Evolutionary characterization of *Ty3/gypsy*-like LTR retrotransposons in the parasitic cestode *Echinococcus granulosus*

YOUNG-AN BAE*

Department of Microbiology, Gachon University College of Medicine, 191 Hambakmoe-ro, Yeonsu-gu, Incheon 21936, Republic of Korea

(*Received 17 April 2016; revised 16 June 2016; accepted 19 July 2016; first published online 30 August 2016*)

SUMMARY

Cyclophyllidean cestodes including *Echinococcus granulosus* have a smaller genome and show characteristics such as loss of the gut, a segmented body plan, and accelerated growth rate in hosts compared with other tissue-invading helminths. In an effort to address the molecular mechanism relevant to genome shrinkage, the evolutionary status of long-terminal-repeat (LTR) retrotransposons, which are known as the most potent genomic modulators, was investigated in the *E. granulosus* draft genome. A majority of the *E. granulosus* LTR retrotransposons were classified into a novel characteristic clade, named *Saci-2*, of the *Ty3/gypsy* family, while the remaining elements belonged to the *CsRn1* clade of identical family. Their nucleotide sequences were heavily corrupted by frequent base substitutions and segmental losses. The ceased mobile activity of the major retrotransposons and the following intrinsic DNA loss in their inactive progenies might have contributed to decrease in genome size. Apart from the degenerate copies, a *gag* gene originating from a *CsRn1*-like element exhibited substantial evidences suggesting its domestication including a preserved coding profile and transcriptional activity, the presence of syntenic orthologues in cestodes, and selective pressure acting on the gene. To my knowledge, the endogenized *gag* gene is reported for the first time in invertebrates, though its biological function remains elusive.

Key words: *Echinococcus granulosus*, long-terminal-repeat retrotransposon, *Saci-2* clade, genome size, domestication of *gag* gene.

INTRODUCTION

Tapeworms (Platyhelminthes, Cestoda) are endoparasites that depend on two different organisms for their life cycle. The parasites' embryos develop into metacestodes (the larval forms) via oncospheres in intermediate invertebrate or vertebrate hosts and these larvae sexually mature in other vertebrates that act as definitive hosts. Four genera of tapeworms infect humans opportunistically during the larval (*Echinococcus* and *Taenia*) or adult (*Hymenolepis* and *Diphyllobothrium*) stage and thus, have a significant medical impact (Smyth and McManus, 1989). Species of *Echinococcus granulosus* complex parasitize the small intestine of canids as an adult, whereas embryos develop into metacestodes in diverse hosts including sheep, cattle, pig, horse and camel, depending on its genotype (Thompson and McManus, 2002). Humans can also act as an intermediate host upon accidental ingestion of eggs, the majority of which develop into cysts in the liver and lungs. The cystic echinococcosis (CE), infection of *E. granulosus* metacestodes, is a widespread and severe zoonosis, which is particularly prevalent in the Mediterranean, central Asia, Northern and Eastern Africa, Australia and southern South America (Jenkins *et al*. 2005; da Silva, 2010). The annual incidence of CE ranges from 1 to 200 per 100 000 inhabitants and more than 20 million people are at risk of infection in endemic areas (Craig *et al*. 2007).

Neodermatan platyhelminthes have evolved endo- (Trematoda and Cestoda) and ecto-parasitic (Monogenea) mode of life. Phylogenetic analyses based on molecular information showed that cestodes and trematodes form a monophyletic clade, and monogeneans form a basal paraphyletic clade with respect to them. This fact might suggest that obligate parasitism has arisen only once during the course of platyhelminth evolution (reviewed in Olson and Tkach, 2005; Littlewood, 2006). Cladistic investigations further assumed that ectoparasitism between free-living flatworms and fish preceded endoparasitism, and that cestodes and trematodes share a common ancestor despite significant differences in their life histories. Compared with other human-invading helminths, cestodes display characteristics such as the complete loss of a gut and a modified, segmented body plan, making them an extreme representative of parasitism (Thompson and McManus, 2002). Therefore, cestodes are likely to have undergone a series of genomic modifications related to their unique phenotypes.

The genome sizes of cyclophyllidean tapeworms examined (haploid DNA content per cell) were ranged from 114·9 (*E. granulosus*) to 141·1 Mb

* Corresponding author: Department of Microbiology, Gachon University College of Medicine, 191 Hambakmoe-ro, Yeonsu-gu, Incheon 21936, Korea. E-mail: yabae03@gmail.com

doi:10.1017/S0031182016001499

(*Hymenolepis microstoma*) (Tsai *et al*. 2013). The values were seemingly much smaller than those of trematodes and turbellarian, *Schistosoma mansoni* (364·5 Mb; Protasio *et al*. 2012), *Clonorchis sinensis* (547·1 Mb; Huang *et al*. 2013) and *Schmidtea mediterranea* (480 Mb; http://genome.wustl.edu/genomes/detail/schmidtea-mediterranea/). The difference in genome size is partly due to changed copy numbers and intron lengths of functional gene families. However, one of the major factors contributing to size variation might include transposable elements (TEs), which have differentially shrunk or expanded in genomes of these parasites (Berriman *et al*. 2009; Wang *et al*. 2011; Olson *et al*. 2012; Tsai *et al*. 2013). In the *E. granulosus* genome, the total length of retrotransposon sequences was approximately 136 kb (0·09%), which was highly contrasted to those of *S. mansoni* (75 Mb, 20%), *Schistosoma japonicum* (78·8 Mb, 19·8%) and *C. sinensis* (60·18 Mb, 11%) (Huang *et al*. 2013; Zheng *et al*. 2013). Based on these findings, it was speculated either that retrotransposons have entered a degenerative phase in the cyclophyllidean cestode genomes or that expansion of these mobile elements is tightly regulated by the donor organisms.

Retrotransposons are class I mobile genetic elements abundantly found in the majority of eukaryotic genomes (Boeke and Stoye, 1997). By replicating their progenies onto novel genomic loci via mRNA intermediates, these elements exert great influence on genome remodelling and generation of intragenomic variation, which can ultimately lead to speciation (McDonald, 1990; Long *et al*. 2000). A change in whole genome size is also highly attributable to the self-replicating activity of TEs (Kidwell, 2002). Of multiple retrotransposon families (Capy, 2005), the *Ty3/gypsy* family appeared to be the major family in the *E. granulosus* genome (Tsai *et al*. 2013). There have been no reports on retrotransposons of Taeniid cestodes, except for a recent study describing a non-autonomous terminal-repeat retrotransposon in miniature (*ta-TRIM*) and an autonomous long-terminal-repeat (LTR) retrotransposon (*lennie*) in *Echinococcus multilocularis* (Koziol *et al*. 2015). In the present study, the structural properties and evolutionary status of these representative LTR retrotransposons were investigated in *E. granulosus* with significantly shrunken genome size. The probable domestication of a *gag* gene, which originated from one of these retroelements, was also examined by analysing the degree of syntenic conservation and type of selective pressure acting on the gene.

MATERIALS AND METHODS

Identification and retrieval of Ty3/gypsy*-like retrotransposons in* E. granulosus *genome*

The draft genome of *E. granulosus* was surveyed with the Gag-Pol sequences of *CsRn1* (*C. sinensis*, AAK07487), *Ty3* (*Saccharomyces cerevisiae*, Q7LHG5), *Penelope* (*Drosophila virilis*, U49102), *Copia* (*Drosophila melanogaster*, X04456), *Bel* (*D. melanogaster*, U23420) and *DIRS1* (*Dictyostelium discoideum*, M11340) using the tBLASTn program in GeneDB (http://www.genedb.org/blast/submitblast/GeneDB_Egranulosus, assembly version 3). Pol proteins encoded in non-LTR *LINE-1* (human, L19088) and *CRE1* elements (*Crithidia fasciculate*, M33009), as well as *E. granulosus* proteins annotated as reverse transcriptases (RTs) in the GeneDB protein database, were also applied in the homology search. Hits with a BLAST score over 100 and a length larger than 100 were selected for further examination (*E*-value < 1·0e-06). Sequences displaying values under the thresholds were also considered significant if the matching segments were sequentially located close or between strong hits in both the query and the subject sequences.

Sequence analyses

Pol protein sequences encoded in *E. granulosus* retrotransposons were determined based on the tBLASTn alignments. Coding profiles of the corresponding nucleotide sequences were also examined by using the open reading frame (ORF) Finder program at National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/gorf/gorf.html). Sequence information between hits and premature stop codons(s) was treated as missing data. Amino acids of the RT domain that were tightly conserved in diverse retroelements (Xiong and Eickbush, 1990), were extracted by homology-based matching of these polypeptides against the *CsRn1* Pol sequence. These amino acid sequences were aligned with those of other LTR retrotransposons representing each of the distinct clades of the *Ty3/gypsy* family (Malik and Eickbush, 1999; Bae *et al*. 2001, 2008; DeMarco *et al*. 2004) using the ClustalX program (Thompson *et al*. 1997). LTR retrotransposons of parasitic platyhelminths including *S. mansoni* were also included in the structural comparison. The alignment was used in a phylogenetic analysis using MEGA (version 6.0; Tamura *et al*. 2013). The analytical options were as follows: the Jones–Taylor–Thornton (JTT) model for amino acid substitution, estimation of invariant site proportion from the input data, and rate heterogeneity with eight gamma category (the gamma parameter was estimated from the dataset by MEGA). The gaps introduced during alignment and the missing data mentioned above were deleted in a pairwise manner, and the neighbour-joining algorithm was adapted for tree construction. The statistical significance of each branching node was estimated by a bootstrap analysis of 1000 replicates. The tree was

https://doi.org/10.1017/S0031182016001499 Published online by Cambridge University Press

displayed with the TreeView program (Page, 1996). A portion of the retrieved elements were employed in the phylogenetic analysis using sequence information supplemented with RNase H (RH) and integrase (IN) domains, if possible, to increase the analytical resolution (Malik and Eickbush, 1999).

