Acta Tropica 159 (2016) 132–141

Contents lists available at ScienceDirect

Acta Tropica

journal homepage: www.elsevier.com/locate/actatropica

De novo assembly and characterization of the *Trichuris trichiura* adult worm transcriptome using Ion Torrent sequencing

Leonardo N. Santos[a], Eduardo S. Silva[a], André S. Santos[a], Pablo H. De Sá[b], Rommel T. Ramos[b], Artur Silva[b], Philip J. Cooper[c,d], Maurício L. Barreto[e,f], Sebastião Loureiro[e], Carina S. Pinheiro[a], Neuza M. Alcantara-Neves[a,*], Luis G.C. Pacheco[a,*]

[a] *Institute of Health Sciences, Federal University of Bahia, Salvador, BA, Brazil*
[b] *Institute of Biological Sciences, Federal University of Pará, Belém, PA, Brazil*
[c] *Institute of Infection and Immunity, St. George's University of London, London, UK*
[d] *Centro de Investigacion en Enfermedades Infecciosas y Cronicas, Pontificia Universidad Catolica del Ecuador, Quito, Ecuador*
[e] *Institute of Public Health, Federal University of Bahia, Salvador, BA, Brazil*
[f] *Centro de Pesquisas Gonçalo Muniz, FIOCRUZ-BA, Salvador, BA, Brazil*

ARTICLE INFO

Article history:
Received 10 July 2015
Received in revised form 23 March 2016
Accepted 30 March 2016
Available online 31 March 2016

Keywords:
Trichuris trichiura
Transcriptome
Next-generation sequencing
Functional genomics

ABSTRACT

Infection with helminthic parasites, including the soil-transmitted helminth *Trichuris trichiura* (human whipworm), has been shown to modulate host immune responses and, consequently, to have an impact on the development and manifestation of chronic human inflammatory diseases. *De novo* derivation of helminth proteomes from sequencing of transcriptomes will provide valuable data to aid identification of parasite proteins that could be evaluated as potential immunotherapeutic molecules in near future. Herein, we characterized the transcriptome of the adult stage of the human whipworm *T. trichiura*, using next-generation sequencing technology and a *de novo* assembly strategy. Nearly 17.6 million high-quality clean reads were assembled into 6414 contiguous sequences, with an N50 of 1606 bp. In total, 5673 protein-encoding sequences were confidentially identified in the *T. trichiura* adult worm transcriptome; of these, 1013 sequences represent potential newly discovered proteins for the species, most of which presenting orthologs already annotated in the related species *T. suis*. A number of transcripts representing probable novel non-coding transcripts for the species *T. trichiura* were also identified. Among the most abundant transcripts, we found sequences that code for proteins involved in lipid transport, such as vitellogenins, and several chitin-binding proteins. Through a cross-species expression analysis of gene orthologs shared by *T. trichiura* and the closely related parasites *T. suis* and *T. muris* it was possible to find twenty-six protein-encoding genes that are consistently highly expressed in the adult stages of the three helminth species. Additionally, twenty transcripts could be identified that code for proteins previously detected by mass spectrometry analysis of protein fractions of the whipworm somatic extract that present immunomodulatory activities. Five of these transcripts were amongst the most highly expressed protein-encoding sequences in the *T. trichiura* adult worm. Besides, orthologs of proteins demonstrated to have potent immunomodulatory properties in related parasitic helminths were also predicted from the *T. trichiura de novo* assembled transcriptome.

1. Introduction

Trichuris trichiura (human whipworm) is a soil-transmitted helminth of high public health relevance and an estimated 5 billion humans are at risk of stable transmission with this parasite, of whom nearly 1 billion are school-aged children (Pullan and Brooker, 2012). Adult whipworm measures between 30 and 50 mm in length and colonizes the human large intestine, where it may

Abbreviations: BLAST, Basic Local Alignment Search Tool; cDNA, complementary DNA; DHPLC, denaturing high-performance liquid chromatography; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; NCBI, National Center for Biotechnology Information; NR, non-redundant; rRNA, ribosomal RNA.

* Corresponding authors.
E-mail addresses: neuza@ufba.br (N.M. Alcantara-Neves), luis.pacheco@ufba.br (L.G.C. Pacheco).

http://dx.doi.org/10.1016/j.actatropica.2016.03.036
0001-706X/© 2016 Published by Elsevier B.V.

live for up to two years, with females laying up to 5000 eggs a day (Bethony et al., 2006). Clinical presentation of trichuriasis typically includes intestinal manifestations (diarrhea, abdominal pain), general malaise, weakness and impaired cognitive and physical development (WHO, 2014).

Infection with helminthic parasites, including whipworms, has been shown to modulate host immune responses and, consequently, to have an impact on the development and manifestation of chronic human inflammatory diseases (Bashi et al., 2015; Maizels et al., 2014; Mishra et al., 2014). Specifically, epidemiological studies have clearly demonstrated an inverse association between *T. trichiura* infection and Type 1 skin hypersensivity to aeroallergens in children (Rodrigues et al., 2008). Besides, the therapeutic potential of *Trichuris suis* (pig whipworm) ova has been demonstrated in some groups of patients with inflammatory bowel diseases, such as Crohn's disease and ulcerative colitis, and in patients with relapsing-remitting multiple sclerosis (Fleming et al., 2011; Summers et al., 2003, 2005a, 2005b).

In an effort to better characterize the immunomodulatory potential of human whipworm proteins, our group conducted a previous study in which several fractions of the *T. trichiura* adult worm somatic extract (hereafter refereed as TtEFs) were evaluated for their immunomodulatory effects on cytokine responses by human peripheral blood mononuclear cells (PBMCs) cultivated in the presence of these antigens (Santos et al., 2013). Six TtEFs were identified that were able to promote greater production of the immune regulatory cytokine interleukin (IL)-10, when cultured with PBMCs, and were also able to inhibit helper T cell 1 (T_H1) and T_H2 cytokine production by PBMCs when co-cultured with optimal stimuli (*e.g.* phytohaemagglutinin for Th2 cytokines) (Santos et al., 2013). Additional characterization of these protein fractions (TtEFs) by nano-liquid chromatography/mass spectrometry identified several proteins, most of which had orthologs in the related parasite *Trichinella spiralis* and were considered to be promising candidates for future evaluation as immunomodulatory molecules (Santos et al., 2013). To take this research a step further to allow us to engineer these parasite molecules as recombinant proteins, we did a transcriptomic study of *T. trichiura* adult worm to identify the genes that code for these immunomodulatory proteins, taking advantage of the recently published draft genome assembly for *T. trichiura* and the reference genome and transcriptome for the mouse parasite *T. muris* (Foth et al., 2014). We report here, to our knowledge, the first transcriptomic analysis of the species *T. trichiura*, which will contribute to the functional annotation of the recently released draft genomic sequence (Foth et al., 2014) and will aid discovery of *T. trichiura* molecules that may possess immunomodulatory properties.

2. Materials and methods

2.1. Collection of parasites and total RNA extraction

Adult *T. trichiura* worms were obtained by treatment of infected children from the province of Esmeraldas–Ecuador with pyrantel pamoate and collection of worms from stool samples for 1–2 days after treatment, as described (Meekums et al., 2015). Ethical approval was provided by the Ethical Committee of the Universidad San Francisco de Quito–Ecuador and written informed consent was provided by parents or guardians. Adult worms were washed carefully in 0.15 M phosphate-buffered saline (pH 7.4) and total RNA was extracted using a standard Trizol (Life technologies) protocol with the aid of zirconium/silica beads (BioSpec Products, Inc.). RNA quality was verified spectrophotometrically (260 nm/280 nm ratio = 1.95) and quantitation was done fluorometrically using the Qubit® RNA Assay kit (Life Technologies).

2.2. Depletion of rRNA by DHPLC and preparation of the cDNA library

One hundred micrograms of total RNA were subjected to depletion of ribosomal RNA by denaturing high-performance liquid chromatography (DHPLC) using a RNASeP™ Cartridge (Transgenomic®) column in a Wave® System 4500 (Transgenomic®) machine, as described in details elsewhere (Castro et al., 2013). rRNA-depleted total RNA was then used for preparation of the cDNA library with the Ion Total RNA-Seq Kit v2 (Life Technologies), according to the manufacturer's recommendations.

2.3. Ion Torrent sequencing and *de novo* assembly of the transcriptome

Transcriptome sequencing was achieved using the Ion Torrent Personal Genome Machine™ (PGM) and the Ion 318™ chip, according to the manufacturer's recommendations (Life Technologies). A total of three replicate sequencing runs were performed and the raw reads were merged into a single FASTQ file; short reads (<50 nt) were removed and the remaining sequences were trimmed at both 5′ and 3′ ends with the FastX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). Quality checking of the sequences was done with the FastQC tool through the RNA-Rocket platform (Warren et al., 2015).

