DNA Research, 2018, 25(5), 499–510
doi: 10.1093/dnares/dsy020
Advance Access Publication Date: 25 June 2018
Full Paper

Full Paper

The genome of tapeworm *Taenia multiceps* sheds light on understanding parasitic mechanism and control of coenurosis disease

Wenhui Li[1,†], Bo Liu[2,†], Yang Yang[1,†], Yuwei Ren[2,†], Shuai Wang[1,†], Conghui Liu[2,†], Nianzhang Zhang[1], Zigang Qu[1], Wanxu Yang[2], Yan Zhang[2], Hongbing Yan[1], Fan Jiang[2], Li Li[1], Shuqu Li[2], Wanzhong Jia[1,3], Hong Yin[1,3], Xuepeng Cai[1], Tao Liu[4], Donald P. McManus[5,]*, Wei Fan[2,]*, and Baoquan Fu[1,3,]*

[1]State Key Laboratory of Veterinary Etiological Biology, Key Laboratory of Veterinary Parasitology of Gansu Province, Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences, Lanzhou 730046, China, [2]Agricultural Genomic Institute, Chinese Academy of Agricultural Sciences, Shenzhen 518124, China, [3]Jiangsu Co-innovation Center for Prevention and Control of Important Animal Infectious Disease, Yangzhou, Jiangsu, China, [4]Annoroad Gene Tech. (Beijing) Co Ltd, Beijing, China, and [5]Molecular Parasitology Laboratory, QIMR Berghofer Medical Research Institute, Brisbane, Australia

*To whom correspondence should be addressed. Tel. +86 0931 8342 675. Email: don.mcmanus@qimrberghofer.edu.au (D.P.M.); fanweiagis@126.com (W.F.); fubaoquan@caas.cn (B.F.)

[†]These authors contributed equally to this work.

Edited by Prof. Takashi Ito

Received 3 November 2017; Editorial decision 25 May 2018; Accepted 30 May 2018

Abstract

Coenurosis, caused by the larval coenurus of the tapeworm *Taenia multiceps*, is a fatal central nervous system disease in both sheep and humans. Though treatment and prevention options are available, the control of coenurosis still faces presents great challenges. Here, we present a high-quality genome sequence of *T. multiceps* in which 240 Mb (96%) of the genome has been successfully assembled using Pacbio single-molecule real-time (SMRT) and Hi-C data with a N50 length of 44.8 Mb. In total, 49.5 Mb (20.6%) repeat sequences and 13, 013 gene models were identified. We found that *Taenia* spp. have an expansion of transposable elements and recent small-scale gene duplications following the divergence of *Taenia* from *Echinococcus*, but not in *Echinococcus* genomes, and the genes underlying environmental adaptability and dosage effect tend to be over-retained in the *T. multiceps* genome. Moreover, we identified several genes encoding proteins involved in proglottid formation and interactions with the host central nervous system, which may contribute to the adaption of *T. multiceps* to its parasitic life style. Our study not only provides insights into the biology and evolution of *T. multiceps*, but also identifies a set of species-specific gene targets for developing novel treatment and control tools for coenurosis.

Key words: *Taenia multiceps*, coenurosis, genome, parasitism

1. Introduction

Coenurosis is an often fatal central nervous system (CNS) parasitic disease mainly in sheep and other ungulates, this frequently leads to their death and huge socioeconomic losses, especially in developing countries. This parasite disease is caused by the larval stage (*Coenurus cerebralis*) of the tapeworm *Taenia multiceps*[1] (Phylum Platyhelminthes). The parasite can also cause zoonotic infections in humans, but has been largely neglected due to a lack of diagnostic techniques and studies.[2] Two hosts are required to complete its life cycle. The adult tapeworm inhabits the small intestine of the definitive host (dog, wolf and fox), while the larval coenurus stage parasitizes the brain or spinal cord of the intermediate host (sheep, goat and cattle),[3] causing a disease called ‘staggers’ or ‘gid’ resulting in pronounced intracranial pressure, leading to headache, ataxia, hypermetria, head deviation, blindness, stumbling, paralysis and even death.[4–6] Coenurosis has spread across most regions of the world, especially in developing countries in Africa and southeast Asia that are involved in the husbandry of sheep or goats, resulting in serious economic losses.[7] Notably, coenurosis is also a zoonosis which can lead to serious pathological conditions in humans, and appears to be more common than previously assumed.[8]

T. multiceps has evolved several strategies to adapt to a parasitic life style including the loss of respiratory organs and a digestive tract so that andnutrients are absorbed through its unique external tegument.[9] It is monoecious, producing diploid eggs by self-fertilization which give rise to infective oncospheres.[10] Further, the germinative neck region of the adult worm of *T. multiceps* can produce hundreds of proglottids (segments), each of which has a set of complete reproductive organs and contains tens of thousands of oncospheres (eggs) after maturation.[9] In addition, one oncosphere can develop into a cysticercus which possesses unusual powers of asexual multiplication, forming a bladder or coenurus which can give rise to hundreds of daughter protoscoleces, each of which has the potential to grow into an adult worm, if ingested by a definitive host. The characteristics of nutrient absorption and reproduction have contributed to the wide-spread, global distribution of *T. multiceps*.

Though treatment and preventative measures are available, the control of coenurosis is still challenging. Burning or burying infected intermediate host offal is the most simple and effective way to block the transmission of *T. multiceps* to canine definitive hosts, but these procedures result in considerable economic loss. Surgery after general anaesthesia of an infected animal has been used for the removal of the coenurus cyst,[11,12] but the procedure is limited under field conditions due to economic constraints. Antiparasitic drugs, such as a combination of fenbendazole, praziquantel and albendazole, are effective against the migrating larvae and represent a more practicable method for expelling *T. multiceps*, but they are often associated with side effects in the host. Development of an effective vaccine can complement interventions for the prevention of coenurosis, and in recent years, the recombinant proteins, Tm16 and Tm18, have been developed as vaccine candidates, having been shown to provide some protection in sheep against experimental infection with *T. multiceps* eggs.[13,14] However, the vaccine efficacy is comparatively modest compared with other recombinant anti-cestode vaccines that have been developed.[15] Therefore, the development of more effective and safe interventions to control coenurosis is still an urgent need.

The recent sequencing of tapeworm genomes has facilitated a better understanding of their genetic makeup, providing new insight on the functional biology and mechanisms of pathogenesis in cestodes as well as providing new targets for control tools leading to improved prevention and treatment of cestodiases. To date, several tapeworm genomes, including two *Echinococcus* species (*E. multilocularis* and *E. granulosus*) and three *Taenia* species (*T. solium*, *T. saginata* and *T. asiatica*) have been sequenced, with assembled genome sizes ranging from 114 to 169 Mb.[16–18] However, there is limited genetic information reported for *T. multiceps*, apart from some limited transcriptome and microRNA data.[2,19] Here, we present a high-quality genome assembly for *T. multiceps* together with transcriptomic information for several of its developmental stages. Our study not only provides insights into the biology and evolution of *T. multiceps*, but also identifies a set of potentially novel drug and vaccine targets, which we anticipate will contribute to the development of much needed new treatment and control measures for coenurosis.

2. Materials and methods

2.1. Samples collection and sequencing

Adult worms of *T. multiceps* were collected 68 days after infection from a dog experimentally infected with protoscoleces. The protoscoleces and cyst wall were separated from the brain of a naturally infected sheep around 3 months infection collected in Gansu province, China. According to the Animal Ethics Procedures and Guidelines of the People’s Republic of China, animals used in this study were cared in accordance with good animal practice and the study was permitted by the Institutional Committee for the Care and Use of Experimental Animals of Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences (no. LVRIAEC 2010-002). Scolex-neck proglottids, immature-mature proglottids and gravid proglottids were obtained by cutting the whole adult worms into three parts based on the tapeworm proglottids characteristics.[10] Oncospheres were hatched by 0.75% sodium hypochlorite and activated under artificial intestinal fluid conditions in vitro. Genomic DNA was extracted from a single adult *T. multiceps* for constructing four Illumina 400-bp insert libraries and sequenced on Illumina HiSeq 2500 platforms. Two PacBio 20-kb insert library were constructed and sequenced on RSII and Sequel, respectively. Total mRNA was extracted from the seven *T. multiceps* materials (activated oncospheres, protoscolices, cyst wall and adult, scolex-neck proglottids, immature-mature proglottids and gravid proglottids) for constructing cDNA libraries (insert 350-bp) (Supplementary Methods and Results section), and sequenced on an Illumina HiSeq 2500 sequencer.

