The mitochondrial genome of *Paragonimus westermani* (Kerbert, 1878), the lung fluke representative of the family Paragonimidae (Trematoda)

Among helminth parasites, *Paragonimus* (zoonotic lung fluke) gains considerable importance from veterinary and medical point of view because of its diversified effect on its host. Nearly fifty species of *Paragonimus* have been described throughout the globe. It is estimated that more than 20 million people are infected worldwide and the best known species is *Paragonimus westermani*, whose type locality is probably India and which infects millions of people in Asia causing disease symptoms that mimic tuberculosis. Human infections occur through eating raw crustaceans containing metacercarie or ingestion of uncooked meat of paratenic hosts such as pigs. Though the fluke is known to parasitize a wide range of mammalian hosts representing as many as eleven families, the status of its prevalence, host range, pathogenic manifestations and its possible survivors in nature from where the human beings contract the infection is not well documented in India. We took advantage of the whole genome sequence data for *P. westermani*, generated by Next Generation Sequencing, and its comparison with the existing data for the *P. westermani* complete mitochondrial genome sequence to design precise and specific primers for amplification of mitochondrial genome sequences from the parasite DNA sample. The Ion torrent next generation sequencing platform was harnessed to completely sequence the mitochondrial genome, and applied innovative approaches to bioinformatically assemble and annotate it. A strategic PCR primer design utilizing the whole genome sequence data from *P. westermani* enabled us to design specific primers capable of amplifying all regions of the mitochondrial genome from *P. westermani*. Assembly of NGS data from libraries enriched in mtDNA sequence by PCR gave rise to a total of 11 contigs spanning the entire 14.7 kb mt DNA sequence of *P. westermani* available at NCBI. We conducted gap-filling by traditional Sanger sequencing to fill in the gaps. Annotation of non-protein coding genes successfully identified tRNA regions for the 24 tRNAs coded in mtDNA and 12 protein coding genes. Bayesian phylogenetic analyses of the concatenaed protein coding genes placed *P. westermani* within the family Opisthorchida. The complete mtDNA sequence of *P. westermani* is 15004 base pair long; the lung fluke is the major etiological agent of paragonimiasis and the first Indian represenative for the family Paragonimidae to be fully sequenced that provides important genetic markers for ecological, population and biogeographical studies and molecular diagnostic of digeneans that cause trematodiases.

The mitochondrial genome of *Paragonimus westermani* (Kerbert, 1878), the lung fluke representative of the family Paragonimidae (Trematoda)

Devendra K Biswal[1], Anupam Chatterjee[2], Alok Bhattacharya[3,*], Veena Tandon[1,4,*]

[1]Bioinformatics Centre, North-Eastern Hill University, Shillong 793022, Meghalaya, India

[2]Department of Biotechnology and Bioinformatics, North-Eastern Hill University, Shillong 793022, Meghalaya, India

[3]School of Life Sciences, Jawaharlal Nehru University, New Delhi 110067, India

[4]Department of Zoology, North-Eastern Hill University, Shillong 793022, Meghalaya, India

[*]Corresponding author

Email addresses:

DKB: devbioinfo@gmail.com

AC: chatterjeeanupam@hotmail.com

AB: alok.bhattacharya@gmail.com

VT: tandonveena@gmail.com

Abstract

Background

Among helminth parasites, *Paragonimus* (zoonotic lung fluke) gains considerable importance from veterinary and medical point of view because of its diversified effect on its host. Nearly fifty species of *Paragonimus* have been described throughout the globe. It is estimated that more than 20 million people are infected worldwide and the best known species is *Paragonimus westermani*, whose type locality is probably India and which infects millions of people in Asia causing disease symptoms that mimic tuberculosis. Human infections occur through eating raw crustaceans containing metacercarie or ingestion of uncooked meat of paratenic hosts such as pigs. Though the fluke is known to parasitize a wide range of mammalian hosts representing as many as eleven families, the status of its prevalence, host range, pathogenic manifestations and its possible survivors in nature from where the human beings contract the infection is not well documented in India. We took advantage of the whole genome sequence data for *P. westermani*, generated by Next Generation Sequencing, and its comparison with the existing data for the *P. westermani* complete mitochondrial genome sequence to design precise and specific primers for amplification of mitochondrial genome sequences from the parasite DNA sample. The Ion torrent next generation sequencing platform was harnessed to completely sequence the mitochondrial genome, and applied innovative approaches to bioinformatically assemble and annotate it.

Results

A strategic PCR primer design utilizing the whole genome sequence data from *P. westermani* enabled us to design specific primers capable of amplifying all regions of the mitochondrial genome from *P. westermani*. Assembly of NGS data from libraries enriched in mtDNA sequence by PCR gave rise to a total of 11 contigs spanning the entire 14.7 kb mt DNA sequence of *P. westermani* available at NCBI. We conducted gap-filling by traditional Sanger sequencing to fill in the gaps. Annotation of non-protein coding genes successfully identified tRNA regions for the 24 tRNAs coded in mtDNA and 12 protein coding genes. Bayesian phylogenetic analyses of the concatenaed protein coding genes placed *P. westermani* within the family Opisthorchida.

Conclusions

The complete mtDNA sequence of *P. westermani* is 15004 base pair long; the lung fluke is the major etiological agent of paragonimiasis and the first Indian represenative for the family Paragonimidae to be fully sequenced that provides important genetic markers for ecological, population and biogeographical studies and molecular diagnostic of digeneans that cause trematodiases.

Introduction

Among about 50 known species of the genus *Paragonimus*, *Paragonimus westermani*, one of the causative agents of paragonimiasis, was first described as early as in 1878 and is the most well-known species within the genus *Paragonimus* because of its wide geographical distribution and medical importance (Blair, XU & Agatsuma, 1999). Typically, paragonimiasis is a disease of the lungs and pleural cavity but extra-pulmonary paragonimiasis also happens to be an important clinical manifestation. It is a neglected disease that has received feeble attention from public health authorities. As per the recent estimates, about 293 million people are at risk, while several millions are infected worldwide (Keiser & Utzinger, 2009). However, this may be an underestimate as there are still many places where the disease burden has yet to be assessed. There has been an increased recognition of the public health importance of paragonimiasis and other foodborne trematodiases in recent times (Fried, Graczyk & Tamang, 2004) and some serious concern for *Paragonimus* species outside endemic areas owing to the risk of infection through food habits in today's globalized food supply. In the case of paragonimiasis, this resurgence of interest can partly be attributed to the common diagnostic confusion of paragonimiasis with tuberculosis, as symptoms of the former closely mimic those of the latter, thereby leading to an inappropriate treatment being administered especially in areas where both tuberculosis and paragonimiasis co-occur and create overlapping health issues (Toscano *et al*., 1995). The state-of-the-art molecular biology techniques, next generation sequencing (NGS) technology and their rapid development in contemporary times may provide additional tools for the differential identification of digenean trematode infections to overcome limitations of current morphology-based diagnostic methods. Owing to their high nucleotide substitution rates, parasitic flatworm mitochondrial (mt) genomes have become very popular markers for diagnostic purposes and for resolving their phylogenetic relationships at different taxonomic ranks. Comparative mitochondrial genomics can provide more reliable results and reveal important informations of mtDNA architectural features such as gene order and structure of non-coding regions.

In our present study, we determined the complete mtDNA nucleotide sequence of *P. westermani*, which was collected from several sites in Changlang District, Arunachal Pradesh in India, using NGS data generated from total genomic DNA extracts. Phylogenetic analyses were carried out using a supermatrix of all the concatenated mt sequences of 12 protein-coding genes of digenean trematode and cestodes, (taking nematode species as an outgroup) available in public domain (GenBank). This newly sequenced Indian isolate *P. westermani* mt genome sequence along with the one in the RefseQ database of NCBI would provide useful information on both genomics and Paragonimidae evolution, including the biogeographic status of the cryptic species of the lung flukes and other mtDNA sequences available for any member of the trematode group.