De novo assembly of the clean reads was generated with the MIRA 4.0 assembler (Chevreux et al., 2004). The *T. trichiura* transcriptome short reads dataset is publicly accessible through the European Nucleotide Archive (ENA), under accession PRJEB12315. The *de novo* assembled transcripts are available as a Supplementary material (Supplementary file S1).

Trimmed reads were also mapped against the recently released draft genome sequence of *T. trichiura* v2.1 (Foth et al., 2014) using the Ion Torrent mapper (TMAP) strategy and the CLC Genomics Workbench. Then, we associated the locus tags with their respective gene products according to *T. trichiura* gene annotation v. 2.2 (Foth et al., 2014).

2.4. Functional annotation of the transcriptome

The Blast2GO tool v.3.1 (http://www.blast2go.org) and the TRAPID (Rapid Analysis of Transcriptome Data) tool (Van Bel et al., 2013) were employed to retrieve functional annotations for the *T. trichiura* assembled transcripts. Firstly, BLASTx similarity searches were performed against the non-redundant (nr) database of NCBI (E-value: <1E^{-05}) using the *de novo* assembled transcripts as queries; conserved protein domains of the predicted proteins and assignment of the sequences to gene families were performed through searches against the OrthoMCLDB 5.0 data source (Van Bel et al., 2013). Then, all assembled transcripts were manually annotated using the Translate tool (http://web.expasy.org/translate/) followed by BLASTp similarity searches against NCBI's nr (E-value: <1E^{-05}) or the HelmDB database (http://gasser-research.vet.unimelb.edu.au/helmdb/). We obtained Gene ontology (GO) functional classifications (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations (Kanehisa et al., 2014) for the entire set of proteins predicted from the manually annotated transcriptome, through the Blast2GO v. 3.1 suite using default parameters and the most recent releases of the GO and KEGG databases as per December 2015. All sequences coding for as-yet-uncharacterized proteins were further characterized through searches against the PROSITE collection of motifs, using ScanProsite (http://prosite.expasy.org/scanprosite/). Prediction of probable signal peptides was performed with the SignalP 4.1 server (http://www.cbs.dtu.dk/services/SignalP/).

2.5. Identification of candidate novel protein-encoding sequences and non-coding transcripts

Two different strategies were used to identify candidate novel transcripts for the species *T. trichiura*: (i) firstly, all *de novo* assembled transcripts that remained unannotated after BLASTx searches against NCBI's NR Protein database were then scanned for ORFs with the AUGUSTUS gene prediction program, as described (Stanke et al., 2006; Chen et al., 2015). Sequences for which no ORFs could be detected were considered as potential non-coding transcripts; all these transcripts were used in BLASTn searches against the complete genome sequence of *T. trichiura* v.2.1 through the recently released WormBase Parasite Platform (Howe et al., 2016) in order to confirm that these novel transcripts do align with the reference genome but do not fall into known gene models for the species. (ii) In a second strategy, following manual annotation of protein encoding transcripts, all proteins predicted from the *T. trichiura* transcriptome in this study were then compared with the entire set of proteins predicted for the species in genome annotation v.2.2, using the CD-HIT program (Fu et al., 2012) with a 70% protein identity cut-off. Sequences that did not match proteins predicted from known *T. trichiura* gene products were then compared with the entire protein set for the related species *T. suis* (Jex et al., 2014), in order to detect potential protein orthologs.

2.6. Analysis of T. trichiura transcripts that code for proteins with immunomodulatory activities

tBLASTn searches (E-value: $<1E^{-05}$) were performed against the *T. trichiura* transcriptome dataset using the amino acid sequences of the proteins that were identified previously through mass spectrometry analysis of six chromatographic fractions of the whipworm somatic extract (TtEF 6, TtEF 8, TtEF 9, TtEF 10, TtEF11 and TtEF 12) and which showed *in vitro* immunomodulatory activities (Santos et al., 2013). These proteins had been identified in the previous study using ProteinLynx Global Server (PLGS) searches against protein databases of related parasitic worms, in particular *T. spiralis* (Santos et al., 2013).

Three-dimensional protein structures were modeled using the SWISS-MODEL server (Biasini et al., 2014) and validation was performed using Prosa-web (Wiederstein and Sippl, 2007) and the STING Millenium Suite (Higa et al., 2004). Multiple structural alignment was obtained with MUSTANG (Multiple Structural Alignment Algorithm) (Konagurthu et al., 2006) and visualization of the protein structures was performed using UCSF Chimera software, version 1.6.1 (Pettersen et al., 2004).

2.7. Transcript abundance estimation and cross-species gene expression analysis

The expression levels of the *T. trichiura* transcripts were estimated both with the CLC Genomics Workbench (CLC bio) and the NextGENe® software v2.4 (SoftGenetics), following mapping of the entire set of trimmed reads against the *ca.* 75 Mbp draft genome assembly of *T. trichiura* v.2.1 and gene annotations v.2.2 (GenBank: GCA_000613005.1). The numbers of reads mapping to the existing gene models for the species were calculated and then normalized to RPKM (Reads Per Kilobase per Million mapped reads) values. A RPKM threshold value of 0.3 was set to confidentially detect the presence of a transcript for a specific protein encoding gene.

A cross-species gene expression analysis was performed for comparing the transcript abundance of gene orthologs shared by *T. trichiura* and the closely related parasites *T. suis* and *T. muris*. For this, the RNA-seq short read files from *T. suis* and *T. muris* adult worm transcriptomes were retrieved from ArrayExpress (accession number: E-ERAD-125) and the Sequence Read Archive (accession numbers: SRR1041654 and SRR1041655). Then, CLC Genomics Workbench (CLC bio) was used to map the reads against a reference composed only of *T. trichiura* predicted gene sequences (Foth et al., 2014). The alignment parameters used for these analyses were as follows: (i) mismatch cost = 2; (ii) insertion cost = 3; (iii) deletion cost = 3; (iv) length fraction = 0.7; and (v) similarity fraction = 0.8. Results were normalized to RPKM values. For confirmation of expression profiles observed for specific transcripts, the *T. trichiura* genes were mapped back to their corresponding orthologs in the other two species, and the specific transcript expression levels were obtained from the works by Jex and collaborators (2014) and Foth et al. (2014).

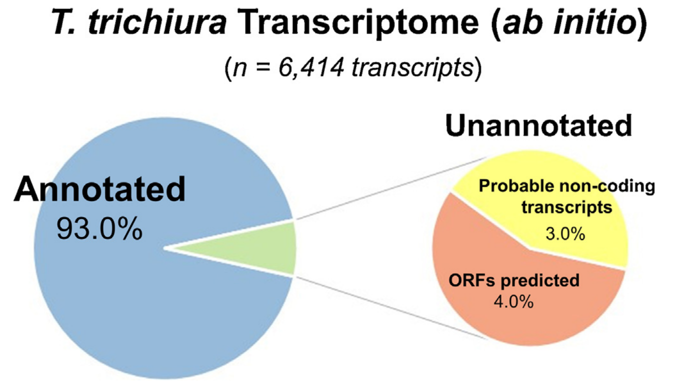

Fig. 1. Automatic annotation of the *T. trichiura de novo* assembled transcripts. All transcript sequences were automatically annotated using BLASTx searches against NCBI's non-redundant protein database. Sequences for which no BLAST hits could be automatically retrieved (unannotated) were then scanned for the presence of ORFs, using AUGUSTUS. All these transcripts were also used in BLASTn searches against the *T. trichiura* draft genome assembly v.2.1, through the WormBase Parasite platform.

3. Results and discussion

3.1. Sequencing and de novo assembly of the T. trichiura transcriptome

Nearly 19 million raw reads (accumulated length of 2,642,638,062 bp) were obtained by sequencing of the *T. trichiura* cDNA library. Trimming and quality assessment resulted in approximately 17.6 million (93.6%) high-quality clean sequences, with most of the reads above the Q20 level (Table 1; Fig. S1). *De novo* transcriptome assembly yielded 6414 contigs with an N50 of 1606 bp. The average contig size of the transcriptome assembled *ab initio* was 1404 bp, and contig length ranged from 183 bp to 9499 bp (Table 1). Trimmed reads were also mapped against the recently released draft genome sequence of *T. trichiura* v2.1 (Foth et al., 2014) and 92.08% of the reads mapped to the reference; 81.64% of the reads mapped to predicted protein coding sequences (Table 1), which demonstrates a good quality of the RNA extraction protocol and of the cDNA library preparation (Sultan et al., 2014). Positional analysis of specific transcripts was obtained through the WormBase Parasite Platform (Howe et al., 2016), following alignment to the *T. trichiura* reference genome v.2.1; this analysis showed good concordance of the *de novo* assembled transcriptome with the existing gene models for the species (Fig. S2A–E).

