susceptible to infection by *S. mansoni* (Davis, 2002). The geographic distribution of the susceptible snail population strongly influences the epidemiology of schistosomiasis (Steinmann *et al.*, 2006). Vertebrate hosts can be from a variety of classes including mammals and birds. Because of their clinical relevance, human-infectious *Schistosoma* spp. are the most widely studied. In the laboratory, the life cycle of *S. mansoni* can be maintained by using small rodents such as mice and hamsters as definitive hosts.

Adult worms of different species have different preferences for their final location in the definitive host: *S. mansoni* and *S. japonicum* stay in the inferior and superior

mesenteric vessels respectively while *S. haematobium* prefers the small venules around the bladder and the ureter (Cook *et al.*, 2003). In the case of *S. japonicum* and *S. mansoni*, once male and female have paired they migrate against the blood flow through the hepatic portal vein and towards the mesenteric branches around the intestine. Once they have reached sexual maturity, male worms are 6-13 mm long and 1 mm wide while females are typically between 10-20 mm long and 0.16 mm wide. Female worms are easily recognisable even to the naked eye as they appear more slender than the males and show a noticeable dark pigmentation in their gut (**Figure 1.2B**). Other distinctive phenotypic characteristics are the oral (for feeding) and ventral (for attachment) suckers in both genders and the gynaecophoric canal in the male, where the female resides. The female lays eggs continuously and in close proximity to the intestine's endothelia to facilitate their migration through the gut wall into the intestinal lumen through which the eggs finally reach the exterior in the excreta. *S. mansoni* and *S. haematobium* females produce up to ~300 eggs a day while *S. japonicum* can produce up to 10 times more (~3,500). Eggs are about 100-150 um long, with each species presenting a characteristic shape (**Figure 1.2A**) commonly used in the diagnostics of the disease [reviewed in (Davis, 2002)]. Approximately 50% of the eggs are not excreted and are retained within the host tissues. Eggs are passively taken by the blood flow towards the liver (*S. japonicum* and *S. mansoni*) or bladder (*S. haematobium*) where granulomas are formed as a consequence of the host's immune response against egg secreted antigens. The formation of a granuloma around the egg in the host tissues is the cause of pathology in all *Schistosoma* infections. Granulomas are organized agglomerations of cells (eosinophils, macrophages, CD4+ T cells) and collagen fibres, whose main objective is to isolate the egg (source of antigen) from the host. The egg eventually dies and the granuloma resolves itself leaving a fibrotic plaque. As infection progresses, more and more granulomas are formed and resolved and the affected tissue becomes fibrotic causing the pathology [reviewed in (Pearce *et al.*, 2002)].

After contact with fresh water, eggs hatch into the second free-living stage called miracidium. Miracidia can actively swim and are infective to snails for 8-12 hours after hatching. Once in the snail, miracidia transform into mother sporocysts and after eight days, germ cells bud off from them and develop into daughter sporocysts. Another round of germinal-cell production generates the cercariae, which are released into the fresh-water in response to environmental cues [reviewed in (Davis, 2002)]. The stage in the snail host represents cycles of asexual reproduction.

As with miracidia, cercariae are short lived and need to find a suitable host within hours. Host encounter is promoted by the active and intermittent swimming of the cercariae [reviewed in (Curwen *et al.*, 2003)]. Once the larvae have found a suitable host, probably aided by the presence of host fatty acids (Haeberlein *et al.*, 2008), proteases are secreted to assist the penetration of the cercarial heads in the host skin (Salter *et al.*, 2000; McKerrow *et al.*, 2002; Curwen *et al.*, 2006; Hansell *et al.*, 2008). At this point, the tails are cast off and the cercarial heads transform into schistosomula. The process of transformation involves mainly the remodelling of the surface membrane, which has consequences both in the parasites' anatomy (i.e., generation of a double outer bilayer) and physiology (i.e., schistosomula are water-intolerant). The passage through the skin can take from few minutes and ends when schistosomula find and penetrate into a venule.

Between four and seven days after infection parasites are found in the vasculature of the lungs where they stay for at least two to three days (Miller *et al.*, 1980). After the lung stage, parasites migrate to the hepatic portal system where they reduce their size and are now phenotypically more similar to the schistosomula at day 0 post transformation. The first parasites to arrive at this location do it around the tenth day after infection. From the moment of host invasion until day 10 to 11, parasites show a decrease of many metabolic markers/parameters, such as wet weight and oxygen consumption (Lawson *et al.*, 1980). At this point, male and female worms are morphologically distinguishable but not yet sexually matured. Worms of opposite sex pair up and migrate together against the blood flow towards the mesenteric veins where they attach and reside.

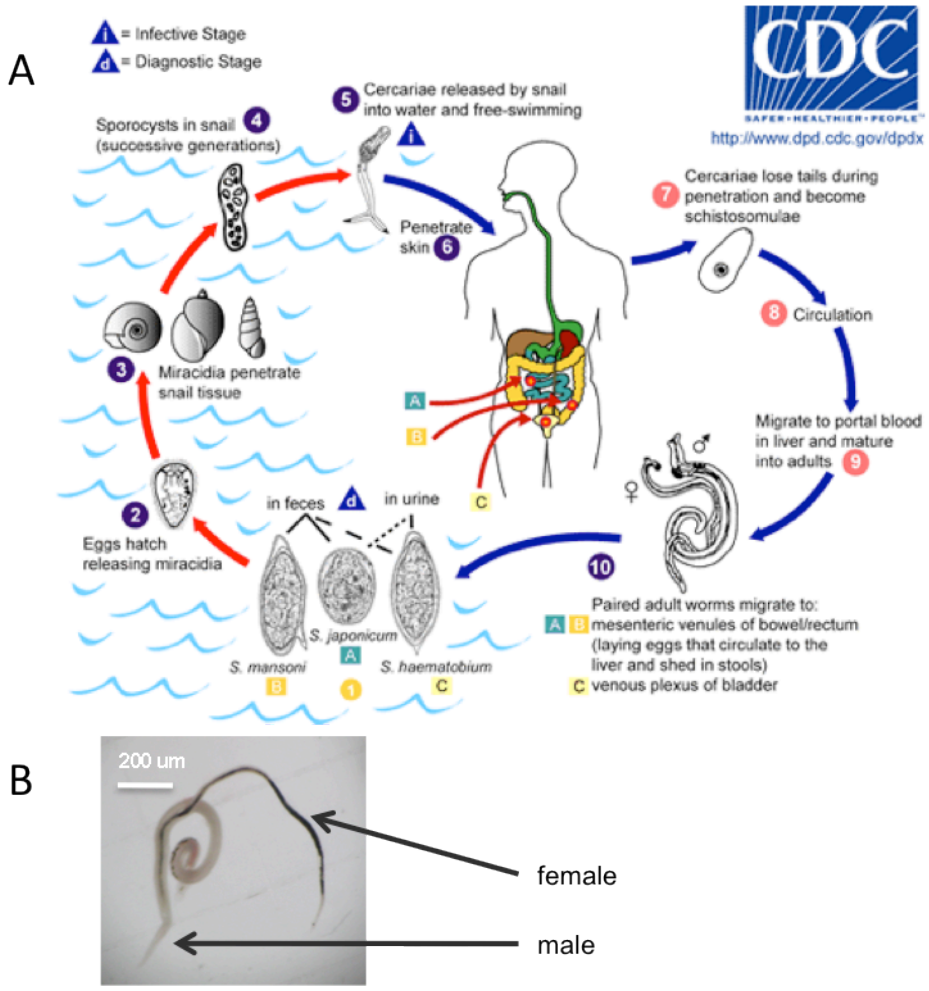


Figure 1.2 – Life cycle of *Schistosoma* spp. A - Life cycle of *S. mansoni, S. japonicum* and *S. haematobium* [reproduced from (CDC, 2011)]. These three parasites have a very similar life cycle. Eggs are released in the urine or faeces of infected humans (1). Upon contact with fresh water, the miracidia are released (2). Once they have infected a susceptible snail host (3), start a series of asexual reproduction stages occur (4) finally generating the human-infective larvae called cercariae (5). These are released by the snail into the freshwater and infect humans and other species by penetrating intact skin (6). The cercariae lose their tails and transform into schistosomula (7). After a short passage through the skin, the parasites reach the circulation (8) migrating to the lungs (9) and later the portal system where they develop. Finally, male and female worms pair up and lodge themselves (10) in the mesenteric veins of the portal system (*S. japonicum* and *S. mansoni*) or the bladder tributaries (*S. haematobium*). Females produce hundreds (*S. mansoni* and *S. haematobium*) to thousands (*S. japonicum*) of eggs. B - *S. mansoni* male and female worms partially attached. The female worm is more slender and appears darker than the male.

1.2.2 Chemotherapy and control of schistosomiasis

The main objective of chemotherapy is to restore the patient's wellbeing and reduce transmission of the infection. Anti-schistosomiasis chemotherapy introduced in the late 1960s fulfilled these criteria. Current methods of treatment of infected individuals produce the death of adult worms and the concomitant reduction in eggs deposited in tissues and hence reduction of pathology. At the same time, reduction in egg laying generates a break in the parasites' life cycle reducing community infections and therefore improving the health of the population. However, in real-life treatment programmes the situation is rarely this idyllic: in mass treatments usually only 50-60% of the population is cured – partly due to non-compliance - and this situation commonly leads to re-infection which is further favoured by poor sanitary conditions (Davis, 2002).

Praziquantel is the drug of choice because it is very effective against the adult worms of the three main *Schistosoma* spp., is also cheap and well tolerated. Patients may suffer from side effects, but these are temporary and have no long-term consequences. There is a common fear amongst *Schistosoma* researchers and public health officers that resistance to praziquantel (PZQ) may develop in the near future (Davis, 2002). Reduced susceptibility of *S. mansoni* worms to PZQ has been reported in the field (Ismail *et al.*, 1996; Melman *et al.*, 2009) and it has been proven that PZQ resistance can be induced in experimental conditions (Ismail *et al.*, 1994); raising the possibility that a similar situation could be also seen in the field.

Antioxidant pathways are known to play an important role in the survival of schistosomes in the oxidative environment of the blood stream where they reside (Loverde, 1998). Research into these pathways led to the identification of one chokepoint in the parasites' antioxidant portfolio, the thioredoxin glutathione reductase, which has been proven a lethal target for all intra-mammalian stages of the life cycle of *S. mansoni* (Sayed *et al.*, 2008). Further studies into essential pathways may lead to the discovery of new targets for intervention.

1.2.3 Insights into the cercariae and skin schistosomula stages

The following sections describe in detail the anatomy and physiology of the cercariae and the schistosomula. These stages are the focus of this work.

1.2.3.1 Cercariae

Cercariae, the human infective stage, emerge from the infected snail in response to light stimulus. They are typically 500 um long but can vary due to their great capacity of contraction/extension (Dorsey *et al.*, 2002). Once in the water, the cercariae move with sudden upwards motions followed by passive sinking (Graefe *et al.*, 1967; Cook *et al.*, 2003). The body of a cercaria consists of a head and a tail (**Figure 1.3**) and is well adapted for the task of invasion: the head can penetrate the host skin whereas the tail is lost.

Anatomically, the head can be divided into three parts: oral, middle and aboral (**Figure 1.3**). The oral part contains the oral sucker, the mouth and a strong musculature structure that is thought to assist migration throughout the skin. A ventral sucker (commonly referred to as acetabulum) is located towards the aboral part [reviewed in (Stirewalt, 1974)]. Pre- and post- unicellular acetabular glands are found anterior and posterior from the acetabulum with their cytoplasmatic processes extend towards the oral sucker. These are used to release peptidase-containing vesicles to the exterior [(Fishelson *et al.*, 1992), see section 1.2.3.1.1].

The tail is a highly specialized organ, which provides motility to the parasite during its free-living stage in the water. This organ is packed with myocytes, neurons, osmoregulatory cells and supporting cells. Myocytes are organised in an inner longitudinal, a subtegumental and three outer circular muscle layers that form the tail musculature structure. The tail must fulfil high-energy demands and is packed with large mitochondria, large numbers of ribosomes and glycogen (Dorsey *et al.*, 2002).

1.2.3.1.1 Cercariae secretions during skin penetration

During cercariae development in the snail, the most prominent differentiating cells are the ones that would give origin to the pre- and post-acetabular glands. These are unicellular structures that have their cell body located either posterior or anterior to the acetabulum or ventral sucker and whose cytoplasmatic processes extend and open at the apical end of the cercarial head (**Figure 1.3B**). Secretory products are packed in vesicles and reach the exterior through the cytoplasmatic processes. These vesicles burst upon contact with the hosts skin and expose proteases, which diffuse through the dermal extracellular matrix. The main function of these secretions is to aid the penetration of the cercariae across the different layers of the skin barrier. Many of the secretion products have been found to elicit an immune response [e.g, the *S. japonicum* paramyosin (Gobert *et al.*, 1997)] and therefore have received much attention from the research community

(Curwen *et al.*, 2003; McKerrow, 2003; Curwen *et al.*, 2006; Hansell *et al.*, 2008) due to their potential as vaccine targets.

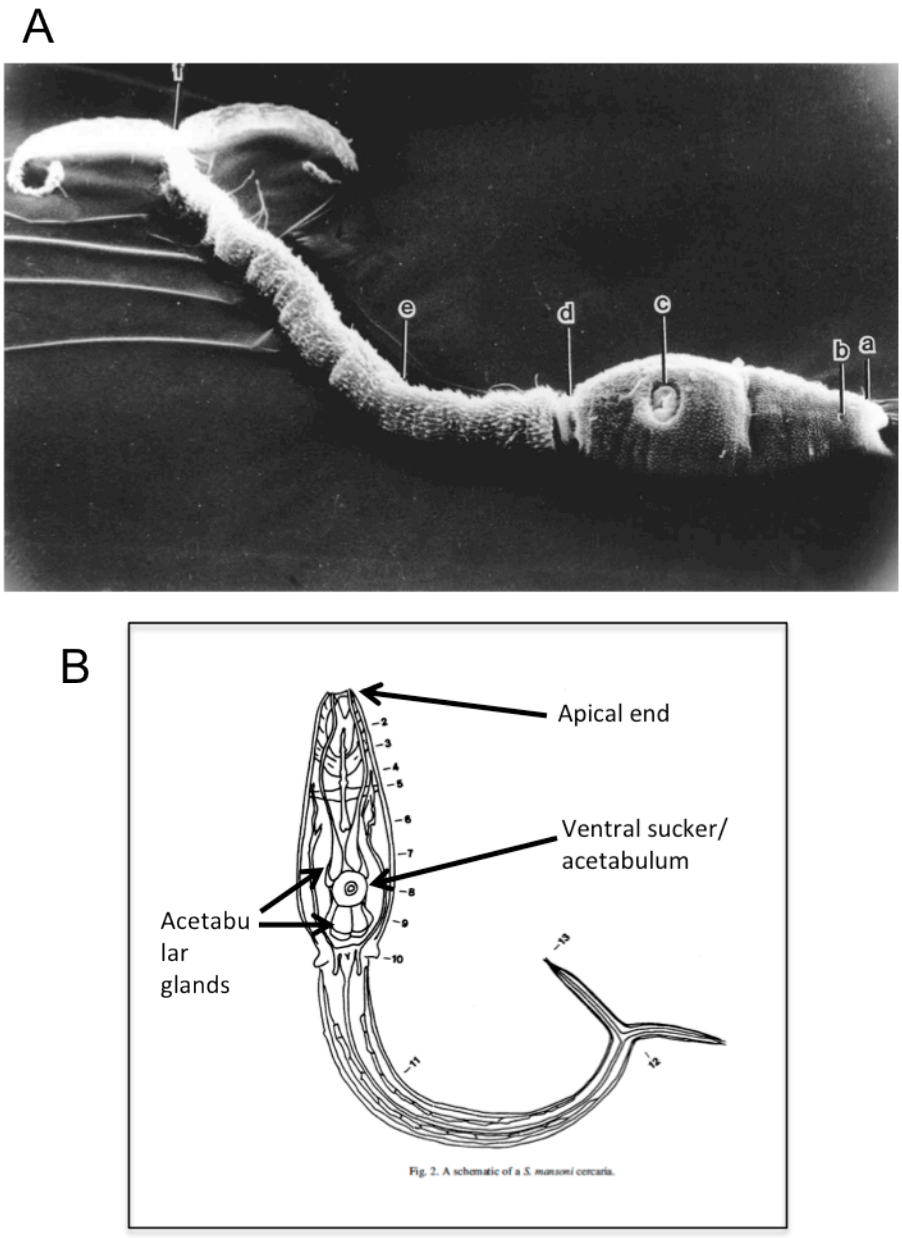


Figure 1.3 – Body organization of the cercariae. A - Scanning electron microscopy [reproduced from (Dorsey *et al.*, 2002)]. a – oral sucker; b – mouth; c – ventral sucker or acetabulum; d - body/tail junctions; e – tail; f – bifurcated tail. B - Schematic representation of the cercariae. Note how the cell bodies of the acetabular glands are located next to the acetabulum but their cytoplasmatic processes extend towards the apical end of the cercarial head [modified from (Dorsey *et al.*, 2002)].

Because the secretions are found to cover the penetration tunnel, they are also found in association with the parasite's surface and they might contribute towards the surface transformation of the parasite (Fishelson *et al.*, 1992).

1.2.3.1.2 Cercarial tegument

The general structure of the tegument (**Figure 1.4**) is similar in the cercariae and the adult worms (Stirewalt, 1974) and covers the entire surface of the parasites. Much of the research done regarding the tegument has been focused around the host-parasite interaction in the adult worms [reviewed in (Skelly *et al.*, 2006)], but most of the general characteristics are also valid for the cercariae. The tegument is located immediately underneath the apical plasma membrane (either cercariae, schistosomula or adult worm) and above the basal membrane. It is a syncytium (a continuum of cytoplasmatic material) with its cell bodies located deep under the muscle layers (Dorsey *et al.*, 2002). The tegument contains lipids, carbohydrates (glycoproteins and glycolipids) and many different proteins including enzymes and receptors. However, there is no evidence of DNA or RNA molecules found in the syncytium, suggesting that protein synthesis occurs exclusively in the cell bodies. Thin cytoplasmatic processes communicate the nucleated cell bodies with the tegument (Skelly *et al.*, 2006).

There are five types of tegumental cells, named I to IV plus the head gland. They connect to the syncytium through their cytoplasmatic processes, which are lined with microtubules. The different types of cells have different collections of vesicles and inclusion bodies. For example, one of the cell types contains multi-laminated vesicles thought responsible for the generation of the double bilayer, while others are packed with biogenic amines or ribosomes (Dorsey *et al.*, 2002).

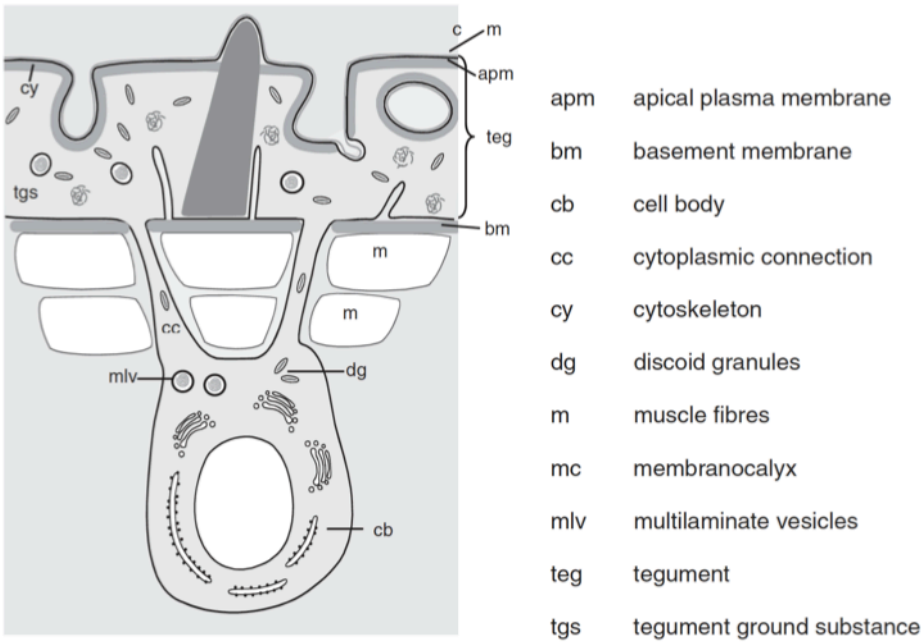


Figure 1.4 – Schematic organization of the tegument in a transversal cut. The cell bodies of the tegumental cells are located underneath the muscle layers and connected to the syncytium through their cytoplasmatic processes The tegument is located between the basal membrane and the apical membrane. The glycocalyx (or membranocalyx) is the outermost layer. [reproduced from (Skelly *et al.*, 2006)]

1.2.3.1.3 The glycocalyx, glycocalyx shedding and tegument dynamics

The glycocalyx is the outer-most layer of the parasite surface and it is tightly connected to the tegument. Transmission microscopy experiments have shown that the glycocalyx forms a layer 1-2 um thick with fibrils 15-30 nm long. It is of very high molecular weight (in the order of millions) and it bears antigenic properties (Samuelson *et al.*, 1985). Different staining techniques have shown that the glycocalyx is of carbohydrate or possibly lipopolysaccharide nature (Samuelson *et al.*, 1985); with fucose and glucose being the major components of the head and tail respectively (Nanduri *et al.*, 1991). The glycocalyx protects the cercariae from the water environment and once inside the host skin, it may protect the transforming schistosomula from the attack of immune agents by trapping host molecules or acting as a physical barrier between the parasite's tegument and the host environment. The loss of the glycocalyx, which happens only after the newly acquired double bilayer membrane is formed, is necessary for the transformation of the cercariae into schistosomula. Hence, it is possible that it has a role in protecting the parasite during membrane remodelling (Skelly *et al.*, 2006).

Three main events are known to contribute towards the loss of the glycocalyx layer. These were described in detailed by Fisherson *et al.*, (1982) and a summary is presented below.

The first event is the thinning of the glycocalyx layer by the action of cercariae secretions (Samuelson *et al.*, 1985). These seem to originate from the secreted vesicles from the pre- and post-acetabular glands. Because these vesicles are released during penetration, probably triggered by the presence of specific fatty acids in the surface of the host (Stirewalt, 1978), the vesicle's contents are found covering the penetration channels and therefore can get in contact with the cercariae' surface (Hansell *et al.*, 2008). The thinning glycocalyx makes way to the exposure of spines covering the surface of the parasite. In transformed schistosomula, these spines can be clearly seen (Crabtree *et al.*, 1980). A similar effect is observed when cercariae are concentrated in aquarium water (Fishelson *et al.*, 1992). Concentration of cercariae by centrifugation is known to trigger the release of the contents of the acetabular glands producing the "artificial" thinning of the glycocalyx as cercarial proteases are found within the cercariae pellet. Second, microvilli start extending from the surface of the cercarial body and through the glycocalyx layer. These extensions may appear as soon as five minutes after transformation and are also shed from the parasite within 60 minutes. They are transient structures; the glycocalyx is anchored to them and when the microvilli are shed, part of

the glycocalyx is shed with them. The molecular structure of the glycocalyx is not damaged during this process (i.e., this is not a degradation event). Thirdly, multi-laminated vesicles originating from the tegumental cells reach the surface and discharge their contents. The migration of the vesicles occurs at the same time as the microvilli formation and it may be linked to the generation of the double bilayer that will replace the single bilayer previously found. The shedding of the glycocalyx and replacement of the membrane occurs asynchronically and 1 hour after transformation, the parasite's surface still has both single and double bilayers distributed in patches (Fishelson *et al.*, 1992). With time, the whole of the membrane transforms into a double bilayer. In the maintenance of the parasites' surface after schistosomula stage, the multi-laminated vesicles continue to provide material for the double bilayer and the glycocalyx (Skelly *et al.*, 2006).

One hour after penetration of the host skin, the transformation of the cercarial surface is almost completed and the parasites show changes in aspects of their physiology. Through close examination of transforming parasites, Stirewalt *et al.,* (1966) described the schistosomula as saline and serum adapted tail-less organisms derived from the cercarial stage. These transformed parasites are water-intolerant and their acetabular glands are emptied after skin penetration. As previously described, the surface of the schistosomula has changed into a double bilayer (Stirewalt *et al.*, 1966) that looks as if heptalaminated under the light microscope.

1.2.3.2 Schistosomula

Schistosomula reach the boundary between the dermis and epidermis sometime between 30 minutes and two and a half hours after skin penetration. By then, the glands from the apical end still contain secretory granules (Dorsey, 1976). Many parasite-derived proteases and proteins are found in the epidermal tunnels and are thought to assist the migration process. Some of these are a serine protease called cercarial elastase/protease, a serpin (serin protease inhibitor), heat shock proteins (HSP) 86 and 70 and a couple of proteins putatively involved in immune evasion [paramyosin and Sm16 – a protein found in the secretions of invading cercariae (Holmfeldt *et al.*, 2007)]. All these are released within the first half of the hour of the skin penetration process (Hansell *et al.*, 2008). Acetabular gland tubes, previously prominent in the cercariae, are no longer found at 40 hours post infection suggesting that certain secretory mechanisms disappear after invasion (Crabtree *et al.*, 1985). This is in agreement with their proposed role in transforming the surface membrane. Although passage through the skin is supposed to be a relatively rapid process, many parasites seem to take longer. Even when a population of

cercariae are let to penetrate live animal skin under experimental conditions, schistosomula were shown to still be in the epidermis after 24 or even 40 hours post infection (Crabtree *et al.*, 1985).

At 48 hours post-penetration, parasites that successfully migrated into the dermal layer of the skin still contain vesicles in the head glands. After locating a blood vessel it will take schistosomula approximately eight hours to penetrate through the blood vessel endothelia and reach the blood stream (Wilson *et al.*, 1980). The fact that head glands have not been completely emptied when the parasite arrives at the blood vessel endothelia suggests that some of the glands' contents may be used to disrupt the extracellular "cement" found in endothelial walls and make way into the venule. The parasite also displays rapid everted/inverted movements, with the head capsule showing a strong and prominent musculature that controls the protrusion of the apical end. Taken altogether these suggest that endothelium penetration is achieved by using a combination of chemical (contents of the head gland) and physical (movements resembling a "battering ram") action (Crabtree *et al.*, 1985).

1.2.3.3 Energy metabolism

In terms of energy production, free-living stages (miracidia and cercariae) use the same typical aerobic metabolism as other higher eukaryotes: glycogen is degraded to pyruvate through the classic glycolysis pathway [reviewed in (Barrett, 1981)]. Pyruvate is then transported to the mitochondria where it enters the tricarboxylic acid cycle (TCA, Kreb's cycle) releasing CO_2 and producing NADH to feed into the respiratory chain [reviewed in (Tielens, 1994)] and generate ATP through oxidative phosphorylation. These mechanisms have been proven for both miracidia and cercariae stages in *Fasciola hepatica* and *S. mansoni* (Barrett, 1981).

During host-larval stages, such as that of the schistosomula, energy metabolism is mainly anaerobic but parasites retain their potential to use aerobic metabolism. It has been shown that approximately half of the production of L-lactic acid in the newly transformed schistosomula comes from oxidative phosphorylation, hence the metabolic switch seems to be an incomplete one (Coles, 1973). The cause for the change in the metabolism is not yet clear: some authors suggest that it is due to the availability of glucose, which increases the glycolytic flux (Barrett, 1981) while others suggest that it is linked to what triggers the transformation of the parasite as a whole [i.e., host fatty acids (Coles, 1973)]. Irrespective of the mechanism, this metabolic switch only occurs in the cercarial heads and not in the tails (Horemans *et al.*, 1991). Although the whole organism

has the same capacity for using different metabolic pathways, the tails use a different one from that used in the cercarial head. Tails have more cytochrome oxidase activity (Coles, 1973; Skelly *et al.*, 1993) to meet the energy demands of this specialized organ. It is commonly understood that the up-regulation of enzymes corresponding to anaerobic metabolism (i.e. lactate dehydrogenase) corroborates "the described switch in the larval metabolism from aerobic to anaerobic pathways during transformation" [from cercariae to schistosomula] (Lawson *et al.*, 1980; Farias *et al.*, 2011).

1.2.4 Pathology and Immunology of schistosomiasis

General characteristics of the pathology of schistosomiasis are directly related to the life cycle of the parasite in the human host (Table 1.1). These are different in non-immune individuals in comparison to individuals leaving in endemic areas where their exposure to continuous infection and re-infection episodes grants them certain level of protection (Cook *et al.*, 2003).

The Katayama fever syndrome is common to all schistosome infections affecting humans. It is most marked in the primary infections of individuals living in non-endemic areas. The period between infection and set off fever varies: for *S. japonicum* it ranges from two to six weeks while in *S. mansoni* is generally from three to seven weeks. Symptoms are basically those of an acute fever episode: continuing high body temperature, shivering/trembling, sweating, general muscles pain, headaches and in less percentage of cases: anorexia, nausea and abdominal discomfort (Cook *et al.*, 2003). These are the characteristics of a dominating T-helper 1 (Th1) response. This pro-inflammatory response features high levels of circulating tumour-necrosis factor (TNF), interleukin-1 (IL-1) and IL-6 produced by peripheral blood mononuclear cells. Approximately six to eight weeks after infection, adult worms are fully matured and egg laying starts. Eggs lodge in the liver where they produce secretions rich in carbohydrates (also known as schistosome egg antigen or SEA). These secretions induce a T-helper 2 (Th2) response which in turn down regulates the effector function of the Th1 pro-inflammatory response. This switch from Th1 to Th2 is vital for the survival of the host. However, long lasting Th2 responses, as occurs in endemic areas of schistosomiasis infections, are also cause of morbidity (Pearce *et al.*, 2002). Eggs trapped in the liver (*S. mansoni* and *S. japonicum*) or bladder (*S. haematobium*) also produce SEA causing the formation of granulomas (Cook *et al.*, 2003). As previously mentioned, these are organized agglomerations of cells such as eosinophils, macrophages, CD4+ T cells and collagen fibres. As infection progresses, more and more granulomas are formed and resolved leaving a significant number of fibrotic

plaques and a fibrotic liver. This leads to the blockage of the portal tracts causing portal hypertension. To compensate, liver capillaries are enlarged and branched leading to enlargement of the liver (hepatomegaly). Other complications arising from hypertension in the portal venous system include the spleen becoming tough, fibrotic and enlarged (splegnomegaly). The spleen may also become hyperactive (hyperslenism) causing the reduction of red and white cells, platellets and anemia [reviewed in (Pearce *et al.*, 2002 and Cook *et al.*, 2003)].

Table 1.1 - Summary of clinical manifestations in a *S. mansoni* infection.

DESCRIPTION	APPEARANCE CLINICAL MANIFESTATIONS (Time post-infection)	CHARACTERISTICS	IMMUNE EFFECTORS
Cercariae infection and schistosomula migration	24 to 48 hours	Cercarial allergic dermatitis	Eosinophilia and antibody-dependent cell-mediated cytotoxic response involving IgG.
Schistosomula maturation and establishment of paired adult worms	2 to 16 weeks	Febrile illness (Katayama syndrome)	Initiation of Th1, TNF, IL-1, IL-6.
Egg laying	From 2 months onwards	Granulomas	Start of Th2 response, down regulation of Th1.
Late staged of infection	Years	Portal hypertension and hepatosplegnomegaly (Sm)*	

In endemic areas, re-infection is more the rule than the exception. Children in school age and until puberty are the most heavily infected subgroup. Older people are usually less susceptible to re-infection and this protection to re-infection has been associated with high levels of immunoglobulin-E (IgE) directed against the adult worm. Interestingly, adult worms are not susceptible to the immune response of the host and therefore the slow development of resistance to re-infection agrees with the long life span of the adult parasites (Pearce *et al.*, 2002).

1.2.5 Vaccine development and vulnerability of schistosomes.

The search for a vaccine against schistosomiasis is based on the need to decrease the disease burden, which has not been reduced by chemotherapy. Additionally, high rates of post-treatment re-infection and the lack of an alternative to praziquantel raise the issue of the inadequacy of the intervention programmes so far used. The realisation of a successful vaccine is further encouraged by evidence of partial immunity against schistosome infections acquired by adults leaving in endemic areas [reviewed in (Hotez *et al.*, 2010)].

Previous efforts using irradiated cercariae have elicited immunity against a subsequent challenge infection in laboratory animals. However, the administration of a live vaccine represents many logistic inconveniences making it not easily viable in the field. Other approaches include the use of recombinant proteins such as the *S. haematobium* 28 kDa GST against urinary schistosomiasis, the *S. mansoni* 14 kDa fatty acid-binding protein, DNA vaccine Sm-p80 as well as several *S. mansoni* tetraspanins, all of which have been or are in early stages of clinical trials. However, these have had either limited success or are awaiting clinical trials that will assay their safety and efficacy [reviewed in (Hotez *et al.*, 2010)].

The debate about which of the life cycle stages is the most vulnerable to intervention has received some attention (Curwen *et al.*, 2003; McKerrow, 2003). However, it has been suggested that the schistosome's weakest point in the life cycle might reside in the early encounter of the parasite with its mammalian host; that is the skin schistosomula stage (Wilson *et al.*, 2009). This was based in the following principles. First, it has been shown that the schistosomula stage is susceptible to the attack of oxidative species rendering certain level of vulnerability (Loverde, 1998). Second, a series of studies reviewed by Capron *et al.,* (1986) showed that schistosomula killing in the rat model is mediated by antibody-dependent cell cytotoxicity involving IgE (Capron *et al.*, 1986). In summary, developmental changes undergone by the migrating schistosomula leave open

opportunities for the interaction with the host defences enabling both recognition and killing (Wilson *et al.*, 2009).

A prophylactic vaccine would have the capacity of priming the host's immune system against a challenging *Schistosoma* infection by introducing one or many substances that resemble structures found in the invading schistosomula. The host's immune system detects and attacks such structures but also remembers them. In the event of a challenge infection, the immune response is ready to act defending the host (Abbas *et al.*, 2000).

One of the current strategies for finding the magical bullet(s) that would lead to a successful vaccine is to look for molecules that would comply with certain characteristics. Apart from being expressed in the skin stage of the parasites, such potential antigens would have to be secreted or exposed in the parasites surface where the host's immune system can detect them. In the last five years, our understanding of which genes are being expressed upon infection has been improved by the many microarray studies targeting the cercariae and schistosomula stages (Chai *et al.*, 2006; Dillon *et al.*, 2006; Fitzpatrick *et al.*, 2009; Gobert *et al.*, 2010; Parker-Manuel *et al.*, 2011). Although these studies achieved good resolution in identifying differential expression of hundreds and thousands of genes within the first days of infection (3 to 7-10 days old schistosomes), only one study focused entirely in the first 3 hours post infection (Gobert *et al.*, 2010) when important changes in the surface of the parasite take place. If the early schistosomula stage is the one chosen to be targeted for intervention, a more thorough and comprehensive study of its gene expression portfolio would need to be achieved.

Current efforts in the development of intervention strategies base their platforms in combining the genome and the transcriptome information together with high-throughput expression studies (TheSchistoVac, 2009) to identify potential drug targets and vaccine candidates. To this end, the accuracy of both genome and transcriptome data is a fundamental aspect of the drug and vaccine development pipeline.

1.3 Genome biology of S. mansoni and current status

The *S. mansoni* karyotype comprises 7 pairs of autosomes and a pair of sex chromosomes. All chromosomes are distinguishable by their unique chromosomal banding patterns (Grossman *et al.*, 1981). Females are the heterogametic sex with both Z and W chromosomes while males have two copies of the Z chromosome.

The haploid genome size had been estimated to be approximately 270 Mb (million DNA base-pairs) (Simpson *et al.*, 1982), much larger than the nematode *Caenorhabditis elegans* [~100 Mb (*C. elegans* Sequencing Consortium, 1998)] and the fruitfly *Drosophila melanogaster* [~117 Mb (Adams *et al.*, 2000)]. The genome has a GC content of 35% and is 40% repetitive (Berriman *et al.*, 2009), with retrotransposons being the most commonly repeated element found in the genome (Simpson *et al.*, 1982).

The quest for large-scale gene discovery started in mid 1990s with the first publication of an Expression Sequence Tags (ESTs) study (Franco *et al.*, 1997). In the early studies, and limited by the technology available at that time, only a restricted number (466 unique genes) of genes could be discovered (Franco *et al.*, 1997). Other ESTs projects both in *S. mansoni* [i.e., (Franco *et al.*, 1997; Santos *et al.*, 1999)] and *S. japonicum* [i.e., (Fan *et al.*, 1998; Fung *et al.*, 2002)] followed and in 2003 probably the most comprehensive EST study in *S. mansoni* was published (Verjovski-Almeida *et al.*, 2003). In this publication, the authors generated ~160,000 ESTs with 31,000 assembled sequences spanning six different life cycle stages (cercariae, 7-day old *in vitro* cultured schistosomula, adult worms, egg, miracidia and germ-balls). With this work it was possible to estimate that the gene complement of *S. mansoni* would be around ~14,000 genes. However, the success of using ESTs for gene finding is limited mainly by the under-representation of low expressed transcripts (specially in normalized libraries such as those used in Verjovski-Almeida *et al.,* 2003) and lack of coverage over the full length of the transcript (specially towards the 5'-end). In addition, some life cycle stages are still poorly represented in existing EST resources. Furthermore, EST sequencing using conventional methods, such as the capillary Sanger method, is nowadays relatively expensive and laborious

The first published draft of the *S. mansoni* genome was assembled into 19,022 scaffolds with a gene complement of 11,809 genes and 13,197 transcripts (Berriman *et al.*, 2009), a similar figure to that obtained by a comprehensive study of *S. mansoni*'s ESTs (Verjovski-Almeida *et al.*, 2003). More recently the genome has been systematically improved. This was done using the original draft data, new capillary sequencing data and

second-generation genomic data produced from DNA obtained from single miracidial-derived worms, resulting in a much less fragmented assembly with only ~885 scaffolds (Protasio *et al.*, 2012). Furthermore, 86% of the improved assembly can now be allocated into physical chromosomes thanks to linkage markers (Criscione *et al.*, 2009) and fluorescence *in situ* hybridization of mapped BACs that had been previously generated (Berriman *et al.*, 2009). A summary of statistics from both assemblies is presented in **Table 1.2**.

Table 1.2 – Characteristics of the old and improved *S. mansoni* genome assemblies [reproduced from (Protasio *et al.*, 2012)].

	Old version[a]	**New version[b]**
Assembly size (Mb)	374.9	364.5
Proportion assigned to chromosome (%)	43	86
Contig statistics		
Number	50,292	9,203
Average length (kb)	7.5	39.4
N50 length (kb)	16.3	78.3
Largest contig (kb)	139.4	460
Scaffold statistics		
Number	19,022	885
Average length (kb)	20	411.9
N50 length (Mb)	0.8	32.1
Largest scaffold (Mb)	4.2	65.5 [c]

[a] Version 4.0 of the *S. mansoni* genome was the published draft genome (Berriman *et al.*, 2009). [b] Version 5.0 (Protasio *et al.*, 2012).

Both EST databases and the genome assembly provided a good tool kit for the investigation of gene expression profiles. The study of the transcriptome originated from EST projects and greatly advanced by the availability of microarrays quickly led to high-throughput studies that looked at genes expressed at different stages of the parasite's life cycle. These advances would help the study of the organism's basic biology and in the identification of pathways that could represent points of weakness for intervention (Hoffmann *et al.*, 2003; Gobert, 2010).

Microarrays have been used to achieve a systematic and quantitative approach to gene expression in *S. mansoni* (Fitzpatrick *et al.*, 2005; Dillon *et al.*, 2006; Fitzpatrick *et al.*, 2006; Vermeire *et al.*, 2006; Jolly *et al.*, 2007; Verjovski-Almeida *et al.*, 2007; Fitzpatrick *et al.*, 2009; Gobert *et al.*, 2010; Parker-Manuel *et al.*, 2011). They have some clear advantages - such as being able to study a large number of genes simultaneously, and for a larger number of individuals than possible with previous techniques. However, they also have certain disadvantages, when compared to high-throughput transcriptome sequencing: given that microarrays are (typically) designed based on known gene models or known genomic sequences, by definition, this type of survey is biased and it does not allow the identification of new transcripts. In addition, the study of related sequences and transcripts derived from alternative splicing is difficult to interrogate with microarrays due to cross-hybridization. Furthermore, microarrays rely on detection of continuous analogue signals that are difficult to quantify and often saturable; they depend on the inclusion of internal standards, and background fluorescence prevents the measurement of transcripts with low expression. Results are usually difficult to normalize between platforms and laboratories because each of them often uses different experimental designs (e.g, array design, glass slide, etc) (Shendure, 2008).

The advent of second generation sequencing technologies (also referred as "new" or "next generation" sequencing technologies) developed by several companies (Illumina/Solexa, 454-pyrosequencing/Roche, ABI/SOLiD and Helicos) has given new horizons to the study of functional genomics [reviewed in (Mardis, 2008; Morozova *et al.*, 2008)]. The year 2008, the same year this thesis work begun, saw the first reports on Illumina sequencing technology applied to gene expression quantification. These reports introduced the direct sequencing of the transcriptome through "RNA-seq" (Marioni *et al.*, 2008; Mortazavi *et al.*, 2008), which involves sampling the whole transcriptome of an organism at a given time point (or several) and subjecting it to high-throughput deep-sequencing [reviewed in (Wang *et al.*, 2009)]. To date, hundreds of publications have

featured the use of RNA-seq for gene calling, gene structure refinement, *de novo* transcriptome assembly and full characterisation of all kind of organisms' transcriptomes [i.e., (Otto *et al.*, 2010; Severin *et al.*, 2010; Chaudhuri *et al.*, 2011; Xia *et al.*, 2011)] including some describing improvement to probably some of the most highly characterised genomes and transcriptomes such as *Caenorhabditis elegans* (Hillier *et al.*, 2009) or *Drosophila melanogaster* (Daines *et al.*, 2011). Even some parasitic worms have been subjected to this type of studies (Laing *et al.*, 2011) including *S. mansoni* (Almeida *et al.*, 2011; Protasio *et al.*, 2012).

Briefly, Illumina uses massive parallel sequencing through the sequencing-by-synthesis technology (Bennett *et al.*, 2005) capable of great depth of coverage and ultra high-throughput. The basis of the reaction chemistry is shown and explained in **Figure 1.5**. Sequencing yield and quality has been dramatically improved over the recent years due to the ongoing research and development that is put into this technology. In January 2008 the state of the art in production pipeline at the Wellcome Trust Sanger Institute was of 3 million 37 bases single end reads per lane. By July 2011, the state of the art sequencing in the same facilities was of 120 million 108 bases paired end reads per lanes.