Selective pressure acting on each codon of the *gag* gene was investigated using the maximum-likelihood algorithm of MEGA by evaluating the differences between non-synonymous (dN) and synonymous substitution (dS), dN – dS (0, neutral; -, negative selection; +, positive selection). The maximum-likelihood tree (bootstrap value 1000) used in the analysis was also constructed by the program based on the nucleotide alignment of cestode *gag* sequences.

Determination of full-length retrotransposons

Nucleotide sequences (approximately 20 kb) encompassing the segmental *pol*-like genes were extracted from each of the scaffolds and used as queries in the BLASTn searches of the draft *E. granulosus* genome. Multiple scaffolds retrieved by a single query were compared with one another and then, a common genetic element was isolated from them by BL2Seq at NCBI, as described previously (Bae *et al*. 2008). Terminal repeats flanking a protein-encoding internal region were determined similarly by the program. Structural integrity of mobile elements was further verified by recognizing duplicated target sequences from the direct upstream and downstream regions of the 5′- and 3′-LTRs. The consensus nucleotide sequence of full-unit retrotransposon was determined by comparing multiple copy sequences using the GeneDoc program and its coding profile was predicted using ORF Finder. The consensus sequence was also applied in the retrieval and characterization of homologous elements from the *E. multilocularis* and *Taenia solium* genomes (BLASTn). The full-unit retrotransposons were used in the phylogenetic analysis as described above and the construction of distance matrix based on the nucleotide sequence homology with the MEGA program.

RESULTS

Retrieval of Ty3/gypsy*-like retrotransposons in* E. granulosus

The draft genome of *E. granulosus* in GeneDB was surveyed with Pol sequences of *CsRn1* and *Ty3* to isolate the *Ty3/gypsy*-like LTR retrotransposons (tBLASTn algorithm, *E*-value < 1e-06). A total of 46 homologous sequences were isolated from 25 scaffolds during this examination (Fig. 1 and Supplementary table 1). The lengths and positions of the matched regions, and the statistical significance levels varied among hits, while a significant fraction of these sequences contained a region for RT (34 elements, 73·9%). ORFs encoded in all retrieved sequences were heavily corrupted by multiple premature stop codons. Otherwise, they displayed discontinuous matching patterns against queries with short gaps and/or changes in reading frames between hits. The average length of the retrieved sequences was 1279·1 ± 728·63 bp (including gapped regions; 420–2492 bp) or 1078·4 ± 521·58 bp (excluding gapped regions; 420–2130 bp) (Supplementary table 1). Searches using Pol proteins in *Xena*- (*Penelope*), *Copia*- (*Copia*) and *Bel*-family (*Bel*) members did not show any significant match, while that of *DIRS1* (*DIRS* family) detected multiple sequences that overlapped to the results with *CsRn1* and *Ty3* (data not shown). No homologous sequence was also retrievable with the Pol protein of *CRE1*; however, human *LINE-1* detected a single scaffold sequence (pathogen_EgG_scaffold_0325, 5397 bp; identities >95% at the amino acid and nucleotide levels), which might have originated from contaminating human DNA.

Six entries named RT were also detected in the GeneDB database of *E. granulosus*. The coding sequences of the corresponding genes were intervened by one (EgrG_000078300, EgrG_000332600, EgrG_002017100 and EgrG_000438600) or seven (EgrG_000864400 and EgrG_000487300) introns. When considering their genomic loci, some of these *rt* genes were overlapped with those in Fig. 1 (EgrG_000078300 = EgG-s-0395, EgrG_000332600 = EgG-s-0031, EgrG_002017100 = EgG-s-0001-F and EgrG_000864400 = EgG-s-0017-A). EgrG_000438600 was matched to an endonuclease (EN)-RT protein of *C. sinensis* (GAA53638, *E* value = 7e-30), which suggested that the protein is highly related to a non-LTR retrotransposon. Inconsistent with its name, EgrG_000487300 did not contain any Pol-related domain, but possessed a YTH domain of YTH family proteins that act in the removal of meiosis-specific gene transcripts expressed in mitotic cells (Harigaya *et al*. 2006).

Classification of E. granulosus *retrotransposons*

The amino acid sequences of 34 RT-containing hits were predicted largely based on their tBLASTn alignments. The RT sequences (153·4 ± 32·02 amino acids in lengths, Supplementary Fig. 1) were combined with those of various retrotransposons representing distinct clades/families of LTR retrotransposons. LTR retrotransposons isolated in parasitic trematodes were also included in the analysis. After alignment, the sequence information was used in a phylogenetic analysis. As shown in Fig. 2, a neighbour-joining tree tightly separated members of *Mag, Mdg1, Gypsy* and, *Athila* clades, although the relationships of other clade members

https://doi.org/10.1017/S0031182016001499 Published online by Cambridge University Press

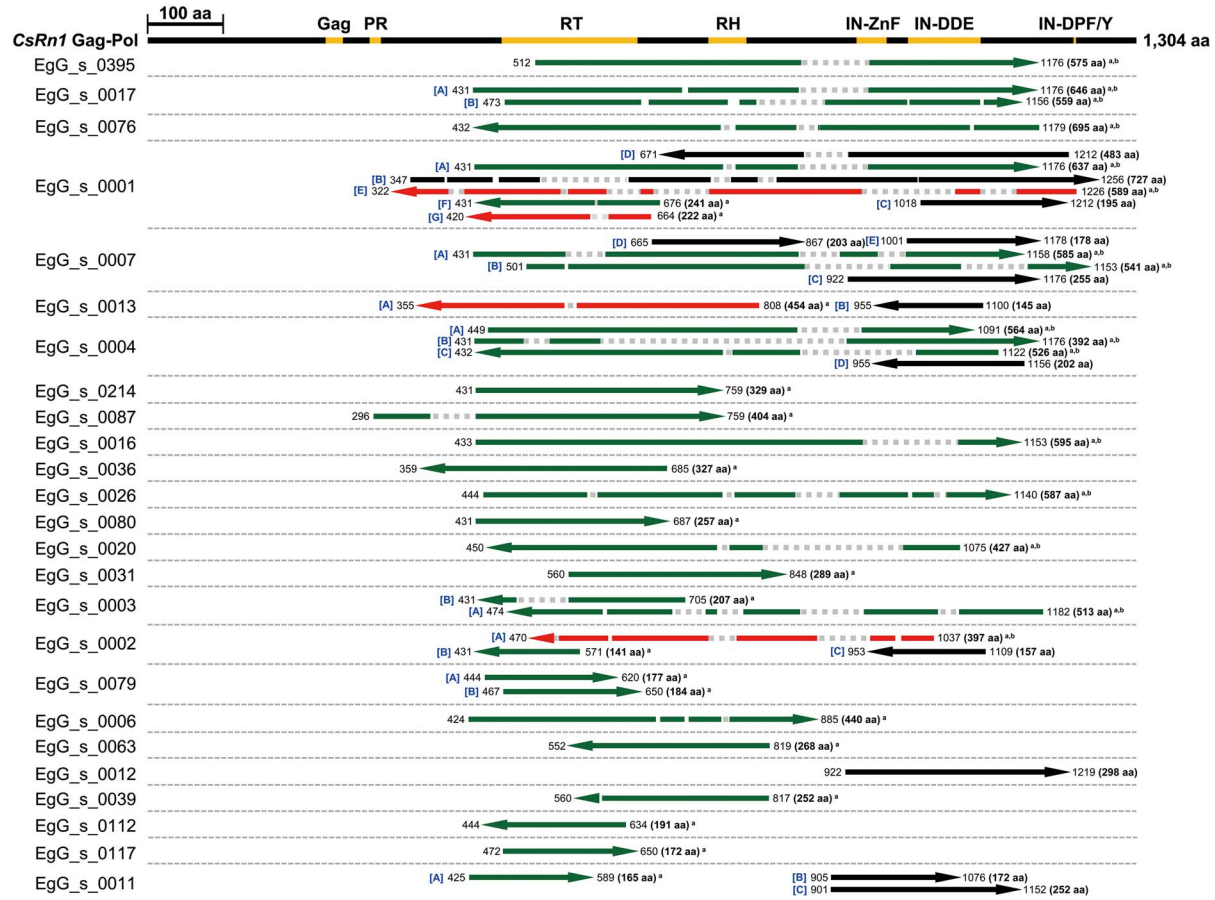

Fig. 1. Similarity patterns in *Echinococcus granulosus* LTR retrotransposons. The predicted amino acid sequences of partial retrotransposons retrieved from the draft genome of *E. granulosus* were mapped against the homologous regions of the *CsRn1* Gag-Pol protein, based on tBLASTn results. The identity of each element was distinguished by the entry number of the respective scaffold and when needed, by a subsequent English alphabet (e.g. EgG-s-00017-A or EgG-s-0017-B, where s abbreviates scaffold). Arrowheads direct the relative orientations of retrotransposons in the scaffolds and the dotted-arrow regions mark the gaps between hits during the BLAST searches. The starting and end points of matched sequences in *CsRn1* Gag-Pol are provided at both ends of each arrow. Numerals in parentheses indicate the length of the matched sequences. PR, protease; RT, reverse transcriptase; RH, RNase H; IN-ZnF, Zn-finger motif of integrase; IN-DDE, DDE motif of IN; IN-DPF/Y, DPF/Y motif of IN. [a]Elements used in the RT-based phylogenetic analysis (Fig. 2) and [b]those used in RT–RH–IN-based examination (Fig. 3). Red and green arrows indicate the *CsRn1*- and *Saci-2*-like elements, respectively.

including those of *Ty3* and *Mag* were somewhat ambiguous, which may be due to limited information in the RT-based phylogeny (Malik and Eickbush, 1999). Interestingly, distribution of *E. granulosus* retrotransposons appeared to be restricted only within *CsRn1* (four elements; red-coloured in Fig. 1) and *Saci-2* (30 elements; green coloured in Fig. 1) clades. Branch nods connecting each of the clade members were well-supported statistically by bootstrap analysis (bootstrap values 96 and 54, respectively).

