

ARTICLE

Received 21 Nov 2015 | Accepted 8 Aug 2016 | Published 22 Sep 2016

DOI: 10.1038/ncomms12845

OPEN

# Comparative genomics reveals adaptive evolution of Asian tapeworm in switching to a new intermediate host

Shuai Wang<sup>1,\*</sup>, Sen Wang<sup>2,3,\*</sup>, Yingfeng Luo<sup>2</sup>, Lihua Xiao<sup>4</sup>, Xuenong Luo<sup>1</sup>, Shenghan Gao<sup>2</sup>, Yongxi Dou<sup>1</sup>, Huangkai Zhang<sup>2,3</sup>, Aijiang Guo<sup>1</sup>, Qingshu Meng<sup>2</sup>, Junling Hou<sup>1</sup>, Bing Zhang<sup>5</sup>, Shaohua Zhang<sup>1</sup>, Meng Yang<sup>2</sup>, Xuelian Meng<sup>1</sup>, Hailiang Mei<sup>2</sup>, Hui Li<sup>1</sup>, Zilong He<sup>2</sup>, Xueliang Zhu<sup>1</sup>, Xinyu Tan<sup>2</sup>, Xing-quan Zhu<sup>1</sup>, Jun Yu<sup>2</sup>, Jianping Cai<sup>1</sup>, Guan Zhu<sup>6</sup>, Songnian Hu<sup>2</sup> & Xuepeng Cai<sup>1</sup>

*Taenia saginata*, *Taenia solium* and *Taenia asiatica* (beef, pork and Asian tapeworms, respectively) are parasitic flatworms of major public health and food safety importance. Among them, *T. asiatica* is a newly recognized species that split from *T. saginata* via an intermediate host switch ~1.14 Myr ago. Here we report the 169- and 168-Mb draft genomes of *T. saginata* and *T. asiatica*. Comparative analysis reveals that high rates of gene duplications and functional diversifications might have partially driven the divergence between *T. asiatica* and *T. saginata*. We observe accelerated evolutionary rates, adaptive evolutions in homeostasis regulation, tegument maintenance and lipid uptakes, and differential/specialized gene family expansions in *T. asiatica* that may favour its hepatotropism in the new intermediate host. We also identify potential targets for developing diagnostic or intervention tools against human tapeworms. These data provide new insights into the evolution of *Taenia* parasites, particularly the recent speciation of *T. asiatica*.

<sup>1</sup>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, Gansu 730046, China. <sup>2</sup>CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China. <sup>3</sup>University of Chinese Academy of Sciences, Beijing 100049, China. <sup>4</sup>Division of Foodborne, Waterborne and Environmental Diseases, National Center for Emerging and Zoonotic Infectious Diseases, Centers for Disease Control and Prevention, Atlanta, Georgia 30329-4018, USA. <sup>5</sup>Core Genomic Facility and CAS Key Laboratory of Genome Sciences & Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China. <sup>6</sup>Department of Veterinary Pathobiology, College of Veterinary Medicine & Biomedical Sciences, Texas A&M University, College Station, Texas 77843-4467, USA. \* These authors contributed equally to this work. Correspondence and requests for materials should be addressed to G.Z. (email: gzhu@cvm.tamu.edu) or to S.H. (email: husn@big.ac.cn) or to X.C. (email: caixuepeng@caas.cn).

Tapeworms (cestodes) infect all major groups of animals, including humans and many economically important species. Among them, *Taenia solium* (pork), *Taenia saginata* (beef) and *Taenia asiatica* (Asian) tapeworms cause taeniasis in humans (definitive host)<sup>1</sup>. Pork and beef tapeworms are globally distributed, each infecting ~50–60 million people around the world<sup>2,3</sup>. Their larvae infect swine and cattle (intermediate hosts), respectively, leading to considerable economic losses and significant burdens in global food trade<sup>1</sup>. Asian tapeworm is confined to only Asian endemics, and was previously confused with beef tapeworm due to their morphological similarity in adult stage until the mid-1980s when pigs were found to be the major intermediate host. It was recognized as a new species in 1993 (refs 4,5).

Asian and beef tapeworms differ in several morphological details and predilection sites in intermediate hosts. Like *T. solium*, the cysticerci of *T. saginata* (~10 mm in diameter) are mainly established in the bovine striated muscles, whereas those of *T. asiatica* (~2 mm) mainly infect pig livers<sup>6,7</sup>. For *T. asiatica*, fully mature cysticerci are developed in 4 weeks (versus ~10–12 weeks for *T. saginata*)<sup>5</sup>. The hepatotropic feature is considered the major reason that *T. asiatica* has been documented only in Asian countries where some populations consume raw or undercooked pork livers<sup>8</sup>. However, the global impact is probably underappreciated because adult Asian tapeworm can only be distinguished from beef tapeworm by molecular techniques that are not routinely used in some regions of the world, and its ability to cause human cysticercosis has not been ruled out<sup>5,8</sup>. Besides morphological similarity, earlier studies indicated that Asian and beef tapeworms share a recent common ancestor that infected hominids and bovids (resembling *T. saginata*) at the Pleistocene period in Africa, suggesting a switch of the preferred intermediate host from bovids to suids during the speciation of *T. asiatica*<sup>9–11</sup>. Their divergence date were estimated between 0.78 and 1.71 Myr ago<sup>9</sup>, or ~1 Myr ago (0.24–1.64)<sup>12</sup>, based on substitution rates of the mitochondrial COI gene. However, it is unclear how *T. asiatica* became adapted to a new intermediate host and infection site, and evolved into a new species in a relatively short evolutionary time.

Among human intestinal tapeworms, only the pork tapeworm genome was recently reported<sup>13</sup>. Here we present the genomes of beef and Asian tapeworms, making the genomes of all three human taeniasis parasites available for comparative analysis to gain insights into their biological features and genome evolutions, and adaptation of *T. asiatica* to a new intermediate host. Our genome-scale analyses reveal that the divergence time

between beef and Asian tapeworms coincided with the migration of *Homo erectus* from Africa to Asia. The frequent gene duplications may have contributed significantly to the speciation processes. We observe higher genome variability and more accelerated adaptive evolution in *T. asiatica*, particularly in genes involved in host–parasite interactions, physiological homeostasis and nutrient uptake. The two genomes also provide an urgently needed resource for identifying molecular targets shared by human tapeworms for developing new therapeutics, as well as species-specific genes for developing molecular diagnostic tools as described below.

## Results

**Genome features and comparison.** We sequenced the *T. saginata* and *T. asiatica* genomes derived from single adult worms to ~95-fold coverage using the Illumina platform, and assembled them into 3,626 (N50 = 583 kb, total size = 169.1 Mb) and 6,904 (N50 = 342 kb, size = 168.0 Mb) scaffolds with lengths >0.5 kb, respectively. (Note: for clarity, parameters will be described in order of *T. saginata* and *T. asiatica* whenever appropriate hereafter.) The draft assemblies are larger than that of *T. solium* (122.3 Mb)<sup>13</sup>. Both genomes have an identical 42.3% GC content, similar to those of *T. solium*, *Echinococcus multilocularis* and *E. granulosus* (41.9–43.5%)<sup>13,14</sup>, but higher than *Schistosoma* spp.<sup>15</sup> and *Caenorhabditis* spp. (34.1–37.9%; Table 1; Supplementary Table 1). The completeness of the two sequenced genomes are similar, at 89.52% and 90.32%, respectively, as estimated using the Core Eukaryotic Genes Mapping Approach<sup>16</sup>, which are comparable with those of the two well-assembled *Echinococcus* genomes (89.11–92.74%; Supplementary Fig. 1). The two tapeworm genomes contain 10.38 and 10.90% repeated sequences, similar to other tapeworms, but lower than flukes, and all major non-coding RNA species, including conserved microRNA and transfer RNA genes (Table 1; Supplementary Tables 2 and 3).

We predicted 13,161 and 13,323 protein-coding genes in the two genomes, 77.2 and 75.7% of which were supported by RNA sequencing (RNA-seq; Supplementary Methods). More than half of these genes were mappable to the gene ontology (GO) terms (1,472 and 1,461 terms), KO identifiers in the Kyoto Encyclopedia of Genes and Genomes database (3,039 and 3,033) and Pfam domains (3,123 and 3,111). Among the predicted proteins, 2,361 (17.93%) and 2,365 (17.75%) contained transmembrane (TM) domains, and 1,094 (8.21%) and 1,048 (7.87%) possessed signal peptide sequences. These genes constitute metabolic pathways

**Table 1 | Genomic features of *T. saginata* and *T. asiatica* in comparison with other worms.**

	<i>T. saginata</i>	<i>T. asiatica</i>	<i>T. solium</i> *	<i>E. multilocularis</i>	<i>H. microstoma</i>	<i>S. mansoni</i>	<i>C. elegans</i>
Assembly size (Mb)	169	168	131	114	141	365	100
GC content (%)	43.2	43.2	43.5	42.2	36.0	35.2	35.4
Coding genes number	13,161	13,323	11,902	10,506	10,141	10,772	20,469
Average gene length (Kb)	6.0	5.9	4.6	5.4	6.1	15.4	3.1
Protein length (aa)	464	466	444	505	490	477	453
Gene density (genes per Mb)	77.9	79.3	90.9	92.2	71.9	29.5	201.0
Number of exons per gene	6.2	6.2	6.6	6.8	6.4	6.5	6.4
Mean length of exons (bp)	237	244	237	220	229	226	212
Number of introns per gene	5.2	5.2	5.6	5.8	5.4	5.5	5.4
Mean length of introns (bp)	864	831	775	684	862	2,460	354
GC content of exon (%)	49.7	49.6	50.2	50.0	44.3	36.0	42.6
GC content of intron (%)	41.5	41.2	40.8	39.9	34.6	34.7	32.5
Repeat content (%)	10.4	10.9	18.1	10.9	7.6	40.0	17.0
tRNA number	339	353	162	856	44	153	966

