Transcriptome analysis of *Taenia solium* cysticerci using Open Reading Frame ESTs (ORESTES)

Carolina R Almeida¹, Patricia H Stoco², Glauber Wagner²,³, Thaís CM Sincero², Gianinna Rotava², Ethel Bayer-Santos², Juliana B Rodrigues², Maísa M Sperandio², Antônio AM Maia⁴, Elida PB Ojopi¹, Arnaldo Zaha⁵, Henrique B Ferreira⁵, Kevin M Tyler⁶, Alberto MR Dávila⁷, Edmundo C Grisard*²,⁶ and Emmanuel Dias-Neto¹,⁸

Address: ¹Laboratório de Neurociências (LIM27), Departamento e Instituto de Psiquiatria, Faculdade de Medicina da Universidade de São Paulo, Rua Dr. Ovídio Pires de Campos 785, CEP 05403-010, São Paulo, SP, Brazil, ²Laboratórios de Protozoologia e de Bioinformática, Departamento de Microbiologia, Imunologia e Parasitologia, Universidade Federal de Santa Catarina (UFSC), Caixa postal 476, CEP 88040-970, Florianópolis, SC, Brazil, ³Laboratório de Biologia Molecular e Bioinformática, Área das Ciências Biológicas e da Saúde, Universidade do Oeste de Santa Catarina (UNOESC), CEP 89600-000, Joaçaba, SC, Brazil, ⁴Departamento de Ciências Básicas, Faculdade de Zootecnia e Engenharia de Alimentos, Universidade de São Paulo (USP), Av. Duque de Caxias Norte, 225, Centro, CEP 13630-000, Pirassununga, SP, Brazil, ⁵Departamento de Biologia Molecular e Biotecnologia and Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves 9500, Prédio 4342131, Setor IV, Campus do Vale, CEP 91501-970, Porto Alegre, RS, Brazil, ⁶Biomedical Research Centre, School of Medicine, Health Policy and Practice, University of East Anglia (UEA), NR4 7TJ, Norwich, Norfolk, UK, ⁷Instituto Oswaldo Cruz, Fundação Oswaldo Cruz (FIOCRUZ), Pavilhão Leônidas Deane, Avenida Brasil 4365, CEP 21204-360, Rio de Janeiro, RJ, Brazil and ⁸Department of Genitourinary Medical Oncology, MD Anderson Cancer Center, 1515 Holcombe Blvd., Unit 1374 Medical Center, 77030 Houston, USA

Email: Carolina R Almeida - Carolina.Almeida@appliedbiosystems.com; Patricia H Stoco - patriciastoco@gmail.com; Glauber Wagner - glauber.wagner@unoesc.edu.br; Thaís CM Sincero - thais@ccbi.ufsc.br; Gianinna Rotava - gianinna_rotava@yahoo.com.br; Ethel Bayer-Santos - ethel.bs@gmail.com; Juliana B Rodrigues - juliberka@gmail.com; Maísa M Sperandio - maisamichels@hotmail.com; Antônio AM Maia - antomaia@usp.br; Elida PB Ojopi - elida@usp.br; Arnaldo Zaha - zaha@cbiot.ufrgs.br; Henrique B Ferreira - henrique@cbiot.ufrgs.br; Kevin M Tyler - k.tyler@uea.ac.uk; Alberto MR Dávila - davila@fiocruz.br; Edmundo C Grisard* - grisard@ccbi.ufsc.br; Emmanuel Dias-Neto - emmanuel@usp.br

* Corresponding author

Published: 31 July 2009

Received: 5 May 2009

Accepted: 31 July 2009

This article is available from: http://www.parasitesandvectors.com/content/2/1/35

© 2009 Almeida et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: Human infection by the pork tapeworm *Taenia solium* affects more than 50 million people worldwide, particularly in underdeveloped and developing countries. Cysticercosis which arises from larval encystation can be life threatening and difficult to treat. Here, we investigate for the first time the transcriptome of the clinically relevant cysticerci larval form.

Results: Using Expressed Sequence Tags (ESTs) produced by the ORESTES method, a total of 1,520 high quality ESTs were generated from 20 ORESTES cDNA mini-libraries and its analysis revealed fragments of genes with promising applications including 51 ESTs matching antigens previously described in other species, as well as 113 sequences representing proteins with potential extracellular localization, with obvious applications for immune-diagnosis or vaccine development.
Conclusion: The set of sequences described here will contribute to deciphering the expression profile of this important parasite and will be informative for the genome assembly and annotation, as well as for studies of intra- and inter-specific sequence variability. Genes of interest for developing new diagnostic and therapeutic tools are described and discussed.

Background

*Taenia solium*, the pork tapeworm, infects around 50 million people worldwide and is one of the foremost public health problems in developing countries [1,2]. The high influx and immigration of people coming from endemic areas to more industrialized nations has produced a complex spreading pattern for cysticercosis which is now a world-wide issue [2].

Cystercercosis arises from the development of *T. solium* cysticerci in soft tissues as a result of ingesting *T. solium* eggs [3-5]. Neurocysticercosis; which can cause epilepticiform attacks, headaches, learning difficulties and convulsions; is considered the primary cause of acquired epilepsy and its clinical/therapeutic management is difficult, highlighting the importance of search for new drug targets [6-8]. In this work, we investigate for the first time the gene expression profile of *T. solium* in the larval form responsible – the cysticerci.