Where possible, amino acids were similarly isolated from the *rh* and *in* domains of *E. granulosus* retrotransposons, as well as the representative *Ty3/gypsy* elements. The sequences were combined with their respective RT sequences and used in a phylogenetic analysis. Members of the previously well-defined clades were strongly co-clustered in the neighbour-joining tree (Fig. 3). As in the RT-based tree, the 16 *E. granulosus* elements selected in this analysis were segregated into *CsRn1* (two elements) and *Saci-2* (14 elements) clades (bootstrap values 100). Another tree constructed with the minimum evolution algorithm showed a topology similar to that of the neighbour-joining tree (data not shown).

Analysis of full-unit retrotransposons and their mobile potential

Full-length retrotransposons encompassing the partial *pol* sequences shown in Fig. 1 were surveyed in the *E. granulosus* genome. A majority of these elements were heavily truncated by segmental deletions and thus, their integral structures with two flanking LTRs were not readily defined. Nonetheless, a

https://doi.org/10.1017/S0031182016001499 Published online by Cambridge University Press

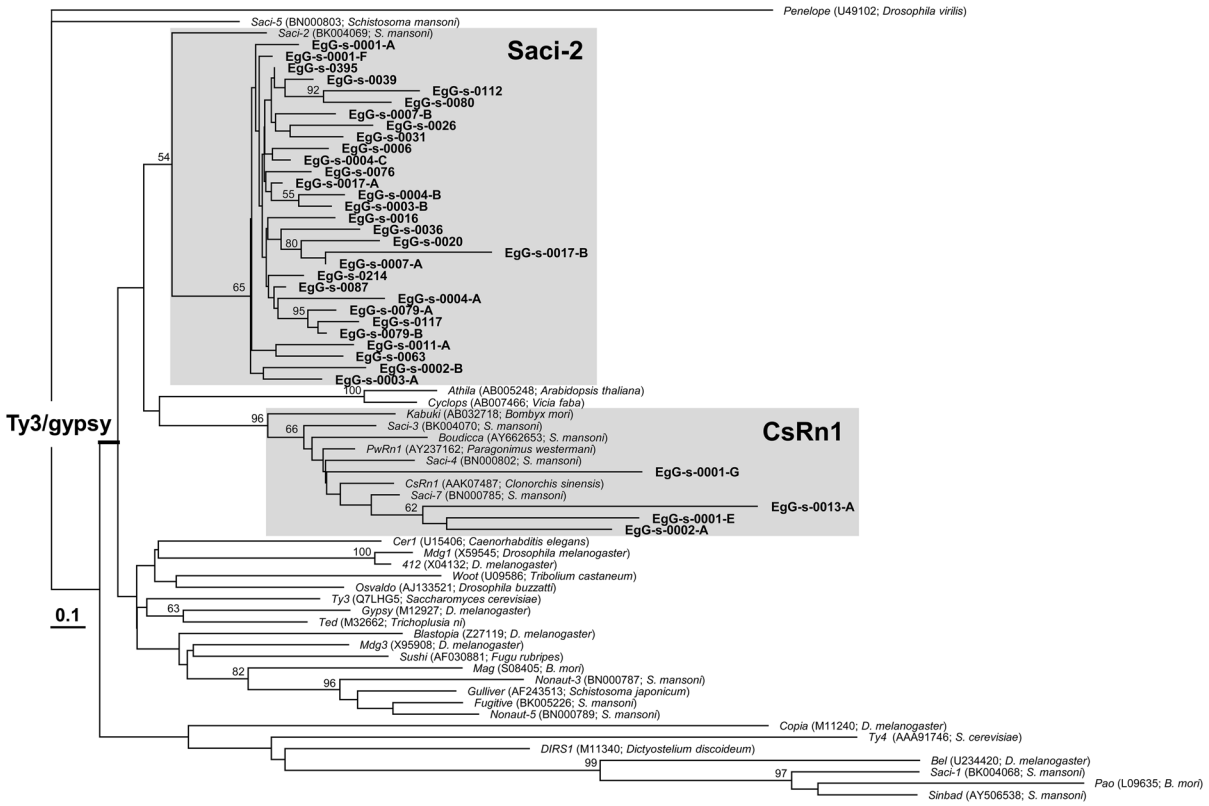

Fig. 2. Phylogenetic positions of *Echinococcus granulosus* retrotransposons against the representative family members. The neighbour-joining tree was constructed with amino acid sequences of RT using MEGA. The identities of *E. granulosus* retrotransposons in bold letters were marked with the respective entry numbers of scaffold sequences occasionally followed by an English alphabet. Numerals at the major branching nodes indicate their percentages of appearance in 1000 bootstrap replicates.

degenerate copy of whole-unit retrotransposon encompassing EgG-s-0017-A could be isolated from a genomic scaffold sequence (pathogen_EgG_scaffold_0017, nucleotide positions 945 896–950 285). The element comprised 4390 bp including 5′- and 3′-LTRs (236 and 235 bp, respectively; 93% similarity), and was bound by a slightly modified 5-bp target site duplication (TSD, CTAGT/CTATT). A consensus 4427-bp sequence was obtained using homologous genomic segments from GenBank (40 entries, 2340 ± 892·3 bp in length) and other full-unit copies, which were similarly determined from the *E. granulosus* genome (Supplementary table 2). The theoretical sequence contained an ORF for a 1327-amino acid Pol protein (Fig. 4A). Homologous elements were also identified in the genomes of *E. multilocularis* and *T. solium* (Fig. 4A and Supplementary table 2). The *E. multilocularis* element appeared to be identical to *lennie* that was previously described by Koziol *et al*. (2015) and thus, the *E. granulosus* and *T. solium* elements were named *Eg_lennie* and *Ts_lennie*, respectively. These taeniid elements were bounded by direct repeats of 4- or 5-base nucleotides known as TSDs and conserved structural/functional motifs including LTRs that initiated as TG and terminated as CA, primer-binding site (PBS) for ^{Leu}tRNA and poly-purine tract (PPT), although none of them retained its coding sequence (Fig. 4B; see also Koziol *et al*. 2015). The degrees of divergence in LTRs and internal region sequences of *lennie* copies were similar in *E. granulosus* (0·574 ± 0·062 and 0·232 ± 0·009) and *T. solium* (0·524 ± 0·032 and 0·261 ± 0·005), while the values were much lower in *E. multilocularis* (0·261 ± 0·005 and 0·076 ± 0·002) (P values < 0·001 in t test of pairwise distance matrices; Fig. 4C and Supplementary table 3). The clustering pattern of *lennie* copies in Fig. 4C further suggested that the element had expanded independently in each of the taeniid genomes. Meanwhile, a single copy in *E. granulosus* (*Eg_lennie-1*) was tightly co-clustered with the *Em_lennie* copies, especially with *Em_lennie-3*, which demonstrated that *lennie* had been inserted in the corresponding locus before divergence of *E. granulosus* and *E. multilocularis*. The nucleotide sequences flanking *Eg_lennie-1* and *Em_lennie-3* showed similarity values larger than 85% (data not shown).