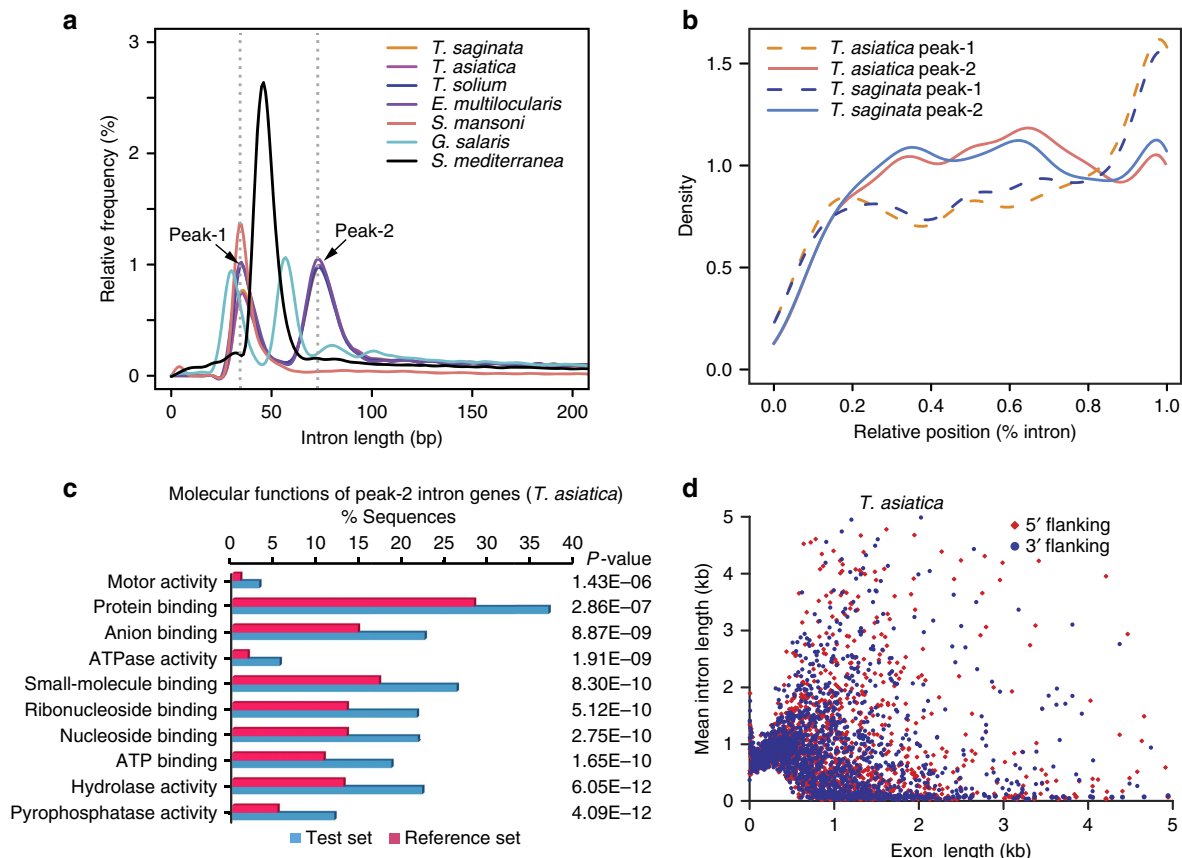
\**T. solium* v2 genome (China isolate) properties and gene models.

that are virtually identical to those in other tapeworms, including the loss of ability to *de novo* synthesize certain nutrients such as most amino acids, steroid hormones and lipids<sup>13,14</sup> (Supplementary Fig. 2).

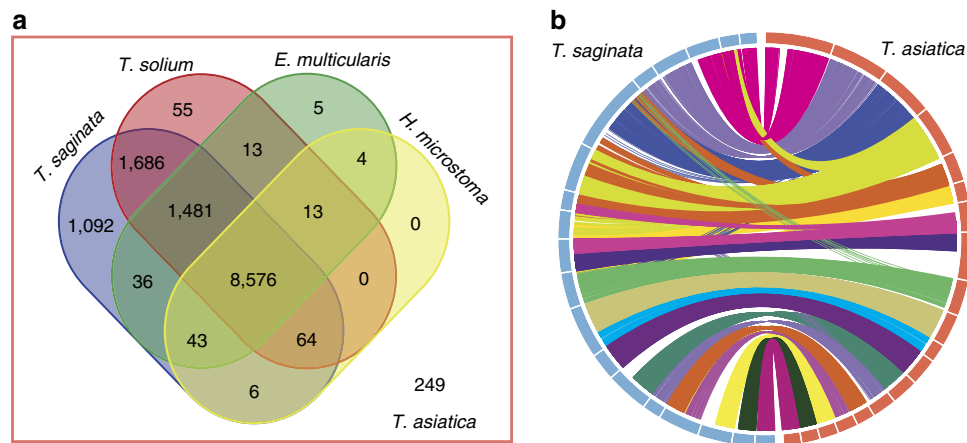
Introns are present in 81.0 and 80.9% of the genes in the two tapeworms. The lengths of short introns follow a bimodal distribution pattern with two major peaks at ~36 bp (peak-1 introns) and ~73 bp (peak-2; Fig. 1a). This feature was also observed in other tapeworms<sup>13</sup> and the monogenean *Gyrodactylus salaris*<sup>17</sup>, but not in the flukes and *Schmidtea mediterranea* (peak-1 only), suggesting the presence of this feature predating the expansion of cestodes (or before the divergence of parasitic flatworms, but lost in the fluke lineage). These peak-1 and 2 introns are located preferentially in the 3'-end and middle regions of genes, respectively (Fig. 1b). Genes containing peak-2 introns tend to possess more introns than those containing peak-1 introns (for example, average 10.52 and 14.80 introns per gene in peak-1 and 2 intron-containing sequences in *T. saginata* ( $P < 0.01$  by two-sided Wilcoxon-rank sum test) (Supplementary Fig. 3). Interestingly, peak-2 intron-containing genes are significantly enriched to certain functional groups (for example, pyrophosphatase activity, hydrolase activity and nucleoside binding) and cellular components (for example, cytoskeletal motor proteins and membrane proteins; Fig. 1c; Supplementary Data 1). A striking but previously unreported feature in the tapeworms is the apparent preference of the mean lengths of neighbouring introns flanking small exons.

The minimal mean lengths of introns flanking small exons (<400 bp) are 502 and 370 bp long in *T. saginata* and *T. asiatica*, respectively. This feature also occurs in some other invertebrate species, for example, *S. mansoni*, *Caenorhabditis elegans* and *Drosophila melanogaster* (Fig. 1d; Supplementary Figs 4 and 5). However, the extensiveness of this feature among eukaryotes and the mechanism behind it remains to be elucidated. Introns in tapeworms are biased towards A/T bases (26.8/31.1% in *T. saginata*; 26.6/31.1% in *T. asiatica*), but not as obvious as in *S. mansoni* (31.3/33.6%) and *G. salaris* (33.8/35.3%; Supplementary Fig. 6). However, no A/T-bias in exons was observed in tapeworms, although it was found in the flukes (that is, 25.2/25.0% in *T. saginata* and 25.3/25.1% in *T. asiatica*, versus 31.5/33.2% in *S. mansoni*; Supplementary Fig. 7).

We compared homologous genes among human tapeworms, and observed that 12,984 (97.5%) and 11,888 *T. asiatica* genes (90.3%) had homologues in *T. saginata* and *T. solium* (BLASTP cutoff:  $1e^{-4}$ ; Fig. 2a). Pair-wise collinearity analysis of orthologous blocks on scaffolds revealed a higher degree of similarity between *T. saginata* and *T. asiatica* ( $n_1 = 7,201$ ,  $n_2 = 7,212$ , 292 blocks) than between the two species and *T. solium* ( $n_1 = 6,055$ ,  $n_2 = 6,058$ , 303 blocks; Fig. 2b). Similarly, a higher nucleotide identity (92.26%) with larger alignable blocks (total 138 Mb, mean length = 5.4 kb) was observed between *T. saginata* and *T. asiatica* (versus ~88.53% identity, ~108 Mb total and 3.8 kb mean lengths between beef/Asian and pork tapeworms). These data agree with their taxonomic affiliations.



**Figure 1 | Unique intron features in the tapeworm genomes.** (a) Bimodal length distributions of short introns in tapeworms (*T. asiatica*, *T. saginata*, *T. solium* and *E. multilocularis*) and monogenean *G. salaris* in comparison with the unimodal distributions in the fluke *S. mansoni* and the free-living flatworm *S. mediterranea*. (b) Preferential distributions of peak-1 and peak-2 introns toward 3'-end and middle regions of genes, respectively. The curves show the relationship between intron densities and relative positions from the 5'-ends of genes. (c) Peak-2 intron-containing genes are highly enriched to certain functional groups (shown for *T. asiatica*, by two-sided Fisher's exact test). (d) Length preference of introns flanking small exons (<400 bp; shown for *T. asiatica*). The minimal mean length of (5' or 3') introns flanking small exons that have a specific length (<400 bp) is 370 bp in *T. asiatica*.