High throughput sequencing for gene discovery and gene expression profiling using traditional [9] or alternative ‘Expressed Sequence Tags’ (ESTs) such as ‘Open Reading Frame ESTs’ (ORESTES) [10,11] has greatly increased our knowledge of the set of expressed genes of some important helminthic parasites, notably *Schistosoma mansoni* [12,13] and its intermediate vector *Biomphalaria glabrata* [14], *S. japonicum* [15] and the cestodes *Echinococcus granulosus* [16], *E. multilocularis* [17] and *Mesocestoides corti* [18].

Recently, Aguilar-Díaz et al. [1] described the *T. solium* genome initiative designed to unravel the parasite’s complete genome. The availability of transcribed sequences, such as those presented here, will be key to the facilitate genome annotation and gene discovery in *T. solium*.

Results

Here we present the sequencing and analysis of 2,857 ORESTES derived from *T. solium* cysticerci, revealing a fraction of the parasite transcriptome. A total of 1,520 high-quality ORESTES generated here were deposited in dbEST database of GenBank http://www.ncbi.nlm.nih.gov/dbEST, being 1,180 annotated as from *T. solium* [GenBank:EX150322 to EX151133 and GenBank:FD661301 to FD661668] and 340 corresponding to pig-derived sequences [GenBank:EX151134 to EX151473]. These sequences are also available at the STINGRAY system on the BiowebDB consortium website http://stingray.biowebdb.org/, together with relevant annotations and additional files. A list of the *T. solium* ORESTES and their respective GenBank accession numbers is presented on the Additional file 1.

General Features

A general overview of the *T. solium* ESTs generated here is presented in Table 1. More detailed analysis of the parasite transcriptome, such as codon usage and G+C content, can be obtained online at the STINGRAY system http://stingray.biowebdb.org/index.cgi?project=TS.

A total of 2,857 clones were sequenced and, after removal of poor quality (Phred<15 and/or less than 100 bp) and less informative sequences (typically rRNA and mtRNA), the remaining 1,520 ORESTES were used for sequence assembly following detailed analysis by the STINGRAY. After assembling, sequences were arranged into distinct sets named ‘Cysticerci’ and ‘Cysticerci PIGS’, which are available at STINGRAY.

The ‘Cysticerci’ project http://stingray.biowebdb.org/index.cgi?project=TS corresponds to the parasite-derived transcriptome and contains a total of 1,180 ESTs clustered in 812 non-redundant sequences (185 clusters + 627 singlets), with an average size of 355 nt, totaling 288,496 nt.

The ‘Cysticerci PIGS’ dataset http://stingray.biowebdb.org/index.cgi?project=TP was determined on the

| Table 1: General features of *Taenia solium* cysticerci cDNA sequences. |
|---------------------------------------------------------------|
| **Total number of valid ORESTES**  | 1,180  |
| **Total number of non-redundant sequences** | 812    |
| **Average length of non-redundant sequences** | 355 bp |
| **Clusters** | 185    |
| **Singlets** | 627    |
| **Non-redundant sequences G+C average content** | 49%    |
| **Number of non-redundant sequences hits on/with:** |
| - Blast | 424 (52.1%) |
| - RPSBlast | 150 (18.4%) |
| - InteProScan | 74 (9.0%) |
| - HHMNER | 97 (12.0%) |
| - *T. solium* Genome ESTs | 107 (13.2%) |
| **Sequences with no hits (blast/Interpro/HMMER)** | 350 (43.0%) |
| **Number of manually validated CDS** | 191 (23.5%) |
| **CDS G+C content average** | 53% |
| **Number of annotated sequences** | 191 (23.5%) |

* After trimming by quality, length and rRNA/mtRNA, as described in the text.
basis of blast similarity analysis with high scores against genomic sequences of *S. scrofa*. It is composed of 340 non-redundant singlets with an average size of 390 nt and about 132,000 nt in total. The stringency criterion used here warrants that most of this subset is certainly composed of the host transcripts, which may include transcripts relevant for the host-parasite interaction.

The parasite transcriptome
Among the parasite's 812 non-redundant sequences (627 singlets + 185 clusters), 462 yielded significant hits with at least one of the databases used for comparative analysis [http://stingray.biowebdb.org/index.cgi?project=TS](http://stingray.biowebdb.org/index.cgi?project=TS). From these sequences with significant hits, 204 showed similarity with sequences of parasitic metazoan species, including sequences from the *T. solium* Genome Project (84), from other *Taenia* species such as *T. saginata* (2), *T. crassiceps* (2), *T. asiatica* (1), as well as from other parasitic cestodes as *E. granulosus* (7) and *E. multilocularis* (4). Hits were also found against sequences from platyhelminths such as *S. mansoni* (88), *S. japonicum* (11), *Clonorchis sinensis* (1) and *Fasciola hepatica* (6). The remaining 263 sequences showed similarity with other metazoan species (see Additional file 2). For 350 sequences no hits were obtained on Blast, InterProScan or HMMER analyses.

After automated and manual annotation of all 812 non-redundant sequences, 191 were validated as coding sequences (CDS) (Table 1), of which 60 were considered hypothetical proteins or hypothetical conserved proteins. The number of ORESTES sequences according to their annotation identifiers is given in Additional file 3. As expected, this dataset enriched for coding sequences and showed a higher G+C content (53%) as compared to the total dataset (49%) (Table 1).

Analysis of the 191 annotated sequences using Gene Ontology (GO) allowed the categorization of 96 sequences, among which 84 were classified according to molecular function, 65 to biological processes and 48 to cellular component, several with multiple categories (Fig. 1). From the 65 sequences with biological processes annotation, the most frequent GO sub-categories were proteins related to cellular processes (40), followed by metabolic processes (10), biological regulation (4) and adhesion (4) (Fig. 2A). Among the GO molecular function sub-categories, binding (34), catalytic activity (24), structural molecule activity (14) and motor activity (7) were the most frequent (Fig. 2B). It is noteworthy that a relevant fraction of the transcripts revealed here appear to be related to structural aspects (such as adhesion, binding or structural molecule activity) that might be involved with the solid constitution of the cysts and their establishment on host tissues (see Additional file 4). A detailed description of each GO sub-category can be found on the annotated database available at the STINGRAY system [http://stingray.biowebdb.org/index.cgi?project=TS](http://stingray.biowebdb.org/index.cgi?project=TS).