As previously mentioned, RNA samples can be sequenced using this approach. In the case of eukaryotic samples extracted RNA, usually performed by standard methods such as extraction with TRIzol® or column based methods, is subjected to a selection process that enrich the samples in polyadenylated molecules, which will contain mainly protein-encoding mRNAs. These are later fragmented and double-stranded DNA is generated based on the RNA sequences (Mortazavi *et al.*, 2008). After ligation of the appropriate adapters the DNA molecules are ready to be sequenced as previously described (**Figure 1.5**).

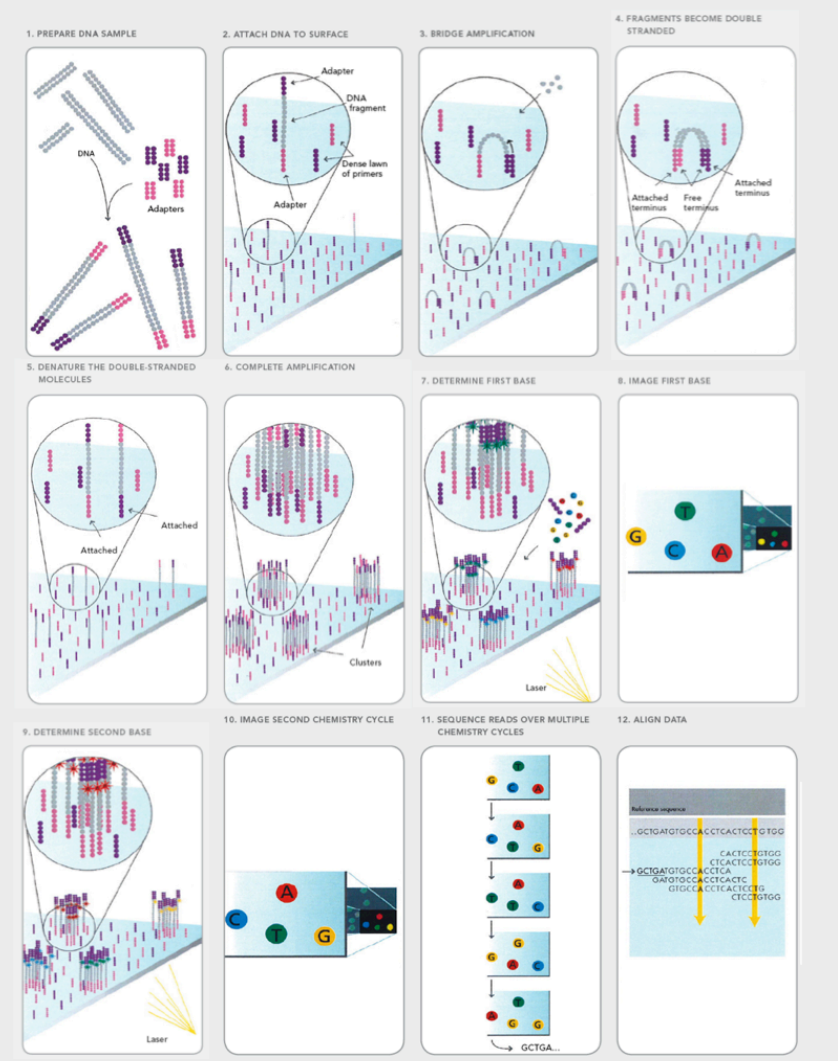


Figure 1.5 – DNA sequencing using Illumina technology. A – This series of panels explains the steps taken to sequence DNA using high-throughput Illumina technology. Double stranded DNA is randomly fragmented, size selected and adaptors are ligated to each end of the molecules (1). These molecules are bound (single-stranded) to a flow cell containing a dense lawn of previously bound primers (with same sequence as the adapters - 2). The amplification reaction occurs through cycles of "bridging" and PCR reactions (usually referred to as bPCR) with unlabeled nucleotides (3-5). Repeated cycles produce a "polony" (cluster or colony of molecules), one for every molecule that was initially fixed to the flow cell (6). Then, the actual sequencing begins: labelled reversible terminators (nucleotides) are added together with primers and all reagents need for an amplification reaction. Given the nature of the terminators, only one base can be added at a time and the rest of the reagents are washed off. The laser excites the labelled molecules unblocking them and a camera captures the emitted fluorescence and coordinates of the emission in the slide. Then, a second round of sequencing can take place (7-10). After a number of cycles, a sequence of "colours" is recorded for a particular position in the slide representing a read (11). Alignment algorithms can be used to map these reads to a reference (12). For paired-end sequencing, a second round of sequencing takes place using the opposite adaptor as primer and therefore sequencing the other end of the same molecule (not shown). Images were reproduced from (Illumina, 2011).

1.4 What this thesis is about

The work of this thesis concentrates on three main topics. The first topic is the application of *S. mansoni* RNA-seq data to assist and improve the structural annotation of genes in the recently upgraded genome assembly. Four time points in the life cycle of the parasite were sampled and sequence. These data was used to resolve gene structures to a single-base resolution. Results obtained from this work are featured in Chapter 3 and have been recently accepted for publication in a peer-reviewed international journal (Protasio *et al.*, 2012).

The second topic that this thesis deals with is the equivalency of skin-transformed and mechanically transformed schistosomula. As previously introduced, cercariae transforms into schistosomula as it penetrates the skin barrier of the definitive host. This transformation can be mimicked in the laboratory by application of shear pressure to a cercarial sample. Because most downstream application use mechanically transformed parasites, it is important to understand what are the differences between these and more naturally transformed parasites that would better resemble a natural infection. Often very low numbers of parasites are obtained from skin samples *in situ*. Hence, the samples analysed here were obtained using a modified skin transformation method that allows recovery of a significant number of parasites that have been transformed by penetrating through host skin. These results are presented in Chapter 4.

The third topic focuses on RNA-seq transcriptome analysis applied to the study of a short time course involving cercariae, 3-hours old and 24-hours old schistosomula life cycle time points. The aim of this analysis was to identify processes that had been previously missed by less sensitive techniques, such as microarrays. Focus is done in aspects of the transcriptome that may have a role in assisting the adaptation of the parasites to the new environment in the mammalian host. These results are presented in Chapter 5.

Finally, Chapter 6 reviews the contribution of this thesis work and placed its main findings in the context of current knowledge of schistosome biology.

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.

A

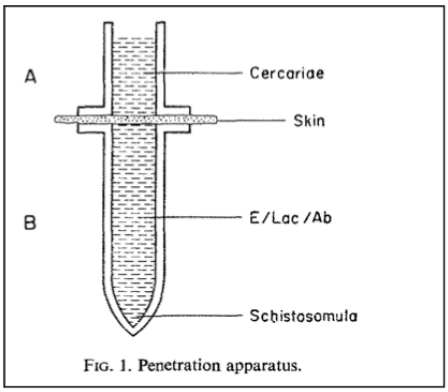
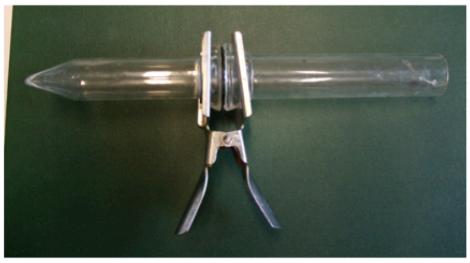


FIG. 1. Penetration apparatus.

B



C

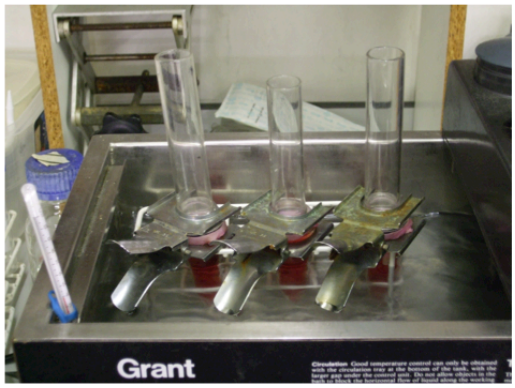


Figure 2.1 - Diagram and photographs of the skin transformation assemblies. A - Graphical representation of a transformation assembly (reproduced from Clegg *et al.,* (1972). B – Photograph of one of the transformation assemblies prior to use. C – Transformation assemblies in use during an experiment 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)

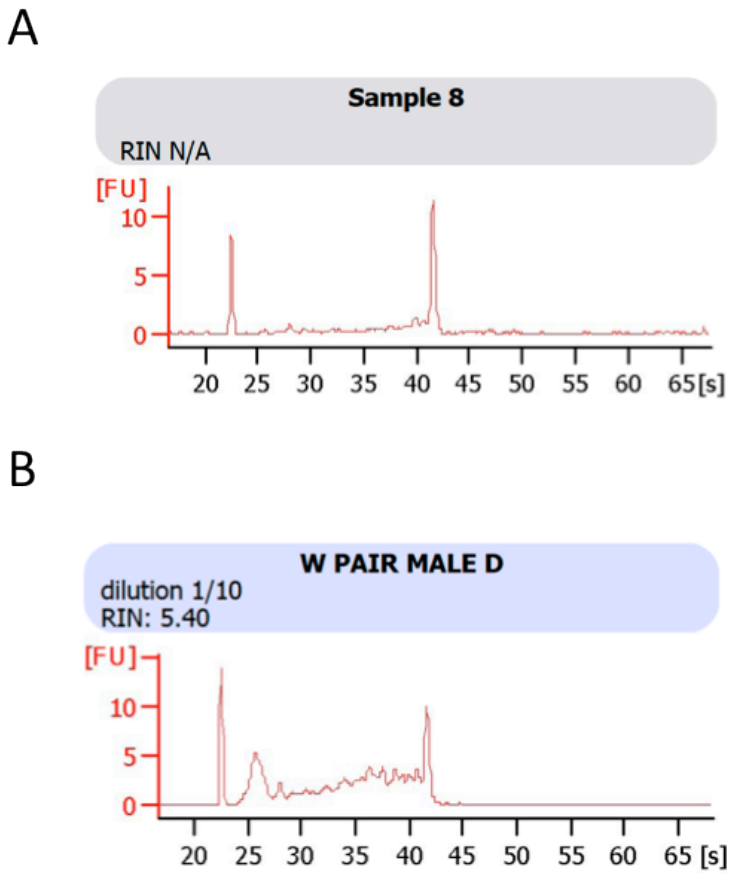


Figure 2.3 – The library preparation procedure requires quality control or RNA samples. A – Electropherogram of total RNA used for the library preparation of somule1 isolated using a QIAgen column. From left to right, the first peak corresponds to a size marker and the second peak corresponds to the 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.

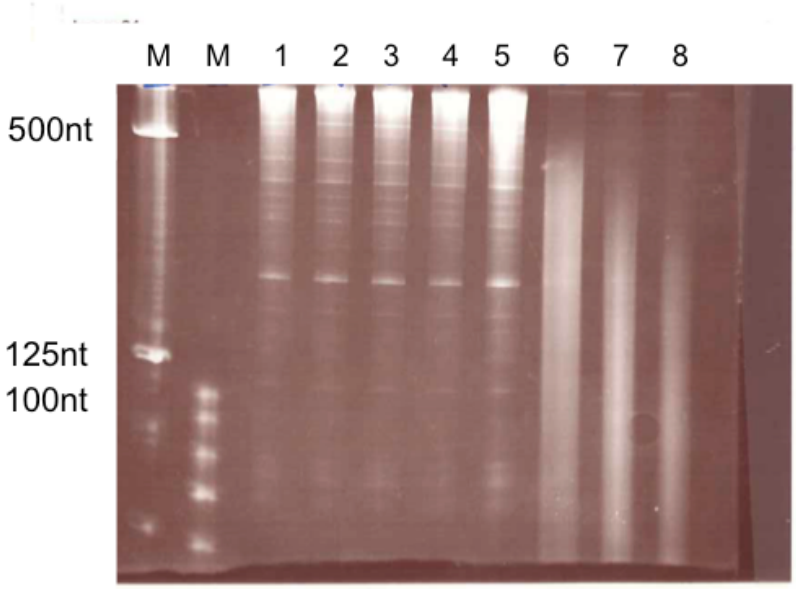


Figure 2.4 – Test on fragmentation of RNA for RNA-seq library preparation. Chemical fragmentation of total RNA is performed in the presence of salts at high temperature (70°C). Lanes 1 to 5 are controls of non-fragmented RNA; lane 6, 7 and 8 are RNA samples incubated at 70°C for 5, 10 and 15 minutes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

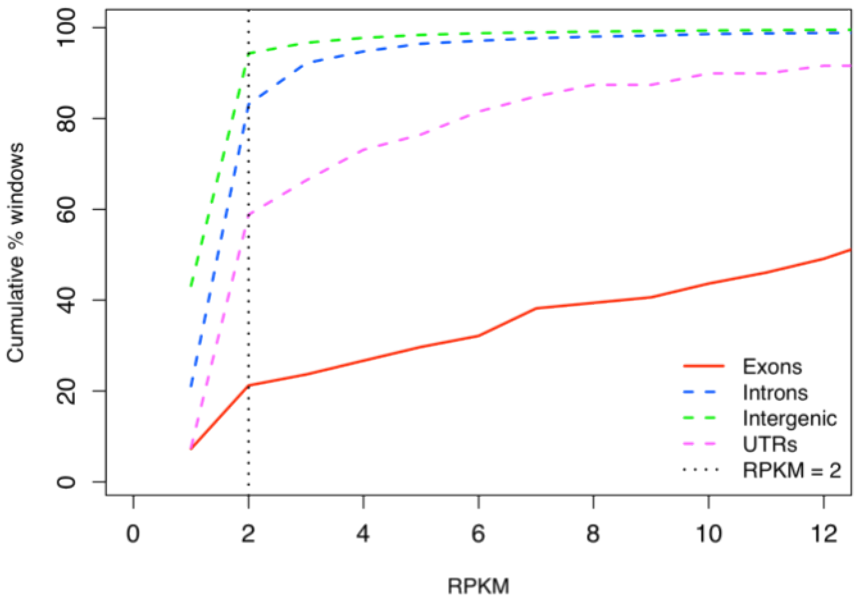


Figure 2.5 – Empirical calculation of minimum RPKM value. Method developed by Dr. Adam Reid (Pathogens Genomics group, WTSI).

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.

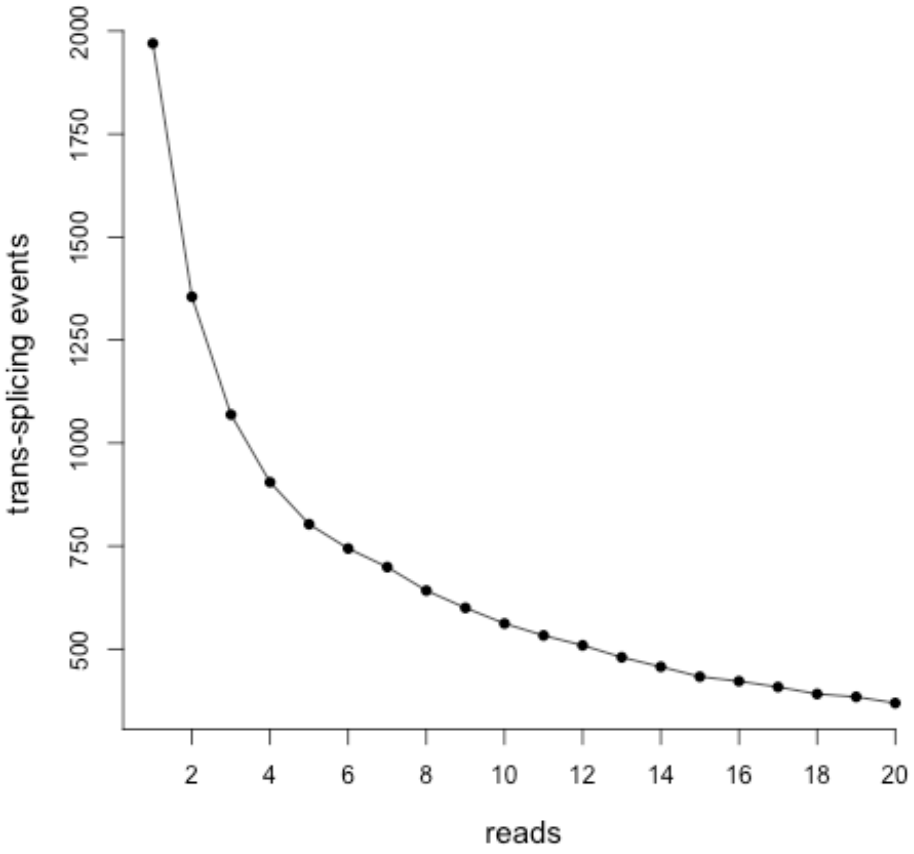


Figure 2.6 – Total number of *trans*-splicing events detected depends on the number of reads supporting such event. Almost half of the *trans*-splicing events are supported by fours reads or less [plot suggested by Mr Ferenc Kiss – University of Wurzburg, Germany].

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.

CHAPTER 3

TRANSCRIPTOME SAMPLING AND ITS IMPACT ON GENE ANNOTATION

3.1 Introduction

Previous to the publication of the improved genome and transcriptome (Protasio *et al.*, 2012) the gene complement present in the *S. mansoni* annotation was a result of *ab initio* predictions and a collection of ESTs mapped to the genome [reviewed in (Haas *et al.*, 2007)].

RNA-seq quantification of gene expression relies heavily on the accuracy of the gene models and therefore it was imperative to improve the gene annotation prior to gene expression analysis.

In the first part of this chapter shows results from the RNA-seq library sequencing, alignment of reads to the genome and analysis of technical and biological replicates are presented.

The second part of the chapter focuses on the contribution of RNA-seq data to the annotation and refinement of gene models and the identification of new coding sequences. Using RNA-seq data it was possible to generate a genome wide map of *trans*-splicing events. Furthermore, these data provided evidence of the transcription of genes as polycistronic units; a phenomenon so far suggested, yet not proven, for only one pair of *S. mansoni* genes.

3.2 Results

3.2.1 Sequencing results

Sequencing of RNA-seq libraries was performed in the Illumina Genome Analyzer IIx using paired-end sequencing with 76 cycles from each end. The yield and average GC content of each sequenced library are shown in **Table 3.1**. Libraries cerc10 and somule1 were sequenced on several occasions. In the case of cerc10 the repeated sequencing runs were performed because lanes 2711_5 and 2844_6 did not fulfil the quality control standards established by the Illumina® pipeline; only the last run of this sample (3012_1) passed the quality control and was considered for analysis. In the case of somule1 a second run of sequencing was required because of the low number of reads obtained in the first sequencing run. In this case both runs passed quality control and therefore are technical replicates (analysed in section 3.2.2.1.1).

Table 3.1 – Summary of sequenced libraries. MT – from mechanically transformation; ST – from skin transformation.

Sample	Life cycle stage	Lane_id	No. of reads	GC%
cerc10a	cercariae	2711_5	38,633,656	40.8
cerc10a	cercariae	2844_6	32,650,412	40.8
cerc10a	cercariae	3012_1	33,692,780	44.7
cerc12	cercariae	4485_5	61,554,460	42.1
cerc13	cercariae	4485_6	43,748,766	42.1
somule1	3-hour old schistosomula MT	3224_2	14,096,330	43.4
somule1	3-hour old schistosomula MT	4441_1	47,454,768	43.4
somule2	3-hour old schistosomula MT	4912_3	58,628,978	38.2
somule3	24-hour old schistosomula MT	4912_5	50,616,612	36.2
somule4	24-hour old schistosomula MT	4912_6	50,441,286	36.8
somule5	24-hour old schistosomula ST	4912_7	44,496,358	34.2
somule6	24-hour old schistosomula ST	4912_8	48,970,182	35.5
adult2	mixed sex 7-week adults	3224_1	14,515,880	35.5

The number of sequencing reads can vary greatly between sequencing runs with a general trend of increasing yield as the technology advances. The variation in the total number of reads per library has no effect on the comparisons between samples.

3.2.2 Transcriptome mapping results

RNA-seq data mapping to the genome was performed using TopHat as described in Chapter 2 section 2.6.1.1. **Table 3.2** shows mapping results for each sequencing run.

Total number of reads per sequencing run varied from 14.5 (adult2) to 61.5 (cerc12) million and the percentages of reads mapped to the genome varied between 41.7 and 61.9%. Variability in the number of reads mapped can be attributed to differences in the sample composition or the quality of the sequencing. For example, the cerc10 library was sequenced three times because the reads obtained in the first two runs did not meet base quality standards. The last run, 3012_1, had a higher percentage of reads mapping to the genome, reflecting a higher read quality. Additionally, another two libraries from cercariae samples were sequenced later on (cerc12 and cerc13) and these showed an even higher percentage of mapped reads (~64%). As an overall trend, sequence read quality improved over time and this was reflected in the percentage of reads mapped to the reference genome. Because normalisation approaches take the total number of mapped reads and not the number of sequenced reads, the variability in the number of reads obtained for each library does not affect any of the aspects of the analyses.

3.2.2.1 Analysis of replicates

As previously stated in Chapter 2, in this study has 3 types of replicates:

Case 1 - technical replicates type 1 – control for the reproducibility of the sequencing

Case 2 - technical replicates type 2 – control for library preparation protocol

Case 3 - biological replicates– control for biological variability in two preparations of parasites from the same time point

Reads per transcript were calculated as described in Chapter 2 section 2.6.1.1. These figures were used to calculate Pearson correlation coefficients between replicates of each type and these results were used to assess the reproducibility of the sequencing, library preparation and sampling of parasites.

Table 3.2 – Summary of TopHat mapping for each library/sequencing run showing total number (top) as well as percentages (bottom) – percentages were calculated based on sequenced reads. Numbers refer to individual (single) reads rather than read-pairs.

Lane_id	Sequenced reads	Total reads mapped	Reads mapped as proper pairs	Reads mapped as pairs	Reads mapped as singletons
3224_1	21,042,510	12,003,412	7,092,630	10,251,146	2,634,550
2844_6	32,650,412	13,619,758	7,396,302	9,836,672	3,783,086
3012_1	33,692,780	16,232,588	10,830,570	14,257,770	1,974,818
2711_5	38,633,656	19,386,692	12,290,106	16,752,142	2,634,550
3224_2	14,096,330	6,842,132	3,780,538	6,008,822	833,310
4441_1	47,454,768	25,869,874	14,477,418	23,103,974	2,765,900
4485_5	61,554,460	39,750,791	25,498,902	35,764,656	3,986,135
4485_6	43,748,766	28,265,977	21,389,366	25,699,466	2,566,511
4912_3	58,628,978	37,774,439	11,420,552	31,709,198	6,065,241
4912_5	50,616,612	31,280,263	9,565,606	25,064,526	6,215,737
4912_6	50,441,286	29,190,170	8,623,524	22,848,236	6,341,934
4912_7	44,496,358	23,639,063	6,723,982	17,209,924	6,429,139
4912_8	48,970,182	28,156,454	8,314,026	21,726,354	6,430,100
TOTAL	572,022,224	328,125,068	155,743,176	274,255,554	54,751,798

Table 3.2 (cont)

Lane_id	Total reads mapped (%)	Reads mapped as proper pairs (%)	Reads mapped as pairs (%)	Reads mapped as singletons (%)
3224_1	57.04	33.71	48.72	12.52
2844_6	41.71	22.65	30.13	11.59
3012_1	48.18	32.15	42.32	5.86
2711_5	50.18	31.81	43.36	6.82
3224_2	48.54	26.82	42.63	5.91
4441_1	54.51	30.51	48.69	5.83
4485_5	64.58	41.42	58.10	6.48
4485_6	64.61	48.89	58.74	5.87
4912_3	64.43	19.48	54.08	10.35
4912_5	61.80	18.90	49.52	12.28
4912_6	57.87	17.10	45.30	12.57
4912_7	53.13	15.11	38.68	14.45
4912_8	57.50	16.98	44.37	13.13
MEAN	56.15	27.69	47.04	9.41

3.2.2.1.1 Case 1 – Technical replicates type 1

With the purpose of increasing the number of reads obtained for one of the libraries, the somule1 library was sequenced twice in independent runs (lanes 3224_2 and 4441_1). Because the sequencing runs were done in different sequencing machines at different times and even differing greatly in yield, they serve as technical replicates. **Figure 3.1A** shows a scatter plot for the comparison of the reads per transcript obtained from sequencing lanes 3224_2 and 4441_1. Pearson's correlation between lanes is very high (0.9997) indicating that technical replicates were highly reproducible; which is in agreement with previous reports (Marioni *et al.*, 2008). Based on these results, it was concluded that further technical replicates would not be necessary for validating other libraries.

3.2.2.1.2 Case 2 – Technical replicates type 2

Libraries somule5 and somule6 were created from the same RNA extraction (24-hour old skin-transformed schistosomula) and therefore their correlation is a measure of the reproducibility of the library preparation method. **Figure 3.1B** shows the correlation for these two libraries. A Pearson's correlation value of 0.9895 indicates that the process of library preparation is highly reproducible introducing almost no variation.

3.2.2.1.3 Case 3 - Biological replicates

Samples somule3 and somule4 were obtained from independent parasite isolations from 24-hour old mechanically transformed schistosomula. By comparing these, biological reproducibility could be assessed. **Figure 3.1C** shows the correlation for these samples. A Pearson's correlation value of 0.9480 indicated that the process of parasites isolation and RNA extraction was highly reproducible. As expected, Pearson's correlation value for biological replicates was lower than that for technical replicates but still very high and also correlated with that obtained in other studies (Hebenstreit *et al.*, 2011).

In all of these test cases, the technical replicates were very highly correlated; indicating that sequencing of technical replicates to assess reproducibility could be avoided. However, biological replicates showed a slight lower correlation value, which could be arising from a combination of biological variation and technical variation from the library preparation process and the sequencing runs. Biological replicates cannot be avoided because they are key in providing statistical power in the differential expression analysis.

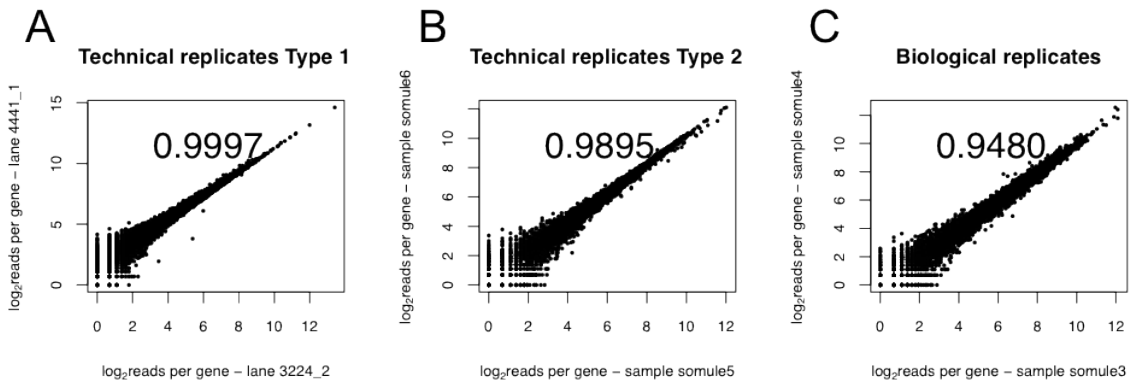


Figure 3.1 - Technical and biological replicates. Scatter plots of reads per gene obtained for pairs of replicates. Axes are in the logarithmic scale. A - Technical replicate type 1 – control for sequencing procedure. The same library (somule1) was sequenced twice (lanes 4444_1 and 3224_2). B – Technical replicate type 2 – control for library preparation protocol. The libraries were prepared from the same RNA sample. C – Biological replicate control for biological variability. Two RNA-seq libraries from different parasites' isolation are compared. Pearson correlarion values are indicated in the plot area.

3.2.3 Correlation with microarrays

As a well-established high-throughput tool, microarrays have been widely used in the study of gene expression in schistosomes (Hoffmann *et al.*, 2003; Fitzpatrick *et al.*, 2005; Chai *et al.*, 2006; Dillon *et al.*, 2006; Fitzpatrick *et al.*, 2006; Vermeire *et al.*, 2006; Jolly *et al.*, 2007; Verjovski-Almeida *et al.*, 2007; Fitzpatrick *et al.*, 2009; Gobert *et al.*, 2010; Parker-Manuel *et al.*, 2011). Since both RNA-seq and microarrays can be used for studying gene expression, the correlation between these two platforms was analysed.

The work of Fitzpatrick *et al.,* (2009) was chosen because, with exception of the skin-transformed schistosomula, it surveyed the same life cycle time points as the RNA-seq data presented in this thesis. Of the 35,078 unique 50-mer probes present in the oligonucleotide array, a total of 16,354 mapped to a unique location (with 100% identity) in the genome (for full description of Methods see Chapter 2 section 2.6.4). After calculating the number of reads for each oligonucleotide location, the correlation between the signal measured with microarray technology and that from RNA-seq was calculated for four time points in the life cycle (**Figure 3.2**). Spearman's rank correlation values are consistent across different life cycle stages and vary between 0.66-0.69.

A second microarray study has been recently published (Parker-Manuel *et al.*, 2011) and was also used to study the RNA-seq *vs.* microarray correlation. This later work featured a more comprehensive array design with a higher density of probes per gene. For the correlation with RNA-seq data, microarray data was processed in the same way as for Fitzpatrick *et al.,* (2009) with the exception that only the correlation with the cercariae sample could be performed. From all the 389,211 60-mer probes included in the array 377,598 were found to match unique locations in the genome with 100% identity. Spearman's rank correlation of the microarray's normalized intensity *vs.* RNA-seq reads is 0.67 and therefore agrees with the correlation value found for Fitzpatrick *et al.,* (2009) (**Figure 3.3A**). What is more, both correlations broadly agree in their distribution (compare **Figure 3.3A and 3.3B**) although some differences can be seen. The data from Fitzpatrick *et al.,* (2009) show a clustering of highly expressed probes (x-axis > 14) compared to the RNA-seq data. This effect is likely to be seen because there is a limit in the signal that can be detected by microarrays. This detection limit is not observed in RNA-seq data due to a much larger dynamic range than microarrays (Shendure, 2008; Wang *et al.*, 2009). In the latter, there are a finite number of molecules that can bind to the DNA probe and therefore there is a limit in the expression that can be measured. For example, lets imagine that a probe A can bind a maximum of 10,000 target molecules but the sample has 20,000 molecules that can hybridise to probe A. The result will be that only 10,000 molecules are detected. On the other hand, RNA-seq finds this limit in the number of reads that can be sequenced, and since the sequencing capacity of the current technologies gets better and better, the limit in the number of reads that can be measured gets higher and higher.

The array designed by Parker-Manuel *et al.,* (2011) did showed less signal saturation suggesting better performance at measuring highly expressed probes than their older counterpart. There is also a cluster of data points where the RNA-seq data, contrary to the microarrays, could not detect expression. This effect is more prominent in the Fitzpatrick *et al.,* (2009) data but it is also present in the Parker-Manuel *et al.,* (2011) data set. In this case, sequences with low expression detected by RNA-seq had microarray $\log_2$ intensity values ranging between 6 and 10, while in the Parker-Manuel *et al.,* (2011) dataset (**Figure 3.3**) these values are found in a narrower window corresponding to ~7 to 9. This is likely to be caused by hybridization issues. For example, non-base complementary hybridization could be caused by a combination of low GC content and non-optimal hybridization temperatures and generating a variable, yet not controlled for, number of

mismatches. This reflects the extent at which experimental conditions can affect the uniformity of a microarray experiment.

In summary, RNA-seq broadly correlates with the probe intensities found through microarrays. However, RNA-seq data showed greater resolution than microarrays in measuring both low and high expression values. The latter is a well-known limitation of microarrays (Shendure, 2008).

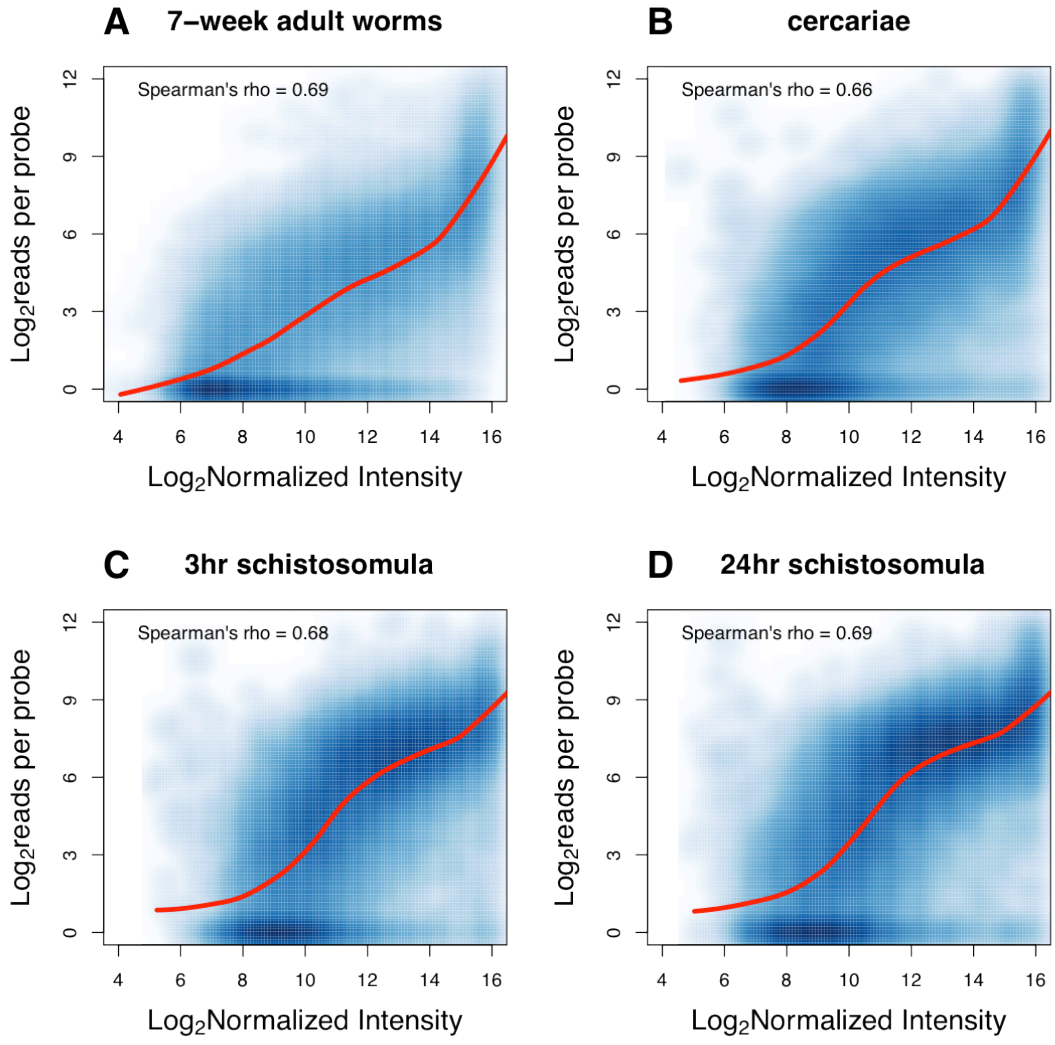


Figure 3.2 – Correlation of RNA-seq expression data with microarray from Fitzpatrick *et al.,* (2009). Each blue dot represents a position in the genome for which microarray expression data (x axis) and RNA-seq expression data (y axis) were calculated. A – 7-week adult worms; B – cercariae; C – 3-hour old MT schistosomula; D – 24-hour old MT schistosomula. The red lines indicate Lowess best-fit curve. Spearman's rank correlation values ranges from 0.66 to 0.69.

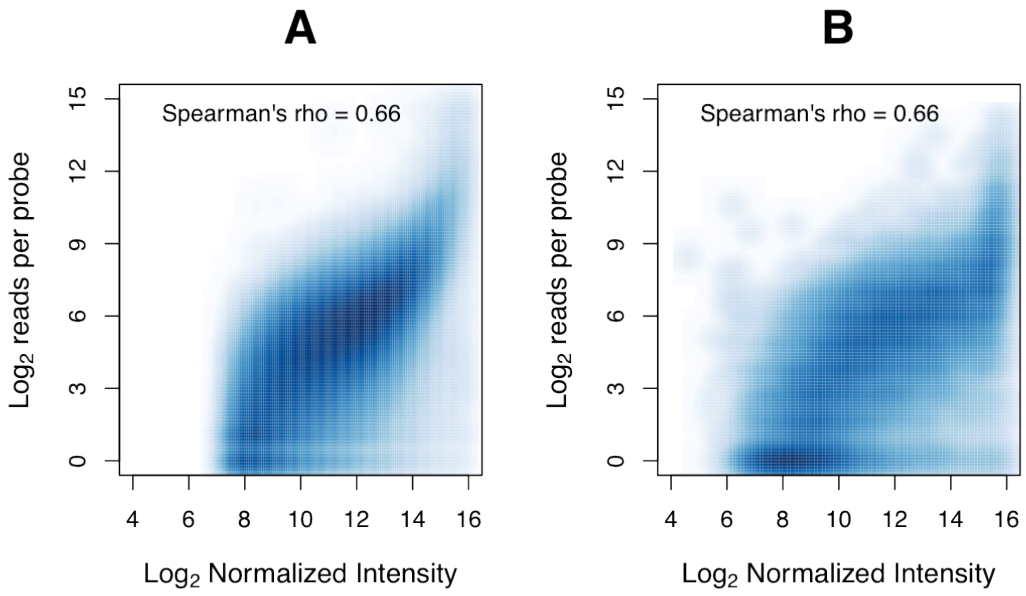


Figure 3.3 – Correlation of RNA-seq expression data for the cercariae sample against two different microarray platforms. Each blue dot represents a position in the genome for which microarray expression data (x axis) and RNA-seq expression data (y axis) were calculated. A – Microarray data from Parker-Manuel *et al.,* (2011); B - Microarray data from Fitzpatrick *et al.,* (2009).

3.2.4 RNA-seq contribution to gene annotation

As previously noted, one of the main advantages of RNA-seq data is that it can be used for gene annotation. The first part of the following section explains the status of the *S. mansoni* annotated gene-set prior to the use of RNA-seq data to refine gene models. The second part explains how RNA-seq data was used to assist manual curation of gene models and the implementation of a semi-automatic approach to scale up this refinement process into a high-throughput process. Annotation improvement led to the identification of new genes, which are also presented in this section. Finally, the identification of *trans*-splicing events and polycistronic transcripts is analysed.

3.2.4.1 The genome before RNA-seq

The previous version of the *S. mansoni* genome (version 4.0) contained 11,809 gene models and 13,197 transcripts. When the genome assembly was improved [version 5.0 - (Protasio *et al.*, 2012)] the software RATT (Rapid Annotation Transfer Tool) (Otto *et al.*, 2011) was used to migrate the gene/transcript structural and functional annotation from the old to the new assembly. Dr. Thomas Dan Otto and Dr. Isheng J. Tsai from the Parasite Genomics group performed the migration work presented here and also in the recently published (Protasio *et al.*, 2012). Their work is presented here as introductory information to provide the necessary context for sections 3.2.4.2 and 3.2.4.3.

RATT software is based on the conserved synteny that may exist between two genome assemblies and uses this to find the new location for a given annotation. Because the old and new *S. mansoni* assemblies are different, some genes present in the old assembly were no longer found in the new one mainly because of the loss of redundant sequence segments during the upgrade of the genome. In other cases, the less repetitive nature of the new version meant that some gene models were found overlapping each other (**Figure 3.4**). As a consequence of the migration, 10,569 unique models were transferred from the old to the new assembly while 841 were not. A total of 418 were partially migrated; in these cases only part of the model was found in the new assembly. These genes typically lacked 5' or 3' end sequences. Further curation of the migrated and partially migrated genes resulted in cases were genes had to be deleted or made obsolete primarily due to redundancy with other better models (**Figure 3.4**). A total of 516 models fell into this category (including partially transferred models) leaving 10,077 models present in the new version prior to the RNA-seq based improvement.

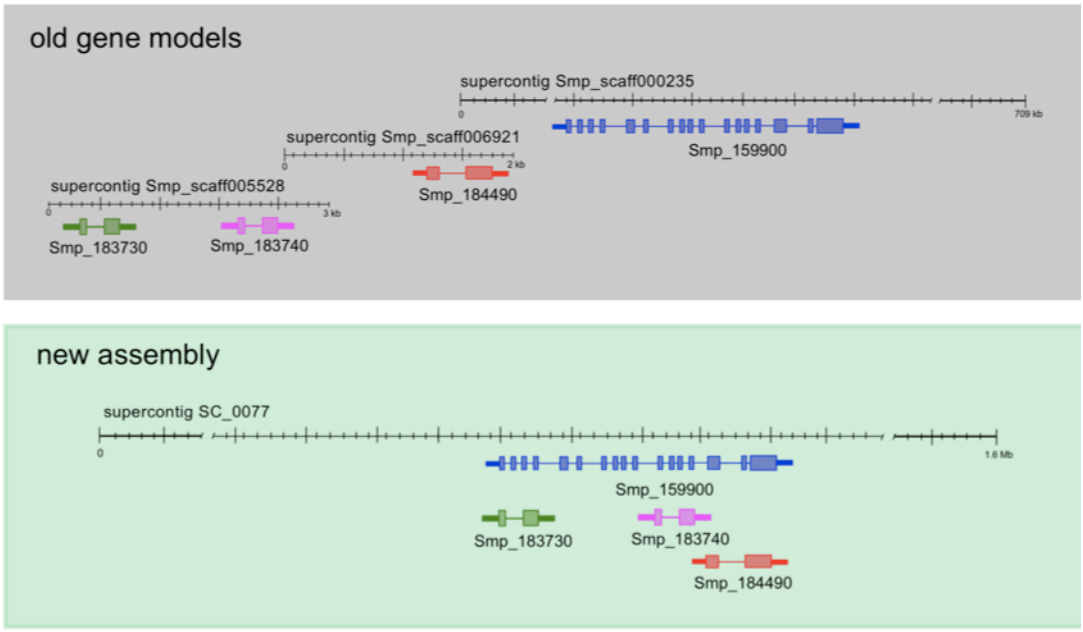


Figure 3.4 - Removal of assembly redundancies produces a more reliable set of gene models [reproduced from (Protasio *et al.*, 2012)]. Gene models were migrated from the previous version of the genome using RATT (Otto *et al.*, 2011). In the new version, many scaffolds converged into one region and hence the gene models contained in them overlap each other. In this example, four supercontigs from the previous version of the assembly collapsed on an unplaced region of Chromosome 3 (superconting SC_0077) in the new assembly. The smaller gene models (green, pink and red) are now obsolete as they were clearly incomplete annotations and their coding region is contained in the exons of a larger gene model (blue).