Methods

Parasite material and DNA extraction

Naturally infected freshwater edible crabs (*Barytelphusa lugubris lugubris*) were collected from Changlang District in Arunachal Pradesh (Altitude - 213 mASL, Longitude - 96°-15'N and Latitude - 27°-30'E). The isolation of Metacercariae from the crustacean host muscle tissues was carried out by digestion technique using artificial gastric juice. The 70% alcohol-fixed metacercariae were further processed for DNA extraction and PCR amplification. The lysed individual worms were subjected to DNA extraction by standard ethanol precipitation technique (Sambrook, Fitsch & Maniatis, 1989) and also extracted from the eggs on FTA cards with aid of Whatman's FTA Purification Reagent. DNA was subjected to a series of enzymatic reactions that repair frayed ends, phosphorylate the fragments, and add a single nucleotide 'A' overhang and ligate adaptors (Illumina's TruSeq DNA sample preparation kit). Sample cleanup was done using Ampure XP SPRI beads. After ligation, ~300–350 bp fragment for short insert libraries and ~500–550 bp fragment for long insert libraries were size-selected by gel electrophoresis, gel extracted and purified using Minelute columns (Qiagen). The libraries were amplified using 10 cycles of PCR for enrichment of adapter-ligated fragments. The prepared libraries were quantified using Nanodrop and validated for quality by running an aliquot on High Sensitivity Bioanalyzer Chip (Agilent). 2X KapaHiFiHotstart PCR ready mix (Kapa Biosystems Inc., Woburn, MA) reagent was used for PCR. The Ion torrent library was made using Ion Plus Fragment library preparation kit (Life Technologies, Carlsbad, US) and the Illumina library was constructed using TruSeqTM DNA Sample Preparation Kit (Illumina, Inc., US) reagents for library prep and TruSeq PE Cluster kit v2 along withTruSeq SBS kit v5 36 cycle sequencing kit (Illumina, Inc., US) for sequencing (Biswal et al., 2013).

Amplification, sequencing and assembly

Ion Torrent Reads

We sheared pooled PCR products to smaller sizes using Bioruptor. Ion Torrent library was constructed as per manufacturer's protocols. PCR products were sonicated, adapter ligated and amplified for x cycles to generate a library and subsequently were sequenced to generate reads of an average of 121 nt SE reads on Ion Torrent. The IonTorrent raw data was processed for 3' low quality bases trimming, and adapter contamination. Since the Ion Torrent data might have host contamination, the processed reads were then aligned to the reference sequence of *Paragonimus westermani* mtDNA (NC_002354) available in GenBank, Department of Environmental Health Science, Kochi Medical School, Oko, Nankoku, Kochi, Japan. The alignment was carried out using Tmap Ion Torrent proprietary tool. The mapped reads were extracted in fastq format using custom perl script. These clean reads were used for further bioinformatics analysis in this study. The processed reads as well as mito mapped reads were quality checked using Genotypic Technology Pvt. Ltd., proprietary tool SeqQC.

Illumina Reads

Illumina reads from our unpublished *P. westermani* whole genome data were mapped to *P.westermani* reference sequence (gi|23957831|ref|NC_002354.2|). The alignment was carried out using Bowtie aligner. The mapped reads were extracted in fastq format using custom perl script. We obtained 62874 paired end reads, which aligned to different intervals in the *P.westermani* mt genome, covering ~ 3 kb of the 15 kb mt genome (NC_002354.2). Accordingly, primers were designed at these regions, using sequence information from reference to ensure optimum primer design (Additional file 1). We conducted PCR using 10 ng of genomic DNA from *P. westermani* with the following PCR conditions: 10 ng of FD-2 DNA with 10 uM Primer mix in 10 ul reaction, PCR thermo cycling conditions – 98C for 3min, 35cycles of 98 °C for 30sec, 60°C for 30sec, 72 °C for 1min 30sec, final extension 72°C for 3 min and 4°C hold. We gel-eluted the bands (Additional file 1) corresponding to different products, pooled these products and proceeded for NGS library construction. These clean single end reads were also further used for bioinformatics analysis in this study. The illumina mito mapped reads were quality checked using Genotypic Technology Pvt. Ltd., proprietary tool SeqQC. The QC reads are outlined in Table 5.

De-novo Assembly

The ion torrent mapped reads were assembled using Newbler (Quinn *et al.,* 2008) software. The illumina mapped reads were subjected to reference assisted denovo assembly using velvet(Zerbino & Birney, 2008) assembler. Quite a few hash lengths were tested for velvetg. Hash length 65 gave the optimal results in terms of total contig length, N50, and maximum contig length. Therefore, k-mer 65 assembly was considered for further analysis. Sanger reads were also added in the final assembly. The draft sequence was generated using IonTorrent reads, Illumina reads, Sanger reads hybrid high-quality denovo assembly and subsequently the denovo-leftout regions were obtained using reference assisted assembly and consensus calling. Extensive manual curation work was carried out to produce the complete sequence. The complete sequence comprises 15004 bases in total. There were a few regions in the mitochondria, namely ~900 bases in the start and ~1500 bases in the end, where there were few or no sequences at 3x depth. In that case, the consensus sequence was retrieved using VCFtools (Danecek *et al.*, 2008). The consensus sequence was introduced at such regions; the sequences in question are represented with lower case of nucleotides, while the confident regions are represented in upper case in the fasta sequence file. Out of 15004 bases in the sequences, 13188 were confident bases (87.88% of the total), while 1818 bases were low quality bases (12.11% of the total). Mapping of assembled mitochondria against the reference was carried out using online Blastn (Altschul *et al.*, 1990). Blastn results show 85% identical bases between the two, with 99% query coverage with the best e-value possible of 0.0 and with maximum score of 12579. Artemis Comparison Tool (Carver *et* *al.,* 2005) (ACT) was used to generate visual output for mapping of the assembled mtDNA sequence against the reference mt genome (NC_002354).

In silico analysis for nucleotide sequence statistics, protein coding genes (PCGs) prediction, annotation and tRNA prediction

Sequences were assembled and edited by using CLC Genome Workbench V.6.02 with comparison to published flatworm genomes and the assembled whole single mtDNA contig was annotated with the aid of ORF finder tool at NCBI (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) and MITOS, which were subsequently used to search for homologous digenean trematode PWGs already housed in REFSEQ NCBI database (http://www.ncbi.nlm.nih.gov/refseq/) by using tBLASTn [Altschul et al., 1990). The program ARWEN (Laslett & Canbäck, 2008) was used to identify the tRNA genes by setting the search to predict secondary structures occasionally with very low Cove scores (<0.5) and, where necessary, also by restricting searches to find tRNAs lacking DHU arms (using the trematode tRNA option). Nucleotide codon usage for each protein-encoding gene was predicted using the program Codon Usage at (http://www.bioinformatics.org/sms2/codon_usage.html).The ORFs and codon usage profiles of PCGs were analyzed. The newly sequenced and assembled *P. westermani* mtDNA was annotated using MITOS and the output file was further used to sketch the newly sequenced genome with GenomeVX at http://wolfe.ucd.ie/GenomeVx/

Phylogenetic analysis

DNA sequences of the 12 protein-coding genes from 13 representative trematodes, cestode and nematode species were retrieved (Table 1), aligned in clustal w and concatenated using MESQUITE (Maddison & Maddison, 2011). The supermatrix was used for generating phylogenetic trees using Bayesian analysis in MrBayes v3.1 (Ronquist and Huelsenbeck, 2003). The mt genome sequence of the nematode *Ascaris suum* and *Ascaris lumbricoides* were used as an outgroup. For the nucleotide alignment, the GTR+I+G model was used and Bayesian analysis was run for 1,000,000 generations and sampled every 1000 generations. The first 25% of trees were omitted as burn-in and the remaining trees were used to calculate Bayesian posterior probabilities reataining the trees with a majority consensus rule of 50.

Results & Discussion

Mitochondrial genome organisation of *P. westermani* mtDNA

The two rRNA genes and 12 protein coding genes, typical of flatworms, were identfied by comparison of their sequence similarity and secondary structures with those of other flatworms.The mt genome lacks atp 8 with no characteristic amino acid signatures. Over a longtime gene order remains stable in animal mtDNAs [Boore, 1999; Saccone *et al*., 1999). Differences in the mtDNA gene order between members of the same family, though rare, can occur in higher taxonomic ranks. A marked difference in the gene order was found among the various trematode, cestode and nematode species as outlined in Fig. 1. The total length for the digenean *P. westermani* (AF219379) is 14,965 bp, and for *Schistosoma japonicum* (NC_002544) and *S. mansoni* (NC_002545) is approximately 14.5 kb as curated by the NCBI staff. Other digeneans possess small mt genomes. The mtDNA sequence of *P. westermani* (Bioproject accession number **PRJNA248332**, Biosample accession sample **SAMN02797822** and SRA **SRX550161**) is 15,004 bp in length and is well within the range of typical metazoan mtDNA sizes (14–18 kb). The mt genome of *P. westermani* is larger than that of other digenean species available in GenBank™ (http://www.ncbi.nlm.nih.gov/genbank/) to date (Table 1). It contains 12 protein-coding genes (cox1-3, nad1-6, nad4L, atp6 and cytb), 24 transfer RNA (tRNA) genes and 2 ribosomal RNA genes (rrnL and rrnS) (Fig. 2)(Table 2). The gene arrangement pact of protein-coding genes in *P. westermani* tallies with that of the *Fasciola hepatica* (Le *et al*., 2000; Le *et al*., 2001), *Opisthorchis felineus* (Shekhovtsov *et al*., 2010), *Fasciola gigantica* (Liu *et al*., 2014), *Fasciolopsis buski* (Biswal *et al.,* 2013), *Paramphistomum cervi* [Yan *et al*., 2013) mt genomes, but different from that seen in *Taenia* and *Ascaris* species (Nakao, Sako & Ito, 2003; Okimoto, Macfarlane & Wolstenholme, 1990) (Fig. 3). An overlapping region spanning nearly 40 bp between 3' nad4L end and nad4 5' end was also seen in *P.westermani*, a feature common to other digenean trematodes. The 12 protein coding genes and their blast hit protein plots are summarised in Fig. 4. The protein plot shows for each gene and each position the quality value if it is above the threshold and the different genes are differentiated with a range of colour codes. Basically, the initial hits used in MITOS [Bernt *et al*., 2013) correspond to the "mountains" in this plot that visualizes the signal from the BLAST (Altschul et al., 1990) searches. The arrows shown on the top of the plot depict the gene order annotation and the quality values are shown on a log scale.