The search for predictive sub-cellular localization of the products related to each annotated CDS was performed using the Wolf-PSORT software [19] and returned 113 hits. Among the 13 sequences with predictions of having extracellular localization with high scores (>25), only seven have a putative function assigned by GO. Among these, three are probably not extracellular ('40S Ribosomal protein S19' [GenBank:EX150987], 'Deoxyribonuclease I' [GenBank:EX151091], and 'Sec61-like protein' [GenBank:EX150487]), two may have extracellular localization ('WD40 repeat' – [GenBank:EX150998] and 'Heat-shock protein' [GenBank:EX151058]), while the 'ToLA protein' [GenBank:EX150587] and the 'Cadherin family member (cdh-4)' [GenBank:EX150991] are probably extracellular (see Additional file 5). Considering the Wolf-PSORT limitations in predicting cellular localization based on short sequences such as ESTs and the fact that none of the 113 proteins predicted as extracellular were annotated as antigens, even though most *Taenia* sp. proteins already reported in the literature have precisely that description, further analysis using full length sequences are necessary to confirm these results.

Conserved Domains and Motifs
The search for protein motifs among the parasite sequences was performed by similarity searches using InterProScan and RPSBlast using all databases available on the STINGRAY system and pointed out 64 distinct motifs distributed in 79 non-redundant sequences (see Additional files 6, 7 and 8). Among these, the 'pistil-specific extensin-like protein motif' [GenBank:IPR003882] was observed in 16 sequences, the 'vinculin/alpha-catenin' [GenBank:IPR006077] in 12, the 'glutelin' [GenBank:IPR000480] in six, 'fibronectin type III-like fold' [GenBank:IPR008957] in five and several others with four or less hits. Detailed categorization of the *T. solium* sequences according to the Eukaryotic Orthologous Groups (KOG) categories is shown on the Additional file 9.

A sensitive search for protein family recognition using multiple alignments was carried out with HMMER software and revealed 92 sequences of our parasite dataset generating at least one hit with the Pfam HMM profiles library. A domain with a still unknown function (DUF1787) was found in 12 sequences, the 'PT-PT repeat' in another nine sequences, the 'Hsp20/alpha crystallin family' (HSP20) and the 'I-set-immunoglobulin' (I-set domains) in four, the 'spectrin repeat' (SPECTRIN) and 'EGF-EGF-like domains' in another three sequences.
Ontology categorization of the distinct ontologies, Biological Process, Molecular Function and Cellular Component.

Comparisons with taeniid sequences

Only 117 of the 812 T. solium cysticerci clustered sequences described in the present study revealed similarity with the T. solium Genome Project ESTs available at GenBank. Among these 117 sequences, 107 showed similarity on tblastx and 100 on blastn analysis with ESTs of the T. solium Genome Project, 39 with exclusive hits to the larval stage sequence, 11 with the adult stage and 67 with genes expressed in both life-cycle stages (see Additional file 10).

Except for nine sequences from Taenia sp. or Echinococcus sp., the remaining cestode-related sequences presenting high score (>90) on blast against the T. solium sequences described in this study, were from Mesocestoides corti (heat shock 70 kDa protein) and from Diphyllobothrium dendriticum (actin). Further 36 low-score (<90) hits with the 28S ribosomal RNA from distinct cestode species were observed.

Comparative analysis against E. granulosus sequences from GenBank mainly revealed constitutive genes such as actin, paramyosin and others metabolic enzymes. However, two clusters [GenBank:EX151048, GenBank:EX151014] showed high similarity with genes coding for ERM family proteins (ezrin, radixin, moesin), exclusively with EST from larval stage of T. solium (see Additional files 6 and 8). Some of these proteins were characterized in Echinococcus species and received distinct names such as EM10, EG10, EM4 and antigen II/3, despite their high nucleotide similarity. In E. granulosus and E. multilocularis these antigens are basically found in the germinal layer of brood capsules and in the tegument of protoscolices, associated with larval stage. Gonzales et al. 2007 [20], showed that the TEG-Tsg gene of T. saginata is homologous to EM10 and EG10 genes of Echinococcus spp. and 97% identical to its T. solium homologue. However, alignment of this T. solium gene with the two clusters sequences described in the present study [GenBank:EX151048, GenBank:EX151014] clearly showed high sequence variability, despite the conserved blocks. The TEG molecules are characterized by an N-terminal FERM domain and a C-terminal ERM domain which are found in a number of cytoskeletal-associated proteins located at the interface between the plasma membrane and the cytoskeleton and in proteins interacting with lipid membranes. Thus TEG protein may play a role in tegument function and interaction with the host.

Genes of interest

A number of transcripts identified here could be of interest for further study (see Additional file 11). At least 30 genes coding for proteins potentially involved in parasite development, including transcriptional factors, component of chromatin remodelling complexes, cell adhesion-related molecules, receptors and other transducing components of signalling pathways have been identified. Moreover, putative orthologues of two proteins possibly associated to invertebrate immunity were identified for the first time in T. solium cysticerci: a ‘heat shock 90 kDa protein’ [GenBank:EL740676] and an ‘anaphylatoxin-like domain protein’ [GenBank:EX150322, GenBank:EX150873].