**Figure 2 | Homologues and synteny between *T. asiatica* and related tapeworms.** (a) Homologous genes shared between *T. asiatica* and other tapeworms (that is, *T. saginata*, *T. solium*, *E. multilocularis* and *H. microstoma*). (b) Gene block linkages between *T. asiatica* and *T. saginata*. The collinear gene blocks determined by MCSScan between genome scaffolds (>1Mb) represent 7,212 and 7,201 genes for *T. asiatica* and *T. saginata*, respectively.

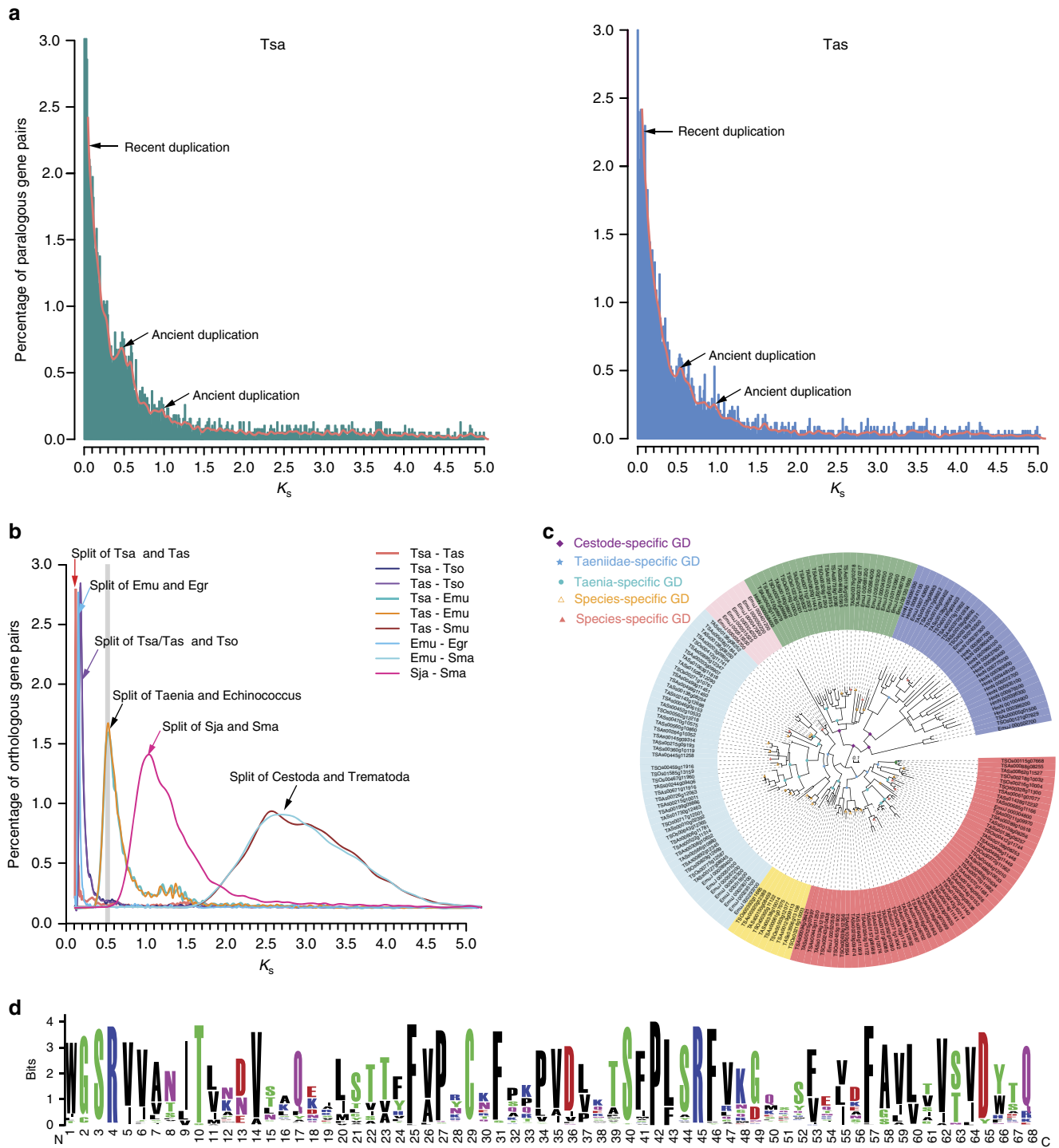
**Gene duplications during divergence of tapeworms.** Gene duplication (GD) is known as a primary source of materials for evolutionary innovations and adaptations<sup>18,19</sup>, in which the age of a GD event is proportional to the number of synonymous substitutions per synonymous site ( $K_s$ ) of paralogous genes. In the genomes of *T. saginata*, *T. asiatica* and other related parasitic flatworms, the  $K_s$  distributions are typically quasi-exponentially L shaped (Fig. 3a,b; Supplementary Fig. 8), which agrees well with the notions that most duplications are young (for example, 4.36 and 4.98% GDs with  $K_s < 0.01$  in the *T. saginata* and *T. asiatica* genomes) because of the continuous loss of duplicated genes over the time. These events are mostly derived from small-scale gene duplications (SSGDs), predominated by dispersed duplications (for example, 71.95% in *T. saginata*, 74.14% in *T. asiatica* and 69.94% in *E. multilocularis*), followed by tandem (16.69, 15.69 and 15.08%) and proximal duplications (10.75, 8.90 and 11.29%; Supplementary Table 4) in the current assemblies.

Although retrotransposons are the major elements subjected to GDs in the *T. saginata* and *T. asiatica* genomes, several functional homologous groups (including surface antigens, HSP70, ubiquitin conjugating enzyme, ryanodine receptor 44f, cyclin-dependent kinase, puromycin-sensitive aminopeptidase and zinc-finger proteins) appear to have also experienced continuous and extensive SSGDs during the evolution history of the tapeworm lineage (Fig. 3c; Supplementary Figs 9 and 10; Supplementary Table 5). These SSGD events resulted in many overrepresented super-families with high sequence diversities accompanying the diversification of their biological functions. Among them, the frequent duplications and retentions of tapeworm-specific surface antigens (for example, Taeniidae antigens and diagnostic antigen gp50) are indicative of their importance in the parasite survival and/or adaptations to new environments. Indeed, Taeniidae antigens are known to play important roles in the evasion of host immunity<sup>20</sup>, while the diagnostic antigen gp50 proteins are glycosyl phosphatidylinositol-anchored membrane glycoproteins also heavily involved in interacting with the host immune system<sup>21</sup>. The gp50 gene family appears to have duplicated more extensively, driven mostly by duplicative transpositions and tandem duplications along the Taeniidae and *Taenia*-specific evolution history (Fig. 3c). These duplicated genes have been differentially retained in different tapeworms, and might have carried out multiple functions through neo-/sub-functionalization with greatly divergent sequences (Fig. 3d), although their precise biological roles remain to be illustrated.

As a major force in evolution, species-specific GDs can lead to the differentiation of gene functions, thus facilitating the species-specific adaptation and divergence<sup>18</sup>. Both *T. saginata* and *T. asiatica* genomes possess a large number of recently duplicated genes (involving 866 and 1,075 in-paralogs after their divergence, respectively) that were derived from 481 and 614 duplicate events along each lineage, respectively. The duplicate genes appear to arise at a high average rate in their genomes (0.0321 and 0.0404 duplicates per gene per Myr for *T. saginata* and *T. asiatica*, respectively), similar to those in estimated in *E. multilocularis* (0.0304) and *C. elegans* (0.0208)<sup>22</sup>, suggesting a probable high degree of plasticity of their genomes (Supplementary Methods, section 8.9). Species-specific GDs were also shown by gene enrichment analysis, in which differential distributions of significantly enriched gene categories in biological processes ( $P < 0.05$  by two-sided Fisher's exact test) were observed between *T. asiatica* (mostly enriched in nucleosome assembly, cilium movement and ribosome localization) and *T. saginata* (mostly enriched in protein glycosylation; Supplementary Table 6), indicative of nonrandom processes of gene retentions in each genome. These observations suggest that the high rate of origin of GDs and preferential retentions of duplicated gene families might have contributed, at least partially, to the divergence of the two closely related tapeworms.

Duplicated genes in the process of acquiring new functions at the time of species separation are likely to contribute to species differentiation<sup>19</sup>. Using out-paralogous genes from *T. solium* as outgroups, we compared the evolutionary rate for each pair of the in-paralogous genes arising after the divergence of *T. saginata* and *T. asiatica*, and identified a number of newly duplicated genes that evolved significantly asymmetrically between paralogous pairs (88/592 in *T. saginata*, 72/804 in *T. asiatica*;  $P < 0.05$  by Tajima's relative rate test), implying that they were prone to diverge functions. Particularly, the cytoskeleton components (for example, actin, dynein heavy chain and kinesin), tegument surface antigens (for example, EG95 and gp50), ion transporters/channels (for example, ryanodine receptors 44f, solute carrier family 12 and multidrug resistance protein) and growth/development-related proteins (for example, fibroblast growth factor receptor 4, round spermatid basic protein 1 and segment polarity protein disheveled) appear to not reach at a stable rate yet in *T. asiatica* (Supplementary Data 2). These proteins are probably important in the speciation of





**Figure 3 | Gene duplications revealed by  $K_s$  analysis.** (a) The percentage of paralogous gene pairs of duplicated genes along with  $K_s$  values are typically L shaped in the *T. saginata* (Tsa) and *T. asiatica* (Tas) genomes, indicating the occurrence of continuous gene duplication events and the losses of duplicated genes over the time. (b) Distribution of  $K_s$  values in orthologous genes with peaks indicating the splits between various flatworms *T. asiatica* (Tas), *T. saginata* (Tsa), *T. solium* (Tso), *E. granulosus* (Egr), *E. multilocularis* (Emu), *S. mansoni* (Sma) and *S. japonicum* (Sja). The grey bar indicates the divergence point between the *Taenia* and *Echinococcus* lineages. (c) Extensive duplications of diagnostic antigen gp50 genes in the tapeworm lineage (*Taenia*, *Echinococcus* and *Hymenolepis*). (d) Sequence logo shows the conserved and distinct sequence characteristics of the gp50 sequences from the tapeworms. The sequence logo was generated from 183 sequences aligned at the conserved blocks selected by Gblocks with a less stringent selection ([http://molevol.cmima.csic.es/castresana/Gblocks\\_server.html](http://molevol.cmima.csic.es/castresana/Gblocks_server.html)).