BLAST searches of the *E. granulosus* expressed sequence tag (EST) database in GenBank with the nucleotide sequence of Eg_lennie_{cons} recognized several ESTs including JZ787251·1 (95%), JZ791468·1 (93%) and JZ779873·1 (93%). Although

https://doi.org/10.1017/S0031182016001499 Published online by Cambridge University Press

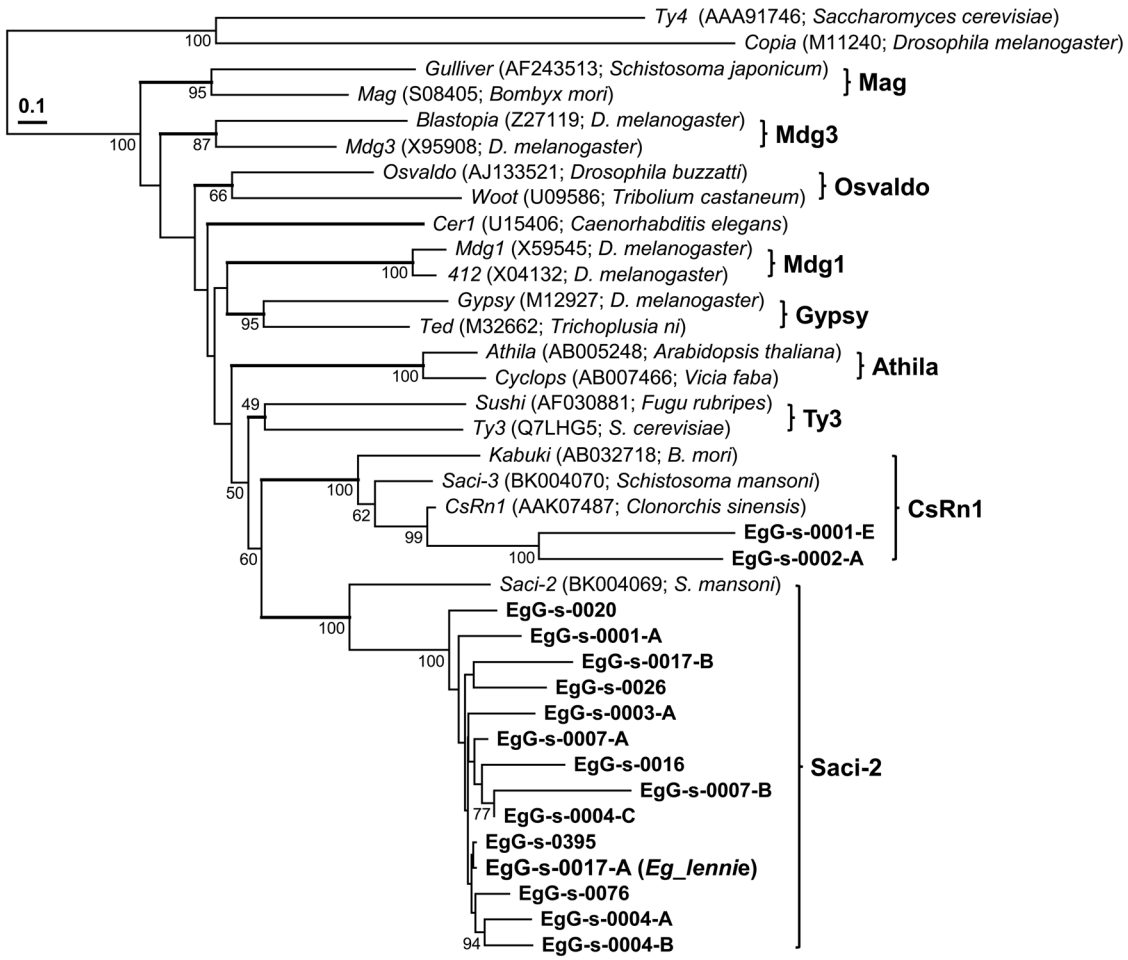

Fig. 3. Pol-based phylogeny of *Echinococcus granulosus* LTR retrotransposons. The amino acid sequences of major Pol domains (RT, RH and IN) were summed to construct the neighbour-joining tree. The tree was rooted with *Copia* and *Ty4*. The identities of *E. granulosus* retrotransposons in bold letters were distinguished with the scaffold entry numbers occasionally followed by an English alphabet. Numerals at the major branching nodes indicate their percentages of appearance in 1000 bootstrap replicates. The major clades of the *Ty3/gypsy* family are marked in the tree.

these ESTs were matched to various regions of *Eg_lennie*$_{cons}$ (arrows in Fig. 4A), several in-frame stop codons were detected in all sequences. The Pol sequences of *Eg_lennie*$_{cons}$, *CsRn1* and *Ty3* (tBLASTn algorithm) showed results similar to that obtained using the *Eg_lennie*$_{cons}$ nucleotide sequences. Screening of *T. solium* EST databases using the *Ts_lennie*$_{cons}$ sequence showed a result similar to that of *Eg_lennie*$_{cons}$. ESTs homologous to *Em_lennie*$_{cons}$ could not be retrieved from *E. multilocularis* ESTs (Fig. 4A).

The amino acid sequences of *lennie* elements showed highest similarities to Gag-Pol of *Saci-2*-like elements isolated from various invertebrates including *Schistosoma* and *Trichinella* species (*E*-values < 2e-177; Fig. 5A). Homologues were also detected in teleost fish such as *Danio rerio* and *Larimichthys crocea*. Gag proteins encoded in these *Saci-2*-like elements conserved the conventional nucleocapsid motif $CX_2CX_4HX_4C$, except for those detected in schistosomes. The motif was slightly modified as CX_2CX_9CXH in the schistosome elements (see also DeMarco *et al*. 2004). However, none of the nucleocapsid motifs known in the LTR retrotransposons could be detected in the taeniid *lennie* elements (Fig. 5B).

Characterization of a gag *gene endogenized in the* E. granulosus *genome*

The Pol-based phylogenetic trees demonstrated that *E. granulosus* LTR elements belong to either the *CsRn1* or *Saci-2* clade (Figs 2 and 3). Since all the partial sequences did not cover gag regions except for *lennie* (Figs 1 and 5), it could not be addressed whether these retrotransposons preserved the unique CHCC (Bae *et al*. 2001) and CCCH (DeMarco *et al*. 2004) Gag motifs, except for *lennie*. A BLAST search with the *CsRn1* Gag isolated only a single scaffold sequence of the *E. granulosus* genome (pathogen_EgG_scaffold_0005, *E*-value = 3·4e-07), but the scaffolds shown in Supplementary table 1 were not retrievable by the Gag sequence. The ORF Finder program predicted

https://doi.org/10.1017/S0031182016001499 Published online by Cambridge University Press

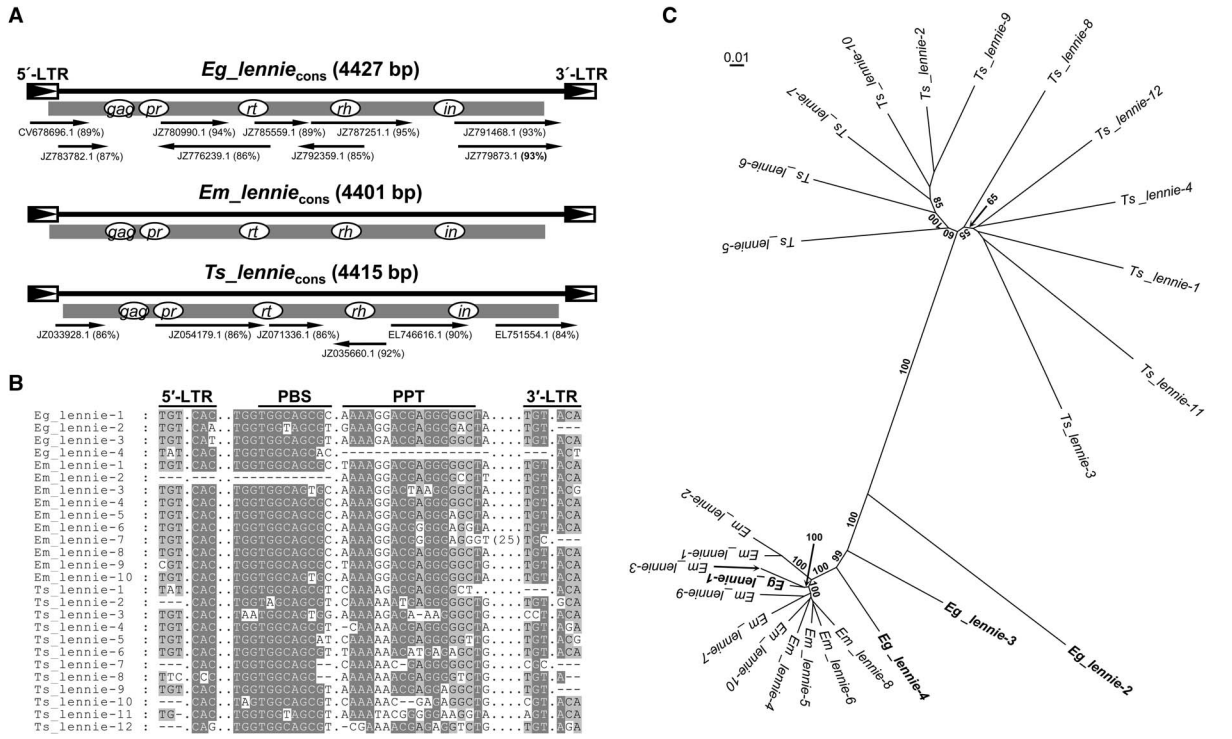

Fig. 4. Characterization of *Echinococcus granulosus lennie*. (A) Overall structure of the theoretically determined consensus sequence of *Eg_lennie* (*Eg_lennie*$_{cons}$). Boxes containing black arrowheads represent the flanking LTRs. The grey bar indicates an ORF containing *gag*, protease (*pr*), reverse transcriptase (*rt*), RNase H (*rh*) and integrase (*in*) regions. Arrows at the bottom mark the positions and orientations of the expressed sequences tags matched to *Eg_lennie*$_{cons}$. Numerals in parentheses are their similarity values. Structures of orthologous elements in *Echinococcus multilocularis* (*Em_lennie*) and *Taenia solium* (*Ts_lennie*) were also presented. (B) Structural/functional motifs of *lennie* elements. After alignment of full-unit *lennie* copies, the boundary nucleotides of 5′- and 3′-LTRs, PBS and PPT were defined. Dashes were introduced during the sequence alignment to increase similarity values and dots were inserted to separate the distinct motifs. (C) Phylogenetic analysis of the full-length *lennie* elements. The analysis was performed with MEGA based on the alignment of internal region sequences of *lennie* elements. The tree was constructed using the neighbour-joining algorithm. Numerals at branch nodes indicate their percentage of appearances in 1000 bootstrap replicates.

an ORF of 233 codons in the matched region (nucleotide positions 5 559 972–5 560 670), of which the transcribed products were detected in the GenBank EST database (JZ785456·1, JZ784648·1 and JZ792189·1). A BLAST search using the theoretical protein sequence provided a result identical to that with *CsRn1* Gag. Meanwhile, no significant match was found in the genome database using *Saci-2* Gag as a query.

