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.

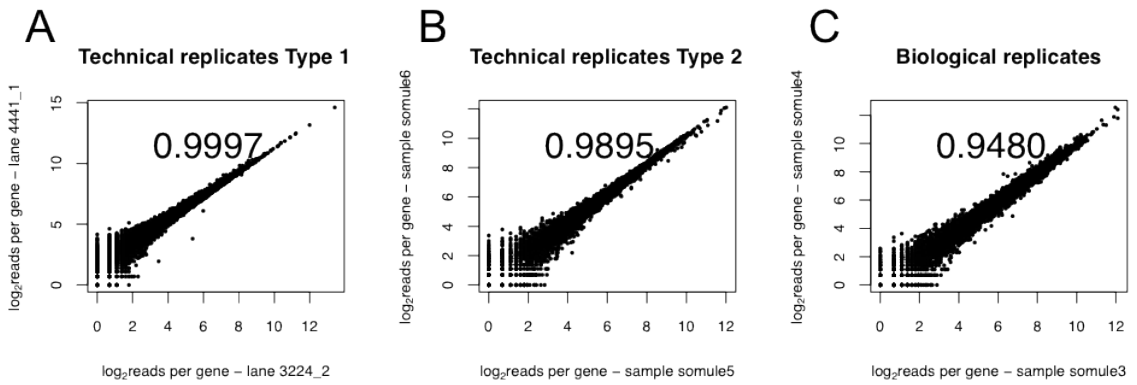


Figure 3.1 - Technical and biological replicates. Scatter plots of reads per gene obtained for pairs of replicates. Axes are in the logarithmic scale. A - Technical replicate type 1 – control for sequencing procedure. The same library (somule1) was sequenced twice (lanes 4444_1 and 3224_2). B – Technical 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).