*T. asiatica*, considering their roles in the tegument maintenance, homeostasis regulation and growth/reproduction. The galactosyltransferase gene family may play a particularly important but yet undefined biological role in *T. saginata*

because of its continuous duplications along tapeworm evolution. In addition, the evolutionary divergence of recently duplicated gene pairs is echoed by the divergences in their expression patterns ( $R^2=0.2462$  in *T. saginata*; 0.0114 in

*T. asiatica*) (Supplementary Fig. 11) and gene structures (accounting for 23.82 and 15.05% in-paralogous pairs in *T. saginata* and *T. asiatica*), indicative of their possible functional divergences.

### Speciation history and accelerated evolution in *T. asiatica*.

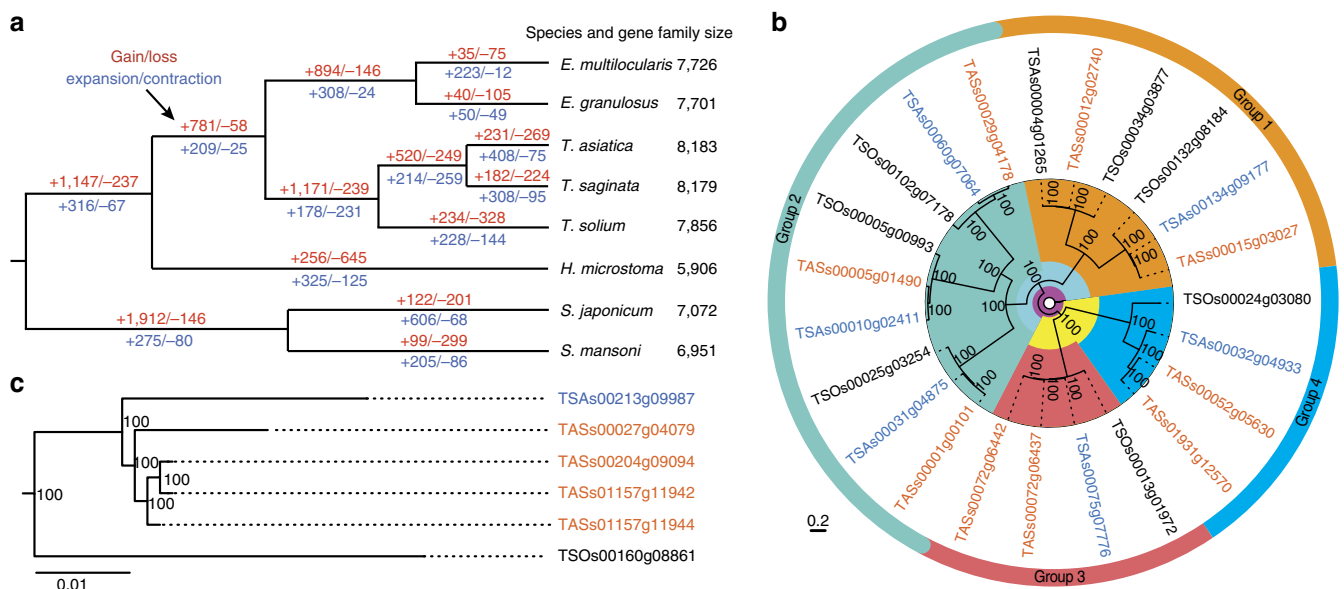
We collected 102 single-copy genes conserved in 10 flatworms and related species for estimating the divergence dates of tapeworms using a relaxed-clock Bayesian approach<sup>23</sup>. The genome-based analysis suggested that beef and Asian tapeworms diverged at 1.142 Myr (0.55–1.43, 95% highest probability density) in the early Pleistocene period (Supplementary Fig. 12), which was close to previously estimated date (~1.0 Myr ago) based on mitochondrial genes<sup>11,12</sup>. This split predated the domestication of pigs and cattle (~10,000 years ago)<sup>24</sup>, and the migration out of Africa of *Homo sapiens* (~100,000 years ago)<sup>25</sup>, but fell within the periods of population expansion and migration of *H. erectus* from Africa to Asia (from ~1.8 to 0.4 Myr ago)<sup>26</sup>. Currently, the closest relatives of *T. saginata* and *T. asiatica* (for example, *T. simbae*) are only found in Africa or adjacent regions<sup>10,11</sup>. If the ‘minimum number of host shifts’ theory was assumed, our finding agrees with an earlier speculation<sup>9–11</sup> that the common ancestor of beef and Asian tapeworms (*T. saginata* like) first colonized humans at or before early Pleistocene, and the hominid ancestor *H. erectus* acquired this tapeworm from bovids in Africa and then transmitted it to suids in Asia. We further speculate that the persistent hunting activity in *H. erectus* permitted long and consistent interactions among three host species, thus mediating the transfer of *T. saginata*-like ancestor from bovids to suids, giving rise to *T. asiatica*.

Despite high molecular and morphological similarities between Asian and beef tapeworms, the nucleotide substitution rate (branch-length) in protein-encoding genes of *T. asiatica* (0.00467 mutation per site) is 1.27-fold higher than that of *T. saginata* (0.00379 mutation per site) (Supplementary Fig. 13; Tajima’s

relative rate test,  $P < 0.01$ ). In addition, the nucleotide mutation rate in the *T. asiatica* genome is at  $4.09 \times 10^{-9}$  ( $3.27\text{--}8.52 \times 10^{-9}$ ) mutations per site per year (versus  $\sim 3.32 \times 10^{-9}$  in *T. saginata* and  $\sim 2.82 \times 10^{-9}$  in *T. solium*), which is  $\sim 10$ -fold higher than that of humans ( $0.33\text{--}0.47 \times 10^{-9}$ )<sup>27</sup>. The higher mutation rate provided a greater genome variation for selection and adaptation needed for the divergence/speciation of *T. asiatica* from *T. saginata*. However, the mechanism leading to the accelerated evolution rate in the *T. asiatica* genome is unclear.

By mapping useful sequence reads of short paired-end libraries to the assembled genomes, we detected substantial numbers of heterozygous single-nucleotide variations (SNVs). The overall SNV rate in *T. asiatica* is 2.97-fold higher than in *T. saginata* (Supplementary Table 7). The genomes of *T. asiatica* and *T. saginata* contained 60,734 (362 sites per Mb) and 20,700 (122 sites per Mb) high-quality heterozygous SNVs, respectively. Among them, 6.90% (*T. asiatica*) and 5.96% (*T. saginata*) were located in protein-coding genes, in which the most significantly enriched genes were transporters in *T. asiatica*, including those for ions ( $n = 32$ ), organic anions ( $n = 11$ ), amino acids ( $n = 5$ ) and sulfates ( $n = 2$ ; Supplementary Data 3). These proteins are mostly involved in maintaining cellular homeostasis and nutrients absorption (Supplementary Fig. 14). In addition, we observed more small indels (size = 1–5 bp) in the *T. asiatica* genome (2,359 indels) than in *T. saginata* (1,014 indels; Supplementary Table 7). Both natural mutations over the time and genetic exchanges between individuals might contribute to heterozygosity, but their individual contributions in *T. asiatica* could not be established here due to the lack of intra-species genetic diversity data.

The *T. asiatica* genome also experienced more gene family gain ( $n = 231$ ) and expansion ( $n = 408$ ) than *T. saginata* ( $n = 182$  and 308; Fig. 4a). The gained gene families are mainly novel domains of unknown functions, possibly related to certain specialized adaptations, while the gene copy-number variations



**Figure 4 | Evolution of gene families in the flatworms and roundworms.** (a) The dynamics of gene family sizes in the genomes of *T. saginata*, *T. asiatica*, *T. solium*, *E. granulosus*, *E. multilocularis*, *H. microstoma*, *S. japonicum* and *S. mansoni*. Numbers above and below the branches indicate gene family gains/losses (red) and the expansions/contractions (blue), respectively. (b) Phylogenetic reconstruction clustered low-density lipoprotein receptor (LDLR) genes from *T. saginata*, *T. asiatica* and *T. solium* into four groups, in which group-4 LDLR genes were expanded only in the *T. asiatica* genome (Supplementary Methods, section 8.8). (c) The expansion of fatty acid desaturases (FADS) in the *T. asiatica* genome, compared with that in the *T. saginata* genome (Supplementary Methods, section 8.8).

are related to functional extension. Intriguingly, significant expansions (branch-specific  $P$  values  $< 0.05$ ) of low-density lipoprotein receptor (LDLR) and fatty acid desaturase (FADS) genes were noticed in *T. asiatica* ( $n = 9$  and  $4$  versus  $n = 7$  and  $1$  in *T. saginata* or *T. solium*; Fig. 4b,c). We speculate that this expansion played an important role in the switch of the intermediate host from cattle to swine by promoting the establishment of cysterici of Asian tapeworm in the lipid-rich liver.