Genetic Code, nucleotide composition and codon usage

It is a well established fact that mt DNA of parasitic flatworms uses AAA to specify ASN (Lys in the universal code), AGA and AGG to specify Ser (Arg in the universal code), and TGA to specify Trp (stop codon in the universal code). ATG is the usual start codon while GTG and other codons are also used as start codons (Le *et al*., 2002). The *P. westermani* mtDNA exhibited ATG and ATA as start codons and TAG and TAA as stop codons (Table 3). mtDNA genomes of invertebrates have a tendency to be AT-rich (Wolstenholme, 1992), a feature common in several parasitic flatworm protein coding genes. However, the nucleotide composition is not uniform among the species. For *Schistosoma mansoni,* the AT-rich percentage is 68.7%, whereas for *Fasciola hepatica* it is 63.5% AT and for *P. westermani* only 54.6% AT (Le *et al*., 2002). The nucleotide composition in the *P. westermani*, Indian isolate was biased towards G and T, which is similar to that of other digeneans, viz. *F. hepatica, O. felineus*, *C. sinensis*, *P. cervi* and unlike *S. japonicum* and other schistosomes, which are more biased towards A and T. The atomic composition in single stranded DNA exhibits Hydrogen with a frequency of 37.5%, Carbon 29.8%, Nitrogen 10.8 %, Oxygen 18.8 % and Phosphorus 3.0% (Table 4).

Transfer and ribosomal RNA genes section

A standard cloverleaf structure is generally seen for most of the tRNAs. There are exceptions that include tRNA(S), in which the paired dihydrouridine (DHU) arm is missing as in all parasitic flatworm species and tRNA(A), in which the paired DHU-arm is missing as in cestodes contrary to trematodes. Previous studies indicate structures for tRNA(C) that somewhat vary among the parasitic flatworms. In some species, a paired DHU-arm is missing (*Schistosoma mekongi* and cestodes), whereas it is present in others (*F.hepatica* and *F. buski*). It is noteworthy that the *P. westermani* Indian isolate exhibited 24 tRNA genes, 1 TV replacement loop tRNA genes and 2 D replacement loop tRNA genes. The tRNA GC range varied from 37.9% to 59.4%. (Fig. 5). Ribosomal large and small subunits in parasitic flatworms are unremarkable. They are smaller than those in most other metazoans but can be folded into a recognizable, conserved secondary structures (Le *et al*., 2001). The rrnL (16S ribosomal RNA) and rrnS (12S ribosomal RNA) genes of *P. westermani* were identified by sequence comparison with those of cloesly related trematodes and these ribosomal genes were separated by tRNA-C (GCA).

Non-coding regions

There are one or two longer non-coding region(s) (NR) in every genome comprising stable stem–loop structures that are associated with genome replication or repeat sequences. Previous studies report repeats in the NR of many animal mt genomes that may be an outcome of slippage-mismatching mechanisms (Le, Blair & McManus, 2001). In parasitic flatworms, NRs vary in length and complexity. The NR is divided by one or more tRNA genes into a SNR and a LNR in digenean trematodes. A common feature of LNRs is the presence of long repeats. In the present study the *P. westermani* mtDNA though didn't exhibit significant demarcation of LNR and SNR, there were regions with repeats with total number of 3158 variants with a total of 1722 SNPs and 1436 INDELS.

Phylogenetic analysis

Several genetic markers from nuclear rDNA regions and mtDNA of flukes have been employed in some systematic and population genetic studies of helminth parasites [7-14]. As of now the full-length mt genomes of 14 digenean, 34 cestode and 70 nematode species have been determined, characterized, and are published in GenBank. It is confirmed that alignments with more than 10,000 nucleotides from mtDNAs can provide ample information for phylogenetic resolution, hypothesis building and evolutionary interpretation of the major lineages of tapeworms. Use of complete mtDNA sequences for phylogenetic analyses are more reliable and informative (Waeschenbach, Webster & Littlewood, 2012). In the present study, a phylogenetic tree inferred from concatenated nucleotide sequences of the 12 protein-coding genes (shown in Fig. 2) is well supported by very high posterior probabilities (100%). Two large clades are visibly informative: one contains members of the Family Schistosomatidae, and the other includes members representing the sequence of families in order of increasingly derived status: Opisthorchiidae, Paragonimidae, Paramphistomidae and Fasciolidae (Trematoda); Ascarididae (Nematoda) and Taeniidae (Cestoda). This arra2ngement was seen in the tree based on nucleotide sequences, in which a clade containing Fasciolidae and Paragonimidae was strongly supported and *P. cervi* was sister to this clade. *P. westermani* claded with *Opisthorchis felineus* and *Clonorchis sinensis*. Members representing Taeniidae served as an outgroup (Fig. 3).

Conclusions

In this study, we took advantage of the whole genome sequence data for *P. westermani*, generated by NGS technology and its comparison to existing data for the *P. westermani* mitochondrial genome sequence for designing precise and specific primers for amplification of mitochondrial genome sequences from the parasite DNA sample. Here we present and discuss the complete sequence of the coding region of the mitochondrial genome of *P. westermani*, the Indian lung fluke isolate, which posesses the same gene order as that of other Digenea (Opisthorchidae and

Paramphistomatidae) and consists of 12 PCGs, 24 tRNAs and 2 rRNAs. There are long repetitive regions in the fluke that can serve as diagnostic markers with phylogenetic signals. The complete mtDNA sequence of *P. westermani* will add to the knowledge of digenean mitochondrial genomics and also provide an important resource for studies of inter- and intra-specific variations, biogepgraphic studies, heteroplasmy of the flukes belonging to Paragonimidae and a resource for comparative mitochondrial genomics and systematic studies of Digenea.

Availability of supporting data

Sequence reads have been deposited at the National Center for Biotechnology Information [Bioproject: PRJNA248332, Biosample : SAMN02797822 and SRA: SRX550161]

Authors' contributions

VT, AB and DKB conceived of the study and participated in its design, coordination and manuscript writing. DKB and AB performed the computational analysis and maintained the computer programs used for the analysis. VT and AC performed the molecular experiments associated with the parasite. All authors have read and approved the final manuscript.

Acknowledgements

We would like to acknowledge Dr. Sudip Ghatani, Department of Zoology, NEHU, Shillong for collecting the biosamples and M/s Genotypic Technologies, Bangalore, India for carrying out NGS sequencing for this project, especially the efforts of Dr. Deepti Saini for the primer design strategy.

References

1. 1 Altschul SF, Gish W, Miller W, Myers,E.W. & Lipman DJ. 1990. **Basic local alignment search tool.** *Journal of Molecular Biology* **215**:403-410

2. Biswal DK, Ghatani S, Shylla JA, Sahu R, Mullapudi N, Bhattacharya A, TandonV. 2013. **An integrated pipeline for next generation sequencing and annotation of the complete mitochondrial genome of the giant intestinal fluke, *Fasciolopsis buski* (Lankester, 1857) Looss, 1899.** PeerJ **1**:e207
3. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, Pütz J, Middendorf M, Stadler PF. 2013. **MITOS: Improved de novo Metazoan Mitochondrial Genome Annotation.** *Molecular Phylogenetics and Evolution* **69**:313-319
4. Blair D, Xu ZB, Agatsuma T. 1999. **Paragonimiasis and genus *Paragonimus***. *Advance Parasitology* **42**:113–222
5. Boore J.L. 1999. **Animal mitochondrial genomes**. *Nucleic Acids Research* **27** :1767–1780
6. Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG, Parkhill J. 2005. ACT: the Artemis Comparison Tool. *Bioinformatics* **16**:3422-3.
7. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE,Lunter G, Marth GT, Sherry ST, McVean G, Durbin R; 1000 Genomes Project Analysis Group. 2011. **The variant call format and VCFtools.** Bioinformatics **27**:2156-8.
8. Fried B, Graczyk TK, Tamang L. 2004. **Food-borne intestinal trematodiases in humans**. *Parasitology Research* **93**:159
9. Keiser J, Utzinger J. 2009. **Food-borne trematodiases**. *Clinical Microbiology Reviews* **3**:466-83
10. Laslett D, Canbäck B. 2008. ARWEN: a program to detect tRNA genes in metazoanmitochondrial nucleotide sequences. *Bioinformatics* **2**:172-5.