3.2.4.2 Manual curation

As previously mentioned, RNA-seq data can be used for gene prediction, refinement of already existing gene models and quantification of gene expression. Experimental transcript evidence, such as that provided by RNA-seq data, is preferred to *ab initio* gene prediction. However, the latter are not entirely obsolete as in most cases RNA-seq based gene models can confirm or complement *ab initio* predictions.

In this section, the impact of RNA-seq data on the gene annotation and refinement is presented. The first approach was to visualize the coverage of RNA-seq reads in the context of the genome. To do this, the genome visualization tool Artemis (Carver *et al.*, 2008) and the alignment visualization tool BamView (Carver *et al.*, 2010) were used. **Figure 3.5** shows an example of an Artemis view of the gene Smp_169190 located in Chromosome 1 and an explanation of how the genome and transcriptome information are displayed in this genome viewer.

The *S. mansoni* gene set has been annotated based on *ab initio* tools and *some* experimental data, mainly ESTs and a handful of contributions from collaborators. *Ab initio* gene models tend to over predict the length of a given gene by joining exons that are far apart in the genome and that may not be part of the same transcript. RNA-seq data provides evidence that this is indeed the case. **Figure 3.6** shows how long genes can be split into smaller ones based on the level of expression shown for different regions of the *ab initio* predicted gene model and the lack of reads spanning the introns that separates the high expressed from the low expressed portions of the gene model.

The nature of paired reads in RNA-seq data and TopHat's ability to split reads, provide a way to link exons belonging to the same transcript. **Figure 3.7** shows an example of two gene models that are joined by RNA-seq reads suggesting that these two models belong to the same physical transcript. Alternatively, they could be part of a polycistronic transcript (see section 3.2.6.2 in this Chapter). However, polycistronic transcripts are often very unstable and only a few reads would be found to span the intergenic region. In the case presented in **Figure 3.7** the number of reads spanning the intergenic region is similar to that found across the full length of both transcripts, suggesting that indeed these are part of the same RNA molecule. BLASTp searches (Altschul *et al.*, 1990) showed that one of the transcripts encode a truncated CNH domain, which is also found in the second model suggesting a functional link between these two transcripts.

It is also possible to identify new exons or splice variants of a transcript using RNA-seq data. The example presented in **Figure 3.8** shows how this was applied to refine the

structure of the gene model Smp_014570. The read coverage suggested that two exons were missing from the 5' end and another two, or possibly three exons were missing at the 3' end. This could represent the actual full structure of this gene or possibly a new splice variant.

Finally, RNA-seq data reveals new gene models, such as the example shown in **Figure 3.9**. Although there used to be a small single exon transcript annotated in the reverse strand (not shown), the main transcript at this locus is present in the forward strand. This is evident from the coverage plots and by the ORF found in the forward strand directly under the peaks of RNA-seq reads. Unfortunately, sequence database searches against Uniprot (Uniprot Consortium, 2009) and Pfam domain database (Finn *et al.*, 2010) revealed no conserved domains in the putative protein encoded by this transcript.

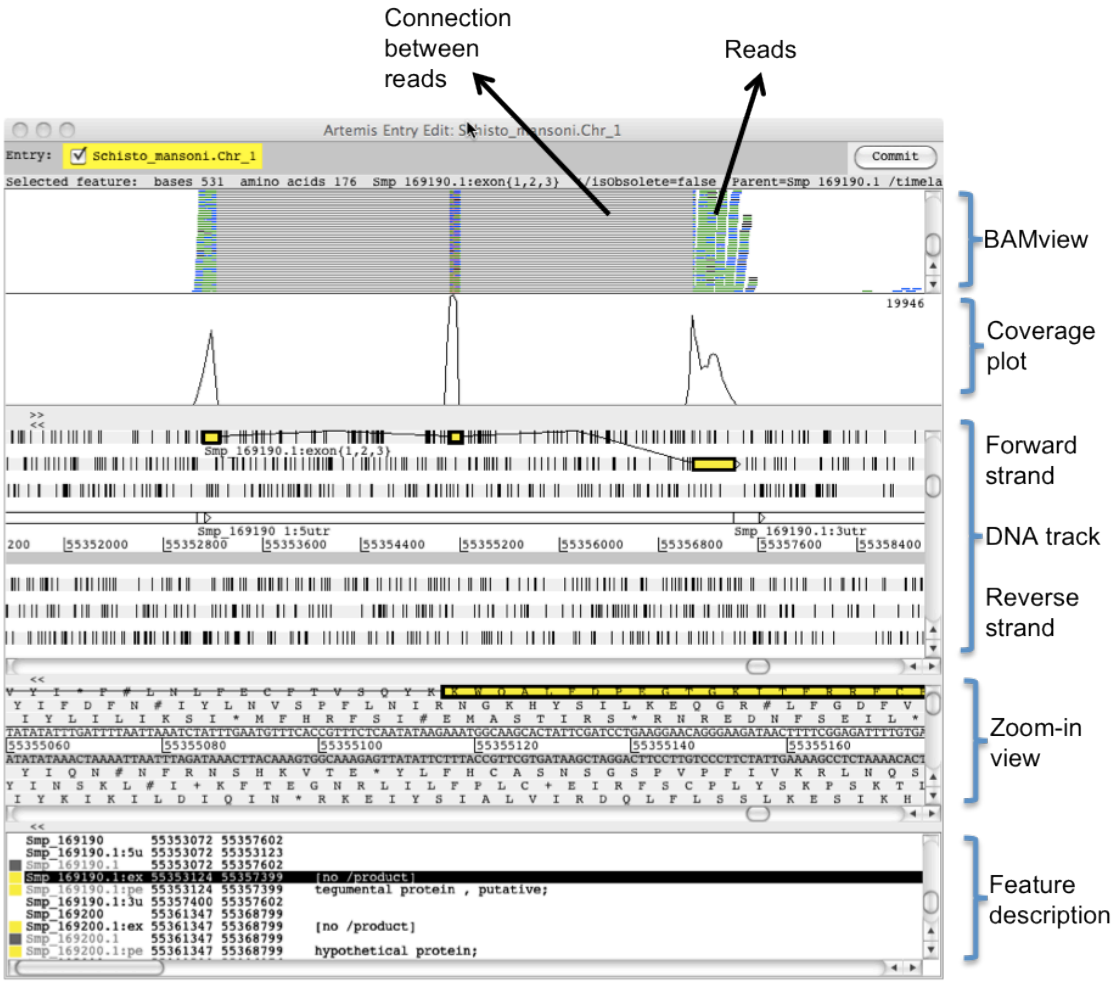


Figure 3.5 – Explanation of Artemis and BamView. Visualization. In this view of Artemis (Carver *et al.*, 2008) and BamView (Carver *et al.*, 2010), the working window is split in five sections. The top most is a graphical representation of the reads' alignment file (BAM file). Reads are represented in blue and green and the grey lines join reads that are either mates or, in the case of a TopHat BAM file, reads that have been split. The second panel shows the same information but in the form of a plot; maximum coverage for the region in display can be found in the top right corner. The third panel shows an overview of the full length of the gene in the context of the genome. The gene chosen for this example is a 3-exon gene (coloured in yellow) and it is found in the forward strand. The three possible frames of translation are shown for both forward and reverse DNA strands. The small vertical black lines represent stop codons. The fourth panel is a zoom-in view of the previous one where the detailed nucleotide and amino acid sequences can be seen. The last bottom panel shows the features annotated in the genome.

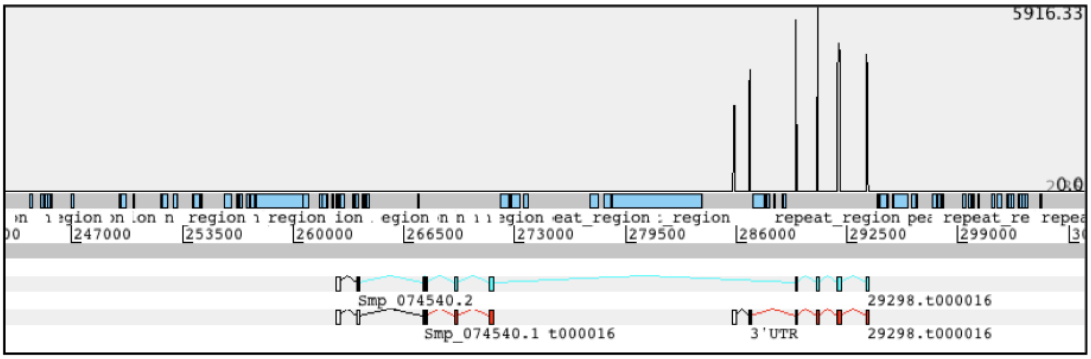


Figure 3.6 –RNA-seq data is used to split gene models. The previous version of model Smp_074540 (blue) and the two current genes (red) are shown. The previous gene model was split based on the differential expression of its 5' end region compared to the 3' end. Both new genes are expressed in this transcriptome sample (cercariae) but one (right) has an astonishing RPKM of 98,619 and while the other (left) has only 1,480. What is more, closer inspection of the reads covering this region provides no evidence of read pairs (or split reads) spanning the intergenic region between the two new gene models therefore confirming that they represent two independent transcription units.

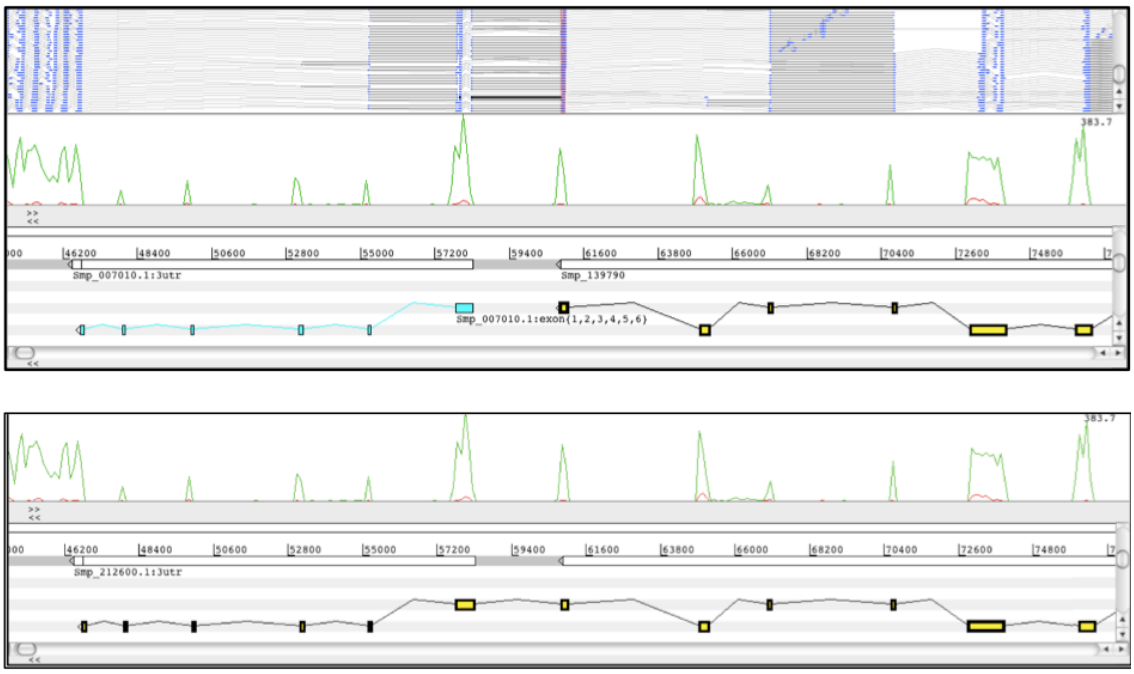


Figure 3.7 – Merging of gene models based on RNA-seq data. Top panel: two gene models (blue and yellow) are shown where RNA-seq data suggest they both belong to the same RNA molecule. Reads spanning the intergenic region between the two gene models (black line) provide evidence that the two transcripts represented are physically connected. Bottom panel: resulting gene model.

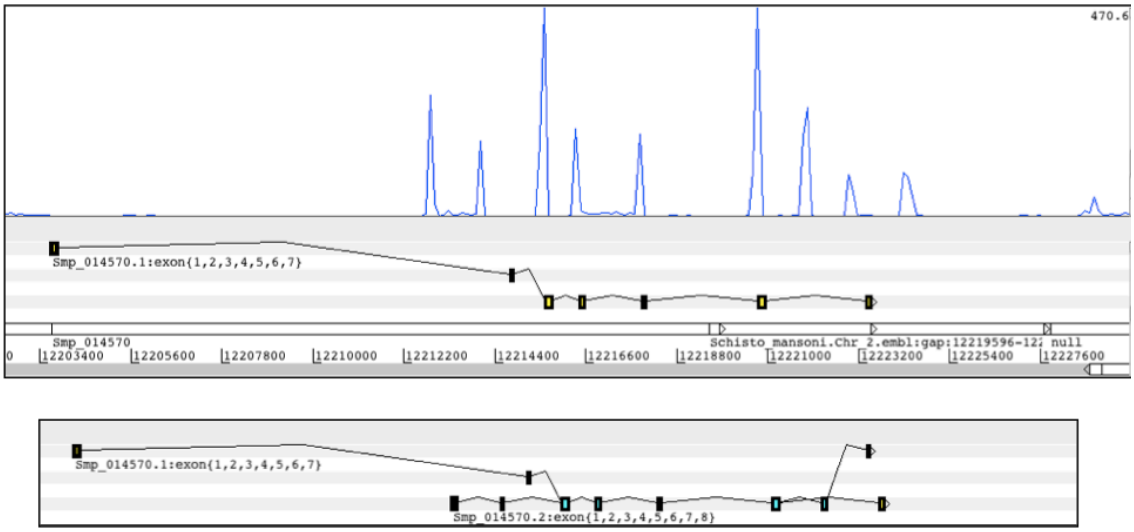


Figure 3.8 – RNA-seq data is used to find new exons and provides evidence of an alternative splicing form of gene Smp_014570. Top panel: the annotated model as a 7-exon transcript with its first exon located relatively far apart form the rest of the coding region. RNA-seq data supports the presence of 5 non-annotated exons and discards the first exon as the start of transcription. Bottom panel: resulting annotation containing both transcript models where some exons are shared.

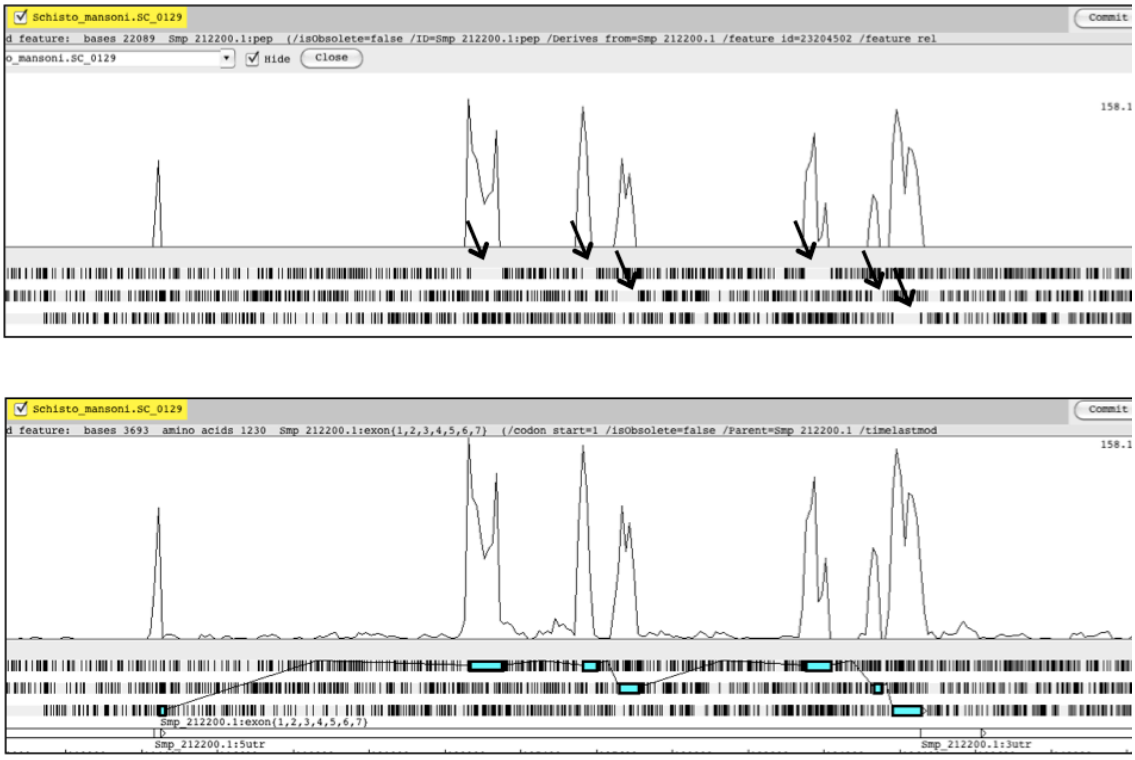


Figure 3.9 -RNA-seq data can reveal new genes. Top panel: each peak of expression corresponds to an exon and ORFs (arrows) are found in the amino-acids track suggesting this is indeed a coding gene. Bottom panel: resulting gene model.

3.2.4.3 Semi-automatic annotation of gene models using RNA-seq data and its merging with previous gene models

Manual curation of transcript models based on RNA-seq data is the most accurate way of structural annotation of gene models. However, this is a very laborious and time-consuming task. An alternative more high-throughput automated approach of doing this is by using the software Cufflinks (Trapnell *et al.*, 2010). Cufflinks takes the TopHat output and generates gene models based on the predicted exons and uses both split reads and read pair information to join exons together into transcriptional units. **Figure 3.10** represents a summary of scenarios comparing gene models from the old assembly and their respective Cufflinks predictions. In some cases, Cufflinks correctly predicted the gene model **(Figure 3.10B**) while in other cases, and due to the small introns present at the 5' end of many *S. mansoni* genes, Cufflinks predicted these to be UTRs (**Figure 3.10C**). Other scenarios included models in which modifications introduced in the assembly caused several exons to be joined in one larger exon (**Figure 3.10D)**.

At this stage, the genome had two sets of gene predictions: the ones migrated from the previous genome assembly (see section 3.2.4.1) and the ones derived from Cufflinks (RNA-seq data). As described before, Cufflinks predictions differ from the already existing gene models making it necessary to merge them into one set of gene predictions. This was done using the software Jigsaw (Allen *et al.*, 2005; Allen *et al.*, 2006). The following modifications were recorded (see also **Table 3.3**):

- Gene models from the old assembly were either split or merged in the new assembly based on RNA-seq coverage (as shown in **Figure 3.6** and **3.7**)
- At least one additional exon was added to some gene models based on RNA-seq data – an example can be seen in **Figure 3.8**.
- Cufflinks-Jigsaw models automatically replaced gene models when they provided a longer CDS than the already annotated model.
- New gene models resulted from putative new transcripts that did not overlap previous predictions and were not similar to previously reported transposable elements.

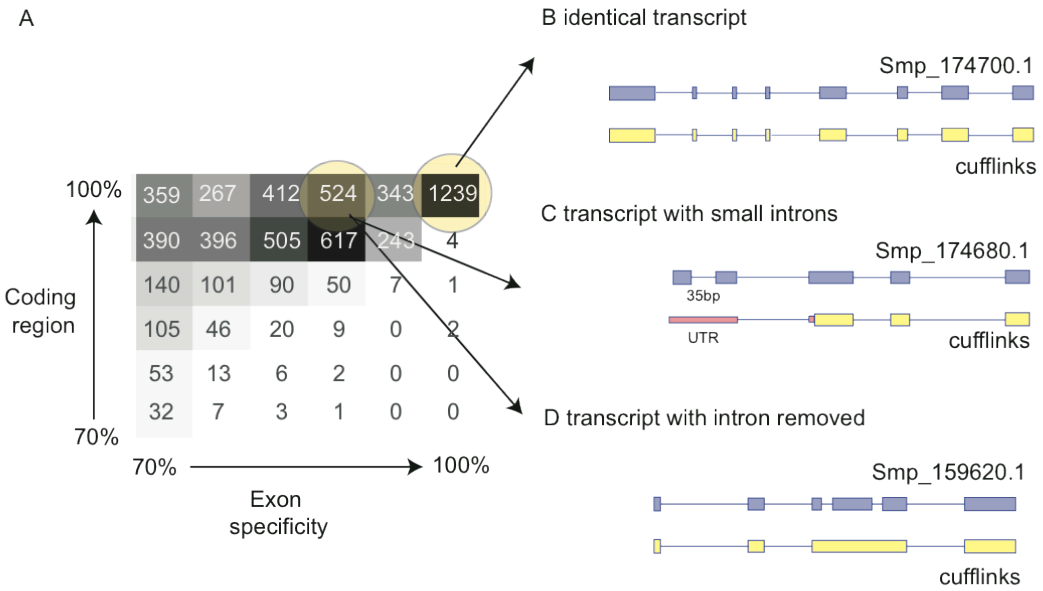


Figure 3.10 – Example scenarios of Cufflinks' models compared with previous gene models [reproduced from (Protasio *et al.*, 2012)]. A - Heatmap displaying comparisons between previous gene models and transcript fragments generated from Cufflinks. For each model, the extent of coding region that overlaps with a Cufflinks' model and the proportion of correctly predicted exon boundaries was calculated and categorised into bins of 70-100%. Models in this plot were excluded with less than 70% of their exon boundaries or coding regions predicted. B, C and D - Example scenarios of Cufflinks' models compared with previous gene models where B the Cufflinks prediction is identical to the 1,239 existing models; C Cufflinks fails to identify small introns; D Cufflinks removes incorrect introns present in the previous gene model, probably due to the improved assembly which, by correcting gaps, produced a longer single exon while the reading frame is preserved.

Table 3.3 – Changes performed on gene models from the old version (4.0) and the current gene count for the new version (5.0). Reproduced from (Protasio *et al.*, 2012). The criteria for each category are described in section 3.2.4.3.

	Number
Total gene models in old genome version	**11,719**
Not transferred	1,088
Deleted models	545
Split or merged models	731
Models with additional exons	3,438
Models that have been automatically replaced	1,116
New genes	504
Genes in new version	**10,852**

3.2.4.3.1 Putative functions of the new genes derived from RNA-seq data

Cufflinks was able to identify 504 new gene models in loci where there was no previous annotation. The mean length of the set of new gene models is 261 nucleotides with the largest model spanning 4,242 bases. Approximately 75% of the new models have just one exon and the rest range from two to five exons with only three outliers of nine and 12 exons (**Figure 3.11**). Of all the new genes, 64 of them have a significant InterProScan (Zdobnov *et al.*, 2001) match (e-value < $1e^{-05}$). What is more, this similarity search led to the assignment of Gene Ontology (see Chapter 2 section 2.6.6) terms to 49 new transcripts. The remaining 440 transcripts could not be assigned a putative function at this stage, and were therefore classed as "hypothetical proteins".

The largest transcript from those that could be assigned a putative function was further characterised. Smp_204750.1 is a 4,242 nt long transcript encoded in four exons. The *in silico* translation produces a 1,413 amino acid polypeptide with a secretory signal peptide, four Immunoglobulin I-set domains towards the N-terminal region and one fibronectin domain in the mid section. Additionally, a *trans*-membrane domain is found immediately after the fibronectin domain towards the C-terminus. These data suggest that the product of Smp_204750.1 is expressed in the cell surface and could have a role in cell-to-cell signalling. This is a good example of the power of RNA-seq data to fully identify previously missed annotations.

To sum up, the latest version of the genome (v5.0) has a total of 10,852 annotated genes. These are a result of combining the annotation from the old assembly (*ab initio* predictions, ESTs and a handful of manually annotated genes) with RNA-seq data produced in this study. Based on the latter, significant changes were made such as identification of new exons, new genes, modification of existing genes by the splitting or merging of gene structures. RNA-seq data identified 504 new gene models many of which (440 genes) could not be assigned a putative function.

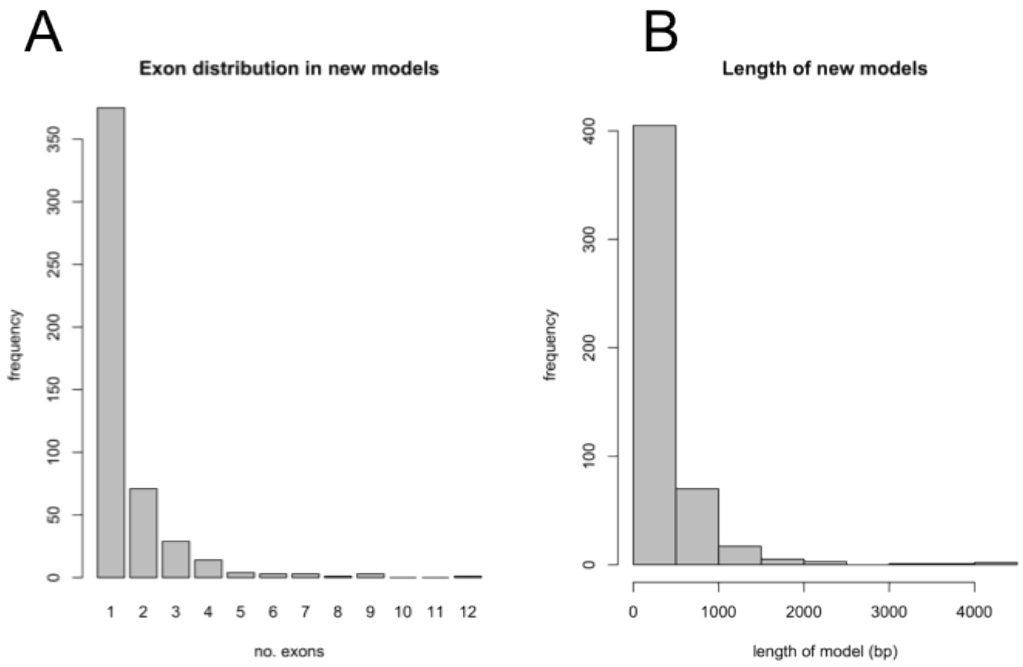


Figure 3.11 – Characteristics of the 504 new genes identified by Cufflinks and based on RNA-seq data. A – Exons distribution of new models. B – Length of the coding regions of new models.

3.2.5 Defining expression

In order to discriminate between signal arising from active transcriptional units and background noise, a background RPKM value was calculated. This calculation was based on the signal originating from intergenic regions, which would represent a measure of the reads mapped to non-expressed regions of the genome. This method was developed by Dr. Adam Reid (Pathogen Genomics group, Wellcome Trust Sanger Institute) and it was described in Chapter 2 section and 2.6.2. Briefly, it estimates the RPKM value corresponding to background transcriptome signal by calculating the RPKM for 500 bp non-overlapping windows of exons, UTRs, intron and intergenic regions. Using this cut-off, expressed and non-expressed genes can be identified for further analysis or filtering.

3.2.5.1 Non-expressed genes.

As shown above, an RPKM value of 2 was established as the cut-off for discriminating expression from transcriptional background. With this cut-off, a total of 1,584 transcripts were found as "not expressed" across all of the life cycle time points studied in this thesis. Of these, 101 are new RNA-seq derived genes (6.4%) without an assigned description and another 1,021 (64.5%) are described as "hypothetical proteins". The percentage of "hypothetical proteins" in the group of non-expressed genes (64.5%) is comparatively higher than that found in the totality of product descriptions (56%). The resulting 462 transcripts that are not expressed during the cercariae or selected intra-mammalian stages have a product description. To test the hypothesis that these genes might be expressed during intra-molluscan stages, a similarity search (BLASTn (Altschul *et al.*, 1990)) using the non-expressed genes against a daughter sporocyst EST collection[1] was performed. It was found that 83 of the non-expressed genes matched at least one EST (e-value < $1e^{-5}$); 25 were among the 462 genes with a product description. Some examples of these are four homologs of the cercarial elastase, two peptidases of unknown function and a putative inositol 1,4,5-trisphosphate receptor among others (a compete list of these 25 genes and their descriptions is shown in **Appendix B**). Additionally, four VAL genes (SmVAL2, 3, 5 and 9), which are expressed only in miracidia or mother sporocyst stages (Chalmers *et al.*, 2008), were also found among the non-expressed genes.

[1] This database is held at the WTSI. Samples to generate the EST libraries were provided by Prof. Alan Wilson, York University, UK.

Data presented here strongly suggest that the population of non-expressed genes is not a random sample from the whole transcriptome and that it contains genes that are expressed in other life cycle stages. However, non-expressed genes are shorter, have less number of exons and a higher percentage "hypothetical proteins" compared to the rest of the transcriptome. It is possible that the lack of experimental evidence to back up these gene models caused the unusual structures. RNA-seq sampling of intra-molluscan stages as well as egg and miracidia could shed light on the structures of these non-expressed genes.

3.2.6 *Trans*-splicing

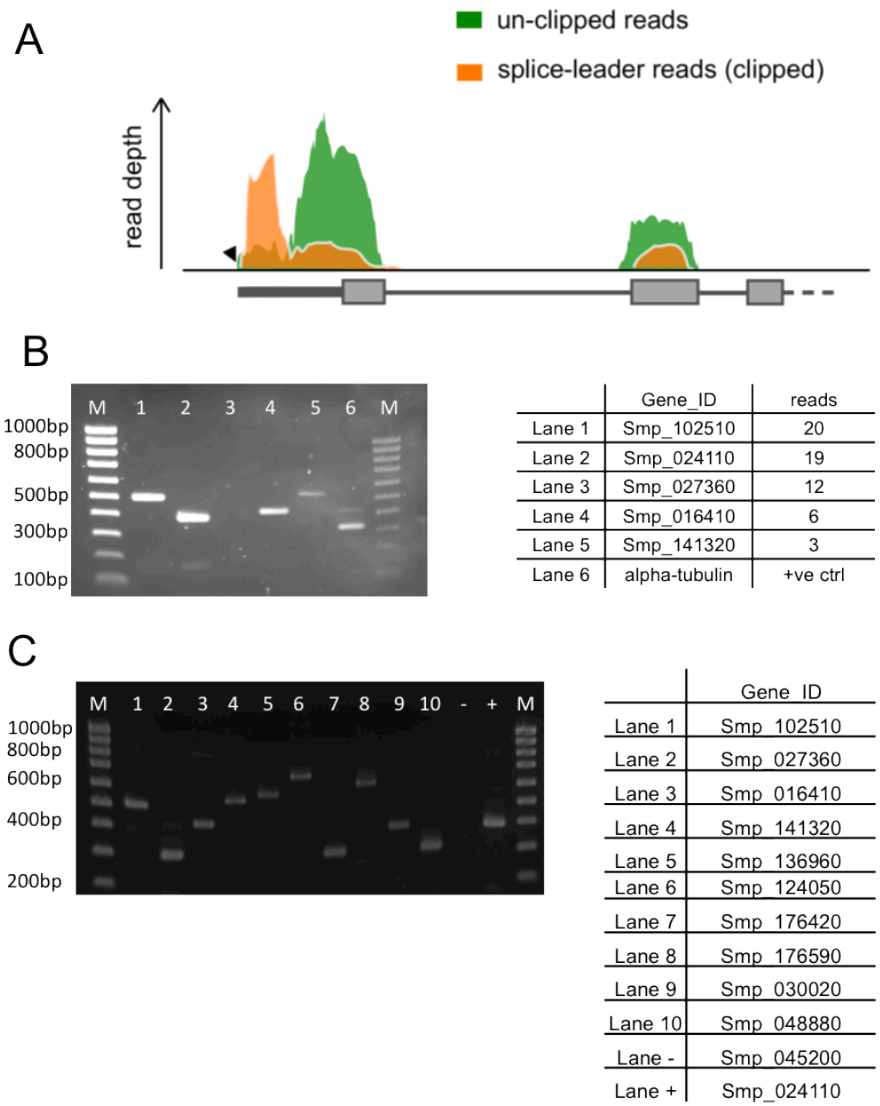
Trans-splicing is a mechanism where two RNA molecules are combined to form a mature RNA. In the case of splice leader (SL) *trans*-splicing, one of the RNA molecules involved is a small nuclear ribonucleoprotein commonly referred to as SL-RNA. The *trans*-splicing process is similar to that of *cis*-splicing and involves an enzymatic complex known as the spliceosome. *Trans*-splicing occurs at a canonical splicing acceptor site with a canonical intron sequence but lacking the (or with a non-conserved) splicing donor site (Conrad *et al.*, 1991). The SL sequence present in the mature mRNA is typically small comprising from 22 nt to up to approximately 53 nt depending on the species.

SL *trans*-splicing was first described in trypanosomes (kinetoplastid protozoan) (Murphy *et al.*, 1986; Sutton *et al.*, 1986) and later in the nematode *C. elegans* (Krause *et al.*, 1987). The first report of *trans*-splicing in platyhelminths was in *S. mansoni* (Rajkovic *et al.*, 1990) followed by *F. hepatica* (Davis *et al.*, 1994), *E. multilocularis* (Brehm *et al.*, 2000) and *T. solium* (Brehm *et al.*, 2002). The percentage of genes that are subjected to *trans*-splicing varies among species. In trypanosomes, all mature mRNAs are *trans*-spliced while in *C. elegans* it occurs in 70% of the transcripts. A previous report has estimated that it affects ~10% of genes in *S. mansoni* and only a small sample of *trans*-spliced transcripts have been described so far (Davis *et al.*, 1995). It is not known whether there is a common function among *trans*-spliced transcripts. In *C. elegans,* there are two conserved SL sequences, 60% of *trans*-splicing events occur by acquisition of a SL1 and ~10% with SL2. The latter SL sequence is reserved for resolving polycistrons. It is possible that the function of the SL in *trans*-spliced transcripts is more closely related to the nature of the *trans*-spliced mRNA than to the function of the encoded protein; for example it may have a role in the regulation of translation and/or transcript stability. Upon *trans*-splicing, the mRNA molecules acquire a 2,2,7-trimethylguanosine (TMG) cap different from that present in mature mRNAs that are not *trans*-spliced. It has been suggested that this TMG is

a required modification for certain processes undergone by *trans*-spliced transcripts (Brehm *et al.*, 2000).

3.2.6.1 "Standard" *trans*-splicing

By filtering RNA-seq reads containing the spliced leader (SL) sequence, the locations where *trans*-splicing events occur could be mapped genome-wide. The procedure involved identifying those reads that contained the SL sequence, trimming this sequence from the read and mapping the remainder of the read to the genome. The locations where these reads map reveal putative *trans*-splicing acceptor sites (see Methods Chapter 2 section 2.6.3). An example of a putative *trans*-spliced transcript is shown in **Figure 3.12A**. In order to validate the sensitivity of this detection approach, *trans*-splicing events with different number of supporting reads were chosen for experimental validation using PCR (**Figure 3.12B**). Results show that *trans*-splicing events supported by as little as three reads could be validated. A total of 944 transcript models (~8.7% of all annotated genes) were found to be potentially *trans*-spliced, a figure in close agreement with the 10% previously predicted by Davis *et al.,* (1995). The criteria for categorising a *trans*-splicing event is that the acceptor site is located in an exon or intron, or is located within 1 kb upstream from the start of a transcript. Further validation experiments were performed by randomly selecting ten putative *trans*-spliced transcripts (**Figure 3.12C**) for PCR using a SL primer and a gene specific primer. All ten experiments confirmed that *trans*-spliced forms of these transcripts exist.



	Gene_ID	reads
Lane 1	Smp_102510	20
Lane 2	Smp_024110	19
Lane 3	Smp_027360	12
Lane 4	Smp_016410	6
Lane 5	Smp_141320	3
Lane 6	alpha-tubulin	+ve ctrl

	Gene_ID
Lane 1	Smp_102510
Lane 2	Smp_027360
Lane 3	Smp_016410
Lane 4	Smp_141320
Lane 5	Smp_136960
Lane 6	Smp_124050
Lane 7	Smp_176420
Lane 8	Smp_176590
Lane 9	Smp_030020
Lane 10	Smp_048880
Lane -	Smp_045200
Lane +	Smp_024110

Figure 3.12 – *Trans*-splicing. A - Schematic view of the 5' end of *trans*-spliced gene Smp_176420. Shaded coverage plots represent non-normalized RNA-seq reads still containing the spliced-leader (SL) sequence (green – unclipped reads) and reads previously found to contain the SL sequence (orange - clipped). In the latter, the SL sequence was removed prior to aligning the reads to the genome; which improved the reads' mapping (lower coverage in the unclipped reads than in the orange reads). B – Validation of the sensitivity of the bioinformatics approach to detect *trans*-splicing events. A PCR positive control was included (lane 6, alpha tubulin). There was no PCR product for the *trans*-spliced validation of Smp_027360 shown in lane 3. This was later repeated and resulted in a positive *trans*-splicing event (see part C). C - RT-PCR validation of 10 putative *trans*-spliced genes with SL1 as forward primer and a gene-specific reverse primer. Smp_024110.1, previously described as *trans*-spliced (Rajkovic *et al.*, 1990), was included as a positive control (lane indicated with '+') while Smp_045200.1 was included as a negative control of *trans*-splicing (lane indicated with '-'). All PCRs but one (Smp_176590.1) show bands corresponding to expected PCR product size.

As a consequence of identifying *trans*-splicing events, correction of the predicted coding sequence of gene models can be done. Typically, a *S. mansoni* gene model would have an ATG as first translated codon in the gene model. However, this might not be the case for some *trans*-spliced transcripts since the SL molecule in *S. mansoni*, if *trans*-spliced in frame with the rest of the transcript's ORF, can provide the starting methionine (Met). This would imply that for these transcripts, the starting Met need not to be present in the gene locus because it could be provided by the Met encoded in the SL. Previous work (Cheng *et al.*, 2006) showed evidence that the 3' end Met from the SL sequence in *S. mansoni* contributes the initial Met to approximately 40% of all *trans*-spliced transcripts. Sequencing results obtained for the PCR product of SL1-Smp_084890 (**Figure 3.13**) show that the *trans*-spliced transcript has an alternative starting Met in frame with the main ORF. Consequently, the identification of *trans*-spliced genes adds an additional layer of complexity to the annotation of the genome where *trans*-splicing potentially modifies ~8% of the sequence of predicted gene models. Analysis of the presence of Kosak consensus sequences flanking the ATG codon either within the SL sequence or in the recipient transcript would provide further evidence in support of the use of one or the other start of translation.

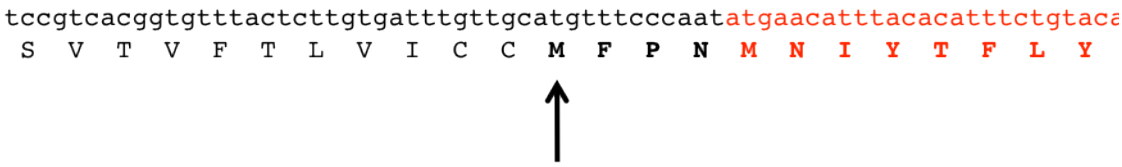


Figure 3.13 – *Trans*-splicing can affect the translation start site of a transcript. *In silico* translation of the PCR product corresponding to the *trans*-spliced form of Smp_084890 indicates that the ATG codon in the SL sequence (arrow) is found in-frame with main ORF of the transcript suggesting this could be used as an alternative initiation of translation.

In many cases, mapping information suggests a second *trans*-splicing acceptor site, usually within 20-50 bases up or downstream from the primary acceptor site. Secondary *trans*-splicing sites also fulfil the acceptor site criteria and therefore are likely to be recognised by the spliceosome. The identification of multiple *trans*-splicing acceptor sites within a single gene could represent "leaky" *trans*-splicing. In such cases, there seems to be one preferred site for *trans*-splicing; although the others are also used but less frequently. This multi-site *trans*-spicing may represent a redundant system put in place to guarantee higher rates of *trans*-splicing for a given transcript.

The small number of genes previously described as *trans*-spliced prevented researchers from identifying common denominators in the functions carried out by *trans*-spliced transcripts (Davis *et al.*, 1995). With a larger number of potentially *trans*-spliced transcripts, it is now possible to investigate whether there is a common function among their products. In order to address this question, GO term enrichment analysis (Alexa *et al.*, 2006) of genes whose transcripts undergo *trans*-splicing was performed. Results shown in Table 3.4 suggest that *trans*-spliced transcripts are enriched in proteins that localise to the endoplasmic reticulum (ER) and the mitochondria.

In terms of biological function, glycosylphosphatidylinositol-anchored protein (GPI-APs) biosynthesis is the most statistically significant term. This led to investigate the relationship between the enzymes from this pathway and the *trans*-splicing phenomenon. It was found that the majority of the enzymes needed to synthesise GPI-APs (starting from palmitoyl-coenzymeA) are encoded in *trans*-spliced transcripts (**Figure 6.1**). These results represent the first indication in platyhelminths of a pathway relying almost entirely on *trans*-spliced genes.

3.2.6.2 *Trans*-splicing in polycistronic transcripts

Polycistronic transcripts originate from a single promoter but are later processed to generate two or more individual mRNAs. This type of transcriptional regulation is characteristic of trypanosomatids (Johnson *et al.*, 1987) and is present in *C. elegans* (Spieth *et al.*, 1993) and other organisms (Douris *et al.*, 2010). It has been suggested that the *S. mansoni* Ubiquinol-cytochrome-c-reductase (UbCRBP) and phosphopyruvate hydratase (Smp_024120 and Smp_024110 respectively) genes might be transcribed as a polycistronic unit and that *trans*-splicing of the phosphopyruvate hydratase transcript might resolve the polycistron into individual transcripts (Davis *et al.*, 1997). However, the authors failed to provide convincing evidence of the existence of such polycistronic transcripts – i.e. a PCR showing the existence of the intergenic region in the pre-mRNA.

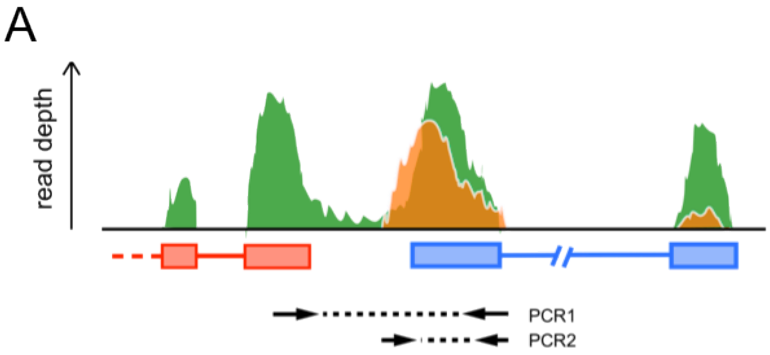
Because intergenic regions within polycistron are short (usually 200 nt but can be up to 2 kb), it is possible to use this information together with the available *trans*-splicing data to identify putative polycistrons in *S. mansoni.* To this end, intergenic distances[1] between genes were calculated. A total of 142 pairs of genes were found separated by at least 200 bp and 46 of them showed evidence of *trans*-splicing in the downstream gene, suggesting these could be polycistronic transcripts. By increasing the intergenic distance cut off to 2 kb, it was found that the number of putative polycistrons increased to 115 (out of a total of 633 genes found within 2 kb distance).

