Poster No. A4.43

The *Schistosoma mansoni* Genome

John Gamble[1], Martin Aslett[1], Brian Haas[2], Al Ivens[1], Paul Mooney[1], Zemin Ning[1], Francisco Prosdocimi[1], Najib El-Sayed[2] and Matt Berriman[1]
[1]Pathogen Sequencing Unit, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK.
[2]The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA.

1. Introduction

Schistosomiasis is a chronic debilitating disease affecting over 200 million people worldwide. The causative agent is the parasitic flatworm, *Schistosoma mansoni*. Although much is known about the biology of *S. mansoni*, there is currently no effective vaccine, and the few chemotherapeutic agents available have met with limited success. The schistosome genome sequencing project was launched with the aim, *inter alia*, of identifying new targets for therapeutic intervention, and is being run as a collaboration between the Wellcome Trust Sanger Institute (WTSI; www.sanger.ac.uk) and The Institute for Genomic Research (TIGR; www.tigr.org).

2. *S. mansoni* genome assembly and gene prediction

S. mansoni has a haploid genome of 270 MB carried on seven pairs of autosomes and one pair of sex chromosomes (female = ZW, male = ZZ). Chromosomes range in size from 18 to 73 MB, and can be distinguished by size, shape and C banding.

A total of 2,720,243 sequencing reads of *S. mansoni* DNA were assembled into 19,032 scaffolds. Repetitive sequences were masked using RepeatScout, and the masked genome then used for gene prediction. Gene models were constructed by combining information from *de novo* gene prediction and sequence similarity methods.

The annotation process is laborious and is complicated by the large number of repeat elements in the schistosome genome (30-40%), as well as by the complex splicing patterns of many of the genes.

De novo gene prediction (Augustus, GlimmerHMM)

Spliced protein alignments

EST/cDNA sequences

17,243 predicted genes

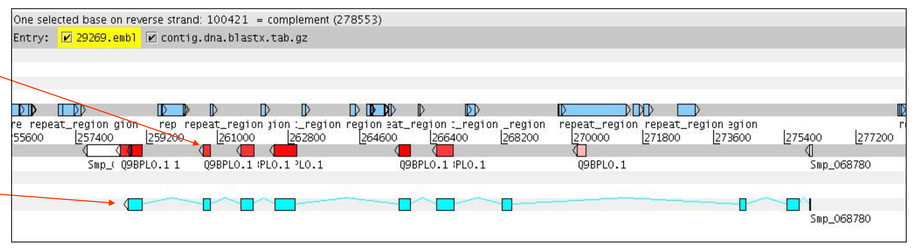
3. Confirmation of predicted gene models

The accuracy of the gene predictions was tested by comparing the published sequences of 100 known *S. mansoni* genes with their corresponding predictions. This is demonstrated below for the *S. mansoni* FTZ-F1 gene (Q9BPL0). The predicted gene is shown in light blue, and the published sequence is shown in red. In this example, the gene model predicts an extra three exons at the 5' end of the gene, as well as an extended 3' exon.

72 predicted genes, out of the 100 examined, were found to match their corresponding published sequences exactly.

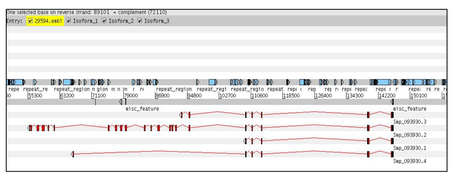
Published gene

Predicted gene

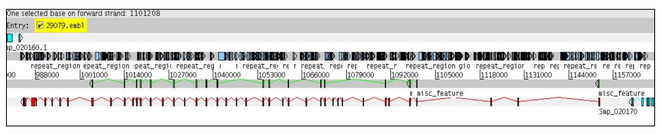


4. Examples of *S. mansoni* gene architecture

Many of the *S. mansoni* genes were found to possess a complex architecture, having numerous small exons and large introns.