11. Le TH, Blair D, Agatsuma T, Humair PF, Campbell NJ, et al. 2000. **Phylogenies inferred from mitochondrial gene orders- a cautionary tale from the parasitic flatworms.** *Molecular Biology and Evolution* **17**: 1123–1125.

12. Le TH, Blair D, McManus DP. 2001. **Complete DNA sequence and gene organization of the mitochondrial genome of the liverfluke, *Fasciola hepatica* L. (Platyhelminthes; Trematoda)**. *Parasitology* **123**:609-21.

13. Le TH, Blair D, McManus DP. 2001. **Complete DNA sequence and gene organization of the mitochondrial genome of the liverfluke, *Fasciola hepatica* L. (Platyhelminthes; Trematoda).** *Parasitology* **123**: 609–621.

14. Le TH, Blair D, McManus DP. 2002. **Mitochondrial genomes of parasitic flatworms.**Trends in Parasitology **18**:206-13

15. Liu GH, Gasser RB, Young ND, Song HQ, Ai L, Zhu XQ. 2014. **Complete mitochondrial genomes of the 'intermediate form' of *Fasciola* and *Fasciola gigantica*, and their comparison with *F. hepatica*.** *Parasites and Vectors*. 2014, **7**:150.

16. Maddison, WP, Maddison DR. 2011. Mesquite: a modular system for incoevolutionary analysis. Version 2.75 Available at http://mesquiteproject.org

17. Nakao M, Sako Y, Ito A. 2003. **The mitochondrial genome of the tapeworm *Taenia solium*: a finding of the abbreviated stop codon U**. *Journal of Parasitology* **89**:633-5.

18. Okimoto R, Macfarlane JL, Wolstenholme DR. 1990. **Evidence for the frequent use of TTG as the translation initiation codon of mitochondrial protein genes in the nematodes, *Ascaris suum* and *Caenorhabditis elegans*.** Nucleic Acids Research **18**:6113-8.

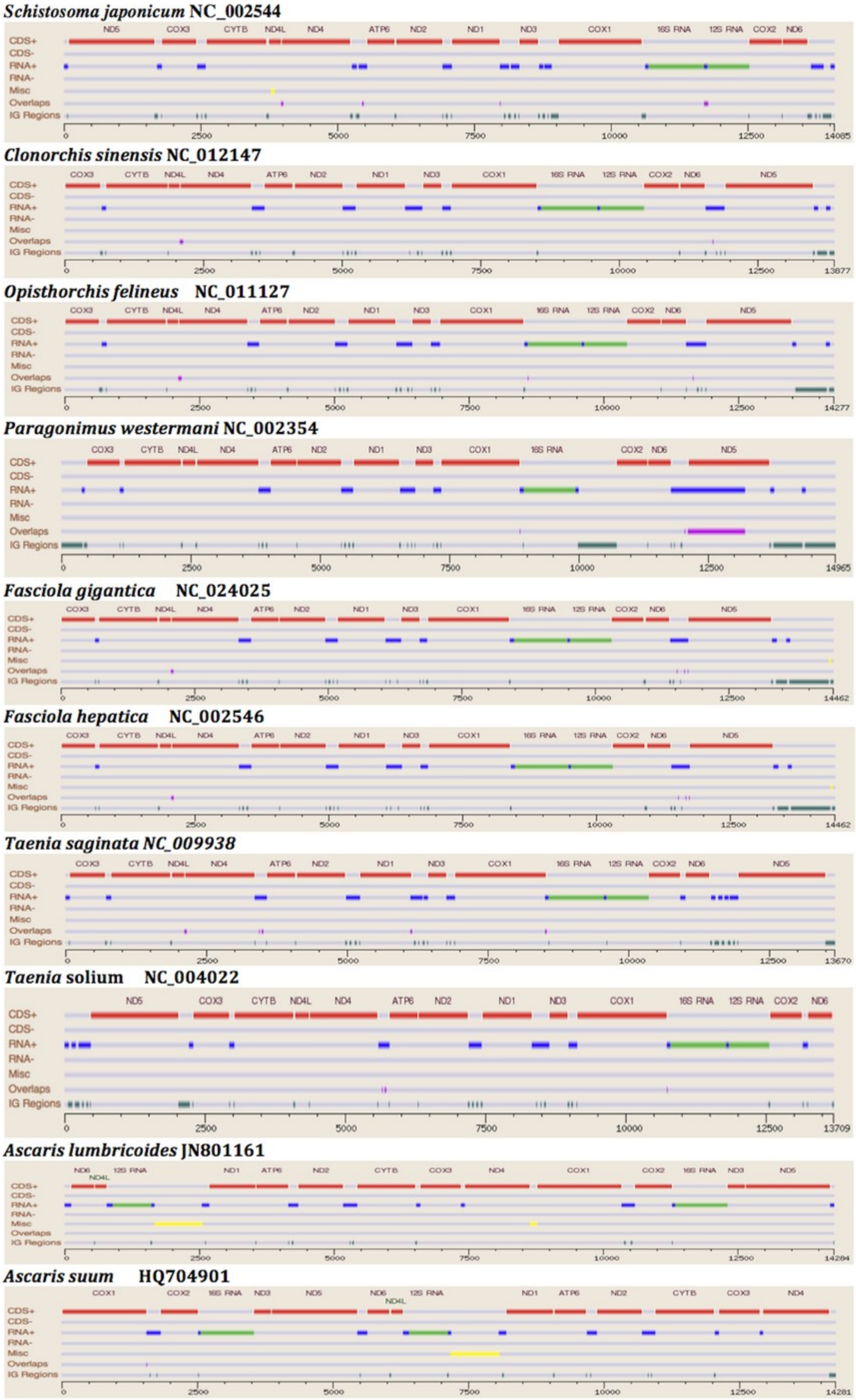
19. Quinn NL, Levenkova N, Chow W, Bouffard P, Boroevich KA, Knight JR, Jarvie TP, Lubieniecki KP, Desany BA, Koop BF, Harkins TT, Davidson WS. 2008. **Assessing the**

feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. *BMC Genomics* **9**:404

20. Ronquist F, Huelsenbeck JP. 2003. **MRBAYES 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* **19**:1572-1574
21. Saccone C, De Giorgi C, Gissi C, Pesole G, Reyes A. 1999. **Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system**. *Gene* **238**:195-209
22. Shekhovtsov SV, Katokhin AV, Kolchanov NA, Mordvinov VA. 2010. **The complete mitochondrial genomes of the liver flukes *Opisthorchis felineus* and *Clonorchis sinensis* (Trematoda)**. *Parasitology International* **59**: 100–103
23. Toscano C, Sen Hai Yu, Nunn P, Mott KE. 1995. **Paragonimiasis and tuberculosis, diagnostic confusion: a review of the literature.** *Tropical Diseases Bulletin* **92**: R1
24. Wolstenholme DR. 1992. **Animal mitochondrial DNA, structure and evolution**. *International Review of Cytology* 141: 173–216
25. Waeschenbach A, Webster BL, Littlewood DT. 2012. **Adding resolution to ordinal level relationships of tapeworms (Platyhelminthes: Cestoda) with large fragments of mtDNA.** Molecular Phylogenetics and Evolution. **63**: 834–847
26. Yan HB, Wang XY, Lou ZZ, Li L, Blair D, Yin H, Cai JZ, Dai XL, Lei MT, Zhu XQ, Cai XP, Jia WZ. 2013. **The mitochondrial genome of Paramphistomum cervi (Digenea), the first representative for the family Paramphistomidae.** *PLoS One*. **8**:e71300
27. Zerbino DR, Birney E. 2008.**Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Research* **18**:821-9.

Figure 1

Comparative Synteny map of the representative species for the helminth mtDNA illustrating the protein coding genes, tRNAs, rRNAs etc.



Schistosoma japonicum NC_002544

Clonorchis sinensis NC_012147

Opisthorchis felineus NC_011127

Paragonimus westermani NC_002354

Fasciola gigantica NC_024025

Fasciola hepatica NC_002546

Taenia saginata NC_009938

Taenia solium NC_004022

Ascaris lumbricoides JN801161

Ascaris suum HQ704901

Figure 2

Circular genome map of Paragonimus westermani mtDNA

The manual and in-silico annotations with appropriate regions for P. westermani mtDNA and annotated GenBank flat file for P. westermani were drawn into a circular graph in GenomeVX depicting the 12 PCGs and 24tRNAs.