An example of the architecture and read coverage for a polycistronic transcript is presented in **Figure 3.14A**. Validation of four of these putative polycistrons was performed using PCR (**Figure 3.14B**) and also by sequencing of the PCR product, which confirmed the presence of sequences from both upstream and downstream transcripts.

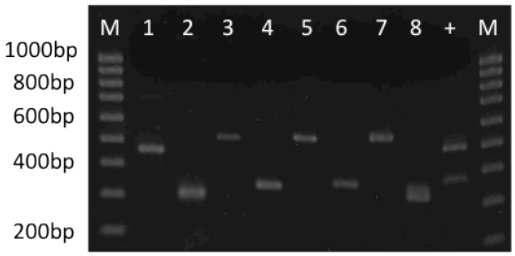
Unlike *C. elegans*, which uses a second spliced leader (SL2) to resolve polycistrons (Spieth *et al.*, 1993) or a tightly control combination of both (Allen *et al.*, 2011), *S. mansoni* seems to use the same SL for both polycistronic- and non-polycistronic *trans*-spliced transcripts. A secondary *trans*-splicing SL sequence has not yet been described for *S. mansoni*.

In *C. elegans*, a promoter located in the 5' end of the polycistron controls the expression of the polycistronic unit. Moreover, it has been hypothesized that polypeptides encoded in these polycistronic transcripts are functionally related, for example they could be part of the same pathway [reviewed in (Blumenthal *et al.*, 2003)]. It is possible that the polycistronic organisation of genes in *S. mansoni* has the architecture of an operon but the presence of a promoter regulating the expression of the whole unit has not yet been demonstrated. In terms of the functions related to proteins encoded in the *S. mansoni* polycistrons, close inspection of the two *in silico* predicted products emerging from each of them failed to reveal a functional link between them. The same comparison was done but looking for shared gene ontology terms between the members of a polycistron but no association could be found.

[1] Intergenic distance is defined as the number of nucleotides found between the end of one coding sequence and the start of another one.



B



	type of test	up stream gene ID	down stream geneID
Lane 1	polycistron	Smp_006980	Smp_006970
Lane 2	trans-spliced		
Lane 3	polycistron	Smp_084900	Smp_084890
Lane 4	trans-spliced		
Lane 5	polycistron	Smp_079750	Smp_079760
Lane 6	trans-spliced		
Lane 7	polycistron	Smp_023160	Smp_023170
Lane 8	trans-spliced		
Lane 9	polycistron +ve ctrl	Smp_024120	Smp_024110

Figure 3.14 – *Trans*-splicing is used to resolve polycistronic transcripts in *S. mansoni.* A - Schematic view of the putative polycistron Smp_079750-Smp_079760. PCR1 represents the amplicon obtained from the *trans*-spliced form of Smp_079760 while PCR2 represents the amplicon obtained from the unprocessed polycistronic transcript containing the intergenic region. B - RT-PCR validation of four putative polycistrons and a positive control (Smp_024110-Smp_024120; lane 9) previously suggested to be a polycistronic unit in (Davis *et al.*, 1997). Each putative polycistron was subjected to two PCR that correspond to PCR1 (e.g lane 1) and PCR2 (e.g lane 2) in (A).

In other species that also use *trans*-splicing to resolve polycistronic units, polycistrons can be resolved in two and up to eight individual transcripts such is the case in *C. elegans*. Most of the polycistronic units found in *S. mansoni* during this study have two resulting transcripts where the downstream transcript (relative to the *trans*-splicing acceptor site) is *trans*-spliced while the upstream transcript is not. However, two exceptions (Smp_170260.1 - Smp_088390.1 -Smp_088380.1 and Smp_038430.1 - Smp_038420.1 - Smp_038410.1) were found in Chromosome W where each polycistron seems to resolve into three transcripts. In these cases, the two 3' most transcripts are *trans*-spliced, while the first one is not. Gene products from these transcripts are all "hypothetical proteins". Experimental validation will be needed to verify these predictions.

3.3 Discussion

The motivation behind this doctoral thesis was the identification of genes that are developmentally up regulated upon *S. mansoni* infection of the human host and which of these may have a role in the adaptation of the parasite to its new environment. The first step was to obtain RNA samples for library generation and sequencing. Because RNA-seq technology was at its infancy during the data collection stage of this work, it was necessary to validate the reproducibility of the method. To this end, the correlation between samples obtained from Illumina sequencing of RNA-seq libraries was analysed (section 3.2.2.1). The high correlation values (~0.99 Pearson's correlation) obtained for the technical replicates, both the library preparation and sample sequencing, suggested that technical reproducibility is indeed very high and agrees with figures reported elsewhere (Marioni *et al.*, 2008; Hebenstreit *et al.*, 2011). Biological replicates were also analysed and the correlation values obtained for those were also very high, which again agrees with previous reports (Hebenstreit *et al.*, 2011). Biological replicates are key components of the statistical analysis [edgeR (Robinson *et al.*, 2010), sections 4.2.2 and 5.2] and are needed to guarantee statistical power in assessing differential expression. In conclusion, it was possible to prescind from technical replicates but biological replicates would be included were possible.

After assessing the reproducibility of the RNA-seq approach in *S. mansoni* samples, a comparison of this technology with other high-throughput methods of gene expression measurement was performed. Several microarray studies have been applied to investigate the transcriptome of several life cycle stages of *S. mansoni* (Hoffmann *et al.*, 2003;

Fitzpatrick *et al.*, 2005; Dillon *et al.*, 2006; Fitzpatrick *et al.*, 2006; Vermeire *et al.*, 2006; Jolly *et al.*, 2007; Verjovski-Almeida *et al.*, 2007; Fitzpatrick *et al.*, 2009; Gobert *et al.*, 2010; Parker-Manuel *et al.*, 2011). The availability of these data provided the opportunity to investigate the correlation between RNA-seq and microarrays. The microarray datasets presented by Fitzpatrick *et al.*, (2009) and Parker-Manuel *et al.*, (2011) were compared to RNA-seq data generated in this study by calculating the Spearman's rank correlation between them. Correlation values between RNA-seq data and each of the microarray studies are very similar (0.66-0.69) but relatively lower than those reported for other systems (Marioni *et al.*, 2008; Otto *et al.*, 2010; Hebenstreit *et al.*, 2011). However, most of the previously reported correlations between RNA-seq data and microarrays use the same source of RNA as starting material to generate the data in both platforms. Since the correlations presented here were generated with biological material isolated in different experimental conditions, it is possible that the lower correlation might be attributed to experimental variation. Nevertheless, these data showed that the RNA-seq approach permits quantification of gene expression over a greater dynamic range than that obtained from microarrays. In the latter, very low expressed probes are usually miss-calculated due to the analogue nature of the signal (Shendure, 2008) while very highly expressed probes can show signal saturation. Additionally, it is expected that future gene expression studies performed with RNA-seq will be directly comparable to samples obtained in this study providing the means of generating larger dataset that can be analysed together.

As introduced in Chapter 1, measuring of transcript abundance is not the only aspect to RNA-seq data. Contrary to microarrays, RNA-seq data is not limited by the sampling of existing features and provides the opportunity of identifying new genes or refining the structural annotation of already existing ones. The genome of *S. mansoni* had been previously annotated based mainly on *ab inito* predictions assisted by limited EST data and a handful of manually curated gene models (Haas *et al.*, 2007; Berriman *et al.*, 2009). Although this contribution represented a landmark in the study of schistosome biology, analysis of the RNA-seq data in the context of the genome provided evidence that many gene models were probably not well represented (section 3.2.4.2). Taken together, the availability of RNA-seq data and specially its reliance on a correct set of genes to accurately measure gene expression, provided the right frame and motivation to generate an improved version of the structural gene annotations. The combination of new evidenced-based transcriptome information derived from RNA-seq data with the previously annotated dataset of genes had a profound effect on the refinement of gene structures. What is more, it generated ~500 new genes of which ~80% could not be

assigned a function based on similarity searches and are therefore catalogued as hypothetical proteins. It is possible that these genes encode proteins with novel schistosome-specific function. Further investigations regarding the nature of these genes would shed light on their potential function; for example, sampling of other stages of the life cycle (egg, miracidium and intra-molluscan stages) with RNA-seq or other methods or the study of non-coding RNAs (Guttman *et al.*, 2010) could improve the current understanding of the genome. In summary, the improvement of the gene annotation represents an important advance in the completeness of the *S. mansoni* genome.

In preparation for the analysis of differential expression (Chapter 4 and 5), an empirical expression cut off was calculated (section 3.2.5). This calculation was based on transcription signal from non-coding regions of the genome. Using this cut off it was possible to find ~1,500 genes across the life cycle stages here analysed (cercariae, 3-hour and 24-hour old schistosomula and adult worms) whose expression was lower than background. There are several reasons why these genes may not be expressed in the studied samples. One of them would be that they are expressed in other life cycle stages such as egg, miracidia or intra-molluscan stages; all of which were not included in this study. Indeed, similarity searches against an intra-molluscan EST library showed that ~80 of these genes had a match in this library and other previously demonstrated intra-molluscan specific genes (SmVAL2, 3, 5 and 9) were also found among the non-expressed transcripts. It is noteworthy that the mean length and number of exons in these genes are significantly different from the rest of the annotated genes. This might reflect that further experimental evidence is needed to generate an even better annotation.

In order to continue the characterisation of the transcriptome, RNA-seq data was used to identify *trans*-splicing events in the genome. By the time this approach was envisaged, there had been no reports of RNA-seq data used for such end; only recently Allen *et al.*, (2011) provided the first report featuring high-throughput identification of *trans*-splicing events in the well-characterised model organism *C. elegans* (Allen *et al.*, 2011). *Trans*-splicing in platyhelminths was identified more than 30 years ago when the first *trans*-spliced transcript was reported (Rajkovic *et al.*, 1990). It was later estimated that 10% of the *S. mansoni* genes could be subjected to *trans*-splicing (Davis *et al.*, 1995) but efforts to identify these in EST databases yielded only a few hundred cases (Cheng *et al.*, 2006) probably because of low sequencing depth of these databases. Results presented in this thesis provided evidence that *trans*-splicing events affect ~8.5% of the annotated genes (section 3.2.6) and PCR validation of ten randomly selected putative *trans*-spliced transcripts validated the approach. What is more, the base-resolution detail of provided by

these data can also be applied to accurately predict the location of the *trans*-splicing acceptor site and therefore generate a more accurate structural annotation of the *trans*-spliced genes (section 3.2.6.1). Taken altogether, the identification of *trans*-splicing events at the genome wide scale is an important contribution to the ongoing gene annotation effort on which many downstream applications (i.e., gene cloning) depend.

In terms of their function, the low number of previously identified *trans*-spliced genes had prevented finding a common denominator among *trans*-spliced genes. Now, with a much larger repertoire, it was possible to identified at least one pathway (glycosylphosphatidyl inositol anchored proteins synthesis) where almost all the enzymes are encoded by genes whose transcripts have been found to undergo *trans*-splicing. These results led to the hypothesis that *trans*-splicing might be functioning as a molecular switch under which the expression, stability or availability of a group of genes can be orchestrated. It would be interesting to test whether the parasite can survive without *trans*-splicing and if so, how does this affect their phenotype. These questions remain open and will required further investigation.

Polycistronic transcripts have been previously identified in other organisms (Johnson *et al.*, 1987; Spieth *et al.*, 1993) including *S. mansoni* (Davis *et al.*, 1997) where the presence of one polycistron has been suggested but not demonstrated. The close association of this phenomenon with that of *trans*-splicing led to investigate whether available RNA-seq data could be applied to study polycistronic transcription in *S. mansoni*. RNA-seq data, in combination with the existing annotation, were used to identify putative loci where *trans*-splicing might be resolving polycistronic units (section 3.2.6.2). PCR validation of these putative polycistrons suggested that this is indeed the case and that the complement of polycistronic transcripts in *S. mansoni* (46 with a maximum intergenic distance of 200 nt) might be much larger than previously thought. According to the collected data, *S. mansoni* seem to encode typically two transcripts in each polycistron. Only two cases could be identified were the polycistron might be resolved into three individual mRNA, however this still requires validation.

It is possible that the number of both *trans*-splicing events and polycistrons is in reality larger than the one reported here, which opens the question of how many genes are actually *trans*-spliced in *S. mansoni*? One possible way of addressing this question would involve creating a SL-specific library. Because this would include only a fraction of the whole RNA population otherwise present in conventional RNA-seq libraries, the SL-containing molecules could be sequenced at a much deeper depth than regular RNA-seq

samples. Nevertheless, any additional RNA-seq experiment in *S. mansoni* could potentially be exploited to enlarge the dataset of *trans*-splicing events in this worm.

CHAPTER 4

COMPARATIVE STUDY OF TRANSCRIPTOME PROFILES OF MECHANICAL- AND SKIN-TRANSFORMED SCHISTOSOMULA

4.1 Introduction

Before the development of a mechanical transformation method (Gazzinelli *et al.*, 1973; Howells *et al.*, 1974; Ramalho-Pinto *et al.*, 1974), schistosomula in large numbers could only be recovered by naturally allowing parasites to penetrate a layer of excised host skin. Such an approach was first reported in 1966 where a protocol using "scraped, dried, plucked skin" from the abdominal area of rats, mice and hamsters was described (Stirewalt *et al.*, 1966). Rat skin proved to be the most convenient given its larger size and availability. Despite providing large quantities of schistosomula (~45,000 per experiment) this protocol was extremely time- and labour-intensive taking 18 hours to dry the skin prior to the actual skin penetration experiment. However, this work allowed the authors to define schistosomula based on experimental observation of parasites obtained *in vitro* (with the described protocol) and *in vivo*. According to the authors, schistosomula are:

> *"saline- and serum-adapted, water-intolerant larval stage which develops from the cercaria after the tail has been cast, the pre- and post-acetabular glands have been evacuated, and there has been a change of surface resulting in altered permeability, and the loss of the precise cercarial shape and the capacity to form pericercarial sero-envelopes"* (Stirewalt *et al.*, 1966)

A later report from the same group (Stirewalt *et al.*, 1969) refined their own protocol (Stirewalt *et al.*, 1966) and tested the influence of different factors (such as temperature, light, number of cercariae applied, etc) on the yield of recovered schistosomula. The authors established that approximately 51% of the applied cercariae successfully transformed into schistosomula and that the final preparation lacked post-penetration larvae (cercariae that had penetrated through the layer of skin but did not comply with the definition of schistosomula). Later reports by Clegg and Smithers (1972) slightly modified the protocol from Stirewalt *et al.,* (1969) by using non-dried freshly excised skin from mouse abdomen. Schistosomula were phenotypically comparable to those obtained using rat skin (Stirewalt *et al.*, 1966) and also to those recovered from infected animals (*in vivo* collection) (Clegg *et al.*, 1972). However, the yield of schistosomula decreased using mouse skin (20-30%) compared to the optimised conditions listed above. In all the experiments and according to the authors, these approaches yielded tail-free and cercariae-free schistosomula preparations.

The *in vitro* mechanical transformation was introduced in 1974 (Ramalho-Pinto *et al.*, 1974). In this approach, the authors used centrifugation of freshly shed cold cercariae followed by incubation in culture media at 30°C for 40 minutes to produce viable schistosomula. Separation of the tails and isolation of cercariae bodies was achieved by sedimentation. Further incubation of the cercariae bodies in culture media and mild shaking induced changes in the parasites surface completing their transformation. This process is much more efficient than the skin penetration approaches,

transforming 50-70% of the initial cercariae subjected to the procedure. This incremental improvement in total number of recovered parasites and the ease of the experimental protocol represented the major attractions of the approach for subsequent investigators. Furthermore, these parasites fulfilled the criteria for schistosomula suggested by Stirewalt *et al.,* (1966). Brink *et al.,* (1977) improved the original mechanical transformation protocol by adding a Ficoll gradient step to obtain a schistosomula preparation virtually free from tails. To validate their results, the authors compared these "mechanically transformed schistosomula" (MT) with preparations obtained using the skin penetration approach (ST) (Clegg *et al.*, 1972) but observed that ST were different from MT in several aspects:

- Pre-acetabluar glands from ST were emptied after penetration, whilst the MT retained the contents in the glands.
- 3-hour old ST schistosomula showed evidence of presence of human A and B blood group-like antigens whilst MT schistosomula did not.
- In infectivity tests, a slightly higher percentage of ST parasites where recovered from mice compared to MT. However, the difference between these two groups was not significant.

On the other hand, several important factors did not differ between ST and MT, such as loss of the glycocalyx (as early as 2 hours) or the replacement of the trilaminate surface membrane by the heptalaminate membrane structure, which is a characteristic of transformation of cercariae into schistosomula. After 12 days, a higher percentage of ST (50-70%) had completed the process of closing the gut[1], while only 25-50% of the MT population had reached the same stage of development. This was not interpreted as a physiological difference but a delay in the development of the parasites. The authors explained this effect by arguing that during the ST there is a selection for the most "fit" cercariae as many of them die during the penetration process. In this case, the percentage of "gut-close" parasites in an entire population of ST is higher than that observed in MT where all cercariae (fit and no so fit) transform into schistosomula. This work showed that ST and MT are morphologically almost equivalent to each other (both change their surface membrane, both shed their glycocalyx) but that a slight "delay" can occur in their progression to the "gut-close" stage in the MT. Interestingly, MT schistosomula produced by the methods employed by Brink *et al.,* (1977) did not empty the contents of their pre-acetabular glands (Brink *et al.*, 1977), a requirement of the suggested definition of schistosomula (Stirewalt *et al.*, 1966). The authors argued that they do not consider the

[1] This stage is reached approximately 15 days after transformation and it describes parasites that have both the ends of the ceca joined. It is often called "gut-close" stage and it was defined in Clegg, (1965).

state of the pre-acetabular glands as a criterion to define a schistosomulum (Brink *et al.*, 1977). More recent studies have suggested that some of the contents of the acetabular gland complex are not emptied until the parasites are located deeper in the skin of the invaded host, probably because different secretion cocktails are needed to penetrate through different barriers (Crabtree *et al.*, 1985). This would support the claim of Brink *et al.,* (1977) that the status of the pre-acetabular glands after transformation is not a criterion for the definition of a schistosmulum.

The implementation of the MT together with the work of several groups that evaluated ST and MT parasites in terms of their phenotype and biochemistry provided enough evidence that these parasites were equivalent.

High throughput gene expression studies (i.e. microarrays) have relied on this equivalency, mainly because obtaining enough material from *in vivo* infections is not always feasible. Therefore, MT schistosomula have been used almost exclusively in high-throughput studies of gene expression (Verjovski-Almeida *et al.*, 2003; Dillon *et al.*, 2006; Fitzpatrick *et al.*, 2009; Gobert *et al.*, 2010; Farias *et al.*, 2011), identification of drug targets (Fitzpatrick *et al.*, 2009) and identification of effective drugs against schistosomes from a compound library (Abdulla *et al.*, 2009). In many cases, high-throughput gene expression studies are used as a "fishing expedition" where interesting genes, or at least those relevant for a particular process (i.e., invasion, host-parasite interaction, etc) can be pointed out based on their pattern of expression. Using inadequate samples could lead to miss interpretation of the results. For example, the mechanical transformation protocol might be subjecting the parasites to artificial sources of stress that could elicit a response from the parasites. In the case of high-throughput studies this could lead to the follow-up of genes expressed as a result of these artefacts instead of signals arising from the natural transformation of the cercariae into schistosomula. Another scenario would be to miss exploitable vulnerabilities that could arise from cues triggered by the skin barrier; such signals would be missed in MT parasites. As an example, cercariae of human-infectious *Schistosoma* spp. are known to react to fatty acids as chemoattractans (Haeberlein *et al.*, 2008) and these induce the release of contents of the acetabular glands (Stirewalt, 1978). It is expected that downstream signals deriving from these chemical cues be seen in the skin-transformed parasites but not in their mechanical counterparts. Therefore, validation at the transcriptional level of the equivalency of *in vitro* (mechanically transformed schistosomula) and *in vivo* obtained samples is important to justify these studies.

So far, there has been only one study reporting such a comparison. The work of Chai *et al.,* (2006) used a microarray platform to investigate the gene expression profile of

S. japonicum parasites from the lung stage (Chai *et al.*, 2006). They studied two lung-stage schistosomula parasite populations: one obtained *in vitro* (mechanical transformation followed by *in vitro* cultivation for three days, this preparation will be referred as MTS) and the other prepared *in vivo* (obtained from lungs from infected mice three days post infection, referred as IVS). A comparison with adult worms was also presented. Their results show that MTS and IVS transformed parasites were phenotypically identical with the exception that some MTS retained part of the cercarial glycocalyx. However, in terms of differential expression, the authors found that a striking 6,662 genes were differentially expressed, with a p-value ≤ 0.001 (but not corrected for multiple hypothesis/multiple testing), between MTS and IVS (3,207 and 3,455 genes with higher expression in the MTS and IVS respectively). As a comparison, the total number of differentially expressed genes observed between the MTS and the adult stage was found to be significantly less (just 3,777 genes) (Chai *et al.*, 2006).

One of the reasons that might have caused the large number of differentially expressed genes between MTS and IVS samples is the heterogeneity of the parasite population. As introduced in Chapter 1, the skin penetration and the passage through the endothelial wall are processes that vary greatly in time (skin penetration can occur between 0 and 40 hours and endothelial passage may take up to 8 hours) and therefore transformation of parasites within a population is not synchronised. This heterogeneity would most likely also be reflected in the repertoire of expressed genes.

On the other hand, mechanically transformed parasites are easier to synchronize and represent a more homogeneous population facilitating the analysis of differentially expressed genes. The large number of genes appearing as differentially expressed between these two populations might be an artefact of the heterogeneous population of *in vivo* recovered parasites used in this experiment. The authors conclude that MTS parasites do not represent their *in vivo* (IVS) age equivalents (Chai *et al.*, 2006).

The identification of transcripts appearing differentially expressed in either mechanically- or skin-transformed schistosomula provides the means to establish the differences between these two ways of preparing schistosomula from the gene expression perspective. Given that the early schistosomula stage represents the onset of infection where the schistosomes start their parasitic life in the mammalian host, it is important to understand and identify to what extent the mechanisms of transformation affect these parasites. In the following sections, 24-hour old skin-transformed and mechanically transformed schistosomula transcriptomes are compared and their differences are established.

.

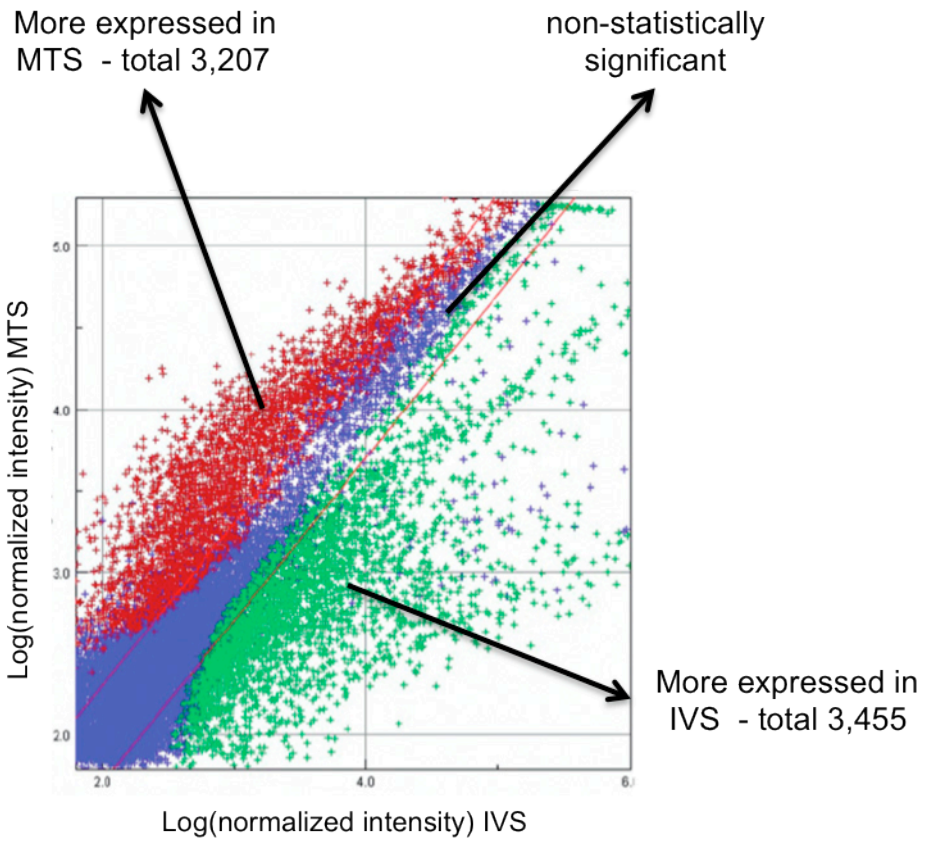


Figure 4.1 – Previous reports of comparison between *in vitro* and *in vivo* samples for *S. japonicum*. This scatter plot was taken from (Chai *et al.*, 2006) to illustrate the differences observed between lung-stage schistosomula obtained *in vivo* (x axis) and schistosomula obtained *in vitro* (y axis). The authors found 3,207 genes over expressed in the MTS (green crosses) and 3,455 genes over expressed in the IVS (red crosses) with p-value < 0.0016.

4.2 Results

4.2.1 Assessment of transformed parasites

The first step in the study of transcriptional differences between 24-hour old parasites recovered both *in vitro* or *in vivo* is the collection of the parasitic material. The protocols used for this were introduced in Chapter 2 sections 2.2. The following three sections in this chapter explain the steps taken to optimise the aforementioned protocols and the experimental approaches performed to evaluate the successful transformation of schistosomula.

4.2.1.1 Optimization of transformation protocols

Both mechanical and skin transformation protocols were subjected to optimization. In the case of the mechanical transformation two main aspects of the sample preparation were taken into close consideration for optimization: presence of contaminating tails plus the number and proportion of "healthy" parasites in the final schistosomula preparation. The concentration and temperature of Percoll were found to be important factors in obtaining tail-free schistosomula preparations. Percoll concentrations ranging from 60% to 80% were tested. The lowest percentage of Percoll tested (60%) resulted in an significant number of tails contaminating the final schistosomula preparation (**Figure 4.2A**) while the higher percentage of Percoll tested (80%) reduced the number of tails but also the number of schistosomula found after separation of tails and cercarial bodies (**Figure 4.2B**). A 70% Percoll solution was found to be a good compromise between the two extremes providing a virtually tail-free schistosomula preparation (~1% tails) with maximum yield of transformed parasites. Additionally, better separations were obtained when using an ice-cold solution of Percoll rather than one kept at room temperature.

The second aspect to be evaluated was the fraction of viable parasites obtained with the mechanical transformation protocol. The criterion used to catalogue viable individuals was introduced in Chapter 2 – section 2.2.4. Initially, 23 passages through a syringe needle were used to generate the separation of the heads and tails. This resulted in a major fraction of parasites showing characteristics of non-viable individuals (**Figure 4.2A,** ~15%): some would be dead, showing a uniform shape with granular appearance and lack of motility; while others would show evidence of morphological damage, also showing a granular appearance and adopting different shapes usually with some level of motility. By

introducing a step of vigorous shaking in a vortex the number of passages through the syringe could be reduced from 23 to 12 obtaining similar numbers of MT parasites with an increased proportion of viable individuals (from ~15% non-viable to just ~1%, **Figure 4.2B**). This improvement was probably due to the reduction in the number of passages through the syringe needle, which may represent a very harsh treatment to the parasites.

Regarding the skin transformation protocol, the main factor affecting the recovery of tail-free preparations of schistosomula was the number of cercariae deposited in the upper compartment of the transformation apparatus. Numbers ranging from 4,000 to 20,000 cercariae were tested and a maximum yield of transformed parasites, virtually free from tail contamination (~ 1 to 4%), could be obtained by placing approximately 14,000 cercariae in the upper compartment of the transformation apparatus. However, tail contamination varied greatly from one experiment to another even among transformation apparatuses within the same experiment. Therefore, schistosomula recovered from individual assemblies were observed under the light microscope to assess contamination and only those with < 4% tails were considered for further experimentation. The sample preparation procedure had to be undertaken approximately 20 times (with 12-13 transformation apparatuses each time) before sufficient number of parasites could be collected.

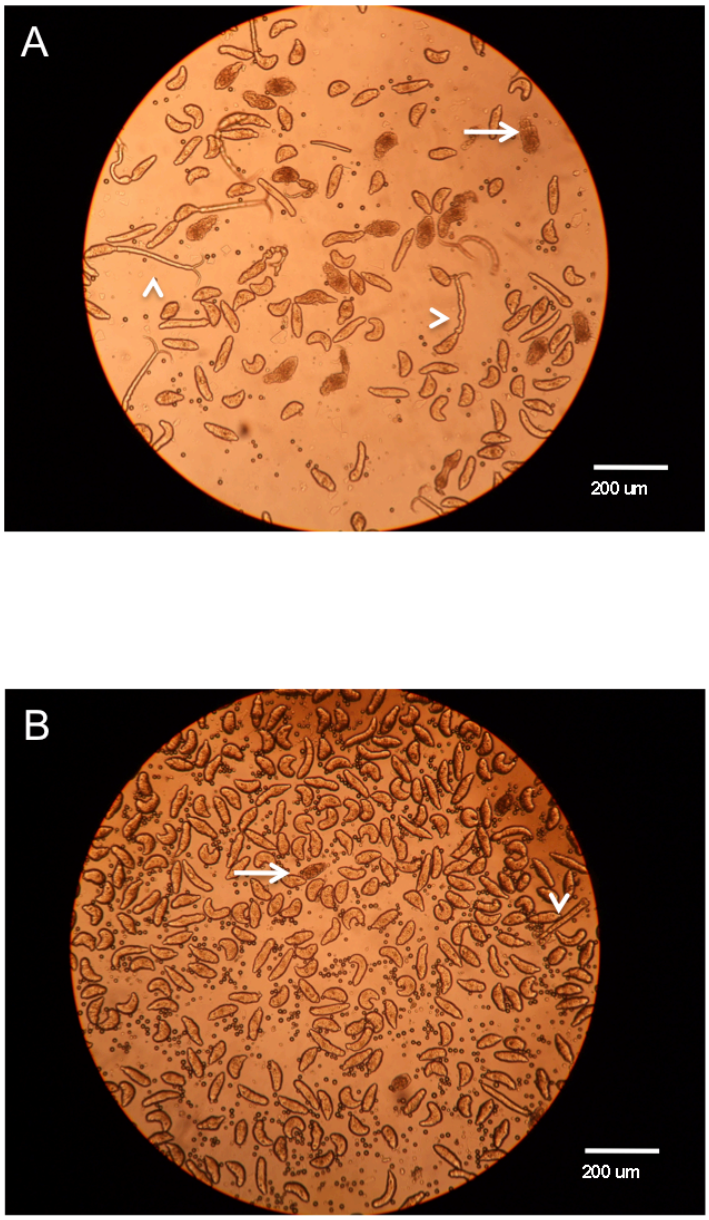


Figure 4.2 – Improving the mechanical transformation protocol. A – Parasites were transformed according to protocol described in Chapter 2 section 2.2.2 except that parasites were subjected to 23 passages through a syringe needle and separation of heads and tails was performed with 60% Percoll solution at room temperature. B - Transformation of parasite was performed using the optimized protocol described in Chapter 2 section 2.2.2 which involved shaking of the cercariae suspension in a vortex mixer followed by only 12 passages through a syringe needle. Separation of cercarial heads from tails was performed using 70% ice-cold Percoll solution. Arrows indicate damaged schistosomula, arrowheads point to contaminating cercaria or tails. Both panels (A and B) show light microscope images of 3-hour old mechanically transformed schistosomula.

4.2.1.2 Parasite viability

The quality of the transformed parasites, obtained either through MT or ST, was tested by phenotypic evaluation in an inverted microscope. Apart from testing for the characteristics of non-viable parasites (introduced in the Section 4.2.1.1), assessing the parasites' capability to remain alive when cultured *in vitro* for a given period of time can also be used as a parameter for evaluating fitness of parasites. Both MT and ST schistosomula were successful in progressing to the 2-weeks old stage. However, at this point both parasite populations were more phenotypically heterogeneous compared to those at 24 hours. These preparations contained parasites whose shape resembled the typical 2-weeks old worm grown in culture (much longer and thinner than the 24 hours schistosmulum) as well as parasites that showed a much more contracted shape resembling 3- or 24-hour old schistosomula. Microscopic evaluation of the phenotypic characteristics of both MT and ST parasites fulfilled the criteria used by others (Crabtree *et al.*, 1980) to define viable parasites at the lung stage indicating that both MT and ST protocols generated viable organisms.

4.2.1.3 Changes to the parasite surface

As introduced in Chapter 1 section 1.2.3, the parasites change the composition of their surface upon transformation. It has been shown that newly transformed schistosomula bind concanavilin A (Con A) to their surface (Samueleson *et al.*, 1989). This characteristic was used to evaluate whether the MT and ST transformation protocols triggered the remodelling of such structures. **Figure 4.3** shows Con A-FITC binding to both MT and ST schistosomula 5 hours after transformation. Consistent with previous reports (Samuelson *et al.*, 1982), Con A binding is positive for schistosomula but negative for the cercarial tails. Both MT and ST parasites bind Con A with the typical differential distribution of the Con A binding sites. Pronounced binding of Con A is observed at the apical end and the ventral sucker of the schistosomula, consistent with previous observations (Samuelson *et al.*, 1982).

A difference in the intensity of the fluorescence between samples (more in the MT parasites) can be seen in **Figure 4.3**. This might be due to differences in the sample itself or in the image acquisition procedure. Nevertheless, both samples positively bind Con A indicating that they have indeed transformed their surface membrane.

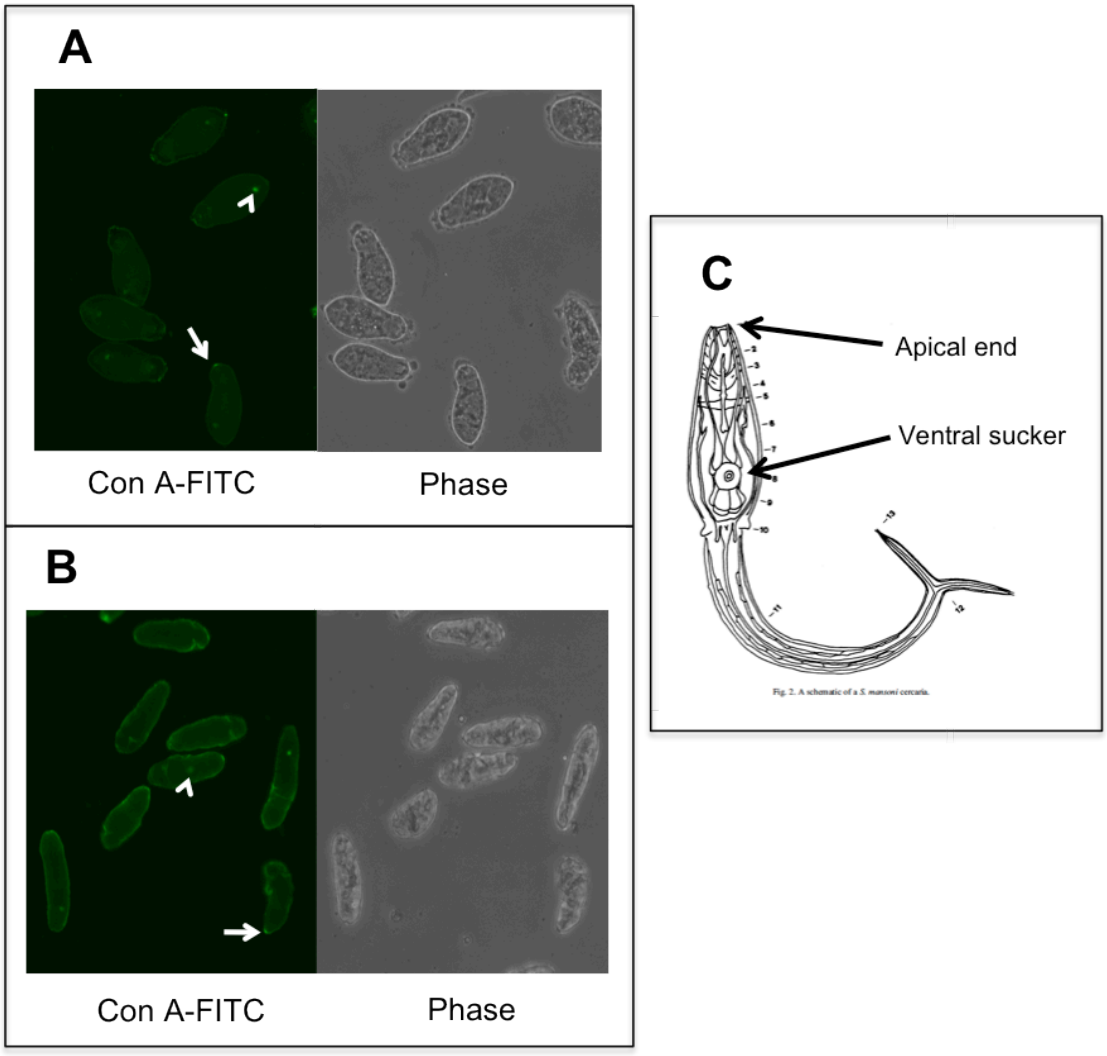


Figure 4.3 – Concanavilin A (Con A) binding to the surface of schistosomula but not to the cercariae (not shown) as a consequence of the uncovering of epitopes as evidence of transformation. Con A binds to both ST (A) and MT (B) schistosomula in similar fashion. (C) is included as a reference to the cercariae anatomy. Increased binding of Con A-FITC is observed in the apical end (arrows) and ventral sucker (arrowheads) of the parasites. Top and bottom panels are replicates of each sample. Note that the shape of the parasites in (A) and (B) are different and this is due to the sample preparation (see text).

A slight difference in the schistosomula size in the two samples can be observed in **Figure 4.3**. This was due to the differential conditions in which the samples were prepared for observation. MT schistosomula were left to move freely therefore appearing more elongated and slender, while the ST parasites were photographed using a cover slip limiting their movement and appearing rounder and shorter than the MT. It is not expected that these differences in specimen's preparation would affect the observations.

In summary, observations obtained from analysis and Con A binding as well as the parasite's capability of surviving for at east 2 weeks in culture led to the conclusion that the methods used to transform cercariae into schistosomula, both MT and ST, provide the necessary cues to trigger the transformation from the free-living cercariae to the parasitic schistosomula.

4.2.2 Differential expression between mechanical and skin transformed schistosomula.

After assessing that both parasite preparations fulfilled the criteria of schistosomula, samples were collected, RNA was extracted and RNA-seq libraries were prepared as described in Chapter 2.

In order to analyse the differences between MT and ST parasites at the transcriptional level, RNA-seq data for 24-hour old MT and ST schistosomula were generated. An overview of the results obtained from sequencing these libraries (number of reads, percentage mapped to genome, etc) were presented in Chapter 3 section 3.2.2. Reads per transcript and RPKM values for each gene were calculated as described in Chapter 2 section 2.6.1.1. Analysis (**Figure 4.4**) of the two 24-hour old schistosomula samples showed that they are highly correlated with Pearson's product and Spearman's rank coefficients of 0.98 and 0.99 respectively. Indeed, these correlation values resemble those obtained for biological replicates as previously shown in Chapter 3 section 3.2.2.1.3. Previous studies, for example the one from Chai *et al.,* (2006) presented in the introduction of this chapter, used deviations from the correlation to identify differentially expressed genes. Such approach cannot be used in this study because of the high correlation of the samples.

The software package edgeR (Robinson *et al.*, 2010) was used to achieve a more detailed analysis of transcripts differentially expressed between these two samples. First, all transcripts with levels of expression that could be attributed to noise were removed from the dataset by filtering out all transcripts with RPKM values lower than the

empirically determined background (see Chapter 2 section 2.6.2), corresponding to an RPKM of 2. This reduced the total number of transcripts from 10,852 to 8,715 (2,137 transcripts had negligible expression in both samples). After filtering the reads and using a statistical significance cut-off of adjusted p-value < 0.01, edgeR revealed 77 differentially expressed transcripts. By loosening the statistical criterion (adjusted p-value < 0.05) a total of 149 differentially expressed transcripts were found. Of these, 93 transcripts were more highly expressed in the ST parasites (**Table 4.1**) while 49 transcripts were more highly expressed in the MT (**Table 4.2**). A graphical representation of differentially expressed transcripts is shown in **Figure 4.5**).

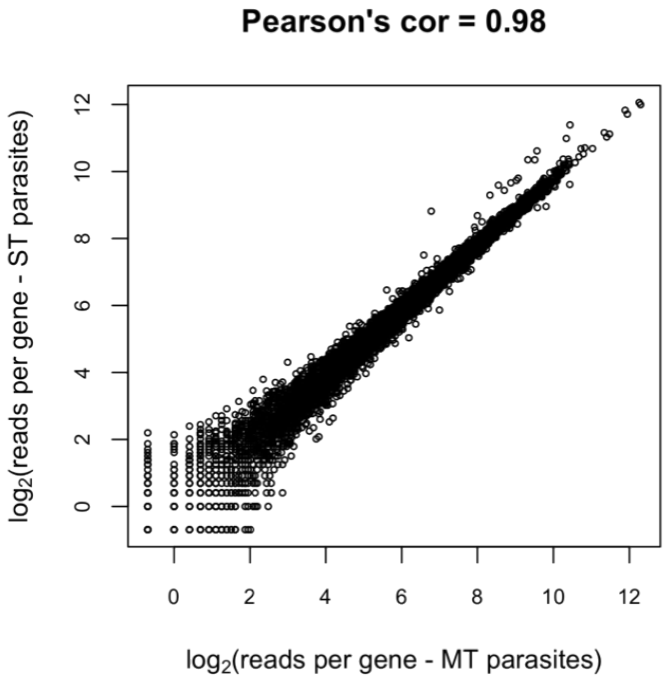


Figure 4.4 – Correlation of expression values of transcripts for the mechanically transformed (x-axis) and skin transformed (y-axis) schistosomula. Both Pearson and Spearman's correlations are high (0.98 and 0.99 respectively) indicating very low variability between these two samples.