Fifty-one T. solium ORESTES revealed similarity with known antigens, including five previously characterized helminth antigens with potential for development of immunodiagnostic and/or vaccines. These are ‘paramyosin’ [GenBank:EL745866, GenBank:EL750686, GenBank:EL762552], ‘major egg antigen’ [GenBank:EL740635, GenBank:EL758824, GenBank:EL760346], ‘cathepsin L-like cysteine proteinase’ [GenBank:EL742569], ‘heat shock 70 kDa protein’ [GenBank:EL740975, GenBank:EL740984, GenBank:EL741400, GenBank:EL744608, GenBank:EL744338, GenBank:EL745376, GenBank:EL747548, GenBank:EL747588] and the ‘H17 g’ or ‘TEG-Tsol surface antigen’ [GenBank:AI581129], which is highly conserved among T. solium and T. saginata.

Discussion

Transcriptome investigations have greatly benefited from the recent maturation of gene expression approaches. Among these, the microarray has evolved as the most prominent high-throughput method to assess a given
Figure 2

(A) Distribution of the *Taenia solium* sequences according to the Gene Ontology sub-categorization concerning Biological Process. The predominant category of the Biological Process ontology was Cellular Process in *T. solium* ORESTES sequences. (B) Distribution of the *Taenia solium* sequences according to Gene Ontology sub-categorization Molecular Function. Among the Molecular Function Ontology it was possible to attribute the Binding and Catalytic activity function terms to sets of sequences, indicative of a close relationship between host and parasite gene regulation during the process of cyst establishment and the host response to infection.
expression profile. However, they are still subjected to hybridization issues such as reaction kinetics and probe mismatches as former methods. Also, microarrays cannot adequately address expression profiles of samples containing mixed species, which are yielded in studies of most parasite interactions. In these situations, the use of short gene tags, such as SAGE [21] is also problematic, due to the ambiguous tag-to-gene assignment and the difficulties of gene identification, especially in situations when the genome and/or the transcriptome of one of the species is not available. By comparison, the generation of longer sequence tags, such as those derived from EST or ORESTES, can facilitate gene discovery and annotation and also provides a much less ambiguous tag-to-gene mapping.

As formerly shown, ORESTES is able to give a normalized transcriptome view, as well as to characterize sequences from the central portion of the genes, including the less-abundant transcript markers [10,11,22-24]. The normalization capability of ORESTES, together with its ability to sample the central portion of genes makes this approach complementary to traditional ESTs, more frequently used in large-scale cDNA sequencing projects. Thus, as we have shown before for other species, including humans [11], S. mansoni [12,13], Drosophila melanogaster [25] or Apis mellifera [26], ORESTES provides a distinct contribution to gene discovery in T. solium. The present study shows the first comparative sequence analysis of the T. solium transcriptome using ORESTES from the larval stage (cysticerci).

Comparison of the T. solium ORESTES generated in this study with all T. saginata and T. solium sequences retrieved from GenBank showed identical hits with both datasets, indicating a high level of conservation in genes like 'Tsp36 small heat shock protein'. Few hits were obtained from other taeniids (T. crassiceps and T. asiatica), which may be due to their small sequence datasets or to the higher distance from T. solium and these other species. As an example of such intra-genus variability, T. asiatica is morphologically similar to T. saginata occurring in almost all Asian countries being capable of infecting pigs and humans [27] possibly leading to cysticercosis, but unlikely neurocysticercosis [28].

Since only 119 of the T. solium sequences described in this work showed similarity to T. solium Genome Project sequences, our results significantly contribute to the knowledge of the parasite expression profile by increasing the number of sequenced transcripts and through functional annotation of several genes. Thus, the present report is complementary to the T. solium genome initiative and may be helpful on the parasite genome assembly and annotation [29], as well as on studies of intra- and inter-specific sequence variability.

Considering the overall picture of the T. solium cysticerci transcriptome presented in this work, comparative sequence analysis revealed 350 sequences (43%) producing hits with a database. Despite the small dataset, it is interesting to note that Aguilar-Diaz et al. [1] found a very similar picture in the analysis of the transcriptome of adult worms, with 40% of the genes showing no hits. A systematic, functional investigation of these unknown genes using postgenomic tools such as "gene knockout" or RNA-mediated "knockdown" is desirable.

Several protein domains related to cell structure, including cell wall organization, were found among the generated sequences. The pistil-specific extensin-like and the vinculin/alpha-catenin motifs found in this study are of special interest due their role in cell wall structure and interaction. According to Interpro, the pistil-specific extensin-like protein motif [Interpro:IPR003882] is frequently found in the cell-wall proteins of many plants, and can account for up to 20% of their dry weight. Interestingly, this motif is also found in metazoans like Brugia malayi [Interpro:A8Q5T0/A8QD88] and C. elegans [Interpro:Q20327]. Since extensin-like proteins in plants are involved in cell wall strengthening in response to mechanical stress, such as attack by pests or plant-bending in the wind, it is reasonable to hypothesize a similar role on the T. solium cysts walls, conferring rigidity with a possible role in parasite defense.

The vinculin and/or alpha-catenin are eukaryotic actin-binding protein motifs, usually containing proline-rich motifs and several ligand-biding motifs. Vinculins are frequently used as markers for cell-cell and cell-extracellular matrix junctions, named as focal adhesions, also interacting with other structural proteins such as talin and alpha-actinins [30]. It is tempting to speculate that proteins containing these motifs may have a function on the organization of the cysticerci wall as well as on the interaction with host’s tissues.