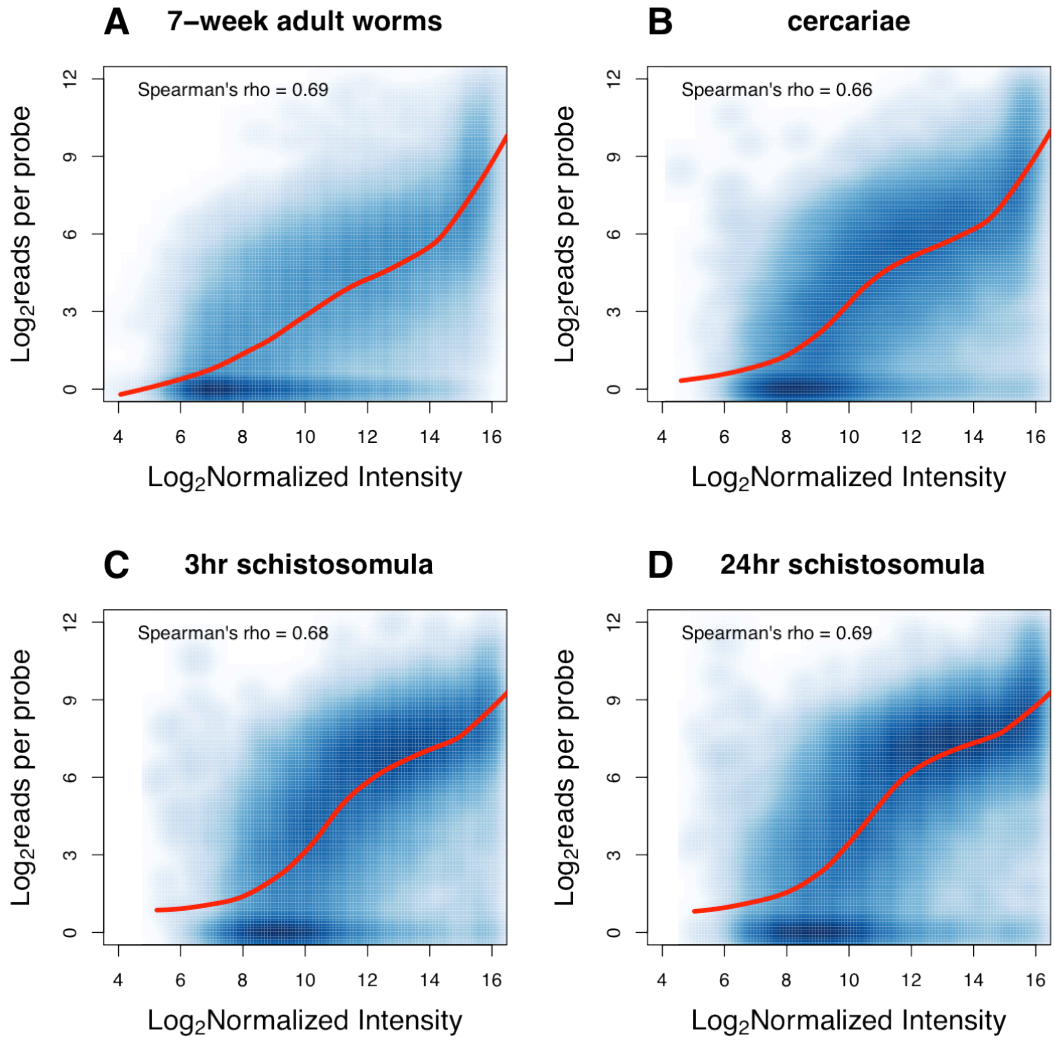


Figure 3.2 – Correlation of RNA-seq expression data with microarray from Fitzpatrick *et al.,* (2009). Each blue dot represents a position in the genome for which microarray expression data (x axis) and RNA-seq expression data (y axis) were calculated. A – 7-week adult worms; B – cercariae; C – 3-hour 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.

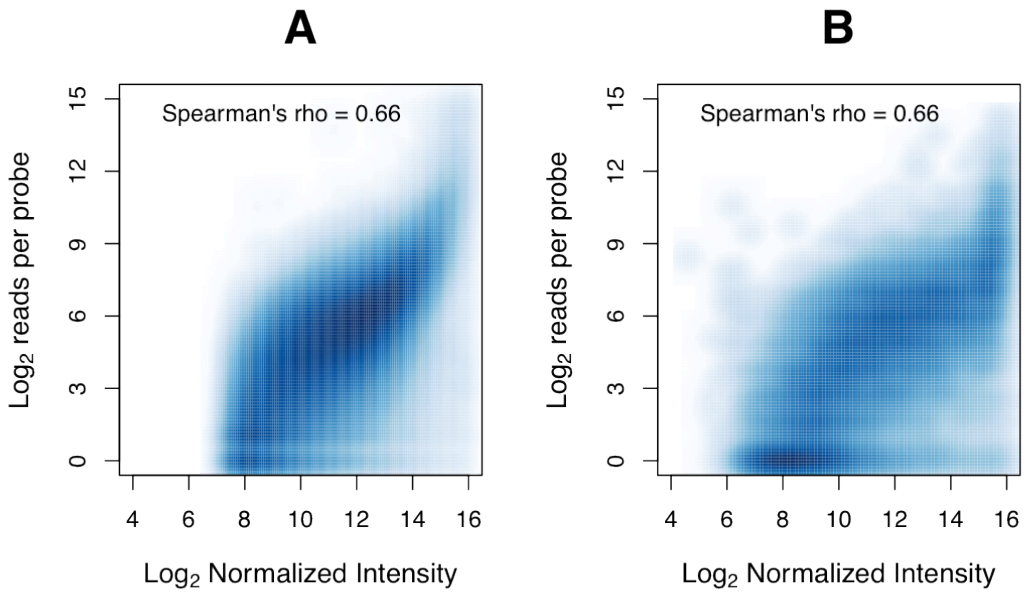


Figure 3.3 – Correlation of RNA-seq expression data for the cercariae sample against two different microarray platforms. Each blue dot represents a position in the genome for which microarray expression data (x axis) and RNA-seq expression data (y axis) were calculated. A – Microarray data from Parker-Manuel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