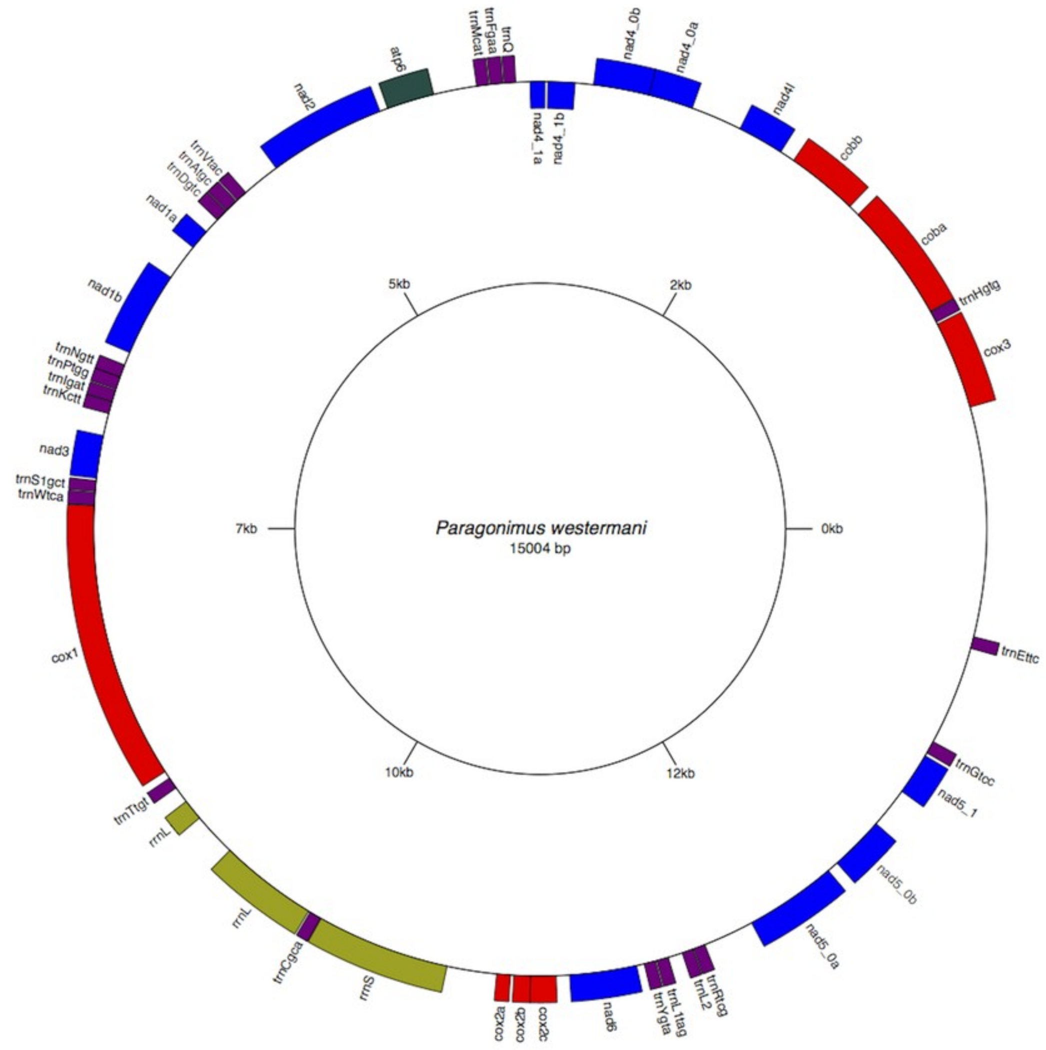

Paragonimus westermani
15004 bp
0kb
2kb
5kb
7kb
10kb
12kb
nad4_1a
nad4_1b
nad4_0b
nad4_0a
nad4l
cobb
coba
trnHgtg
cox3
trnEttc
trnGtcc
nad5_1
nad5_0b
nad5_0a
trnRtcg
trnL2
trnL1tag
trnYgta
nad6
cox2c
cox2b
cox2a
rrnS
trnCgca
rrnL
rrnL
trnTtgt
cox1
trnWtca
trnS1gct
nad3
trnKctt
trnIgat
trnPtgg
trnNgtt
nad1b
nad1a
trnDgtc
trnAtgc
trnVtac
nad2
atp6
trnMcat
trnFgaa
trnQ

Figure 3

Inferred Phylogenetic relationship among the representative helminth mtDNA species of the concatenated 12 protein coding genes

Trees were inferred using MrBayes v3.1. A, tree inferred from concatenated nucleotide sequences of 12 protein-coding genes, using the cestode Echinococcus granulosus as the outgroup. Posterior support values are given at nodes. Differences in the gene order in the mitochondrial genomes of parasitic flatworms from the Trematoda and Cestoda and taking Nematoda (Ascaridida) as an outgroup are indictaed on the phylogenetic leaf nodes. See text for more details.