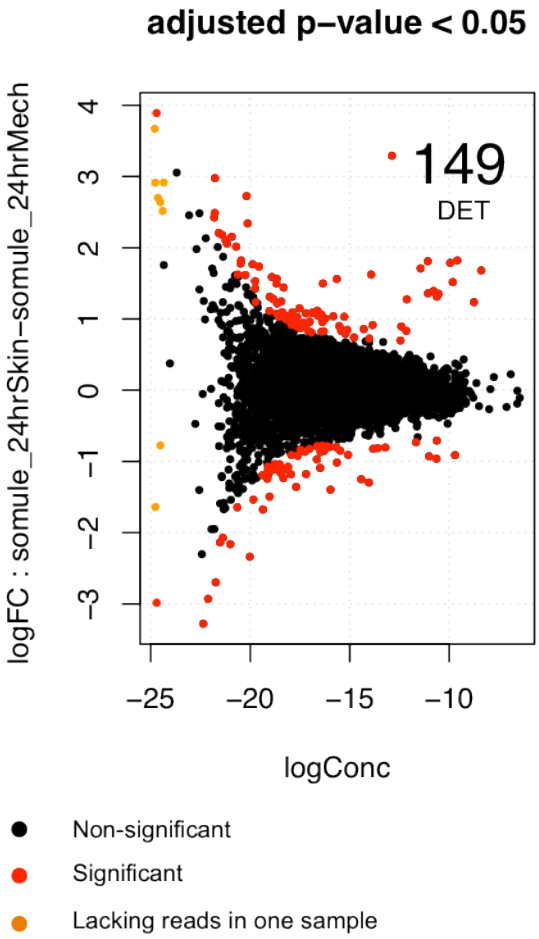


Figure 4.5 - Differential expression of transcripts. The comparison between MT and ST parasites at 24 hours after transformation (adjusted p-value < 0.05) is shown. Relative concentration (x axis) is plotted against fold change values (y axis) in the $\log_2$ scale. Positive $\log_2$ fold changes represent transcripts more highly expressed in ST schistosomula while negative $\log_2$ fold changes represent transcript more highly expressed in MT schistosomula. Red dots indicate differentially expressed transcripts below the statistic significance cut off value and therefore statistically differentially expressed; black dots represent transcripts above the statistic significance cut off value of differential expression; orange dots represent transcripts expressed in one sample but not in the other in which cases both the fold change value and the relative concentration are estimated. DET = differentially expressed transcripts.

Table 4.1 - Transcripts over expressed in ST *vs*. MT

91 genes; adjusted p-value cut-off is 0.05

GeneDB_ID	log_2FC	Product description
Smp_197430.1	28.57	Hypothetical protein
Smp_900100.1	3.29	NADH dehydrogenase subunit 3
Smp_199840.1	2.99	Nucleolar protein c7b, putative
Smp_204970.1	2.64	Na
Smp_144640.1	2.48	Hypothetical protein
Smp_172770.1	2.31	Hypothetical protein
Smp_200080.1	2.31	Na
Smp_119730.1	2.23	Hypothetical protein
Smp_205470.1	2.17	Na
Smp_202510.1	2.10	Na
Smp_139420.1	2.07	Hypothetical protein
Smp_177710.1	2.04	Hypothetical protein
Smp_205950.1	1.99	Na
Smp_900040.1	1.82	NADH dehydrogenase subunit 2
Smp_202050.1	1.82	Na
Smp_900020.1	1.81	NADH dehydrogenase subunit 6
Smp_900110.1	1.79	NADH dehydrogenase subunit 1
Smp_029780.1	1.76	Hypothetical protein
Smp_197440.1	1.75	Hypothetical protein
Smp_149340.1	1.73	Hypothetical protein
Smp_900060.1	1.71	Cytochrome c oxidase subunit 3
Smp_900050.1	1.68	NADH dehydrogenase subunit 5
Smp_028850.2	1.62	Hypothetical protein
Smp_151800.1	1.62	Hypothetical protein
Smp_159800.1	1.61	MEG-2 (ESP15) family
Smp_202920.1	1.60	Na
Smp_202120.1	1.56	Na
Smp_096750.1	1.54	Hypothetical protein
Smp_900090.1	1.52	NADH dehydrogenase subunit 4
Smp_169830.1	1.51	Hypothetical protein
Smp_159810.1	1.51	MEG-2 (ESP15) family
Smp_127860.1	1.46	Hypothetical protein
Smp_205870.1	1.46	Na
Smp_131830.1	1.40	Hypothetical protein
Smp_067800.1	1.39	Fibrillin 2, putative
Smp_900030.1	1.36	Atpase subunit 6
Smp_900070.1	1.35	Cytochrome B
Smp_900010.1	1.31	Cytochrome c oxidase subunit 2

Table 4.1 - Transcripts over expressed in ST *vs*. MT (cont)		
Smp_201190.1	1.31	Na
Smp_157330.1	1.30	Hypothetical protein
Smp_900080.1	1.28	NADH dehydrogenase subunit 4L
Smp_203200.1	1.27	Na
Smp_056260.1	1.26	Beta-1,4-galactosyltransferase, putative;with=uniprot:Q9GUM2
Smp_900000.1	1.24	Cytochrome c oxidase subunit 1
Smp_170650.1	1.23	Hypothetical protein
Smp_170630.1	1.15	Hypothetical protein
Smp_028840.1	1.13	Hypothetical protein
Smp_047400.1	1.11	Hypothetical protein
Smp_146940.1	1.09	Innexin, putative;with=Pfam:PF00876
Smp_146760.1	1.09	Hypothetical protein
Smp_131730.1	1.08	Hypothetical protein
Smp_107750.1	1.08	Hypothetical protein
Smp_133340.1	1.08	Hypothetical protein
Smp_007950.1	1.06	Beta-1,4-galactosyltransferase, putative;with=uniprot:Q80WN7
Smp_008400.1	1.06	Adenosine kinase, putative
Smp_195130.1	1.04	KRAB-A domain-containing protein
Smp_133660.1	1.04	Lin-9, putative
Smp_192220.1	1.04	Hypothetical protein
Smp_111640.1	1.02	Hypothetical protein
Smp_165250.1	1.02	Hypothetical protein
Smp_200450.1	1.02	Na
Smp_177250.1	1.02	Histone deacetylase, putative
Smp_204360.1	1.00	Na
Smp_183870.1	1.00	Hypothetical protein
Smp_142120.1	0.99	Achaete-scute transcription factor-related
Smp_200110.1	0.98	Na
Smp_168400.1	0.97	Hypothetical protein
Smp_169680.1	0.97	G-protein coupled receptor fragment, putative;with=uniprot:Q18179
Smp_194280.2	0.95	Hypothetical protein
Smp_193700.1	0.95	Hypothetical protein
Smp_200940.1	0.95	Na
Smp_051110.1	0.94	Hypothetical protein
Smp_144480.1	0.93	Hypothetical protein
Smp_117340.1	0.92	Hypothetical protein
Smp_022290.1	0.91	Hypothetical protein
Smp_131710.1	0.91	Hypothetical protein
Smp_151600.1	0.90	Neuronal calcium sensor, putative
Smp_057860.1	0.90	Hypothetical protein
Smp_126290.1	0.88	Hypothetical protein

Table 4.1 - Transcripts over expressed in ST *vs*. MT (cont)		
Smp_195030.1	0.86	ABC transporter subunit, putative
Smp_163800.1	0.85	Hypothetical protein
Smp_028860.1	0.83	Hypothetical protein
Smp_132670.1	0.83	Myosin regulatory light chain, putative
Smp_032970.1	0.83	Calmodulin, putative
Smp_164550.1	0.83	Hypothetical protein
Smp_039590.1	0.82	Hypothetical protein
Smp_166020.1	0.81	Hypothetical protein
Smp_155320.1	0.79	Hypothetical protein
Smp_144910.1	0.74	Collagen alpha-1(V) chain precursor, putative
Smp_163630.1	0.72	Expressed protein 10.3; MEG-4 (10.3) family
Smp_211020.1	0.70	Na

Na – Not assigned

Table 4.2 – Transcripts over expressed in MT vs. ST

58 genes; adjusted p-value cut-off is 0.05

GeneDB_ID	log_2FC	Product_description
Smp_180340.1	28.89	MEG-2 (ESP15) family
Smp_156200.1	3.30	Hypothetical protein
Smp_200150.1	2.78	Na
Smp_199820.1	2.66	Serine/threonine kinase;with=uniprot:P16912
Smp_204890.1	2.37	Na
Smp_116960.1	2.26	Hypothetical protein
Smp_172960.1	2.08	Kunitz-type protease inhibitor, putative;with=uniprot:P00978
Smp_199900.1	1.99	Phospholipid transport protein;with=uniprot:Q99J08
Smp_070940.1	1.62	Hypothetical protein
Smp_113760.1	1.61	Anti-inflammatory protein 16, putative
Smp_052760.1	1.56	TFIIH basal transcription factor complex TTD-A subunit (General transcription factor IIH polypeptide 5) (TFB5 ortholog), putative
Smp_203150.1	1.46	Na
Smp_047680.1	1.40	Ferritin, putative;with=uniprot:P25320
Smp_203400.1	1.34	Na
Smp_147730.1	1.29	Kunitz-type protease inhibitor, putative;with=uniprot:P00978
Smp_200290.1	1.28	Na
Smp_047660.1	1.25	Ferritin, putative;with=uniprot:P25320
Smp_013600.1	1.22	Nk homeobox protein, putative
Smp_201470.1	1.20	Na
Smp_018510.1	1.19	Hypothetical protein
Smp_159390.1	1.19	Hypothetical protein

Table 4.2 – Transcripts over expressed in MT vs. ST (cont)

Smp_195090.1	1.17	Hypothetical protein
Smp_195180.1	1.17	Surface membrane antigen;with=uniprot:Q04171
Smp_168610.1	1.17	Myelin transcription factor 1, myt1, putative
Smp_053950.1	1.16	Heterogeneous nuclear ribonucleoprotein, putative
Smp_205660.1	1.12	Na
Smp_146230.1	1.10	Hypothetical protein
Smp_125130.1	1.09	Hypothetical protein
Smp_170380.1	1.09	Hypothetical protein
Smp_100350.1	1.07	Hypothetical protein
Smp_198540.1	1.06	Hypothetical protein
Smp_120280.1	1.06	Hypothetical protein
Smp_079120.1	1.05	Hypothetical protein
Smp_204260.1	1.02	Na
Smp_147380.1	0.98	Hypothetical protein
Smp_113660.1	0.97	Cleavage and polyadenylation specificity factor, putative
Smp_002150.1	0.96	subfamily S1A unassigned peptidase (S01 family);with=UniProt:P20918
Smp_200850.1	0.93	Na
Smp_143190.1	0.92	hypothetical protein
Smp_089670.1	0.91	macroglobulin/complement, putative;with=UniProt:Q63041
Smp_134870.1	0.91	early growth response protein, putative
Smp_139500.1	0.90	hypothetical protein
Smp_089550.1	0.86	Eif5b-like protein, putative
Smp_158480.1	0.86	AMP dependent ligase, putative
Smp_179170.1	0.85	family C13 non-peptidase homologue (C13 family);with=UniProt:P09841
Smp_161650.1	0.84	hypothetical protein
Smp_125710.1	0.83	hypothetical protein
Smp_146360.1	0.82	hypothetical protein
Smp_063330.1	0.82	hypothetical protein
Smp_194050.1	0.81	Clumping factor A precursor (Fibrinogen-binding protein A) (Fibrinogen receptor A), putative
Smp_133830.1	0.80	hypothetical protein
Smp_064280.1	0.80	hypothetical protein
Smp_184280.1	0.79	hypothetical protein
Smp_158650.1	0.78	hypothetical protein
Smp_000790.1	0.78	actin binding LIM protein family member 2-related;with=UniProt:Q6H8Q1
Smp_151640.1	0.77	hypothetical protein
Smp_201910.1	0.73	Na
Smp_124000.1	0.71	MEG-14

Na – Not assigned

4.2.2.1 Genes more highly expressed in skin-transformed schistosomula

4.2.2.1.1 Mitochondrial transcripts

The results shown in Table 4.1 indicate that the skin-transformed parasites have higher expression of mitochondrial genes. These are also shown in **Figure 4.6A** where they cluster together at relative high concentrations and more highly expressed in the ST parasites. Their linear fold change values range from 2.3 to 9.8 and they are all statistically significant.

In order to investigate whether this difference in the rate of expression of the mitochondrial genes had any consequence on the metabolic activity of the parasites, the AlamarBlue® (AB) assay was used. As introduced before (Chapter 2 section 2.3.9), AB is a good indicator of mitochondrial activity (Springer *et al.*, 1998) through the measurement of redox species generated by the respiratory electron chain.

As a first step, a titration was performed to find the minimum number of parasites needed to identify a significant difference in absorbance between wells containing parasites and a blank. This is necessary because the number of parasites obtained from ST is very low and optimization of the protocol is a critical step in order not to waste valuable samples. For simplicity, these titration experiments were only performed using MT parasites. Although a statistical difference was found between wells with parasites and the blanks, no significant difference in absorbance could be detected among wells with different numbers of 3-hour old schistosomula (**Figure 4.6B**). However, experiments using 24-hour old parasites showed a significant difference between each population of parasites (250, 500 and 1000 parasites per well) and their blank. It was concluded that 250 parasites per well are sufficient to detect a difference in absorbance reflecting the metabolic activity in parasites compared to blanks at 24 hours after transformation.

Second, MT and ST parasites were tested (**Figure 4.6C**). A significant difference (t-test, p-value < 0.01) could be observed between 6-hour old parasites originating from MT and ST (red boxplots). Both 6-hour old MT and ST parasites were different from the blank and from each other, suggesting that parasites at this stage are already distinguishable in terms of metabolic activity. Interestingly, MT and ST parasites at 24 hours after transformation do not show significant difference between them (green boxplots) when incubated in AB for 3 hours. However, they are significantly different from their blank indicating that AB has reacted with species present in the sample. Increasing the

incubation time in AB is known to provide better resolution simply because there is a higher concentration of species that react with AB (manufacturer instructions). As expected, when parasites were left to grow for 24 hours in the presence of AB they showed a much more pronounced increment in the accumulation of metabolites (compared to blank wells) that reflected in a greater difference in the absorbance with respect to the blank. What is more, a significant difference between MT and ST could be established, suggesting that these two populations of parasites are indeed metabolically different with the ST parasites being more "metabolically active" than the MT after 24 hours in *in vitro* culture.

In summary, the differential expression analysis indicated an over expression of mitochondrial transcripts in the ST compared to the MT parasites. Increased numbers of mitochondrial transcripts could be responsible for the observed increment of mitochondrial activity, either by increased protein levels or because a higher number of active mitochondria are found in this parasites. This observation agrees with the explanation proposed by Brink *et al.,* (1977) – discussed earlier – that ST parasites are a selection of the most "fit" cercariae. If MT preparations contain a mixture of fit and less-fit parasites not all of them would be expected to develop at their maximum rate and, as seen here, their measurable metabolic activity would be reduced.

4.2.2.1.2 Other biologically relevant genes with higher expression in ST parasites

In order to identify other biological processes determined by the genes with higher expression in the ST parasites compared to the MT parasites, a GO term enrichment analysis of genes more highly expressed in the ST schistosomula was performed. These results are presented in **Table 4.3** and according to their description they can be further grouped as: mitochondrial function, G-protein couple receptors, microtubule movement, retrotransposon related functions, betaine synthesis and skeletal muscle development. An analysis of the higher rate of mitochondrial function inferred from the expression of mitochondrial genes was already presented in section 4.2.2.1.1.

Transcripts encoding putative G-protein couple receptors (GPCR - Smp_144640.1, Smp_117340.1 and Smp_169680.1) are more highly expressed in the skin-transformed parasites. Smp_144640.1 encodes a seven-trans-membrane domain and can be considered a putative GPCR receptor. The last two are better characterised: Smp_117340 is a PROF1 receptor and Smp_169680.1 is a neuropeptide receptor (Zamanian, 2011). PROF1 are a

group of GPCRs that are found exclusively in platyhelminths (Zamanian, 2011). Knowledge of the role of these GPCRs in helminths is still rudimentary; further experiments are needed to assess their roles in either cell-to-cell signalling or host-parasite interactions. Results presented here suggest that some of these GPCRs are more highly expressed when the parasites are transformed through a layer of skin.

Other GO enriched terms are microtubule-based movement. Two transcripts (Smp_131710.1 and Smp_170650.1) represent this term: the first of these transcripts is a small one-exon gene encoding a 891 amino acids polypeptide with no predicted conserved domains.

The second is a multi-exon gene encoding a conserved kinesin domain. Kinesin domains are involved in movement along microtubules and are often associated with proteins that have a role in transporting organelles along microtubules (Vale *et al.*, 2000). As previously described in the introduction of this chapter, chemoattractants found in the host skin induce the release of the contents of the acetabular glands (Stirewalt, 1978). Since these glands are unicellular, it would be expected that the vesicles containing secretory products (see Chapter 1 section 1.2.3.1) travel to the glands' openings transported by microtubules. Because the chemoattractant cues would be absent during the mechanical transformation the expression of these kinesin proteins may be induced at a lower level in MT parasites.

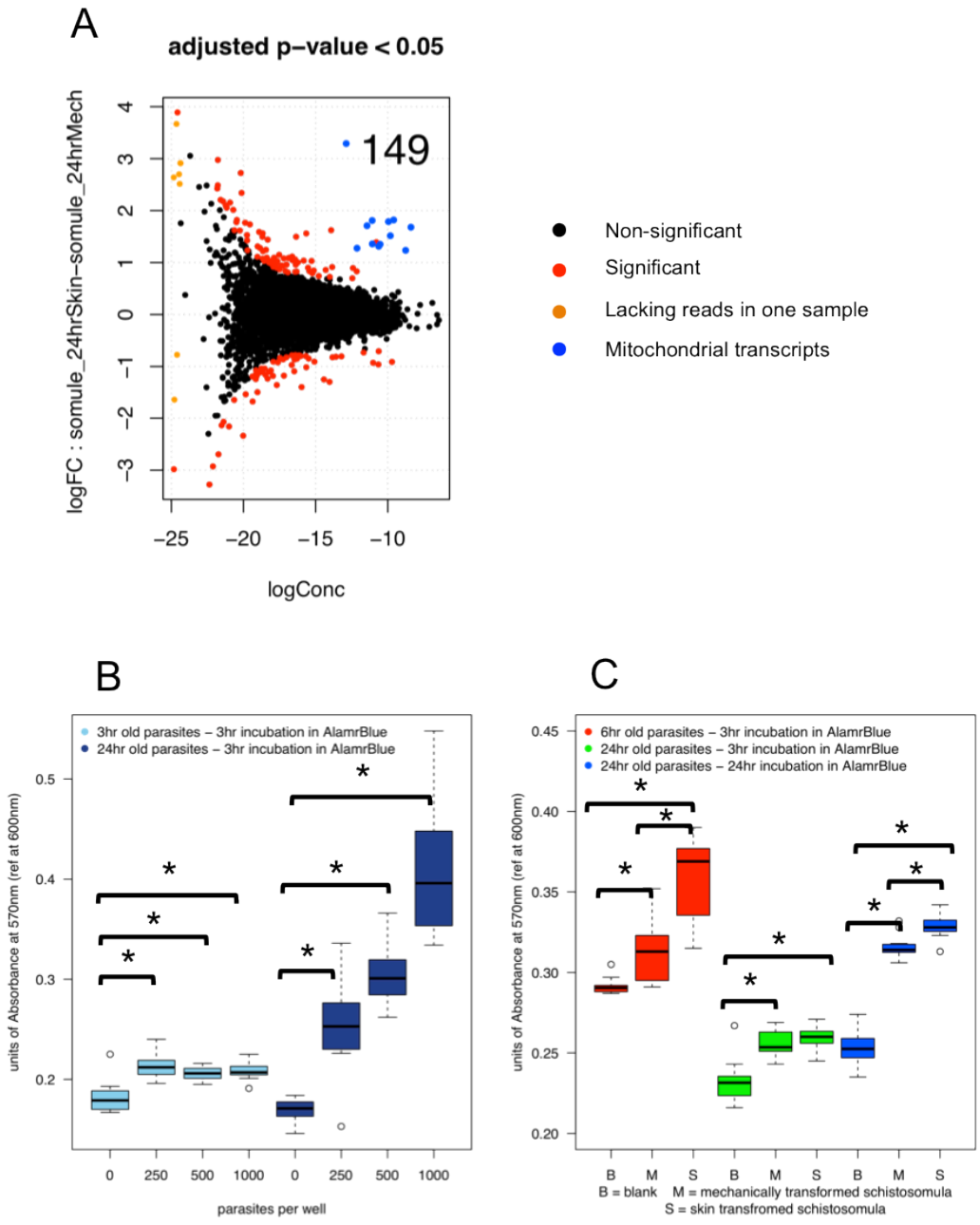


Figure 4.6 – Mitochondrial genes are differentially expressed in ST compared to MT. A – Mitochodrial transcripts (blue dots) are clearly separated from the rest of the differentially expressed transcripts. B – Titration of AlamarBlue® reactivity against different number of MT parasites at 3 hours after transformation (light blue) and 24 hours after transformation (dark blue). C – AlamarBlue® reactivity of 6-hour old and 24-hour old MT and ST schistosomula. Six-hours old with 3 hours incubation in AlamarBlue® (red boxplots) and 24-hours old schistosomula with 24 hours incubation in AlamarBlue® (blue boxplots) samples were statistically different from each other and compared to the blank. Twenty four-hour old schistosomula with 3 hours incubation in AlamarBlue® (green boxplots) were both statistically different from the blank but not between them (*t*-test, p.value = 0.25). Stars represent statistically different samples (*t*-test, p-value < 0.01).

Table 4.3 - Enriched Gene Ontology terms (Biological Processes).

A - 24 hours old SKIN transformed schistosomula

GO.ID	Term	p-value
GO:0022904	Respiratory electron transport chain	0
GO:0042773	ATP synthesis coupled electron transport	0
GO:0046797	Viral procapsid maturation [a]	0.008
GO:0015074	DNA integration [a]	0.011
GO:0032196	Transposition [a]	0.013
GO:0019285	Glycine betaine biosynthetic process from choline	0.014
GO:0006310	DNA recombination [a]	0.024
GO:0007018	Microtubule-based movement	0.032
GO:0019079	Viral genome replication [b]	0.037
GO:0006410	Transcription, RNA-dependent [b]	0.037
GO:0007186	G-protein coupled receptor protein signalling	0.040
GO:0042775	Mitochondrial ATP synthesis coupled electron transport	0.041
GO:0007519	Skeletal muscle tissue development	0.046

B - 24 hours old MECHANICALLY transformed schistosomula

GO.ID	Term	p-value
GO:0006879	Cellular iron ion homeostasis [c]	0
GO:0006826	Iron ion transport [c]	0.001
GO:0007548	Sex differentiation [d]	0.015
GO:0055114	Oxidation reduction [c]	0.018
GO:0006526	Arginine biosynthetic process	0.022
GO:0006269	DNA replication, synthesis of RNA primer	0.029
GO:0007530	Sex determination [d]	0.033
GO:0030154	Cell differentiation	0.037

Categories with the same super index letter are represented by the same transcripts.

Other differentially expressed transcripts not identified through the GO term analysis were also found biologically relevant in the process of transformation. EF-hand domains are calcium-binding domains and some can be associated to tegument antigens [such as Sm22.6/SmTAL1 (Dunne *et al.*, 1992; Fitzsimmons *et al.*, 2004)] in schistosomes. What is more, calcium-binding proteins have been associated in mechanisms of larval adaptation to the mammalian host (Kusel *et al.*, 2007). ST parasites show the over expression of four EF-hand encoding transcripts (see **Table 4.1** -Smp_151600.1 neuronal calcium sensor, putative; Smp_032970.1 calmodulin, putative; Smp_132670.1 myosin regulatory light chain, putative; Smp_137750.1 calbindin-32, putative). If these calcium-binding proteins have a role in the mechanisms of adaptation, it is possible that such mechanisms are triggered upon contact with the host skin. The lack/reduced expression of such transcripts in MT parasites might be associated with the lack of the appropriate signals.

4.2.2.2 Genes more highly expressed in mechanically-transformed schistosomula

In this section a description and analysis of the transcripts that are more highly expressed in the mechanically transformed schistosomula is presented. As done previously for the highly expressed transcripts in the ST parasites, GO term enrichment analysis was also used here as a guide for identifying biological processes among this group of genes. However, the study of the list of genes more highly expressed in the MT parasites alongside with their product description (**Table 4.2**) provided a good resource to associate these transcripts with processes that might be occurring in the parasites. Results emerging from these two sources are presented here.

4.2.2.2.1 Ferritins

The dataset of transcripts more highly expressed in the MT parasites present two transcripts associated with iron homeostasis and oxidation-reduction functions. These are ferritins and their main function is to bind iron and release it in a controlled way, avoiding the generation of reactive oxygen species (Chiancone *et al.*, 2004). In *S. mansoni* there are four ferritin genes (Smp_047640.1, Smp_047650.1, Smp_047660.1, Smp_047680.1) found in one cluster in Chromosome 2. The two ferritin genes found differentially expressed in the MT *vs.* ST (Smp_047660.1 and Smp_047680.1) share the higher percentage of similarity (98%) compared to the others in the cluster. Because 24-hours old schistosomula can ingest plasma molecules (Bennett *et al.*, 1991) it is possible that the role of ferritins is associated with uptake of iron from these sources or with later uptake of blood cells. This hypothesis is in agreement with the presence of ferritins in the gut vomit (Hall *et al.*, 2011), which suggests a role of these proteins in the process of feeding. However, it is not possible to infer why these transcripts are more highly expressed in the MT parasites.

4.2.2.2.2 Proteases

MT parasites show higher expression of transcripts encoding proteases. Proteases have a recognised important role in schistosomes. For example, adult worms use a set of aspartic proteases called cathepsins for the purpose of feeding (Brinkworth *et al.*, 2001) while cercariae use elastase and other proteases during the process of skin invasion (McKerrow *et al.*, 2002; McKerrow, 2003). Two proteases are shown more highly expressed in MT than in ST parasites.

The first one is encoded in the transcript Smp_002150.1, which is twice as much expressed in the MT parasites than in the ST parasites. The polypeptide product is a putative secreted serine protease (S1 family) from the trypsin family characterised by the conserved catalytic triad His/Asp/Ser (chymotrypsionogen numbering) towards their C-termini. Similarity searches using BLASTp (Altschul *et al.*, 1990) against the Uniprot database (Uniprot Consortium, 2009) showed similarity with other serine proteases in *S. japonicum, T. solium* and the *E. granulosus* antigen 5 precursor (Ag5). The latter is a well-characterised secreted antigen present in large quantities in the hydatid cyst fluid in the infected intermediate host. Interestingly, the serine in the catalytic triad of the *E. granulosus* Ag5 has been substituted by a threonine and probably compromises the serine protease activity of the enzyme (Lorenzo *et al.*, 2003). Contrary to Ag5, Smp_002150.1 has all residues from the catalytic site.

The second protease is Smp_179170.1, encodes a legumain C13 asparginyl endopeptidase domain. It is highly similar to another *S. mansoni* asparginyl endopeptidase Sm32 (Smp_075800 – former hemoglobinase), which is known to participate in activating gut-related proteins [cathepsin B and F (Dalton *et al.*, 1995)]. These two genes are located adjacent to each other approximately 8 kB distance in Chromosome 3. SignalP results (Emanuelsson *et al.*, 2007) suggest that Sm32 is secreted (0.96 probability) while the same prediction for the product of Smp_179170.1 resulted in inconclusive (0.5 probability). Closer analysis of their protein sequence revealed that they are highly similar in both protein (98%) and mRNA (93%) sequences. Their intronic sequences are different suggesting this is not an assembly error but a possible gene duplication. As these transcripts share only the last half (corresponding to the last 177 aa) of Sm32, Smp_179170.1 seems to be a truncated version of Sm32. However, it encodes the protein domain characteristic of these endopeptidases. Because of their high similarity, it is not possible to say which of the transcripts is actually more highly expressed in the MT schistosomula. However, it is possible to say that mRNA encoding this particular endopeptidase domain is differentially expressed (whether it comes from the expression of Smp_179170 or Sm32) indicating that, provided the protein in synthesised, this asparaginyl endopeptidase function is present in the parasites. Sm32 has been localized to the acetabular glands of the cercariae, the gut lumen and gut epithelium of the adult worm using northern and western blotting, suggesting a role of both nutrition and defense/invasion for this endopeptidase (el Meanawy *et al.*, 1990). Given the high sequence similarity of Smp_179170.1, it is speculated it could carry out similar functions.

Because proteases are important in the process of invasion, one of the hypotheses is that these proteases are remnants of those used by the cercariae during the skin penetration process. The pattern of expression for both transcripts indicates that they are developmentally up regulated after the transformation in schistosomula (**Figure 4.7**) discarding the former hypothesis. Smp_002150.1, the putative trypsin protease, shows an impressive 30 times fold increase (linear scale) between 3-hours and 24-hours old parasites; while the asparaginyl endopeptidase Smp_179170.1 is up regulated almost 4 times between 3-hours old schistosomula and cercariae.

As for the ferritins, no hypothesis can be generated as this point that would explain why these proteases are more highly expressed in the MT parasites.

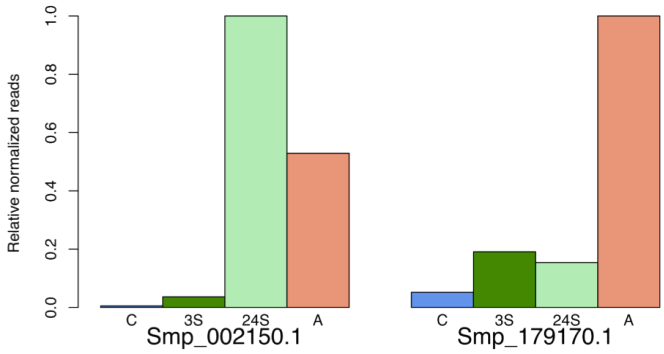


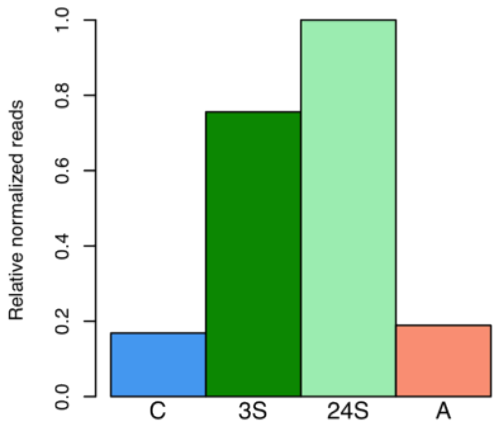
Figure 4.7 – Differential expression of two proteases with higher expression in ST parasites.

4.2.2.2.3 Protease inhibitors

It is interesting to find that protease inhibitors are also differentially expressed alongside proteases. Protease inhibitors can neutralise the action of host- and/or parasite-derived proteases. There are two functional classes of protease inhibitors: the "active-site" inhibitors which bind to the active site of the protease with such affinity that it inactivates their activity, and the alpha-macroglobulins, which modulate the protease activity through an entrapment mechanism [reviewed in (Armstrong, 2006)]. Macroglobulins are also capable of inhibiting coagulation by the inhibition of thrombin (de Boer *et al.*, 1993). ST schistosomula show higher expression of two protease inhibitors, one from each type.

Smp_089670.1 encodes a 1,800 amino acids polypeptide with high similarity to an alpha-macroglobulin protein. Closer examination of the protein sequence showed that it has the basic structure of an alpha-macroglobulin protein: it encodes a signal peptide and

5 functional domains: 2 macroglobulin domains, a thiol-ester bond-forming region with the sequence GCGEQNM strictly conserved among alpha-macroglobulins (Armstrong, 2006), an A-macroglobulin complement component and a macroglobulin receptor binding site. An interesting difference between the *S. mansoni* alpha-macroglobulin and its homolog proteins in other organisms is the gene organisation: in *S. mansoni* this gene is encoded in only two exons while in rat and in *C. elegans* this gene is encoded in 36 and 17 exons respectively.



comp	log2FC	adj–p.val
3S/C	2.1648	1.085E–04
24S/3S	0.4037	7.260E–01
A/24S	–2.4025	1.187E–02
24S/C	2.5686	1.342E–06

Figure 4.8 – The alpha-macroglobulin transcript Smp_089670.1 is highly expressed in the schistosomula stage (mechanically transformed parasires only). The barplot (top) shows the relative expression in each life cycle stage. C, cercariae; 3S, 3 hours old schistosomula; 24S, 24 hours old schistosomula; A, mix sex adult. The table (bottom) shows Log_2 fold changes for each of the comparisons and the corresponding adjusted p-values.

The mechanism of action of macroglobulins is complex but its explanation is necessary to understand why this protease inhibitor may have a *regulatory* rather than inhibitory effect. Macroglobulins function as a cage for proteases: proteases cleave the exposed "bait site" and this produces a rapid conformational change. This has two consequences; first the conformational change is such that the protease is now trapped inside the cage formed by the macroglobulin. The entrapment of the protease regulates its activity in the sense that the protease is still active but only on substrates that are small enough to enter the cage and interact with the protease. Second, the conformational change causes the macroglobulin receptor-binding site to be now exposed; this guarantees that the "used" macroglobulins are taken out from circulation [reviewed in (Armstrong, 2006)].

It is worth mentioning that Parker-Manuel *et al.,* (2011) found this transcript up regulated in the 3-days old schistosomula compared to cercariae and Hall *et al.,* (2011) found this protein in proteome analysis of the gut contents of adult worms. RNA-seq data presented in this thesis indicate that Smp_089670.1 has its peak of expression in the 24-hour old MT schistosomula where it is up regulated compared to the cercariae stage (**Figure 4.8**).

Another inhibitor of protease activity is Smp_147730.1. This is a kunitz-type inhibitor and it is just one of a larger family of kunitz domain containing proteins. These are known for their capability to inhibit serine proteases and are thought to have a role in the host parasite-interaction in the cestode *E. granulosus* (Gonzalez *et al.*, 2009). Smp_147730.1 fulfils the criteria that define a kunitz inhibitor protein: it is small (~150 amino acids) and contains both a signal peptide and a kunitz inhibitor domain with six conserved cysteines that form disulfide bonds. Kunitz-inhibitors act by competing for the active site of serine proteases. Although it has expression below background (RPKM < 2) in cercariae, it shows rapid up regulation as early as 3 hours after transformation. At 24 hours after transformation, its expression reaches 55-fold (linear scale) compared to cercariae. The early up regulation at the onset of infection suggests an important role of the kunitz inhibitor in the development of the schistosomula. Similar increments are observed for the ST sample compared to cercariae but with slightly less magnitude: the fold change between MT over ST being 2.3 fold. Interestingly, serine protease inhibitors (also called serpins), as well as the previously described alpha-macroglobulin, have been identified as up regulated in 3 days old parasites (Parker-Manuel *et al.*, 2011) and the corresponding protein has been found present in the contents of the gut (Hall *et al.*, 2011). However, both

these records correspond to another serine protease (Smp_090080) and not a kuntiz-type inhibitor.

Results presented here suggest that some of the genes more highly expressed in the MT parasites may have a role in feeding or tissue invasion process. Ferritins, proteases and protease inhibitors are thought to have a role in uptake of nutrients or perhaps a role in host tissue invasion. These results complement those reported by Parker-Manuel *et al.,* (2011) - where the authors described up regulation of the same transcripts in 3-days old schistosomula – and agrees with functional studies where schistosomula were seen to actively ingest plasma proteins only 24-hours after transformation (Bennett *et al.*, 1991).

The reason why these transcripts are differentially more highly expressed in the MT compared to the ST parasites is not clear. One of the possibilities could be that the MT population has more parasites in synch with each other than the ST population. Therefore, all the transcripts that show rapid up regulation in the early schistosomula stage would appear as "not so expressed" in the ST parasites. Generating a time course experiment that would reflect several time points in the development o these parasites my help answer this question. Independently of whether digestive-related transcripts are differentially expressed between "types" of parasites, it is important to note that they are expressed in very early stages of the schistosomula transformation; and that their expression is independent from the presence of a skin barrier. Further research into the localization of this proteins in the schistosomula, for example by using WISH (whole *in situ* hybridization), would shed more light on the functions of these genes.

4.2.2.3 Microexon genes are expressed at 24 hours post-transformation

Study of the differentially expressed transcripts presented in **Table 4.1** showed six microexon genes appearing as differentially expressed in both the MT and ST parasites.

Microexons are small exons of < 36 bp that are found in genes present in many eukaryotes (Volfovsky *et al.*, 2003). Typically, microexons contribute a small proportion of the coding sequence of genes but in *S. mansoni* microexons are found as mayor components (~75%) of the coding sequence of some genes; these genes are therefore called microexon genes (MEGs). They were first reported in Berriman *et al.*, (2009) with a more detailed description in DeMarco *et al.*, (2010). The most prominent characteristics of these genes are:

- Approximately 75% of the coding region is encoded in microexons.
- Microexons are typically < 36 bp, with bases in multiples of 3.
- MEGs have conventional (> 36 bp) exons and /or UTRs in their 3' and 5' ends.

- They share little or no sequence similarity among them.

Despite the last point, microexon genes are grouped in families based on the criterion that members of a family should have a BLASTp e-value < $1e^{-4}$ (Berriman *et al.*, 2009). No MEG homologs could be found in other species except for *S. japonicum* and *S. haematobium*.

MEGs show expression of several transcript variants for each locus at a given time. This is achieved by microexon skipping that, because each microexon is a multiple of three, does not change the ORF of the rest of the sequence. This has led to the hypothesis that splice variants change along the life cycle of the parasites providing them with the potential to generate antigenic variation of MEG product (DeMarco *et al.*, 2010).

Six MEGs from three different families appeared as differentially expressed in the comparison of MT *vs.* ST schistosomula at 24 hours after transformation. A summary is presented in **Table 4.4** and **Figure 4.9A**.

Table 4.4 – Summary of MEGs found as differentially expressed in ST and MT.

GeneDB_id	Gene name	Log2-FC	RPKM MT	RPKM ST
Smp_159800.1	MEG-2.2	1.50	4.7	14.2
Smp_159810.1	MEG-2.4	1.33	20.8	59.2
Smp_180340.1	MEG-2.?	-28.93	3.27	< 2
Smp_124000.1	MEG-14	-0.72	2275.5	1396.1
Smp_163630.1	MEG-4	0.93	120.5	197.8
Smp_138080.1	MEG-3	0.96	245.8	161.8

Smp_124000 is a member of the MEG-14 family (DeMarco *et al.*, 2010) where Smp_124000.1 is the only annotated splice variant in GeneDB. Although TopHat reads alignment for this model broadly agrees with the current annotation (**Figure 4.9B**), the profile of microexons expressed in this variant is different from the one present in the database, suggesting this is a different variant compared to that previously identified. This new splice variant shows three extra microexons (black arrows) and lacks the expression of others (not shown). Interestingly, the transcript expressed in the ST sample differs from that expressed in the MT sample. There are three exons clearly expressed in the ST but

absent in the MT schistosomula sample (**Figure 4.10**). Analysis of the transcript expression level showed that this is high in the schistosomula stage and relatively low in cercariae, the same pattern as previously reported (DeMarco *et al.*, 2010). The sampling of close time points in this experiment suggests that this MEG-14 variant is up-regulated as early as 3 hours after transformation (**Figure 4.9A**). Its expression peaks, at least within the samples here studied, at 24 hours post transformation and it is more highly expressed in the MT compared to the ST. Adult worms present low levels of expression of this gene. From the group of MEGs found to be differentially expressed in MT and ST, this is the only one with higher expression in MT.

Smp_163630.1 and Smp_138080.1 are members of the MEG-4 and MEG-3 families respectively. They both have higher expression in the ST samples with similar fold changes (nearly twice as much expression in the ST compared to MT). As in the previous example, close inspection of the reads alignment suggests that although the reads broadly support the current gene models, some differences are noted and these suggest the presence of a different variant from the one annotated in GeneDB (GeneDB, 2011).

The current annotation for Smp_163630.1 shows 16 exons, of which 14 (all except the 3'- and 5' most exons) are microexons. RNA-seq data showed expression of eight exons of the current annotated model for Smp_163630 plus two new exons. This new variant has its peak of expression at 3 hours post transformation, representing another example of an early up regulated transcript. Its expression is down regulated at 24 hours post transformation and goes up again in adult worms to a level comparable to that seen at 3 hours (**Figure 4.9A**).

Smp_138080.1 has 20 exons of which only 15 are expressed in this life cycle stage; RNA-seq data suggest that the 5 exons at the 5'end of the models are not expressed. This member of the MEG-3 family is up regulated in 3-hours old schistosomula [(DeMarco *et al.*, 2010) and **Figure 4.9A**] and has its maximum expression (within the samples here analysed) at 24 hours after transformation.

The other three differentially expressed transcripts are members of the MEG-2 family. Close inspection of the reads alignment showed that the coverage for these genes neither support the current gene models nor identify a new variant. The coverage plot and read alignment for Smp_159800.1 is illustrated in **Figure 4.10** as an example; Smp_159810.1 and Smp_180340.1 show the same pattern of read coverage (not shown). In order to investigate whether expression of the members of the MEG-2 family could be representing non-coding RNA species, the intronic regions with RNA-seq reads coverage were used to

query the Rfam database[1] (Gardner *et al.*, 2011) but no homology with known non-coding RNAs could be found.

According to the data presented here, only three (Smp_124000.1, Smp_163630.1 and Smp_138080.1) out of six MEGs identified as differentially expressed between MT and ST parasites show read alignments that support the annotated gene model. Data suggest that in two out of these three cases, the variants expressed are novel splice forms. What is more, in the case of Smp_124000, the variant expressed in the ST sample has three extra exons compared to the one expressed in the MT sample suggesting that the differential treatment received by these two populations of parasites may have an effect in the regulation of the expression of different isoforms.

4.2.2.4 Differentially expressed transcripts likely to be false positives

The functions represented by these transcripts vary greatly. In the case of genes more highly expressed in the ST parasites, transcripts encoding proteins related to retrotransposon functions, betaine synthesis and muscle development were identified. Other potentially interesting transcripts are those encoding an innexin, a betagalactosyltranserase and an ornithine decarboxylase.

In the genes found more highly expressed in the MT parasites, transcripts encoding proteins with DM domains (domain involved in sexual differentiation in *D. melanogaster*), a myelin transcription factor and a protein involved in arginine biosynthesis.

The gene structures, RNA-seq read coverage and similarity searches were performed for all the transcripts mentioned above. With exception of those encoding retrotransposon elements, all the rest of the transcripts showed either a model that could not be confirmed with RNA-seq data (usually because the reads coverage was very low) or they had no similarity with any other protein that could provide any information about their function. A summary of these transcripts and the reasons why they are not regarded as biological relevant is presented in **Table 4.5**.

