High-quality reference genome of *Fasciola gigantica*: Insights into the genomic signatures of transposon-mediated evolution and specific parasitic adaption in tropical regions

Xier Luo*[1,2], Kuiqing Cui*[1],Zhiqiang Wang*[1], Lijuan Yin[2], Zhipeng Li[1], Zhengjiao Wu[1], Tong Feng[1,2], Xiaobo Wang[1,2], Weikun Jin[1], Wenda Di[1], Dongying Wang[1], Saif ur Rehman[1], Weiyi Huang[1], Xingquan Zhu[3], Weiyu Zhang[†1], Jue Ruan[†2], Qingyou Liu[†1]

1. State Key Laboratory for Conservation and Utilization of Subtropical Agro- bioresources,Guangxi University, Nanning 530004, China

2. Guangdong Laboratory for Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China.

3. College of Veterinary Medicine, Shanxi Agricultural University, Taigu, 030801, Shanxi, People's Republic of China.

*These authors contributed equally:Xier Luo, Kuiqing Cui, Zhiqiang Wang.

†Corresponding author. E-mail: zweiyu@gxu.edu.cn; ruanjue@caas.cn; qyliu-gene@gxu.edu.cn.

Abstract

Fasciola gigantica and *Fasciola hepatica* are causative pathogens of *fascioliasis*, with the widest latitudinal, longitudinal, and altitudinal distribution; however, among parasites, they have the largest sequenced genomes, hindering genomic research. In the present study, we used various sequencing and assembly technologies to generate a new high-quality *Fasciola gigantica* reference genome. We improved the integration of gene structure prediction, and identified two independent transposable element expansion events contributing to (1) the speciation between Fasciola and Fasciolopsis during the Cretaceous-Paleogene boundary mass extinction, and (2) the habitat switch to the liver during the Paleocene-Eocene Thermal Maximum, accompanied by gene length increment. Long interspersed element (LINE) duplication contributed to the second transposon-mediated alteration, showing an obvious trend of insertion into gene regions, regardless of strong purifying selection. Gene ontology analysis of genes with long LINE insertions identified membrane-associated and vesicle secretion process proteins, further implicating the functional alteration of the gene network. We identified 852 excretory/secretory proteins and 3300 protein-protein interactions between *Fasciola gigantica* and its host. Among them, copper/zinc superoxide dismutase genes, with specific gene copy number variations, might play a central role in the phase I detoxification process. Analysis of 559 single-copy orthologs suggested that *Fasciola gigantica* and *Fasciola hepatica* diverged at 11.8 Ma near the Middle and Late Miocene Epoch boundary. We identified 98 rapidly evolving gene families, including actin and aquaporin, which might explain the large body size and the parasitic adaptive character resulting in these liver flukes becoming epidemic in tropical and subtropical regions.

Introduction

Fasciola gigantica and *Fasciola hepatica*, known as liver flukes, are two species in the genus *Fasciola*, which cause *fascioliasis* commonly in domestic and wild ruminants, but also are causal agents of *fascioliasis* in humans. *Fascioliasis* reduces the productivity of animal industries, imposes an economic burden of at least 3.2 billion dollars annually worldwide [1], and is a neglected zoonotic tropical disease of humans, according the World Health Organization’s list [2]. *F. gigantica*, the major fluke infecting ruminants in Asia and Africa, has been a serious threat to the farming of domesticated animals, such as cows and buffaloes, and dramatically reduces their feed conversion efficiency and reproduction [3]. The prevalence *F. gigantica* infection has greatly affected subsistence farmers, who have limited resources to treat their herds, and has hindered economic development and health levels, especially in developing countries.

The various omics technologies provide powerful tools to advance our understanding of the molecules that act at the host-parasite interface, and allow the identification of new therapeutic targets against *fascioliasis* [4]. To date, four assemblies for *F. hepatica* and two assemblies for *F. gigantica* have been deposited at the NCBI [5-8]. These assemblies reveal a large genome with a high percentage of repeat regions in *Fasciola* species, and provided valuable insights into features of adaptation and evolution. However, these assemblies are based on the short read Illumina sequencing or hybrid sequencing methods, with limited ability to span large families of repeats. Various limitations have led to the current assemblies in the genus *Fasciola* being fragmented (8 kb to 33 kb and 128 kb to 1.9 Mb for contig and scaffold N50s, respectively). Subsequent gene annotation analysis using current assemblies were also challenging, with abundant transposition events occurring over evolutionary history, which significantly increased the repeat components in intron regions, resulting in considerable fragmentation in gene annotation.

Infection by Fasciola causes extensive damage to the liver, and excretory/secretory (E/S) proteins play an important role in host-parasite interactions. Parasite-derived molecules interact with proteins from the host cell to generate a protein interaction network, and these proteins partly contribute to Fasciola's striking ability to avoid and modulate the host's immune response [9]. Previous proteomics of E/S proteins have highlighted the importance of secreted extracellular vesicles (EVs) and detoxification enzymes to modulate host immunity by internalizing with host immune cells [10, 11]. The anthelminthic drug, triclabendazole (*TCBZ*), is currently the major drug available to treat *fascioliasis* at the early and adult stages, which acts by disrupting β-tubulin polymerization [1]; however, over-reliance on *TCBZ* to treat domesticated ruminants has resulted in selection for resistance to liver flukes [12]. Drug and vaccine targets for molecules associated with reactive oxygen species (ROS)-mediated apoptosis have recently been validated as an effective tools in multiple helminth parasites [13]. Increased understanding of host-parasite and drug-parasite interactions would facilitate the development of novel strategies to control *fascioliasis*.

In recent years, there have been increasing numbers of human cases of *fascioliasis*, becoming a major public health concern in many regions [14, 15]. However, high quality genome assemblies for liver flukes are still insufficient. In the present study, we combined multiple sequencing technologies to assemble a chromosome-level genome for *F. gigantica* and provided integrated gene annotation. Protein-protein interactions were analyzed between the predicted *F. gigantica* secretome and host proteins expressed in the small intestine and liver. In addition, gene family analysis identified a series of genes expansions in *F. gigantica*. Interestingly, the distribution of repeat sequences in the genome exhibit an excess of long interspersed element (LINE) duplications inserted into intronic regions, potentially helping to explain the duplications of transposable element (TE) plasticizing gene structures and possibly acting as long-term agents in the speciation of Fasciola.

Results

Pacbio long reads-based *de novo* assembly and gene annotation

The *F. gigantica* genome contains abundant repeat sequences that are difficult to span using short read assembly methods, and the complex regions also hinder integrated gene annotation of the genome. Therefore, in the present study, multiple sequencing technologies, have been applied: (1) Single-molecule sequencing long reads (~91× depth) using the Pacbio Sequel II platform; (2) paired-end reads (~66× depth) using the Illumina platform; and (3) chromosome conformation capture sequencing (Hi-C) data (~100× depth) (Supplementary Table 1). The initial assembly was performed using the Pacbio long reads, followed by mapping using single-molecule sequencing and Illumina sequencing reads to polish assembly errors and sequencing mistakes, resulting in a contig N50 size of 4.89 Mb (**Fig. 1A**). The Hi-C data were used to build final super-scaffolds, resulting in a total length of 1.35 Gb with a scaffold N50 size of 133 Mb (**Fig. 1B, Table 1, Supplementary Table S2-3, Supplementary Fig. S1**). The final assembly consists of 10 pseudo-chromosomes covering more than 99.9% of the *F. gigantica* genome, and the length distribution was approximate equal to the estimation by karyotype in previous research (**Supplementary Fig. S2, Supplementary Table S4**) [16]. The assessment of nucleotide accuracy shows that the error rate was 5.7×10^{-6} in the genome. QUAST analysis [17] showed a high mapping and coverage rate using both Illumina short reads and Pacbio long reads, in which 99.73% of reads mapped to 99.85% of the genome with more than 10× depth (**Supplementary Table S5**).

Combing *de novo*/homolog/RNA-seq prediction, a total of 12,503 protein coding genes were annotated in the *F. gigantica* genome. BUSCO assessment [18] indicated that the genome is 90.4% complete and 5.6% fragmented, underscoring the significant improvement of the genome continuity and gene-structure predictions compared with previous assemblies (**Supplementary Table S6**). Specifically, the average gene length in the annotated data is 28.8 kb, nearly twice the length of that in other digenean species, but contrasted with the similar average length of the coding sequences (CDSs). Through functional annotation, we found that 8569 of the genes could be characterized in the InterPro database [19, 20], 7892 of them were mapped to the gene ontology (GO) terms, and 5353 of them were identified by the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways database (**Supplementary Fig. S3-4, Supplementary Table S7**).

The unique repeat duplications in Fasciola

TEs are insertional mutagens and major drivers of genome evolution in eukaryotes, and replication of these sequences, resulting in variation of gene structure and expression, have been extensively documented [21, 22]. Besides, TEs are molecular fossils, being remnants of past mobilization waves that occurred millions of years ago [23]. In the present study, we identified repeat sequences combined the analysis from RepeatModeler [24] and RepeatMasker [25], and detected a significant proportion of them neglected by previous studies. In the *F. gigantica* genome, we identified 945 Mb of repeat sequences, which was approximate 20% more than that identified in other assemblies in *Fasciola* species, while the lengths of non-repeat sequences were nearly identical. The most convincing explanation for the additional assembled repeat sequences was that the contigs constructed from Pacbio long reads spanned longer repeat regions, which were compressed in previous assemblies. Among these repeat sequences, there were 408 Mb of LINEs (corresponding to 30.3% of the assembled genome), 285 Mb of long terminal repeats (LTRs, corresponding to 21.2% of the assembled genome), and 162 Mb of unclassified interspersed repeats (corresponding to 12.0% of the assembled genome) (**Supplementary Fig. S5, Supplementary Table S8**). According to the repeat landscapes, we found that there were two shared expansion events for LINEs and LTRs that occurred approximately 12 million years ago (Ma) and 65 Ma, and an additional expansion event at 33 Ma for LTRs (**Supplementary Fig. S6-7)**. The abundant repeat sequences in the Fasciola genomes aroused our interest concerning the role of repeats in evolution (**Fig. 2A**), and inspired us to hypothesize that the expansion of TEs enlarged the genome size of an ancestor of Fasciola to gain a new advantage by rewiring gene networks. To test this hypothesis, we focused on the genome-wide repeats distribution and test signatures of selection.

