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.

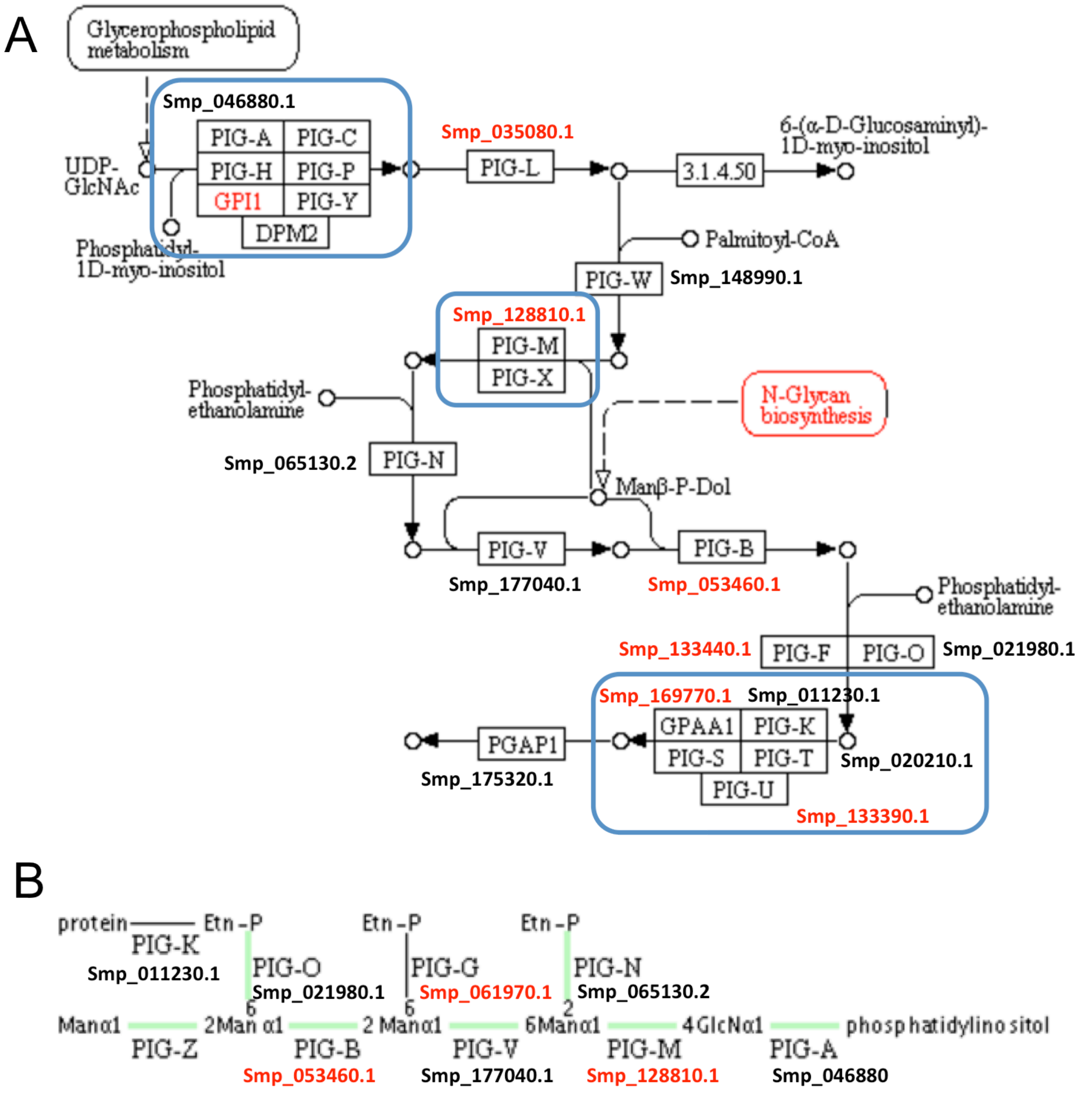

Figure 6.1 – Pathway map of glycosylphosphatidylinositol-anchored protein (GPI-APs) biosynthesis obtained from KEGG database (Wixon *et al.*, 2000). Where orthologs could be found (BLASTp e-value < 1e^{-10}), *S. mansoni* gene identification names have been added. Red names represent genes whose transcripts 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.

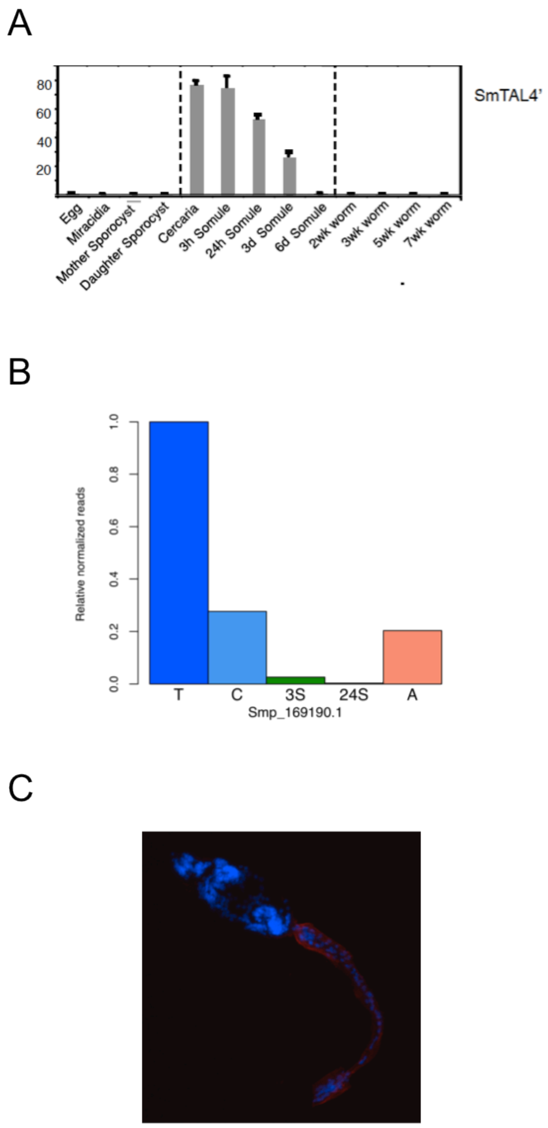

Figure 6.2 – Expression profile of SmTAL4. A – Relative expression retrieved from microarray data (Fitzpatrick *et al.*, 2009) for SmTAL4. Note the high level of expression in the 3- and 24-hours old schistosomula. B – RNA-seq relative normalized expression of SmTAL4. T, tail; C, cercariae; 3S, 3-hours 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)