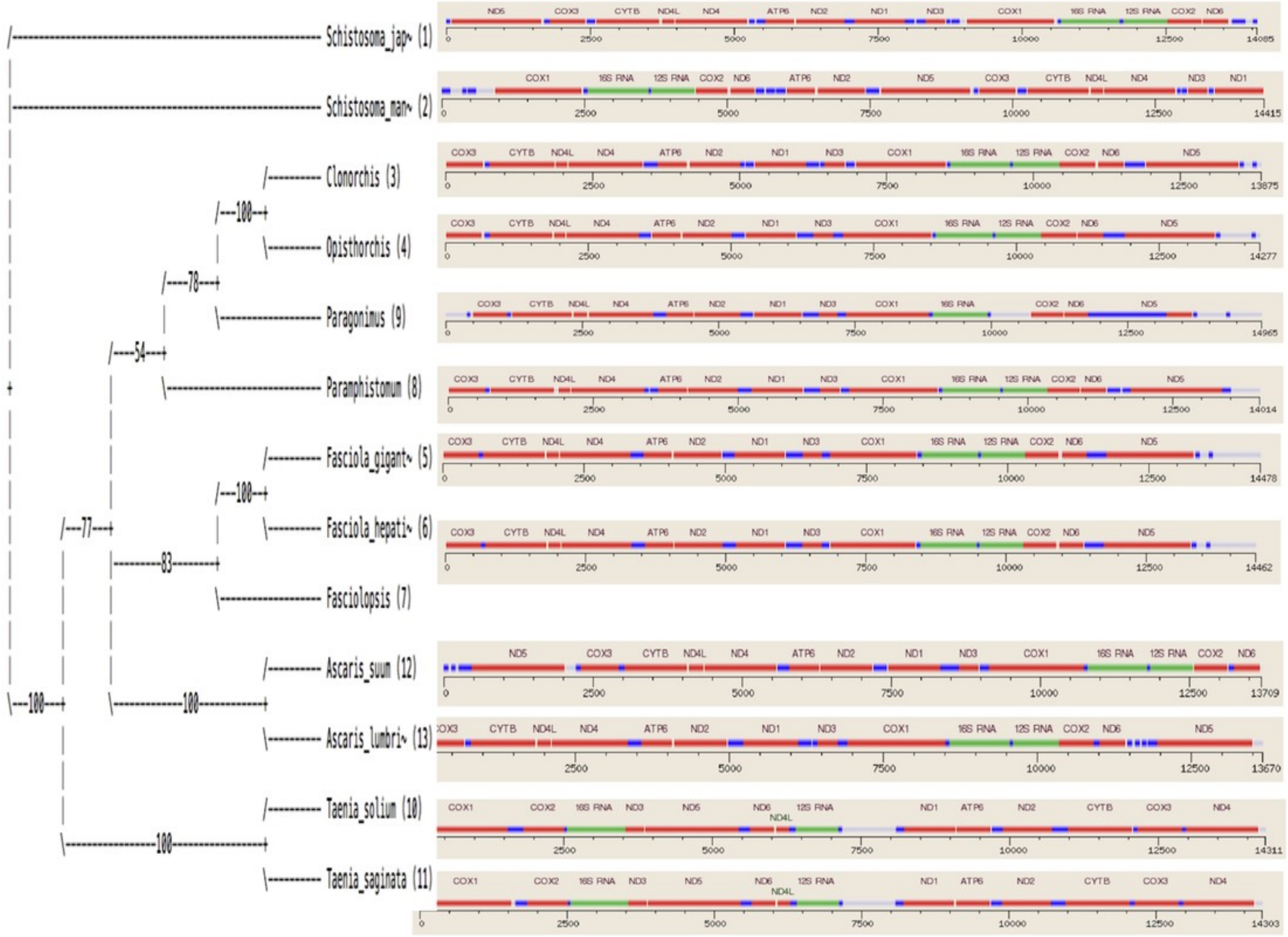

Schistosoma_jap~ (1)
ND5 COX3 CYTB ND4L ND4 ATP6 ND2 ND1 ND3 COX1 16S RNA 12S RNA COX2 ND6
0 2500 5000 7500 10000 12500 14085
Schistosoma_man~ (2)
COX1 16S RNA 12S RNA COX2 ND6 ATP6 ND2 ND5 COX3 CYTB ND4L ND4 ND3 ND1
0 2500 5000 7500 10000 12500 14415
Clonorchis (3)
COX3 CYTB ND4L ND4 ATP6 ND2 ND1 ND3 COX1 16S RNA 12S RNA COX2 ND6 ND5
0 2500 5000 7500 10000 12500 13875
100
Opisthorchis (4)
COX3 CYTB ND4L ND4 ATP6 ND2 ND1 ND3 COX1 16S RNA 12S RNA COX2 ND6 ND5
0 2500 5000 7500 10000 12500 14277
78
Paragonimus (9)
COX3 CYTB ND4L ND4 ATP6 ND2 ND1 ND3 COX1 16S RNA COX2 ND6 ND5
0 2500 5000 7500 10000 12500 14965
54
Paramphistomum (8)
COX3 CYTB ND4L ND4 ATP6 ND2 ND1 ND3 COX1 16S RNA 12S RNA COX2 ND6 ND5
0 2500 5000 7500 10000 12500 14014
Fasciola_gigant~ (5)
COX3 CYTB ND4L ND4 ATP6 ND2 ND1 ND3 COX1 16S RNA 12S RNA COX2 ND6 ND5
0 2500 5000 7500 10000 12500 14478
100
77
Fasciola_hepati~ (6)
COX3 CYTB ND4L ND4 ATP6 ND2 ND1 ND3 COX1 16S RNA 12S RNA COX2 ND6 ND5
0 2500 5000 7500 10000 12500 14462
83
Fasciolopsis (7)
Ascaris_suum (12)
ND5 COX3 CYTB ND4L ND4 ATP6 ND2 ND1 ND3 COX1 16S RNA 12S RNA COX2 ND6
0 2500 5000 7500 10000 12500 13709
100
100
Ascaris_lumbri~ (13)
COX3 CYTB ND4L ND4 ATP6 ND2 ND1 ND3 COX1 16S RNA 12S RNA COX2 ND6 ND5
2500 5000 7500 10000 12500 13670
Taenia_solium (10)
COX1 COX2 16S RNA ND3 ND5 ND6 ND4L 12S RNA ND1 ATP6 ND2 CYTB COX3 ND4
2500 5000 7500 10000 12500 14311
100
Taenia_saginata (11)
COX1 COX2 16S RNA ND3 ND5 ND6 ND4L 12S RNA ND1 ATP6 ND2 CYTB COX3 ND4
0 2500 5000 7500 10000 12500 14303

Figure 4

Summarized 12 protein coding genes and their blast hit protein plots

The protein plot depicts the quality value for each gene and each position if it is above the threshold and the different genes are differentiated with a range of colour codes. The hits used in MITOS correspond to the "mountains" in this protein plot that visualizes the signal from the BLAST searches. The arrows shown on the top of the plot depict gene order annotation and the quality values are shown on a log scale.

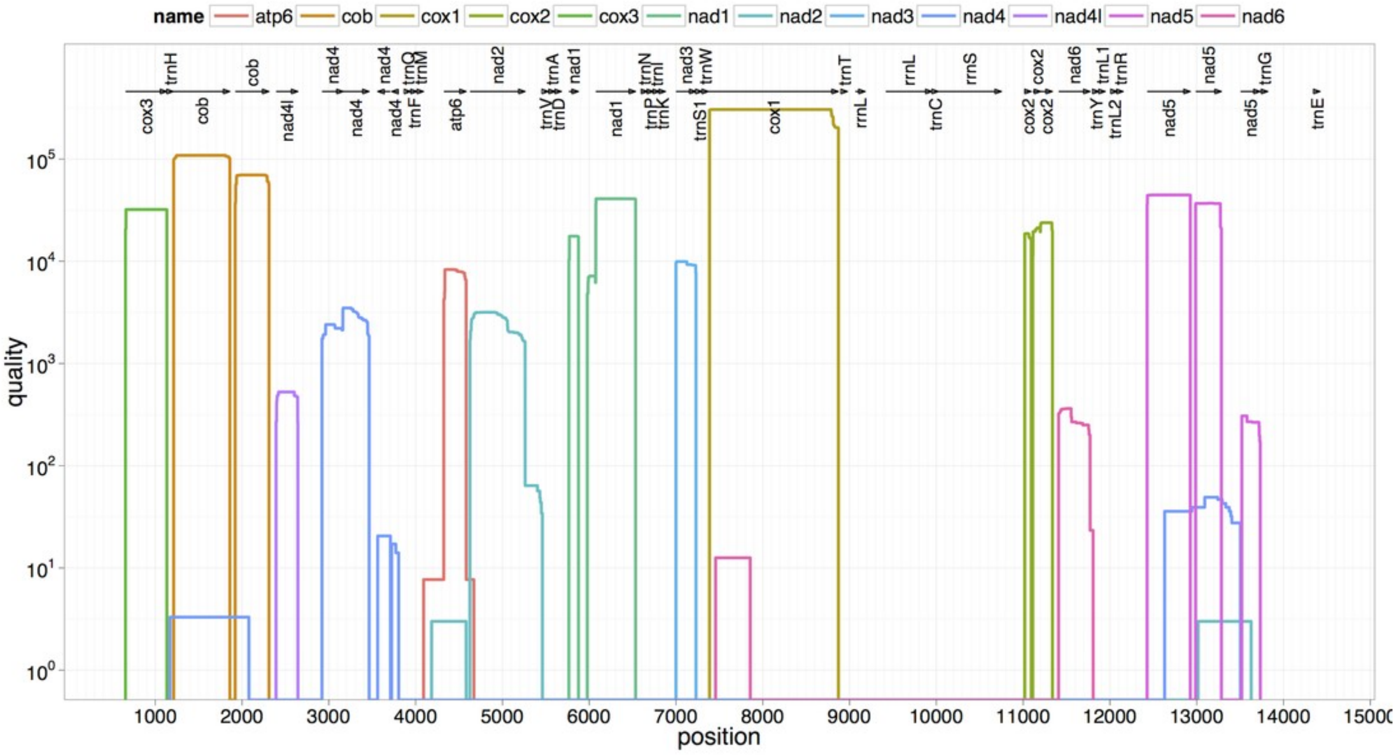
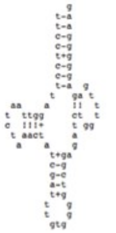
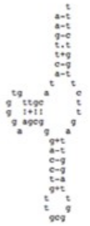


Figure 5

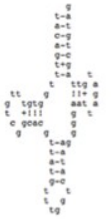
24 tRNA secondary structures predicted using ARWEN



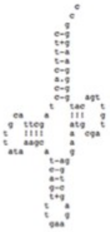
```
mtRNA-His(gtg)
65 bases, %GC = 46.2
Sequence [1146,1210]
```



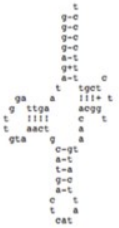
```
TV-loop mtRNA-Arg(gcg)
57 bases, %GC = 47.4
Sequence c[3653,3709]
```



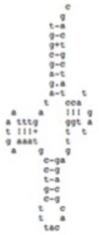
```
mtRNA-?(Gln|Pro)(tg)
63 bases, %GC = 41.3
Sequence [3882,3944]
```



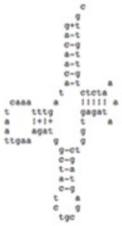
```
mtRNA-Phe(gaa)
74 bases, %GC = 45.9
Sequence [3950,4023]
```



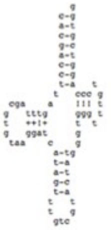
```
mtRNA-Met(cat)
66 bases, %GC = 42.4
Sequence [4027,4092]
```



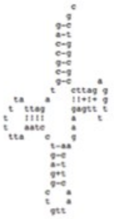
```
mtRNA-Val(tac)
65 bases, %GC = 44.6
Sequence [5469,5533]
```



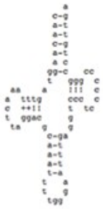
```
mtRNA-Ala(tgc)
73 bases, %GC = 38.4
Sequence [5539,5611]
```



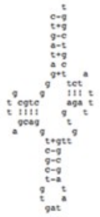
```
mtRNA-Asp(gtc)
69 bases, %GC = 49.3
Sequence [5614,5682]
```



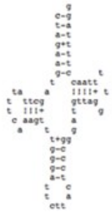
```
mtRNA-Asn(gtt)
71 bases, %GC = 43.7
Sequence [6606,6676]
```