For new TE insertions to persist through vertical inheritance, transposition events must be under strong purifying selection among gene loci to avoid disturbing their biological function. However, we observed many intronic repeat elements in Fasciola, resulting in a larger intron size per gene. If there are equal selection effects on newly inserted TEs in intronic and intergenic regions, there would a high correlation between the distribution of insertion time and retained TE lengths between these two regions. By contrast, there would be fewer accumulated repeat sequences existing under purifying selection. In this study, we use the relative proportion of TEs between intronic and intergenic regions as a simple indicator, and use the inferred size of intronic and intergenic regions over evolutionary history as a control to estimate the signatures of selection. The results showed that TE insertions into intronic regions are under persistent intense purifying selection, except for LINEs. There was an excess of persistent LINE insertions into intronic regions between 41 Ma and 62 Ma, indicating different modes of accumulating LINEs into intronic regions compared with that in other periods (**Fig. 2B**). Specifically, the time of the ancient intronic LINE expansion (~51.5 Ma) was different to the genome-wide LINE expansion time (~68.0 Ma), whereas the time was coincident with two important environmental change events, the Cretaceous-Paleogene boundary (KPB) mass extinction (~66.0 Ma) and the Paleocene-Eocene Thermal Maximum (PETM) (~55.8 Ma). Both the PETM and KPB events recorded extreme and rapid warming climate changes; however, rapid evolutionary diversification followed the PETM event, as opposed to near total mass extinction at the KPB [26]. Therefore, we selected genes with different LINE lengths, derived between 41 Ma and 62 Ma, and expected to identify a transposon-mediated alterative gene network contributing to the host switch and the shift from intestinal to hepatic habitats.

LINE-mediated alterative gene network

We identified a substantial proportion of genes with LINE insertions, derived between 41 Ma and 62 Ma, indicating a universal effect of the gene network. We selected 1288 genes with the LINE insertions of more than 10 kb, representing more than one third of the average gene length, and annotated the genes using Gene Ontology (GO) terms and processes and Kyoto encyclopedia of genes and genomes (KEGG) pathways (**Fig. 2C, Supplementary Table S9-11**). These genes involve molecules internalizing substances from their external environment, including membrane-associated and vesicle secretion process proteins. Meanwhile, the gene network was likely adapted to the evolution of protein biosynthesis and modification of histones.

Enrichment analysis of GO terms showed that membrane and membrane-associated proteins are over-represented, involving “synaptic membrane” (*P* = 3.52E-04), “clathrin-coated vesicle membrane” (*P* = 1.08E-03), and “synaptic vesicle” (*P* = 3.02E-03), as well as vesicles secretion processes, such as “endocytosis” (*P* = 7.06E-06), “Golgi organization” (*P* = 7.45E-05), “COPII vesicle coating” (*P* = 2.72E-04), “intracellular signal transduction” (*P* = 5.16E-04), and “endosomal transport” (*P* =

2.47E-03). The over-representation of genes involved in membrane transport was particularly interesting because helminth parasites interfere with the host immune system by secreting molecules from surface tegument or gut. The *TMED10* gene in *F. gigantica* (encoding transmembrane P24 trafficking protein 10) was used as an example. *TMED10* is a cargo receptor involved in protein vesicular trafficking along the secretory pathway [27, 28], and the genes has an 11.1 kb LINE insertion in the third intron, resulting in an over three-fold increment in the gene length (**Fig. 2D**). The enrichment suggests that the gene network related to secretion could have experienced adaptive evolution during LINE transposition events. We further compared our dataset with the proteome result from *F. hepatica* extracellular vesicles (EVs) [29], and found 21 proteins that were also identified as surface molecules associated with EV biogenesis and vesicle trafficking (*IST1*, *VPS4B*, *TSG101*, *MYOF*, *ATG2B*, *STXBP5L*, and 15 Rho GTPase-activating related proteins). Specifically, *IST1*, *VPS4B*, and *TSG101* are members of the endosomal sorting complex required for transport (ESCRT) pathway, which promotes the budding and release of EVs. *TSG101*, a crucial member of the ESCRT-I complex, has an important role in mediating the biogenesis of multi-vesicular bodies, cargo degradation, and recycling of membrane receptors. Besides, the ESCRT pathway promotes the formation of both exosomal carriers for immune communication. During the formation of the immunological synapse between T-cells and antigen-presenting B cells, *TSG101* ensures the ubiquitin-dependent sorting of T-Cell Receptor (*TCR*) molecules to exosomes that undergo *VPS4*-dependent release into the synaptic cleft[30].

The most significant KEGG pathway was aminoacyl-tRNA biosynthesis (*P* = 7.16E-04), containing 15 out of 38 annotated aminoacyl tRNA synthetases (*AAASs*). *AARSs* are the enzymes that catalyze the aminoacylation reaction by covalently linking an amino acid to its cognate tRNA in the first step of protein translation. The large-scale insertion of LINEs reside in *AAAS* genes suggested that the ancestor of Fasciola may have profited from the effect of transposition, with changes to protein biosynthesis and several metabolic pathways for cell viability. In addition, a significant number of genes are strongly associated with histone modulation, including “histone deacetylase complex” (*P* = 1.89E-03), “histone methyltransferase activity (H3-K36 specific)” (*P* = 1.08E-03), and “methylated histone binding” (*P* = 2.37E-03). Histone modifications play fundamental roles in the manipulation and expression of DNA. We found nine histone deacetylases and Histone methyltransferases in the gene set (*HDAC4*, *HDAC8*, *HDAC10*, *KMT2E*, *KMT2H*, *KMT3A*, *KDM8*, *NSD1*, and *NSD3*). Histone modifications can exert their effects by influencing the overall structure of chromatin and modifying and regulating the binding of effector molecules [31, 32]; therefore, the variation of these genes might bring about evolution from a disturbed gene structure to a mechanism of genome stabilization to tackle a continuous genome amplification process in evolutionary history.

Genome-wide host-parasite interaction analysis

In the Fasciola genome, we predicted genes encoding 268 proteases, 36 protease inhibitors (PIs), and 852 excretory/secretory (E/S) proteins that are commonly involved in interacting with hosts and modulating host immune responses. The largest class of proteases was cysteine peptidases (n = 113), which was also identified in the *F. hepatica* genome (**Fig. 3A, Supplementary Table S12**). The largest (n = 19, 52.8% of PIs) PI family was the I02 family of Kunitz-BPTI serine protease inhibitors, which bind to Cathepsin L with a possible immunoregulatory function [33] (**Supplementary Table S13**). GO enrichment analysis of E/S proteins showed that proteins related to "activation of cysteine-type endopeptidase activity" (P = 6.14E-19), "peroxidase activity" (P = 3.79E-07) and "protein disulfide isomerase activity" (P = 3.75E-06) are over-represented (**Fig. 3B**, **Supplementary Table S14-15**). Indeed, there were 38 cysteine peptidases identified as E/S proteins, including cathepsin L-like, cathepsin B-like, and legumain proteins, which participate in excystment, migration through gut wall, and immune evasion [34].

In parasites, as in mammalian cells, ROS are produced as a by-product of cell metabolism and from the metabolism of certain pharmacological agents. The ability of a parasite to survive in its host has been directly related to its antioxidant enzyme content [35]. To further analyze host-parasite interactions, we identified the protein-protein interactions (PPIs) between the *F. gigantica* secretome and human proteins expressed in the small intestine and liver. In total, we identified 3300 PPIs, including rich interactions that directly or indirectly participated in the two phases of detoxification pathways (**Fig. 3C**). Superoxide dismutase [Cu-Zn] (*SOD*, PPIs = 49) was first highlighted because of its important role on phase I detoxification against ROS, in which it catalyzes the dismutation of the superoxide radical to molecular oxygen and hydrogen peroxide (H_2O_2) [36]. Gene family analysis identified six *SOD* paralogs in *F. gigantica*, and two of them contained a signal peptide (**Fig. 4D**). Previous enzyme activity assays also confirmed a significant difference between *SOD* activities and concentration in E/S proteins of two *Fasciola* species [37], suggesting an intense ability to resist superoxide radical toxicity. Meanwhile, the metabolite of phase I, H_2O_2, can also damage parasites, which requires detoxification enzymes, including glutathione-dependent enzymes *GPx*, glutathione reductase, and other peroxidases. Protein disulfide-isomerase (*P4HB*, PPIs = 132) and phospholipid hydroperoxide glutathione peroxidase (*GPX4*, PPIs=28) were as functioning in phase II detoxification. GPx catalyzes the reduction of hydroperoxides (ROOH) to water, using glutathione (*GSH*) as the reductant. P4HB also participates in the process by mediating homeostasis of the antioxidant glutathione [38]. However, we did not identify E/S proteins in the Cytochrome P450 (*CYP450*) family in phase III detoxification. Therefore, we speculated that successful parasite defense against *F. gigantica* is mainly depends on the strong superoxide activity and efficient hydrogen peroxide detoxification.