Oxidative and other types of stress are inherent to the host environment to which a parasite is exposed. Therefore, proteins that allow the cysticerci to cope with stress may be important in infection maintenance. In this study several heat shock proteins (hsp16, hsp20, hsp25, hsp70, hsp86, and hsp90) and other stress response-related proteins have been identified as being transcribed by this developmental stage. Previous studies with T. solium cysticerci showed that the expression of 70 and 80 kDa heat shock proteins was highly induced under temperature stress [31]. Recently, another heat shock protein of 35.4 kDa was described for T. solium cysticerci and points out the importance of such proteins for the parasite life cycle [32].
The host immune response to tissue parasitism is an important aspect to the establishment and development of the neurocysticercosis pathology. In this study, the 'heat shock 90 kDa' protein and the 'anaphylatoxin-like domain' (a complement-associated protein in vertebrates) – which are described for the first time for the *T. solium* cysticerci – have been associated to a possible immune response in invertebrates [33,34] and along with the host immunity may be involved on the host-parasite immunological interplay.

Among several genes related to the antigenic coat of the parasite, the TEG-Tsol gene is of major importance for both immune diagnostic and vaccine development, due to its high antigenicity, strong similarity (~97%) between *T. solium* and *T. saginata* paralogs, conservation among other taeniid species and reactivity to distinct animal sera [20]. TEG-Tsol was found among the ORESTES in the other taeniid species and reactivity to its high antigenicity, strong similarity (~97%) between immune diagnostic and vaccine development, due to its high antigenicity, strong similarity (~97%) between *T. solium* and *T. saginata* paralogs, conservation among other taeniid species and reactivity to distinct animal sera [20]. TEG-Tsol was found among the ORESTES in the other taeniid species and reactivity to distinct animal sera [20].

Despite some encouraging results on vaccine development [39-41], several studies have pointed out intra- and inter-specific variability of taeniid species at both genotypic and phenotypic levels [26,42-50], which may represent a problem for the global-scale use of single- or multi-antigen recombinant vaccines. Thus, genome and transcriptome sequences – especially when derived from parasites collected at different endemic areas – are of major importance to address such variability and to point new vaccine and diagnostic/prognostic candidate markers. In this context, different from genomic markers, ESTs are powerful tools not only to indicate potentially relevant candidates, but also to provide experimental evidence of expression specific developmental stages.

**Conclusion**
The sequencing effort presented here is complementary to the *T. solium* Genome Project, having described several unknown genes for this species, which may have direct and immediate applications on diagnosis, therapeutics and/or vaccine development. Furthermore, this database represents a key resource to understanding aspects of the cysticerci biology and host-parasite interaction. Considering the ongoing efforts to sequence the hydatid disease agents (*E. granulosus* and *E. multilocularis*) along with the *T. solium* Genome Project [1,28], we hope our results can contribute to the development of comparative parasitic metazoon genomics, yielding new molecular diagnosis targets [51] and new insights into the pathogenesis of cysticercosis and taeniasis.

**Methods**

**Cysticerci collection**

Taenia solium cysticerci were collected from a naturally infected, landrace, bred pig (*Sus scrofa*). The animal was humanely sacrificed and cysticerci, spontaneously detached from abdominal and thoracic muscles were recovered and carefully micro-dissected to remove any tissue fragments that remained attached. Cysts were extensively washed with phosphate-buffered saline and immediately stored at -80°C. The study was previously approved by the Ethics Committee on Animal Research of the Faculty of Animal Science and Food Engineering (FZEA) of Universidade de São Paulo (USP), and was carried out following the institution's guidelines for animal husbandry.

**RNA extraction, RT-PCR and cDNA libraries preparation**

Total RNA was obtained from cysticerci using the Trizol® (Invitrogen, Carlsbad). Messenger RNA (mRNA) was purified using the μMACs mRNA isolation kit (Miltenyi Biotech, Bergisch Gladbach), following manufacturer's directions, as described [52]. mRNA concentration was evaluated by spectrophotometry (U-3010 Hitachi, Tokyo, Japan) and 25 ng mRNA aliquots were frozen for the posterior generation of ORESTES amplification profiles as described [12]. Briefly, cDNA was synthesized and amplified with some of the oligonucleotide primers previously used in the *S. mansoni* transcriptome project [12]. Twenty cDNA mini-libraries were constructed using ORESTES and a set of different oligonucleotide primers (see Additional file 12). The amplification profiles were evaluated in ethidium bromide-stained agarose gels, cloned in pGEM-T-Easy plasmids (Promega Corporation, Madison, USA) and used for *Escherichia coli* (strain DH10β) transformation. Recombinant clones were obtained by selective growth (X-Gal, IPTG and ampicillin), screened by PCR amplification of the insert using primers pGEM-F (5′-ACG CCA AGC TAT TTA GGT GAC ACT ATA-3′) and EXCEL-R (5′-GTT GTA AAA CGA CGG CCA GTG AAT-3′) and stored as glycerol stocks at -80°C. For sequencing, the bacterial clones were grown in LB medium for 20 hours at 37°C, followed by plasmid DNA extraction by alkaline lysis according to standard protocols [53].

**DNA sequencing and analysis**

ORESTES sequencing was carried out by two laboratories located at UFSC and USP using the DYEnamic® ET Dye Terminator kit (GE Healthcare, Fairfield) or ThermoSequenase II dye terminator cycle sequencing kit (Amerham-Pharmacia Biotech) in a MegaBace 1000® DNA Analysis System (GE Healthcare) and on an ABI PRISM® 3100 Genetic Analyzer (Applied Biosystems, Foster City), respectively. Briefly, each sequencing reaction used 5 pmol of pGEM-F or EXCEL-R oligonucleotides, and PCR products [54] or plasmid DNA as templates. The labeling
conditions were: 95°C/25 sec., 35 cycles of 95°C/15 sec., 50°C/20 sec. and 60°C/90 sec. The products were then precipitated (70% isopropanol), injected at 2 KV for 100 sec. and electrophoresed for 140 min. at 7 KV.