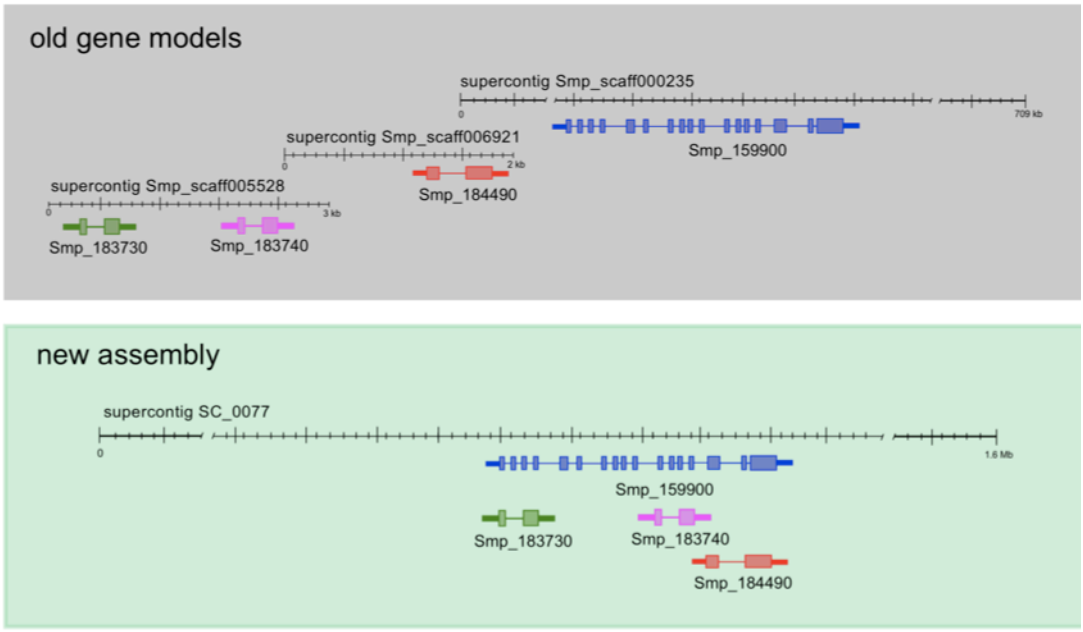


Figure 3.4 - Removal of assembly redundancies produces a more reliable set of gene models [reproduced from (Protasio *et al.*, 2012)]. Gene models were migrated from the previous version of the genome using RATT (Otto *et al.*, 2011). In the new version, many scaffolds converged into one region and hence 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.