Gene family analysis

Gene family analysis was performed using eight taxa (*F. gigantica*, *F. hepatica*, *Fasciolopsis buski*[39], *Clonorchis sinensis* [40], *Schistosoma mansoni*)[41], *Taenia multiceps* [42], swamp buffalo [43], and human [44], which identified 17,992 gene families (**Fig. 4A**). Phylogeny analysis of 559 single-copy orthologs showed that *F. gigantica* and *F. hepatica* shared a common ancestor approximately 11.8 million years ago (2.2-22.5 Ma, 95% highest posterior density [HPD]) near the Middle and Late Miocene Epoch boundary. The Miocene warming began 21 million years ago and continued until 14 million years ago, when global temperatures took a sharp drop at the Middle Miocene Climate Transition (MMCT). The divergence of the two *Fasciola* species may have resulted from the consequences of rapid climate changes, such as migration of the host causing geographic isolation. Our estimation is between the previously suggested date of 5.3 Ma based on 30 nuclear protein-coding genes [45], and 19 Ma based on cathepsin L-like cysteine proteases [46]. Although we used a more integrative gene dataset, the wide HPD interval could not be neglected, raising possible uncertainty from the complex process of speciation or inappropriate protein sequence alignment between members of the genus *Fasciola*.

The distribution of gene family size among different species is used to estimate which lineages underwent significant contractions or expansions. Compared with *F. hepatica*, *F. gigantica* shows more gene family expansion events (643 compared to 449) and a similar number of gene family contractions (713 compared to 672). The result emphasize the general trend that, relative to the common ancestor of Fasciola, the ancestor of *F. gigantica* apparently underwent a higher extent of gene-expansion than did the ancestor of *F. hepatica*. Gene duplication is one of the primary contributors to the acquisition of new functions and physiology [47]. We identified 98 gene families, including 629 genes, as rapidly evolving families specific to *F. gigantica*. Family analysis showed a fascinating trend of gene duplication, with substantial enrichment for the "structural constituent of cytoskeleton" (*P* = 3.52E-24), "sarcomere organization" (*P* = 2.29E-14), "actin filament capping" (*P* = 6.19E-13), and "spectrin" (*P* = 3.03E-11) in *F. gigantica* (**Supplementary Table S16**). There were 24 actin paralogs in *F. gigantica*, in contrast to 8 actin paralogs in *F. hepatica*. Actin is one of the most abundant proteins in most cells, and actin filaments, one of the three major cytoskeletal polymers, provide structure and support internal movements of organisms [48]. They are also highly conserved, varying by only a few amino acids between algae, amoeba, fungi, and animals [49]. We observed three types of actin proteins in flukes, according to their identity from human actin family. Seventeen of the 24 actin proteins in *F. gigantica* are highly conserved (Identity > 95%) (**Fig. 4B**). Consistent with the accepted role of the epidermal actin cytoskeleton in embryonic elongation [50, 51], we speculated that the significant expansion of actin and spectrin genes increased the body size of *F. gigantica* via cell elongation or proliferation during morphogenesis. Another rapidly evolving family is the aquaglyceroporin subfamily in the membrane water channel family. We found six aquaglyceroporin paralogs in *F. gigantica*, which were over-represented in the GO term "water transport" (*P* = 2.10E-06) (**Fig. 4C**). Aquaglyceroporins are highly permeated by glycerol and other solutes, and variably permeated by water, as functionally validated by several studies [52, 53]. The mammalian aquaglyceroporins regulate glycerol content in epidermal, fat, and other tissues, and appear to be involved in skin hydration, cell proliferation, carcinogenesis, and fat metabolism. A previous study showed that *F. gigantica* could withstand a wider range of osmotic pressures compared with *F. hepatica* [54], and we speculated that a higher aquaglyceroporin gene copy number might help explain this observation.

It is worth mentioning that 57.6% of rapidly evolving expansion genes specific to the *F. gigantica* genome were driven by tandem duplication, such that the newly formed duplicates preserved nearly identical sequences to the original genes. The newly formed genes would accumulate non-functionalizing mutations, or develop new functions over time. We found only few tandem duplicated genes that had non-functionalizing mutations, suggesting that adaptive evolution could have an important role in the consequences of these genes via a dosage effect or neo-functionalization.

Discussion

The genome of *Fasciola* species contains a large percentage of repeat sequences, making them the largest parasite genomes sequenced to date. Since the first assembly of *F. hepatica* was submitted in 2015 [6], several studies have aimed to improve the quality of assembly and gene annotation [5, 7, 8]. With advances in long read sequencing assembly and Hi-C scaffolding technologies, it is now viable to resolve the genomic "dark matter" of repetitive sequences, and other complex structural regions at relatively low cost [55]. Therefore, we present the highest quality genome and gene annotation for *F. gigantica* to date, and provide long-awaited integrated genome annotation for *fascioliasis* research.

Our research determined the TE sequences among intronic and intergenic regions. TE sequences of *F. gigantica* experienced massive expansion through the genome via a 'copy-and-paste' model of transposition [56]. Especially, the speciation between Fasciola and Fasciolopsis was most likely caused by a Fasciola-specific whole genome repeat expansion event during the KPB mass extinction, and similarly, the speciation between the Fasciola and Fascioloides—a habitat switch from the small intestine to the liver in the host—occurred during the PETM, accompanied by LINE expansion biased toward intronic regions (**Fig. 5**). These synchronous events informed a new hypothesis of adaptive evolution driven by transposition events and will prompt investigations of how such differences contribute mechanistically to the morphological phenotypes of liver flukes and related species. This hypothesis could be tested by targeted genome assembly of Fascioloides species and estimating whether they had a different pattern of LINE duplication among intronic regions. There are also many studies in other species supporting the hypothesis that TE invasions endured by organisms have catalyzed the evolution of gene-regulatory network [57]. For example, Eutherian-specific TEs have the epigenetic signatures of enhancers, insulators, and repressors, and bind directly to transcription factors that are essential for pregnancy and coordinately regulate gene expression [58]. Similarly, genes with large-scale insertion of TEs in *Fasciola* species identified here, represent a signature of Fasciola-specific evolutionary gene network to distinguish other flukes of the family Fasciolidae. These genes overlap significantly with host-parasite interaction genes, including proteases and E/S proteins, and are enriched in the pathways of EV biogenesis and vesicle trafficking.

The data from genomic, transcriptomic, and proteomic studies can form a good complementary relationship to further our understanding of helminth parasites and their interaction with their hosts. Previous studies have identified a rich source of stage-specific molecules of interest using transcriptomic and proteomic analysis [59, 60]. Here, we provided a comprehensive list of predicted E/S proteins in *F. gigantica* and predicted 3300 PPIs at the host-parasite interface, extending our understanding of how the phase I and phase II detoxification enzymes counteract the effect of ROS. The ability of *Fasciola* species to infect and survive in different tissue environments is underpinned by several key E/S protein gene duplications. Both *Fasciola* species have a common expansion in the secretion of papain-like cysteine peptidase family (Clan A, family C1) [6]. Besides, *F. gigantica* has a specific variation in the *SOD* gene copy number, allowing it to regulate the catalytic activity of the superoxide radical released by the host. The effect of specific gene duplications can also be reflect in the increased body size of *F. gigantica*, which is an important morphometric character to distinguish *Fasciola* species and has a decisive influence on the final host species [61], although a gene level study of this phenotype is barely reported.

Overall, our study demonstrated that the combination of long-read sequencing with Hi-C scaffolding produced a very high-quality liver fluke genome assembly and gene annotation. Additionally, identification of the repeat distribution among the gene regions extended our understanding of the evolutionary process in Fasciola species. Further detailed functional studies of secretion might be of great scientific significance to explore their potential application in *fascioliasis* treatment.

Materials and Methods

Sample collection and *de novo* sequencing.

All animal work was approved by the Guangxi University Institutional Animal Care and Use Committee. For the reference genome sequencing, *F. gigantica* was derived from infected buffalo in the Guangxi Zhuang Autonomous Region. Nucleic acids were extracted using a QIAGEN DNeasy (DNA) kit (Qiagen Hilden, Germany). Three *de novo* genome sequencing methods were performed on the liver fluke: We generated (1) 122.4 Gb (~88× depth) PacBio Sequel II single-molecule long reads, with an average read length of 15.8 kb (PacBio, Menlo Park, CA, USA); (2) 89.5 Gb (~66× depth) Illumina HiSeq PE150 pair-end sequencing to correct errors (Illumina, San Diego, CA, USA); and (3) 134 Gb (~100× depth) chromosome conformation capture sequencing (Hi-C) data (sequenced by Illumina platform).

De novo assembly and assessment of the genome quality.

A PacBio-only assembly was performed using Canu v2.0 [62, 63] using new overlapping and assembly algorithms, including an adaptive overlapping strategy based on *tf-idf* weighted MinHash and a sparse assembly graph construction that avoids collapsing diverged repeats and haplotypes. To remove haplotigs and contig overlaps in the assembly, we used Purge_Dups based on the read depth [64]. Arrow (https://github.com/PacificBiosciences/GenomicConsensus) was initially used to reduce the assembly error in the draft assembly, with an improved consensus model based on a more straightforward hidden Markov model approach. Pilon [65] was used to improve the local base accuracy of the contigs via analysis of the read alignment information based on paired-end bam files (thrice). As a result, the initial assembly resulted had an N50 size of 4.89 Mb for the *F. gigantica* reference genome. ALLHiC was capable of building chromosomal-scale scaffolds for the initial genome using Hi-C paired-end reads containing putative restriction enzyme site information [66].

Three methods were used to evaluate the quality of the genomes. First, we used

QUality ASsessment Tool (QUAST) [67] to align the Illumina and PacBio raw reads to the *F. gigantica* reference genome to estimate the coverage and mapping rate. Second, all the Illumina paired-end reads were mapped to the final genome using BWA [68], and single nucleotide polymorphisms (SNPs) were called using Samtools and Bcftools [69]. The predicted error rate was calculated by the homozygous substitutions divided by length of the whole genome, which included the discrepancy between assembly and sequencing data. Thirdly, we assessed the completeness of the genome assemblies and annotated the genes using BUSCO [18].