**Adaptive selection in the *T. asiatica* genome.** Positive selection is an important source of evolutionary innovation and one of the major forces driving species divergence. To evaluate the role of positive selection in the evolution of *T. asiatica*, we selected 1:1 orthologous genes from six tapeworms for branch-site model analysis by Phylogenetic Analysis by Maximum Likelihood (PAML)<sup>28</sup>, and identified 134 and 102 positively selected genes (PSGs) in the *T. asiatica* and *T. saginata* genomes (likelihood ratio test,  $P < 0.05$ ; Supplementary Data 4). We analysed the PSGs in *T. asiatica* to examine whether these rapidly evolving genes were enriched for specific functions after the divergence from *T. saginata*. We observed evolutionary pressures on some essential genes in cellular processes, including those involved in transcription, translation and regulating protein degradation (for example, various ribosomal proteins, tRNA guanine N7 methyltransferase, bifunctional aminoacyl-tRNA synthetase, small subunit processome component 20, transcription factors and ubiquitination-associated proteins; Supplementary Data 4).

Adaptive selection signals were also observed in some genes associated with specialized survival environment for *T. asiatica*, but not in *T. saginata*. For instance, selection was detected in genes involved in pH maintenance and ion homeostasis (for example, carbonic anhydrase, glutamate receptor ionotropic kainite and amiloride-sensitive cation channel 4), implying adaptation to the new host internal environment. In addition to *LDLR* and *FADS* genes, we observed PSGs involved in lipid scavenger (for example, Niemann Pick C1 protein, fatty acid-binding protein and glycolipid transfer protein) and glycolysis (for example, pyruvate kinase and fructose-2,6-bisphosphatase) that were probably beneficial to *T. asiatica* in establishing infection in the lipid/sugar-rich liver in pigs. The tegument is essential for protecting parasitic flatworms from attacks by host defence systems<sup>29–31</sup>. We observed strong selection signals in genes responsible for maintaining body surface integrity in Asian tapeworm, including cytoskeleton-associated proteins<sup>32,33</sup> (for example, myotubularin protein, myosin heavy chain, dynein heavy and light chains, kinesin-related proteins, intraflagellar transport protein, calponin, katanin and 4.1 protein, ezrin, radixin, moesin (FERM) domain-containing protein), cell adhesion/junction (for example, tight junction protein,  $\beta$ -catenin protein, protocadherin gamma and FRAS1-related extracellular matrix protein), and a glycosyltransferase gene probably involved in forming the thick glycocalyx layer on the tegumental surface<sup>13,34</sup>.

**Proteins involved in host–parasite interactions.** We analysed proteases, protease inhibitors (PIs) and excretory/secretory (E/S) proteins that are commonly involved in interacting with hosts and modulating host immune responses<sup>35,36</sup>. Particularly, secreted proteases can modulate host Th2 immune responses against helminths<sup>37</sup>. We predicted 157 and 161 proteases, plus 142 and 155 non-protease homologues, in the *T. saginata* and *T. asiatica* genomes, which were comparable to those in *E. multilocularis*<sup>13</sup> (Supplementary Data 5 and 6). They belong to five major classes (aspartic, cysteine, metallo, serine and

threonine), predominated by metallo- ( $n = 46–48$ ), cysteine ( $n = 41–44$ ) and serine proteases ( $n = 27–30$ ; Supplementary Fig. 15). In addition, 23 and 26 proteases are encoded by the top 10% highly transcribed genes in larval *T. saginata* and adult *T. asiatica* (Supplementary Table 8). The two genomes encode 70–71 PIs that are all serine, cysteine and metalloprotease inhibitors, including 139 family PIs that interact with endopeptidases regardless of the catalytic type (Supplementary Data 7). The largest PI family is I02 ( $n = 22$ , 14.0% in *T. saginata* and  $n = 20$ , 12.4% in *T. asiatica*), which are serine PIs (aka Kunitz inhibitors; Supplementary Table 9). Several families of proteases and PIs were among the most enriched E/S proteins (Supplementary Tables 10 and 11). By comparison with PIs in *S. mansoni*, we identified several tapeworm-specific inhibitor families (for example, I87, I21 and I93), suggesting that tapeworms and flukes employ lineage-specific mechanisms to regulate protease activities.

The secretomes are large in *T. asiatica* ( $n = 824$ , 6.18%) and *T. saginata* ( $n = 885$ , 6.72%; Supplementary Tables 10 and 11; Supplementary Fig. 16). Many of them are proteases and PIs as described above. The two genomes also encode a large set of other classes of E/S proteins that may be involved in modulating host immune responses. For instance, some ‘Taeniidae antigens’ ( $n = 24$  in *T. asiatica* and  $n = 39$  in *T. saginata*) could impair neutrophil chemotaxis and/or modulate Th2 polarization<sup>38</sup>. Another large E/S family is venom allergen-like proteins, which are known to modulate host immune function and regulate sexual development of parasites in the host<sup>39</sup>.

**Molecular targets for intervention and diagnosis.** Together with the previously reported *T. solium* and *Echinococcus* genomes, the availability of *T. saginata* and *T. asiatica* genome sequences allowed us to identify potential targets shared by all human tapeworms but divergent or absent in hosts for developing therapeutics. We identified 75–78 G-protein-coupled receptors (GPCRs) and 353–355 protein kinases in *T. asiatica* and *T. saginata* that are well-known classic drug targets (Supplementary Data 8 and 9). Most GPCRs are rhodopsin family proteins ( $n = 63$ ), while protein kinases cover  $\sim 10$  major classes, in which 180 kinase groups (Supplementary Data 9) could not be clustered with those from the reference species (that is, human, *D. melanogaster* and *C. elegans*), thus may serve as potential ideal drug or vaccine targets against the parasitic helminthes. Ligand-gated ion channels (LGICs) are validated targets for many current antihelminthic drugs. We identified at least 33 members of three major LGIC families (glutamate-activated cationic channels, cys-loop LGIC and ATP-gated ion channels), and  $\sim 20$  members of related families (that is, cyclic-nucleotide-gated cation channel and amiloride-sensitive sodium channel related; Supplementary Data 10). Most of these drug targets are conserved among tapeworm genomes, thus might potentially serve as broad-spectrum drug targets.

We further searched for parasite-specific sequences from the potential drug targets and host–parasite interaction-associated proteins, and identified 34 sequences (*T. saginata*) and 45 sequences (*T. asiatica*) with no homologues in mammals (Supplementary Data 11 and 12). Most of these sequences were supported by transcription data, including several homologues of known drug targets. Among them, cystatin and phytochelatin synthase are the top drug target candidates because they are present in all tapeworms and critical in interacting with hosts and heavy-metal detoxification, respectively<sup>13,40</sup>.

The tapeworm genomes were searched for species-specific genes potentially valuable in developing molecular and/or immunological diagnostic tools, particularly those for specific



detection of Asian tapeworm. We collected all single-copy genes in the *T. saginata* and *T. asiatica* genomes, and ranked them by nucleotide sequence divergence as a community resource for developing molecular detection tools (Supplementary Data 13). We also recovered 15 (versus *T. saginata*) and 110 (versus *T. solium*) high-confidence species-specific protein-coding genes in the *T. asiatica* genome (Supplementary Table 12). However, the feasibility of these genes in developing immunological assays needs experimental evaluations.

## Discussion

*T. asiatica* is morphologically indistinguishable from *T. saginata* in adult stage, but shares the same intermediate host with *T. solium*. Its speciation and switch of the intermediate host are intriguing evolutionary questions. Our comparative analysis revealed that the three human taeniasis parasites share many common genomic features but differ from each other in the evolutions and diversifications of certain specialized gene families, and reaffirmed the sister relationship between *T. asiatica* and *T. saginata*. Genome-based analysis suggests that Asian and beef tapeworms diverged ~1.14 Myr ago, which coincides the migration of *H. erectus* to Asia, rather than the more recent migration of *H. sapiens*<sup>9,11</sup>. Thus, the speciation of *T. asiatica* predated that of *H. sapiens*. The divergence between the Asian and beef tapeworms might have been driven by the differential GDs in their genomes that display typically L-shaped distribution patterns. The high rates of extensive and continuous duplications, differential retentions and subsequent functional diversifications of gene families in the two tapeworm genome (for example, families associated with the cytoskeleton components, tegument surface antigens, ion transporters/channels in *T. asiatica* and glycosylation in *T. saginata*) might have significantly contributed to the speciation of *T. asiatica*.

We observed accelerated evolution in *T. asiatica* in mutation rate, heterozygosity and gene family gain/expansion that are all at higher rates than in *T. saginata*, suggesting that this parasite is of high evolution vigour in adaptation to new host environments. These observations challenge the hypothesis that *T. asiatica* species is at risk of extinction due to its minimal genetic diversity and limited geographical distribution<sup>41–43</sup>. Given that those studies were only based on analyses of a limited number of mitochondrial or nuclear genes, the real intra-species diversity within the *T. asiatica* populations may still need further large-scale investigations.