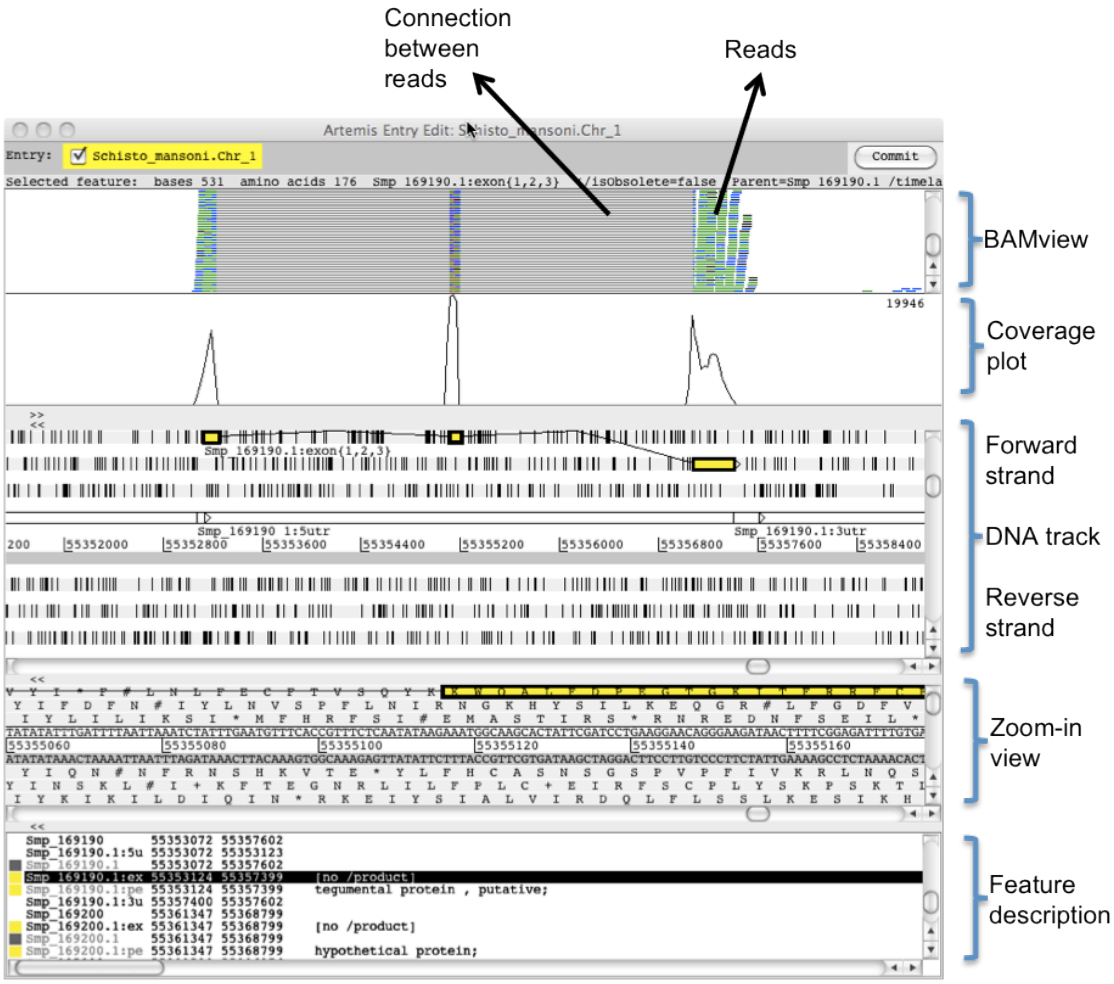


Figure 3.5 – Explanation of Artemis and BamView. Visualization. In this view of Artemis (Carver *et al.*, 2008) and BamView (Carver *et al.*, 2010), the working window is split in five sections. The top most is a graphical representation of the reads' alignment file (BAM file). Reads are represented 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.

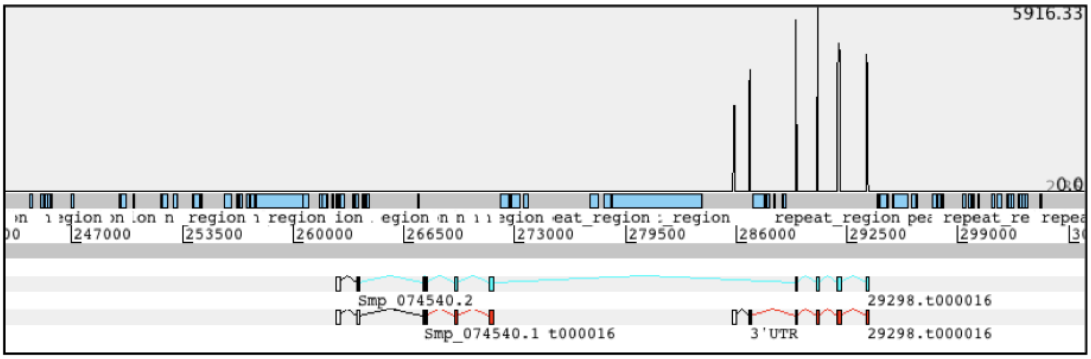


Figure 3.6 –RNA-seq data is used to split gene models. The previous version of model Smp_074540 (blue) and the two current genes (red) are shown. The previous gene model was split based on the differential expression of its 5' end region compared to the 3' end. Both new genes are expressed in this transcriptome 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.