Genome annotation

Three gene prediction methods, based on *de novo* prediction, homologous genes, and transcriptomes, were integrated to annotate protein-coding genes. RNA-seq data of *F. gigantica* were obtained from the NCBI Sequence Read Archive, SRR4449208 [70]. RNA-seq reads were aligned to the genome assembly using HISAT2 (v2.2.0) [71] and subsequently assembled using StringTie (v2.1.3) [72]. PASA (v2.4) [73] was another tool used to assemble RNA-seq reads and further generated gene models to train *de novo* programs. Two *de novo* programs, including Augustus (v3.0.2) [74] and SNAP (v2006-07-28) [75], were used to predict genes in the repeat-masked genome sequences. For homology-based prediction, protein sequences from UniRef100 [76] (plagiorchiida-specific, n = 75,612) were aligned on the genome sequence using TBLASTn [77] (e-value < 10^{-4}), and GeneWise (version 2.4.1) [78] was used to identify accurate gene structures. All predicted genes from the three approaches were combined using MAKER (v3.1.2) [79] to generate high-confidence gene sets. To obtain gene function annotations, Interproscan (v5.45) [80] was used to identify annotated genes features, including protein families, domains, functional sites, and GO terms from the InterPro database. SwissProt and TrEMBL protein databases were also searched using BLASTp [81] (e-value < 10^{-4}). The best BLASTp hits were used to assign homology-based gene functions. BlastKOALA [82] was used to search the KEGG ORTHOLOGY (KO) database. The subsequent enrichment analysis was performed using clusterProfiler using total annotated genes as the background with the “enricher” function [83].

Repeat annotation and analysis

We combined *de novo* and homology approaches to identify repetitive sequences in our assembly and previous published assemblies, including *F. gigantica*, *F. hepatica*, and *Fasciolopsis buski*. RepeatModeler (v2.0.1) [24] was first used to construct the *de novo* identification and accurate compilation of sequence models representing all of the unique TE families dispersed in the genome. Then, RepeatMasker (v4.1.0) [25] was run on the genome using the combination of *de novo* libraries and a library of known repeats (Repbase-20181026). The relative position between a repeat and a gene was identified using bedtools [84], and the type of repeat was further divided to intronic and intergenic origin. The repeat landscape was constructed using sequence alignments and the complete annotations output from RepeatMasker, depicting the Kimura divergence (Kimura genetic distances between identified repeat sequences and their consensus) distribution of all repeats types. The most notable peak in the repeat landscapes was considered as the most convincing time of repeat duplication in that period. The transition between the Kimura divergence and age was performed by dividing the divergence by the two-fold mutation rate per year (T = d/2mu). The mutation rate (mu = 1.73×10^{-9}) was calculated using MCMCTree [85] based on the CDS sequence alignment of single-copy gene families.

Genome-wide host-parasite protein interaction analysis

In addition to the genome data that we generated for *F. gigantica*, we downloaded genome annotation information for human (GCA_000001405.28), swamp buffalo (GWHAAJZ00000000), *F. hepatica (*GCA_002763495.2), *Fasciolopsis buski (*GCA_008360955.1), *Clonorchis sinensis* (GCA_003604175.1), *Schistosoma mansoni* (GCA_000237925.2), and *Taenia multiceps* (GCA_001923025.3) from the NCBI database and BIG Sub (China National Center for Bioinformation, Beijing, China). Proteases and protease inhibitors were identified and classified into families using BLASTp (e-value < 10^{-4}) against the MEROPS peptidase database (merops_scan.lib; (European Bioinformatics Institute (EMBL-EBI), Cambridge, UK)), with amino acids at least 80% coverage matched for database proteins. These proteases were divided into five major classes (aspartic, cysteine, metallo, serine, and threonine proteases). E/S proteins (i.e., the secretome) were predicted by the programs SignalP 5.0 [86], TargetP [87], and TMHMM [88]. Proteins with a signal peptide sequence but without a transmembrane region were identified as secretome proteins, excluding the mitochondrial sequences. Genome-wide host-parasite protein interaction analysis was perform by constructing the PPIs between the *F. gigantica* secretome and human proteins expressed in the tissues related to the liver fluke life cycle. For the hosts, we selected human proteins expressed in the small intestine and liver, and located in the plasma membrane and extracellular region. The gene expression and subcellular location information were obtained from the TISSUES [89] and Uniprot (EMBL-EBI) databases, respectively. For *F. gigantica*, secretome molecules were mapped to the human proteome as the reference, using the reciprocal best-hit BLAST method. These two gene datasets were used to construct host-parasite PPI networks. We downloaded the interaction files (protein.links.v11.0) in the STRING database [90], and only highly credible PPIs were retained by excluding PPIs with confidence scores below 0.7. The final STRING network was plotted using Cytoscape [91].

Gene family analysis

We chose the longest transcript in the downloaded annotation dataset to represent each gene, and removed genes with open reading frames shorter than 150 bp. Gene family clustering was then performed using OrthoFinder (v 2.3.12) [92], based on the predicted gene set for eight genomes. This analysis yielded 17,992 gene families. To identify gene families that had undergone expansion or contraction, we applied the CAFE (v5.0.0) program [93], which inferred the rate and direction of changes in gene family size over a given phylogeny. Among the eight species, 559 single-copy orthologs were aligned using MUSCLE (v3.8.1551) [94], and we eliminated poorly aligned positions and divergent regions of the alignment using Gblock 0.91b [95]. RAxML (v 8.2.12) was then used with the PROTGAMMALGF model to estimate a maximum likelihood tree. Divergence times were estimated using PAML MCMCTREE [85]. A Markov chain Monte Carlo (MCMC) process was run for 2,000,000 iterations, with a sample frequency of 100 after a burn-in of 1,000 iterations under an independent rates model. Two independent runs were performed to check the convergence. The fossil-calibrated eukaryote phylogeny was used to set the root height for the species tree, taken from the age of Animals (602–661 Ma) estimated in a previous fossil-calibrated eukaryotic phylogeny [96] and the divergence time between the euarchontoglires and laurasiatheria: (95.3–113 Ma) [97].

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Fund (U20A2051, 31760648 and 31860638), and Guangxi Natural Science Foundation (AB18221120), and Guangxi Distinguished scholars Program (201835), Science and Technology Major Project of Guangxi (Guike AA17204057).

DATA AND MATERIALS AVAILABILITY

The whole genome assembly (contig version) and gene annotation reported in this paper have been deposited in the Genome Warehouse in BIG Data Center [98], Beijing Institute of Genomics (China National Center for Bioinformation), Chinese Academy of Sciences, under accession number GWHAZTT00000000 that is publicly accessible at http://bigd.big.ac.cn/gwh. The AGP file for Hi-C was uploaded as supplement file. The Pacbio sequencing reads has been deposited into the genome sequence archive (GSA) in BIG under accession code CRA003783. The whole genome assembly also can be obtained in the National Center for Biotechnology Information (NCBI) under Bioproject PRJNA691688.

AUTHOR CONTRIBUTIONS

Q.L., J.R. and Y.W. conceived and designed the project. Q.L., K.Q, Z.Q., Z.J., Z.P., W.K., W.D. and D. W. collected the samples and performed experiments. J.R., Q.L., X.L., Z.Q., X.W., and T.F. analyzed the data. X.L., Q.L., K.Q. and Z.Q. drafted the manuscript. J.R., Y.W., W.Y., X.Q., and J.Y revised the manuscript.

References

1. Spithill TW, Smooker PM, Copeman DB. "Fasciola gigantica": Epidemiology, control, immunology and molecular biology. Fasciolosis. Oxon UK: CABI; 1999. p. 465-525.

2. World Health O. Accelerating work to overcome the global impact of neglected tropical diseases - a roadmap for implementation. Accelerating work to overcome the global impact of neglected tropical diseases - a roadmap for implementation. 2012:37 pp.- pp. PubMed PMID: CABI:20123334373.

3. Yadav SC, Sharma RL, Kalicharan A, Mehra UR, Dass RS, Verma AK. Primary experimental infection of riverine buffaloes with Fasciola gigantica. Veterinary Parasitology. 1999;82(4):285-96. doi: 10.1016/s0304-4017(99)00005-9. PubMed PMID: WOS:000080591400004.

4. Cwiklinski K, Dalton JP. Advances in Fasciola hepatica research using 'omics' technologies. International Journal for Parasitology. 2018;48(5):321-31. doi: https://doi.org/10.1016/j.ijpara.2017.12.001.

5. McNulty SN, Tort JF, Rinaldi G, Fischer K, Rosa BA, Smircich P, et al. Genomes of Fasciola hepatica

from the Americas Reveal Colonization with Neorickettsia Endobacteria Related to the Agents of Potomac Horse and Human Sennetsu Fevers. Plos Genetics. 2017;13(1). doi: 10.1371/journal.pgen.1006537. PubMed PMID: WOS:000394147700015.

6. Cwiklinski K, Dalton JP, Dufresne PJ, La Course J, Williams DJL, Hodgkinson J, et al. The Fasciola hepatica genome: gene duplication and polymorphism reveals adaptation to the host environment and the capacity for rapid evolution. Genome Biology. 2015;16. doi: 10.1186/s13059-015-0632-2. PubMed PMID: WOS:000353190400001.

7. Choi Y-J, Fontenla S, Fischer PU, Thanh Hoa L, Costabile A, Blair D, et al. Adaptive Radiation of the Flukes of the Family Fasciolidae Inferred from Genome-Wide Comparisons of Key Species. Mol Biol Evol. 2020;37(1):84-99. doi: 10.1093/molbev/msz204. PubMed PMID: WOS:000515121200009.