3.2. Protein encoding sequences and candidate novel transcripts

The great majority of the 6414 *T. trichiura de novo* assembled transcripts represented sequences for which it was possible to retrieve automatic annotations following BLASTx searches against known protein sequences in public databases (Fig. 1). Among the remaining 440 transcripts with no BLASTx hits, it was possible to

Table 1
Summary of results of *T. trichiura* adult transcriptome sequencing and assembly.

Statistics	*T. trichiura* transcriptome dataset
Number of raw reads	18,838,891
Number of filtered reads	17,640,541
Total base pairs (bp) (raw data)	2,642,638,062
Total base pairs (bp) (filtered data)	2,360,442,157
Sequence length range in nucleotides	50–362
Average GC content	49.0%
Number of contigs assembled *ab initio*	6414
Contig length range	183 bp–9499 bp
Average contig length	1404 bp
N50 contig size	1606 bp
Percent of reads above Q20 level	97.6%
Reads mapped to protein coding regions[a]	81.64%
Reads mapped to non-coding regions[a]	10.44%
Number of transcripts assigned to specific gene families[b]	4935 (76.9%)
Number of transcripts coding for recognizable protein domains[b]	4154 (64.7%)
Number of proteins predicted from the manually curated transcriptome	5673
Number of trancripts coding for proteins identified in immunomodulatory fractions of the *T. trichiura* protein extract[c]	20

[a] Reads were mapped against the draft genome assembly of *T. trichiura* v2.1, in which 9650 genes were predicted for this species (Foth et al., 2014).
[b] As predicted by the TRAPID tool (Van Bel et al., 2013), using a OrthoMCLDB 5.0 data source.
[c] Based on the immunomodulatory protein fractions identified in the work by Santos et al. (2013).

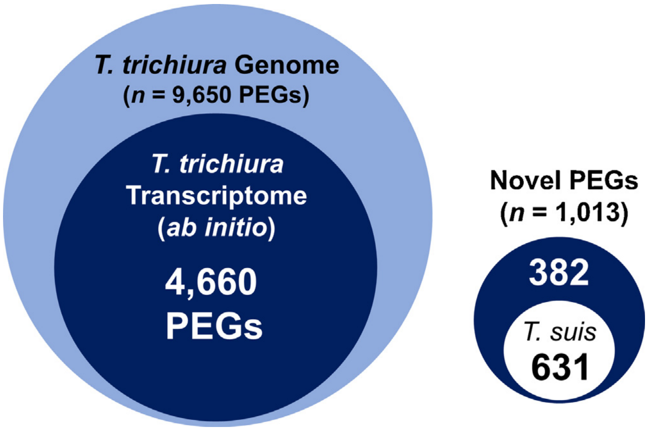

Fig. 2. Proteins predicted from the *T. trichiura* transcriptome and candidate novel protein-encoding genes (PEGs). The *de novo* assembled transcripts were manually inspected for the presence of protein coding sequences using the Expasy's Translate tool, followed by BLASTp similarity searches against NCBI's NR or the HelmDB database. All predicted proteins were also matched with correspondent proteins predicted from the *T. trichiura* annotated genomic sequence. Unmatched proteins were then analyzed against proteins predicted for the related parasite *T. suis*.

predict novel open reading frames (ORFs) for 249 sequences (Fig. 1; Table S1). One hundred and ninety-one transcripts rendered no hits with known protein sequences but aligned with high identities (mean = 97.0%) with the *T. trichiura* draft genome assembly, representing probable novel non-coding transcripts (Fig. 1; Table S2). Of these, only seven were in the neighborhoods of known protein encoding genes, considering a 100-nucleotides window size (Table S2). Five sequences produced significant alignments with non-coding RNAs present in the NONCODE2016 database (Zhao et al., 2016) (Table S2). One hundred and seventy-one of these transcripts are longer than 200 nucleotides, representing then candidate long non-coding RNAs of the species (Table S2).

After manual curation of the *de novo* assembled transcripts, 5673 protein encoding sequences were identified in the *T. trichiura* transcriptome. Of these, 4660 matched known proteins predicted from the *T. trichiura* genome v.2.1 (Fig. 2). Of the remaining 1013 protein encoding sequences identified in the manually curated transcriptome, 631 matched proteins predicted from the genome of the related species *T. suis* (Fig. 2) (Supplementary file S2).

The discovery of potential novel protein encoding genes (PEGs) from the *T. trichiura* transcriptome, when compared to closely related species, is not unexpected given the fact that only 9650 genes have been predicted from the *T. trichiura* draft genome assembly v.2.1, whilst 14,261 and 11,004 PEGs where predicted for *T. suis* and *T. muris*, respectively (Jex et al., 2014; Foth et al., 2014). A total of 7431 gene orthologs are shared by the latter two species as predicted in the current gene models available in WormBase Parasite release 5, whereas only *ca*. 5700 PEGs of the current *T. trichiura* gene annotation are shared with the other two species. Importantly, transcriptomic studies were already available for *T. suis* and *T. muris* (Cantacessi et al., 2011; Jex et al., 2014; Foth et al., 2014), and then aided prediction of PEGs in the newly generated genomic sequences of these species.

3.3. Functional annotation and classification

Most of the 6414 *T. trichiura* assembled transcripts represented sequences that code for proteins predicted from the recently available genomic sequences for the species *T. trichiura* and *T. suis* (Fig. 3A), with various orthologs shared by the related parasite *T. spiralis* and by filarial nematodes (*Loa loa* and *Brugia malayi*) (Fig. 3B). Among these, 4935 transcripts could be assigned to 2683 unique gene families (Table 1). A total of 22,799 GO functional annotations were obtained for the predicted protein coding sequences;in the categories representing biological processes (73.0%) and molecular functions (27.0%) (Fig. 4A and B). Among the sequences annotated at the biological processes level, the most highly scored categories were cellular and metabolic processes (3001 and 2885 sequences, respectively), whereas 114 predicted proteins were classified in immune system processes and 194 sequences in reproductive processes (Fig. 4A). At the molecular functions level, the *T. trichiura* sequences were distributed in eleven categories, with 378 predicted proteins involved with transporter activities and most of the remaining proteins (2842) involved in binding functions (Fig. 4B).

T. trichiura predicted proteins were also submitted to biological pathway analysis by mapping against unique KEGG orthologs. We could identify several transcripts that code for enzymes involved in metabolic modules that are conserved among parasitic nematodes and that have been demonstrated to be highly activated in adult worms (Tyagi et al., 2015), such as reactions involved in purine metabolism. In total, the predicted proteins could be assigned to 88 different biochemical pathways (Table S3). The predominant pathways were: “Purine metabolism” (ko00230, 162 sequences), “Pyrimidine metabolism” (ko00240, 65 sequences),