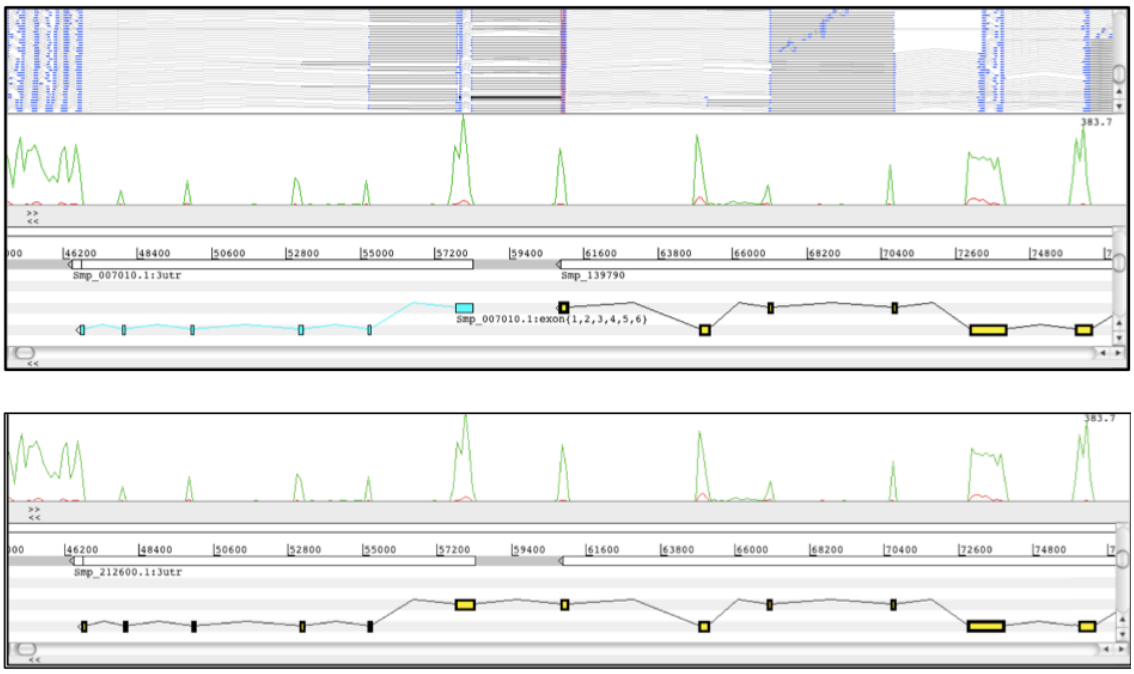


Figure 3.7 – Merging of gene models based on RNA-seq data. Top panel: two gene models (blue and yellow) are shown where RNA-seq data suggest they both belong to the same RNA molecule. Reads spanning the intergenic region between the two gene models (black line) provide evidence that the two transcripts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

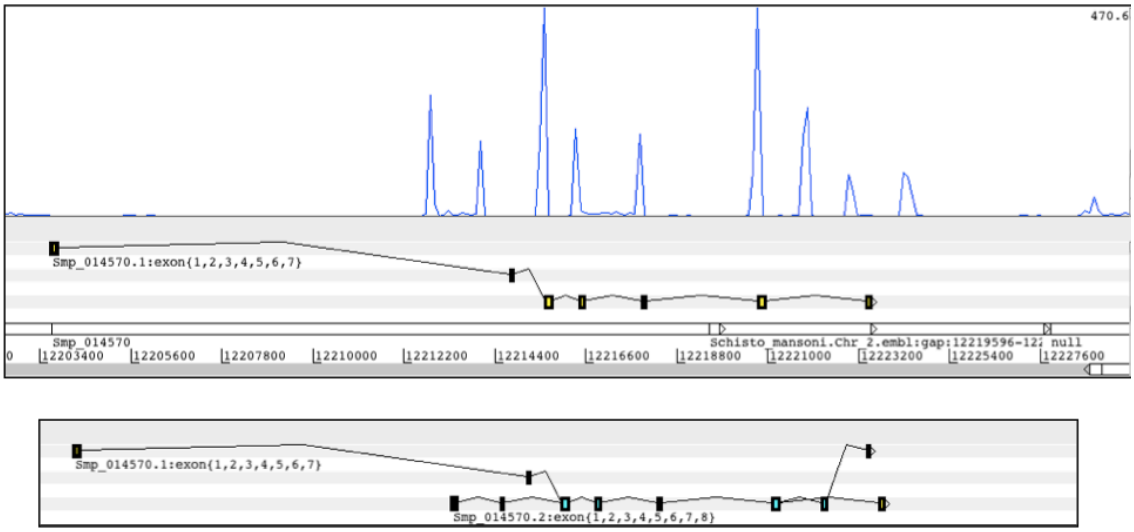


Figure 3.8 – RNA-seq data is used to find new exons and provides evidence of an alternative splicing form of gene Smp_014570. Top panel: the annotated model as a 7-exon transcript with its first exon located relatively far apart form the rest of the coding region. RNA-seq data supports the presence of 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.