2.2. Genome assembly and annotation

The Illumina raw reads were filtered by trimming the adapter sequence and low-quality part, resulting in a clean and high-quality reads data with average error rate < 0.001. For the PacBio raw data, the short subreads (<2 kb) and low-quality (error rate > 0.2) subreads were filtered out, and only one representative subread was retained for each PacBio read. The clean PacBio reads were assembled by the software canu (http://canu.readthedocs.io/en/stable/index.html#), then Illumina reads were aligned to the contigs by Burrows-Wheeler Aligner-MEM (BWA-MEM),[20] and single base errors in the contigs were corrected by Pilon (v1.16)[21] (Supplementary Methods and Results section).

For the Hi-C-based proximity-guided assembly, we removed duplications and kept reads that uniquely mapped to the reference genome. The assembly package, Lachesis was applied to do clustering, ordering and orienting. Based on the agglomerative hierarchical clustering algorithm,[22] we clustered the scaffolds into 100 N. Then the longest acyclic spanning tree which be called 'trunk' was build according to the relations between the normalized HiC interactions and the scaffolds that were excluded from the trunk were sites that maximized the amount of linkage between adjacent scaffolds. For each chromosome cluster, we got an exact scaffold order of the internal groups and traversed all the directions of the scaffolds through a weighted directed acyclic graph to predict orientation for each of the scaffolds.

The gene models in *T. multiceps* genome were predicted by EVidence Modeler v1.1.1,[23] integrating evidences from *ab initio* predictions, homology-based searches and RNA-seq alignments (Supplementary Methods and Results section). Then, the protein-coding sequences were mapped by RNA-seq data and functionally annotated using UniProt[24] and InterProScan (5.16-55.0) databases.[25] Finally, the gene models were retained if they had at least one supporting evidence from UniProt database, InterProScan domain and RNA-seq data. Gene functional annotation was performed by aligning the protein sequences to NCBI NR, UniProt, eggNOG and KEGG databases using BLASTP v2.3.0+ with E-value cut-off of 10^{-5}. The pathway analysis and functional classification were conducted based on KEGG database.[26] InterProScan[25] was used to assign preliminary gene ontology (GO) terms, Pfam domains and IPR domains to the gene models.

A *de novo* repeat library for *T. multiceps* was constructed by RepeatModeler (v1.0.4; http://www.repeatmasker.org/RepeatModeler.html). transposable elements (TEs) in the *T. multiceps* genome were also identified by RepeatMasker (v4.0.6; http://www.repeatmasker.org/) using both Repbase library and the *de novo* library (Supplementary Methods and Results section). Tandem repeats were predicted using Tandem Repeats Finder v4.07b.[27] The divergence rates of TEs were calculated between the identified TE elements in the genome and their consensus sequence at the repeat family level.

2.3. Evolution analysis

Duplicated genomic fragments were identified by MCscanX,[28] requiring at least 10 paralogous gene pairs per collinear block, and the duplicate_gene_classifier in MCscanX was implemented to classify the origins of the duplicate genes into different types. Orthologous and paralogous gene families were assigned from 14 species (*T. multiceps*, *T. saginata*, *T. asiatica*, *T. solium*, *E. granulosus*, *E. multilocularis*, *Hymenolepis microstoma*, *Brugia malayi*, *Caenorhabditis elegans*, *Clonorchis sinensis*, *Pristionchus pacificus*, *Schistosoma mansoni*, *Fasciola hepatica* and *Trichinella spiralis*) by OrthoFinder[29] with default parameters.

Gene families that contain only one gene for each species were selected to construct the phylogenetic tree. The protein sequences of each gene family was independently aligned by muscle v3.8.31[30] and then concatenated into one super-sequence. The phylogenetic tree was constructed by maximum likelihood (ML) using PhyML v3.0[31] with best-fit model (MtMAN) that was estimated by ProtTest3.[32] The Bayesian-Relaxed Molecular Clock approach was adopted to estimate the neutral evolutionary rate and species divergence time using the programme MCMCTree, implemented in PAML v4.9 package.[33] The calibration time (fossil record time) interval (27–29 Mya) of *Echinococcus* species was adopted from previous results.[34]

Paralogous gene pairs with a length of more than 100 amino acids and E value of 1 e^{-10} were used to calculate synonymous mutation rate (*Ks*) by KaKs_Calculator 2.0 with default parameter.[35] *Ks* values between two species were calculated using the syntenic gene pairs.

The positively selected genes (PSGs) were identified based on the 3,316 one-to-one orthologous gene groups from genomes of the six tapeworms from the results by OrthoFinder.[29] Multiple alignments of protein-coding DNA sequence were generated as referenced by the protein alignments using ParaAT (v1.0) and MAFFT (v7.147b). The gaps in the alignment results were removed, and likelihood ratio tests (LRTs) for selection ($P < 0.05$) on the *T. multiceps* lineage of the phylogenetic tree were performed using Codeml with a modified branch-site model A (model = 2, *N* sites = 2) implemented in the PAML package (v4.8).

2.4. Identification of potential drug targets

The enzyme drug targets from US Food and Drug Administration (FDA) were used to identify homologues in *T. multiceps* based on NR and UniProt databases. Furthermore, the G-protein coupled receptors (GPCRs) proteins were identified based on Pfam annotation. The transmembrane domain of these GPCR proteins in *T. multiceps* was re-annotated by Prediction of transmembrane helices and topology of proteins (HMMTOP).[36] Then, For the GPCR proteins containing more than two transmembrane domains were selected as potential drug targets.[16] On the other hand, the genes of *T. multiceps* involved in serotonin (5-HT) and acetylcholine (Ach) synaptic metabolic pathway were also selected as potential drug targets.[16] Besides, previous report antigens, including antigen B/secreted antigen Ts8B1, EMY162, Tsol15, oncosphere protein, STARP (sporozoite threonine and asparagine-rich protein), cytoplasmic antigen 1, cytoplasmic antigen 1, GP50 and 8 kDa antigen, in other tapeworm species[16] were used to align against the proteins in *T. multiceps* by BLASTP with *E*-value <1 e^{-5}. Finally, to reduce the side-effect of potential drug target for host, the selected drug target genes were aligned to the host protein sequences, including dog, sheep and human respectively, by BLASTP with *E*-value < 1 e^{-5}. The species-specific drug target genes were identified based on the BLASTP hit.

3. Results

3.1. High-quality genome assembly and gene annotation

We generated 31 Gigabase (Gb) PacBio Single-molecule real-time (SMRT) sequences with an average read length of 9.5 kb and 51 Gb Illumina paired-end sequences with a read length of 250 bp using DNA extracted from a single diploid adult worm of *T. multiceps* (Supplementary Fig. S1 and Supplementary Table S1). The PacBio SMRT sequences were assembled by Canu,[37] giving rise to an assembly of 2,050 contigs with a total length of 240 Mb, a N50 length of 756 kb, and a N90 length of 34 kb. The GC content of the genome is 43.7%, which is similar to those of other reported Taeniidae species ranging from 41.9 to 43.2%[16–18] (Table 1), but higher than that of *H. microstoma* (35.9%). Estimated from the k-mer frequency distribution, the genome sizes of sequenced *Taenia* species, including *T. saginata*,[17] *T. asiatica*[17] and *T. multiceps*, are all nearly 250 Mb, larger than those of *Echinococcus* species estimated from 154 to 195 Mb (Table 1 and Fig. 1a; Supplementary Fig. S2). The long continuity information resulting from the PacBio SMRT data indicated that ~96% of the genome has been successfully assembled in *T. multiceps*, which is much higher than those from the Illumina/454 based assembly of other tapeworm species that range from 60 to 75% (Table 1). Based on Hi-C technology, a total of 217.0 Mb (90.4%) assembled contigs were anchored and oriented to the seven linkage groups (Fig. 1b), with the longest 80.1 and shortest 8.5 Mb, respectively, suggesting that the assembly quality was much better than the published genomes of *Taenia* sepecies so far (Table 1).

Table 1. Summary of assembly and annotation of tapeworm genomes

Assembly feature	*T. multiceps*	*T. solium*	*T. saginata*	*T. asiatica*	*E. multilocularis*	*E. granulosus*	*H. microstoma*
Estimated genome size (Mb)	250	—	260[a]	260[a]	154	195	—
Assembled sequences (Mb)	240	122	169	168	115	114	182
Assembly coverage (%)	96	—	65	65	75	60	—
Gaps Ratios (%)	0.05	0.1	1.6	2.5	0.3	2.4	10.5
Longest scaffold size (Mb)	10.5	0.7	7.3	4.2	20.1	15.9	22.3
N50 size of scaffold (Mb)	44.8	0.1	0.6	0.3	13.8	5.2	7.6
N90 size of scaffold (kb)	8,527.5	5.3	29.4	14.3	2,924.3	213.5	40.7
GC content in genome (%)	43.7	42.9	43.2	43.1	42.2	41.9	35.9
Gene annotation							
Number of gene models	13,013	12,481	13,161	13,323	10,663	10,245	12,368
BUSCO complete gene (ratio)	351 (81.8%)	382 (89.0%)	364 (84.8%)	360 (83.9%)	393 (91.6%)	388 (90.4%)	381 (88.8%)
Coding sequence size (Mb)	18.5	15.5	13.3	13.5	15.7	15.2	16.9
Average CDS size (bp)	1,424	1,242	1,011	1,013	1,472	1, 484	1, 366
Average exon number	6.6	5.6	4.4	4.3	6.7	6.8	6.0
Average exon size (bp)	215	222	232	234	218	218	228
GC content in coding region (%)	50.9	50.1	50.1	50.1	50.0	50.0	44.4

The genome data of six other tapeworm species (*T. solium*, *T. saginata*, *T. asiatica*, *E. multilocularis*, *E. granulosus* and *H. microstoma*) were downloaded from WormBase and the NCBI database. The N50 and N90 sizes were calculated based on the assembled genome size.