Orthologues of the *E. granulosus* solo Gag protein, which had been previously annotated in GenBank (CDS20277·1), were isolated in other cestode parasites including *E. multilocularis* (CDI98622·1), *H. microstoma* (CDS32873·1) and *T. solium* (GeneDB, TsM_000811700·1..pep). These proteins exhibited identity values of 19–95% to one another or to the *CsRn1*-clade Gags. All proteins, except for the *H. microstoma* Gag, contained the tightly conserved CHCC nucleocapsid motif in their C-terminal regions (arrowheads in Fig. 6). To better understand the evolutionary origin of the cestode solo *gag*s, the syntenic relationship of neighbouring genes was examined in the *gag*-containing genomic scaffolds. As shown in Fig. 7, six genes (glutamyl tRNA synthase, D111 G patch, multidrug resistance associated protein 4, phospholipase D, phospholipase D1 and microtubule-associated protein xmap215) were clustered with the *gag* gene within a 100 kb region on the pathogen_EgG_2_scaffold_0005. This gene cluster was also detected in the genomic scaffolds of *E. multilocularis* (pathogen_EmW_scaffold_04), *T. solium* (pathogen_TSM_contig_00051, 236 kb) and *H. microstoma* (pathogen_HYM_scaffold_0004), although the synteny covered approximately 150 kb in *H. microstoma* and phospholipase D1 was not detected in the *T. solium* sequence. The relative orientations of these syntenic genes were also fully conserved in the cestodes. Of the 231 consensus codons in the cestode *gag* genes, 144 (62·34%) codons showed $dN - dS < 0$ indicating purifying selection, while 35 codons (15·15%) diverged neutrally without selection pressure (Supplementary Fig. 2).

DISCUSSION

Expansion and deletion of retrotransposon-related sequences has been considered one of the major

https://doi.org/10.1017/S0031182016001499 Published online by Cambridge University Press

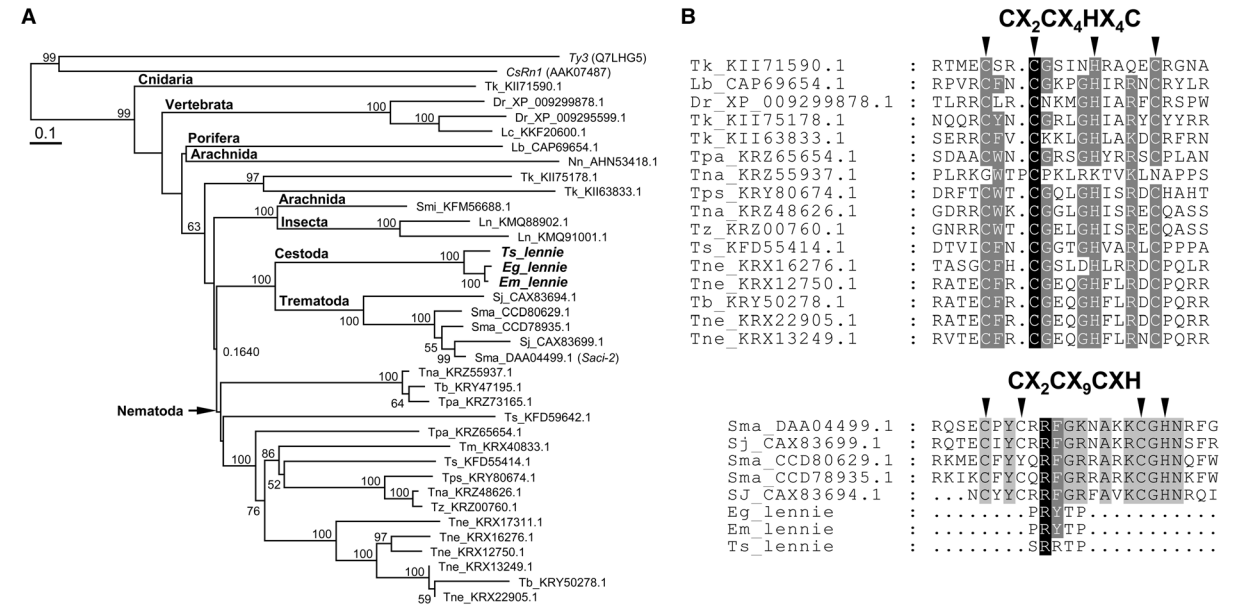

Fig. 5. Phylogenetic analysis of *lennie* elements detected in *Echinococcus granulosus* (*Eg_lennie*), *Echinococcus multilocularis* (*Em_lennie*) and *Taenia solium* (*Ts_lennie*). (A) The neighbour-joining tree of *lennie* and other *Saci-2*-like elements was constructed using the sum of amino acids comprising RT, RH, and IN. *Ty3* of *Saccharomyces cerevisiae* and *CsRn1* of *Clonorchis sinensis* were included in the analysis as outgroup members. The identities of elements were given as species names of donor organisms and their GenBank accession numbers. Tk, *Thelohanellus kitauei*; Lb, *Lubomirskia baicalensis*; Nn, *Nuttalliella namaqua*; Ts, *Trichuris suis*; Tpa, *Trichinella papuae*; Tne, *Trichinella nelson*; Tb, *Trichinella britovi*; Tm, *Trichinella murrelli*; Tps, *Trichinella pseudospiralis*; Tna, *Trichinella native*; Tz, *Trichinella zimbabwensis*; Smi, *Stegodyphus mimosarum*; Ln, *Lasius niger*; Sj, *Schistosoma japonicum*; Sma, *Schistosoma mansoni*; Dr, *Danio rerio*; Lc, *Larimichthys crocea*. (B) Nucleocapsid motifs conserved in the Gag proteins of *Saci-2*-like elements. Dots were introduced in the alignment to increase similarity values. Arrowheads indicate the conventional CCHC or unique CCCH signatures.

factors that change the size of eukaryotic genomes. Cyclophyllidean cestodes possess genomes with much smaller sizes compared to closely related trematode or more basal turbellarian species. In an effort to address the shrinkage process in cestode genomes, structural and evolutionary features of LTR retrotransposons were investigated by surveying the *E. graulosus* genome. The *E. granulosus* retrotransposons examined belonged to *CsRn1* or *Saci-2* clades of the *Ty3/gypsy* family. The structural integrity and coding potential of these elements were heavily corrupted by sporadic base substitutions and segmental deletions, even though they seemed to be the major TEs occupying the parasite's genome. Apart from the degenerate and inactive copies, a solo active *gag* gene homologous to those of *CsRn1* clade members was also detected in the genomes of *E. granulosus* and other taeniid cestodes. The gene was likely to be 'endogenized' in the genomes of cestode parasites including *E. granulosus*, to play a role(s) beneficial to host organisms.

Unlike prokaryotes, eukaryotic organisms have varied in genome size over 6000-fold (approximately 20–130 000 Mb; Animal Genome Size Database, http://www.genomesize.com), with a distribution bias skewed toward smaller values <5000 Mb (Oliver *et al*. 2007; Dufresne and Jeffery, 2011). Genome size tends to increase in proportion to the size of cell/nucleus and duration of cell division, both of which negatively influence the developmental rate of donor organisms (reviewed in Dufresne and Jeffery, 2011). Therefore, rapidly growing or invasive organisms have evolved genomes with a smaller size (Lavergne *et al*. 2010; Dufresne and Jeffery, 2011). The overall genome sizes were strikingly reduced in cyclophyllidean tapeworms compared to their evolutionary neighbours mainly due to decreases in intron lengths and retrotransposon ratios (Tsai *et al*. 2013; Zheng *et al*. 2013). In *E. granulosus*, the average intron length and retrotransposon ratio were 726 bp and 0·09%, respectively, whereas they were 1692 bp and 20·0% in *S. mansoni* (Zheng *et al*. 2013). However, the average number of introns in the coding genes was similar in both species (approximately 5). Considering that TEs play a central role in expansion of intron size (Jiang and Goertzen, 2011), it could be suggested that withdrawal of TEs from these cestode genomes is the most potent mechanism, if not the only one, associated with the greatly shrunken genome sizes.