[1] The Rfam database Gardner, P. P., J. Daub, J. Tate, B. L. Moore, I. H. Osuch, S. Griffiths-Jones, . . . A. Bateman (2011). "Rfam: Wikipedia, clans and the "decimal" release." Nucleic Acids Res **39**(Database issue): D141-145. holds a catalogue of known non-coding RNA motifs.

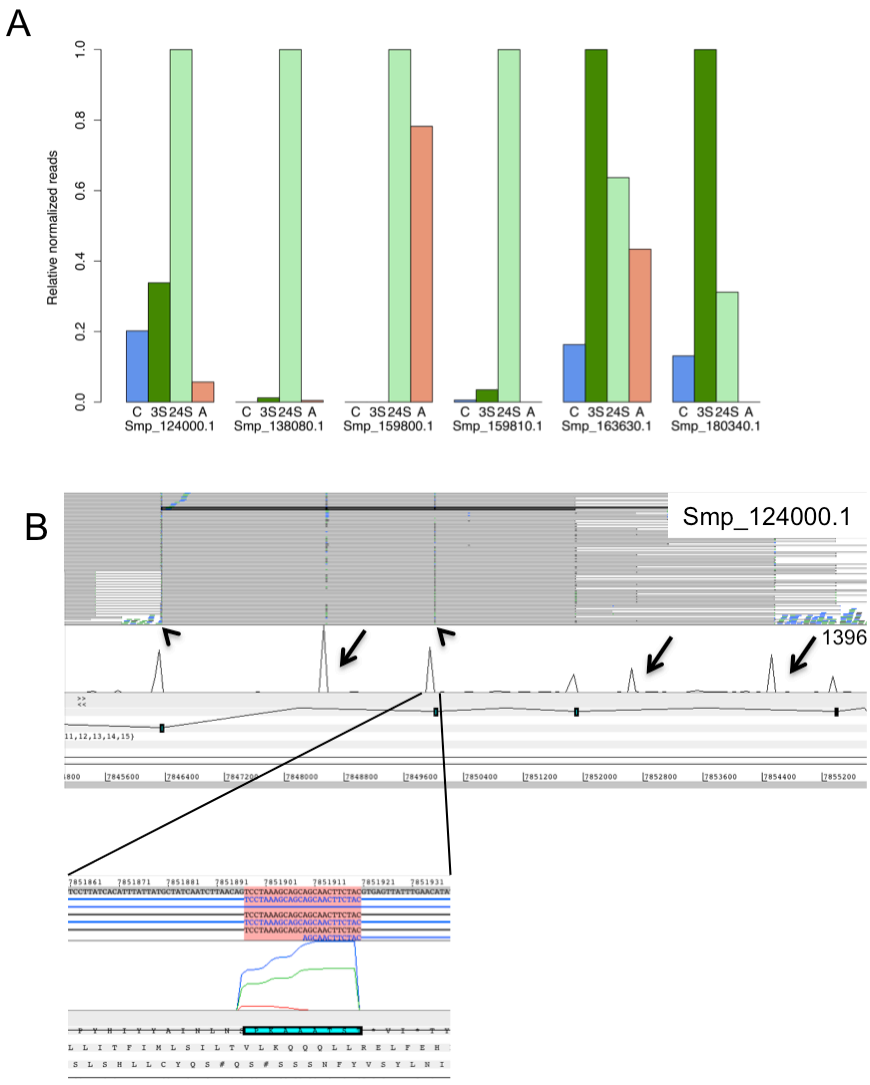


Figure 4.9 – Some microexon genes are differentially expressed in ST and MT parasites. A - Relative expression of microexon genes found differentially expressed in the MT *vs.* ST comparison; the bar plots show their relative expression across the life cycle time points surveyed in this thesis. C: cercariae, 3S: 3-hour old schistosomula, 24S: 24-hour old schistosomula (MT), A: 7-week old mixed sex adult worms. B - Artemis view of a region of the microexon gene Smp_124000. Top panel shows blue and green lines representing reads and grey lines representing split/paired reads (alignment of reads performed using TopHat); middle panel shows a plot representing the coverage of reads (maximum value of 1,512 reads). Arrows indicate the presence of novel exons expressed in this variant. Note how split reads are found mapped to micro exons (arrowheads). In some cases the coverage plot appears to be offset with the annotated exon; closer inspection of the alignment (zoom-in inset) shows that the offset is just an artefact of the visualization tool.

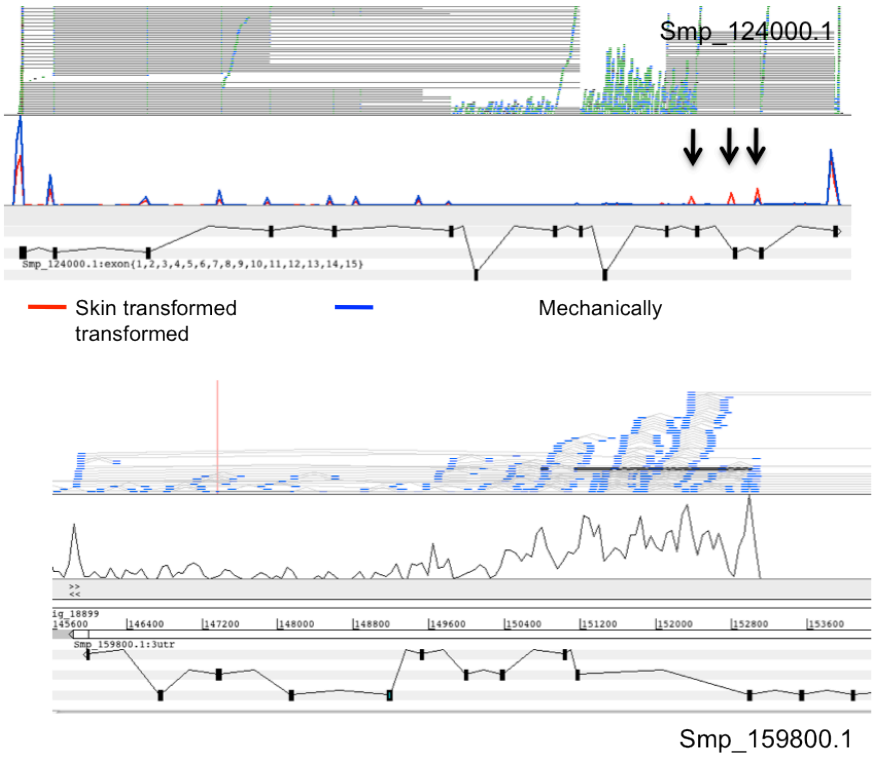


Figure 4.10 - Some microexon genes are differentially expressed in ST and MT parasites. Top panel: Artemis view of a region of the microexon gene Smp_124000 where the ST sample shows expression of three exons that are not present in the MT sample (arrows). Bottom panel: Artemis view of a region of the microexon gene Smp_159800. Top panel shows blue lines representing reads and grey lines representing split or paired reads. The middle panel shows a graphical representation of the coverage of reads. Note how there is no clear peaks of increased expression for the exons of this microexon gene compared to introns.

Table 4.5 –Some transcripts found differentially expressed between ST and MT samples have inaccurate functional annotation and are therefore not considered for the analysis. A summary of these is presented in this table.

Transcripts with higher expression in ST schistosomula		
GeneBD_ID	Function	Description
Smp_142120.1	Muscle development	DNA binding domain but no significant matches in protein database – function cannot be inferred.
Smp_212180.1	Glycine betaine biosynthesis	Encodes a choline dehydrogenase without any simililarities in protein databases except for a hit in *Clonorchis sinensis*.
Smp_007950.1	Glycosylation	Read coverage does not support the gene model – new gene model cannot be inferred
Smp_146940.1	Innexin	Gene model is truncated - encodes a trans-membrane domain of the four that are characteristic of innexin proteins
Smp_067800.1	Ornithine decarboxylase	No significant similarities; possible wrong functional annotation

Transcripts with higher expression in MT schistosomula		
GeneBD_ID	Function	Description
Smp_143190.1	Sex determination	Encodes DM domain (sex differentiation domain in *D. melanogaster*); read coverage does not support the gene model, new gene model cannot be inferred
Smp_168610.1	Myelin transcription factor	Read coverage does not support the gene model, new gene model cannot be inferred

4.2.2.5 Transcripts of unknown function

From the total of 142 differentially expressed transcripts (adjusted p-value < 0.05) between ST and MT parasites, many of them have no associated function and therefore are described as "hypothetical proteins". The percentage of "hypothetical proteins" found within the differentially expressed transcripts (44%) is similar to that found for the rest of the transcriptome.

As presented in Chapter 3, the newer version of the genome annotation contains 504 entirely new genes derived from RNA-seq evidence. These gene models represent 5% of the total gene complement and most of them have no function associated with them. Within the significant differentially expressed genes found in the comparison of ST and MT it was found that 17% of them (26 in total) are new RNA-seq derived predictions. Of these, only two of them have a significant match in the InterPro database; both these transcripts have their peak of expression at 24 hours post transformation and they are approximately 2 fold over expressed in the ST compared to the MT parasites. Smp_202120.1 found in Chromosome 1 has a conserved homebox domain with a putative transcription factor function. Smp_200450.1 is found in Chromosome 3 and has a conserved galactosyltransferase domain suggesting it may have a role in the biosynthesis of polysaccharides. Another related enzyme, a beta-1,4-galactosyltransferase, is also more highly expressed in the ST schistosomula. Because reorganization of the outer membrane is very important in the early stages of the schistosomula (see Chapter 1 sections 1.2.3.1.3), it is possible that the expression of these genes is related to a higher activity in the generation of polysaccharides for the schistosomula coating.

4.3 Discussion

The process of cercarial invasion and schistosomula migration, especially during the early stages of infection, are relevant in the study of intervention strategies against *Schistosoma* parasites. As introduced in Chapter 1, the skin schistosomula stage is regarded as a vulnerable stage for parasite killing (Wilson *et al.*, 2009). Understanding the biology of the parasites at this stage would improve the chances of finding the appropriate targets for intervention.

Many high-throughput gene expression studies have been performed in which different schistosomula stages were analysed (Chai *et al.*, 2006; Dillon *et al.*, 2006; Fitzpatrick *et al.*, 2009; Gobert *et al.*, 2010; Parker-Manuel *et al.*, 2011). With one

exception (Chai *et al.*, 2006), all studies used mechanically transformed parasites. Different schistosomula stages (i.e., skin, lung, etc) were obtained by varying the time parasites were incubated after transformation or after recovery from the host. Chai *et* al., (2006) showed that indeed lung stage (3-days post infection) obtained from infected animals and mechanically transformed parasites have different profiles of gene expression. These findings prompted the question of whether skin transformed and mechanically transformed schistosomula in the skin stage (24-hours post invasion) would also present differential expression of genes.

In order to address this question, transcriptome samples from both mechanically transformed and skin transformed schistosomula were sequenced and transcriptomes quantitatively compared.

The experimental approach presented in this thesis differs from that of Chai *et al.,* (2006) in many aspects. First, the work of Chai *et al.,* (2006) was performed in another species. The authors used *S. japonicum*, where significant differences in the speed of migration through the skin have been reported between this species and *S. mansoni* (Wang *et al.*, 2005). Second, the authors focus on 3-days old schistosomula residing in the lung tissue. Many publications have feature the differences between the skin- and lung-stages of the parasite (Crabtree *et al.*, 1980; Wilson *et al.*, 1980; Crabtree *et al.*, 1985; Crabtree *et al.*, 1986) and therefore these cannot be easily compared. Thirdly, the authors used a microarray platform where oligonucleotides sequences were generated from are *S. japonicum* and *S. mansoni* EST databases.

I the work here presented, the first challenge was to obtain the biological material. As shown before, (see section 4.2.1.1), mechanical transformed preparations of schistosomula can present a significant number of damaged parasites or contaminating tails. Therefore, optimization of the protocol was required to obtain healthy non-contaminated populations of transformed schistosomula. This was achieved by replacing part of the harsh needle-passage step by vigorous shaking of parasites in a vortex mixer. At the same time, optimization of the percentage of Percoll® used in the separation of cercarial heads and tails yielded a preparation virtually free from tails and cercariae. These modifications to the protocol resulted in fewer number of damaged parasites and less contamination of tails in the schistosomula preparation. The skin transformation protocol was also subjected to optimization; yet the schistosomula obtained from these experiments were usually contaminated with tails or cercariae and assessment of individual preparations was performed in order to obtain tail-free and cercariae-free schistosomula preparations.

The skin transformation protocol presents some disadvantages: 1) it is laborious and time consuming, taking approximately 9 hours to take the experiment to completion compared to the mechanical transformation (~ 3 hours); 2) it is possible that skin-transformed parasites are slightly off synchronization because transformation can occur at any time within a 3-hour window. On the other hand, skin transformation provides a more natural way of transforming the cercariae into schistosomula. Compared to the more widely used mechanical transformation, the skin transformation protocol allows the parasite to penetrate a layer of skin; which mimics the natural barrier present in the human host. The advantages of the mechanical transformation are evident: it is a quick protocol that allows researchers to obtain typically hundred of thousands of schistosomula per experiment. Once the protocol is optimized, parasites are generally healthy and phenotypically closely resemble the skin-transformed parasites even though they have been subjected to artificially imposed environmental variables such as low temperatures, centrifugation forces, shaking and squeezing through a syringe needle. In spite of being very different, both transformation protocols yielded healthy schistosomula that were capable of surviving *in vitro* for several days.

Although many studies have focused on the differences and similarities between these two types of schistosomula, none before had focused on the comparison of their transcriptome profiles. As mentioned earlier, the work of Chai *et al.*, (2006) is the closest to the experiment presented in this chapter. In their work, the authors found that IVS[1] parasites express transcripts encoding protaglandins, glutathione-S-transferase Sm28GST, paramyosin, stress related proteins and transcripts related to markers of anti-inflammatory and immunomodulatory processes. In the case of MTS, the authors show that these schistosomula show higher expression of transcripts involved in glucose transport, and fatty acids transport and haemoglobin digestion (Chai *et al.*, 2006).

The RNA-seq differential expression analysis of MT *vs.* ST parasites performed in this thesis showed that 149 transcripts are differentially expressed (adjusted p-value < 0.05). Transcripts encoded in the mitochondrial genome (mitochondrial genes) were found among the most differentially expressed transcripts being all of them more highly

[1] IVS refers to schistosomula obtained from lungs from infected mice three days post infection; MTS refers to mechanically transformed schistosomula followed by *in vitro* cultivation for three days Chai, M., D. P. McManus, R. McInnes, L. Moertel, M. Tran, A. Loukas, . . . G. N. Gobert (2006). "Transcriptome profiling of lung schistosomula,in vitro cultured schistosomula and adult Schistosoma japonicum." Cell Mol Life Sci **63**(7-8): 919-929..

expressed in the ST parasites. This over expression of mitochondrial transcripts was correlated with higher metabolic rate in the ST parasites compared to the MT measured using the AlamarBlue® essay (section 4.2.2.1.1) suggesting that skin transformed parasites could be regarded as more metabolically active than their mechanical counterparts. The reasons behind the increased abundance of mitochondrial transcripts could be many. One possibility could be that the mitochondrial genes are subjected to higher transcription rate in the ST parasites. Another possibility could be that the ST parasites simply have more functional mitochondria than the MT parasites. Counting the mitochondria in both schistosomula preparations using a mitochondrial marker could test this hypothesis. Contraction and extension movements generated by the musculature structure demands important quantities of ATP and the increased metabolic activity could be a reflection of such need. Unfortunately, it was not possible to quantitatively evaluate the rate of movement of MT and ST parasites, data from such experiments would provide more evidence to back up this hypothesis.

In order to study which other processes were represented among transcripts more highly expressed in ST parasites, Gene Ontology enrichment was performed. This analysis showed a potpourri of transcripts associated to different functions such as GPCR signalling, microtubule-related movement, retrotransposon related functions, betaine synthesis and muscle development. In the first two functional categories, the transcripts products description correlated well with the associated GO terms, providing a back up of the functional annotation of such genes. Additionally, their up-regulation in the ST schistosomula could be associated with processes likely to be taking place in these parasites. The rest of the categories were discarded based on the lack of reliability of either the gene models or their annotation. In the case of retrotransposon related function it was not possible to associate these transcripts to any known biological process in the schistosomula. Their function in this or any other stages of development remains unknown.

Mechanically transformed parasites showed higher expression of ferritins, proteases and protease inhibitors. Ferritins (Hall *et al.*, 2011) and certain proteases [reviewed in (Caffrey *et al.*, 2004)] are thought to be involved in uptake of nutrients from the host. At the same time, proteases have been linked to host tissue invasion (Salter *et al.*, 2000; Hansell *et al.*, 2008). Chai *et al.*, (2006) reported the differential expression of transcripts involved in nutrients uptake. However, these transcripts are not the same as those reported in this chapter. Nevertheless, it is interesting to notice that results presented here agree with those from Chai *et al.*, (2006) in that proteases are more highly expressed

in mechanically transformed parasites and the authors attribute this to culturing conditions. This explanation cannot be extrapolated to the results presented in this thesis regarding the differential expression of ferritins and proteases at 24 hours post transformation because both types of parasites were exposed to the same culturing conditions.

Independently from their higher expression in MT or ST parasites, it is noteworthy that all these transcripts are being up-regulated in the early schistosomula and their higher expression in MT parasites may reflect the higher synchronization in their development compared to the ST parasites. The expression of these transcripts in such an early stage of development is very interesting because of their possible implications in the survival of the parasite.

Some members of the microexon family 2 (MEG-2) were found over expressed in the ST parasites. What is more, the variants found here are different from the ones previously reported confirming that different splice variants from the same loci are expressed at different time points. Interestingly, one of the variants expressed in the ST parasites has a different exon profile than in the MT transcript. This is an important finding because it suggests that different cues from the environment might be triggering the differential expression of certain exons in the microexon gene repertoire. The function of microexon genes remains to be unknown.

Both populations of parasites showed expression of "false positives" that could be discarded based on the crosscheck of the putative functions associated to the products. Future upgrades in the annotation of the genome promise the opportunity to further improve the interpretation of these results.

CHAPTER 5

TIME COURSE ANALYSIS OF DIFFERNTIAL EXPRESSION OF GENES: CERCARIAE, 3 HOUR OLD & 24 HOUR OLD SCHISTOSOMULA

5.1 Introduction

During the process of infection of the human host, *S. mansoni* parasites change from the cercariae free-living stage into a parasitic life form, the schistosomula. In natural conditions, this transition is triggered by the penetration of the skin barrier upon which the cercarial head invades the host while the cercarial tail is lost. This situation can be reproduced in the laboratory by the application of shear pressure as described in Chapter 2 section 2.2.2. Detailed descriptions of the changes that accompany these transformations were introduced in Chapter 1 sections 1.2.3.1 to 1.2.3.2. It is also known that this transformation is not dependent on the presence of the skin barrier. A brief recount of these changes is summarised here.

Once in the skin, the parasites loose their glycocalyx, the outer membrane changes from a single to a double bilayer, the contents of the acetabular glands are gradually emptied to assist the penetration process across the outermost layers (Holmfeldt *et al.*, 2007) and the deeper layers of skin (Wilson *et al.*, 1980).

Many groups have attempted to describe the gene expression changes during the transition from the cercariae to schistosomula. So far, approaches used to this end involved quantitative ESTs (Farias *et al.*, 2011) and high-throughput studies using microarrays (Chai *et al.*, 2006; Dillon *et al.*, 2006; Fitzpatrick *et al.*, 2009; Gobert *et al.*, 2010; Parker-Manuel *et al.*, 2011). The work of Fitzpatrick *et al.,* (2009) spanned 15 different life cycle time points including cercariae, 3-hour old and 24-hour old schistosomula; the same time points considered in this thesis. With a p-value cut off (non-adjusted) of < 0.05 the authors showed that 159 genes were differentially expressed at 3 hours after transformation (114 up regulated and 45 down regulated) and 321 genes at 24 hours after transformation (202 up regulated and 119 down regulated) compared to cercariae. However, after correcting the p-values for multiple testing [adjusted p-value (Benjamini *et al.*, 2001)] the number of differentially expressed genes was reduced to zero. These results are surprising considering cercariae and schistosomula are very structurally different stages of the parasite.

In a more recent microarray study performed by Gobert *et al.,* (2010), in which the cercariae *vs.* 3-hour old schistosomula comparison was also considered, the authors identified 2,791 differentially expressed genes (1,608 up regulated and 1,183 down regulated) between these two life cycle stages. The up-regulated dataset included genes related to the structure of the tegument (i.e. tetraspanins and calcium-binding proteins),

stress response proteins (i.e. HSP70), development (i.e. frizzle related protein) and enzymes involved in the blood meal digestion (i.e., range of cathepsins) among others (Gobert *et al.*, 2010). Unfortunately, no test statistic was applied to the analysis of fold changes in this report making the interpretation and comparison with other studies a difficult task.

The main motivation to re visit the question of which genes are differentially expressed during the cercariae to schistosomula transformation lies in the application of novel techniques (RNA-seq) and statistical approach [edgeR, (Robinson *et al.*, 2010)] that would allow a better resolution and improved understanding of the genes that have a role in shaping the adaptation of the parasite to the new environment. As exposed before, the microarrays studies that tried to tackle this question had some limitations that the RNA-seq approach can overcome. In the particular case of *Schistosoma* microarrays, these have relied on the existence of previously identified genes and genome sequences to generate the probes that form the array [i.e., (Fitzpatrick *et al.*, 2005)]. On the contrary, the generation of RNA-seq data is independent from previous sequence knowledge and offers the possibility of exploring coding sequences previously unknown. What is more, and as it was shown in Chapter 3 section 3.2.3, the dynamic range of RNA-seq surpasses that of microarrays providing a better tool for measuring very high and very low levels of expression with improved resolution.

This chapter present the transcriptome changes of the cercariae and the early stages of mechanically transformed schistosomula (3-hours and 24-hours old parasites). Briefly, transcriptome sampling of the cercariae, 3-hours old and 24-hours old schistosomula were taken and sequenced using the Illumina second-generation sequencing technology. The first part of this chapter offers an overview of the time course analysis including four time points in the parasites' life cycle: cercariae, 3-hours old and 24-hours old schistosomula and 7-week old mix sex adult worms. The latter sample was used as a reference to extrapolate the expression of genes found differentially expressed in the schistosomula stage. This is of particular importance because it helps to recognise genes that are expressed after transformation but are no longer needed in the parasites' adult life. The second part of this chapter studies the different biological processes found to be up regulated or down regulated in the schistosomula compared to the cercaria stage through the analysis of differentially expressed genes. The level of resolution achieved by these data has no precedent in any other parasitic worm.

5.2 Results – time course analysis of transcriptome changes.

The first step in the time course analysis was to identify transcripts that are differentially expressed between different life cycle time points. To this end, the edgeR package (Robinson *et al.*, 2010) was used (see Chapter 2 section 2.6.5). Briefly, edgeR takes the number of reads per transcript as input and uses the biological replicates to estimate the dispersion of the samples. The calculated dispersion serves to normalise the libraries and make them comparable to each other. A normalised list of reads per gene is generated and from it, differential expression for each gene/entry can be calculated together with a significance value (p-value).

In preparation for the edgeR analysis, all non-expressed genes (as defined in Chapter 3 section 3.2.5) were removed from the dataset leaving a total of 9,096 transcripts for which differential expression could be calculated. Four time points in the parasite's life cycle were analysed (**Table 5.1** and **Figure 5.1**).

- Cercariae vs. 3-hours old schistosomula
- 3-hours old schistosomula vs. 24-hours old schistosomula
- 24-hours old schistosomula vs. adult
- Cercariae vs. 24-hours old schistosomula

Table 5.1 - Number of differentially expressed genes (adjusted p-value <0.01).

Stage comparison	Up regulated§	Down regulated§	Total
Cercariae - 3 hour schistosomula	964	522	1,486
3 hour schistosomula - 24 hour schistosomula	470	560	1,030
24 hour schistosomula - adult	1,120	960	2,080
Cercariae - 24 hour schistosomula	1,652	1,198	2,850

§ Fold change threshold is 2-fold

A total of 3,300 non-redundant transcripts (excluding alternative spliced forms) were found differentially expressed (adjusted (Benjamini *et al.*, 2001) p-value < 0.01) within the cercariae *vs.* 3-hours old schistosomula, 3-hours old schistosomula *vs.* 24-hours old schistosomula and 24-hours old schistosomula *vs.* adult comparisons. In this study, more differentially expressed transcripts are found than in the work of Fitzpatrick *et al.,* (2009) but less than in Gobert *et al.,* (2010).

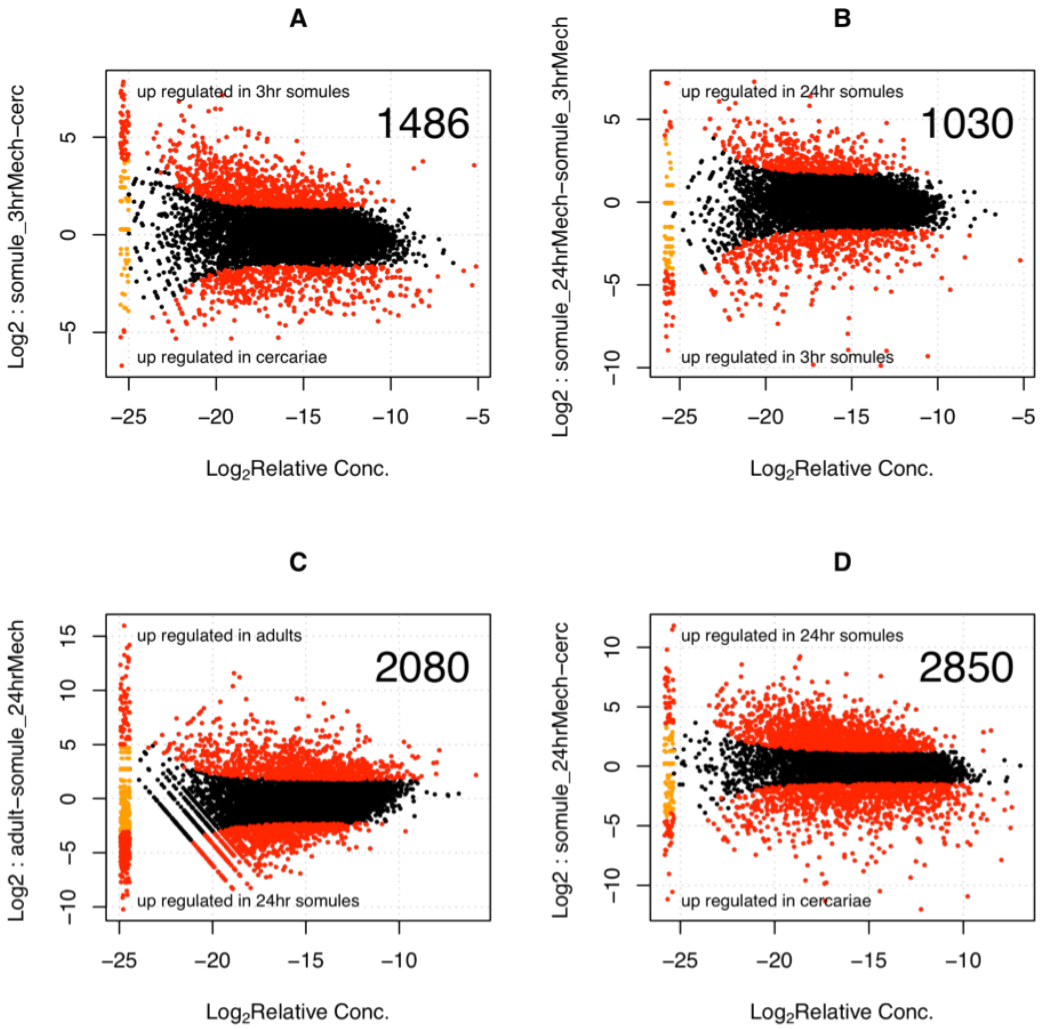


Figure 5.1 - Differential expression of transcripts across four time points in the life cycle of *S. mansoni*. Relative concentration (x axis) is plotted against fold change values (y axis) in the logarithmic scale in base 2. Legend: Red dots, significantly different (adjusted p-value <0.01); black dots, non-significant. Data points grouped in the left of each plot (red and orange) represent transcripts that have reads in one sample but not in the other and therefore the relative concentration and fold change cannot be calculated. Numbers in the top right corner of each plot represents the total number of significant differentially expressed transcripts. A – Differential expression between cercariae and 3-hours old schistosomula. B - Differential expression between 3-hours old and 24-hours old schistosomula. C - Differential expression between 24-hours old schistosomula and mix sex 7-week old adults. D - Differential expression between cercariae and 24-hours old schistosomula.

5.2.1 Genes with no change in expression

Fitzpatrick *et al.,* (2009) identified a set of 355 microarray probes with less than 1% variability in their expression values across 15 life cycle time points (Fitzpatrick *et al.*, 2009). The authors regarded genes represented by these probes as "constitutively expressed genes". It was found that 192 of these probes have an unambiguous match against the transcripts dataset (for Methods see Chapter 2 section 2.6.4) and 75 of them are differentially expressed at least in one of the comparisons. **Figure 5.2** shows an example of the distribution of the 192 transcripts in the cercariae to 3–hour old schistosomula comparison where only 25 transcripts were found differentially expressed. The variability between the results shown in this study and those shown in microarrays for the same genes might be attributed to the higher resolution that can be achieved through RNA-seq measurement. It is worth mentioning that most of the constitutively expressed genes/transcripts are found at relatively high concentration values meaning that they are relatively highly expressed compared to the rest of the transcripts. As shown before in Chapter 3 section 3.2.3, the capacity of microarrays in identifying differential expression in highly expressed genes is much reduced compared to RNA-seq. This limitation of the microarray platform could explain why these genes are found differentially expressed in the RNA-seq approach but not in the microarray.

5.2.2 Validation using known genes

In order to validate the RNA-seq approach in *S. mansoni*, genes with known expression profile were compared to the RNA-seq relative expression values. The control genes were chosen based on differences in their changes of expression shown by northern blotting. These genes are an 8 kDa calcium binding protein (Smp_033000.1), associated with tegument remodelling during cercariae transformation into schistosomula (Ram *et al.*, 1989; Ram *et al.*, 1994); a heat shock protein 70 (HSP70 - Smp_106930.1), active in very early (~3-hours old) schistosomula (Hedstrom *et al.*, 1987; Neumann *et al.*, 1992; Neumann *et al.*, 1993); and the tegument antigen Sm22.6 (Stein *et al.*, 1986) also known as SmTAL1 (for *S. mansoni* tegument allergen-like protein 1 - Smp_045200.1), associated with resistance to re-infection in adult patients of endemic areas (Dunne *et al.*, 1992). RNA-seq results broadly agreed (**Figure 5.3**) with relative gene expression measurements obtained through other approaches. The only case in which the data did not correlate was in the comparison of 24-hours old schistosomula *vs.* adult expression of HSP70 –

Smp_106930.1. The source of this discrepancy is unknown, although it might be related to the rapid control in the regulation of expression of this transcript.

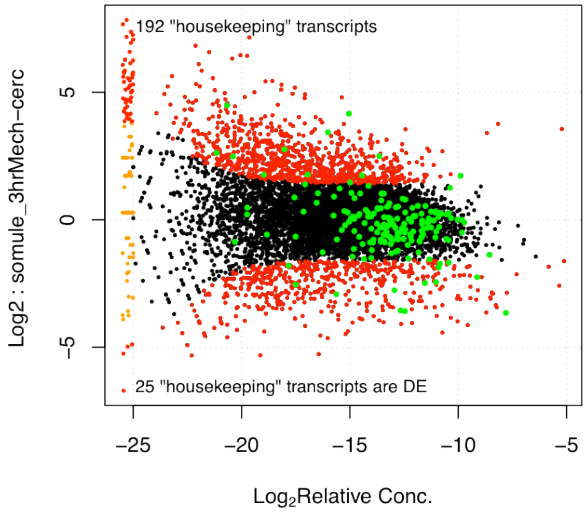


Figure 5.2 – Differential expression of transcripts in the cercariae *vs.* 3-hour old schistosomula. Green dots represent constitutively expressed as reported by Fitzpatrick *et al.,* (2009).

5.2.3 Analysis of differential expression in cercariae, 3 hour and 24 hour old schistosomula

As pointed out in section 5.2, there are ~3,300 non-redundant differentially expressed transcripts (adjusted p-value < 0.01) in the comparisons considered in this study. A breakdown of the biological processes over-represented among the groups of up- or down- regulated genes in the cercariae to 3-hours old and 24-hours old schistosomula transitions is presented.

5.2.3.1 Down regulated transcripts/processes in the schistosomula stage

In order to identify down regulated processes in the passage from the cercariae to the schistosomula stage, GO term enrichment analysis was performed using the dataset of down regulated genes from three pair-wise comparisons: the cercariae *vs.* 3-hour old schistosomula, 3-hour old schistosomula *vs.* 24-hour old schistosomula and cercariae *vs.*

24-hour old schistosomula. Results are shown in **Table 5.2**, **5.3** and **5.4** respectively. With few exceptions, down regulated processes do no differ much between the comparison cercariae *vs.* 3-hour old schistosomula and cercariae *vs.* 24-hour old schistosomula.

During the passage through the host skin, the cercariae lose their tails, and their heads transform into schistosomula. Tails are multi-cellular structures characterised by tissues resembling striated muscle fibres and packed with mitochondria. Tails are anatomically and physiologically different from the cercarial head (see Chapter 1 section 1.2.3.1) and therefore it is expected that they have a very different transcript repertoire. Because of this, transcripts that appear as down regulated in the schistosomula stage compared to the cercariae could arise from two scenarios: either transcripts could be truly down regulated in the schistosomula compared to the cercarial head or they could be exclusively or highly expressed in the cercarial tail and therefore appear as down regulated or are undetectable in the schistosomula stage.

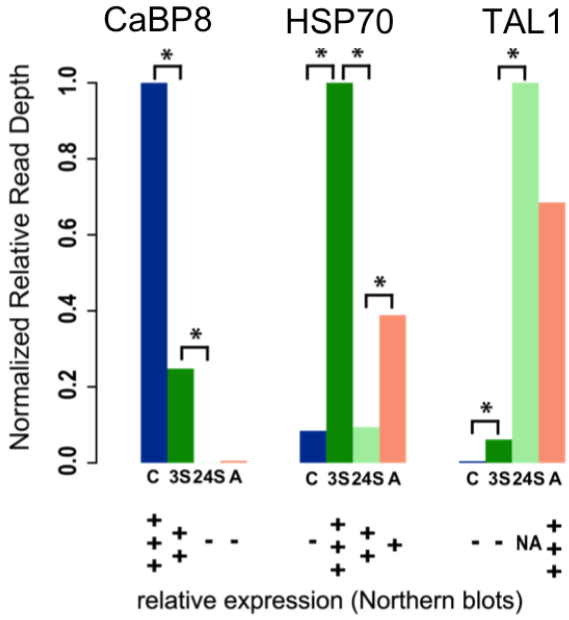


Figure 5.3 - Comparison of expression of genes previously identified to be developmentally regulated. Barplots represent relative normalized reads (from RNA-seq data) for 3 transcripts, asterisks represent comparisons where differential expression is significant (adjusted p-value <0.01). Relative expression reported in the literature (Stein *et al.*, 1986; Ram *et al.*, 1989; Neumann *et al.*, 1992) is shown at the bottom (+++, high expression, ++ medium expression, + some expression, - not expressed, NA, no information available). C = cercariae, 3S = 3-hour old schistosomula, 24S = 24-hour old schistosomula, A = adult.

Because of the cercariae losing the tails, many of the transcripts down regulated in the schistosomula are related to biological processes known to occur mainly in the tail. This was shown by the GO enrichment analysis (see **Table 5.2** and **5.4)**. Higher expression of transcripts involved in aerobic respiration (ATP biosynthesis couple proton transport, TCA, glycolysis and electron transport chain) is detected in cercariae reflecting the metabolic pathway used to generate energy (Skelly *et al.*, 1993).

Consistently with loosing the mitochondria, the expression of solute and protein transporters related to the mitochondrial function are also down regulated. Transcripts included in this "transport" category (see "transport" in **Table 5.2**) encode a range of carrier proteins mostly with trans-membrane domains associated to the traffic of solutes across the mitochondrial membrane and therefore, expected to appear as down regulated once the cercarial tail is lost in the transformation.

As previously mentioned, the cercarial tail has a complex muscle structure formed by longitudinal and circular muscle layers. These types of structures are not found neither in the cercarial head nor the schistosomula. In accordance to this, results show down regulation of three calponin proteins and six transcripts containing collagen (grouped under phosphate transport, which could indicate wrong annotation). A calponin-like protein has been localised to the cercarial tail muscle in *S. japonicum* (Jones *et al.*, 2001) and proteins containing the calponin domain have been shown to bind actin and other components of the smooth muscle contraction apparatus in vertebrates (el-Mezgueldi, 1996). Collagen is a principal component of the connective tissue, also very abundant in the cercarial tail (Dorsey *et al.*, 2002). The differential expression pattern of genes related to the tail musculature is in agreement with loss of the tail.

Observations of the *in vivo* and *in vitro* schistosomula have suggested that there is no cell division until the fourth day after transformation (Clegg *et al.*, 1972). In this context, transcripts involved in DNA replication and cell division where investigated to asses whether the observations from Clegg *et al.,* (1972) were also valid at the transcriptional level. Histone protein expression is a good indicator of cell division. Histone mRNA accumulates during the DNA replication phase of the cell cycle and histone proteins are synthesised during the S phase [reviewed in (Osley, 1991)]. A decrease in the histone mRNA would be a good marker for absence of DNA replication. Consistent with the observations made by Clegg *et al.,* (1972) almost 40 years ago, RNA-seq data show that histone mRNAs (h1/h5, H2A, H2B, H3, H4 and linker histone H1) are down regulated (up to 40 times in linear scale) at 24 hours post transformation compared to the cercariae (**Table 5.5**). It is noteworthy that this effect is not detected when comparison cercariae

with 3-hours old schistosomula, indicating that this is not an effect of loosing the tail. It is possible that histone mRNA is still present in the cercariae and 3-hour old schistosomula as a reminiscent of the high rate of cell division observed in the germ ball (Parker-Manuel *et al.*, 2011); suggesting that the schistosomula shuts down it cell division machinery only after 3 hours post-transformation. Once the parasites has reached day 3 post-transformation the expression of histones mRNA is higher than that recorded for the cercariae stage (Parker-Manuel *et al.*, 2011) suggesting that cell cycle is resumed some time between 24-hours and 3-days old schistosomula.

Table 5.2 – Gene Ontology enrichment (Biological processes) among transcripts down regulated in the cercariae *vs.* 3-hours old schistosomula.

GO term ID	GO term description	p-value
GO:0015986	ATP synthesis coupled proton transport	0
GO:0000226	Microtubule cytoskeleton organization	0
GO:0006099	Tricarboxylic acid cycle	0
GO:0042773	ATP synthesis coupled electron transport	0.002
GO:0006810	Transport	0.003
GO:0006096	Glycolysis	0.003
GO:0022900	Electron transport chain	0.004
GO:0009098	Leucine biosynthetic process	0.004
GO:0007283	Spermatogenesis	0.007
GO:0006094	Gluconeogenesis	0.008
GO:0006108	Malate metabolic process	0.008

Table 5.3 – Gene Ontology enrichment (Biological processes) among transcripts down regulated in the 3-hours old schistosomula *vs.* 24-hours schistosomula.

GO term ID	GO term description	p-value
GO:0006412	Translation	0
GO:0006817	Phosphate transport	0
GO:0006334	Nucleosome assembly	0
GO:0006278	RNA-dependent DNA replication	0.001
GO:0008380	RNA splicing	0.001
GO:0008272	Sulfate transport	0.005
GO:0006542	Glutamine biosynthetic process	0.005
GO:0009097	Isoleucine biosynthetic process	0.005
GO:0051258	Protein polymerization	0.005

Table 5.4 – Gene Ontology enrichment (Biological processes) among transcripts down regulated in the cercariae *vs.* 24-hours old schistosomula.

GO term ID	GO term description	p-value
GO:0006096	Glycolysis	0
GO:0006810	Transport	0.001
GO:0006817	Phosphate transport	0.002
GO:0007283	Spermatogenesis	0.003
GO:0042775	Mitochondrial ATP synthesis coupled electron transport	0.004
GO:0031032	Actomyosin structure organization	0.004
GO:0008380	RNA splicing	0.005
GO:0006334	Nucleosome assembly	0.006
GO:0015986	ATP synthesis coupled proton transport	0.006
GO:0000398	Nuclear mrna splicing, via spliceosome	0.006
GO:0022904	Respiratory electron transport chain	0.008

Other processes that seem to accompany this halt in the cell cycle are also identified as down regulated. A group of five transcripts encoding proteins known to participate in microtubule formation (grouped under the GO biological process of "microtubule cytoskeleton organization") are down regulated in the schistosomula. These proteins contain highly conserved tektin domains, which are involved in formation and stabilization of microtubules and centrioles (Amos, 2008). The down regulation of these transcripts in the schistosomula stage may be a consequence of the cell cycle arrest status. Transcripts related to the spliceosome, protein translation machinery and protein polymerisation are also down regulated (**Table 5.3** and **5.4**).

Because some enzymatic reactions appear in more than one pathway, it is common to find that the GO term analysis show certain processes as enriched while in reality they are just part of down stream steps of other processes. Leucine biosynthesis, malate metabolic process and gluconeogenesis (**Table 5.3**), and glutamine and isoleucine biosynthesis (**Table 5.4**) are shown as down regulated after transformation. In this case, these processes are shown in the GO analysis because the enzymes encoded by these transcripts are also part of the TCA cycle or the glycolysis pathway. Since the functional annotation of these genes is correct (data not shown) it would not be appropriate to consider them false positives.

However, true sources of false positives in GO enrichment analysis are transcripts associated with the wrong annotation. Spermatogenesis is a good example: there are five spermatogenesis-associated transcripts up regulated in this stage. Three of them have no

homology with any informative conserved domains (one of them has homology to one "domain of unknown function"). The other two transcripts encode a calponin domain and a DNA-binding zinc-finger domain respectively. An association between these transcripts and the process of spermatogenesis could not be found.

Table 5.5 – Nucleosome components are down regulated in the schistosomula stage

GeneDB	FC§ 3-hours old schistosomula	FC§ 24-hours old schistosomula	product description
Smp_002930.1	-11.48	-40.55	histone H2A, putative
Smp_003770.1	NS	-16.27	histone h1/h5, putative
Smp_036220.1	-3.71	-5.45	histone H2B, putative;with=UniProt:Q8I1N0
Smp_053390.1	-6.12	-9.55	histone H4, putative
Smp_054230.1	NS	-7.15	hypothetical protein
Smp_074610.1	-6.56	-11.57	histone H3, putative
Smp_082240.1	-9.24	-14.92	histone H3, putative
Smp_162370.1	-12.02	-34.11	Linker histone H1

§ Linear fold change of the gene expression in the schistosomula stage compared to cercariae. NS: Non-significantly.