[a]The estimated genome sizes were reported in Wang *et al.* (2016); For the *E. multilocularis* and *E. granulosus*, the estimated genome sizes were used the Illumina reads by distribution of kmer frequency (Supplementary Fig. S2). We did not estimate the genome sizes of *T. solium* and *H. microstoma*, because of Illumina reads not be found.

The accuracy and completeness of the assembly were evaluated by mapping the Illumina shotgun reads to the assembled reference genome. Significantly, on average 99.8 and 98.2% of the genome-derived and transcriptome-derived reads could be aligned to the reference genome (Supplementary Table S2), respectively, indicating that the majority of the genomic sequence is contained in the current assembly. Furthermore, ~98% of Illumina assembled contigs were mapped to ~50% of the reference genome with 99.2% identity, emphasizing the considerable advantage of using the long-reads assembly (Fig. 1c). Moreover, a total of 13 telomeres were identified located on one end of the contigs with an average size of 10 kb, compared with 12 identified telomeres in *E. multilocularis* and only 1–3 telomeres in other tapeworm genomes, further supporting the high physical coverage of the current assembly (Supplementary Table S4).

The protein-coding genes were predicted on the reference genome by evidence modeler,[23] integrating evidence from *de novo* prediction, transcriptome and homology data. In total, 13,013 gene models were predicted as the reference gene set, with coding regions spanning ~18.5 Mb (7.7%) of the genome (Table 1 and Supplementary Table S5). The distribution of exon number and CDS length in *T. multiceps* is similar to the closely related *E. multilocularis* and *E. granulosus* (Supplementary Fig. S3). 81.8% of eukaryote core genes from OrthoDB (http://www.orthodb.org) were identified as complete in the reference gene set by BUSCO[38] (Table 1). For functional annotation, a total of 11,914 (92%) coding proteins were annotated by functional databases (Supplementary Table S6). On the other hand, a total of 521 non-coding RNA genes (379 tRNA and 142 rRNA) were identified in the *T. multiceps* genome (Supplementary Table S7). Together, these demonstrate that the accuracy of gene annotation we achieved is comparable to or better than the other published tapeworm genomes.[16,17]

3.2. Recent expansion of transposable elements in *Taenia* species

The high assembly coverage we obtained enabled a comprehensive analysis of TEs which have multiple roles in driving genome evolution in eukaryotes.[39] In total, we identified 28.9 Mb (12.0%) TE sequences in the assembled *T. multiceps* genome (Fig. 2a and Supplementary Table S8). Notably, the most abundant LTR retrotransposons present in *T. multiceps* are *Gypsy* elements, accounting for 49.7% of the LTR elements, followed by *Copia* (12.2%) and *ERV1* (3.7%) (Fig. 2b). Furthermore, we used the same strategy to annotate and compare TEs in the published *T. saginata*, *T. asiatica*, *E. multilocularis* and *E. granulosus* genomes. The total TE content in the *Taenia* species is almost twice that found in the *Echinococcus* species, but the overall TE content and classification are similar within the former species (Fig. 2a and Supplementary Table S8).

Next, we analysed the divergence rate of TEs among the available sequenced tapeworm genomes. The divergence rate was calculated by comparing all TE sequences in each subfamily to its corresponding consensus sequence. The results showed that each *Taenia* species had a peak at ~20% divergence rate, containing approximately half of TE copies. In contrast, the *Echinococcus* species had a peak at ~36% divergence rate (Fig. 2c and Supplementary Fig. S4).

Tandem repeat is another kind of repetitive sequence abundant in the eukaryote genome, which plays an important role in maintaining the genome structure. For the sequenced tapeworms, most of the tandem repeats were not correctly assembled and thus missing from the current reference sequences based on Illumina sequencing. In contrast, in the reference assembly of *T. multiceps*, we successfully annotated a total of 20.5 Mb (8.5%) tandem repeat sequences (Fig. 2a), ~92% of which locate in the fragmentally assembled gene-free sequences. These assembled tandem repeat sequences provide a valuable resource for studying the structure of centromere, telomere, as well as heterochromatic regions.

3.3. Gene expansions mostly derived from small-scale gene duplications

To gain insights into an evolutionary perspective for *T. multiceps*, we built a phylogenetic tree based on 271 high-confidence single copy orthologs of 14 helminth worm species by PhyML[40] and estimated divergence time using mcmctree[41] (Fig. 3a). The results

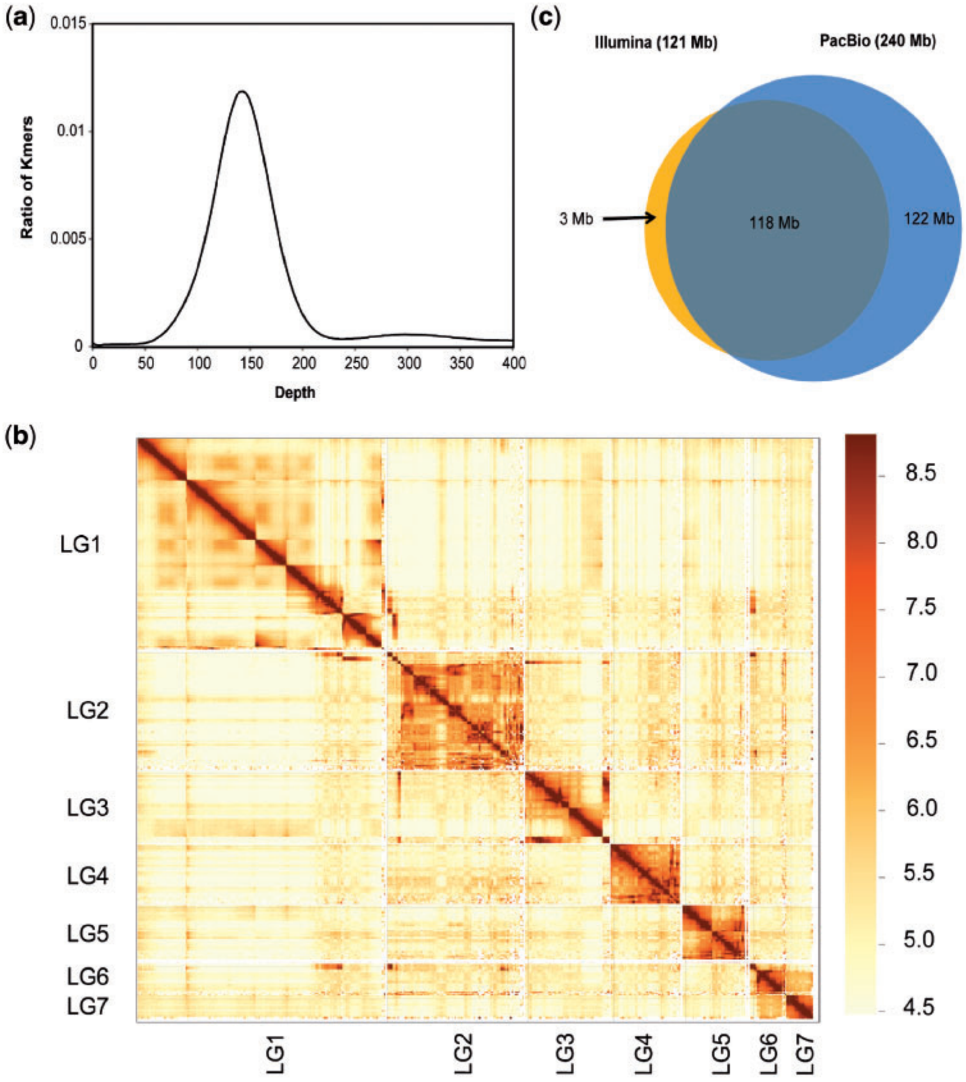


Figure 1. The genome characteristics of *T. multiceps.* (a) Distribution of 19-mer frequency. Error corrected Illumina reads were used to calculate kmer frequency. (b) Hi-C produces a genome-wide contact matrix with 500 kb window between in seven-linkage group (LG). (c) A Venn diagram showing the unique and shared size.