8. Pandey T, Ghosh A, Todur VN, Rajendran V, Kalita P, Kalita J, et al. Draft Genome of the Liver Fluke Fasciola gigantica. Acs Omega. 2020;5(19):11084-91. doi: 10.1021/acsomega.0c00980. PubMed PMID: WOS:000537145000049.

9. Soyemi J, Isewon I, Oyelade J, Adebiyi E. Inter-Species/Host-Parasite Protein Interaction Predictions Reviewed. Current Bioinformatics. 2018;13(4):396-406. doi: 10.2174/1574893613666180108155851. PubMed PMID: WOS:000437860800010.

10. de la Torre-Escudero E, Gerlach JQ, Bennett APS, Cwiklinski K, Jewhurst HL, Huson KM, et al. Surface molecules of extracellular vesicles secreted by the helminth pathogen Fasciola hepatica direct their internalisation by host cells. PLoS Negl Trop Dis. 2019;13(1). doi: 10.1371/journal.pntd.0007087. PubMed PMID: WOS:000457398700049.

11. Jaikua W, Kueakhai P, Chaithirayanon K, Tanomrat R, Wongwairot S, Riengrojpitak S, et al. Cytosolic superoxide dismutase can provide protection against Fasciola gigantica. Acta Tropica. 2016;162:75-82. doi: https://doi.org/10.1016/j.actatropica.2016.06.020.

12. Kelley JM, Elliott TP, Beddoe T, Anderson G, Skuce P, Spithill TW. Current Threat of Triclaoencazole Resistance in Fasciola hepatica. Trends in Parasitology. 2016;32(6):458-69. doi: 10.1016/j.pt.2016.03.002. PubMed PMID: WOS:000377730100007.

13. Rehman A, Ullah R, Gupta D, Khan MAH, Rehman L, Beg MA, et al. Generation of oxidative stress and induction of apoptotic like events in curcumin and thymoquinone treated adult Fasciola gigantica worms. Experimental Parasitology. 2020;209:107810. doi: https://doi.org/10.1016/j.exppara.2019.107810.

14. Le TH, De NV, Agatsuma T, Thi Nguyen TG, Nguyen QD, McManus DP, et al. Human fascioliasis and the presence of hybrid/introgressed forms of Fasciola hepatica and Fasciola gigantica in Vietnam. International Journal for Parasitology. 2008;38(6):725-30. doi: https://doi.org/10.1016/j.ijpara.2007.10.003.

15. Ashrafi K, Valero MA, Panova M, Periago MV, Massoud J, Mas-Coma S. Phenotypic analysis of adults of Fasciola hepatica, Fasciola gigantica and intermediate forms from the endemic region of Gilan, Iran. Parasitology International. 2006;55(4):249-60. doi: https://doi.org/10.1016/j.parint.2006.06.003.

16. Rhee JK, Eun GS, Lee SB. Karyotype of Fasciola sp. obtained from Korean cattle. Kisaengch'unghak chapchi The Korean journal of parasitology. 1987;25(1):37-44. PubMed PMID: MEDLINE:12886080.

17. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072-5. doi: 10.1093/bioinformatics/btt086. PubMed PMID: WOS:000318109300015.

18. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol.

2018;35(3):543-8. doi: 10.1093/molbev/msx319. PubMed PMID: WOS:000427260700002.

19. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Research. 2001;29(1):37-40. doi: 10.1093/nar/29.1.37. PubMed PMID: WOS:000166360300007.

20. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9):1236-40. doi: 10.1093/bioinformatics/btu031. PubMed PMID: WOS:000336095100007.

21. Lanciano S, Cristofari G. Measuring and interpreting transposable element expression. Nature Reviews Genetics. 2020;21(12):721-36. doi: 10.1038/s41576-020-0251-y.

22. Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. Microbiol Spectr. 2015;3(2):MDNA3-2014. doi: 10.1128/microbiolspec.MDNA3-0061-2014. PubMed PMID: 26104698.

23. Bejerano G, Lowe C, Ahituv N, King B, Siepel A, Salama S, et al. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature. 2006;441:87-90.

24. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences of the United States of America. 2020;117(17):9451-7. Epub 2020/04/16. doi: 10.1073/pnas.1921046117. PubMed PMID: 32300014.

25. Smit AFA. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Current Opinion in Genetics & Development. 1999;9(6):657-63. doi: 10.1016/s0959-437x(99)00031-3. PubMed PMID: WOS:000084277900007.

26. Keller G, Mateo P, Punekar J, Khozyem H, Gertsch B, Spangenberg J, et al. Environmental changes during the Cretaceous-Paleogene mass extinction and Paleocene-Eocene Thermal Maximum: Implications for the Anthropocene. Gondwana Research. 2018;56:69-89. doi: https://doi.org/10.1016/j.gr.2017.12.002.

27. Pastor-Cantizano N, Montesinos JC, Bernat-Silvestre C, Marcote MJ, Aniento F. p24 family proteins: key players in the regulation of trafficking along the secretory pathway. Protoplasma. 2016;253(4):967-85. doi: 10.1007/s00709-015-0858-6.

28. Montesinos JC, Sturm S, Langhans M, Hillmer S, Marcote MJ, Robinson DG, et al. Coupled transport of Arabidopsis p24 proteins at the ER-Golgi interface. J Exp Bot. 2012;63(11):4243-61. Epub 2012/05/10. doi: 10.1093/jxb/ers112. PubMed PMID: 22577184.

29. de la Torre-Escudero E, Gerlach JQ, Bennett APS, Cwiklinski K, Jewhurst HL, Huson KM, et al. Surface molecules of extracellular vesicles secreted by the helminth pathogen Fasciola hepatica direct their internalisation by host cells. PLoS Negl Trop Dis. 2019;13(1):e0007087-e. doi: 10.1371/journal.pntd.0007087. PubMed PMID: 30657764.

30. Juan T, Fürthauer M. Biogenesis and function of ESCRT-dependent extracellular vesicles. Seminars in Cell & Developmental Biology. 2018;74:66-77. doi: https://doi.org/10.1016/j.semcdb.2017.08.022.

31. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Research. 2011;21(3):381-95. doi: 10.1038/cr.2011.22.

32. Oda H, Okamoto I, Murphy N, Chu J, Price SM, Shen MM, et al. Monomethylation of histone H4-lysine 20 is involved in chromosome structure and stability and is essential for mouse development. Mol Cell Biol. 2009;29(8):2278-95. Epub 2009/02/17. doi: 10.1128/MCB.01768-08. PubMed PMID: 19223465.

33. Muiño L, Perteguer MJ, Gárate T, Martínez-Sernández V, Beltrán A, Romarís F, et al. Molecular and

immunological characterization of Fasciola antigens recognized by the MM3 monoclonal antibody. Molecular and Biochemical Parasitology. 2011;179(2):80-90. doi: https://doi.org/10.1016/j.molbiopara.2011.06.003.

34. Dalton JP, Neill SO, Stack C, Collins P, Walshe A, Sekiya M, et al. Fasciola hepatica cathepsin L-like proteases: biology, function, and potential in the development of first generation liver fluke vaccines. International Journal for Parasitology. 2003;33(11):1173-81. doi: https://doi.org/10.1016/S0020-7519(03)00171-1.

35. Batra S, Chatterjee RK, Srivastava VML. Antioxidant system of Litomosoides carinii and Setaria cervi: effect of a macrofilaricidal agent. Veterinary Parasitology. 1992;43(1):93-103. doi: https://doi.org/10.1016/0304-4017(92)90052-B.

36. McGonigle S, Dalton JP. ISOLATION OF FASCIOLA-HEPATICA HEMOGLOBIN. Parasitology. 1995;111:209-15. doi: 10.1017/s0031182000064969. PubMed PMID: WOS:A1995RR28600011.

37. Farahnak A, Golestani A, Eshraghian M. Activity of Superoxide Dismutase (SOD) Enzyme in the Excretory-Secretory Products of Fasciola hepatica and F. gigantica Parasites. Iran J Parasitol. 2013;8(1):167-70. PubMed PMID: 23682275.

38. Okada K, Fukui M, Zhu B-T. Protein disulfide isomerase mediates glutathione depletion-induced cytotoxicity. Biochemical and Biophysical Research Communications. 2016;477(3):495-502. doi: https://doi.org/10.1016/j.bbrc.2016.06.066.

39. Biswal DK, Roychowdhury T, Pandey P, Tandon V. De novo genome and transcriptome analyses provide insights into the biology of the trematode human parasite Fasciolopsis buski. Plos One. 2018;13(10). doi: 10.1371/journal.pone.0205570. PubMed PMID: WOS:000447430800025.

40. Wang X, Chen W, Huang Y, Sun J, Men J, Liu H, et al. The draft genome of the carcinogenic human liver fluke Clonorchis sinensis. Genome Biology. 2011;12(10). doi: 10.1186/gb-2011-12-10-r107. PubMed PMID: WOS:000301176900010.

41. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, et al. The genome of the blood fluke Schistosoma mansoni. Nature. 2009;460(7253):352-U65. doi: 10.1038/nature08160. PubMed PMID: WOS:000267979000029.

42. Li W, Liu B, Yang Y, Ren Y, Wang S, Liu C, et al. The genome of tapeworm Taenia multiceps sheds light on understanding parasitic mechanism and control of coenurosis disease. DNA Research. 2018;25(5):499-510. doi: 10.1093/dnares/dsy020. PubMed PMID: WOS:000456004800005.