The EGF receptor (Q26566), showing four distinct isoforms.

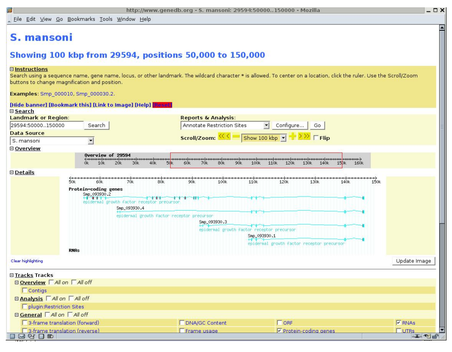


High voltage-activated calcium channel Cav2a (Q963L8) has 40 exons, and is the largest gene so far annotated in *S. mansoni*. The coding sequence alone spans 168,622 bp, and encodes a protein of 2203 amino acids.

5. *S. mansoni* genome assembly in GeneDB

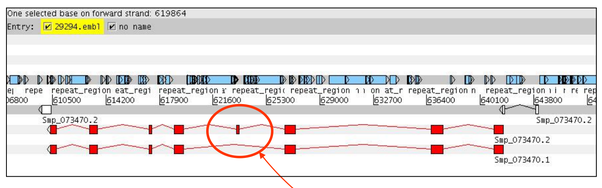
The *S. mansoni* genome sequence and all annotated gene models can be accessed from GeneDB (www.genedb.org).

Specific genes and their products can be searched for using GeneDB, and viewed within a genomic context using the integral genome browser, GBrowse, as shown opposite.



6. Applications of genomic sequence data

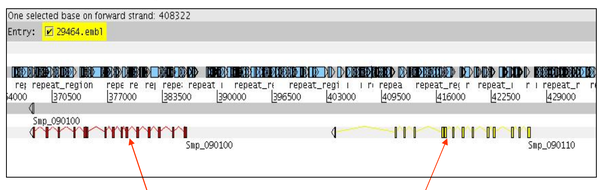
Fig 1. Prediction of a novel SmRXR-2 isoform



Novel SmRXR-2 isoform has no DNA-binding domain

SmRXR-2 is a member of the nuclear receptor superfamily[2,3]. These proteins act as transcriptional regulators by forming homo- / heterodimers which bind to regulatory DNA sequences within their target genes. Our annotation pipeline predicts a new SmRXR-2 isoform that lacks the DNA-binding domain. This isoform would be able to dimerize as normal, but the resulting dimers would be unable to bind DNA and hence would be unable to affect gene transcription. This suggests a mechanism by which SmRXR-2, and possibly other nuclear receptors, may be regulated.

Fig 2. Prediction of a novel protease – possibly involved in entry of cercariae to human host

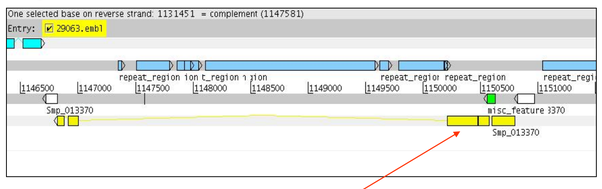


SmPepM8 metalloprotease

SmPepM8-related novel protease.

Schistosomiasis is contracted through direct skin penetration by *S. mansoni* cercariae. This process is mediated by cercarial secretions that digest the outer layers of the skin. A novel metalloprotease (SmPepM8) was recently discovered in secretions of *S. mansoni* cercariae[1], and is believed to facilitate entry of cercariae into the human host. Aligning the amino acid and genomic DNA sequences (BlastX) revealed not only the architecture of the SmPepM8 gene, but also the presence of a second closely related protease gene.

Fig 3. Potential novel vaccine candidates.