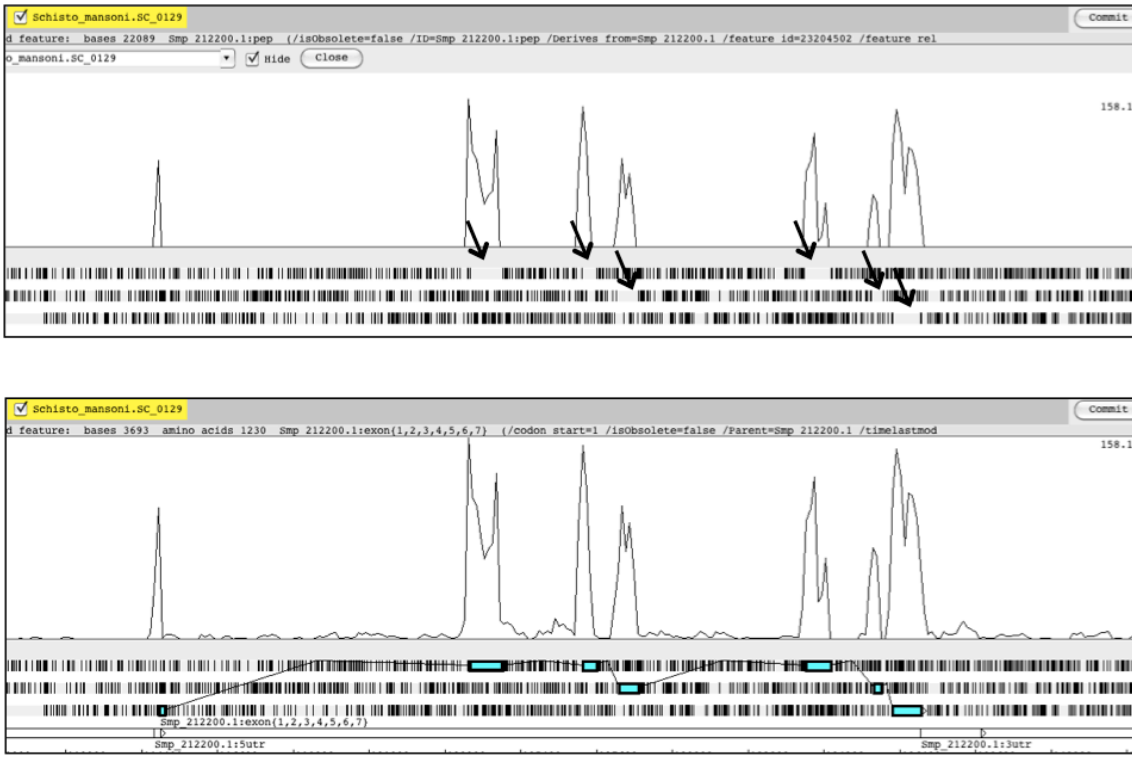


Figure 3.9 -RNA-seq data can reveal new genes. Top panel: each peak of expression corresponds to an exon and ORFs (arrows) are found in the amino-acids track suggesting this is indeed a coding gene. Bottom panel: resulting gene model.

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.

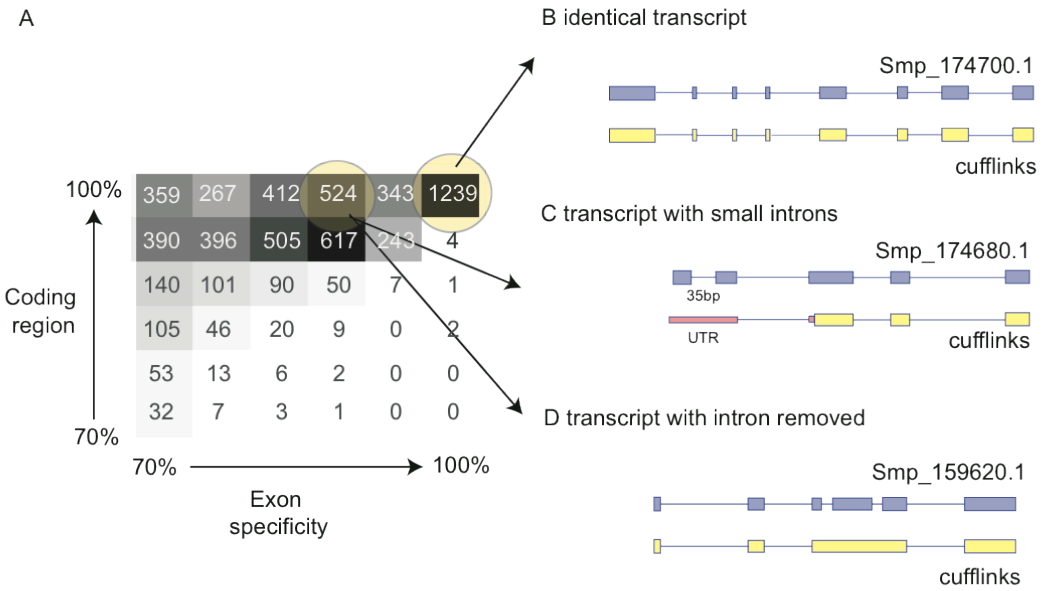


Figure 3.10 – Example scenarios of Cufflinks' models compared with previous gene models [reproduced from (Protasio *et al.*, 2012)]. A - Heatmap displaying comparisons between previous gene models and transcript fragments generated from Cufflinks. For each model, the extent of coding region that overlaps 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.