Petrov *et al*. (1996) have proposed that deletion of nucleotides occurring in dead-on-arrival (DOA) copies of retrotransposons is essential to maintain or generate a smaller genome. In *E. granulosus*, *Ty3/gypsy*-family retrotransposons appeared to be

https://doi.org/10.1017/S0031182016001499 Published online by Cambridge University Press

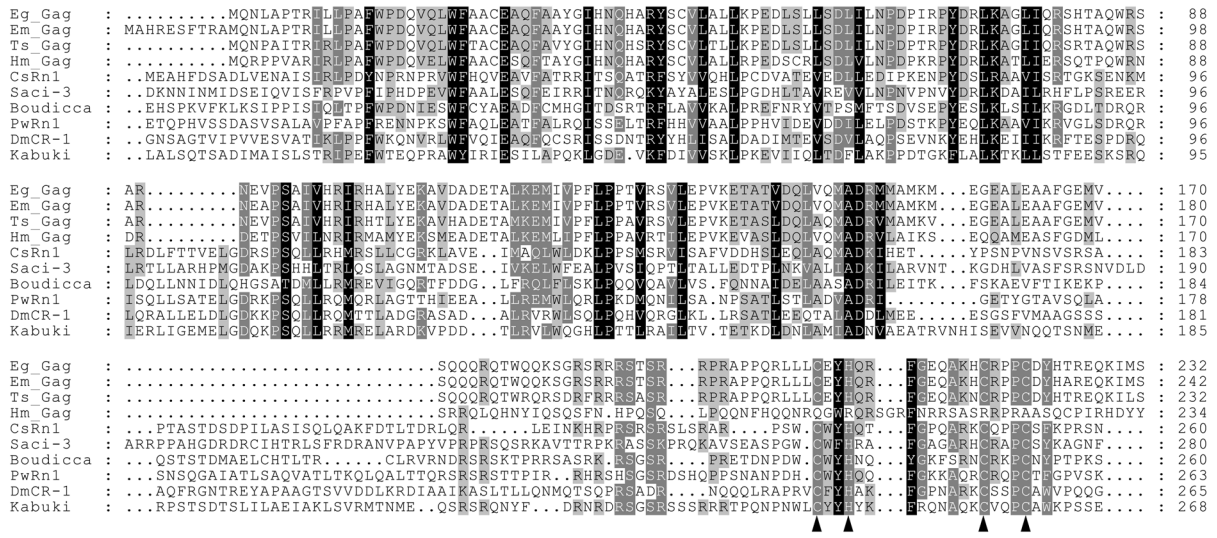

Fig. 6. Comparison of primary structures between *Echinococcus granulosus* solo Gag and its orthologues. The amino acid sequences of cestode Gag orthologues in *E. granulosus* (Eg_Gag, CDS20277·1), *Echinococcus multilocularis* (Em_Gag, CDI98622·1), *Taenia solium* (Ts_Gag, TsM_000811700·1..pep) and *Hymenolepis microstoma* (Hm_Gag, CDS32873·1) were aligned with those of *CsRn1*-like LTR retrotransposons. Dots represent the gaps introduced in the alignment to increase similarity values. Arrowheads indicate the unique CHCC signature conserved in the nucleocapsid domain of *CsRn1*-like elements. The retrotransposons used in the comparison are as follows: *CsRn1* of *Clonorchis sinensis* (AAK07487); *Saci-3* of *Schistosoma mansoni* (BK004070); *Boudicca* of *S. mansoni* (DAA04496); *PwRn1* of *Paragonimus westermani* (AY237162); *DmCR-1* of *Drosophila melanogaster* (AE003787); *Kabuki* of *Bombyx mori* (AB032718).

heavily corrupted, especially by segmental DNA loss. Thus, it was difficult to isolate their full-unit sequences (only that of *Eg_lennie* was successfully defined in this study). Given the sizes of trematode (five species, 911·5 ± 396·02 Mb) and turbellarian (61 species, 2018·3 ± 3092·80 Mb) genomes in the Animal Genome Size Database (release 2·0), cestodes might also have had genomes with sizes similar to those of their neighbours in an early evolutionary period, as is the case of *Spirometra erinaceieuropaei* (1260 Mb; Bennett *et al*. 2014). Thereafter, the genomes were likely to have shrunken gradually, which followed the inactivation of retrotransposons and/or accumulation of their DOA copies. The genetic/genomic damages and metabolic costs generated by the TE explosion are known to invoke natural selection against proliferating TEs (Charlesworth and Langley, 1989). The approved mechanisms that effectively repress TEs include CpG island methylation, heterochromatin formation, repeat-induced point mutation and action of the *piwi* gene (Blumenstiel, 2011; Dufresne and Jeffery, 2011). To date, there have been no reports on the molecular machinery regulating retrotransposons in parasitic platyhelminths including cestodes, although a neodermatan-specific argonaute gene has been suggested to play a role similar to that of the *piwi* gene in *S. mansoni* (Collins *et al*. 2013; Wang *et al*. 2013). Nevertheless, it could be suggested that smaller genomes in parasitic cestodes provide advantages to adapt to or overcome the physicochemical barriers in host environments, such as accelerated growth/development rate and reduced metabolic costs. As an example, small genome size is likely to be highly advantageous in reducing the metabolic cost of DNA replication in *E. granulosus*, which lacks enzymes involved in *de novo* synthesis of purine and pyrimidine bases (Zheng *et al*. 2013).

A phylogenetic analysis of multiple *lennie* copies demonstrated that these copies had expanded independently in each of their host genomes, similar to *ta-TRIM* (Fig. 4C; Koziol *et al*. 2015). The genomic positions of only a single pair (*Eg_lennie-1* and *Em_lennie-3*) were found to be orthologous in *E. granulosus* and *E. multilocularis* (>85% identities in their flanking nucleotide sequences), suggesting its insertion before divergence of the host species. Like the genomic copies, the ESTs of *E. granulosus* and *T. solium*, which were homologous to *lennie*, contained multiple premature stop codons. Therefore, *Eg_lennie* and *Ts_lennie* might also have lost their mobile potential, similar to *Em_lennie* (Koziol *et al*. 2015). Considering the lowest distance values, *lennie* seemed to have amplified most recently in *E. multilocularis* (Supplementary table 3). Investigations on the characterization of *lennie* homologs in diphyllobothroid tapeworms including *S. erinaceieuropaei* and role(s) of the mRNAs transcribed from the DOA copies of *lennie* might be helpful in understanding the genomic implication of *lennie* in association with the evolution of cestode parasites.

LTR retrotransposons identified in *E. granulosus* were classified exclusively as *Saci-2*- or *CsRn1*-clade members (Figs 2 and 3). *Saci-2* described in

https://doi.org/10.1017/S0031182016001499 Published online by Cambridge University Press

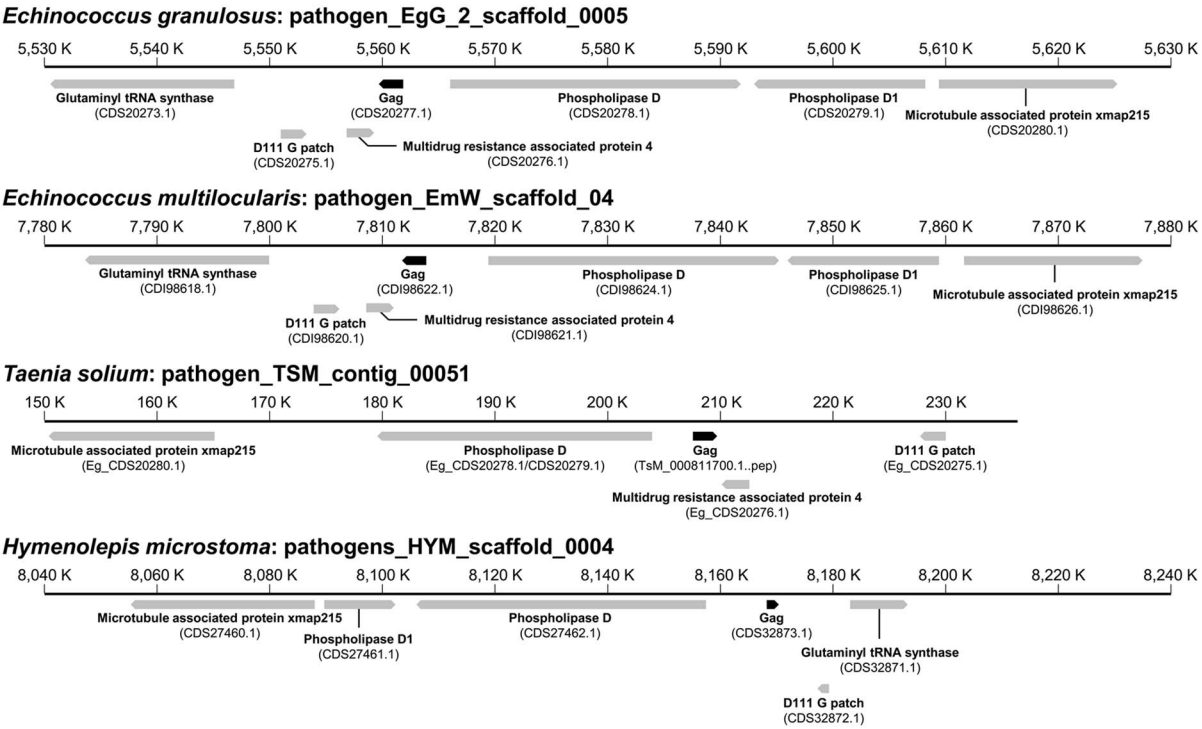

Fig. 7. Comparative structures of cestode genomic loci encompassing orthologues of the solo *gag* gene identified in *Echinococcus granulosus*. The position of the *gag* gene (Black) was mapped on each genomic scaffold in the 150 kb (*E. granulosus* and *Echinococcus multilicularis*) or 200 kb (*Hymenolepis mircostoma*) scale. The protein products of the neighbouring genes are indicated in grey with their GenBank accession numbers (in case of *Taenia solium* genes, identity numbers in GeneDB are used). The relative transcriptional directions are represented by arrowheads.

