The *Hymenolepis* genome and transcriptome

Magdalena Zarowiecki[1,2], Alejandro Sanchez-Flores[1], Natalia Pouchkina-Stantcheva[2], Nancy Holroyd[1], Matt Berriman[1] & Peter D. Olson[2]

[1]The Wellcome Trust Sanger Institute, Cambridge, UK
[2]The Natural History Museum, London, UK

mz3@sanger.ac.uk

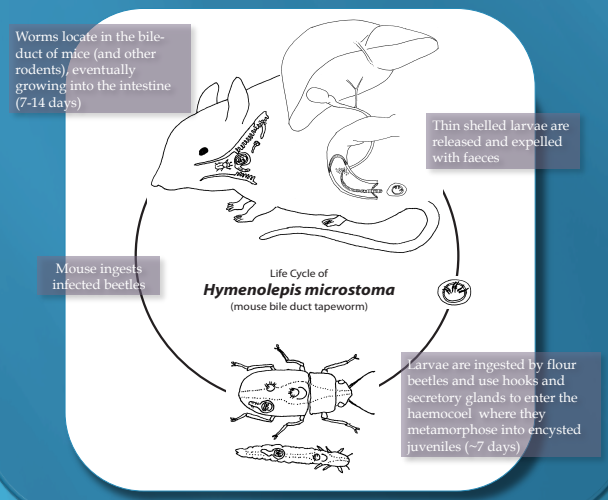
Plathelminthes (from the Greek πλατύ, *platy*, meaning "flat" and ἕλμινς (root: ἑλμινθ-), *helminth-*, meaning worm)

Introduction

In 2009, the Parasite Genomics Group at Wellcome Trust Sanger Institute began characterization of the genome and transcriptome of the model tapeworm *Hymenolepis microstoma* in collaboration with Pete Olson of the Natural History Museum, London.

Why *Hymenolepis*?

- *Hymenolepis* species (e.g. *H. diminuta*, *H. nana* and *H. microstoma*) have been used as laboratory models since the 1950s, and thus much of our basic understanding of tapeworm biology stems from work on these species.
- The entire life cycle can be kept in the lab, making them more practical models than groups having medical or veterinary importance, such as *Taenia* and *Echinococcus*.
- The highly inbred 'Nottingham' strain is expected to show reduced variability, resulting in fewer assembly problems.

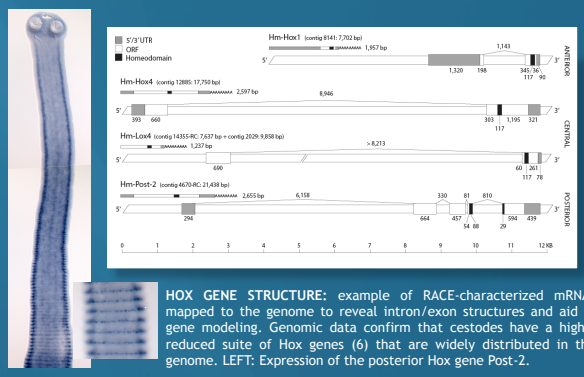


Results

Estimated Genome Size = 242 MB

Genome assembly	Solexa	454	All data
Total length	164,454,905	144,216,446	140,918,059
No. of contigs	477,156	68,588	13,772
N50	9611	3707	31,535
No of contigs >N50	4366	10767	1290
Mean contig length	345	2103	10,232
Longest contig	119,539	62,308	189,659

Transcriptome	ADULT		LARVAE	
Paired-end seq	6.3E+7		4.8E+7	
Reads mapped to genome	4.8E+7	76%	3.7E+7	76%
Reads paired	3.0E+7	48%	3.2E+7	45%
Perfect reads mapped	3.9E+7	63%	3.0E+7	63%



HOX GENE STRUCTURE: example of RACE-characterized mRNAs mapped to the genome to reveal intron/exon structures and aid in gene modeling. Genomic data confirm that cestodes have a highly reduced suite of Hox genes (6) that are widely distributed in the genome. LEFT: Expression of the posterior Hox gene Post-2.

Materials & Methods

Data derive from a laboratory strain of *H. microstoma* (ie. 'Nottingham') maintained *in vivo* at the NHM using outbred conventional mice and flour beetles (*Tribolium confusum*). Specimens were removed from the bile ducts and intestines of mice and genomic DNA extracted from somatic tissues (ie. anterior parts of worm). RNA was extracted from whole adult worms as well as from larvae during mid-metamorphosis in the beetle host. Representing development during both phases of the life cyle, these samples give insights into the transcriptome during both segmentation and larval metamorphosis.

The genome was assembled from 5 full Roche 454 Titanium runs (3 un-paired, 2 paired with ~3 Kbp inserts) and 3 Illumina GAII lanes, with a read-length of 76 bp. Illumina insert sizes for 2 lanes range from 300-400 bp, and for the third ~3000 bp. A *de novo* assembly of the genome was made using the software Newbler 2.3 (for Roche/454) and ABySS 1.2.0 (for Illumina), and contigs then merged using minimus2 from the AMOS Pipeline. The transcriptome was sequenced using separate lanes of Illumina data (ie. adult vs. larva), and mapped to the genome using BWA after screening for contaminants such as host RNA/DNAs. Gene models are currently being constructed using Augustus, SNAP and Jig-saw.

Discussion

After the first year of the project we have generated 42x coverage of the genome and comparisons with similar amounts of data from *Echinococcus* show that the *Hymenolepis* assembly is highly efficient. These data will provide a platform for further research in all areas of tapeworm biology, including development and drug targets.

The finished project will result in:

- A high-quality whole genome-assembly
- Transcriptomic data from multiple life-stages
- Annotated gene-predictions for a majority of genes, based on transcriptomic data and gene-prediction algorithms.

The latest genome assembly is available from:
http://www.sanger.ac.uk/Projects/Pathogens/

Acknowledgements

All primary data generation and bioinformatic infrastructure sponsored by the Sanger Institute. Additional funding for MZ thanks to a CoSyst grant to PDO, MB & MZ which is a joint UK Research Council initiative. NPS and PDO are supported by a BBSRC grant to PDO.