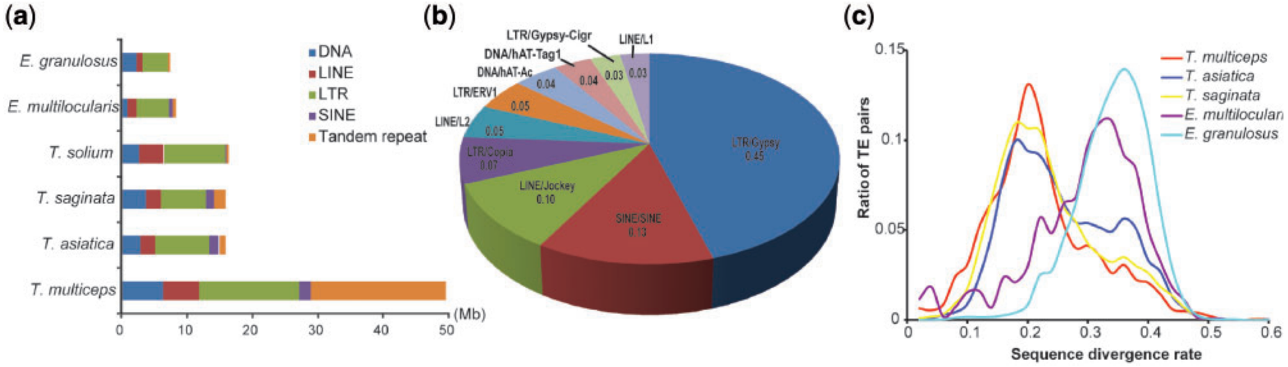


Figure 2. Repetitive sequences content and TE divergence rate in tapeworm genomes. (a) Classification and contents of repetitive sequences in *T. multiceps* compared with *T. saginata*, *T. asiatica*, *T. solium*, *E. granulosus* and *E. multilocularis*. (b) The ratio of top 10 abundant TEs families in *T. multiceps* genome. (c) Distribution of TE divergence rate in tapeworm genomes.

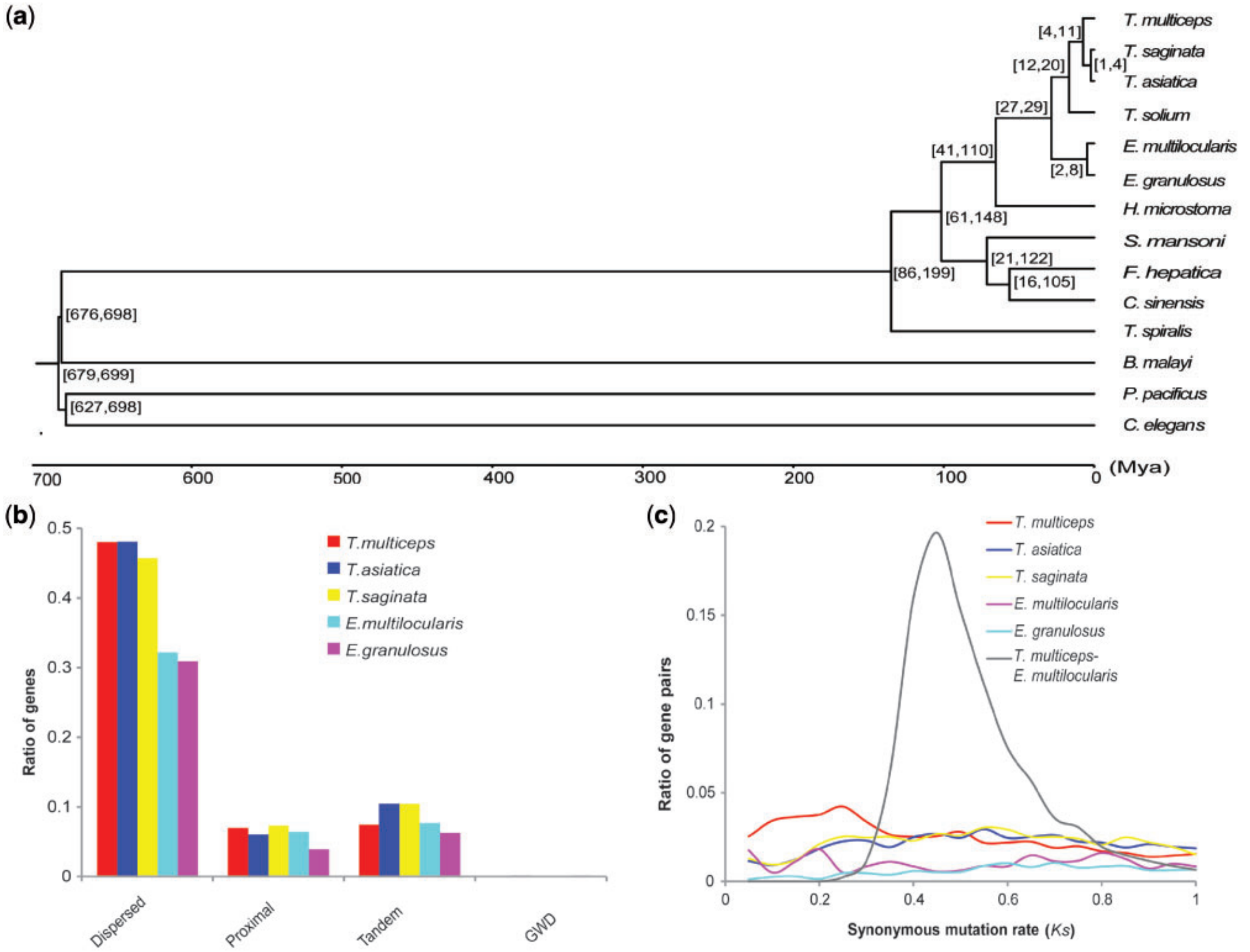


Figure 3. Evolution of the *T. multiceps* genome. (a) Dated tree for 14 species. The age of each node is indicated by 95% CI. (b) Distribution of different types of gene duplication in tapeworm genomes. WGD means whole genome duplication. (c) Distribution of synonymous mutation rate (K_s) values for paralogous gene pairs in each tapeworm genome. K_s values between *T. multiceps* and *E. multilocularis* are calculated using the syntenic ortholog gene pairs.

indicate that *T. multiceps* diverged from *T. saginata* and *T. asiatica* 4–11 million years ago (Mya) (95% CI, and from *T. solium* 13–20 Mya. Analysis of gene families in the high-confidence gene set and 14 sequenced helminth worm genomes identified 122 *T. multiceps*-specific gene families (Supplementary Fig. S5).

To investigate the evolutionary dynamics of genome structure, we performed comparative genomic analysis among *T. multiceps*, *T. saginata* and *E. multilocularis*, using proteins as markers to identify syntenic genes by MCScan pipeline with parameter –w 10. In the *T. multiceps* genome, 42.5% genes were identified synteny relationships with *T. saginata*; whereas 5.15% genes were identified synteny relationships with *E. multilocularis*. Scrutiny of the local scale synteny, we found that the ratio of syntenic genes between *T. multiceps* and *T. saginata* was ~71, versus 68% between *T. multiceps* and *E. multilocularis*, which is reasonable considering their phylogenetic distance (Supplementary Fig. S6).

In *T. multiceps*, there were 6,300 dispersed duplicated genes (48.4% of total genes), which were much more than proximal duplicated genes 1,446 (11.1%) and tandem duplicated genes 1,128 (8.7%). The proportion of dispersed duplicated genes in the *T. multiceps* was similar to that in the *T. saginata* and *T. asiatica* (45.8 and 48.1%, respectively), but was significantly higher than that in the *E. multilocularis* and *E. granulosus* (32.2 and 30.9%, respectively; $P < 0.001$; Fig. 3b), consistent with previously reports for the *T. saginata* and *T. asiatica* genomes.[17] These results suggest that gene expansions in *Taenia* species are mostly derived from small-scale gene duplications, instead of whole genome duplication.

Using the data obtained for the duplicated genes, we analysed the synonymous mutation rate (K_s) of these paralogous gene pairs. No obvious K_s peaks were observed for any of the five tapeworms, reinforcing our inference of no recent whole genome duplication having occurred. The duplicated genes were likely generated gradually over time with relatively more recent duplication events in the *Taenia spp.* compared with *Echinococcus* (Fig. 3c). In total, 3,697 genes with $K_s < 0.3$ were used to analyse the gene function. Results showed that the Hsp70, fibronectin type III, transcription factor families (T-box, homeobox and zinc finger) were significantly enriched in these duplicated genes of *T. multiceps* ($P < 0.001$), and the genes involved in the MAPK signalling pathway (map04010) were significantly enriched ($P = 0.0071$). These observations indicate that genes underlying environmental adaptability and dosage effect tend to be conserved in the *T. multiceps* genome.