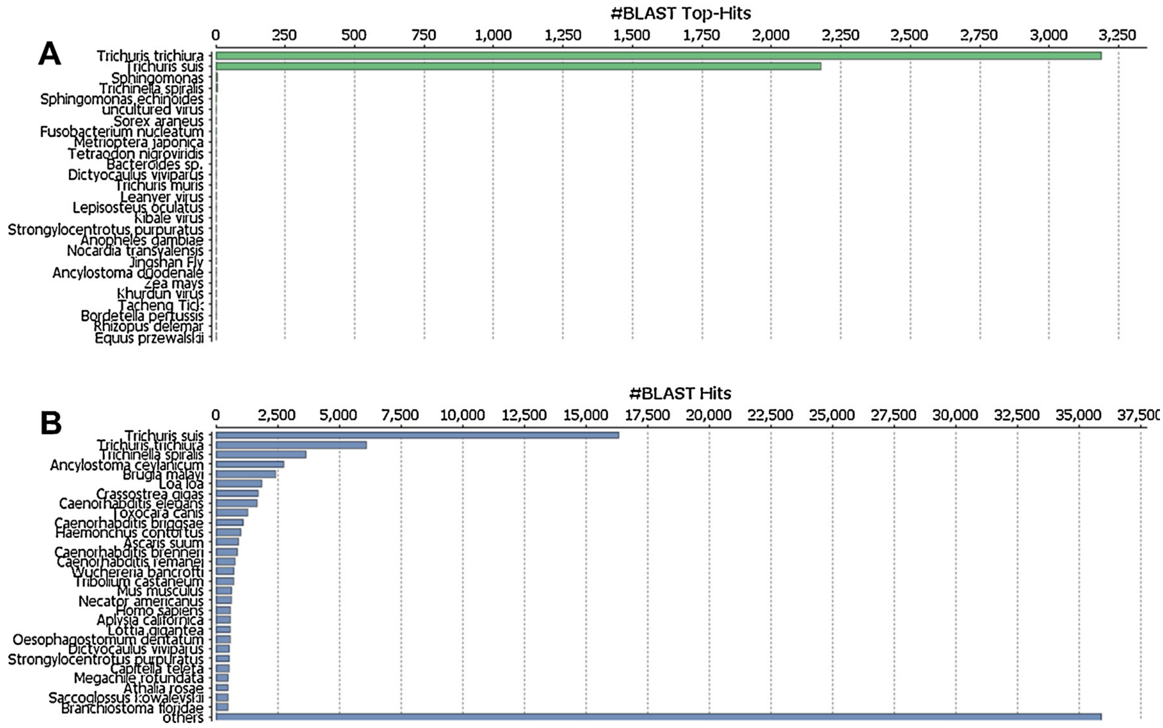

Fig. 3. Distribution of top BLAST hits found for proteins predicted from the *T. trichiura* transcriptome. (A) Species distribution of the top BLAST hits found for each *T.trichiura* protein, as determined through BLASTp similarity searches of the translated transcriptome sequences against the non-redundant database of NCBI (E-value cut-off: <1E^{-05}) (B) distribution of the total hits found for *T.trichiura* proteins.

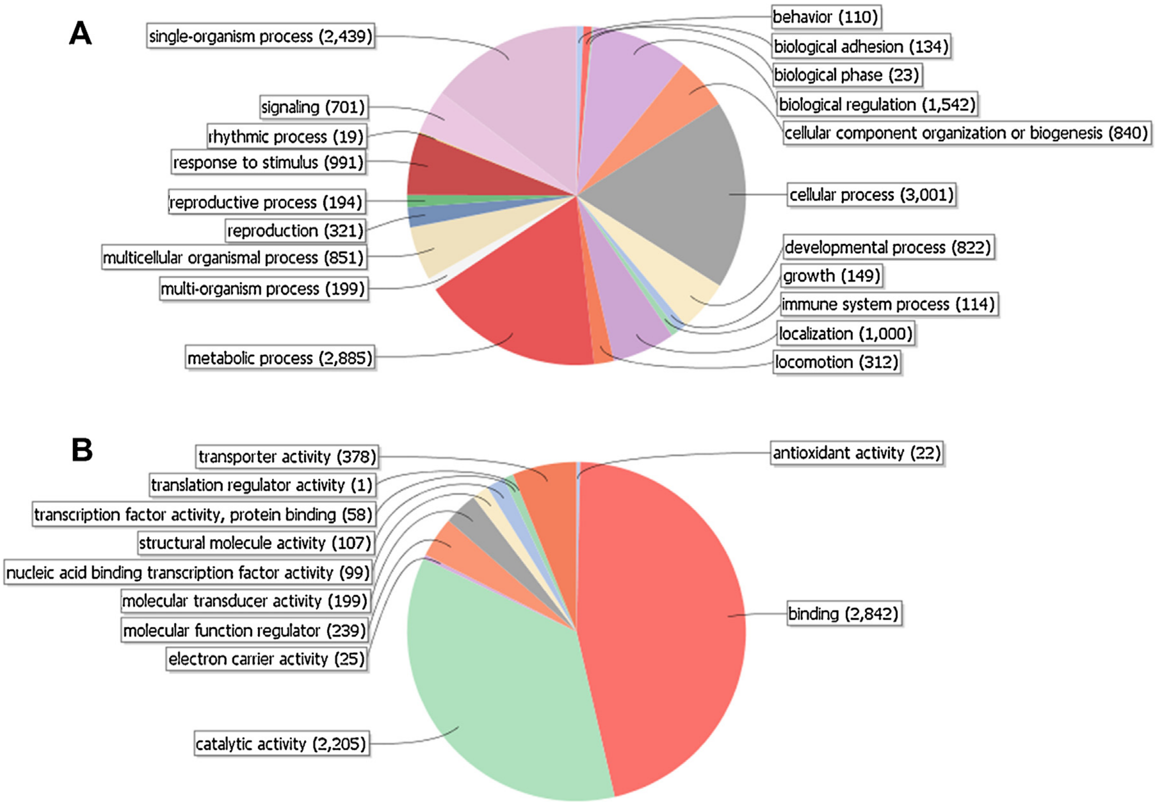

Fig. 4. Gene Ontology (GO) functional annotations of proteins predicted from the *T. trichiura* transcriptome. GO classifications of the predicted proteins are summarized in two main categories, at second level: (A) biological processes and (B) molecular functions.

“Aminoacyl-tRNA biosynthesis” (ko00970, 58 sequences), and “Phosphatidylinositol signaling system” (ko04070, 41 sequences) (Table S3).

3.4. Annotation of transcripts that code for proteins of uncharacterized functions

In total, 2096 (*ca.* 35.0%) of the protein-encoding transcripts identified in the *T. trichiura* transcriptome code for uncharacterized proteins (annotated as either ‘unknown-function proteins’ or ‘hypothetical proteins’). This proportion of uncharacterized proteins reflects the overall numbers predicted form the recently available genomic sequences of *T. trichiura* and *T. muris* (Foth et al., 2014). Eighty-four of these proteins were seen to possess signal peptides. By searching for PROSITE signatures, using the ScanProsite approach (de Castro et al., 2006), we were able to obtain additional functional annotations for 607 of the predicted *T. trichiura* uncharacterized proteins. It was possible to identify 1522 PROSITE hits in these proteins, belonging to 337 different domains, families and functional sites. Twenty-seven of these hits were present in more than 5 uncharacterized proteins, as shown in Fig. S3A. The zinc-finger domain was the most prevalent and was found in 48 different uncharacterized proteins (Fig. S3A). Zinc-finger domains are nucleic acid-binding protein structures mostly involved in transcription regulation. Many classes of zinc-fingers are characterized according to the number and positions of the histidine and cysteine residues involved in the zinc atom coordination (Rosenfeld and Margalit, 1993). The zinc-finger C2H2-type domain (PDOC00028), with DNA or RNA binding property, and the zinc finger RING-type signature (PDOC00449), which can play a key role in the ubiquitination pathway (Ito et al., 2001), were the most commonly found in the *T. trichiura* proteins (Fig. S3A).

The protein kinase domain (PDOC00100) was identified in 34 *T. trichiura* uncharacterized proteins; proteins with the highest probabilities of being true protein kinases will commonly present two specific signatures: (i) serine/threonine residues, involved in ATP binding, and (ii) tyrosine protein kinases, with catalytic activity (Hanks and Hunter, 1995; Knighton et al., 1991). Six of the predicted *T. trichiura* proteins, out of 34 proteins identified with protein kinase domains, displayed both proteins signatures (Fig. S3B).

3.5. Expression levels of *T. trichiura* transcripts and cross-species comparisons

The abundance of the *T. trichiura* transcripts was estimated by normalized RPKM values, following mapping of the transcriptome reads against the recently released draft genome assembly for the species (Foth et al., 2014). A stringent RPKM threshold of 0.3 was defined to confidentially detect the presence of a transcript for a particular gene.

There is a high number of transcripts coding for unknown-function proteins (hypothetical proteins) among the top 40 most abundant protein-encoding transcripts identified in the transcriptome of the *T. trichiura* adult worm (Table 2), similarly to what has been demonstrated for the transcriptionally most highly expressed genes of *T. muris* (Foth et al., 2014). Several of these highly expressed transcripts that code for hypothetical proteins are predicted to be part of a gene family that contains 8 paralogs annotated in *T. trichiura* (Fig. 5); there are 6 orthologs identified in the related parasites *T. suis* and *T. muris*, according to the WormBase Parasite database (Fig. 5). Fig. S4 shows an analysis of unique reads identified for each of the paralogs in the *T. trichiura* transcriptome. Noteworthy, the gene orthologs found in *T. suis* and *T. muris* were also seen amongst the most highly expressed protein-encoding genes according to previous transcriptomic studies of these species (Jex et al., 2014; Foth et al., 2014).

Considering that adult female whipworms can lay thousands of eggs per day, it is not unexpected to find a high transcriptional activity of genes that code for proteins involved in reproductive processes. Accordingly, among the most abundant transcripts identified in the *T. trichiura* adult worm transcriptome we found sequences that code for proteins involved in lipid transport, such as vitellogenins (which are major components of yolk), and several chitin-binding proteins (Table 2). Interestingly, some of these protein-encoding transcripts were also seen to be more expressed in female adult worms (particularly in the posterior body) in the *T.muris* transcriptome (Foth et al., 2014). Besides, transcripts coding for these proteins were also identified amongst the 20 most highly expressed protein-encoding transcripts in separate *T. suis* transcriptomic studies (Cantacessi et al., 2011; Jex et al., 2014).