The PSGs in *T. asiatica* are more concentrated to gene families involved in internal homeostasis (for example, carbonic anhydrase), tegumental development (for example, cytoskeletal proteins, cell junction proteins and glycosyltransferases) and lipid uptake (for example, LDLR and fatty acid-binding protein), which are probably associated with the establishment of new immune-evasion and nutrient uptake strategies at the lipid-rich infection site (liver) in a new intermediate host (pigs). Our analysis indicates that tegument and membrane proteins in tapeworms are under particularly high evolutionary pressure in adaptation to new hosts, as evidenced by the rapidly evolved new genes/adaptive evolution/gene expansion associated with the recent host switch or speciation in *T. asiatica*. The tegument surface antigens (for example, Taeniidae antigens and diagnostic antigen gp50) may be of particular importance in tapeworm's survival and/or adaptations to new environments due to their critical roles in interacting with the host immune systems.

We identified several sets of proteins that might serve as broad-spectrum drug targets in tapeworms, including kinases, GPCRs and ion channels, as well as two proteins (cystatin and phytochelatin synthase) that could serve as potentially ideal

targets in *T. saginata* and *T. asiatica*. Due to the morphological similarities, misdiagnosis between *T. saginata* and *T. asiatica* is not uncommon. Recently, several molecular approaches using sequence-specific DNA probes, PCR-based RFLP and multiplex PCR based on mitochondrial sequences were explored for differential diagnosis of the two *Taenia* tapeworms<sup>44–46</sup>. We also provide here a list of nuclear genes based on sequence divergence between the two closely related tapeworms for development of new molecular diagnostic tools.

## Methods

**Samples and preparations.** Adult worms of *T. saginata* and *T. asiatica* were isolated from two patients (one worm per patient) in Yunnan Province, China. The study was approved by the ethics committee of Affiliated Hospital of Dali University (Yunnan, China) and patients have given written consent to publication of this study with the exclusion of any personal identifiers. Larvae (cysticerci) of *T. saginata* were obtained from the skeletal muscle of an experimentally infected calf 7 weeks after infection. The animal was cared in accordance with good animal practice according to the Animal Ethics Procedures and Guidelines of the People's Republic of China, and the study was approved by the Institutional Committee for the Care and Use of Experimental Animals of Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences (no. LVRIAEC2010-002). Genomic DNA was extracted from freshly collected middle proglottids for constructing two paired-end libraries (300 and 500 bp inserts) and three mate-pair libraries (1, 5 and 10 kb) for *T. saginata*, and two paired-end libraries (500 bp) and seven mate-pair libraries (2, 2, 3.5, 5, 5, 7 and 10 kb) for *T. asiatica* (Supplementary Methods, section 1). Messenger RNA was isolated from *T. saginata* larvae and the middle proglottids of *T. asiatica* for the construction of paired-end cDNA libraries (300 bp). For each worm, the clonality was confirmed by the distribution of minor allele frequencies of the heterozygous sites (Supplementary Fig. 17).

**High-throughput sequencing and assembly.** Paired-end and mate-pair sequencing was performed using the Illumina Sequencing Systems. Adaptor sequences, PCR duplicates, contaminants and low-quality sequences were removed, and high-quality sequences were assembled into contigs using ABySS (v1.3.5)<sup>47</sup> and SOAPde novo (v1.05)<sup>48</sup>. Scaffolds were constructed from contigs using SSPACE (v-PREMIUM-2.3)<sup>49</sup>. Gaps in the scaffolds were closed using GapFiller (v1.10)<sup>50</sup>. Assembly completeness and redundancy were assessed with Core Eukaryotic Genes Mapping Approach<sup>16</sup> and RNA-seq data (Supplementary Methods, section 2). Genome sizes were estimated by the k-mer-based method (Supplementary Fig. 18). For transcriptome analysis, high-quality RNA-seq reads were mapped to the genomes using TopHat (v2.0.12)<sup>51</sup>. Transcripts were built by Cufflinks (v2.0.2)<sup>52</sup> with default settings. *De novo* assembly of RNA-seq reads was performed using Trinity (v2.0.3)<sup>53</sup>. Expression levels were evaluated by fragments per kilobase of transcript per million fragments of mapped genes, using Cufflinks (v2.0.2) (-G parameter) referenced to the final EVM integrated GFF files (Supplementary Methods, section 4).

**Gene prediction and annotation.** We combined several approaches to predict protein-encoding genes, including homology-based searches, *ab initio* prediction and transcriptome-based prediction methods. Individual predictions were merged by EvidenceModeler (v1.1.1)<sup>54</sup>. Untranslated regions were added with PASA (v2.0.0)<sup>55</sup>. Predicted proteins were searched by BLAST algorithms for homologues in the National Center for Biotechnology Information non-redundant protein databases and Uniprot database. InterproScan5 (ref. 56) was used for identifying domains, mapping GO terms and assigning functional classifications. KO terms were assigned and pathway mapping was performed using the annotation tools at the Kyoto Encyclopedia of Genes and Genomes server (<http://www.genome.jp/tools/kaas/>) (see Supplementary Methods, section 5). Non-coding RNA and repeat elements were also predicted (Supplementary Methods, section 3.2).

**Comparative genomics and SNV analysis.** Protein similarities were determined by all-against-all BLASTP searches ( $1e^{-4}$ ) using predicted proteome sequences of *T. asiatica* as queries against those of *T. saginata* and *T. solium*. The *T. solium* genome assemblies (version 2) available at <http://taenia.big.ac.cn/taenia/index.html> were used in all analyses or otherwise as specified. The resulting *T. asiatica*-specific genes were further searched by BLASTP ( $1e^{-3}$ , 80% length coverage) against genome assemblies of *T. saginata* and *T. solium* to identify high-confident-specific genes. The assemblies of the three human tapeworms were aligned using MUMmer (v3.22)<sup>57</sup> to identify genome syntenic blocks. The collinearity analysis for orthologous genes on scaffolds was conducted using MCScanX<sup>58</sup> (Supplementary Methods, section 6). High-quality reads were mapped to the reference assemblies using Bowtie2 (v2.2.3)<sup>59</sup>. Reads corresponding to PCR duplicates were removed by MarkDuplicates from PICARD (v1.119) (<http://picard.sourceforge.net>), followed by base quality recalibration and indel realignment by GATK (v3.5)<sup>60</sup>. SNVs and indels in the *T. asiatica* and *T. saginata* genomes were detected by HaplotypeCaller from GATK and filtered by the



coverage, mapping quality score, FisherStrand (FS) value (Phred-scaled  $P$  value using Fisher's exact test) and other filters. To make the result comparable, a similar coverage ( $\sim 65\times$ ) of reads for each genome was used to calculate the proportion of variants removed by each filter (Supplementary Methods, section 3.3).

**Gene family construction and phylogeny analysis.** Gene families were identified using protein sequences of *T. saginata*, *T. asiatica* and 8 other worms (that is, *T. solium*, *E. granulosus*, *E. multilocularis*, *Hymenolepis microstoma*, *Schistosoma japonicum*, *S. mansoni*, *Ascaris suum* and *C. elegans*; Supplementary Table 13) by OrthoMCL (v2.0.9)<sup>61</sup> (Supplementary Figs 19 and 20). A maximum likelihood phylogenetic tree was inferred from a concatenated nucleotide data set containing 747 single-copy orthologous genes shared by all 10 species by RAxML (v8.0.24)<sup>62</sup> with the best fit model (GTR+I+4-rate  $\Gamma$  for CDS; Supplementary Methods, section 7). The divergence dates were estimated with the relaxed-clock model using BEAST2 (v2.1.3)<sup>23</sup> (Supplementary Methods, section 8).

**Gene family analysis.** Gene family expansion and contraction were determined using the CAFÉ (v3.0)<sup>63</sup> based on the phylogenetic tree constructed by RAxML (v8.0.24). The minimum ancestral gene families were determined using DOLLOP program included in the PHYLIP package (v3.695)<sup>64</sup> to estimate gain/loss evolutions of gene families.

**Paralogous gene groups and GDs.** An all-against-all protein sequence similarity search was performed using BLASTP ( $e$  value  $\leq 1e^{-10}$ ), followed by clustering the paralogous groups within each genome using Markov Clustering (MCL) (mclblastline pipeline)<sup>65</sup>. For each gene family, a protein alignment was constructed using MAFFT (v7.147b)<sup>66</sup>. This alignment was used as a guide for aligning the DNA sequences of gene family pairs, using ParaAT (v1.0)<sup>67</sup>. Paralogous gene pairs were retained if the two sequences were alignable over a length of >150 amino acids with an identity score of at least 30%.  $K_s$  value was calculated with the maximum likelihood estimation method using the program codeml (CodonFreq = 2, runmodel = -2) in the PAML package (v4.8). Only  $K_s$  values  $\leq 5$  were retained for further analysis. An average linkage clustering approach was used to correct the redundancy of  $K_s$  values that correspond to GD events. The GD modes of each genome were estimated by MSCANX. Recently, duplicated genes (in-paralogs) along *T. asiatica* and *T. saginata* that arise after their divergence were determined by Inparanoid (v4.1)<sup>68</sup>. The Tajima's relative rate test for in-paralogous gene pairs was performed, using orthologous genes of *T. solium* as outgroup (supplementary Methods, section 8).