3.4. PSGs imply extended evolutionary pressures in the *Taenia* lineage

Positive selection provides evolutionary innovation during adaptation to a new environment. We evaluated the role of positive selection in the evolution of *T. multiceps* by branch-site model analysis in PAML[33] based on orthologous genes from the Taeniidae tapeworms, and identified 204 PSGs (LRT, false discovery rate < 0.05) (Supplementary Material S1). In fact, we found PSGs of essential genes in many fundamental cellular processes, particularly those containing molecular functions of the DNA replication and repairing-related genes, transcription regulation, translation, and protein kinase (Supplementary Material S1). *T. multiceps* has a life cycle involving a tissue tropism in the brain of ungulates, which, in terms of nutrient availability, represents a different environment compared with most of the other *Taenia* tapeworms. Strong positive selection signals were observed in energy-metabolism related genes, such as genes involved in carbohydrate metabolism and uptake, including Glucosidase II beta subunit, glucan (1, 4 alpha), branching enzyme 1 and nicotinamide adenine dinucleotid ubiquinone oxidoreductase sgdh subunit, guanosine diphosphate fucose transporter. Notably, homologous genes probably involved in the speciation of *T. asiatica*[17] were also included in the PSGs of *T. multiceps* (e.g. carbonic anhydrase, amiloride sensitive cation channel 4A, dynein light chain), suggesting the frequent involvement of these genes in adaptation to a new host or tissue environment during tapeworm evolution.

3.5. Expansion and up-regulation of the T-box 6 genes produce more proglottids in strobilization

Both asexual and sexual reproduction occurs in *T. multiceps*. Strobilization is a process of asexual reproduction whereby the adult tapeworm buds off new segments from the undifferentiated and germinative neck region, forming a series of proglottids which when mature give rise sexually to ovoid eggs.[42] The number of proglottids varies among the different tapeworms, with hundreds of segments produced by *Taenia* species but with only three to five being produced by in the *Echinococcus* species.[10]

The T-box family of transcription factors shows broad participation in the development of all metazoans; the subfamily T-box 6 (Tbx6), especially plays a central role in somite polarization in the mouse.[43] In this study, the number of T-box transcription factors was shown to be significantly expanded in the Taeniidae, representing on average 31 genes in the 4 *Taenia* species and 5 genes in the 2 *Echinococcus* species ($P = 1.02\ e^{-4}$) (Supplementary Table S11). Specifically, there were 31, 23, 20 and 12 Tbx6 genes in *T. multiceps*, *T. asiatica*, *T. saginata* and *T. solium*, respectively, significantly more than in *E. granulosus* and in *E. multilocularis*, (each having only one Tbx6 gene ($P < 0.001$). Based on a phylogenetic analysis of the Tbx6 genes in the Taeniidae, we found that the genes have been expanded in the ancestors of the *Taenia* species after their divergence from *Echinococcus*. Furthermore, the sequences of the Tbx6 genes have been divergent along the *T. multiceps* lineage which contained different motifs on each clade (Fig. 4a and b). These results imply that the Tbx6 genes might gain rapid functional divergence after gene duplication. On the other hand, to investigate the gene expression pattern of Tbx6, we analysed genes expressed with a cut off fragments per kilobase per million (FPKM) > one in the four samples of three different *T. multiceps*, life cycle stages (one from oncospheres, two from the coenurus and one from the adult). It was surprising that 24 of 31 Tbx6 genes were expressed in the adult stage, whereas no or very limited expression was apparent in the other stages (Fig. 4c). Overall, the different number of proglottids produced in the *Taenia* species compared with the *Echinococcus* species may be as a result of the expansion of the Tbx6 genes in the former group which may play an important role in the strobilization process in adult tapeworms.

3.6. Coenurus occupy nutrient and release toxicity to the host CNS

Coenurosis, as a parasitic zoonosis, could cause potential health threat to the human beings and serious economic losses in sheep

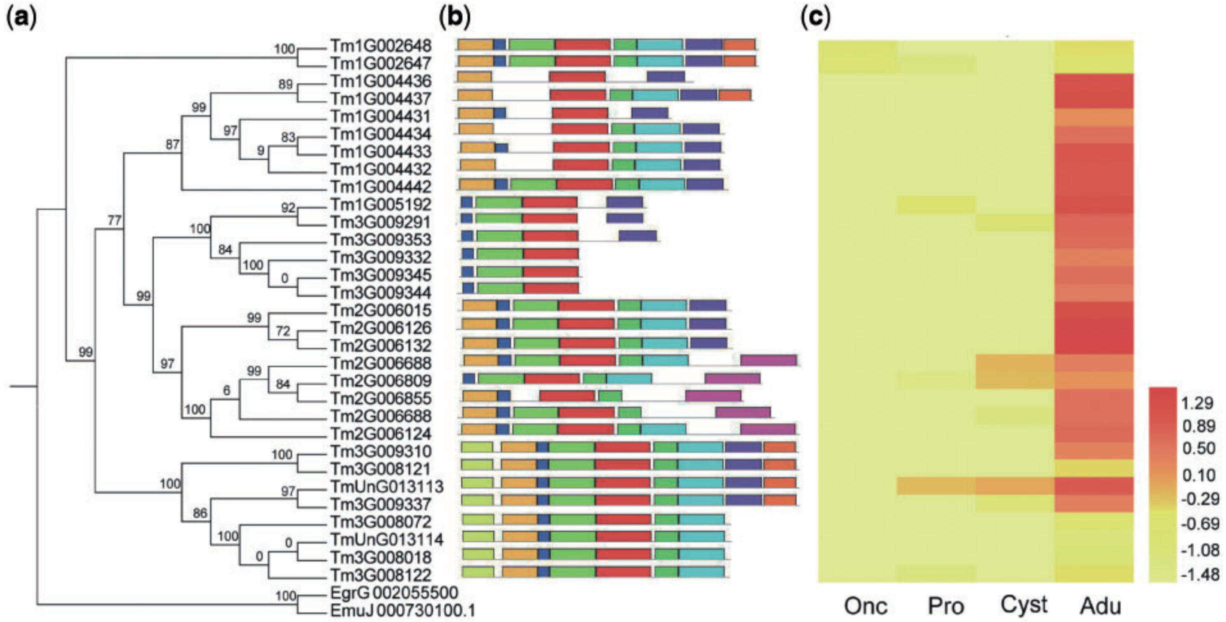


Figure 4. Expansion and divergence of the Tbx6 subfamily in *T. multiceps*. (a) ML Phylogenetic tree of Tbx6 genes in *T. multiceps*. Two Tbx6 genes of *Echinococcus* were used as outgroup. (b) Motifs of Tbx6 genes in *T. multiceps*. Each colour of rectangle represents a motif. (c) Expression patterns of Tbx6 genes in stages (Onc, Pro, Cyst and Adu) of *T. multiceps*.

farms.[7,8] It is a fatal CNS disease to the intermediate host sheep. The coenurus cyst contains hundreds of protoscolices and can grow to the size of a hen's egg, caused coenurosis by pressing the brain to induce neural lesion (Fig. 5a). Moreover, the nervous system toxicity related with wastes from metabolic process and neurodegenerative disorder protein of parasite (Fig. 5b), may also occupy an important position in the causation of sheep CNS disease.

To proliferate and grow, the coenurus larva needs to exchange nutrients including glucose, amino acids, lipids and other materials between the cyst wall and host brain. In this respect, glutamate is not only able to provide metabolic energy but also acts as a neurotransmitter to regulate signal transmissions between the synaptosomes and neurons.[44,45] Several members of the SLC1A glutamate transporter family have been reported to be involved in glutamate transmission.[46] We identified 7 SLC1A gene family members in *T. multiceps*, which was considerably more than found present in the five other tapeworms (mean 4.4; median 4; $P < 0.05$), suggesting an increased glutamate transmission capability. Among the SLC1A genes, SLC1A4 composed a major percentage of 50%, which showed a specific expression in different stages (Supplementary Material S2). It is noteworthy that three members of SLC1A4 were abundantly expressed in the coenurus, indicating that the SLC1A4 genes might be functional divergent in *T. multiceps*. Sugars are also centrally involved in metabolism and energy production and storage, and we identified in the *T. multiceps* genome one glucose and four sugar transporters having a high level of expression in the cyst wall or protoscoleces. Moreover, two important genes mediating glucose metabolism and energy production (glucose-6-phosphate-dehydrogenase, Tm1G004355 and glucose-6-phosphate isomerase, Tm3G008866), were highly expressed in protoscoleces (Supplementary Material S2). Besides obtaining nutrients from the mammalian host, the coenurus stage causes neurologic dysfunction by the production and release of neurotoxic substances. During metabolism, the consumption of glucose and oxygen by protoscoleces results in the production of carbon dioxide that is released outside the cyst wall, resulting in host cell toxicity. Additionally, there is high expression of sodium/potassium-transporting ATPase (Tm1G003772) and calcium pump (Tm1G002964) in protoscoleces and this may impact on ion homeostasis in the host neuron microenvironment. It has been shown that the imbalance of glutamate diffusing over synaptosomes could change synaptic function and interfere with signal transmission, causing toxicity in the host nervous system.[46] The expression of glutamate dehydrogenase 1 (Tm3G009148; FPKM = 577) in protoscoleces was significantly higher than in the other stages ($P < 0.001$), and this could result in an imbalance in the concentration of glutamate in the host brain, resulting in synaptic dysfunction. Moreover, two gene homologues of neurodegenerative disorder were highly expressed in protoscoleces (Supplementary Material S2), and these may also contribute to nervous system toxicity.