Sequence analysis was performed using the STINGRAY system (System for Integrated Genomic Resources and Analysis), an improved version of the formerly published GARSA system (Genomic Analysis Resources for Sequence Annotation) [55]. Briefly, the system workflow initially performs evaluation of the quality of the obtained chromatograms (cut-off Phred ≥ 15) following removal of vector sequences through Phred and Cross-match [56,57] and then clustering the sequences using CAP3 [58]. Following similarity searches performed by Blast (Basic Local Alignment Tool), Psi-Blast (Position-Specific Integrated Blast), RPSBlast (Reverse Position-specific Blast) [59], InterProScan [60] and HMMER (Hidden Markov Models for sequence profile analysis) [61] packages against local pre-formatted databases, blast analysis was also performed using all EST sequences from the *T. solium* Genome Project and the *Sus scrofa* genome available at GenBank. After removal of ribosomal RNA (rRNA) sequences, blast analysis against the *Sus scrofa* genome was used to separate parasite sequences from host sequences, creating two datasets that were evaluated separately. Functional annotation was performed using the Gene Ontology (GO) vocabulary as described by Jones et al. [62] and putative sub-cellular localization of each coding sequence was performed through the Wolf-PSORT program [63]. The G+C content of singlets and clusters was estimated by the GeeCee program (EMBOSS – European Molecular Biology Open Software Suite – package) and the tRNA sequences were predicted by tRNAscan-SE [64].

The results were then individually and manually checked during annotation, when sequences were validated as CDS when presenting i) high similarity values (e-value < e-15 and similarity>75%) with protein databases (uniprot_swissprot, uniprot_trembl, uniref90, refseq_protein) or with protein sequences from phylogenetically related organisms (Cestoda and/or Trematoda) available on GenBank, ii) the presence of conserved domains as revealed by RPS-Blast against CDD (see Additional file 6), COG (see Additional file 7) and KOG databases (see Additional file 8); iii) the presence of protein domains as revealed by InterProScan and HMMER and iv) annotations on Gene Ontology analysis, when available. *T. solium* sequences having no protein domain and showing exclusive hits with high similarity values (e-value < e-15 and similarity>75%) with ‘hypothetical proteins’ or ‘hypothetical conserved proteins’ from GenBank were annotated accordingly.

The *T. solium* cysticerci annotated transcripts, the host-parasite transcribed sequences, all databases used for comparative analysis as well as the additional material to this work are available online at the STINGRAY system http://stingray.biowebdb.org/index.cgi?project=TS.

**Abbreviations**

ESTs: Expressed Sequence Tags; ORESTES: Open Reading frame Expressed Sequence Tags; PCR: Polymerase Chain Reaction.

**Competing interests**
The authors declare that they have no competing interests.

**Authors’ contributions**

CRA and PHS are the main authors. CRA, PHS, GW, AAMM, ECG and EDN have equally contributed to this work. TCMS, GR, EBS, JBR, MMS, AZ, HBF, KT and AMRD have participated on the sequence analysis. All authors have participated on the manuscript preparation.

**Authors’ informations**

GW, EBS, GR and MMS are recipients of CNPq scholarships. PHS and TCMS are recipients of CAPES scholarships. ECG is currently a CNPq Post-Doctoral Fellow at BMRC/UEA, UK. EDN is a visiting scientist at MD Anderson Cancer Center, Houston TX, USA.

**Additional material**

### Additional file 1

**Table S1.** List of *Taenia solium* ESTs and their GenBank accession numbers.

Click here for file [http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S1.xls](http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S1.xls)

### Additional file 2

**Table S3.** Distribution of the *Taenia solium* EST according to their characteristics and similarity analysis.

Click here for file [http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S2.xls](http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S2.xls)

### Additional file 3

**Table S4.** Number of *Taenia solium* ORESTES according to their annotation as coding sequences (CDS).

Click here for file [http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S3.xls](http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S3.xls)

### Additional file 4

**Table S5.** Distribution of *Taenia solium* EST hits according to the Gene Ontology (GO) classification.

Click here for file [http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S4.xls](http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S4.xls)
Additional file 5
Table S6. Taenia solium ESTs with predicted extra-cellular sub-localization by Wolf-PSORT.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S5.xls]

Additional file 6
Table S7. Number of Taenia solium EST elicting hits with the CDD database.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S6.xls]

Additional file 7
Table S8. Number of Taenia solium EST elicting hits with the COG database.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S7.xls]

Additional file 8
Table S9. Number of Taenia solium EST elicting hits with the KOG database.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S8.xls]

Additional file 9
Picture S1. Distribution of the most frequent Eukaryotic Orthologous Groups categories observed for the Taenia solium sequences.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S9.ppt]

Additional file 10
Table S10. List of Taenia solium EST with similarity to the parasite genome project sequences.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S10.xls]

Additional file 11
Table S11. List of interesting genes found in Taenia solium ESTs.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S11.xls]

Additional file 12
Table S2. List of cDNA mini-libraries, primers used for generation of the Taenia solium EST profiles and the number of sequences obtained in each library.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1756-3305-2-35-S12.xls]

Acknowledgements
This work was supported by Associação Beneficente Alzira Denise Hertzog da Silva (ABADHS) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). The funders had no role in study design, data collection and analysis, decision to publish, or on the preparation of the manuscript.