**Likelihood ratio tests for PSGs.** A total of 6,581 one-to-one orthologous gene groups were extracted from genomes of the six tapeworms (*T. asiatica*, *T. saginata*, *T. solium*, *E. granulosus*, *E. multilocularis* and *H. microstoma*) to identify PSGs. Multiple protein-coding DNA sequence alignments were generated using ParaAT (v1.0)<sup>67</sup> and MAFFT (v7.147b)<sup>66</sup>. All gaps in the alignments were removed, and likelihood ratio tests for selection ( $P < 0.05$ ) on each branch of the phylogenetic tree were performed using Codeml implemented in the PAML package (v4.8)<sup>28</sup> with a modified branch-site model A (model = 2, NSites = 2; Supplementary Methods, section 9).

**Identification of potential drug targets.** Putative proteases and PIs were detected using the MEROPS batch BLAST server ([http://merops.sanger.ac.uk/cgi-bin/batch\\_blast](http://merops.sanger.ac.uk/cgi-bin/batch_blast)) ( $E$  value  $< 1e^{-4}$ ). LGIC genes were identified by searching tapeworm homologues against the LGIC database (<http://www.ebi.ac.uk/compneur-srv/LGICdb/>) and annotated LGIC proteins from *E. multilocularis* and *S. mansoni*<sup>13,15</sup>. The resulting hits were used as BLASTP queries against the National Center for Biotechnology Information non-redundant database. Sequences homologous to LGIC-related proteins or having no hits in the non-redundant database were retained as putative LGICs. Protein kinase domain-containing proteins were extracted from InterProScan5 domain annotations. The corresponding domains were clustered with a reference domain data set (Human, fly and *C. elegans*; KINBASE; <http://kinase.com/kinbase/FastaFiles/>) using OrthoMCL. The non-assignable domains during the clustering were searched against other tapeworm kinases for classification. TM domains were first predicted by Phobius algorithm<sup>69</sup> (length > 250 aa). The resulting proteins with  $\geq 3$  and  $\leq 15$  TM domains were retained and searched by HMMER (v3.1b1)<sup>70</sup> with HMMs against annotated GPCR sequences from Pfam database and from other tapeworms. Hits were further filtered by BLAST search against the non-redundant protein database (Supplementary Methods, section 11). Several bioinformatics tools were employed to identify classic and non-classic E/S proteins in the *T. asiatica* and *T. saginata* genomes (Supplementary Methods, section 10).

**Data availability.** All sequence data that support the findings of this study have been deposited in GenBank with the following accession numbers: LWMK00000000 and LWMJ00000000 for Whole Genome Shotgun projects of *T. saginata* and *T. asiatica* under BioProject accession PRJNA71493 and PRJNA299871, respectively; SRR2890401, SRR2890402, SRR2890403, SRR2890404

and SRR2890405 for the *T. saginata* genome sequencing data; SRR2890205, SRR2890207, SRR2890209, SRR2890210, SRR2890211, SRR2890213, SRR2890214, SRR2890215 and SRR2890216 for the *T. asiatica* genome sequencing data; SRR2895139 and SRR2895068 for the transcriptome data of *T. saginata* and *T. asiatica*, respectively. The genome assemblies and annotations used in this study are also available at <http://taenia.big.ac.cn/taenia/index.html>. All other data supporting the findings of this study are available within the article and its Supplementary Information files, or from the corresponding authors on request.

## References

- Food and Agriculture Organization of the United Nations & World Health Organization. *Multicriteria-based Ranking for Risk Management of Food-borne Parasites*. Available at [http://www.who.int/foodsafety/publications/mra\\_23/en/](http://www.who.int/foodsafety/publications/mra_23/en/) (2014).
- Hawk, M. W., Shahlaie, K., Kim, K. D. & Theis, J. H. Neurocysticercosis: a review. *Surg. Neurol.* **63**, 123–132 (2005).
- Craig, P. & Ito, A. Intestinal cestodes. *Curr. Opin. Infect. Dis.* **20**, 524–532 (2007).
- Jeon, H. K., Kim, K. H. & Eom, K. S. Complete sequence of the mitochondrial genome of *Taenia saginata*: comparison with *T. solium* and *T. asiatica*. *Parasitol. Int.* **56**, 243–246 (2007).
- Galán-Puchades, M. T. & Fuentes, M. V. Lights and shadows of the *Taenia asiatica* life cycle and pathogenicity. *Trop. Parasitol.* **3**, 114–119 (2013).
- Flisser, A. et al. Portrait of human tapeworms. *J. Parasitol.* **90**, 914–916 (2004).
- Eom, K. S. & Rim, H.-J. Natural infections of Asian *Taenia saginata* metacestodes in the livers of Korean domestic pigs. *Korean J. Parasitol.* **30**, 15–20 (1992).
- Galán-Puchades, M. T. & Fuentes, M. V. *Taenia asiatica*: the most neglected human *Taenia* and the possibility of cysticercosis. *Korean J. Parasitol.* **51**, 51–54 (2013).
- Hoberg, E. P., Alkire, N. L., de Queiroz, A. & Jones, A. Out of Africa: origins of the *Taenia* tapeworms in humans. *Proc. Biol. Sci.* **268**, 781–787 (2001).
- Hoberg, E. P. Phylogeny of *Taenia*: species definitions and origins of human parasites. *Parasitol. Int.* **55**, S23–S30 (2006).
- Michelet, L. & Dauga, C. Molecular evidence of host influences on the evolution and spread of human tapeworms. *Biol. Rev.* **87**, 731–741 (2012).
- Michelet, L. et al. The pig tapeworm *Taenia solium*, the cause of cysticercosis: biogeographic (temporal and spatial) origins in Madagascar. *Mol. Phylogenet. Evol.* **55**, 744–750 (2010).
- Tsai, I. J. et al. The genomes of four tapeworm species reveal adaptations to parasitism. *Nature* **496**, 57–63 (2013).
- Zheng, H. J. et al. The genome of the hydatid tapeworm *Echinococcus granulosus*. *Nat. Genet.* **45**, 1168–1302 (2013).
- Berriman, M. et al. The genome of the blood fluke *Schistosoma mansoni*. *Nature* **460**, 352–358 (2009).
- Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
- Hahn, C., Fromm, B. & Bachmann, L. Comparative genomics of flatworms (platyhelminthes) reveals shared genomic features of ecto- and endoparasitic neodermata. *Genome Biol. Evol.* **6**, 1105–1117 (2014).
- Zhang, J. Z. Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**, 292–298 (2003).
- Ting, C. T. et al. Gene duplication and speciation in *Drosophila*: evidence from the Odysseus locus. *Proc. Natl Acad. Sci. USA* **101**, 12232–12235 (2004).
- Rigano, R. et al. Modulation of human immune response by *Echinococcus granulosus* antigen B and its possible role in evading host defenses. *Infect. Immun.* **69**, 288–296 (2001).
- Hancock, K. et al. Characterization and cloning of GP50, a *Taenia solium* antigen diagnostic for cysticercosis. *Mol. Biochem. Parasitol.* **133**, 115–124 (2004).
- Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
- Bouckaert, R. et al. BEAST 2: a software platform for bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).
- Bradley, D. G., MacHugh, D. E., Cunningham, P. & Loftus, R. T. Mitochondrial diversity and the origins of African and European cattle. *Proc. Natl Acad. Sci. USA* **93**, 5131–5135 (1996).
- Stringer, C. B. & Andrews, P. Genetic and fossil evidence for the origin of modern humans. *Science* **239**, 1263–1268 (1988).
- Klein, R. G. Darwin and the recent African origin of modern humans. *Proc. Natl Acad. Sci. USA* **106**, 16007–16009 (2009).
- Langergraber, K. E. et al. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc. Natl Acad. Sci. USA* **109**, 15716–15721 (2012).
- Yang, Z. H. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
- Van Hellemond, J. J. et al. Functions of the tegument of schistosomes: clues from the proteome and lipidome. *Int. J. Parasitol.* **36**, 691–699 (2006).