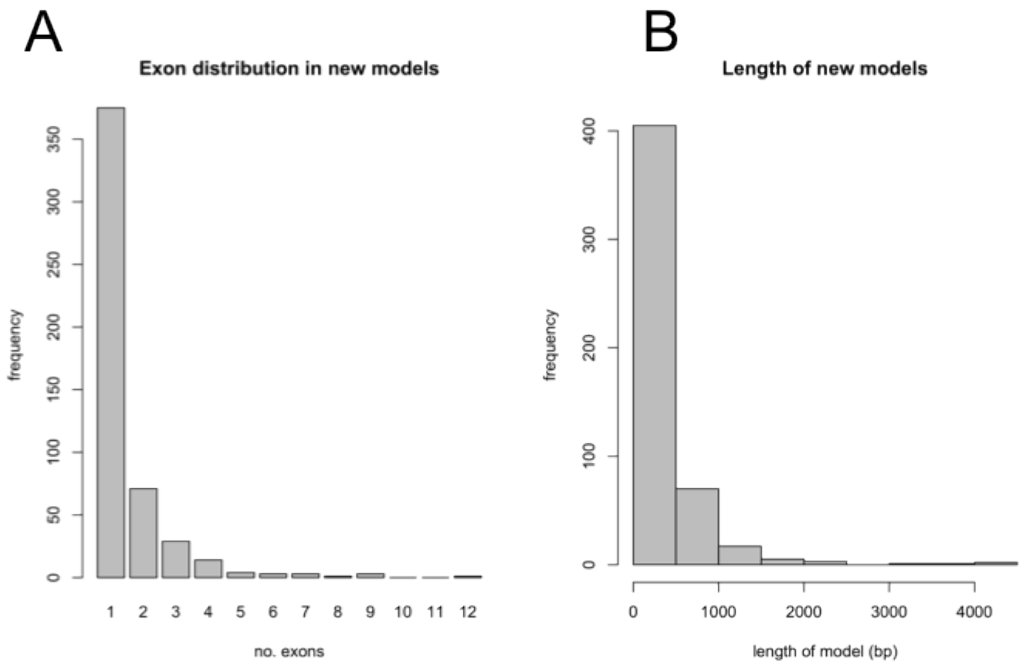


Figure 3.11 – Characteristics of the 504 new genes identified by Cufflinks and based on RNA-seq data. A – Exons distribution of new models. B – Length of the coding regions of new models.

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

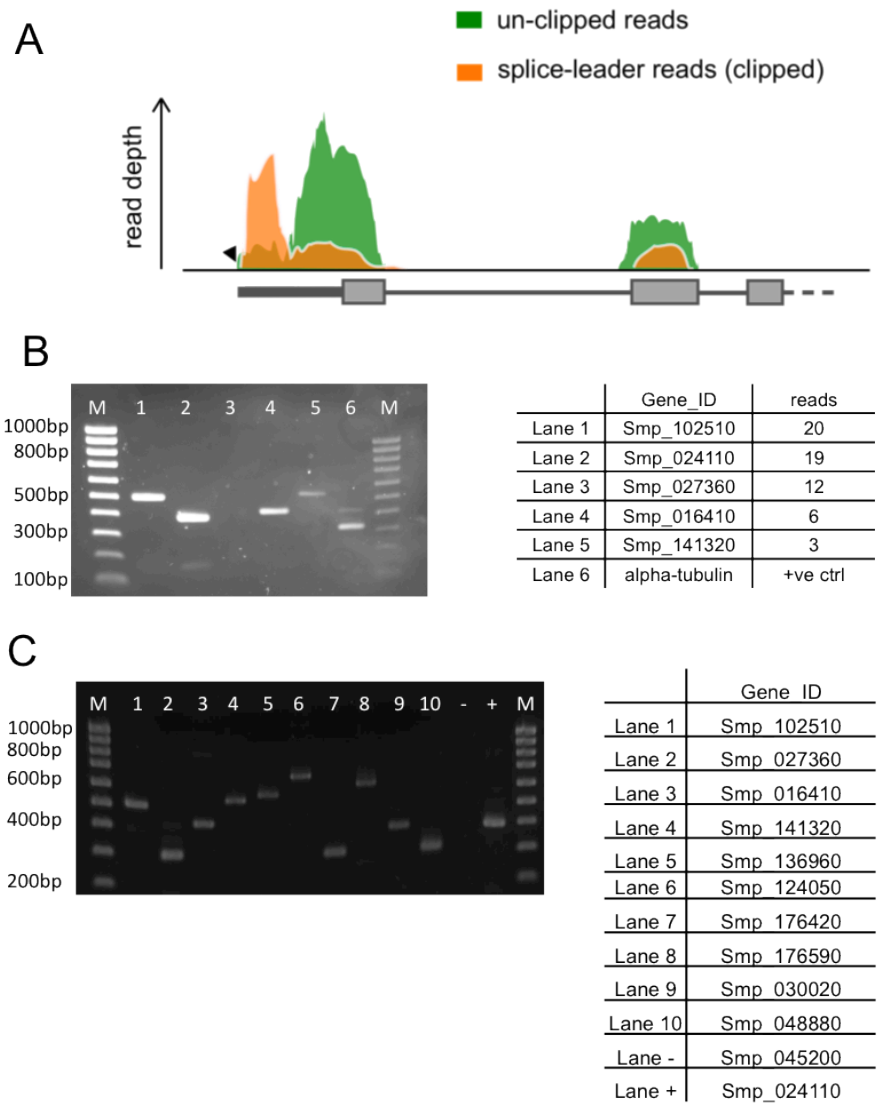
Trans-splicing is a mechanism where two RNA molecules are combined to form a mature RNA. In the case of splice leader (SL) *trans*-splicing, one of the RNA molecules involved is a small nuclear ribonucleoprotein commonly referred to as SL-RNA. The *trans*-splicing process is similar to that of *cis*-splicing 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.