3.7. Environment adaptation and immune evasion contributed by horizontal gene transfer

Horizontal gene transfer (HGT), or the passage of genetic material between non-mating species, is a ubiquitous process that increases the divergence and improves the physiological metabolism of recipient organisms.[47,48] We identified HGT genes in *T. multiceps* obtained from the host using a published determination pipeline with modifications.[49]

Eleven genes were identified as HGT candidates in the *T. multiceps* genome from three hosts, including sheep (*Ovis aries*), dog (*Canis lupus familiaris*) and goat (*Capra hircus*) (Supplementary Table S12). In phylogenetic analysis, the 11 candidate genes displayed a close relationship with host orthologs but were much more distant from the orthologs of species close to *T. multiceps* (Fig. 6a and Supplementary Fig. S7a–j). Among the 11 HGT candidates, all have high identity (>70%) and coverage (>80%) with the orthologs from mammalian hosts, such as Tm1G003541 (Fig. 6b). To confirm that the HGT genes were transferred from the host rather than due to assembly error, the Illumina pair-end reads were mapped to the region flanking a HGT candidate, Tm1G003541. The results showed that the pair-end reads could cross the HGT candidate region continuously in which the flanking genes represented syntenic relationships with *E. multilocularis*, excepting the HGT candidate (Fig. 6c). These results indicated that the HGT candidate (Tm1G003541), which might bind to the MHCII from the host and regulate the release of Type I interferon, was probably transferred from host and play important role to interfere in the immune recognition and following immune responses. Furthermore, 8 of 11 candidates had unigene evidence from a previous shotgun assembled transcriptome of *T. multiceps*.[2] In our transcriptome data on stages and tissues of *T. multiceps*, the expression of HGT candidates showed that a protein related to translational elongation (Tm1G005508) and a Hsp90 (Tm1G003541) had high expression levels, especially in the cyst wall and protoscoleces, whereas five Hsp70 proteins showed low expression patterns (Fig. 6d). The GO analysis of the 11 candidates in *T. multiceps* showed that the HGT could be clustered into 4 pools, including protein modification, adaptation, reproduction and heat shock (Fig. 6e). In summary, these results suggest that the *T. multiceps* genes transferred from hosts may play important roles in environment adaptation and immune evasion.

3.8. Genome-based new molecular targets for drug intervention

Coenurosis is a devastating disease and difficult to treat and the *T. multiceps* genome project makes feasible the identification of novel targets for drug discovery. Historically, enzymes have been a key target group for drug discovery.[50] Based on Food and Drug Administration (FDA) and clinical studies, we identified a total of 210 known drug target genes involved with 56 enzyme systems in *T. multiceps* (Supplementary Table S13). In addition to enzymes, kinases, GPCRs and ligand-gated ion channels (LGICs) may also be important drug targets for the treatment of *T. multiceps*. We identified a total of 284 kinases and 38 GPCRs in *T. multiceps* (Supplementary Tables S14 and S15) and 45 members of four major LGIC families. Several effective anthelminthic drugs interfere with one of the multiple forms of neural control and communication.[51] Therefore, we searched the signalling pathways of neurotransmitters and identified 11 and 8 conserved genes in the Ach and serotonin pathways respectively (Supplementary Tables S16 and S17), which may serve as broad-spectrum drug targets for treating coenuriasis and other cestodiases.

To identify species-specific genes in *T. multiceps*, we compared the above potential targets to host proteins by BLASTP (cutoff: $1\ e^{-5}$). In total, no homologues of 218 specific targets genes in *T. multiceps* were found in humans, canines and sheep (Supplementary Material S3), indicating that these species-specific targets not only have potential for developing as diagnostics or as vaccines but may also have limited side effects on mammalian hosts.

4. Discussion

SMRT sequencing offers longer length reads than second-generation sequencing technologies, making it well suited for unsolved problems in genome research.[52] *De novo* assemblies using PacBio sequencing,

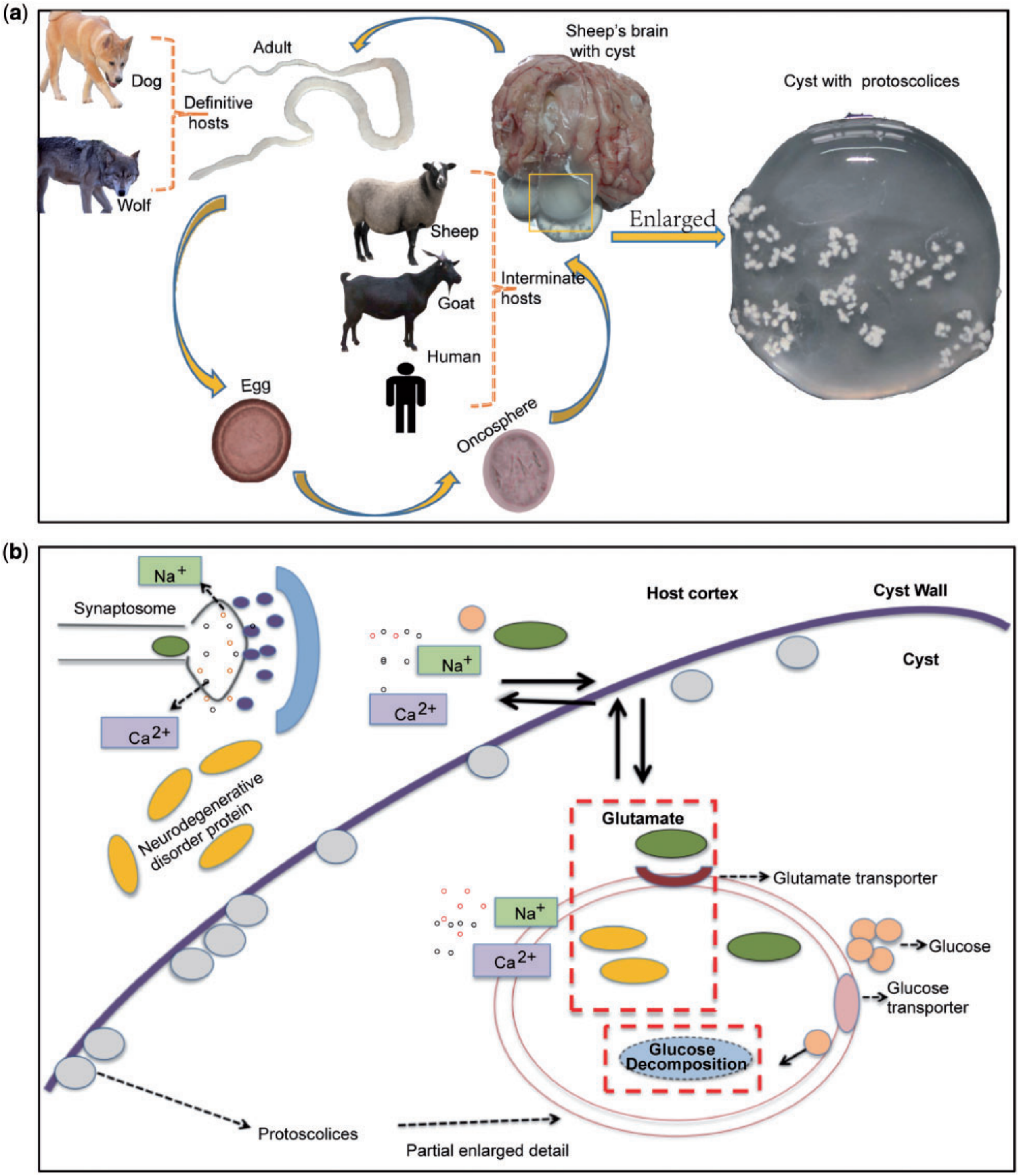
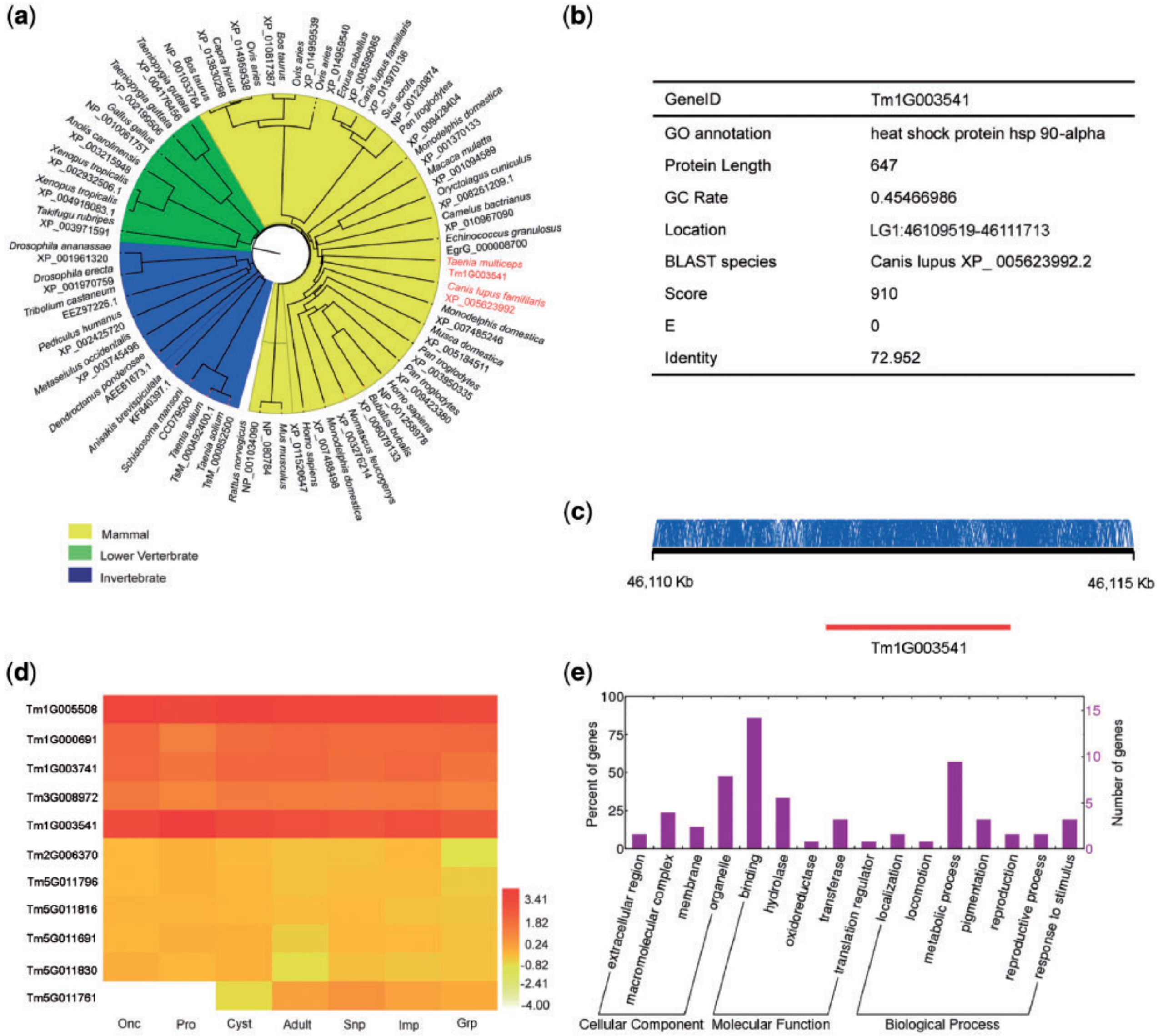