References
1. Aguilar-Diaz H, Bobes RJ, Carrero JC, Camacho-Carranza R, Cervantes C, Cevallos MA, Davila G, Rodriguez-Dorantes M, Escobedo G, Fernandez JL, Fragoso G, Gaytan P, Garciaubio A, Gonzalez VM, Gonzalez L, Jose MV, Jimenez L, Laclette JP, Landa A, Larralde C, Morales-Montor J, Morett E, Ostoa-Saloma P, Sciutto E, Santamaria RJ, Soberon X, De La Torre P, Valdes V, Yanetz J. The genome project of Taenia solium. Parasitol Int 2006, 55:S127-130.
2. Scolozio F, Fragoso G, Laclette JP, Sotelo J, Aluja A, Vargas L, Larralde C. Taenia solium disease in humans and pigs: an ancient parasitosis disease rooted in developing countries and emerging as a major health problem of global dimensions. Microbes Infect 2000, 2:1875-1890.
3. Campbell G, Garcia HH, Nakao M, Ito A, Craig PS. Genetic variation in Taenia solium. Parasitol Int 2006, 55:S121-126.
4. Garcia HH, Gonzalez AE, Evans CA, Gilman RH. Taenia solium cysticercosis. Lancet 2003, 362:547-556.
5. Ito A, Nakao M, Wandra T. Human Taeniasis and cysticercosis in Asia. Lancet 2003, 362:1918-1920.
6. Román G, Sotelo J, Del Brutto O, Fliesser A, Dumas M, Wadía N, Botero D, Cruz M, Garcia H, De Bittencourt PR, Trelles L, Arriagada C, Lorenzana P, Nash TE, Spina-Franca A. A proposal to declare neurocysticercosis an international reportable disease. Bull World Health Organ 2000, 78:239-240.
7. Sinha S, Sharma BS. Neurocysticercosis: A review of current status and management. J Clin Neuroscience 2009, 16:867-876.
8. Fleury A, Dessein A, Preux PM, Dumas M, Tapia G, Larralde C, Sciutto E. Symptomatic human neurocysticercosis—age, sex and exposure factors relating with disease heterogeneity. J Neuro 2004, 251:830-837.
9. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merrill CR, Wu A, Olde B, Moreno RF, Kerlavay AR, McCombie WR, Venter JC. Complementary DNA sequences: expression sequence tags and human genome project. Science 1991, 252:1611-1616.
10. Dias-Neto E, Harrop R, Correa-Oliveira R, Wilson RA, Pena SD, Simpson AJ. Minilibraries constructed from cDNA generated by arbitrarily primed RT-PCR: an alternative to normalized libraries for the generation of ESTs from nanogram quantities of mRNA. Gene 1997, 186:33-42.
11. Dias-Neto E, Correa RG, Verjovski-Almeida S, Briones MR, Nagai MA, Da Silva W Jr, Zago MA, Bordin S, Costa FF, Goldman GH, Carvalho AF, Matsukuma A, Baia GS, Simpson DH, Brunstein A, Oliveira PS, Bucher P, Jongeneel CV, O'hare MJ, Soares F, Brentani RR, Reis LF, Souza S, Simpson AJ. Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. Proc Natl Acad Sci USA 2000, 97:3491-3496.
12. Verjovski-Almeida S, Demarco R, Martins EA, Guimaraes PE, Oippio EP, Paquola AC, Piazza JP, Nishiyama MY, Kitajima JP, Adamson RE, Ashton RD, Bonaldo MF, Coulson PS, Dillon GP, Farias LP, Gregorio SP, Ho PL, Leite RA, Malagias LC, Marques RC, Miyasato PA, Nascimento AL, Ohlweiler FP, Reis EM, Ribeiro MA, Sa RG, Stukart GC, Soares MB, Gargioni C, Kawano T, Rodrigues V, Madeira AM, Williams RA, Menck CF, Setubal JC, Leite LC, Dias-Neto E. Transciptome analysis of the acoelomate human parasite Schistosoma mansoni. Nature Genet 2003, 35:148-157.
13. Verjovski-Almeida S, Leite LC, Dias-Neto E, Menck CF, Wilson RA. Schistosome transcriptome: insights and perspectives for functional genomics. Trends Parasitol 2004, 20:304-308.
14. Lockyer AE, Spinks JN, Walker AJ, Kane RA, Rollinson D, Neuwald AF, Matzke MA, Dujon B, Ashburner M, Flicek P, Birney E. ORESTES (Organism Resource for ESTs): identification of cell-signalling, transcriptional control and immune-related genes from open reading frame expressed sequence tags (ORESTES). Developmental and Comparative Immunology 2007, 31:763-782.
15. Hu W, Yan Q, Shen DK, Liu F, Zhu ZD, Song HD, Xu XR, Wang ZJ, Rong YP, Zeng LC, Wu J, Zhang X, Wang Jj, Xu XN, Wang G, Zhang XL, Wang ZQ, Brindley PJ, Mcmanus DP, Xue CL, Feng Z, Chen Z, Han ZG. Evolutionary and biomedical implications of a Schistosoma japonicum complementary DNA resource. Nature Genet 2003, 35:139-147.
Page 10 of 11

http://www.parasitesandvectors.com/content/2/1/35
40. Flisser A, Guacci CG, Zoli A, Martínez-Ocana J, Garza-Rodríguez A, Domínguez-Alpizar JL, Maravilla P, Rodriguez-Canul R, Avila-G, Aguilar V, Vega L, Kyongdon C, Geerts T. Lightowers MVW. Induction of protection against porcine cysticercosis by vaccination with recombinant oncosphere antigens. Infect Immun 2004, 72:5292-5297.