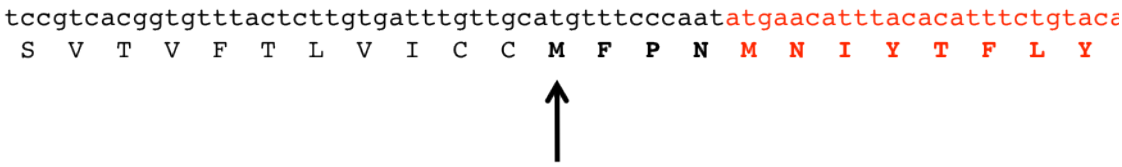


Figure 3.13 – *Trans*-splicing can affect the translation start site of a transcript. *In silico* translation of the PCR product corresponding to the *trans*-spliced form of Smp_084890 indicates that the ATG codon in the SL sequence (arrow) is found in-frame with main ORF of the transcript suggesting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

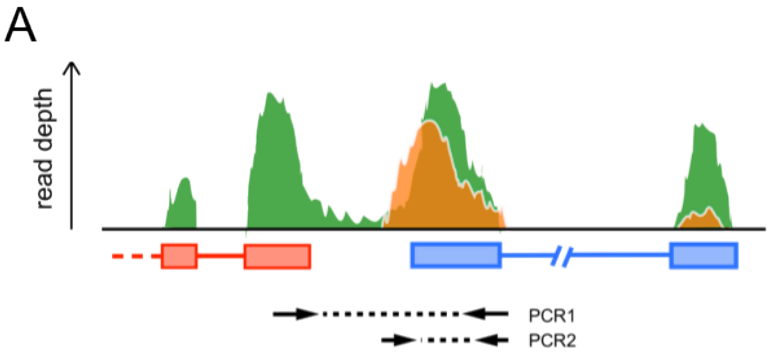
Because intergenic regions within polycistron are short (usually 200 nt but can be up to 2 kb), it is possible to use this information together with the available *trans*-splicing data to identify putative polycistrons in *S. mansoni.* To this end, intergenic distances[1] between genes were calculated. 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.

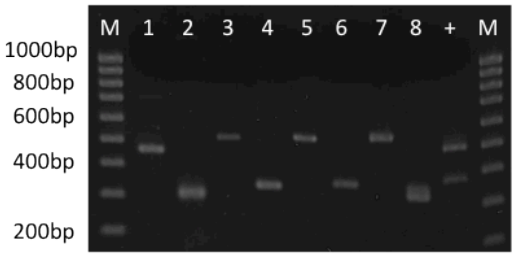
Unlike *C. elegans*, which uses a second spliced leader (SL2) to resolve polycistrons (Spieth *et al.*, 1993) or a tightly control combination of both (Allen *et al.*, 2011), *S. mansoni* seems to use the same SL for both polycistronic- and non-polycistronic *trans*-spliced transcripts. A secondary *trans*-splicing 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.