Figure 5. Effect of the coenurus on the sheep CNS. (a) The diagrammatic map of *T. multiceps* life cycle. The enlarged photo is a growing coenurus cyst with clusters of protoscoleces. (b) Possible mechanisms of the effects of the coenurus on the intermediate host brain. Nutrients and waste materials, are transported outside and inside. The word 'Partial enlargement' means the big circle at the right of the arrow is the enlarged model of protoscoleces. Various shapes exist in the figure, and the same shape with the same colour refers to the same protein or ion, and two squares drawn with a red dotted line indicate the TCA cycle and neurotoxic materials, respectively. The two-way arrows with thickened black line show the double-direction transmission of proteins and ions, whereas the single-way arrow indicates one direction; the single-way arrow with a dotted line points to the name of the shape.

as applied here, can close gaps in current reference assemblies, which are mostly composed of large repeat sequences.[52] Using this high coverage genome assembly, we have identified much more TEs and tandem repeat sequences in *T. multiceps* than in the previously published *Taenia* genomes, enabling a comprehensive analysis of the content of each type of repeat sequence as well as telomere structures. The high-quality assembly allows fine annotation of gene models, facilitating evolutionary and functional analysis.



GeneID	Tm1G003541
GO annotation	heat shock protein hsp 90-alpha
Protein Length	647
GC Rate	0.45466986
Location	LG1:46109519-46111713
BLAST species	Canis lupus XP_ 005623992.2
Score	910
E	0
Identity	72.952

Figure 6. Predicted host-origin HTGs in *T. multiceps*. (a) Phylogenetic tree of Tm1G003541 and its homologues. The clades highlighted in yellow, green and blue represent mammals, lower vertebrates and invertebrates, respectively, while the font in red represents tapeworm and dog. (b) The characteristic of sequence alignment between HTG candidate (Tm1G003541). (c) The local pair-end reads mapping of the flanking region of HTG candidate (Tm1G003541). The red box indicates the HTG gene. (d) Expression patterns of *T. multiceps* HTGs based on transcriptome analysis. (e) GO annotation of the HGT candidates.

The genome and transcriptome data we generated provide a valuable resource for better understanding the molecular biology, physiology, phylogeny and mechanisms of pathogenesis in *T. multiceps* and coenurosis. In this study, we found expansion of Tbx6 genes and their up-regulation in the adult stage of *T. multiceps*, which may be associated with the increased number of proglottids. We also found several genes involved in interaction with the host CNS, and several genes likely derived from horizontal transfer, as a result of adaption to the parasitic life style. Furthermore, the species-specific gene targets identified in *T. multiceps* provide candidatesfor development as novel immunological diagnostic markers, as well as preventative vaccines and therapeutic drugs. This high-quality *T. multiceps* reference genome provides not only a comprehensive foundation for deeper understanding of the genetic and evolution of tapeworms, but also an invaluable resource to accelerate the development of novel treatments and tools for the control of coenurosis.

Acknowledgements

We thank W.J. Tian, G. Yao and F.S. Cai from Lanzhou Veterinary Research Institute, CAAS for helping us to collect parasite materials. We also thank J. Ruan and D.M. Fang from Agricultural Genomic Institute at Shenzhen, CAAS for providing suggestions on genome assembly.

Accession numbers

The assembly and raw data have been deposited in GenBank with the following accession numbers: MRSW00000000 for Whole Genome Shotgun project of *T. multiceps* under Bio Project accession PRJNA307624; The version described in this paper is version MRSW02000000; SRR3103080, SRR3219054, SRR3219055 and SRR3219056 for the *T. multiceps* genomic Illumina DNA data; SRX1531851, SRX3851533, SRX3851532 and SRX3851531 for the *T. multiceps* genomic PacBio SMRT data; SRR5032913, SRR5032914, SRR5032918, SRR5032916, SRR5032912, SRR5032915 and

SRR5032917 for the *T. multiceps* RNA-seq data. SRX3856587 for the *T. multiceps* genomic Hi-C data.

Funding

The study was funded by the State Key Laboratory of Veterinary Etiological Biology of Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences; Agricultural Science and Technology Innovation Program (ASTIP) of Chinese Academy of Agricultural Sciences (CAAS); and was also partially supported by the National Natural Science Foundation of China (Grant No. 31702228); the National Beef Cattle and Yak Industrial Technology System, NBCITS, MOA (No. CARS-37); the Elite Youth Program of Chinese Academy of Agricultural Sciences and National Key Research and Development Program of China (2016YFC1200600); Shenzhen science and technology program (JCYJ20150630165133395).

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at *DNARES* online.