30. Smyth, J. D. & McManus, D. P. *The Physiology and Biochemistry of Cestodes* (Cambridge Univ. Press, 1989).
31. Lumsden, R. D. The tapeworm tegument: a model system for studies on membrane structure and function in host-parasite relationships. *Trans. Am. Microsc. Soc.* **94**, 501–507 (1975).
32. Jones, M. K., Gobert, G. N., Zhang, L. H., Sunderland, P. & McManus, D. P. The cytoskeleton and motor proteins of human schistosomes and their roles in surface maintenance and host-parasite interactions. *Bioessays* **26**, 752–765 (2004).
33. Poddubnaya, L. G., Scholz, T., Kuchta, R., Levron, C. & Brunanska, M. Ultrastructure of the proglottid tegument (neodermis) of the cestode *Echinophallus wagneri* (Pseudophyllidea: Echinophallidae), a parasite of the bathypelagic fish *Centrolophus niger*. *Parasitol. Res.* **101**, 373–383 (2007).
34. Diaz, A., Casaravilla, C., Allen, J. E., Sim, R. B. & Ferreira, A. M. Understanding the laminated layer of larval *Echinococcus* II: immunology. *Trends Parasitol.* **27**, 263–272 (2011).
35. Jex, A. R. *et al.* *Ascaris suum* draft genome. *Nature* **479**, 529–533 (2011).
36. McKerrow, J. H., Caffrey, C., Kelly, B., Loke, P. & Sajid, M. Proteases in parasitic diseases. *Annu. Rev. Pathol.* **1**, 497–536 (2006).
37. McSorley, H. J., Hewitson, J. P. & Maizels, R. M. Immunomodulation by helminth parasites: defining mechanisms and mediators. *Int. J. Parasitol.* **43**, 301–310 (2013).
38. Rott, M. B. *et al.* Comparative analysis of two different subunits of antigen B from *Echinococcus granulosus*: gene sequences, expression in *Escherichia coli* and serological evaluation. *Acta Trop.* **75**, 331–340 (2000).
39. Chalmers, I. W. & Hoffmann, K. F. Platyhelminth venom allergen-like (VAL) proteins: revealing structural diversity, class-specific features and biological associations across the phylum. *Parasitology* **139**, 1231–1245 (2012).
40. Gregory, W. F. & Maizels, R. M. Cystatins from filarial parasites: evolution, adaptation and function in the host-parasite relationship. *Int. J. Biochem. Cell Biol.* **40**, 1389–1398 (2008).
41. Ale, A. *et al.* Epidemiology and genetic diversity of *Taenia asiatica*: a systematic review. *Parasite Vector* **7**, 45 (2014).
42. Jeon, H. K. & Eom, K. S. *Taenia asiatica* and *Taenia saginata*: genetic divergence estimated from their mitochondrial genomes. *Exp. Parasitol.* **113**, 58–61 (2006).
43. Gonzalez, L. M. *et al.* Genetic variability of the 18 kDa/HP6 protective antigen in *Taenia saginata* and *Taenia asiatica*: implications for vaccine development. *Mol. Biochem. Parasitol.* **176**, 131–134 (2011).
44. Bowles, J. & McManus, D. Genetic characterization of the Asian *Taenia*, a newly described taeniid cestode of humans. *Am. J. Trop. Med. Hyg.* **50**, 33–44 (1994).
45. Yamasaki, H. *et al.* DNA differential diagnosis of human taeniid cestodes by base excision sequence scanning thymine-base reader analysis with mitochondrial genes. *J. Clin. Microbiol.* **40**, 3818–3821 (2002).
46. McManus, D. P. Molecular discrimination of taeniid cestodes. *Parasitol. Int.* **55**, S31–S37 (2006).
47. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
48. Li, R. Q. *et al.* *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010).
49. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
50. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome Biol.* **13**, R56 (2012).
51. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
52. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
53. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
54. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVIDENCEModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
55. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
56. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
57. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
58. Wang, Y. P. *et al.* MCSanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
59. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* **9**, 357–359 (2012).
60. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
61. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).
62. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006).
63. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1271 (2006).
64. Felsenstein, J. *PHYLIP: Phylogenetic Inference Program, Version 3.6* (Univ. of Washington, 2005).
65. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584 (2002).
66. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
67. Zhang, Z. *et al.* ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.* **419**, 779–781 (2012).
68. O'Brien, K. P., Remm, M. & Sonnhammer, E. L. L. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33**, D476–D480 (2005).
69. Kall, L., Krogh, A. & Sonnhammer, E. L. L. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* **35**, W429–W432 (2007).
70. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).

## Acknowledgements

We thank H. Li from Dali University for providing parasite material; Y. Liu, M. Sun, W. Li, J. Yang and S. Li from Beijing Institute of Genomics, Chinese Academy of Sciences for preparing genomic DNA, constructing libraries and generating raw sequencing reads; and Y. Zheng for helpful discussions. This work was funded by the 'National Key Basic Research Program (973 Program) of China' (grant no. 2015CB150300), 'International Science & Technology Cooperation Program of China' (grant no. 2013DFA31840) and State Key Laboratory of Veterinary Etiological Biology of Lanzhou Veterinary Research Institute, Chinese Academy of Agricultural Sciences.

## Author contributions

Sh.W., J.H., S.Z., X.L., H.L., Y.D., X.Z. and X.M. collected, cultivated and prepared parasite material for sequencing. B.Z., X.T., Q.M. and J.Y. purified nucleic acids and coordinated sequencing. Se.W., Sh.W., G.Z., S.H. and X.C. planned and undertook the assembly, annotation and comparative analyses of genomic and transcriptomic data. Y.L. and S.G. improved the genome assemblies and curated genes. Sh.W., Se.W. and G.Z. performed the phylogenetic and adaptive evolution analyses. H.Z., S.G., H.M., Z.H. and M.Y. created the *Taenia* genome database. Sh.W., A.G., J.C. and X.-Q.Z. analysed drug targets. Sh.W., Se.W., G.Z. and L.X. wrote the manuscript. X.C., G.Z. and S.H. directed the project.

## Additional information

**Supplementary Information** accompanies this paper at <http://www.nature.com/naturecommunications>

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**How to cite this article:** Wang, S. *et al.* Comparative genomics reveals adaptive evolution of Asian tapeworm in switching to a new intermediate host. *Nat. Commun.* **7**:12845 doi: 10.1038/ncomms12845 (2016).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016

# Erratum: Comparative genomics reveals adaptive evolution of Asian tapeworm in switching to a new intermediate host

Shuai Wang, Sen Wang, Yingfeng Luo, Lihua Xiao, Xuenong Luo, Shenghan Gao, Yongxi Dou, Huangkai Zhang, Aijiang Guo, Qingshu Meng, Junling Hou, Bing Zhang, Shaohua Zhang, Meng Yang, Xuelian Meng, Hailiang Mei, Hui Li, Zilong He, Xueliang Zhu, Xinyu Tan, Xing-quan Zhu, Jun Yu, Jianping Cai, Guan Zhu, Songnian Hu & Xuepeng Cai

*Nature Communications* 7:12485 doi: 10.1038/ncomms12845 (2016); Published 22 Sep 2016; Updated 3 Nov 2016

This Article contains errors in Fig. 3. In panel b, the axes are displaced relative to the data, and in panel c, the label 'Species-specific GD' in red should read 'Species-specific GD (tandem)'. The correct version of Fig. 3 appears below as Fig. 1.

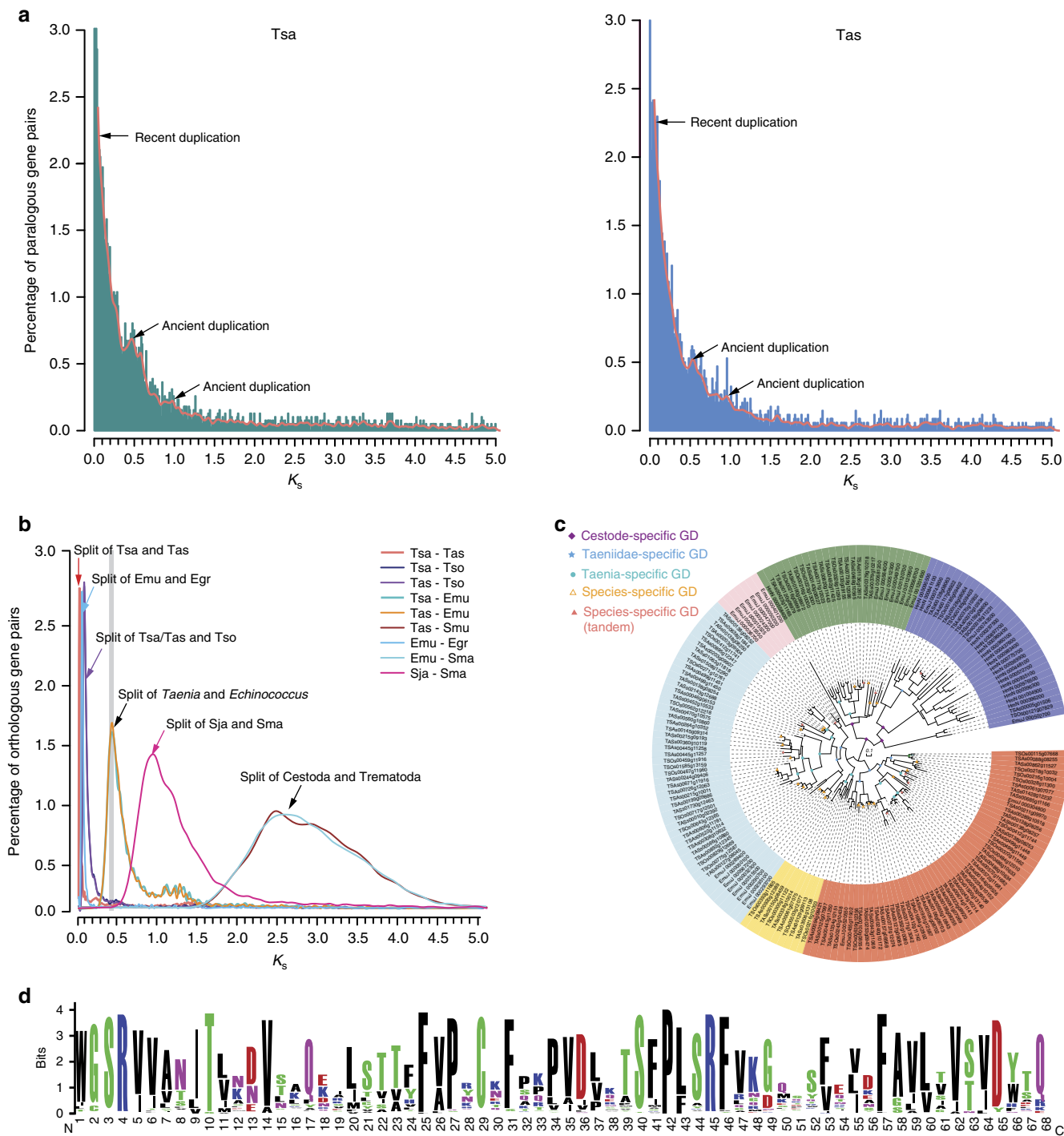


Figure 1

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

© The Author(s) 2016