43. Luo X, Zhou Y, Zhang B, Zhang Y, Wang X, Feng T, et al. Understanding divergent domestication traits from the whole-genome sequencing of swamp- and river-buffalo populations. National Science Review. 2020;7(3):686-701. doi: 10.1093/nsr/nwaa024. PubMed PMID: WOS:000537425800024.

44. de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ, et al. Initial sequencing and analysis of the human genome (vol 409, pg 860, 2001). Nature. 2001;412(6846):565-6. PubMed PMID: WOS:000170202900052.

45. Choi Y-J, Fontenla S, Fischer PU, Le TH, Costábile A, Blair D, et al. Adaptive Radiation of the Flukes of the Family Fasciolidae Inferred from Genome-Wide Comparisons of Key Species. Mol Biol Evol. 2020;37(1):84-99. doi: 10.1093/molbev/msz204. PubMed PMID: 31501870.

46. Irving JA, Spithill TW, Pike RN, Whisstock JC, Smooker PM. The Evolution of Enzyme Specificity in Fasciola spp. Journal of Molecular Evolution. 2003;57(1):1-15. doi: 10.1007/s00239-002-2434-x.

47. Näsvall J, Sun L, Roth JR, Andersson DI. Real-time evolution of new genes by innovation, amplification, and divergence. Science. 2012;338(6105):384-7. doi: 10.1126/science.1226521. PubMed PMID: 23087246.

48. Pollard TD. Actin and Actin-Binding Proteins. Cold Spring Harb Perspect Biol. 2016;8(8):a018226.

49. Dominguez R, Holmes KC. Actin Structure and Function. Annual Review of Biophysics. 2011;40(1):169.

50. Priess JR, Hirsh DI. Caenorhabditis elegans morphogenesis: the role of the cytoskeleton in elongation of the embryo. Developmental biology. 1986;117(1):156-73. doi: 10.1016/0012-1606(86)90358-1. PubMed PMID: MEDLINE:3743895.

51. McKeown C, Praitis V, Austin J. sma-1 encodes a beta(H)-spectrin homolog required for Caenorhabditis elegans morphogenesis. Development. 1998;125(11):2087-98. PubMed PMID: WOS:000074337100011.

52. de Almeida A, Martins AP, Mosca AF, Wijma HJ, Prista C, Soveral G, et al. Exploring the gating mechanisms of aquaporin-3: new clues for the design of inhibitors? Molecular Biosystems. 2016;12(5):1564-73. doi: 10.1039/c6mb00013d. PubMed PMID: WOS:000374936700015.

53. Soveral G, Casini A. Aquaporin modulators: a patent review (2010-2015). Expert Opinion on Therapeutic Patents. 2016;27(1):49.

54. Geadkaew A, von Bülow J, Beitz E, Grams SV, Viyanant V, Grams R. Functional analysis of novel aquaporins from Fasciola gigantica. Molecular and Biochemical Parasitology. 2011;175(2):144-53. doi: https://doi.org/10.1016/j.molbiopara.2010.10.010.

55. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nature Reviews Genetics. 2018;19(6):329-46. doi: 10.1038/s41576-018-0003-4.

56. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008;9(5):397-405. doi: 10.1038/nrg2337. PubMed PMID: 18368054.

57. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics. 2017;18(2):71-86. doi: 10.1038/nrg.2016.139.

58. Lynch VJ, Leclerc RD, May G, Wagner GP. Transposon-mediated rewiring of gene regulatory networks contributed to the evolution of pregnancy in mammals. Nature Genetics. 2011;43(11):1154-9. doi: 10.1038/ng.917.

59. Zhang F-K, Zhang X-X, Elsheikha HM, He J-J, Sheng Z-A, Zheng W-B, et al. Transcriptomic responses of water buffalo liver to infection with the digenetic fluke Fasciola gigantica. Parasites & Vectors. 2017;10(1):56. doi: 10.1186/s13071-017-1990-2.

60. Zhang F-K, Hu R-S, Elsheikha HM, Sheng Z-A, Zhang W-Y, Zheng W-B, et al. Global serum proteomic changes in water buffaloes infected with Fasciola gigantica. Parasites & Vectors. 2019;12(1):281. doi: 10.1186/s13071-019-3533-5.

61. Valero MA, Darce NAn, Panova M, Mas-Coma S. Relationships between host species and morphometric patterns in Fasciola hepatica adults and eggs from the northern Bolivian Altiplano hyperendemic region. Veterinary Parasitology. 2001;102(1):85-100. doi: https://doi.org/10.1016/S0304-4017(01)00499-X.

62. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Research. 2017;27(5):722-36. doi: 10.1101/gr.215087.116. PubMed PMID: WOS:000400392400007.

63. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nature Biotechnology. 2015;33(6):623-30. doi: 10.1038/nbt.3238.

64. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic

duplication in primary genome assemblies. Bioinformatics (Oxford, England). 2020;36(9):2896-8. doi: 10.1093/bioinformatics/btaa025. PubMed PMID: 31971576.

65. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Plos One. 2014;9(11). doi: 10.1371/journal.pone.0112963. PubMed PMID: WOS:000345533200052.

66. Zhang X, Zhang S, Zhao Q, Ming R, Tang H. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nature Plants. 2019;5(8):833-45. doi: 10.1038/s41477-019-0487-8.

67. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34(13):i142-i50. Epub 2018/06/29. doi: 10.1093/bioinformatics/bty266. PubMed PMID: 29949969; PubMed Central PMCID: PMCPMC6022658.

68. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589-95. doi: 10.1093/bioinformatics/btp698. PubMed PMID: WOS:000274973800001.

69. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. PubMed PMID: WOS:000268808600014.

70. Zhang X-X, Cwiklinski K, Hu R-S, Zheng W-B, Sheng Z-A, Zhang F-K, et al. Complex and dynamic transcriptional changes allow the helminth Fasciola gigantica to adjust to its intermediate snail and definitive mammalian hosts. BMC Genomics. 2019;20(1):729-. doi: 10.1186/s12864-019-6103-5. PubMed PMID: 31606027.

71. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature biotechnology. 2019;37(8):907-15. Epub 2019/08/02. doi: 10.1038/s41587-019-0201-4. PubMed PMID: 31375807.

72. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology. 2015;33(3):290-5. Epub 2015/02/18. doi: 10.1038/nbt.3122. PubMed PMID: 25690850.

73. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Jr., Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic acids research. 2003;31(19):5654-66. doi: 10.1093/nar/gkg770. PubMed PMID: 14500829.

74. Stanke M, Steinkamp R, Waack S, Morgenstern B. AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Research. 2004;32(suppl_2):W309-W12. doi: 10.1093/nar/gkh379.

75. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5(1):59. doi: 10.1186/1471-2105-5-59.

76. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic acids research. 2017;45(D1):D158-D69. Epub 2016/11/29. doi: 10.1093/nar/gkw1099. PubMed PMID: 27899622.

77. Gertz EM, Yu Y-K, Agarwala R, Schäffer AA, Altschul SF. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol. 2006;4:41-. doi: 10.1186/1741-7007-4-41. PubMed PMID: 17156431.

78. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome research. 2004;14(5):988-95. doi: 10.1101/gr.1865504. PubMed PMID: 15123596.

79. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics. 2011;12:491-. doi: 10.1186/1471-2105-12-491. PubMed PMID: 22192575.

80. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucleic acids research. 2005;33(Web Server issue):W116-20. doi: 10.1093/nar/gki442. PubMed PMID: 15980438.

81. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997;25(17):3389-402. doi: 10.1093/nar/25.17.3389. PubMed PMID: 9254694.

82. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. Journal of Molecular Biology. 2016;428(4):726-31. doi: https://doi.org/10.1016/j.jmb.2015.11.006.

83. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. Omics-a Journal of Integrative Biology. 2012;16(5):284-7. doi: 10.1089/omi.2011.0118. PubMed PMID: WOS:000303653300007.

84. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014;47:11.2.1-.2.34. doi: 10.1002/0471250953.bi1112s47. PubMed PMID: 25199790.

85. Yang Z, Rannala B. Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. Mol Biol Evol. 2006;23(1):212-26. doi: 10.1093/molbev/msj024.

86. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods. 2011;8(10):785-6. doi: 10.1038/nmeth.1701. PubMed PMID: WOS:000295358000004.

87. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of Molecular Biology. 2000;300(4):1005-16. doi: 10.1006/jmbi.2000.3903. PubMed PMID: WOS:000088508500026.

88. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology. 2001;305(3):567-80. doi: 10.1006/jmbi.2000.4315. PubMed PMID: WOS:000167760800017.

89. Santos A, Tsafou K, Stolte C, Pletscher-Frankild S, O'Donoghue SI, Jensen LJ. Comprehensive comparison of large-scale tissue expression datasets. Peerj. 2015;3. doi: 10.7717/peerj.1054. PubMed PMID: WOS:000357321300006.

90. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8-a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research. 2009;37:D412-D6. doi: 10.1093/nar/gkn760. PubMed PMID: WOS:000261906200075.

91. Doncheva NT, Morris JH, Gorodkin J, Jensen LJ. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. Journal of Proteome Research. 2019;18(2):623-32. doi: 10.1021/acs.jproteome.8b00702. PubMed PMID: WOS:000457947700007.

92. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology. 2019;20(1):238. doi: 10.1186/s13059-019-1832-y.

93. Han MV, Thomas GWC, Lugo-Martinez J, Hahn MW. Estimating Gene Gain and Loss Rates in the Presence of Error in Genome Assembly and Annotation Using CAFE 3. Mol Biol Evol. 2013;30(8):1987-97. doi: 10.1093/molbev/mst100. PubMed PMID: WOS:000321820400022.

94. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. Bmc Bioinformatics. 2004;5:1-19. doi: 10.1186/1471-2105-5-113. PubMed PMID: WOS:000223920500001.

95. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously

aligned blocks from protein sequence alignments. Systematic Biology. 2007;56(4):564-77. doi: 10.1080/10635150701472164. PubMed PMID: WOS:000248359900002.

96. Parfrey LW, Lahr DJG, Knoll AH, Katz LA. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proceedings of the National Academy of Sciences. 2011;108(33):13624. doi: 10.1073/pnas.1110633108.

97. Benton MJ, Donoghue PCJ. Paleontological evidence to date the tree of life (vol 24, pg 26, 2007). Mol Biol Evol. 2007;24(3):889-91. doi: 10.1093/molbev/msm017. PubMed PMID: WOS:000244662000027.

98. Members C-N, Partners. Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2021. Nucleic Acids Research. 2021;49(D1):D18-D28. doi: 10.1093/nar/gkaa1022.

Fig. 1 Landscape of the *Fasciola gigantica* genome.

(A) Comparisons of the assembled contigs and scaffold lengths (y-axis) and tallies (x-axis) in *Fasciola* species. (B) Hi-C interactive heatmap of the genome-wide organization. The effective mapping read pairs between two bins were used as a signal of the strength of the interaction between the two bins. (C) Integration of genomic and annotation data using 1 Mb bins in 10 Hi-C assembled chromosomes. (a) Distribution of the GC content (GC content > 39% and < 52%); (b) distribution of the long interspersed element (LINE) percentage > 0% and < 50%; (c) distribution of the long terminal repeat (LTR) percentage > 0% and < 50%; (d) distribution of the gene percentage > 0% and < 70%; (e) distribution of the heterozygosity density of our sample (percentage > 0% and < 1%); (f) distribution of the heterozygosity density of SAMN03459319 in the NCBI database. Hi-C, chromosome conformation capture sequencing;

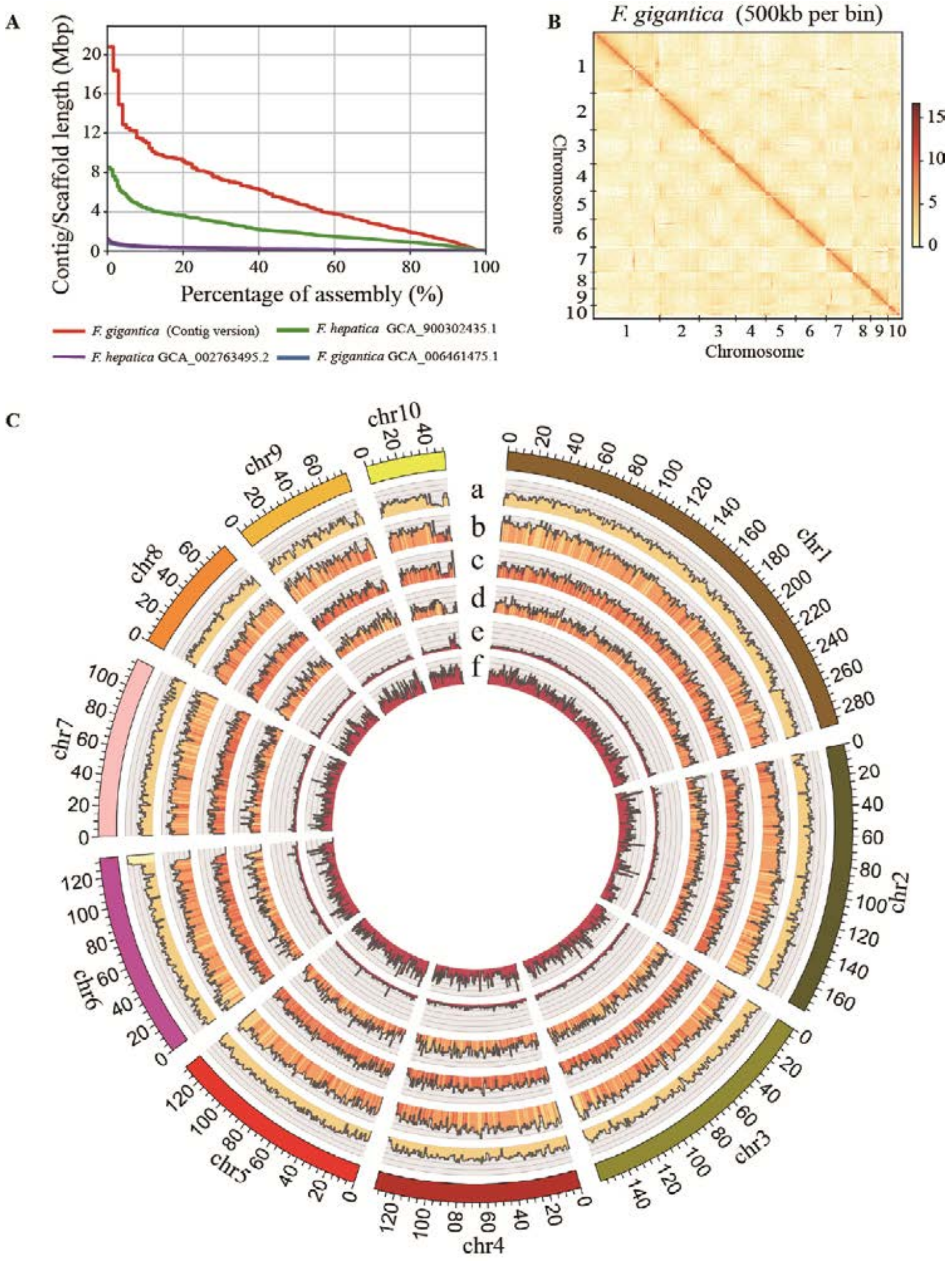

A
Contig/Scaffold length (Mbp)
0
4
8
12
16
20
0
20
40
60
80
100
Percentage of assembly (%)
F. gigantica (Contig version)
F. hepatica GCA_900302435.1
F. hepatica GCA_002763495.2
F. gigantica GCA_006461475.1
B
F. gigantica (500kb per bin)
Chromosome
1 2 3 4 5 6 7 8 9 10
Chromosome
1 2 3 4 5 6 7 8 9 10
15
10
5
0
C
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
a
b
c
d
e
f

Fig. 2 Identification of repeat expansion and alternative gene networks in the *Fasciola gigantica* genome.

(A) The distribution of repetitive sequence length among the genomes of six flatworms and the human genome. (B) Landscape of LINEs and LTRs distribution in the *Fasciola gigantica* genome. The x-axis shows the expansion time of TEs calculated by the divergence between repeat sequences. The mutation rate was set as 1.73 × 10^{-9} per year. The orange line represents the repeat length ratio, used to estimate the signatures of selection, which was corrected by the total length of intronic and intergenic regions in history. (C) The functional enrichment of genes with more than 10 kb LINE insertions between 41 Ma and 62 Ma by Gene Ontology (GO) classification. The GO terms related to vesicle secretion are marked in red. (D) *TMED10* gene structure map. LINEs original between 41 Ma and 62 Ma and longer than 500 bp identified by RepeatMasker were plotted. LTRs longer than 500 bp were plotted. Long interspersed element, LINE; long terminal repeat, LTR; TE, transposable element; TMED10, transmembrane P24 trafficking protein 10.

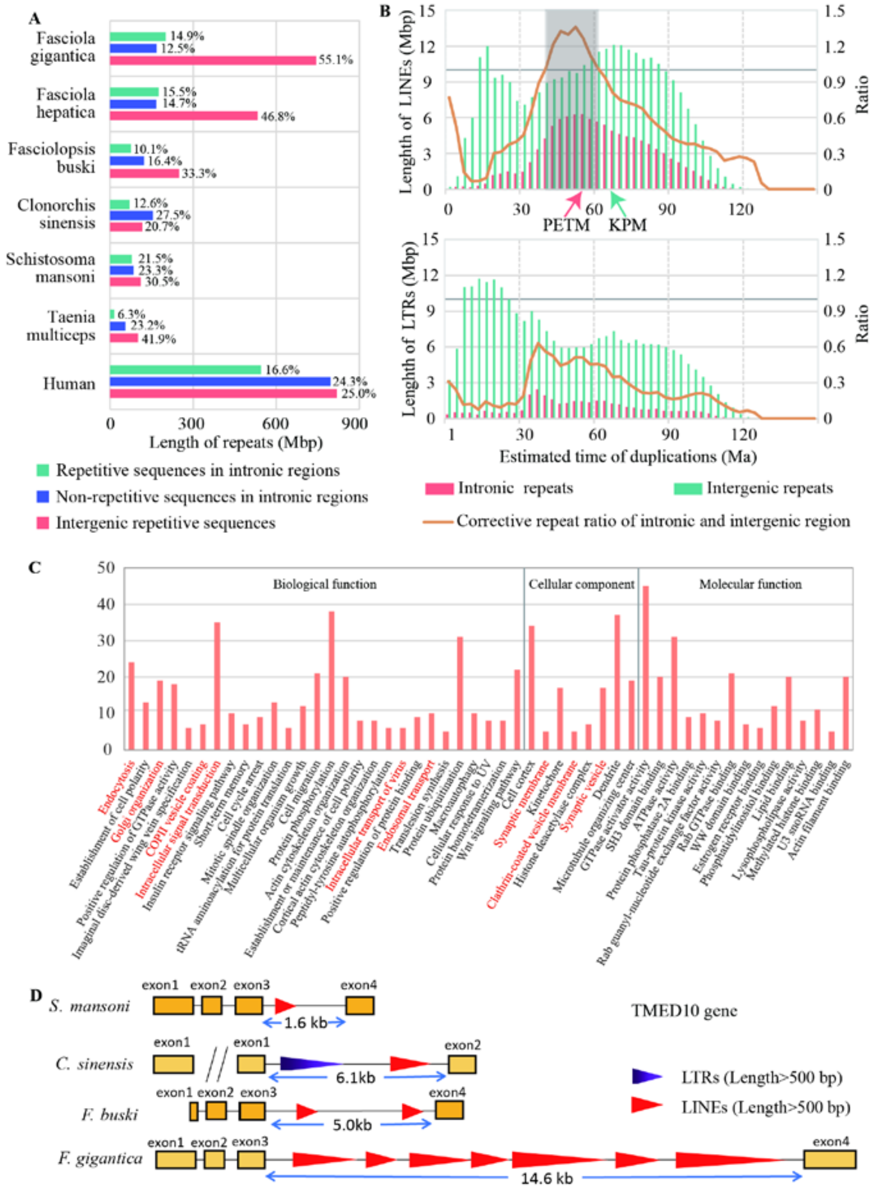


Fig. 3 Genome-wide host-parasite interaction analysis.