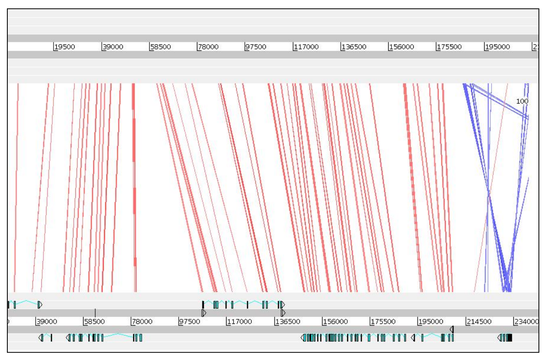
Predicted novel tetraspanin gene. (TmHMM-predicted transmembrane domain in green.)

Tetraspanins are a family of integral membrane proteins that show sequence similarity to surface receptors of B and T cells and are believed to play a role in cell-cell interactions and maintenance of membrane structure. Recently, two members of this family – TSP-1 and TSP-2 – were shown to be promising candidate vaccines against schistosomiasis[4]. Our automatic annotation pipeline has revealed the existence of five additional putative tetraspanin family members in *S. mansoni*, one of which is shown here. These new family members may also warrant testing for their ability to generate an anti-schistosome immune response.

7. Comparative genomics

The Artemis Comparison Tool (ACT) allows sequence comparisons between large genomic sequences to be visualized intuitively, together with their associated annotation[5]. In the analyses shown here, corresponding regions of repeat-masked *S. mansoni* and *S. japonicum* DNA were compared using TBlastX and the output visualised in ACT. Coloured bands, representing conserved open reading frames, allowed identification of conserved genes between the two species of schistosome (figs 4 and 5). In addition, the method can reveal the presence of additional exons in *S. mansoni* genes not predicted by the first-pass annotation (fig. 5). Another possible application of this method would be to confirm the order of contigs in each genome assembly using BlastN as the comparison file generator.

Fig 4. Comparing genome sequences.



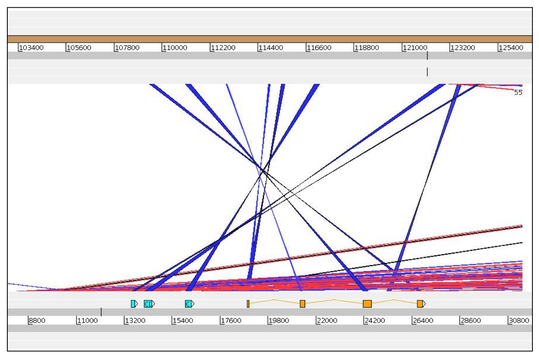
S. japonicum

S. mansoni

Possible uses of genome comparison:

1. Transferring annotation of related genes from *S. mansoni* to *S. japonicum*.
2. Ordering assembly contigs

Fig 5. Extending gene models



S. japonicum

S. mansoni

Comparing corresponding regions of genomic DNA can confirm and extend the annotation of existing gene models. The exons shown in light blue were not predicted in first pass annotation.

References

1. Curwen, R. et al. (2006). Mol. Cell. Proteomics 5: 835-844.
2. Freebern, W. et al. (1999). Gene 233: 33-38.
3. de Mendonca, RL. et al. (2000). Eur. J. Biochem. 267: 3208-3219.
4. Tran, MH. et al. (2006). Nature Medicine 12(7): 835-840.
5. Carver, TJ. et al. (2005). Bioinformatics 21(16): 3422-3423.

Acknowledgements

The *Schistosoma mansoni* genome sequencing project is funded by The Wellcome Trust (WTSI core funding) and NIAID (PI: Phil LoVerde).

The sequence of the SmPepM8 metalloprotease was kindly provided by Alan Wilson and Rachel Curwen, Department of Biology, University of York, PO Box 373, York, YO10 5YW, UK.

Schistosoma japonicum genome sequence was provided by Shengyue Wang and Guoping Zhao, of the Chinese National Human Genome Center, Shanghai, China.

The authors wish to acknowledge the contribution of the numerous staff at WTSI and TIGR who contributed to the sequencing and preliminary analysis of the *S. mansoni* genome.