References

1. Oryan, A., Akbari, M., Moazeni, M. and Amrabadi, O. R. 2014, Cerebral and non-cerebral coenurosis in small ruminants, *Trop. Biomed.*, **31**, 1–16.
2. Wu, X., Fu, Y., Yang, D., et al. 2012, Detailed transcriptome description of the neglected cestode *Taenia multiceps*, *PLoS One*, **7**, e45830.
3. Scala, A. and Varcasia, A. 2006, Updates on morphobiology, epidemiology and molecular characterization of coenurosis in sheep, *Parassitologia*, **48**, 61–3.
4. Bussell, K. M., Kinder, A. E. and Scott, P. R. 1997, Posterior paralysis in a lamb caused by a coeneurus cerebralis cyst in the lumbar spinal cord, *Vet. Rec.*, **140**, 560.
5. Achenef, M., Markos, T., Feseha, G., Hibret, A. and Tembely, S. 1999, Coenurus cerebralis infection in Ethiopian highland sheep: incidence and observations on pathogenesis and clinical signs, *Trop. Anim. Health Prod.*, **31**, 15–24.
6. Al-Riyami, S., Ioannidou, E., Koehler, A. V., et al. 2016, Genetic characterisation of *Taenia multiceps* cysts from ruminants in Greece. *Infect. Genet. Evol.*, **38**, 110–6.
7. Sharma, D. K. and Chauhan, P. P. S. 2006, Coenurosis status in Afro-Asian region: a review, *Small Ruminant Res.*, **64**, 197–202.
8. El-On, J., Shelef, I., Cagnano, E. and Benifla, M. 2008, Taenia multiceps: a rare human cestode infection in Israel, *Vet. Ital.*, **44**, 621–31.
9. Stunkard, H. W. 1962, The organization, ontogeny, and orientation of the Cestoda, *Quart. Rev. Biol.*, **37**, 23–34.
10. Roberts, L. S. S. and Janovy, G. D. 2009, *Gerald D. Schmidt & Larry S. Roberts' Foundations of parasitology*.
11. Pau, A., Perria, C., Turtas, S., Brambilla, M. and Viale, G. 1990, Long-term follow-up of the surgical treatment of intracranial coenurosis, *Br. J. Neurosurg.*, **4**, 39–43.
12. Pau, A., Turtas, S., Brambilla, M., Leoni, A., Rosa, M. and Viale, G. L. 1987, Computed tomography and magnetic resonance imaging of cerebral coenurosis, *Surg. Neurol.*, **27**, 548–52.
13. Gauci, C., Vural, G., Öncel, T., et al. 2008, Vaccination with recombinant oncosphere antigens reduces the susceptibility of sheep to infection with *Taenia multiceps*, *Int. J. Parasitol.*, **38**, 1041–50.
14. Varcasia, A., Tosciri, G., Coccone, G. N., et al. 2009, Preliminary field trial of a vaccine against coenurosis caused by *Taenia multiceps*, *Vet. Parasitol.*, **162**, 285–9.
15. Lightowlers, M. 2006, Cestode vaccines: origins, current status and future prospects, *Parasitology*, **133**, S27–42.
16. Tsai, I. J., Zarowiecki, M., Holroyd, N., et al. 2013, The genomes of four tapeworm species reveal adaptations to parasitism, *Nature*, **496**, 57–63.
17. Wang, S., Wang, S., Luo, Y., et al. 2016, Comparative genomics reveals adaptive evolution of Asian tapeworm in switching to a new intermediate host, *Nat. Commun.*, **7**, 12845.
18. Zheng, H., Zhang, W., Zhang, L., et al. 2013, The genome of the hydatid tapeworm *Echinococcus granulosus*, *Nat. Genet.*, **45**, 1168–75.
19. Wu, X., Fu, Y., Yang, D., et al. 2013, Identification of neglected cestode *Taenia multiceps* microRNAs by illumina sequencing and bioinformatic analysis, *BMC Vet. Res.*, **9**, 162.
20. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754–60.
21. Walker, B. J., Abeel, T., Shea, T., et al. 2014, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement, *PloS One.*, **9**, e112963.
22. Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O. and Shendure, J. 2013, Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions, *Nat. Biotechnol.*, **31**, 1119–25.
23. Haas, B. J., Salzberg, S. L., Zhu, W., et al. 2008, Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments, *Genome Biol.*, **9**, R7.
24. Wu, C. H., Apweiler, R., Bairoch, A., et al. 2006, The Universal Protein Resource (UniProt): an expanding universe of protein information, *Nucleic Acids Res.*, **34**, D187–91.
25. Quevillon, E., Silventoinen, V., Pillai, S., et al. 2005, InterProScan: protein domains identifier, *Nucleic Acids Res.*, **33**, W116–20.
26. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. 2004, The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, **32**, D277–80.
27. Benson, G. 1999, Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Res.*, **27**, 573–80.
28. Tang, H. B., Bowers, J. E., Wang, X. Y., Ming, R., Alam, M. and Paterson, A. H. 2008, Perspective - synteny and collinearity in plant genomes, *Science*, **320**, 486–8.
29. Emms, D. M. and Kelly, S. 2015, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, *Genome Biol.*, **16**, 14.
30. Edgar, R. C. 2004, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.*, **32**, 1792–7.
31. Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. 2010, New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0, *Syst. Biol.*, **59**, 307–21.
32. Darriba, D., Taboada, G. L., Doallo, R. and Posada, D. 2011, ProtTest 3: fast selection of best-fit models of protein evolution, *Bioinformatics* , **27**, 1164–5.
33. Yang, Z. H. 2007, PAML 4: phylogenetic analysis by maximum likelihood, *Mol. Biol. Evol.*, **24**, 1586–91.
34. Knapp, J., Nakao, M., Yanagida, T., et al. 2011, Phylogenetic relationships within *Echinococcus* and *Taenia* tapeworms (Cestoda: taeniidae): An inference from nuclear protein-coding genes, *Mol. Phylogenet. Evol.*, **61**, 628–38.
35. Wang, D., Zhang, Y., Zhang, Z., Zhu, J. and Yu, J. 2010, KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies, *Genomics Proteomics Bioinformatics*, **8**, 77–80.
36. Tusnady, G. E. and Simon, I. 2001, The HMMTOP transmembrane topology prediction server, *Bioinformatics* , **17**, 849–50.
37. Berlin, K., Koren, S., Chin, C.S., Drake, J.P., Landolin, J.M. and Phillippy, A.M. 2015, Assembling large genomes with single-molecule sequencing and locality-sensitive hashing, *Nat. Biotechnol.*, **33**, 623–30.
38. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. and Zdobnov, E. M. 2015, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, *Bioinformatics*, **31**, 3210–2.

39. Feschotte, C., Jiang, N. and Wessler, S. R. 2002, Plant transposable elements: where genetics meets genomics, *Nat. Rev. Genet.*, **3**, 329–41.
40. Guindon, S. and Gascuel, O. 2003, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, *Syst. Biol.*, **52**, 696–704.
41. dos Reis, M. and Yang, Z. H. 2011, Approximate likelihood calculation on a phylogeny for bayesian estimation of divergence times, *Mol. Biol. Evol.*, **28**, 2161–72.
42. Marr, M. 2011, *Developmental Regulatory Genes in Parasitic Flatworms*. Doctoral dissertation, Department of Life Sciences and the Natural History Museum, Imperial College London.
43. Papaioannou, V. E. 2014, The T-box gene family: emerging roles in development, stem cells and cancer, *Development*, **141**, 3819–33.
44. Kanamori, K. 2017, In vivo N-15 MRS study of glutamate metabolism in the rat brain. *Anal. Biochem.*, **529**, 179–192.
45. Natesan, V., Mani, R. and Arumugam, R. 2016, Clinical aspects of urea cycle dysfunction and altered brain energy metabolism on modulation of glutamate receptors and transporters in acute and chronic hyperammonemia, *Biomed. Pharmacother.*, **81**, 192–202.
46. Macrez, R., Stys, P. K., Vivien, D., Lipton, S. A. and Docagne, F. 2016, Mechanisms of glutamate toxicity in multiple sclerosis: biomarker and therapeutic opportunities, *LancetNeurol.*, **15**, 1089–102.
47. Choi, I. G. and Kim, S. H. 2007, Global extent of horizontal gene transfer, *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 4489–94.
48. Gilbert, C., Schaack, S., Pace, J. K., Brindley, P. J. and Feschotte, C. 2010, A role for host-parasite interactions in the horizontal transfer of transposons across phyla, *Nature*, **464**, 1347., 1347-U1344.
49. Keeling, P. J. 2009, Functional and ecological impacts of horizontal gene transfer in eukaryotes, *Curr. Opin. Genet. Dev.*, **19**, 613–9.
50. Alexander, S. P. H., Benson, H. E., Faccenda, E., et al. 2013, The concise guide to pharmacology 2013/14: enzymes, *Br. J. Pharmacol.*, **170**, 1797–867.
51. McVeigh, P., Atkinson, L., Marks, N. J., et al. 2012, Parasite neuropeptide biology: seeding rational drug target selection? *Int. J. Parasitol. Drug*, **2**, 76–91.
52. Rhoads, A. and Au, K. F. 2015, PacBio sequencing and its applications, *Genomics. Proteomics Bioinformatics.*, **13**, 278–89.