5.2.3.2 Up regulated transcripts/processes in the schistosomula stage.

As previously pointed out, the establishment of the parasitic life inside the mammalian host starts with the penetration of the cercarial head in the host skin. In the previous section, transcriptional signals lost or down regulated in the schistosomula compared to the cercariae were discussed. In this section the focus is made in up regulated transcripts. As mentioned in the introduction of this chapter, there are only two previous high-throughput microarray studies that looked at the transcriptional changes in this very short time points during the transition. Among the 1,608 up regulated genes reported by Gobert *et al.,* (2010), the authors pointed out some examples of up regulated genes related to the tegument, gut function, stress and development (Gobert *et al.*, 2010). RNA-seq data could detect up regulated transcripts in all of these categories.

The GO term enrichment analysis applied in Gobert *et al.,* (2010) indicated genes enriched in rather general GO categories. The higher resolution of provided by the RNA-seq approach allowed to identify four main biological processes occurring in the schistosomula. These processes group genes involved in signal transduction (GPCRs, Wnt receptor signalling pathway, potassium/ion transport), carbohydrate transport (and monosaccharide transport), "regulation of transcription" and "tissue development"

(homophilic cell adhesion, cell differentiation, cell adhesion, cell-matrix adhesion, integrin-mediated signalling pathway among others). The full list of categories and how they are represented in the three comparisons analysed is presented in **Figure 5.4** and **Appendix C**.

Schistosoma G-protein couple receptors (GPCR) have been proposed as an important candidate group to search for new drug targets (Berriman *et al.*, 2009) mainly because they have been validated as such in a range of other organisms (Overington *et al.*, 2006). GPCRs are significantly enriched among up regulated transcripts in the 3-hour old schistosomula (32 transcripts) and 24-hours old schistosomula (61 transcripts) compared to cercariae. Closer inspection of these genes [performed in collaboration with Dr. Mostafa Zamanian, (Zamanian, 2011)] resulted in their classification into nine categories: Peptide, PROF1[1], Amine, Frizzled, Secretin, Glutamate, Other, "Unlikely GPCR" and "partial sequence". Among these, neuropeptide receptors are significantly over represented; they are discussed in following sections in the context of neuropeptide signalling.

Voltage-gated ion channels are the fourth most drugable targets in the pharmacological industry (Overington *et al.*, 2006). *S. mansoni* encodes a significant number of ion channels of which 81 are annotated as potassium channels (Berriman *et al.*, 2009). It is noteworthy that one third of them are up regulated in the schistosomula stage. Potassium ion channels are expressed in a wide range of tissues (i.e. nervous system, muscle) and their function is mainly to set the electrical resting potential of cells (Alberts *et al.*, 2002). Schistosomula express a number of genes involved in nervous system development. It is possible that the expression of potassium ion channels is related to the development of the parasite's nervous system; which is discussed later in this section

[1] PROF1 stands for "Platyhelminth Rhodopsin Orphan Family 1"; which constitute a divergent flatworm specific family of receptors with some resemblance to the rhodopsin receptor family.

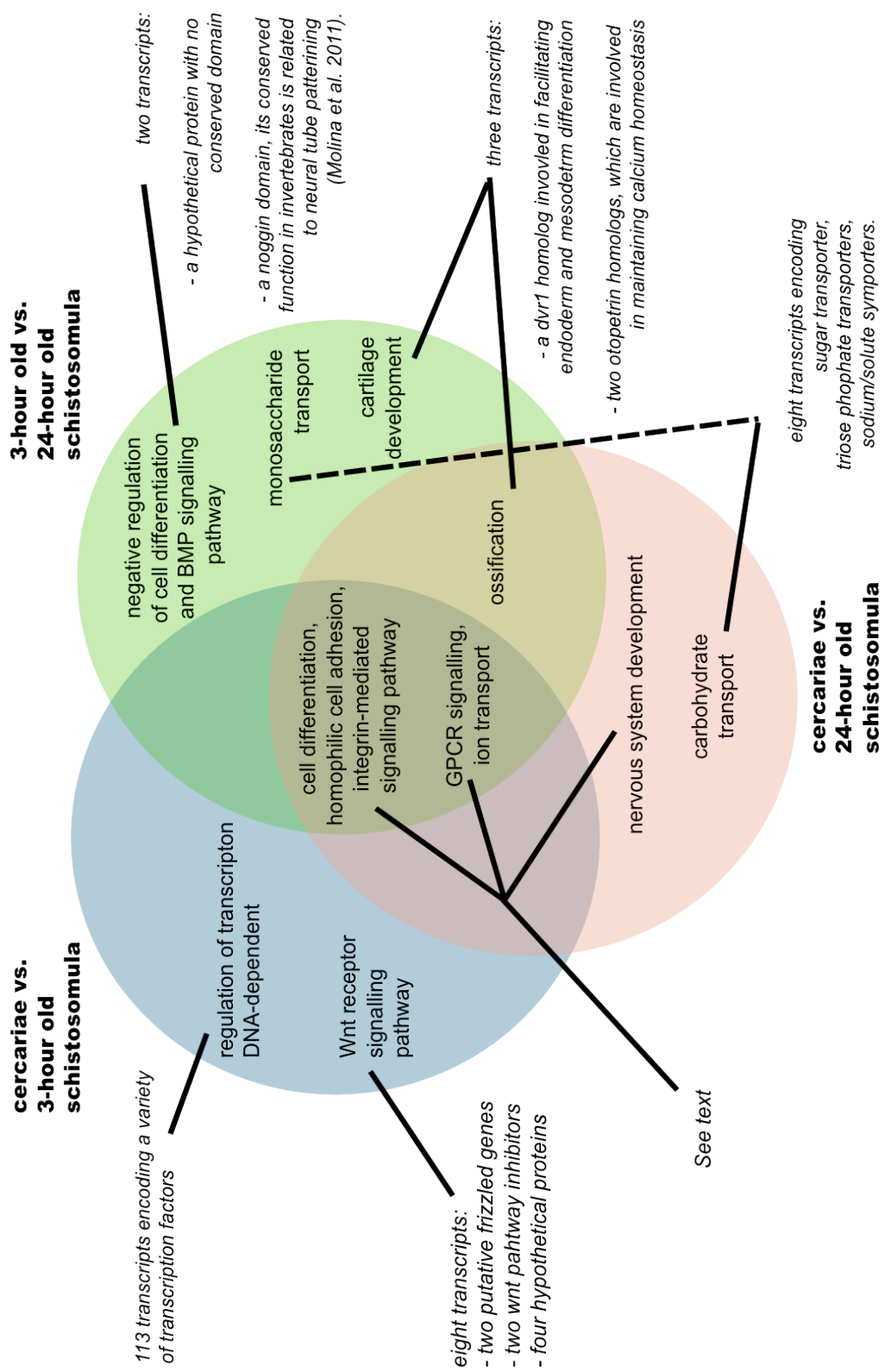


Figure 5.4 – Venn diagram representing biological processes enriched among up regulated genes in the cercariae to 3- and 24-hours old schistosomula.

Up regulated genes under the "transcription regulation DNA-dependent" category comprise a total of 113 transcripts with a range of fold change values between 2.7 and ~27 fold (linear scale) up regulated at 3 hours after transformation. Expression of these genes is either maintained or slightly up or down regulated at 24 hours, but none of them is down regulated with respect to the cercariae (**Figure 5.5**). These transcripts encode transcription factor (TF) domains that range from very conserved eukaryotic domains, such as the POU domain (see section 5.2.3.2.1), to less characterised ones such as DNA- or RNA-binding domains. Seventeen of the up regulated TF encode homeodomains. Genes encoding homeodomains are called homeobox or Hox genes; these are responsible for inducing cellular differentiation through the co-regulation of genes required for tissue and organ development and are key players in the establishment of body axis. Their study in parasitic worms is only but starting (Olson, 2008); therefore these findings are relevant to the advancement of the field.

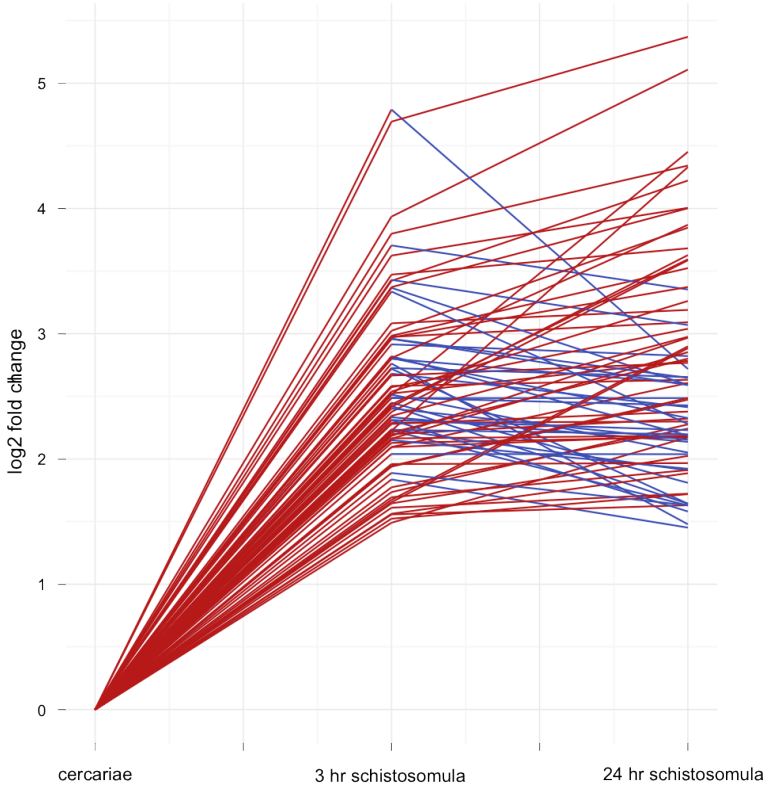


Figure 5.5 – Fold change values of up regulated genes (red) in the category "regulation of transcription DNA-dependent". These genes encode a range of transcription factors and some are slightly down regulated at 24-hours compared to 3-hours old schistosomula (bule).

Consistently with the up regulation of transcriptome factors involved in development, genes involved in many aspects of cell differentiation and tissue development (multicellular organismal development, homophilic cell adhesion, cell differentiation, cell adhesion, cell-matrix adhesion, integrin-mediated signalling pathway among others) are also up regulated indicating that schistosomula have started the programmatic chain of events that could lead to the development of tissues and organs as early as 3 hours after transformation. As mentioned earlier, mitosis does not occur in schistosomula until these are at least 3-4 days old (Clegg *et al.*, 1972); which is reflected in the inhibition of DNA replication. Hence, it is not possible that cell proliferation is actively occurring at this stage. However, as indicated by the GO term enrichment analysis, there is a profound up regulation (15 out of the 61 annotated in the genome) of transcripts encoding putative cadherins, integrins and laminins. Moreover, all these transcripts have their peak of expression at the schistosomula stage (either 3 hours or 24 hours old) suggesting that mRNAs necessary for the formation of such structures are ready and probably put "on hold" during the cell arrest period. Once this is gone, the whole machinery could be quickly activated to generate tissue development.

It was previously noticed that the spliceosome machinery is down regulated suggesting that pre-mRNA processing might be compromised or occur at a slower rate than in the schistosomula stage. Hence, it is possible to hypothesise that the up regulated TF might be inhibiting expression rather then inducing it. However, it was not possible to identify transcription repressors among these genes. It is worth mentioning that the down regulation of the spliceosome machinery seems to occur after the schistosomula have initiated transformation and therefore cannot be attributed to the loss of the tail ("spliceosome" category does not show in the Table 5.2 where the cercariae and 3-hours old schistosomula comparison is presented). The relationship between the down regulation of spliceosome machinery and the up regulation of TF remains to be explored.

5.2.3.2.1 Nervous system development is triggered in the early schistosomula

Consistent with the tissue developmental programme, transcripts involved nervous system development are up regulated in schistosomula. These are grouped in three GO categories: "nervous system development", "negative regulation of cell differentiation" and "negative regulation of BMP (bone morphogenic pathway) signalling pathway". The last two terms are closely related as they hold the same transcripts; one of them encodes a

hypothetical protein while the other encodes a *noggin* domain. The latter has a role in regulating neural development in invertebrates (Molina *et al.*, 2011) and therefore was considered within the nervous system development.

Evidence of the up regulation of genes involved in the nervous system development led to investigate whether any of the up regulated transcription factors were related to neural tissue development. Of the 20 genes involved in "early neural patterning" present in the of *S. mansoni* genome (Berriman *et al.*, 2009), RNA-seq data showed that 15 of them are expressed above background in at least one of the life cycle time points studied. What is more, 12 of them have their peak of expression in either 3-hours old or 24-hours old schistosomula suggesting a role in the development of the parasite at this stage. As an example, a well-conserved eukaryotic transcription factor (POU domain) responsible for specifying the identity of individual neural cells [reviewed in (Hobert, 2010)] is found among the up regulated TF, suggesting that genes expressed at this stage play a key role in determining the fate of neural cells in *S. mansoni*.

This wave of activation of transcription of genes involved in nervous system development led to investigate whether other aspects of neuron proliferation and function were also up regulated. One important aspect to the physiology of the nervous system is communication between neuronal cells, which is mediated in part by neuropeptides and neuropeptide receptors.

S. mansoni has approximately 30 GPCR belonging to the neuropeptide receptors sub-classification [(Berriman *et al.*, 2009) and (Zamanian, 2011)], 18 of which are up regulated in 24-hours old parasites with fold change values ranging from 2.6 to 55 times (linear scale). What is more, all except one are down regulated in adult worms, suggesting an important role during the schistosomula stage. The up regulation of these neuropeptide receptors occurs at the same time with the expression of nervous system development transcripts providing possible grounds for transcriptional co-regulations and further evidence of the active development of this tissue.

The distinct profile of expression of the neuropeptide receptors led to investigate their potential ligands. McVeigh *et al.,* (2009) reported a very comprehensive analysis of the neuropeptide precursor (npp) sequences found in platyhelminths including *S. mansoni* (McVeigh *et al.*, 2009). The authors found 14 npp that they grouped into 11 families. Five of the 14 sequences had already been associated to a gene model in GeneDB (GeneDB, 2011); here the other 9 matches are reported (for full description of Methods see Chapter 2 section 2.6.9). The rest of the npp reported for other platyhelminth species by McVeigh *et al.,* (2009) could not be found in the genome.

Of all the neuropeptide precurosrs, Sm-npp-20b is the only significantly up regulated - approximately 3 times (adjusted p-value < 0.01) compared to cercariae - in 3-hours old schistosomula and rapidly down regulated at 24 hours after transformation remaining low in the adult. This pattern of expression suggests that Sm-npp-20b has a role in the developing schistosomula; which remains so far unknown. Sm-npp-20b belongs to the neuropeptide F (NPF) subfamily of neuropeptidedes; whose individual peptide products are similar to those found in vertebrates [reviewed in (McVeigh *et al.*, 2009)] – these are called NPYs. The function of vertebrate NPYs involves inhibiting the accumulation of cyclic adenosine monophosphate (cAMP) in a concentration-dependent manner. The structural similarity between NPFs and NPYs suggest a similar role in *S. mansoni* (Humphries *et al.*, 2004; McVeigh *et al.*, 2009). However, whether Sm-npp-20b acts as a ligand of any of the neuropeptide receptors up regulated at 3 or 24 hours after transformation will require further investigations. Given the large number of neuropeptide receptors expressed in the schistosomula, it is possible that *S. mansoni* has a larger battery of npp than the one identified so far and perhaps the npp acting upon the neuropeptide receptor identified in the skin schistosomula stage are among those missing. These neuropeptide receptors could also have a role in sensing host-derived neuropeptides.

5.3 Discussion

Previous high-throughput studies aimed at describing transcriptional changes occurring during the transformation of the cercariae into the skin stage schistosomula (Fitzpatrick *et al.*, 2009; Gobert *et al.*, 2010). Although some insightful results could be drawn from these works, they were limited by the inherent disadvantages of microarrays; some of which can be overcome by the use of sequencing - instead of hybridisation - in the analysis of gene expression. The improvement of the *S. mansoni* genome (Protasio *et al.*, 2012) together with the advancement of RNA-seq as the cutting edge technique for analysing messenger RNA samples made it possible to address the question of which genes are developmentally regulated in the early skin schistosomula stage without the restriction of surveying only known features. Moreover, the greater dynamic range and digital nature of the RNA-seq approach allow more accurate measurements of gene expression at both ends of the spectrum.

In this chapter, focus was made on identifying signals that are important for the developmental stage of the parasite. To this end, gene expression from schistosomula

stage was compared to that of cercariae and adult worms. This allowed the identification of genes that are triggered by the transformation process and that might be relevant for establishment of infection.

As a first step, data from RNA-seq data were compared to genes regarded as constitutively expressed (Fitzpatrick *et al.*, 2009). In this case, data broadly agrees with the results from Fitzpatrick *et al.,* (2009); discrepancies may be due to the known limitation of microarrays in discriminating signals from highly expressed genes (as demonstrated in Chapter 3 section 3.2.3). Then, genes known as differentially expressed [8 kDa CaBP (Ram *et al.*, 1989), HSP70 (Hedstrom *et al.*, 1987) and Sm22.6/SmTAL1 (Stein *et al.*, 1986)] were also investigated. Results showed that data obtained from the measurement of expression of individual mRNAs from independent experiments correlate with those of RNA-seq (section 5.2.2). Apart from correlation of individual genes, some of the processes found to change in the cercariae to schistosomula transformation, such as the down regulation of transcripts involved in aerobic metabolism (Skelly *et al.*, 1993; Parker-Manuel *et al.*, 2011) gut function and gut tegument (Gobert *et al.*, 2010) were also found.

Most importantly, with the RNA-seq differential expression approach it was possible to identify groups of genes involved in other processes that were either unknown to happen in the schistosomula or not so well explored. Some examples were presented in this chapter and are discussed here.

Firstly, many signals indicating a state of cell arrest were identified in this study. The most prominent of these signals is the down regulation of all the constituents of the nucleosome. Results shown in this chapter suggest that schistosomula down regulate histone mRNA synthesis probably as a consequence of the state of cell cycle arrest (Clegg *et al.*, 1972). Other signals related to cell division and cell function are also down regulated.

A battery of transcription factors found up regulated in 3-hour old schistosomula, whose expression is maintained in the 24 hours-old parasites and decays in adult worms. Among these, transcription factors involved in the nervous system development were found, indicating that among tissue development signals, neural development is important in this early stage of the schistosomula. If tissue development is important at this stage so cell communication should be. This led to find that many G-protein couple receptors were up regulated in the 3- and 24-hours old parasites. It is known that these receptors have a central role in the cell-to-cell communication in all metazoan organisms (Alberts *et al.*, 2002) and therefore their expression is expected in all stages of the parasite development.

However, the preferential up regulation of the expression of neuropeptide receptors in 24-hours old schistosomula has not been reported previously and open new questions regarding the role of these receptors in the development of the skin stage schistosomula. On the other hand, GPCR proteins and more specifically neuropeptide receptors have been tagged as potential drug targets in other systems (Overington *et al.*, 2006) and also in helminths (Greenwood *et al.*, 2005). Knowing when and which of these receptors are expressed at any time represents an opportunity to understand their function as well as it narrows the list of potential drug targets that would need to be empirically tested.

Regarding the potential ligands for these receptors, only one of the known neuropeptide precursors was found coregulated with the neuropeptide receptors. The participation of other neuropeptides so far unknown as well as the interaction with neuropeptides derived from host proteins cannot be discarded.

On the one hand, parasites seem to be in a state of cell cycle arrest; which has been previously described and can be detected here at the molecular level. Together with cell cycle arrest, other functions related to cell division, such as generation of microtubules and protein synthesis in general seem to be also at halt. On the other hand, the massive up regulation of transcriptome factors, many of them related to tissue development, suggests that there is an underlying programmed fate for these parasite to undergo development. Reduced protein synthesis suggest that this process is regulated post transcriptionally but before translation, probably with the objective of saving resources in case they are needed. Having said that, it is hypothesised that during the parasites first 24 hours in the mammalian host, these organisms suppress parasites' growth during the early skin stage, and this is reflected in reduction of elements needed for cell division. Nevertheless, parasites are getting ready for swift transition into development by preparing part of the machinery (i.e., transcription factors, GPCRs) needed to undergone development.

CHAPTER 6

CONCLUDING REMARKS

Schistosomiasis is a human parasitic disease caused by infection with platyhelminths from the genus *Schistosoma*. This disease is endemic in many countries; especially in poor settings were resources for both prevention and treatment are scarce. Parasites invade the human host by penetrating through healthy skin during water contact. The infectious agent, the free-living cercariae, transforms into a parasitic form, the schistosomula, which migrates through the skin, reaching the circulatory system, then the lungs and the portal system. Finally, parasites develop into adult worms: males and females. The latter lays hundreds of eggs each day, which are the cause of the pathology. Infection can be treated by administration of praziquantel, a well-tolerated and cheap drug. However, wide spread treatment with this drug and reports of laboratory-induced resistance in worms and reduced susceptibility in the field have raised concern among researchers and public heath agencies about the emergence of resistance. In this context, the identification of novel drug targets and the development of a vaccine that would prevent infection are key aspects of schistosome research.

During the early stages of infection the schistosomula are located in the host's skin. This stage is thought to be the most vulnerable for parasite killing. However, only a couple of studies have focused on describing the gene expression landscape present in these early stages of the parasites' development in the human host. The work presented in this thesis focused on describing the gene expression changes occurring to the parasites during the transformation from the cercariae into the schistosomula. Characterisation and understanding of this transformation and the pathways that lead to the adaptation of the parasite to its host are key aspects to the development of intervention strategies. In order to tackle this, four time points of the life cycle of *S. mansoni* were sequenced using RNA-seq technology. This approach offers many advantages in comparison with the previously used methods to investigate gene expression, such as microarrays.

RNA-seq improved the gene annotation of *S. mansoni.*

As the objective of this project was the study of gene expression, it was imperative to revisit the existing structural annotation of the gene models. RNA-seq data was used to improve the current annotation by providing whole transcriptome experimental evidence of the existence and boundaries of transcribed regions. This new RNA-seq derived annotation was evaluated against previous *in silico* predictions (**Figure 3.10**) providing enough evidence that the new predictions were indeed more representative of the transcriptional landscape of the parasites. Many examples of fusion, splitting and addition of exons were presented in **Figure 3.6-3.10** and illustrated how these contributions affect

the actual structure of genes with profound effects on downstream applications such as gene expression, cloning, etc. As an example, **Figure 3.6** (section 3.2.4.2) showed how a very long gene model was resolved into two smaller ones based on RNA-seq data. Because the length of the coding region is used to normalise the expression of genes, the new and older versions of this gene model would have yielded very different results. These individual examples are scalable to bigger projects. For instance, current studies into possible vaccine target candidates [such as the SchistoVac consortium (TheSchistoVac, 2009)] use GeneDB gene models to predict gene function and it is the combination of these and gene expression data that lead researchers' decisions on a priority list of genes worth further investigation. Starting with an incomplete or inaccurate dataset would slow research on such vaccine candidates and delay the arrival of the so needed alternative treatments and prophylaxis. Additionally, proteomic analyses that use mass-spectrometry datasets rely on accurate and complete gene models to correlate the short peptides with full-length *in silico* predictions of polypeptides. For example, the work of Hansell *et al.,* (2008) relied on the older version of gene models and it is recommended these data be revisited and evaluated against the new dataset of gene models.

Trans-splicing events affect 9% of coding sequences

Another finding that led into further modification of gene structures is the description of a comprehensive list of *trans*-splicing events and hence a list of putative *trans*-spliced transcripts. Previous efforts of genome-wide identification of *trans*-splicing events have used a limited number of ESTs (Davis *et al.*, 1995). Interestingly, the efforts to uncover *all trans*-spliced transcripts in *S. mansoni* were abandoned probably due to the difficulty of obtaining a larger dataset of transcripts undergoing *trans*-splicing. The RNA-seq data presented in this thesis was used to generate a high-resolution map of *trans*-splicing events in genome-wide fashion. Similar approaches have already been exploited in other systems (Kolev *et al.*, 2010; Allen *et al.*, 2011). In the case of *S. mansoni,* the description of a comprehensive list of *trans*-spliced transcripts has many consequences. For example, most *trans*-splicing events occur in the 5' most region of the gene model, sometimes not even within the coding regions. This would mean that before adding the information of *trans*-splicing, this model would have had an incorrect start of transcription. Having the correct start of transcription is fundamental for many aspects of research, such as generating the correct primers for PCR amplification and posterior cloning, predict whether a given gene would encode a secreted protein, etc. What is more, whether one transcript is *trans*-spliced or not would define which translation machinery it would use.

Trans-spliced transcripts acquire a m(2,2,7)G-cap or TMG-cap whereas the non-*trans*-spliced transcripts have a "standard" m(7)G-mRNA cap. The effective translation of TMG-capped transcripts depends on the presence of a stem loop formed in the SL sequence; which in turn depends on the assembly of a specific translation initiation complex (Wallace *et al.*, 2010). Thus, *trans*-spliced transcripts use an exclusive translational machinery and might be subjected to a non-conventional or at least different regulation of translation. This represents an opportunity for potential metabolic chokepoints leading to the development of new intervention drugs.

Additionally, the identification of genome wide *trans*-splicing events gave the opportunity to revisit an old question: are *trans*-splicing transcripts related to a given function? Previous reports have rendered inconclusive *trans*-splicing-to-function associations, probably due to the low number of confirmed *trans*-spliced transcripts. By analysing the extended dataset of *S. mansoni trans*-splicing events it was possible to identify one particular pathway, the glycosylphosphatidylinositol-anchored protein (GPI-APs) biosynthesis, in which some of the core enzymes are subjected to *trans*-splicing (**Figure 6.1**).

It has also been suggested that *trans*-spliced transcripts are part of the kit of constitutively expressed genes also known as "housekeeping" genes. These are thought to be essential for the survival and/or development of the organism and therefore can represent potential targets for intervention. Due to their conserved function, it is likely that their sequences and protein configuration are also very conserved even between the parasite and the host; representing a challenge for drug design. The dataset of *trans*-spliced transcripts reported in this thesis included a significant number of "hypothetical proteins" that could have core or housekeeping function in the parasites but that might not be found in the host. These represent good candidates for further drug target development.

Trans-splicing is not an on/off phenomenon. Most *trans*-spliced transcripts show a percentage of *trans*-spliced transcripts while the other are not *trans*-spliced at all (Matsumoto *et al.*, 2010). Deeper sequencing coverage could reveal other less frequent cases of *trans*-splicing and therefore it may be more accurate to refer to *trans*-spliced transcripts as frequently or non-frequently *trans*-spliced. This concept already introduced by Matsumoto *et al.,* (2010) will facilitate the accurate interpretation of the function and prevalence of *trans*-splicing.

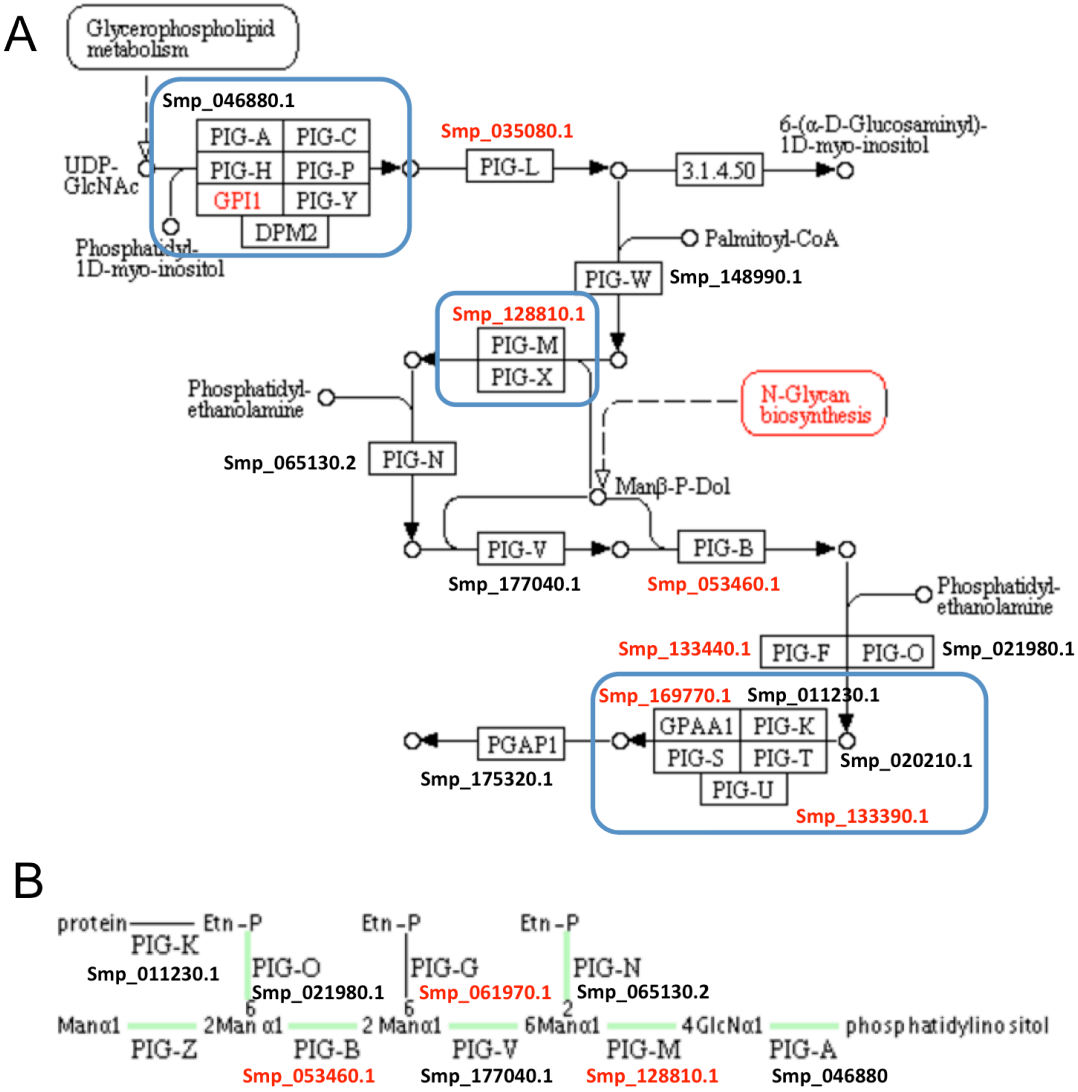


Figure 6.1 – Pathway map of glycosylphosphatidylinositol-anchored protein (GPI-APs) biosynthesis obtained from KEGG database (Wixon *et al.*, 2000). Where orthologs could be found (BLASTp e-value < $1e^{-10}$), *S. mansoni* gene identification names have been added. Red names represent genes whose transcripts are *trans*-spliced. A- Full pathway described in the KEGG database, Protein complexes are marked in blue squares. B – Minimum number of steps and enzymes required to generate a GPI-AP.

Polycistronic transcription

Polycistrons are clusters of two or more individual coding sequences that are transcribed as part of one large pre-mRNA molecule. In eukaryote systems, polycistrons are resolved by *trans*-splicing, which results in individual mRNAs. Because of this organization, one promoter may regulate the polycistron's transcription (making it an operon) and therefore all the transcripts in the same polycistron will be in principle subjected to the same regulation of transcription. Using the dataset of *trans*-splicing events in combination with the improved genome assembly and improved annotation of coding sequences, it was possible to identify a group of putative polycistronic transcripts. What is more, it was possible to provide for the first time *bona fide* experimental evidence of the presence of transient polycistronic transcripts as well as their *trans*-spliced products (**Figure 3.11**). It has previously been suggested that the individual transcripts of a polycistron might be functionally related, for example they would participate in the same pathway. Although it was not possible to find a functional link between the transcripts encoded in the so far identified *S. mansoni* transcripts, it is predicted that deeper sequencing of RNA species and better functional annotation of the gene product would shed light on this question. This information in conjunction with the increasing number of sequencing projects targeting helminths and parasitic nematodes will provide more information about the origins, prevalence and evolution of polycistronic transcription.

Skin *vs.* mechanical transformation of schistosomula.

Due to the complexities of the schistosomes' life cycle, an artificial mechanism to obtain schistosomula was developed almost 40 years ago (Brink *et al.*, 1977). The aim of this approach was to facilitate the collection of large numbers of schistosomula required for experimentation. Many works on the comparison of the "mechanically transformed" and schistosomula obtained from infected animals provided evidence that in most aspects of the physiology and anatomy, these two schistosomula preparations would be equivalent. Given this, many transcriptome projects based their experimental approach solely on mechanically transformed schistosomula. However, it was not until now that a thorough comparison of the transcriptomes of mechanically and skin-transformed schistosomulum was made. The results presented in chapter 4 of this thesis resolves this long-standing controversy and it is now possible to establish that 24-hours old skin- and mechanically transformed schistosomula are transcriptionally equivalent except for 149

genes that show differential expression. This experiment validates previous finding resulting from the use of mechanically transformed schistosomula.

Mitochondrial transcripts are found among differentially expressed genes and it was possible to show that their higher expression in skin-transformed schistosomula has consequences in their metabolic rate. It is possible to hypothesise that the observed lower expression of mitochondrial transcripts in the mechanically transformed schistosomula is a result of these parasites being more heterogeneous than the skin transformed ones. In this scenario, the skin might be acting as a selective barrier that can only be trespassed by the fittest individuals, maybe representing an instance of natural selection. Because investigation in drug development usually relies on survival rates of parasites (Mansour *et al.*, 2010), it is important to remember that the population of mechanically transformed schistosomula usually used for drug testing might be a mixture of fit and no-so fit parasites and that this can affect the outcome of the drug assay because no-so fit parasites can be more susceptible to perish due to drug treatment.

It is worth mentioning that genes related to stress were not found among the differentially expressed transcripts. It is concluded that the schistosomula is a much tougher organism than previously thought and that its gene expression portfolio is barely affected by the transformation method applied.

Gene expression in the skin-stage schistosomula

The main objective of this thesis project was to investigate the transcriptional changes occurring to the schistosomula during the first 24 hours of life in the host. Once the improvement of gene models and the validation of skin- and mechanically transformed parasites were performed, it was possible to proceed to the analysis of skin-stage schistosomula transcriptome in comparison to the cercariae and adult worms. Many microarray studies have investigated different time points in the development of schistosomes (Fitzpatrick *et al.*, 2005; Vermeire *et al.*, 2006; Jolly *et al.*, 2007; Verjovski-Almeida *et al.*, 2007). In particular, many have focused on the cercariae and several post-transformation time points (Dillon *et al.*, 2006; Gobert *et al.*, 2010; Parker-Manuel *et al.*, 2011). However, the only existing report on transcriptional analysis of 3-hours old and 24-hours old schistosomula lacked statistical power and no conclusions could be drawn (Fitzpatrick *et al.*, 2009). What is more, it is suspected that samples used to profile the transcriptome of 3-hours old schistosomula may have a significant amount of tail contamination, which resulted in the misinterpretation of results regarding potential a potential allergen molecule (**Figure 6.2**). The results presented in chapter 5 bridge the

existing gap between reliable transcriptome data obtained from cercariae and early-skin stage schistosomula (3- and 24-hours old). Due to the existence of tail specific markers, it was possible to assess that the schistosomula samples were virtually free from contaminating tails. The conclusions found during the analysis of the schistosomula transcriptome presented in this thesis suggest that there is much to find out about the development of the parasite organs and tissues during the skin-stage. The identification of a battery of transcription factors together with effector proteins such as integrins and cadherins will open new opportunities in the search of targets of intervention. Finding which of these players are key in the development of the parasites will require further research but this can now be narrowed to the study of those gene products found to be expressed during the first hours of the parasite life within the mammalian host.

Finally, the combination of lack of translation and mitosis at this stage in the development of the parasite together with the active transcription of transcription factors and genes known to be key in the development of the nervous system suggest that the parasites might be arming its molecular machinery for the time when environmental factors are favourable for its development.

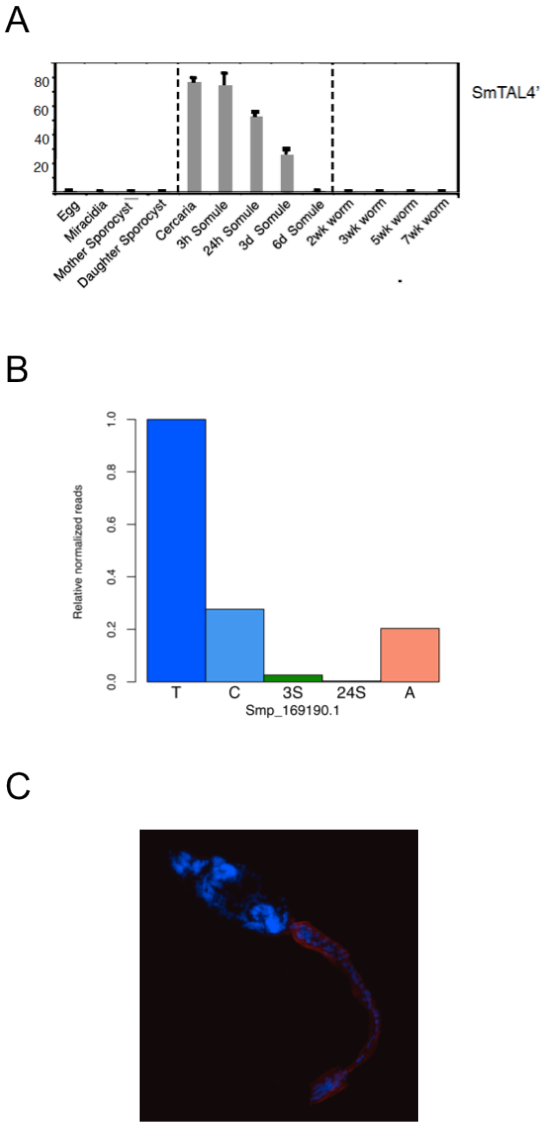


Figure 6.2 – Expression profile of SmTAL4. A – Relative expression retrieved from microarray data (Fitzpatrick *et al.*, 2009) for SmTAL4. Note the high level of expression in the 3- and 24-hours old schistosomula. B – RNA-seq relative normalized expression of SmTAL4. T, tail; C, cercariae; 3S, 3-hours old schistosomula, 24S, 24-hours old schistosomula; A, adult. C – Immunohistochemical staining of SmTAL4 protein. Note how the SmTAL4 protein is located exclusively in the tail of the cercariae and absent from the head or tail junction. Courtesy of Jakub Wawrzyniak (Dunne group – Dep. Pathololgy, U of Cambridge, UK)

CHAPTER 7

APPENDIXES

7.1 Appendix A

Primer name	Primer sequence
SL1	CCGTCACGGTTTTACTCTTG
Smp_102510_R	GCAAATGCTGTCCATAAAGC
Smp_024110_R	CAGCCAAAGACACTCCCAAT
Smp_027360_R	TTTTACTATTAGGACTTTGTGGTGATG
Smp_016410_R	GACTACACGGCGGTACAGGA
Smp_141320_R	TGAATGATGAGGTGTTGGACA
Smp_136960_R	TTCACTTTCCCGCAGTTTT
Smp_124050_R	CATTGCATTTCCATATTGTTCA
Smp_176420_R	AAAACCTTCTGTCTTAATTGTGGTG
Smp_176590_R	AATGCGGGTACGTCTGATATG
Smp_030020_R	CAGCACCAAGTGGAACTGAA
Smp_048880_R	ATTCTTCTGCAGCCTCGTTG
Smp_045200_R	AACTACTTGCCATACACGACCA
Smp_045200_F	ATGGCAACCGAGACGAAAT
enolase_poly_F	TGTTCGGATTCAACAATGCT
enolase_poly_R	TCCACCTCAACTGTGGGATT
Smp_006980-70_F	ATGAGGGGTGCACTTACGAC
Smp_006980-70_R	ATTTCAAACTCGGGACATGC
Smp_084900-890_F	5' GGCTGATGTTGAAGCACAAA
Smp_084900-890_R	5' AGGAGTGAAACGCTGCAAAT
Smp_084890_R	GATCGACTGGATGACGACCT
Smp_079750-60_F	GCAGATTTGACGGAAAATTCA
Smp_079750-60_R	CGTTAATACGAGCTCCACGA
Smp_079760_R	TCATTTGCAGCATCCACATT
Smp_023160-70_F	ACTCCCACAATGTTGCCATA
Smp_023160-70_R	CCGAAATCCCAGACTGACTC
Smp_023170_R	AGGCTCGCACATCCTTAAAA
VAL6_F	CCGGATCAGCAAATAATGACA
VAL6_R	TGATCCCAGTAACATTTGCATC

7.2 Appendix B

GeneDB_ID	Product description
Smp_152660.1	glial cells missing related
Smp_199740.1	Conserved oligomeric Golgi complex component 4, putative;with=UniProt:Q5R7R6
Smp_182910.1	hypothetical protein
Smp_191250.1	hypothetical protein
Smp_159840.1	dynein heavy chain, putative
Smp_156080.1	dynein heavy chain, putative
Smp_194890.1	subfamily S1A non-peptidase homologue (S01 family);with=UniProt:P12546
Smp_006510.1	cercarial elastase (S01 family);with=UniProt:P12546
Smp_112090.1	cercarial elastase (S01 family);with=UniProt:P12546
Smp_167120.1	leishmanolysin-2 (M08 family);with=UniProt:Q8BMN4
Smp_185230.1	arginine/serine-rich splicing factor, putative
Smp_190370.1	rab2, putative
Smp_107700.1	hypothetical protein
Smp_179410.1	hypothetical protein
Smp_185190.1	cercarial elastase (S01 family);with=UniProt:P12546
Smp_119130.1	cercarial elastase (S01 family);with=UniProt:P12546
Smp_190590.1	hypothetical protein
Smp_136830.1	subfamily A1A unassigned peptidase (A01 family);with=UniProt:Q05744
Smp_135320.1	Hypothetical protein, putative
Smp_173950.1	hypothetical protein
Smp_147890.1	Rootletin (Ciliary rootlet coiled-coil protein), putative
Smp_151860.1	inositol 1,4,5-trisphosphate receptor, putative
Smp_133640.1	hypothetical protein
Smp_177900.1	hypothetical protein
Smp_151260.1	Adenine phosphoribosyltransferase, putative

7.3 Appendix C

Gene Ontology enrichment (Biological processes) among transcripts up regulated in the cercariae *vs.* schistosomula comparison. The comparisons are indicated in each table

Table C.1 – cercariae *vs.* 3-hours old schistosomula

GO term ID	GO term description	Annot.	Sig.	p-value
GO:0006355	regulation of transcription DNA-dependent	881	113	0
GO:0007275	multicellular organismal development	301	59	0
GO:0007186	G-protein coupled receptor protein signalling	165	32	0
GO:0007156	homophilic cell adhesion	61	15	0
GO:0030154	cell differentiation	82	19	0
GO:0016055	Wnt receptor signaling pathway	26	8	0.001
GO:0007155	cell adhesion	235	42	0.001
GO:0051258	protein polymerization	44	8	0.001
GO:0007229	integrin-mediated signaling pathway	8	4	0.001
GO:0006813	potassium ion transport	81	13	0.004

Table C.2 – cercariae *vs.* 3-hours old schistosomula.