S. mansoni is known to encode Gag with a unique CCCH nucleocapsid motif (DeMarco *et al*. 2004). A major fraction of *E. granulosus* retrotransposons formed a novel clade in the phylogeny of the *Ty3/gypsy* family together with *Saci-2*. The *Saci-2* clade further included elements identified in a broad range of donor organisms from cnidarians to vertebrates (Fig. 5A). It was apparent that the characteristic CCCH Gag motif specifically emerged in *Schistosoma* retrotransposons, although its biological implications remain unclear. *Saci-2*-like elements detected in other taxa conserved the canonical CCHC motif in their corresponding regions. Interestingly, nucleotide sequences of taeniid $lennie_{cons}$ encoding the probable C-terminus of Gag were highly divergent and therefore, none of the Gag motifs was defined (Fig. 5B). It was also apparent that *lennie* with the degenerate *gag* sequence had independently multiplied in each taeniid genome (Fig. 4C). Gag function is critical for the assembly of viral-like particles (VLPs) during retrotransposition (Boeke and Stoye, 1997). Therefore, it seems likely that the indispensable activity is complemented in *trans* by a Gag or Gag-like protein expressed by another retrotransposon or chromosomal gene (Schulman, 2012). Alternatively, the degenerate form of *lennie* Gag is possibly engaged in VLP formation, even if the efficiency is reduced.

The occasional domestication of TE genes including retroviruses has been demonstrated with the *env* gene of endogenous retroviruses (ERVs) found in eutherian mammals, as well as the universal meiotic recombinase, DNA repair protein Rad51 and telomerase RT. Fixation of *gag*s originating from LTR retrotransposons/retroviruses was also reported in mammals (Volff, 2006). Unlike parasitic TE sequences, the captured gene is subject to strong selective pressure exerted by the host due to its beneficial effect(s). In eutherian mammals, the Env protein plays a key role during placental development (Sinzelle *et al*. 2009; Haig, 2012). Gag proteins are also involved in diverse biological processes such as the control of cell proliferation and apoptosis, early embryonic angiogenesis, and restriction of viral replication (reviewed in Volff, 2006). The ORFs of *Ty3/gypsy*-like LTR retrotransposons in the *E. granulosus* genome were heavily corrupted by insertions/deletions and/or base substitutions, which demonstrated loss of mobile potential and thus, underwent neutral evolution. However, a *gag* gene with the CHCC nucleocapsid motif, which might have originated from a *CsRn1*-like retrotransposon (Bae *et al*. 2001), was found to preserve its coding potential for 232-aa polypeptide. Orthologous genes were further detected in the tightly conserved synteny of other cestode parasites (Figs 6 and 7). The mRNA sequences of the *gag* gene were readily detected in *E. granulosus* (fragments per kilobase of transcript per million mapped reads [FPKM] 6 in protoscolex),

https://doi.org/10.1017/S0031182016001499 Published online by Cambridge University Press

E. multilocularis (FPKM 1 and 4 in pre-gravid and gravid, respectively) and *H. microstoma* (FPKM 11 in adult) during RNA-Seq data analysis (Tsai *et al.* 2013). Collectively, these data suggest that the *gag* gene has been captured as a host gene in cestodes.

Retroviral Gag is the primary structural protein functioning in viral assembly. Though being highly diversified among viral groups, Gag is separated into three functional polypeptides, named matrix (MA), capsid (CA) and nucleocapsid (NC) proteins, by proteolytic cleavage. In retroviruses, these segmental proteins play specific roles in membrane targeting of Gag polyproteins for capsid assembly, formation of the hydrophobic core of the virion (Gag–Gag interaction) and RNA packaging, respectively (Boeke and Stoye, 1997; Maldonado *et al.* 2014). Two zinc-finger motifs conserved in NC are essential for the recognition of and/or binding to a specific region of the viral RNA genome (Gorelick *et al.* 1988; Muriaux *et al.* 2004). NC is also known to be involved in nucleolar trafficking/accumulation of viral Gag (Lochmann *et al.* 2013). Information of Gag function(s) encoded in LTR retrotransposons, especially on those with the un-conventional nucleocapsid motifs, is limited. Therefore, it is not easy to predict the physiological implication of the retrotransposon-derived *gag* gene in *E. granulosus*. Biochemical characterization of the Gag protein including its spatiotemporal expression pattern and DNA/RNA binding activity would provide important clues elucidating its role(s) in relation to parasitic cestode-specific function.

The total length of retrotransposon-related sequences analysed in this study was approximately 58·84 kb, which comprised 43·3% of the total retrotransposons identified in the *E. granulosus* genome (136 kb; Zheng *et al.* 2013). Since Pol, especially RT, of retrotransposons can be used to detect one another in homology-based gene fishing, the present data would represent the real aspects of autonomous *E. granulosus* LTR retrotransposons preserving an analysable length. In general, eukaryotic genome with a smaller size harbours less copies but more diverse clades of LTR retrotransposons and *vice versa* (Volff *et al.* 2003). *E. granulosus*, which has evolved a smaller genome than those of its phylogenetic neighbours, contained numerous copies of degenerate LTR retrotransposons belonging to only two clades (*Saci-2* and *CsRn1*) of the *Ty3/gypsy* family. Therefore, parasitic cestodes might have had a larger genome, where less diverse LTR retrotransposons had over-expanded, during an early evolutionary stage and then, the mobile elements became inactive in parallel in cyclophyllidean species. The intrinsic DNA losses of inactive copies were likely to have contributed to decrease in genome size. Investigation of retrotransposons in pseudophyllidean cestodes with larger genomes such as *S. erinaceieuropaei* is required to address issues regarding the evolutionary history of retrotransposons and their genomic implications in cestode parasites.

SUPPLEMENTARY MATERIAL

The supplementary material for this article can be found at http://dx.doi.org/10.1017/S0031182016001499.

FINANCIAL SUPPORT

This work was supported by the Basic Science Research Programme of the National Research Foundation of Korea (NRF), which was funded by the Ministry of Science, ICT and Future Planning (Grant no. NRF-2013R1A1A2012011).

REFERENCES

Bae, Y. A., Moon, S. Y., Kong, Y., Cho, S. Y. and Rhyu, M. G. (2001). *CsRn*1, a novel active retrotransposon in a parasitic trematode, *Clonorchis sinensis*, discloses a new phylogenetic clade of *Ty3/gypsy*-like LTR retrotransposons. *Molecular Biology and Evolution* **18**, 1474–1483.

Bae, Y. A., Ahn, J. S., Kim, S. H., Rhyu, M. G., Kong, Y. and Cho, S. Y. (2008). *PwRn*1, a novel *Ty3/gypsy*-like retrotransposon of *Paragonimus westermani*: molecular characters and its differentially preserved mobile potential according to host chromosomal polyploidy. *BMC Genomics* **9**, 482.

Bennett, H. M., Mok, H. P., Gkrania-Klotsas, E., Tsai, I. J., Stanley, E. J., Antoun, N. M., Coghlan, A., Harsha, B., Traini, A., Ribeiro, D. M., Steinbiss, S., Lucas, S. B., Allinson, K. S., Price, S. J., Santarius, T. S., Carmichael, A. J., Chiodini, P. L., Holroyd, N., Dean, A. F. and Berriman, M. (2014). The genome of the sparganosis tapeworm *Spirometra erinaceieuropaei* isolated from the biopsy of a migrating brain lesion. *Genome Biology* **15**, 510.

Berriman, M., Haas, B. J., LoVerde, P. T., Wilson, R. A., Dillon, G. P., Cerqueira, G. C., Mashiyama, S. T., Al-Lazikani, B., Andrade, L. F., Ashton, P. D., Aslett, M. A., Bartholomeu, D. C., Blandin, G., Caffrey, C. R., Coghlan, A., Coulson, R., Day, T. A., Delcher, A., DeMarco, R., Djikeng, A., Eyre, T., Gamble, J. A., Ghedin, E., Gu, Y., Hertz-Fowler, C., Hirai, H., Hirai, Y., Houston, R., Ivens, A., Johnston, D. A. *et al.* (2009). The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460**, 352–358.

Blumenstiel, J. P. (2011). Evolutionary dynamics of transposable elements in a small RNA world. *Trends in Genetics* **27**, 23–31.

Boeke, J. D. and Stoye, J. P. (1997). Retrotransposons, endogenous retroviruses, and the evolution of retroelements. In *Retroviruses* (ed. Coffin, J. M., Hughes, S. H. and Varmus, H. E.), pp. 343–435. Cold Spring Harbor Laboratory Press, New York.

Capy, P. (2005). Classsification and nomenclature of retrotransposable elements. *Cytogenetic and Genome Research* **110**, 457–461.

Collins, J. J., 3rd, Wang, B., Lambrus, B. G., Tharp, M. E., Iyer, H. and Newmark, P. A. (2013). Adult somatic stem cells in the human parasite *Schistosoma mansoni*. *Nature* **494**, 476–479.

Charlesworth, B. and Langley, C. H. (1989). The population genetics of *Drosophila* transposable elements. *Annual Review of Genetics* **23**, 251–287.

Craig, P. S., McManus, D. P., Lightowlers, M. W., Chabalgoity, J. A., Garcia, H. H., Gavidia, C. M., Gilman, R. H., Gonzalez, A. E., Lorca, M., Naquira, C., Nieto, A. and Schantz, P. M. (2007). Prevention and control of cystic echinococcosis. *Lancet Infectious Diseases* **7**, 385–394.

da Silva, A. M. (2010). Human echinococcosis: a neglected disease. *Gastroenterology Research and Practice* **2010**, 583297.