Previous studies have demonstrated that it might be possible to achieve successful measurements of mRNA expression levels through mapping of RNA-seq reads to the predicted coding sequences from a reference genome of a closely related species (Hornett and Wheat, 2012). The human and porcine whipworms are very closely related; in fact, these species have only recently been demonstrated to be separate species and unarguably distinguished at the genetics level (Cutillas et al., 2009; Liu et al., 2012). The relatedness between *T. trichiura* and the mouse parasite *T. muris* has also been recently demonstrated in the work by Foth et al. (2014), where they show that these two species share a very high number of gene orthologs, with an average amino acid identity of 79.0% between protein orthologs (Foth et al., 2014).

Therefore, we have hypothesized in this study that it might be possible to use the recently released *T. trichiura* annotated genome as a close reference genome for mapping *T. suis* and *T. muris* RNA-seq short reads, in order to discover gene orthologs that are highly expressed across all the three species (for details; please refer to item 2.7). Fig. 6A shows the numbers of gene orthologs found as highly expressed (calculated RPKM > 200) for each species. It was possible to identify 26 protein-encoding genes that are consistently highly expressed in the adult stages of the three helminth parasite species (Fig. 6A and B); these genes are highly conserved among the *Trichuris* species, and their high expression levels have been confirmed from previous transcriptomic studies of *T. suis* and *T. muris* (Table S4). Several of these 26 genes had been identified in this study among the 40 most highly expressed protein-encoding genes of *T. trichiura* (Fig. 6B and Table 2). Noteworthy, genes belonging to the gene family shown in Fig. 5 were seen as highly expressed in all three species (Fig. 6B). Besides, we have noticed that transcripts coding for proteins involved in reproductive functions in female parasites are also particularly highly represented in the transcriptomes of adult worms (Fig. 6B).

3.6. *T. trichiura* transcripts encoding proteins with immunomodulatory activities

In the present analysis, it was possible to identify 20 transcripts that code for proteins previously detected in the *T. trichiura* immunomodulatory protein fractions (TtEFs) described in the study by Santos et al. (2013) (please refer to item 2.6; Table 3). Five of these transcripts were amongst the 200 most highly expressed, as measured by normalized transcript abundance values (RPKM) (Table 3).

TtEF9 was the protein fraction presenting the greatest immunoregulatory response inducing a high level of IL-10 when cultured with PBMCs and showing strong inhibitory effects against the production of T_H1 and T_H2 cytokines when co-cultured with optimal immune stimuli (Santos et al., 2013). The proteins

Table 2
The 40 most highly transcribed protein-encoding sequences identified in the *T. trichiura* adult worm transcriptome.

Accession number[a]	Description[b]	Orthologs[c]	RPKM[d] (Log$_2$)	Gene Ontology (GO) Annotations[e]		
				Biological Process	Molecular Function	Cellular Component
CDW58692.1	Hypothetical protein TTRE_0000701701	TMUE_s0052001100	14.31			
CDW52980.1	Hypothetical protein TTRE_0000124301	Tsui7105291	14.23			
CDW52979.1	Hypothetical protein TTRE_0000124101	M513_04084	14.06			
CDW58325.1	Hypothetical protein TTRE_0000663201	M514_19266	14.03			
CDW56655.1	Hypothetical protein TTRE_0000493701	TMUE_s0052001100	13.83			Membrane
CDW54233.1	Hypothetical protein TTRE_0000250301	TMUE_s0024002900	13.48			
CDW58129.1	Vitellogenin N and VWD and DUF1943 domain containing protein	Tsui7113118	13.32	Lipid transport	Lipid transporter activity	
CDW54610.1	Spindle-pole body protein (Pcp1)	M513_09656	13.11			
CDW61263.1	Hypothetical protein TTRE_0000971101	TMUE_s0052001000	13.06			
CDW61299.1	Hypothetical protein TTRE_0000974801	Tsui7323554	12.92			
CDW61002.1	Cell wall-associated hydrolase	Asuu8446313	12.66	Metabolic process	Hydrolase activity	
CDW61247.1	Cell wall-associated hydrolase	Asuu8446313	12.55	Metabolic process	Hydrolase activity	
CDW61305.1	Hypothetical protein TTRE_0000975601	Tsui7125196	12.47			
CDW60883.1	Hypothetical protein TTRE_0000928701	Tsui7105291	12.00			
CDW61059.1	Hypothetical protein TTRE_0000948201	Tsui7121137	11.62		Lipid binding	
CDW60789.1	Cell wall-associated hydrolase	Asuu8446313	10.70	Metabolic process	Hydrolase activity	
CDW54015.1	Poly-cysteine and histidine tailed protein isoform 2	Tsui7117112	10.46			
CDW54307.1	Heat shock protein 90	Tsui7117373	10.20	Protein folding; Response to stress	ATP binding; Unfolded protein binding	
CDW59881.1	Hypothetical protein TTRE_0000822501	Tsui7115052	10.08			
CDW54937.1	Polyadenylate binding protein 1	Tsui7108405	9.87		Nucleotide binding; RNA binding	Cytoplasm
CDW61304.1	hypothetical protein TTRE_0000975501	Asuu8446313	9.80			
CDW56365.1	Polyubiquitin	Tsui7109600	9.74			Cytoplasm
CDW58714.1	TIL and CBM 14 domain containing protein	Tsui7119446	9.49	Chitin metabolic process	Chitin binding	Extracellular region
CDW57445.1	CBM 14 domain containing protein	Tsui7137747	9.22	Chitin metabolic process	Chitin binding	Extracellular region
CDW61148.1	Hypothetical protein TTRE_0000958101	Tsui7242031	9.20	Protein folding	Calcium ion binding; Unfolded protein binding	Endoplasmic reticulum
CDW53632.1	Hypothetical protein TTRE_0000189701	Tsui7129116	9.03			
CDW52289.1	Hypothetical protein TTRE_0000054801	Tsui7134747	8.98			
CDW53617.1	Angiotensin converting enzyme	Tsui7118907	8.97	Proteolysis	Peptidase activity; Metallopeptidase activity	Membrane
CDW56602.1	Tyrosinase domain containing protein	Tsui7146002	8.94	Oxidation-reduction process	Oxidoreductase activity	
CDW54940.1	CBM 14 domain containing protein	Tsui7110776	8.88	Chitin metabolic process	Chitin binding	Extracellular region
CDW59986.1	NADH dehydrogenase subunit 5	Tsui7107151	8.79	Oxidation-reduction process	Oxidoreductase activity	Membrane
CDW53952.1	Fatty acid synthase	Tsui7130170	8.79	Biosynthetic process	Fatty acid synthase activity	
CDW55082.1	TSP 1 and CBM 14 domain containing protein	Tsui7387368	8.67	Chitin metabolic process	Chitin binding	Extracellular region
CDW57937.1	Calreticulin	Tsui7272154	8.65	Protein folding	Unfolded protein binding	Endoplasmic reticulum
CDW57439.1	CBM 14 domain containing protein	Tsui7118549	8.65	Chitin metabolic process	Chitin binding	Extracellular region
CDW56497.1	Phosphoenolpyruvate carboxykinase GTP	Tsui7136493	8.59	Gluconeogenesis	Kinase activity	
CDW61309.1	Piwi domain protein	Tsui7129624	8.57		Nucleic acid binding	
CDW60206.1	Trypsin domain containing protein	Tsui7201621	8.55	Proteolysis	Serine-type endopeptidase activity	
CDW56764.1	Bifunctional 3′ phosphoadenosine	Tsui7201551	8.52	Phosphorylation	Adenylylsulfate kinase activity	
CDW59236.1	Hypothetical protein TTRE_0000756701	Tsui7167102	8.44			

Tsui = *Trichuris suis*; Asuu = *Ascaris suum*; TMUE = *Trichuris muris.*

[a] NCBI's PROTEIN Database. Several of the most highly expressed sequences are part of a gene family with 8 paralogs and 6 orthologs annotated in WormBase Parasite database (Howe et al., 2016). These are marked bold.

[b] According to version 2.2 of the *T. trichiura* genome annotation.

[c] Orthologs retrieved from the HelmDB database (Mangiola et al., 2013) or the WormBase Parasite database (Howe et al., 2016), using BLAST.

[d] Normalized transcript abundances obtained using CLC Genomics Workbench (CLC bio). Results are expressed in Log2 of Reads Per Kilobase per Million mapped reads (RPKM). All these measurements were also corroborated by analysis with NextGENe® software v2.4 (SoftGenetics).