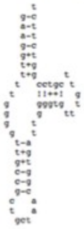
```
mtRNA-Pro(tgg)
68 bases, %GC = 48.5
Sequence [6676,6743]
```



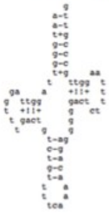
```
mtRNA-Ile(gat)
62 bases, %GC = 50.0
Sequence [6750,6811]
```



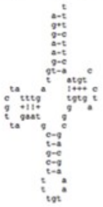
```
mtRNA-Lys(ctt)
66 bases, %GC = 37.9
Sequence [6815,6880]
```



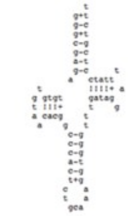
```
D-loop mtRNA-Ser(gct)
59 bases, %GC = 54.2
Sequence [7244,7302]
```



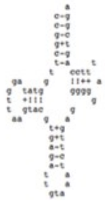
```
mtRNA-(Stop|Trp)(tca)
68 bases, %GC = 44.1
Sequence [7308,7375]
```



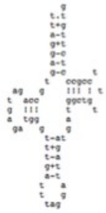
```
mtRNA-Thr(tgt)
66 bases, %GC = 39.4
Sequence [8913,8978]
```



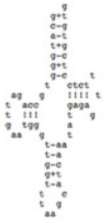
```
mtRNA-Cys(gca)
65 bases, %GC = 50.8
Sequence [9961,10025]
```



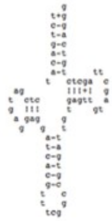
```
mtRNA-Tyr(gta)
61 bases, %GC = 47.5
Sequence [11815,11875]
```



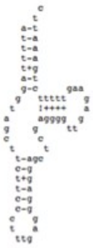
```
mtRNA-Leu(tag)
65 bases, %GC = 46.2
Sequence [11883,11947]
```



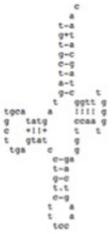
```
mtRNA-?(Leu|Leu)(aa)
62 bases, %GC = 45.2
Sequence [12025,12086]
```



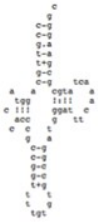
```
mtRNA-Arg(tcg)
68 bases, %GC = 51.5
Sequence [12089,12156]
```



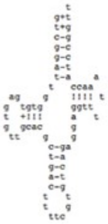
```
D-loop mtRNA-Gln(ttg)
64 bases, %GC = 43.8
Sequence [12182,12245]
```



```
mtRNA-Gly(tcc)
73 bases, %GC = 45.2
Sequence [13750,13822]
```



```
mtRNA-Thr(tgt)
64 bases, %GC = 59.4
Sequence [13874,13937]
```



```
mtRNA-Glu(ttc)
65 bases, %GC = 49.2
Sequence [14358,14422]
```

Table 1(on next page)

mt DNA nucleotide sequence statistics information of representative helminth parasites

Table 1. Mitochondrial DNA Nucleotide sequence statistics information of selected digenean trematodes, cestodes and nematodes

Sequence type	DNA	DNA	DNA	DNA	DNA	DNA	DNA	DNA	DNA	DNA	DNA	DNA	DNA
Length	14,118 bp circular	14,462 bp circular	15,004 bp circular	14,277bp circular	14014 bp circular	13,875bp circular	14,478 bp circular	14,415bp circular	14,085bp circular	13,670bp circular	13,709 bp circular	14,281 bp circular	14,284 bp circular
Organism Name	***Fasciolopsis buski***	***Fasciola hepatica***	***Paragonimus westermani***	***Opisthorchis felineus***	***Paramphistomum cervi***	***Clonorchis sinensis***	***Fasciola gigantica***	***Schistosoma mansoni***	***Schistosoma japonicum***	***Taenia saginata***	***Taenia solium***	***Ascaris lumbricoides***	***Ascaris suum***
Accession	Submitted to GenBank	NC_002546	NC_002354	EU921260	NC_023095	FJ381664	NC_024025	NC_002545	NC_002544	NC_009938	NC_004022	JN801161	NC_001327
Modification Date	submitted	01-FEB-2010	submitted	18-AUG-2010	14-JAN-2014	01-JUL-2010	01-MAY-2014	14-APR-2009	01-FEB-2010	14-APR-2009	01-FEB-2010	01-DEC-2011	11-MAR-2010
Weight (single-stranded)	4396.507	4,499.496 kDa	4,666.455 kDa	4,437.683 kDa	4,363.551 kDa	4,311.834 kDa	4,504.913 kDa	4,482.165 kDa	4,371.002 kDa	4,242.425 kDa	4,251.992 kDa	4,428.619 kDa	4,429.981 kDa
Weight (double-stranded)	8721.667	8,934.244 kDa	9,270.244 kDa	8,820.283 kDa	8,657.348 kDa	8,571.888 kDa	8,944.06 kDa	8,904.302 kDa	8,700.11 kDa	8,443.711 kDa	8,467.723 kDa	8,443.711 kDa	8,822.899 kDa
Annotation table													
Featutre type	**Count**	**Count**	**Count**	**Count**	**Count**	**Count**	**Count**	**Count**	**Count**	**Count**		**Count**	**Count**
CDS	12	12	12	12	12	12	12	12	12	12	12	12	12
Gene	12	12	12	12	12	12	12	12	12	12	12	12	12
Misc. feature	1	1	-	-	-	-		1	1	-	-	1	2
rRNA	2	2	2	2	2	2	2	2	2	2	2	2	2
tRNA	22	22	24	22	22	22	22	23	23	22	22	22	22

Table 2(on next page)

P. westermani mtDNA annotations showing PCGs and tRNA in dot bracket format

Name	Start	Stop	Strand	Length	Structure
cox3	658	1134	+	477	
trnH(gtg)	1147	1209	+	63	(((((((..((((......)))).(((((.......)))))....((.(..).)))))))))).
cob-a	1213	1860	+	648	
cob-b	1922	2311	+	390	
nad4l	2393	2644	+	252	
nad4_0-a	2922	3167	+	246	
nad4_0-b	3163	3465	+	303	
nad4_1-b	3564	3710	-	147	
nad4_1-a	3725	3805	-	81	
trnQ(---)	3882	3944	+	63	(((((((..((((......)))).(((((......)))))....((.......)))))))))).
trnF(gaa)	3951	4020	+	70	((((.((..((((........)))).(((((.......)))))....(((.........))))).)))).
trnM(cat)	4027	4092	+	66	(((((((..((((........)))).(((((.......)))))....((((...))))))))))).
atp6	4326	4583	+	258	
nad2	4627	5262	+	636	
trnV(tac)	5470	5531	+	62	(((((.(..((((.....)))).(((((.......)))))....(((....)))).))))).
trnA(tgc)	5539	5610	+	72	(((((((..((((............)))).(((((.......)))))....(((((...)))))))))))).
trnD(gtc)	5615	5681	+	67	(((((((..((((.........)))).(((((.......)))))....(((.....)))))))))).
nad1-a	5767	5877	+	111	
nad1-b	6077	6535	+	459	
trnN(gtt)	6606	6675	+	70	(((((((..((((........)))).(((((.......)))))....(((((.....)))))))))))).
trnP(tgg)	6676	6743	+	68	((((.(((..((((.......)))).(((((.......)))))....(((.......)))))))))).
trnI(gat)	6749	6812	+	64	((((((.(..((((....)))).(((((.......)))))....(((((..)))))))))))).
trnK(ctt)	6815	6880	+	66	(((((((..((((......)))).(((((.......)))))....(((((...)))))))))))).
nad3	7001	7231	+	231	
trnS1(gct)	7244	7302	+	59	(((((((.......(((((.......)))))....(((((......)))))))))))).
trnW(tca)	7308	7375	+	68	(((((((..((((......)))).(((((.......)))))....((((.......))))))))))).
cox1	7379	8872	+	1494	
trnT(tgt)	8914	8977	+	64	((((((...((((.......)))).(((((.......)))))....(((....))).)))))).
rrnL	9067	9181	+	115	((...((((((((.....((.(((((((.((((...))))...((..(((((.....)))))..))..).)))))).))....((.....)).......))))))))......))
rrnL	9417	9951	+	535	................(((....)))....(.....)..................((................))..((....))..((...................(...)......(((........))).....))..........((((((........))))))......(((.(((.((.....(((((((((((.(((((...)))))..((((...)))).((.............))........((((...))))....))))).)))))))).......(((((((((((((...((.......))...)))))))))).)))..........(((((.((((((..(((.((((((......)))))).)))..(((((.....)))))....)))))..)...)))))......(((((((....))).))))....))..))))))....((...............))...........
trnC(gca)	9961	10025	+	65	(((((((..((((...))))...(((((.......)))))....(((((...)))))))))))).
rrnS	10028	10751	+	724	...(((((.......))))).(((((((...((((((((....((..(....).......(((..................(((..((...)).))).)))))).....(((.(..(((((....)))))))))).)))))))...(((((((((..................)))...............)))))))))))....((((.....((((.(.(((..........(((.((......))...)))............((....))...))).).))))....(((((...((((..........))))...)))))............)))).((((....))))........((((.(((((((((.......(((((..((((((((((....(((.........)))..........(((((((.....((.((..(((((((((((....(((.....(....).((....)).)))..............))).)))...))))))))....))))))).)).))))))))............(((((..........)))))..........)))))......((((((...........))))))...........))))))))).))..............((. (((....))).))................(((((((.((....)).))))))))...........
cox2-a	11020	11094	+	75	
cox2-b	11112	11204	+	93	
cox2-c	11201	11338	+	138	
nad6	11410	11772	+	363	
trnY(gta)	11814	11876	+	63	.((((((..((((.......)))).(((((.......)))))....((((.))))))))))..