DeMarco, R., Kowaltowski, A. T., Machado, A. A., Soares, M. B., Gargioni, C., Kawano, T., Rodrigues, V., Madeira, A. M., Wilson, R. A., Menck, C. F., Setubal, J. C., Dias-Neto, E., Leite, L. C. and Verjovski-Almeida, S. (2004). *Saci-1*, *-2*, and *-3* and *Perere*, four novel retrotransposons with high transcriptional activities from the human parasite *Schistosoma mansoni*. *Journal of Virology* **78**, 2967–2978.

Dufresne, F. and Jeffery, N. (2011). A guided tour of large genome size in animals: what we know and where we are heading. *Chromosome Research* **19**, 925–938.

Gorelick, R. J., Henderson, L. E., Hanser, J. P. and Rein, A. (1988). Point mutants of Moloney murine leukemia virus that fail to package viral RNA: evidence for specific RNA recognition by a "zinc finger-like" protein sequence. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 8420–8424.

Haig, D. (2012). Retroviruses and the placenta. *Current Biology* **22**, R609–R613.

Harigaya, Y., Tanaka, H., Yamanaka, S., Tanaka, K., Watanabe, Y., Tsutsumi, C., Chikashige, Y., Hiraoka, Y., Yamashita, A. and Yamamoto, M. (2006). Selective elimination of messenger RNA prevents an incidence of untimely meiosis. *Nature* **442**, 45–50.

Huang, Y., Chen, W., Wang, X., Liu, H., Chen, Y., Guo, L., Luo, F., Sun, J., Mao, Q., Liang, P., Xie, Z., Zhou, C., Tian, Y., Lv, X., Huang, L., Zhou, J., Hu, Y., Li, R., Zhang, F., Lei, H., Li, W., Hu, X., Liang, C., Xu, J., Li, X. and Yu, X. (2013). The carcinogenic liver fluke, *Clonorchis sinensis*: new assembly, reannotation and analysis of the genome and characterization of tissue transcriptomes. *PLoS ONE* **8**, e54732.

Jenkins, D. J., Romig, T. and Thompson, R. C. (2005). Emergence/re-emergence of *Echinococcus* spp.-a global update. *International Journal for Parasitology* **35**, 1205–1219.

Jiang, K. and Goertzen, L. R. (2011). Spliceosomal intron size expansion in domesticated grapevine (*Vitis vinifera*). *BMC Research Notes* **4**, 52.

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63.

Koziol, U., Radio, S., Smircich, P., Zarowiecki, M., Fernández, C. and Brehm, K. (2015). A novel terminal-repeat retrotransposon in miniature (TRIM) is massively expressed in *Echinococcus multilocularis* stem cells. *Genome Biology and Evolution* **7**, 2136–2153.

Lavergne, S., Muenke, N. J. and Molofsky, J. (2010). Genome size reduction can trigger rapid phenotypic evolution in invasive plants. *Annals of Botany* **105**, 109–116.

Littlewood, D. T. J. (2006). The evolution of parasitism in flatworms. In *Parasitic Flatworms: Molecular Biology, Biochemistry, Immunology and Physiology* (ed. Maule, A. G. and Marks, N. J.), pp. 1–36. CABI Pub., Cambridge.

Lochmann, T. L., Bann, D. V., Ryan, E. P., Beyer, A. R., Mao, A., Cochrane, A. and Parent, L. J. (2013). NC-mediated nucleolar localization of retroviral Gag proteins. *Virus Research* **171**, 304–318.

Long, A. D., Lyman, R. F., Morgan, A. H., Langley, C. H. and Mackay, T. F. (2000). Both naturally occurring insertions of transposable elements and intermediate frequency polymorphisms at the achaete-scute complex are associated with variation in bristle number in *Drosophila melanogaster*. *Genetics* **154**, 1255–1269.

Maldonado, J. O., Martin, J. L., Mueller, J. D., Zhang, W. and Mansky, L. M. (2014). New insights into retroviral Gag-Gag and Gag-membrane interactions. *Frontiers in Microbiology* **5**, 302.

Malik, H. S. and Eickbush, T. H. (1999). Modular evolution of the integrase domain in the *Ty3/Gypsy* class of LTR retrotransposons. *Journal of Virology* **73**, 5186–5190.

McDonald, J. F. (1990). Macroevolution and retroviral elements. *Bioscience* **40**, 183–191.

Muriaux, D., Costes, S., Nagashima, K., Mirro, J., Cho, E., Lockett, S. and Rein, A. (2004). Role of murine leukemia virus nucleocapsid protein in virus assembly. *Journal of Virology* **78**, 12378–12385.

Oliver, M. J., Petrov, D., Ackerly, D., Falkowski, P. and Schofield, O. M. (2007). The mode and tempo of genome size evolution in eukaryotes. *Genome Research* **17**, 594–601.

Olson, P. D. and Tkach, V. V. (2005). Advances and trends in the molecular systematics of the parasitic platyhelminthes. *Advances in Parasitology* **60**, 165–243.

Olson, P. D., Zarowiecki, M., Kiss, F. and Brehm, K. (2012). Cestode genomics-progress and prospects for advancing basic and applied aspects of flatworm biology. *Parasite Immunology* **34**, 130–150.

Page, R. D. (1996). Tree View: an application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* **12**, 357–358.

Petrov, D. A., Lozovskaya, E. R. and Hartl, D. L. (1996). High intrinsic rate of DNA loss in *Drosophila*. *Nature* **384**, 346–349.

Protasio, A. V., Tsai, I. J., Babbage, A., Nichol, S., Hunt, M., Aslett, M. A., De Silva, N., Velarde, G. S., Anderson, T. J., Clark, R. C., Davidson, C., Dillon, G. P., Holroyd, N. E., LoVerde, P. T., Lloyd, C., McQuillan, J., Oliveira, G., Otto, T. D., Parker-Manuel, S. J., Quail, M. A., Wilson, R. A., Zerlotini, A., Dunne, D. W. and Berriman, M. (2012). A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. *PLoS Neglected Tropical Diseases* **6**, e1455.

Schulman, A. H. (2012). Hitching a ride: nonautonomous retrotransposons and parasitism as a lifestyle; Plant transposable elements: impact on genome structure and function. *Topics in Current Genetics* **24**, 71–88.

Sinzelle, L., Izsvák, Z. and Ivics, Z. (2009). Molecular domestication of transposable elements: from detrimental parasites to useful host genes. *Cellular and Molecular Life Sciences* **66**, 1073–1093.

Smyth, J. D. and McManus, D. P. (1989). *The Physiology and Biochemistry of Cestodes*. Cambridge University Press, Cambridge.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Molecular Biology and Evolution* **30**, 2727–2729.

Thompson, R. C. and McManus, D. P. (2002). Towards a taxonomic revision of the genus *Echinococcus*. *Trends in Parasitology* **18**, 452–457.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. and Higgins, D. G. (1997). The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* **25**, 4876–4882.

Tsai, I. J., Zarowiecki, M., Holroyd, N., Garciarrubio, A., Sanchez-Flores, A., Brooks, K. L., Tracey, A., Bobes, R. J., Fragoso, G., Sciutto, E., Aslett, M., Beasley, H., Bennett, H. M., Cai, J., Camicia, F., Clark, R., Cucher, M., De Silva, N., Day, T. A., Deplazes, P., Estrada, K., Fernández, C., Holland, P. W., Hou, J., Hu, S., Huckvale, T., Hung, S. S., Kamenetzky, L., Keane, J. A., Kiss, F. *et al*. (2013). The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **496**, 57–63.

Volff, J. N. (2006). Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* **28**, 913–922.

Volff, J. N., Bouneau, L., Ozouf-Costaz, C. and Fischer, C. (2003). Diversity of retrotransposable elements in compact pufferfish genomes. *Trends in Genetics* **19**, 674–678.

Wang, X., Chen, W., Huang, Y., Sun, J., Men, J., Liu, H., Luo, F., Guo, L., Lv, X., Deng, C., Zhou, C., Fan, Y., Li, X., Huang, L., Hu, Y., Liang, C., Hu, X., Xu, J. and Yu, X. (2011). The draft genome of the carcinogenic human liver fluke *Clonorchis sinensis*. *Genome Biology* **12**, R107.

Wang, B., Collins, J. J., 3rd and Newmark, P. A. (2013). Functional genomic characterization of neoblast-like stem cells in larval *Schistosoma mansoni*. *Elife* **2**, e00768.

Xiong, Y. and Eickbush, T. H. (1990). Origin and evolution of retroelements based upon their reverse transcriptase sequences. *EMBO Journal* **9**, 3353–3362.

Zheng, H., Zhang, W., Zhang, L., Zhang, Z., Li, J., Lu, G., Zhu, Y., Wang, Y., Huang, Y., Liu, J., Kang, H., Chen, J., Wang, L., Chen, A., Yu, S., Gao, Z., Jin, L., Gu, W., Wang, Z., Zhao, L., Shi, B., Wen, H., Lin, R., Jones, M. K., Brejova, B., Vinar, T., Zhao, G., McManus, D. P., Chen, Z., Zhou, Y. *et al*. (2013). The genome of the hydatid tapeworm *Echinococcus granulosus*. *Nature Genetics* **45**, 1168–1175.