[e] Gene Ontology (GO) annotations were retrieved from UniProtKB (Huntley et al., 2015); annotations were last updated in December 2015.

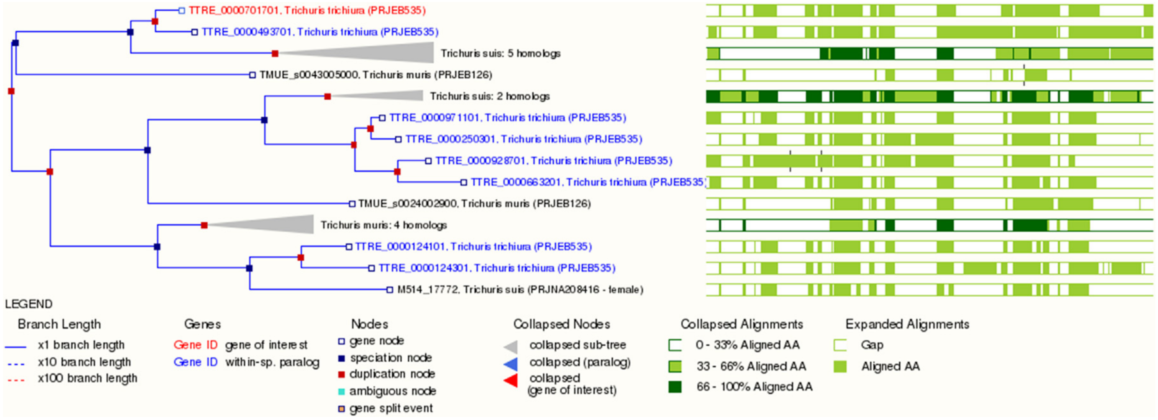
TTRE_0000701701, Trichuris trichiura (PRJEB535)
TTRE_0000493701, Trichuris trichiura (PRJEB535)
Trichuris suis: 5 homologs
TMUE_s0043005000, Trichuris muris (PRJEB126)
Trichuris suis: 2 homologs
TTRE_0000971101, Trichuris trichiura (PRJEB535)
TTRE_0000250301, Trichuris trichiura (PRJEB535)
TTRE_0000928701, Trichuris trichiura (PRJEB535)
TTRE_0000663201, Trichuris trichiura (PRJEB535)
TMUE_s0024002900, Trichuris muris (PRJEB126)
Trichuris muris: 4 homologs
TTRE_0000124101, Trichuris trichiura (PRJEB535)
TTRE_0000124301, Trichuris trichiura (PRJEB535)
M514_17772, Trichuris suis (PRJNA208416 - female)
LEGEND
Branch Length
x1 branch length
x10 branch length
x100 branch length
Genes
Gene ID gene of interest
Gene ID within-sp. paralog
Nodes
gene node
speciation node
duplication node
ambiguous node
gene split event
Collapsed Nodes
collapsed sub-tree
collapsed (paralog)
collapsed (gene of interest)
Collapsed Alignments
0 - 33% Aligned AA
33 - 66% Aligned AA
66 - 100% Aligned AA
Expanded Alignments
Gap
Aligned AA

Fig. 5. A gene family coding for unknown-function proteins with high expression levels in the *T. trichiura* transcriptome. Gene tree showing eight *T. trichiura* paralogs and six orthologs found in related species, according to gene models present in WormBase Parasite.

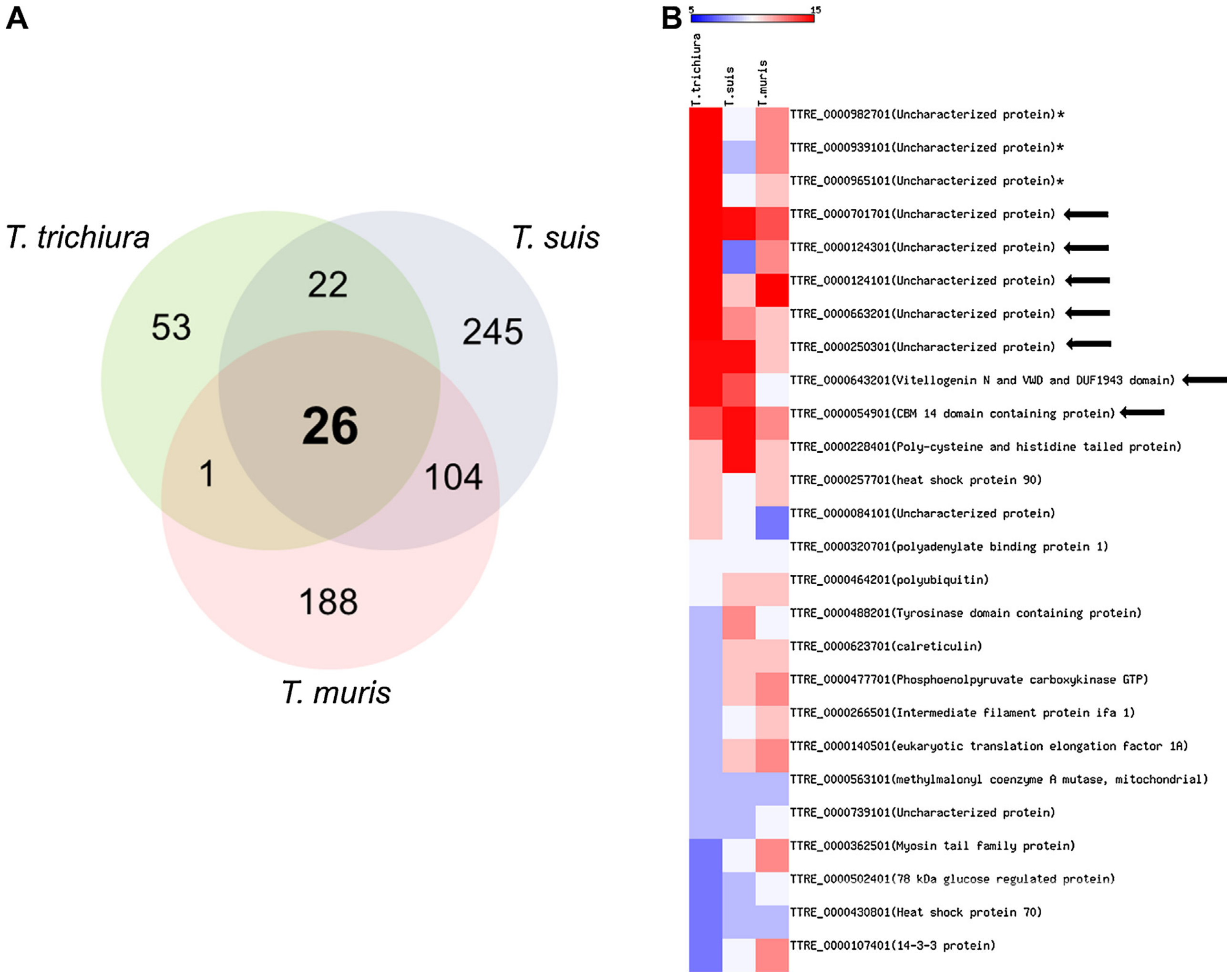

Fig. 6. Cross-species comparisons of transcript abundances. RNA-seq short reads of the adult stages of the three *Trichuris* species were mapped against the *T. trichiura* annotated genes (for details, please refer to the methods section). (A) Venn diagram showing the distribution of the most highly expressed (RPKM > 200) transcripts across the three species. (B) Expression levels of 26 gene orthologs consistently highly expressed in the three *Trichuris* species. Heat map was generated with Log2 transformed RPKM values, using the Matrix2png tool (Pavlidis and Noble, 2003). An asterisk (*) indicates sequences that code for ribosomal RNAs, as shown in Fig. S6. The arrows indicate protein encoding genes whose orthologs have been shown to be highly expressed in female worms in previous transcriptomic studies of the species *T. muris* and *T. suis* (Table S4), including genes in Fig. 5.

Table 3
Expression analysis of transcripts encoding *T. trichiura* proteins with immunoregulatory activities.