GO term ID	GO term description	A	S	p-value
GO:0007156	homophilic cell adhesion	61	13	0
GO:0007186	G-protein coupled receptor protein signalling	165	18	0
GO:0007229	integrin-mediated signaling pathway	8	4	0
GO:0001503	ossification	4	3	0
GO:0007155	cell adhesion	235	31	0
GO:0051216	cartilage development	5	3	0
GO:0007160	cell-matrix adhesion	7	3	0.001
GO:0006811	ion transport	389	27	0.003
GO:0045596	negative regulation of cell differentiattion	3	2	0.003
GO:0030514	negative regulation of BMP signaling pathway	3	2	0.003
GO:0006813	potassium ion transport	81	8	0.006
GO:0015749	monosaccharide transport	4	2	0.007

Table C.3 – cercariae *vs.* 3-hours old schistosomula.

GO term ID	GO term description	A	S	p-value
GO:0007156	homophilic cell adhesion	61	40	0
GO:0007186	G-protein coupled receptor protein signa...	165	61	0
GO:0007275	multicellular organismal development	301	94	0
GO:0007155	cell adhesion	235	96	0
GO:0006811	ion transport	389	86	0
GO:0006814	sodium ion transport	67	22	0
GO:0007229	integrin-mediated signaling pathway	8	6	0
GO:0030154	cell differentiation	82	24	0.001
GO:0007160	cell-matrix adhesion	7	5	0.001
GO:0006813	potassium ion transport	81	22	0.001
GO:0007399	nervous system development	27	9	0.006
GO:0001503	ossification	4	3	0.009
GO:0008643	carbohydrate transport	36	12	0.009

8 REFERENCES

Abbas, A. K., A. H. Lichtman and J. S. Pober (2000). Cellular and molecular immunology. Philadelphia, Saunders.

Abdulla, M. H., D. S. Ruelas, B. Wolff, J. Snedecor, K. C. Lim, F. Xu, . . . C. R. Caffrey (2009). "Drug discovery for schistosomiasis: hit and lead compounds identified in a library of known drugs by medium-throughput phenotypic screening." PLoS Negl Trop Dis **3**(7): e478.

Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, . . . J. C. Venter (2000). "The genome sequence of *Drosophila melanogaster*." Science **287**(5461): 2185-2195.

Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts and P. Walter (2002). Molecular Biology of the Cell. New York, Garland Science.

Alexa, A., J. Rahnenfuhrer and T. Lengauer (2006). "Improved scoring of functional groups from gene expression data by decorrelating GO graph structure." Bioinformatics **22**(13): 1600-1607.

Allen, J. E., W. H. Majoros, M. Pertea and S. L. Salzberg (2006). "JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions." Genome Biol **7 Suppl 1**: S9 1-13.

Allen, J. E. and S. L. Salzberg (2005). "JIGSAW: integration of multiple sources of evidence for gene prediction." Bioinformatics **21**(18): 3596-3603.

Allen, M. A., L. W. Hillier, R. H. Waterston and T. Blumenthal (2011). "A global analysis of *C. elegans* trans-splicing." Genome Res **21**(2): 255-264.

Almeida, G. T., M. S. Amaral, F. C. Beckedorff, J. P. Kitajima, R. Demarco and S. Verjovski-Almeida (2011). "Exploring the *Schistosoma mansoni* adult male transcriptome using RNA-seq." Exp Parasitol.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). "Basic local alignment search tool." J Mol Biol **215**(3): 403-410.

Amos, L. A. (2008). "The tektin family of microtubule-stabilizing proteins." Genome Biol **9**(7): 229.

Armstrong, P. B. (2006). "Proteases and protease inhibitors: a balance of activities in host-pathogen interaction." Immunobiology **211**(4): 263-281.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, . . . G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Barrett, J. (1981). Biochemistry of parasitic helminths. London, Macmillan.

Benjamini, Y., D. Drai, G. Elmer, N. Kafkafi and I. Golani (2001). "Controlling the false discovery rate in behavior genetics research." Behav Brain Res **125**(1-2): 279-284.

Bennett, M. W. and J. P. Caulfield (1991). "*Schistosoma mansoni*: ingestion of dextrans, serum albumin, and IgG by schistosomula." Exp Parasitol **73**(1): 52-61.

Bennett, S. T., C. Barnes, A. Cox, L. Davies and C. Brown (2005). "Toward the 1,000 dollars human genome." Pharmacogenomics **6**(4): 373-382.

Berriman, M., B. J. Haas, P. T. LoVerde, R. A. Wilson, G. P. Dillon, G. C. Cerqueira, . . . N. M. El-Sayed (2009). "The genome of the blood fluke *Schistosoma mansoni*." Nature **460**(7253): 352-358.

Blumenthal, T. and K. S. Gleason (2003). "*Caenorhabditis elegans* operons: form and function." Nat Rev Genet **4**(2): 112-120.

Brehm, K., K. Hubert, E. Sciutto, T. Garate and M. Frosch (2002). "Characterization of a spliced leader gene and of trans-spliced mRNAs from *Taenia solium*." Mol Biochem Parasitol **122**(1): 105-110.

Brehm, K., K. Jensen and M. Frosch (2000). "mRNA trans-splicing in the human parasitic cestode *Echinococcus multilocularis*." J Biol Chem **275**(49): 38311-38318.

Brink, L. H., D. J. McLaren and S. R. Smithers (1977). "*Schistosoma mansoni*: a comparative study of artificially transformed schistosomula and schistosomula recovered after cercarial penetration of isolated skin." Parasitology **74**(1): 73-86.

Brinkworth, R. I., P. Prociv, A. Loukas and P. J. Brindley (2001). "Hemoglobin-degrading, aspartic proteases of blood-feeding parasites: substrate specificity revealed by homology models." J Biol Chem **276**(42): 38844-38851.

Caffrey, C. R., J. H. McKerrow, J. P. Salter and M. Sajid (2004). "Blood 'n' guts: an update on schistosome digestive peptidases." Trends Parasitol **20**(5): 241-248.

Capron, M. and A. Capron (1986). "Rats, mice and men - models for immune effector mechanisms against schistosomiasis." Parasitol Today **2**(3): 69-75.

Carver, T., M. Berriman, A. Tivey, C. Patel, U. Bohme, B. G. Barrell, . . . M. A. Rajandream (2008). "Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database." Bioinformatics **24**(23): 2672-2676.

Carver, T., U. Bohme, T. D. Otto, J. Parkhill and M. Berriman (2010). "BamView: viewing mapped read alignment data in the context of the reference sequence." Bioinformatics **26**(5): 676-677.

CDC Centers for Disease, Control and Prevention - Parasites - Schistosomiasis.(2011) from http://www.cdc.gov/parasites/schistosomiasis/biology.html.

C. elegans Sequencing Consortium., (1998). "Genome sequence of the nematode *C. elegans*: a platform for investigating biology." Science **282**(5396): 2012-2018.

Chai, M., D. P. McManus, R. McInnes, L. Moertel, M. Tran, A. Loukas, . . . G. N. Gobert (2006). "Transcriptome profiling of lung schistosomula,in vitro cultured schistosomula and adult *Schistosoma japonicum*." Cell Mol Life Sci **63**(7-8): 919-929.

Chalmers, I. W., A. J. McArdle, R. M. Coulson, M. A. Wagner, R. Schmid, H. Hirai and K. F. Hoffmann (2008). "Developmentally regulated expression, alternative splicing and distinct sub-groupings in members of the *Schistosoma mansoni* venom allergen-like (SmVAL) gene family." BMC Genomics **9**: 89.

Chaudhuri, R. R., L. Yu, A. Kanji, T. T. Perkins, P. P. Gardner, J. Choudhary, . . . A. J. Grant (2011). "Quantitative RNA-seq analysis of the transcriptome of *Campylobacter jejuni*." Microbiology.

Cheng, G., L. Cohen, D. Ndegwa and R. E. Davis (2006). "The flatworm spliced leader 3'-terminal AUG as a translation initiator methionine." J Biol Chem **281**(2): 733-743.

Chiancone, E., P. Ceci, A. Ilari, F. Ribacchi and S. Stefanini (2004). "Iron and proteins for iron storage and detoxification." Biometals **17**(3): 197-202.

Clegg, J. A. (1965). "In Vitro Cultivation of *Schistosoma mansoni*." Exp Parasitol **16**: 133-147.

Clegg, J. A. and S. R. Smithers (1972). "The effects of immune rhesus monkey serum on schistosomula of *Schistosoma mansoni* during cultivation in vitro." Int J Parasitol **2**(1): 79-98.

Coles, G. C. (1973). "Further studies on the carbohydrate metabolism of immature *Schistosoma mansoni*." Int J Parasitol **3**(6): 783-787.

Conrad, R., J. Thomas, J. Spieth and T. Blumenthal (1991). "Insertion of part of an intron into the 5' untranslated region of a *Caenorhabditis elegans* gene converts it into a trans-spliced gene." Mol Cell Biol **11**(4): 1921-1926.

Cook, G. C., A. Zumla and P. S. T. d. Manson (2003). Manson's tropical diseases. Edinburgh, Saunders.

Cousin, C. E., M. A. Stirewalt and C. H. Dorsey (1981). "*Schistosoma mansoni*: ultrastructure of early transformation of skin- and shear-pressure-derived schistosomules." Exp Parasitol **51**(3): 341-365.

Crabtree, J. E. and R. A. Wilson (1980). "*Schistosoma mansoni*: a scanning electron microscope study of the developing schistosomulum." Parasitology **81**(Pt 3): 553-564.

Crabtree, J. E. and R. A. Wilson (1985). "*Schistosoma mansoni*: an ultrastructural examination of skin migration in the hamster cheek pouch." Parasitology **91 (Pt 1)**: 111-120.

Crabtree, J. E. and R. A. Wilson (1986). "*Schistosoma mansoni*: an ultrastructural examination of pulmonary migration." Parasitology **92 (Pt 2)**: 343-354.

Criscione, C. D., C. L. Valentim, H. Hirai, P. T. LoVerde and T. J. Anderson (2009). "Genomic linkage map of the human blood fluke *Schistosoma mansoni*." Genome Biol **10**(6): R71.

Curwen, R. S., P. D. Ashton, S. Sundaralingam and R. A. Wilson (2006). "Identification of novel proteases and immunomodulators in the secretions of schistosome cercariae that facilitate host entry." Mol Cell Proteomics **5**(5): 835-844.

Curwen, R. S. and R. A. Wilson (2003). "Invasion of skin by schistosome cercariae: some neglected facts." Trends Parasitol **19**(2): 63-66; discussion 66-68.

Daines, B., H. Wang, L. Wang, Y. Li, Y. Han, D. Emmert, . . . R. Chen (2011). "The *Drosophila melanogaster* transcriptome by paired-end RNA sequencing." Genome Res **21**(2): 315-324.

Dalton, J. P., L. Hola-Jamriska and P. J. Brindley (1995). "Asparaginyl endopeptidase activity in adult *Schistosoma mansoni*." Parasitology **111 (Pt 5)**: 575-580.

Davis, A. (2002). Schistosomiasis. Manson's tropical diseases. P. S. Manson, G. C. Cook and A. Zumla. Edinburgh, Saunders**:** 1434-1469.

Davis, R. E., C. Hardwick, P. Tavernier, S. Hodgson and H. Singh (1995). "RNA trans-splicing in flatworms. Analysis of trans-spliced mRNAs and genes in the human parasite, *Schistosoma mansoni*." J Biol Chem **270**(37): 21813-21819.

Davis, R. E. and S. Hodgson (1997). "Gene linkage and steady state RNAs suggest trans-splicing may be associated with a polycistronic transcript in *Schistosoma mansoni*." Mol Biochem Parasitol **89**(1): 25-39.

Davis, R. E., H. Singh, C. Botka, C. Hardwick, M. Ashraf el Meanawy and J. Villanueva (1994). "RNA trans-splicing in *Fasciola hepatica*. Identification of a spliced leader (SL) RNA and SL sequences on mRNAs." J Biol Chem **269**(31): 20026-20030.

de Boer, J. P., A. A. Creasey, A. Chang, J. J. Abbink, D. Roem, A. J. Eerenberg, . . . F. B. Taylor, Jr. (1993). "Alpha-2-macroglobulin functions as an inhibitor of fibrinolytic, clotting, and neutrophilic proteinases in sepsis: studies using a baboon model." Infect Immun **61**(12): 5035-5043.

DeMarco, R., W. Mathieson, S. J. Manuel, G. P. Dillon, R. S. Curwen, P. D. Ashton, . . . R. A. Wilson (2010). "Protein variation in blood-dwelling schistosome worms generated by differential splicing of micro-exon gene transcripts." Genome Res **20**(8): 1112-1121.

Dillon, G. P., T. Feltwell, J. P. Skelton, P. D. Ashton, P. S. Coulson, M. A. Quail, . . . A. C. Ivens (2006). "Microarray analysis identifies genes preferentially expressed in the lung schistosomulum of *Schistosoma mansoni*." Int J Parasitol **36**(1): 1-8.

Dorsey, C. H. (1976). "*Schistosoma mansoni*: description of the head gland of cercariae and schistosomules at the ultrastructural level." Exp Parasitol **39**(3): 444-459.

Dorsey, C. H., C. E. Cousin, F. A. Lewis and M. A. Stirewalt (2002). "Ultrastructure of the *Schistosoma mansoni* cercaria." Micron **33**(3): 279-323.

Douris, V., M. J. Telford and M. Averof (2010). "Evidence for multiple independent origins of trans-splicing in Metazoa." Mol Biol Evol **27**(3): 684-693.

Dunne, D. W., A. E. Butterworth, A. J. Fulford, H. C. Kariuki, J. G. Langley, J. H. Ouma, . . . R. F. Sturrock (1992). "Immunity after treatment of human schistosomiasis: association between IgE antibodies to adult worm antigens and resistance to reinfection." Eur J Immunol **22**(6): 1483-1494.

el Meanawy, M. A., T. Aji, N. F. Phillips, R. E. Davis, R. A. Salata, I. Malhotra, . . . A. H. Davis (1990). "Definition of the complete *Schistosoma mansoni* hemoglobinase mRNA sequence and gene expression in developing parasites." Am J Trop Med Hyg **43**(1): 67-78.

el-Mezgueldi, M. (1996). "Calponin." Int J Biochem Cell Biol **28**(11): 1185-1189.

Emanuelsson, O., S. Brunak, G. von Heijne and H. Nielsen (2007). "Locating proteins in the cell using TargetP, SignalP and related tools." Nat Protoc **2**(4): 953-971.

Fan, J., D. J. Minchella, S. R. Day, D. P. McManus, W. U. Tiu and P. J. Brindley (1998). "Generation, identification, and evaluation of expressed sequence tags from different developmental stages of the Asian blood fluke *Schistosoma japonicum*." Biochem Biophys Res Commun **252**(2): 348-356.

Farias, L. P., C. A. Tararam, P. A. Miyasato, M. Y. Nishiyama, Jr., K. C. Oliveira, T. Kawano, . . . L. C. Leite (2011). "Screening the *Schistosoma mansoni* transcriptome for genes differentially expressed in the schistosomulum stage in search for vaccine candidates." Parasitol Res **108**(1): 123-135.

Finn, R. D., J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, . . . A. Bateman (2010). "The Pfam protein families database." Nucleic Acids Res **38**(Database issue): D211-222.

Fishelson, Z., P. Amiri, D. S. Friend, M. Marikovsky, M. Petitt, G. Newport and J. H. McKerrow (1992). "*Schistosoma mansoni*: cell-specific expression and secretion of a serine protease during development of cercariae." Exp Parasitol **75**(1): 87-98.

Fitzpatrick, J. M. and K. F. Hoffmann (2006). "Dioecious *Schistosoma mansoni* express divergent gene repertoires regulated by pairing." Int J Parasitol **36**(10-11): 1081-1089.

Fitzpatrick, J. M., D. A. Johnston, G. W. Williams, D. J. Williams, T. C. Freeman, D. W. Dunne and K. F. Hoffmann (2005). "An oligonucleotide microarray for transcriptome analysis of *Schistosoma mansoni* and its application/use to investigate gender-associated gene expression." Mol Biochem Parasitol **141**(1): 1-13.

Fitzpatrick, J. M., E. Peak, S. Perally, I. W. Chalmers, J. Barrett, T. P. Yoshino, . . . K. F. Hoffmann (2009). "Anti-schistosomal intervention targets identified by lifecycle transcriptomic analyses." PLoS Negl Trop Dis **3**(11): e543.

Fitzsimmons, C. M., T. J. Stewart, K. F. Hoffmann, J. L. Grogan, M. Yazdanbakhsh and D. W. Dunne (2004). "Human IgE response to the Schistosoma haematobium 22.6 kDa antigen." Parasite Immunol **26**(8-9): 371-376.

Franco, G. R., E. M. Rabelo, V. Azevedo, H. B. Pena, J. M. Ortega, T. M. Santos, . . . S. D. Pena (1997). "Evaluation of cDNA libraries from different developmental stages of *Schistosoma mansoni* for production of expressed sequence tags (ESTs)." DNA Res **4**(3): 231-240.

Fung, M. C., M. T. Lau and X. G. Chen (2002). "Expressed sequence tag (EST) analysis of a *Schistosoma japonicum* cercariae cDNA library." Acta Trop **82**(2): 215-224.

Gardner, P. P., J. Daub, J. Tate, B. L. Moore, I. H. Osuch, S. Griffiths-Jones, . . . A. Bateman (2011). "Rfam: Wikipedia, clans and the "decimal" release." Nucleic Acids Res **39**(Database issue): D141-145.

Gazzinelli, G., C. C. de Oliveira, E. A. Figueiredo, L. H. Pereira, P. M. Coelho and J. Pellegrino (1973). "*Schistosoma mansoni*: biochemical evidence for morphogenetic change from cercaria to schistosomule." Exp Parasitol **34**(2): 181-188.

GeneDB (2011). *Schistosoma mansoni* GeneDB, Wellcome Trust Sanger Institute.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, . . . J. Zhang (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.

Gobert, G. N. (2010). "Applications for profiling the schistosome transcriptome." Trends Parasitol **26**(9): 434-439.

Gobert, G. N., D. J. Stenzel, M. K. Jones, D. E. Allen and D. P. McManus (1997). "*Schistosoma japonicum*: immunolocalization of paramyosin during development." Parasitology **114 (Pt 1)**: 45-52.

Gobert, G. N., M. H. Tran, L. Moertel, J. Mulvenna, M. K. Jones, D. P. McManus and A. Loukas (2010). "Transcriptional changes in *Schistosoma mansoni* during early schistosomula development and in the presence of erythrocytes." PLoS Negl Trop Dis **4**(2): e600.

Gonzalez, S., M. Flo, M. Margenat, R. Duran, G. Gonzalez-Sapienza, M. Grana, . . . C. Fernandez (2009). "A family of diverse Kunitz inhibitors from Echinococcus granulosus potentially involved in host-parasite cross-talk." PLoS One **4**(9): e7009.

Graefe, G., W. Hohorst and H. Drager (1967). "Forked tail of the cercaria of *Schistosoma mansoni*--a rowing device." Nature **215**(5097): 207-208.

Greenwood, K., T. Williams and T. Geary (2005). "Nematode neuropeptide receptors and their development as anthelmintic screens." Parasitology **131 Suppl**: S169-177.

Grossman, A. I., R. B. Short and G. D. Cain (1981). "Karyotype evolution and sex chromosome differentiation in Schistosomes (Trematoda, Schistosomatidae)." Chromosoma **84**(3): 413-430.

Guttman, M., M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, . . . A. Regev (2010). "Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs." Nat Biotechnol **28**(5): 503-510.

Haas, B. J., M. Berriman, H. Hirai, G. G. Cerqueira, P. T. Loverde and N. M. El-Sayed (2007). "*Schistosoma mansoni* genome: closing in on a final gene set." Exp Parasitol **117**(3): 225-228.

Haeberlein, S. and W. Haas (2008). "Chemical attractants of human skin for swimming *Schistosoma mansoni* cercariae." Parasitol Res **102**(4): 657-662.

Hall, S. L., S. Braschi, M. Truscott, W. Mathieson, I. M. Cesari and R. A. Wilson (2011). "Insights into blood feeding by schistosomes from a proteomic analysis of worm vomitus." Mol Biochem Parasitol **179**(1): 18-29.

Hansell, E., S. Braschi, K. F. Medzihradszky, M. Sajid, M. Debnath, J. Ingram, . . . J. H. McKerrow (2008). "Proteomic analysis of skin invasion by blood fluke larvae." PLoS Negl Trop Dis **2**(7): e262.

Hebenstreit, D., M. Fang, M. Gu, V. Charoensawan, A. van Oudenaarden and S. A. Teichmann (2011). "RNA sequencing reveals two major classes of gene expression levels in metazoan cells." Mol Syst Biol **7**: 497.

Hedstrom, R., J. Culpepper, R. A. Harrison, N. Agabian and G. Newport (1987). "A major immunogen in *Schistosoma mansoni* infections is homologous to the heat-shock protein Hsp70." J Exp Med **165**(5): 1430-1435.

Hillier, L. W., V. Reinke, P. Green, M. Hirst, M. A. Marra and R. H. Waterston (2009). "Massively parallel sequencing of the polyadenylated transcriptome of *C. elegans*." Genome Res **19**(4): 657-666.

Hobert, O. (2010). Neurogenesis in the nematode Caenorhabditis elegans. WormBook**:** 1-24.

Hoffmann, K. F. and D. W. Dunne (2003). "Characterization of the Schistosoma transcriptome opens up the world of helminth genomics." Genome Biol **5**(1): 203.

Hoffmann, K. F., D. A. Johnston and D. W. Dunne (2002). "Identification of *Schistosoma mansoni* gender-associated gene transcripts by cDNA microarray profiling." Genome Biol **3**(8): RESEARCH0041.

Holmfeldt, P., K. Brannstrom, M. E. Sellin, B. Segerman, S. R. Carlsson and M. Gullberg (2007). "The *Schistosoma mansoni* protein Sm16/SmSLP/SmSPO-1 is a membrane-binding protein that lacks the proposed microtubule-regulatory activity." Mol Biochem Parasitol **156**(2): 225-234.

Horemans, A. M., A. G. Tielens and S. G. van den Bergh (1991). "The transition from an aerobic to an anaerobic energy metabolism in transforming *Schistosoma mansoni* cercariae occurs exclusively in the head." Parasitology **102 Pt 2**: 259-265.

Hotez, P. J., J. M. Bethony, D. J. Diemert, M. Pearson and A. Loukas (2010). "Developing vaccines to combat hookworm infection and intestinal schistosomiasis." Nat Rev Microbiol **8**(11): 814-826.

Howells, R. E., F. J. Ramalho-Pinto, G. Gazzinelli, C. C. de Oliveira, E. A. Figueiredo and J. Pellegrino (1974). "*Schistosoma mansoni*: mechanism of cercarial tail loss and its significance to host penetration." Exp Parasitol **36**(3): 373-385.

Humphries, J. E., M. J. Kimber, Y. W. Barton, W. Hsu, N. J. Marks, B. Greer, . . . T. A. Day (2004). "Structure and bioactivity of neuropeptide F from the human parasites *Schistosoma mansoni* and Schistosoma japonicum." J Biol Chem **279**(38): 39880-39885.

Hunter, S., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, . . . C. Yeats (2009). "InterPro: the integrative protein signature database." Nucleic Acids Res **37**(Database issue): D211-215.

Illumina. The Genome Analyzer IIx(2011) from http://www.illumina.com/.

Ismail, M., A. Metwally, A. Farghaly, J. Bruce, L. F. Tao and J. L. Bennett (1996). "Characterization of isolates of *Schistosoma mansoni* from Egyptian villagers that tolerate high doses of praziquantel." Am J Trop Med Hyg **55**(2): 214-218.

Ismail, M. M., S. A. Taha, A. M. Farghaly and A. S. el-Azony (1994). "Laboratory induced resistance to praziquantel in experimental schistosomiasis." J Egypt Soc Parasitol **24**(3): 685-695.

Johnson, P. J., J. M. Kooter and P. Borst (1987). "Inactivation of transcription by UV irradiation of T. brucei provides evidence for a multicistronic transcription unit including a VSG gene." Cell **51**(2): 273-281.

Jolly, E. R., C. S. Chin, S. Miller, M. M. Bahgat, K. C. Lim, J. DeRisi and J. H. McKerrow (2007). "Gene expression patterns during adaptation of a helminth parasite to different environmental niches." Genome Biol **8**(4): R65.

Jones, M. K., W. Yang and D. P. McManus (2001). "Immunolocalization of the 38.3 kDa calponin-like protein in stratified muscles of the tail of *Schistosoma japonicum* cercariae." Parasitol Int **50**(2): 129-133.

King, C. H. (2010). "Parasites and poverty: the case of schistosomiasis." Acta Trop **113**(2): 95-104.

Kolev, N. G., J. B. Franklin, S. Carmi, H. Shi, S. Michaeli and C. Tschudi (2010). "The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution." PLoS Pathog **6**(9): e1001090.

Krause, M. and D. Hirsh (1987). "A trans-spliced leader sequence on actin mRNA in *C. elegans*." Cell **49**(6): 753-761.

Kusel, J. R., B. H. Al-Adhami and M. J. Doenhoff (2007). "The schistosome in the mammalian host: understanding the mechanisms of adaptation." Parasitology **134**(Pt 11): 1477-1526.

Laing, R., M. Hunt, A. V. Protasio, G. Saunders, K. Mungall, S. Laing, . . . J. S. Gilleard (2011). "Annotation of Two Large Contiguous Regions from the *Haemonchus contortus* Genome Using RNA-seq and Comparative Analysis with Caenorhabditis elegans." PLoS One **6**(8): e23216.

Langmead, B., C. Trapnell, M. Pop and S. L. Salzberg (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.

Lawson, J. R. and R. A. Wilson (1980). "Metabolic changes associated with the migration of the schistosomulum of *Schistosoma mansoni* in the mammal host." Parasitology **81**(2): 325-336.

Letunic, I., T. Doerks and P. Bork (2009). "SMART 6: recent updates and new developments." Nucleic Acids Res **37**(Database issue): D229-232.

Lorenzo, C., G. Salinas, A. Brugnini, C. Wernstedt, U. Hellman and G. Gonzalez-Sapienza (2003). "*Echinococcus granulosus* antigen 5 is closely related to proteases of the trypsin family." Biochem J **369**(Pt 1): 191-198.

Loverde, P. T. (1998). "Do antioxidants play a role in schistosome host-parasite interactions?" Parasitol Today **14**(7): 284-289.

Mansour, N. R. and Q. D. Bickle (2010). "Comparison of microscopy and Alamar blue reduction in a larval based assay for schistosome drug screening." PLoS Negl Trop Dis **4**(8): e795.

Mardis, E. R. (2008). "Next-generation DNA sequencing methods." Annu Rev Genomics Hum Genet **9**: 387-402.

Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens and Y. Gilad (2008). "RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays." Genome Res **18**(9): 1509-1517.

Matsumoto, J., K. Dewar, J. Wasserscheid, G. B. Wiley, S. L. Macmil, B. A. Roe, . . . K. E. Hastings (2010). "High-throughput sequence analysis of *Ciona intestinalis* SL trans-spliced mRNAs: alternative expression modes and gene function correlates." Genome Res **20**(5): 636-645.

McKerrow, J. H. (2003). "Invasion of skin by schistosome cercariae: some neglected facts: Response from James J. McKerrow." Trends Parasitol **18**(2): 66-68.

McKerrow, J. H. and J. Salter (2002). "Invasion of skin by Schistosoma cercariae." Trends Parasitol **18**(5): 193-195.

McVeigh, P., G. R. Mair, L. Atkinson, P. Ladurner, M. Zamanian, E. Novozhilova, . . . A. G. Maule (2009). "Discovery of multiple neuropeptide families in the phylum Platyhelminthes." Int J Parasitol **39**(11): 1243-1252.

Melman, S. D., M. L. Steinauer, C. Cunningham, L. S. Kubatko, I. N. Mwangi, N. B. Wynn, . . . E. S. Loker (2009). "Reduced susceptibility to praziquantel among naturally occurring Kenyan isolates of *Schistosoma mansoni*." PLoS Negl Trop Dis **3**(8): e504.

Miller, P. and R. A. Wilson (1980). "Migration of the schistosomula of *Schistosoma mansoni* from the lungs to the hepatic portal system." Parasitology **80**(2): 267-288.

Molina, M. D., A. Neto, I. Maeso, J. L. Gomez-Skarmeta, E. Salo and F. Cebria (2011). "Noggin and noggin-like genes control dorsoventral axis regeneration in planarians." Curr Biol **21**(4): 300-305.

Morgan, J. A., R. J. Dejong, G. O. Adeoye, E. D. Ansa, C. S. Barbosa, P. Bremond, . . . E. S. Loker (2005). "Origin and diversification of the human parasite *Schistosoma mansoni*." Mol Ecol **14**(12): 3889-3902.

Morozova, O. and M. A. Marra (2008). "Applications of next-generation sequencing technologies in functional genomics." Genomics **92**(5): 255-264.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.

Murphy, W. J., K. P. Watkins and N. Agabian (1986). "Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for trans splicing." Cell **47**(4): 517-525.

Nanduri, J., J. E. Dennis, T. L. Rosenberry, A. A. Mahmoud and A. M. Tartakoff (1991). "Glycocalyx of bodies versus tails of *Schistosoma mansoni* cercariae. Lectin-binding, size, charge, and electron microscopic characterization." J Biol Chem **266**(2): 1341-1347.

Neumann, S., E. Ziv, F. Lantner and I. Schechter (1992). "Cloning and sequencing of an hsp70 gene of *Schistosoma mansoni*." Mol Biochem Parasitol **56**(2): 357-360.

Neumann, S., E. Ziv, F. Lantner and I. Schechter (1993). "Regulation of HSP70 gene expression during the life cycle of the parasitic helminth *Schistosoma mansoni*." Eur J Biochem **212**(2): 589-596.

Ning, Z., A. J. Cox and J. C. Mullikin (2001). "SSAHA: a fast search method for large DNA databases." Genome Res **11**(10): 1725-1729.

Olson, P. D. (2008). "Hox genes and the parasitic flatworms: new opportunities, challenges and lessons from the free-living." Parasitol Int **57**(1): 8-17.

Osley, M. A. (1991). "The regulation of histone synthesis in the cell cycle." Annu Rev Biochem **60**: 827-861.

Otto, T. D., G. P. Dillon, W. S. Degrave and M. Berriman (2011). "RATT: Rapid Annotation Transfer Tool." Nucleic Acids Res **39**(9): e57.

Otto, T. D., D. Wilinski, S. Assefa, T. M. Keane, L. R. Sarry, U. Bohme, . . . M. Llinas (2010). "New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq." Mol Microbiol **76**(1): 12-24.

Overington, J. P., B. Al-Lazikani and A. L. Hopkins (2006). "How many drug targets are there?" Nat Rev Drug Discov **5**(12): 993-996.

Parker-Manuel, S. J., A. C. Ivens, G. P. Dillon and R. A. Wilson (2011). "Gene Expression Patterns in Larval *Schistosoma mansoni* Associated with Infection of the Mammalian Host." PLoS Negl Trop Dis **5**(8): e1274.

Pearce, E. J. and A. S. MacDonald (2002). "The immunobiology of schistosomiasis." Nat Rev Immunol **2**(7): 499-511.

Protasio, A. V., I. J. Tsai, A. Babbage, S. Nichol, M. Hunt, M. A. Aslett, . . . M. Berriman (2012). "A Systematically Improved High Quality Genome and Transcriptome of the Human Blood Fluke *Schistosoma mansoni*." PLoS Negl Trop Dis **6**(1): e1455.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Rajkovic, A., R. E. Davis, J. N. Simonsen and F. M. Rottman (1990). "A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*." Proc Natl Acad Sci U S A **87**(22): 8879-8883.

Ram, D., Z. Grossman, A. Markovics, A. Avivi, E. Ziv, F. Lantner and I. Schechter (1989). "Rapid changes in the expression of a gene encoding a calcium-binding protein in *Schistosoma mansoni*." Mol Biochem Parasitol **34**(2): 167-175.

Ram, D., B. Romano and I. Schechter (1994). "Immunochemical studies on the cercarial-specific calcium binding protein of *Schistosoma mansoni*." Parasitology **108 (Pt 3)**: 289-300.

Ramalho-Pinto, F. J., G. Gazzinelli, R. E. Howells, T. A. Mota-Santos, E. A. Figueiredo and J. Pellegrino (1974). "*Schistosoma mansoni*: defined system for stepwise transformation of cercaria to schistosomule in vitro." Exp Parasitol **36**(3): 360-372.

Robinson, M. D. and A. Oshlack (2010). "A scaling normalization method for differential expression analysis of RNA-seq data." Genome Biol **11**(3): R25.

Salter, J. P., K. C. Lim, E. Hansell, I. Hsieh and J. H. McKerrow (2000). "Schistosome invasion of human skin and degradation of dermal elastin are mediated by a single serine protease." J Biol Chem **275**(49): 38667-38673.

Samueleson, J. C. and L. D. Stein (1989). "*Schistosoma mansoni*: increasing saline concentration signals cercariae to transform to schistosomula." Exp Parasitol **69**(1): 23-29.

Samuelson, J. C. and J. P. Caulfield (1985). "The cercarial glycocalyx of *Schistosoma mansoni*." J Cell Biol **100**(5): 1423-1434.

Samuelson, J. C., J. P. Caulfield and J. R. David (1982). "Schistosomula of *Schistosoma mansoni* clear concanavalin A from their surface by sloughing." J Cell Biol **94**(2): 355-362.

Santos, T. M., D. A. Johnston, V. Azevedo, I. L. Ridgers, M. F. Martinez, G. B. Marotta, . . . S. D. Pena (1999). "Analysis of the gene expression profile of *Schistosoma mansoni* cercariae using the expressed sequence tag approach." Mol Biochem Parasitol **103**(1): 79-97.

Sayed, A. A., A. Simeonov, C. J. Thomas, J. Inglese, C. P. Austin and D. L. Williams (2008). "Identification of oxadiazoles as new drug leads for the control of schistosomiasis." Nat Med **14**(4): 407-412.

Severin, A. J., J. L. Woody, Y. T. Bolon, B. Joseph, B. W. Diers, A. D. Farmer, . . . R. C. Shoemaker (2010). "RNA-Seq Atlas of Glycine max: A guide to the soybean transcriptome." BMC Plant Biol **10**(1): 160.

Shendure, J. (2008). "The beginning of the end for microarrays?" Nat Methods **5**(7): 585-587.

Sigrist, C. J., L. Cerutti, E. de Castro, P. S. Langendijk-Genevaux, V. Bulliard, A. Bairoch and N. Hulo (2010). "PROSITE, a protein domain database for functional characterization and annotation." Nucleic Acids Res **38**(Database issue): D161-166.

Simpson, A. J., A. Sher and T. F. McCutchan (1982). "The genome of *Schistosoma mansoni*: isolation of DNA, its size, bases and repetitive sequences." Mol Biochem Parasitol **6**(2): 125-137.

Skelly, P. J. and R. Alan Wilson (2006). "Making sense of the schistosome surface." Adv Parasitol **63**: 185-284.

Skelly, P. J., L. D. Stein and C. B. Shoemaker (1993). "Expression of *Schistosoma mansoni* genes involved in anaerobic and oxidative glucose metabolism during the cercaria to adult transformation." Mol Biochem Parasitol **60**(1): 93-104.

Spieth, J., G. Brooke, S. Kuersten, K. Lea and T. Blumenthal (1993). "Operons in *C. elegans*: polycistronic mRNA precursors are processed by trans-splicing of SL2 to downstream coding regions." Cell **73**(3): 521-532.

Springer, J. E., R. D. Azbill and S. L. Carlson (1998). "A rapid and sensitive assay for measuring mitochondrial metabolic activity in isolated neural tissue." Brain Res Brain Res Protoc **2**(4): 259-263.

Stein, L. D. and J. R. David (1986). "Cloning of a developmentally regulated tegument antigen of *Schistosoma mansoni*." Mol Biochem Parasitol **20**(3): 253-264.

Steinmann, P., J. Keiser, R. Bos, M. Tanner and J. Utzinger (2006). "Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk." Lancet Infect Dis **6**(7): 411-425.

Stirewalt, M. (1978). "Quantitative collection and proteolytic activity of preacetabular gland enzyme (s) of cercariae of *Schistosoma mansoni*." Am J Trop Med Hyg **27**(3): 548-553.

Stirewalt, M. A. (1974). "*Schistosoma mansoni*: cercaria to schistosomule." Adv Parasitol **12**: 115-182.

Stirewalt, M. A., D. R. Minnick and W. A. Fregeau (1966). "Definition and collection in quantity of schistosomules of *Schistosoma mansoni*." Trans R Soc Trop Med Hyg **60**(3): 352-360.

Stirewalt, M. A. and A. Uy (1969). "*Schistosoma mansoni*: cercarial penetration and schistosomule collection in an in vitro system." Exp Parasitol **26**(1): 17-28.

Sutton, R. E. and J. C. Boothroyd (1986). "Evidence for trans splicing in trypanosomes." Cell **47**(4): 527-535.

R Development Core Team. " R: A language and environment for statistical computing. " (2011). R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Tenniswood, M. P. and A. J. Simpson (1982). "The extraction, characterization and in vitro translation of RNA from adult *Schistosoma mansoni*." Parasitology **84**(Pt 2): 253-261.

TheSchistoVac, 2009. The targeted development of a new generation vaccine for schistosomiasis. http://ec.europa.eu/research/health/infectious-diseases/neglected-diseases/projects/011_en.htmlfrom.

TSjGS. The *Schistosoma japonicum* Genome Sequencing and Functional Annotation Consortium, (2009). "The Schistosoma japonicum genome reveals features of host-parasite interplay." Nature **460**(7253): 345-351.

Tielens, A. G. (1994). "Energy generation in parasitic helminths." Parasitol Today **10**(9): 346-352.

Trapnell, C., L. Pachter and S. L. Salzberg (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics **25**(9): 1105-1111.

Trapnell, C., B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, M. J. van Baren, . . . L. Pachter (2010). "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation." Nat Biotechnol **28**(5): 511-515.

Uniprot Consortium. (2009). "The Universal Protein Resource (UniProt) 2009." Nucleic Acids Res **37**(Database issue): D169-174.

Untergasser, A., H. Nijveen, X. Rao, T. Bisseling, R. Geurts and J. A. Leunissen (2007). "Primer3Plus, an enhanced web interface to Primer3." Nucleic Acids Res **35**(Web Server issue): W71-74.

Vale, R. D. and R. A. Milligan (2000). "The way things move: looking under the hood of molecular motor proteins." Science **288**(5463): 88-95.

van der Werf, M. J., S. J. de Vlas, S. Brooker, C. W. Looman, N. J. Nagelkerke, J. D. Habbema and D. Engels (2003). "Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa." Acta Trop **86**(2-3): 125-139.

Verjovski-Almeida, S., R. DeMarco, E. A. Martins, P. E. Guimaraes, E. P. Ojopi, A. C. Paquola, . . . E. Dias-Neto (2003). "Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*." Nat Genet **35**(2): 148-157.

Verjovski-Almeida, S., T. M. Venancio, K. C. Oliveira, G. T. Almeida and R. DeMarco (2007). "Use of a 44k oligoarray to explore the transcriptome of *Schistosoma mansoni* adult worms." Exp Parasitol **117**(3): 236-245.

Vermeire, J. J., A. S. Taft, K. F. Hoffmann, J. M. Fitzpatrick and T. P. Yoshino (2006). "*Schistosoma mansoni*: DNA microarray gene expression profiling during the miracidium-to-mother sporocyst transformation." Mol Biochem Parasitol **147**(1): 39-47.

Volfovsky, N., B. J. Haas and S. L. Salzberg (2003). "Computational discovery of internal micro-exons." Genome Res **13**(6A): 1216-1221.

Wallace, A., M. E. Filbin, B. Veo, C. McFarland, J. Stepinski, M. Jankowska-Anyszka, . . . R. E. Davis (2010). "The nematode eukaryotic translation initiation factor 4E/G complex works with a trans-spliced leader stem-loop to enable efficient translation of trimethylguanosine-capped RNAs." Mol Cell Biol **30**(8): 1958-1970.

Wang, L., Y. L. Li, Z. Fishelson, J. R. Kusel and A. Ruppel (2005). "*Schistosoma japonicum* migration through mouse skin compared histologically and immunologically with *S. mansoni.*" Parasitol Res **95**(3): 218-223.

Wang, Z., M. Gerstein and M. Snyder (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet **10**(1): 57-63.

Working to overcome the global impact of neglected tropical diseases - WHO/NTD report update(2011)http://www.who.int/neglected_diseases/2010report/2011_update_report/en/index.html

Wilson, R. A. and P. S. Coulson (2009). "Immune effector mechanisms against schistosomiasis: looking for a chink in the parasite's armour." Trends Parasitol **25**(9): 423-431.

Wilson, R. A. and J. R. Lawson (1980). "An examination of the skin phase of schistosome migration using a hamster cheek pouch preparation." Parasitology **80**(2): 257-266.

Wixon, J. and D. Kell (2000). "The Kyoto encyclopedia of genes and genomes--KEGG." Yeast **17**(1): 48-55.

Xia, Z., H. Xu, J. Zhai, D. Li, H. Luo, C. He and X. Huang (2011). "RNA-Seq analysis and de novo transcriptome assembly of *Hevea brasiliensis*." Plant Mol Biol.

Zamanian, M. (2011). Personal Communicatio to A. V. Protasio - Subclassficiation of G-protein couple receptors in *Schisotosma mansoni*. Department of Biomedical Sciences and Neuroscience Program, Iowa State University, Ames, Iowa, United States of America.

Zdobnov, E. M. and R. Apweiler (2001). "InterProScan--an integration platform for the signature-recognition methods in InterPro." Bioinformatics **17**(9): 847-848.