trnL1(tag)	11883	11947	+	65	.((((((..(((.......))).(((((.......)))))....(((.(...).)))))))))..
trnL2(---)	12025	12086	+	62	(((((((..(((.......))).((((........))))....(((((.)))))))))))).
trnR(tcg)	12091	12154	+	64	(((((((((......)))).(((((.......)))))....(((((.......)))))))))).
nad5_0-a	12430	12927	+	498	
nad5_0-b	12989	13285	+	297	
nad5_1	13506	13733	+	228	
trnG(tcc)	13751	13820	+	70	(((((((..((((..........)))).(((.(.......).)))....((((.....))))))))))).
trnE(ttc)	14358	14422	+	65	(((((((..((((.......)))).(((((.......)))))....((((...))))))))))).

Table 2. *P. westermani* mtDNA annotations showing PCGs and tRNA in dot bracket format

Table 3(on next page)

Codon usage for Paragonimus westermani mt DNA

Table 3. Codon usage for *Paragonimus westermani* mt DNA

AmAcid	Codon	Number	/1000	Fraction
Ala	GCG	57.00	11.40	0.27
Ala	GCA	38.00	7.60	0.18
Ala	GCT	75.00	15.00	0.36
Ala	GCC	39.00	7.80	0.19
Cys	TGT	208.00	41.59	0.76
Cys	TGC	67.00	13.40	0.24
Asp	GAT	91.00	18.20	0.72
Asp	GAC	36.00	7.20	0.28
Glu	GAG	111.00	22.20	0.69
Glu	GAA	51.00	10.20	0.31
Phe	TTT	310.00	61.99	0.74
Phe	TTC	109.00	21.80	0.26
Gly	GGG	168.00	33.59	0.34
Gly	GGA	89.00	17.80	0.18
Gly	GGT	166.00	33.19	0.34
Gly	GGC	66.00	13.20	0.13
His	CAT	44.00	8.80	0.61
His	CAC	28.00	5.60	0.39
Ile	ATT	97.00	19.40	0.71
Ile	ATC	40.00	8.00	0.29
Lys	AAG	66.00	13.20	1.00
Leu	TTG	226.00	45.19	0.34
Leu	TTA	110.00	22.00	0.17
Leu	CTG	92.00	18.40	0.14
Leu	CTA	35.00	7.00	0.05
Leu	CTT	147.00	29.39	0.22
Leu	CTC	56.00	11.20	0.08
Met	ATG	89.00	17.80	0.80
Met	ATA	22.00	4.40	0.20
Asn	AAA	55.00	11.00	0.45
Asn	AAT	44.00	8.80	0.36
Asn	AAC	24.00	4.80	0.20
Pro	CCG	34.00	6.80	0.26
Pro	CCA	20.00	4.00	0.15
Pro	CCT	57.00	11.40	0.43
Pro	CCC	22.00	4.40	0.17
Gln	CAG	42.00	8.40	0.64
Gln	CAA	24.00	4.80	0.36
Arg	CGG	51.00	10.20	0.35
Arg	CGA	26.00	5.20	0.18
Arg	CGT	51.00	10.20	0.35
Arg	CGC	19.00	3.80	0.13
Ser	AGG	125.00	25.00	0.21
Ser	AGA	57.00	11.40	0.09
Ser	AGT	76.00	15.20	0.13
Ser	AGC	36.00	7.20	0.06
Ser	TCG	56.00	11.20	0.09
Ser	TCA	52.00	10.40	0.09
Ser	TCT	134.00	26.79	0.22
Ser	TCC	68.00	13.60	0.11
Thr	ACG	37.00	7.40	0.29
Thr	ACA	20.00	4.00	0.16
Thr	ACT	43.00	8.60	0.34
Thr	ACC	27.00	5.40	0.21
Val	GTG	156.00	31.19	0.29
Val	GTA	58.00	11.60	0.11
Val	GTT	256.00	51.19	0.48
Val	GTC	65.00	13.00	0.12
Trp	TGG	159.00	31.79	0.58
Trp	TGA	113.00	22.60	0.42
Tyr	TAT	74.00	14.80	0.57
Tyr	TAC	55.00	11.00	0.43

End	TAG	66.00	13.20	0.50
End	TAA	66.00	13.20	0.50

Table 4(on next page)

Atomic composition and nucleotide distribution Table of *Paragonimus westermani* mtDNA

Table 4. Atomic composition and Nucleotide distribution Table of *Paragonimus westermani* mtDNA

Atomic composition

As single-stranded

Atom	Count	Frequency
Hydrogen (H)	185664	0.375
Carbon (C)	147756	0.298
Nitrogen (N)	53610	0.108
Oxygen (O)	93068	0.188
Phosphorus (P)	15004	0.03

As double-stranded

Atom	Count	Frequency
Hydrogen (H)	368285	0.374
Carbon (C)	293261	0.298
Nitrogen (N)	111847	0.114
Oxygen (O)	180050	0.183
Phosphorus (P)	30008	0.031

Nucleotide distribution table

Nucleotide	Count	Frequency
Adenine (A)	2571	0.171
Cytosine (C)	2284	0.152
Guanine (G)	4535	0.302
Thymine (T)	5614	0.374
C + G	6819	0.454
A + T	8185	0.546

Counts of di-nucleotides

1.pos\2.pos	A	C	G	T
A	562	389	857	763
C	391	430	505	958
G	882	592	1492	1568
T	736	873	1680	2325

Frequency of di-nucleotides

1.pos\2.pos	A	C	G	T
A	0.037	0.026	0.057	0.051
C	0.026	0.029	0.034	0.064
G	0.059	0.039	0.099	0.105
T	0.049	0.058	0.112	0.155

Table 5(on next page)

Summary of illumina and Ion-Torrent quality control reads

Table 5. Ion Torrent and Illumina reads

Ion torrent reads		
S.No	1	2
Fastq file name	processed_reads.fastq	mapped_mito.fastq
Fastq file size	239.71 MB	71.55 MB
Time taken for Analysis	8.75 Seconds	2.76 Seconds
Maximum Read Length	260	260
Minimum Read Length	35	35
Mean Read Length	121	117
Total Number of Reads	890504	292832
Total Number of HQ Reads 1*	890442	292822
Percentage of HQ Reads	99.993%	99.997%
Total Number of Bases	107866584 bases	34145801 bases
Total Number of Bases in Mb	107.8666 Mb	34.1458 Mb
Total Number of HQ Bases 2*	105216008 bases	33218357 bases
Total Number of HQ Bases in Mb	105.2160 Mb	33.2184 Mb
Percentage of HQ Bases	97.543%	97.284%
Total Number of Non-ATGC Characters	0 bases	0 bases
Total Number of Non-ATGC Characters in Mb	0.000000 Mb	0.000000 Mb
Percentage of Non-ATGC Characters	0.000%	0.000%
Number of Reads with Non-ATGC Characters	0	0
Percentage of Reads with Non-ATGC Characters	0.000%	0.000%
Illumina reads		
S.No		1
Fastq file name		SE_ill.fastq
Fastq file size		14.56 MB
Time taken for Analysis		0.48 Seconds
Maximum Read Length		100
Minimum Read Length		50
Mean Read Length		96
Total Number of Reads		62874
Total Number of HQ Reads 1*		62874
Percentage of HQ Reads		100.000%
Total Number of Bases		6053872 bases
Total Number of Bases in Mb		6.0539 Mb
Total Number of HQ Bases 2*		5982733 bases
Total Number of HQ Bases in Mb		5.9827 Mb
Percentage of HQ Bases		98.825%
Total Number of Non-ATGC Characters		410 bases
Total Number of Non-ATGC Characters in Mb		0.000410 Mb
Percentage of Non-ATGC Characters		0.007%
Number of Reads with Non-ATGC Characters		240
Percentage of Reads with Non-ATGC Characters		0.382%