Protein ID[a]	Description	Transcript abundance (RPKM)[b]	TtEF6	TtEF8	TtEF9	TtEF10	TtEF11	TtEF12
CDW56935.1	Actin	**179.54**			**X**			
CDW53601.1	Actin depolymerizing factor 2, isoform c	7.00	**X**					
CDW53275.1	ATP synthase subunit beta	89.03			**X**	**X**	**X**	
CDW53384.1	Beta hexosaminidase	22.47		**X**				
CDW57885.1	Gut specific cysteine proteinase	6.85	**X**	**X**	**X**			
CDW60307.1	NADH dependent fumarate reductase	12.15	**X**					
CDW53142.1	Eukaryotic translation elongation factor 1A	**249.53**	**X**					
CDW54435.1	Fructose bisphosphate aldolase class I	76.75	**X**	**X**	**X**		**X**	**X**
CDW53185.1	Glutamate dehydrogenase	10.49	**X**					
CDW56034.1	Heat shock protein 70	**196.20**		**X**		**X**		
CDW53003.1	Malate dehydrogenase	64.28			**X**			
CDW54347.1	Protein retinal degeneration B	5.46	**X**					
CDW58553.1	Protein kinase C, brain isozyme	3.68			**X**			
CDW56947.1	Laminin G 2 and Cadherin and EGF CA domain containing protein	10.01		**X**				
CDW57784.1	Ufd2 P core and U-box domain containing protein	17.06	**X**					
CDW56963.1	Nebulin and LIM and SH3 1 domain containing protein	4.87	**X**					
CDW52208.1	Girdin	36.50			**X**			
CDW54989.1	tRNA (cytosine(34) C(5)) methyltransferase	11.39		**X**				
CDW56497.1	Phosphoenolpyruvate carboxykinase GTP	**363.89**	**X**	**X**				
CDW54395.1	Intermediate filament protein ifa 1	**281.45**				**X**		

[a] Proteins identified previously through mass spectrometry analysis of six chromatographic fractions of the whipworm somatic extract (TtEF 6, TtEF 8, TtEF 9, TtEF 10, TtEF11 and TtEF 12), which showed *in vitro* immunomodulatory activities (Santos et al., 2013). NCBI's PROTEIN Database.

[b] Normalized transcript abundances in Reads Per Kilobase per Million mapped reads (RPKM), as determined by analysis with the NextGENe® software v2.4 (SoftGenetics). Transcripts that are amongst the most highly expressed in the *T. trichiura* transcriptome are marked bold.

predicted in this study from the translated *T. trichiura* transcriptome which are found in TtEF9 include: CDW56935.1 (Actin Fragment); CDW53275.1 (ATP synthase subunit β); CDW57885.1 (Gut specific cysteine proteinase); CDW54435.1 (Fructose bisphosphate aldolase); CDW53003.1 (Malate dehydrogenase); CDW58553.1 (Protein kinase C, brain isozyme); and CDW52208.1 (Girdin).

The protein 'Fructose bisphosphate aldolase' (CDW54435.1) was identified in all but one (TtEF 10) of the immunomodulatory fractions previously described (Santos et al., 2013) (Table 3). Likewise, 'Gut specific cysteine proteinase' (CDW57885.1) and 'Heat shock protein 70 (HSP70)' (CDW56034.1) were also present in more than one of the TtEF immunomodulatory fractions (Table 3). Therefore, these sequences represent suitable candidates for future expression using recombinant DNA technology. Three-dimensional modeling of these *T. trichiura* proteins demonstrated high structural similarity with their characterized orthologs from *T. spiralis* (Fig. S5).

In addition to these proteins, we investigated sequences in the *T. trichiura* transcriptome that encode proteins demonstrated to have potent immunomodulatory properties in related parasitic helminths, including a serine-type endopeptidase from *T. suis* first stage larvae (Tsui7304731) (Ebner et al., 2014) and the Cathepsin L1 from *Fasciola hepatica* (Dowling et al., 2010). The protein Tsui7304731, along with the proteins of unknown function, Tsui7583957 and Tsui7234544, have been previously identified by mass spectrometry in the excretory/secretory extract from L1 *T. suis* larvae, and have demonstrated immunomodulatory effects in a murine model of allergy (Ebner et al., 2014). tBLASTn searches against the *T. trichiura* transcriptome using the amino acid sequence of Tsui7304731 identified the coding transcript Trichuris_c27 (E-value: $2.8E^{-136}$). Structural alignment of the two predicted proteins also demonstrated high similarity (Fig. S5). Transcripts encoding proteins with similarities to Tsui7583957 and Tsui7234544 were also found in the *T. trichiura* transcriptome: Trichuris_c4299 (E-value: $3E^{-092}$) and Trichuris_c2345 (E-value: $2E^{-037}$), respectively. The Cathepsin L1 of *F. hepatica* is another helminthic protein with immunoregulatory activity (Dowling et al., 2010). A tBLASTn search against the *T. trichiura* transcriptome identified the coding transcript Trichuris_c3964, which codes for a probable protein with high similarity to the *F. hepatica* counterpart (Fig. S5).

4. Conclusions

In conclusion, we present the first transcriptomic exploration of the adult stage of the human whipworm *T. trichiura*, using next-generation sequencing technology and a *de novo* assembly strategy. We were able to identify a number of transcripts that code for potential newly discovered proteins for the species *T. trichiura* and also for previously unannotated non-coding transcripts. Besides, we identified transcripts that code for proteins previously reported to possess immunomodulatory activities. Our findings will now allow us to produce recombinant *T. trichiura* proteins encoded by these transcripts that could be evaluated as potential therapeutic molecules for the treatment of chronic inflammatory conditions in humans.

Acknowledgements

This study was supported by the following research grants: INCT/MCT/CNPq Programme – Contract no. 5737862008; MCTI/CNPq/FNDCT Ação Transversal N° 79/2013 (RENORBIO); FAPESB (PET0005/2010 and JCB017/2013); and CAPES (project n°077_2012). LNS was recipient of a scholarship from FAPESB. NMAN and MLB were recipients of research fellowships from CNPq.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.actatropica.2016.03.036.

References

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat. Genet. 25, 25–29.

Bashi, T., Bizzaro, G., Ben-Ami Shor, D., Blank, M., Shoenfeld, Y., 2015. The mechanisms behind helminth's immunomodulation in autoimmunity. Autoimmun. Rev. 14, 98–104.

Bethony, J., Brooker, S., Albonico, M., Geiger, S., Loukas, A., Diemert, D., Hotez, P., 2006. Soil-transmitted helminth infections: ascariasis, trichuriasis, and hookworm. Lancet 367, 1521–1532.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., Kiefer, F., Cassarino, T.G., Bertoni, M., Bordoli, L., Schwede, T., 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 42, W252–258.

Cantacessi, C., Young, N.D., Nejsum, P., Jex, A.R., Campbell, B.E., Hall, R.S., Thamsborg, S.M., Scheerlinck, J.P., Gasser, R.B., 2011. The transcriptome of *Trichuris suis*–first molecular insights into a parasite with curative properties for key immune diseases of humans. PLoS One 6, e23590.

Castro, T., Seyffert, N., Ramos, R., Barbosa, S., Carvalho, R., Pinto, A., Carneiro, A., Silva, W., Pacheco, L., Downson, C., Schneider, M., Miyoshi, A., Azevedo, V., Silva, A., 2013. Ion Torrent-based transcriptional assessment of a *Corynebacterium pseudotuberculosis* equi strain reveals denaturing high-performance liquid chromatography a promising rRNA depletion method. Microbiol. Biotechnol. 6, 168–177.

Chen, M., Hu, Y., Liu, J., Wu, Q., Zhang, C., Yu, J., Xiao, J., Wei, F., Wu, J., 2015. Improvement of genome assembly completeness and identification of novel full-length protein-coding genes by RNA-seq in the giant panda genome. Sci. Rep. 5, 18019.

Chevreux, B., Pfisterer, T., Drescher, B., Driesel, A.J., Müller, W.E., Wetter, T., Suhai, S., 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res. 14, 1147–1159.

Cutillas, C., Callejon, R., de Rojas, M., Tewes, B., Ubeda, J.M., Ariza, C., Guevara, D.C., 2009. *Trichuris suis* and *Trichuris trichiura* are different nematode species. Acta Trop. 111, 299–307.

Dowling, D.J., Hamilton, C.M., Donnelly, S., La Course, J., Brophy, P.M., Dalton, J., O'Neill, S.M., 2010. Major secretory antigens of the helminth *Fasciola hepatica* activate a suppressive dendritic cell phenotype that attenuates Th17 cells but fails to activate Th2 immune responses. Infect. Immun. 78, 793–801.

Ebner, F., Hepworth, M., Rausch, S., Janek, K., Niewienda, A., Kühl, A., Henklein, P., Lucius, R., Hamelmann, E., Hartmann, S., 2014. Therapeutic potential of larval excretory/secretory proteins of the pig whipworm *Trichuris suis* in allergic disease. Allergy 69, 1489–1497.

Fleming, J.O., Isaak, A., Lee, J.E., Luzzio, C.C., Carrithers, M.D., Cook, T.D., Field, A.S., Boland, J., Fabry, Z., 2011. Probiotic helminth administration in relapsing-remitting multiple sclerosis: a phase 1 study. Mult. Scler. 17, 743–754.