41. Flisser A, Rodriguez-Canul R, Willingham AL 3rd: Control of the taeniosis/cysticercosis complex: future developments. Vet Parasitol 2006, 139:283-292.

42. Fernandez S, Costa AC, Katsuyama AM, Madeira AM, Gruber A. A survey of the inter- and intraspecific RAP markers of Eimeria spp. of the domestic fowl and the development of reliable diagnostic tools. Parasitol Res 2003, 89:437-445.

43. Ferrer E, Bonay P, Foster-Cuevas M, Gonzalez LM, Davila I, Cortez MM, Harrison LJ, Parkhouse RM, Garate T. Molecular cloning and characterization of Ts8B1, Ts8B2 and Ts8B3, three new members of the Taenia solium metacestode 8 kDa diagnostic antigen family. Mol Biochem Parasitol 2007, 152:90-100.

44. Gonzalez LM, Bonay P, Benitez L, Ferrer E, Harrison LJ, Parkhouse R, Garate T. Molecular and functional characterization of a Taenia adhesion gene family (TAF) encoding potential protective antigens of Taenia saginata oncospheres. Parasitol Res 2007, 100:519-528.

45. Gonzalez LM, Montero E, Sciutto E, Harrison LJ, Parkhouse RM, Garate T. Differential diagnosis of Taenia saginata and Taenia solium infections: from DNA probes to polymerase chain reaction. Transac Roy Soc Trop Med Hyg 2002, 96:5243-250.

46. Jeon HK, Eom KS: Taenia asiatica and Taenia saginata: genetic divergence estimated from their mitochondrial genomes. Exp Parasitol 2006, 113:58-61.

47. Maravilla P, Souza V, Valera A, Romero-Valdivinos M, Lopez-Vidal Y, Domínguez-Alpizar JL, Ambrosio J, Kawa S, Flisser A: Detection of genetic variation in Taenia solium isolates from Mexico and Colombia. Mem Inst O Cruz 2003, 98:1049-1050.

48. Maravilla P, Valera A, Souza V, Martinez-Gordillo M, Flisser A: Isozyme analysis of Taenia solium isolates from Mexico and Colombia. Mem Inst O Cruz 2003, 98:1049-1050.

49. Sciutto E, Ross G, Hernandez M, Morales J, Cruz-Revilla C, Toledo A, Manoutcharian K, Gevorkian G, Blancas A, Acero G, Hernandez B, Cervantes J, Boves RJ, Goldbaum FA, Huerta M, Diaz-Orea A, Fleury A, De Aluja AS, Cabrera-Ponce JL, Herrera-Estrella L, Fragozo G, Larrañeta C. Improvention of the synthetic tri-peptide vaccine (S3Pvac) against porcine Taenia solium cysticercosis in search of a more effective, inexpensive and manageable vaccine. Vaccine 2007, 25:1368-1378.

50. Vega R, Pinero D, Ramananandrasana B, Dumas M, Bouteille B, Fleury A, Sciutto E, Larrañeta C, Fragozo G. Population genetic structure of Taenia solium from Madagascar and Mexico: implications for clinical profile diversity and immunological technology. Int J Parasitol 2003, 33:1479-1485.

51. Almeida CR, Ojopi EP, Nunes CM, Machado LR, Takayanagi OM, Livramento JA, Abraham R, Gattaz WF, Vaz AJ, Dias-Neto E. Taenia solium DNA is present in the cerebrospinal fluid of neurocysticercosis patients and can be used for diagnosis. Eur Arch Psychiat Clin Neurosc 2006, 256:307-310.

52. Ojopi EP, Oliveira PS, Nunes DN, Paquola A, Demarco R, Gregorio SP, Aires KA, Menck CF, Leite LC, Verjovski-Almeida S, Dias-Neto E. A quantitative view of the transcriptome of Schistosoma mansoni adult-worms using SAGE. BMC Genomics 2007, 8:186.

53. Sambrook J, Russel D: Molecular Cloning: A Laboratory Manual Cold Spring Harbor Laboratory Press; 2001.

54. Silva WA, Costa CA, Valente V, Sousa JF, Paco-Larson ML, Esprefaco EM, Camargo SS, Monteiro E, Holanda AJ, Zago MA, Simpson AJ, Dias-Neto E. PCR template preparation for capillary DNA sequencing. Biotechniques 2001, 30:537-542.

55. Davila AM, Lorenzini DM, Mendes PN, Satake TS, Sousa GR, Campos LM, Maczczoni CJ, Wagen G, Pires PF, Grissard EC, Cavalcanti MC. Campos ML: GARSA: genomic analysis resources for sequence annotation. Bioinformatics 2005, 21:4302-4303.

56. Ewing B, Green P: Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998, 8:186-194.

57. Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. III. Accuracy assessment. Genome Res 1998, 8:175-185.

58. Huang X, Madan A: CAP3: A DNA sequence assembly program. Genome Res 1999, 9:886-877.

59. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 1997, 25:3389-3402.

60. Mulder N, Apweiler R: InterPro and InterProScan: tools for protein sequence classification and comparison. Meth Mol Biol 2007, 349:59-70.

61. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL: Pfam 3.1: 1313 multiple alignments and profile HHMMS match the majority of proteins. Nucl Acids Res 1999, 27:260-262.

62. Jones CE, Baumann U, Brown AL: Automated Methods of predicting the function of biological sequences using GO and BLAST. BMC Bioinformatics 2005, 6:272.

63. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: WoLF PSORT: protein localization predictor. Nucl Acids Res 2007, 35:W585-587.

64. Schattner P, Brooks AN, Lowe TM: The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucl Acids Res 2005, 33:W686-689.