Foth, B.J., Tsai, I.J., Reid, A.J., Bancroft, A.J., Nichol, S., Tracey, A., Holroyd, N., Cotton, J.A., Stanley, E.J., Zarowiecki, M., Liu, J.Z., Huckvale, T., Cooper, P.J., Grencis, R.K., Berriman, M., 2014. Whipworm genome and dual-species transcriptome analyses provide molecular insights into an intimate host-parasite interaction. Nat. Genet. 46, 693–700.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152.

Hanks, S.K., Hunter, T., 1995. Protein kinases 6: the eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. FASEB J. 9, 576–596, Review.

Higa, R.H., Togawa, R.C., Montagner, A.J., Palandrani, J.C., Okimoto, I.K., Kuser, P.R., Yamagishi, M.E., Mancini, A.L., Neshich, G., 2004. STING Millennium Suite: integrated software for extensive analyses of 3d structures of proteins and their complexes. BMC Bioinformatics 5, 107.

Hornett, E.A., Wheat, C.W., 2012. Quantitative RNA-seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. BMC Genomics 13, 361.

Howe, K.L., Bolt, B.J., Cain, S., Chan, J., Chen, W.J., Davis, P., Done, J., Down, T., Gao, S., Grove, C., Harris, T.W., Kishore, R., Lee, R., Lomax, J., Li, Y., Muller, H.M., Nakamura, C., Nuin, P., Paulini, M., Raciti, D., Schindelman, G., Stanley, E., Tuli, M.A., Van Auken, K., Wang, D., Wang, X., Williams, G., Wright, A., Yook, K., Berriman, M., Kersey, P., Schedl, T., Stein, L., Sternberg, P.W., 2016. WormBase 2016: expanding to enable helminth genomic research. Nucleic Acids Res. 44, D774–780.

Huntley, R.P., Sawford, T., Mutowo-Meullenet, P., Shypitsyna, A., Bonilla, C., Martin, M.J., O'Donovan, C., 2015. The GOA database: gene ontology annotation updates for 2015. Nucleic Acids Res. 43, D1057–1063.

Ito, K., Adachi, S., Iwakami, R., Yasuda, H., Muto, Y., Seki, N., Okano, Y., 2001. N-Terminally extended human ubiquitin-conjugating enzymes (E2s) mediate the ubiquitination of RING-finger proteins, ARA54 and RNF8. Eur. J. Biochem. 268, 2725–2732.

Jex, A.R., Nejsum, P., Schwarz, E.M., Hu, L., Young, N.D., Hall, R.S., Korhonen, P.K., Liao, S., Thamsborg, S., Xia, J., Xu, P., Wang, S., Scheerlinck, J.P., Hofmann, A., Sternberg, P.W., Wang, J., Gasser, R.B., 2014. Genome and transcriptome of the porcine whipworm *Trichuris suis*. Nat. Genet. 46, 701–706.

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2014. Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. 42, D199–205.Knighton, D.R., Zheng, J.H., Ten Eyck, L.F., Ashford, V.A., Xuong, N.H., Taylor, S.S., Sowadski, J.M., 1991. Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. Science 26, 407–414.

Konagurthu, A., Whisstock, J., Stuckey, P., Lesk, A., 2006. MUSTANG: a multiple structural alignment algorithm. Proteins 64 (3), 559–574.

Liu, G.H., Gasser, R.B., Su, A., Nejsum, P., Peng, L., Lin, R.Q., Li, M.W., Xu, M.J., Zhu, X.Q., 2012. Clear genetic distinctiveness between human- and pig-derived Trichuris based on analyses of mitochondrial datasets. PLoS Negl. Trop. Dis. 6, e1539.

Maizels, R.M., McSorley, H.J., Smyth, D.J., 2014. Helminths in the hygiene hypothesis: sooner or later? Clin. Exp. Immunol. 177, 38–46.

Mangiola, S., Young, N., Korhonen, P., Mondal, A., Scheerlinck, J., Sternberg, P., Cantacessi, C., Hall, R., Jex, A., Gasser, R., 2013. Getting the most out of parasitic helminth transcriptomes using HelmDB: implications for biology and biotechnology. Biotechnol. Adv. 31, 1109–1119.

Meekums, H., Hawash, M.B., Sparks, A.M., Oviedo, Y., Sandoval, C., Chico, M.E., Stothard, J.R., Cooper, P.J., Nejsum, P., Betson, M., 2015. A genetic analysis of *Trichuris trichiura* and *Trichuris suis* from Ecuador. Parasites Vectors 8, 168.

Mishra, P.K., Palma, M., Bleich, D., Loke, P., Gause, W.C., 2014. Systemic impact of intestinal helminth infections. Mucosal Immunol. 7, 753–762.

Pavlidis, P., Noble, W.S., 2003. Matrix2png: a utility for visualizing matrix data. Bioinformatics 19, 295–296.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., Ferrin, T.E., 2004. UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. 25, 1605–1612.

Pullan, R.L., Brooker, S.J., 2012. The global limits and population at risk of soil-transmitted helminth infections in 2010. Parasit Vectors 5, 81.

Rodrigues, L., Newcombe, P., Cunha, S., Alcantara-Neves, N., Genser, B., Cruz, A., Simoes, S., Fiaccone, R., Amorim, L., Cooper, P., Barreto, M., Social Change, A.a.A.i.L.A., 2008. Early infection with *Trichuris trichiura* and allergen skin test reactivity in later childhood. Clin. Exp. Allergy 38, 1769–1777.

Rosenfeld, R., Margalit, H., 1993. Zinc fingers: conserved properties that can distinguish between spurious and actual DNA-binding motifs. J. Biomol. Struct. Dyn. 11, 557–570.

Santos, L.N., Gallo, M.B., Silva, E.S., Figueiredo, C.A., Cooper, P.J., Barreto, M.L., Loureiro, S., Pontes-de-Carvalho, L.C., Alcantara-Neves, N.M., 2013. A proteomic approach to identify proteins from *Trichuris trichiura* extract with immunomodulatory effects. Parasite Immunol. 35, 188–193.

Stanke, M., Schoffmann, O., Morgenstern, B., Waack, S., 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7, 62.

Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., Balzereit, D., Lehrach, H., Yaspo, M., 2014. Influence of RNA extraction methods and library selection schemes on RNA-seq data. BMC Genomics 11 (15), 675.

Summers, R., Elliott, D., Qadir, K., Urban, J.J., Thompson, R., Weinstock, J., 2003. *Trichuris suis* seems to be safe and possibly effective in the treatment of inflammatory bowel disease. Am. J. Gastroenterol. 98, 2034–2041.

Summers, R., Elliott, D., Urban, J.J., Thompson, R., Weinstock, J., 2005a. *Trichuris suis* therapy for active ulcerative colitis: a randomized controlled trial. Gastroenterology 128, 825–832.

Summers, R., Elliott, D., Urban, J.J., Thompson, R., Weinstock, J., 2005b. *Trichuris suis* therapy in Crohn's disease. Gut 54, 87–90.

Tyagi, R., Rosa, B.A., Lewis, W.G., Mitreva, M., 2015. Pan-phylum comparison of nematode metabolic potential. PLoS Negl. Trop. Dis. 9, e0003788.

Van Bel, M., Proost, S., Van Neste, C., Deforce, D., Van de Peer, Y., Vandepoele, K., 2013. TRAPID: an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. Genome Biol. 14, R134.

WHO, 2014. Soil-transmitted helminth infections, Fact sheet N(366 ed. World Health Organization).

Warren, A., Aurrecoechea, C., Brunk, B., Desai, P., Emrich, S., Giraldo-Calderón, G., Harb, O., Hix, D., Lawson, D., Machi, D., Mao, C., McClelland, M., Nordberg, E., Shukla, M., Vosshall, L., Wattam, A., Will, R., Yoo, H., Sobral, B., 2015. RNA-Rocket: an RNA-Seq analysis resource for infectious disease research. Bioinformatics 31 (1), 1496–1498, http://dx.doi.org/10.1093/bioinformatics/btv002, Epub 2015 Jan 7.

Wiederstein, M., Sippl, M.J., 2007. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 35, W407–410.

de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Petra, S., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., Hulo, N., 2006. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 34, 362–365.

Zhao, Y., Li, H., Fang, S., Kang, Y., Wu, W., Hao, Y., Li, Z., Bu, D., Sun, N., Zhang, M.Q., Chen, R., 2016. NONCODE 2016: an informative and valuable data source of long non-coding RNAs. Nucleic Acids Res. 44, D203–208.