(A) Pie chart for proteases identified in *Fasciola gigantica*. (B) The interaction mode between the adult *Fasciola gigantica* and the host. (C) The protein-protein interaction (PPI) network of redox-related pathways in *Fasciola gigantica* with host proteins. The genes indicated in the three gene ontology (GO) terms were significantly enriched and have their encoded proteins have PPIs with excretory/secretory (E/S) proteins.

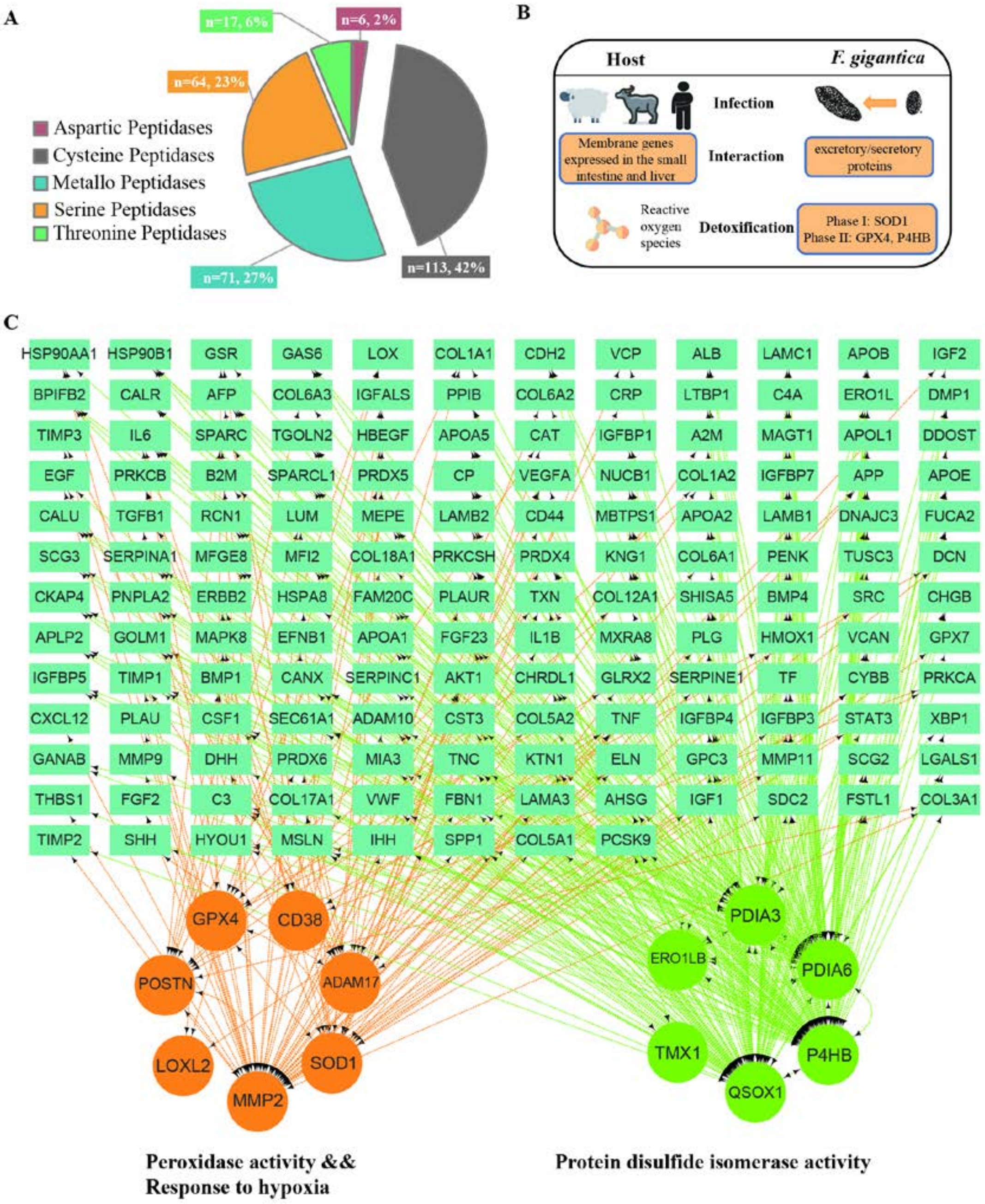


Fig. 4 Phylogenetic tree and gene family analysis.

(A) A phylogenetic tree generated using 559 single-copy orthologous genes. The numbers on the species names are the expanded (+) and contracted (-) gene families. The numbers on the nodes are the divergence time between species. (B) A phylogenetic tree of actin genes in flatworms and humans. All human homologue genes are selected as outgroup. (C) Phylogenetic tree of aquaglyceroporin (AQP) family genes in flatworms and humans. The human homolog genes (*AQP11*, *AQP12A*, and *AQP12B*) were selected as the outgroup. (D) A phylogenetic tree of copper/zinc superoxide dismutase (*SOD*) genes in flatworms and humans. The midpoint was selected as the root node.

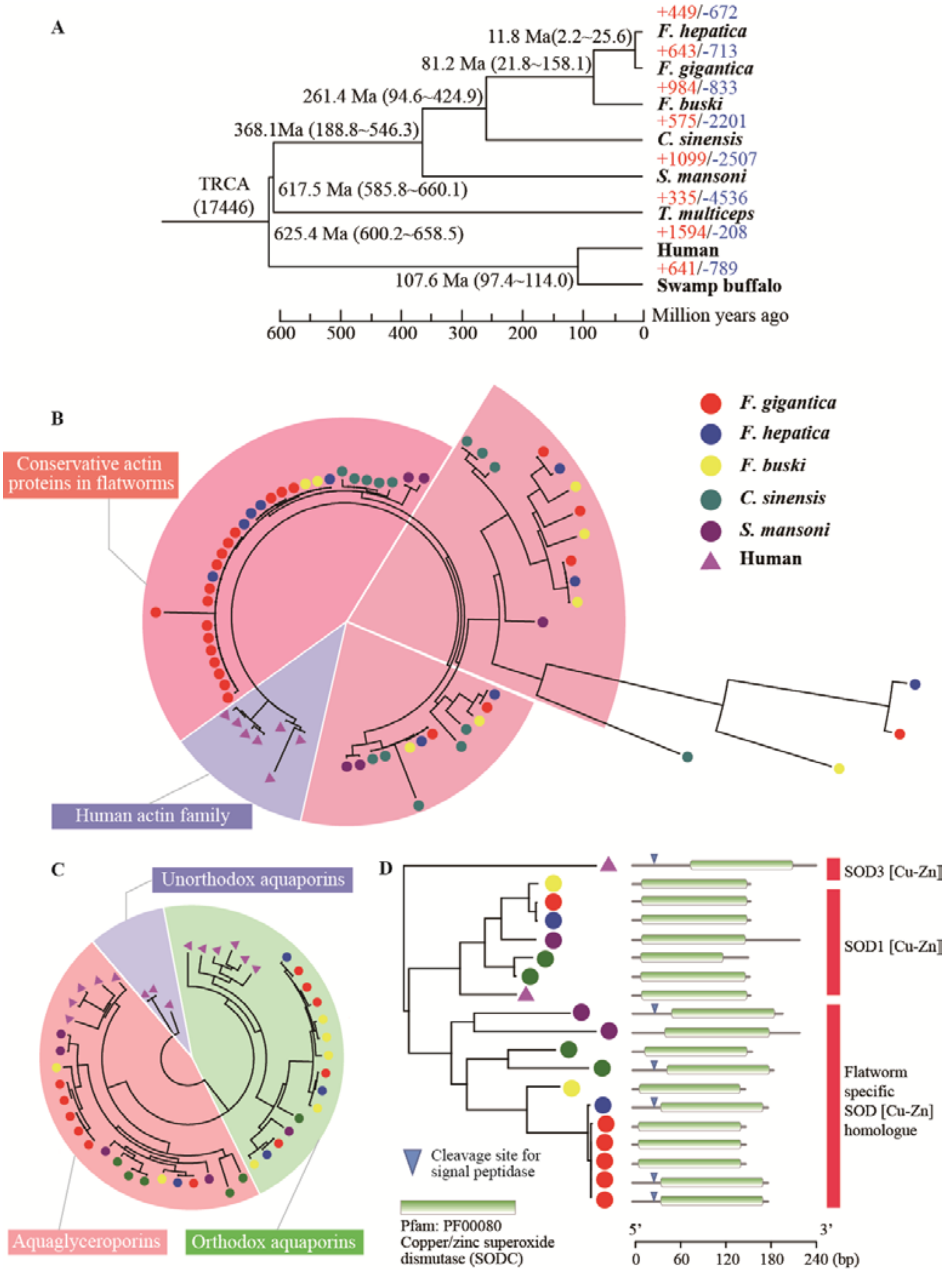


Fig. 5 Schematic diagram of the process of Fasciola-specific repeat expansion during evolution.

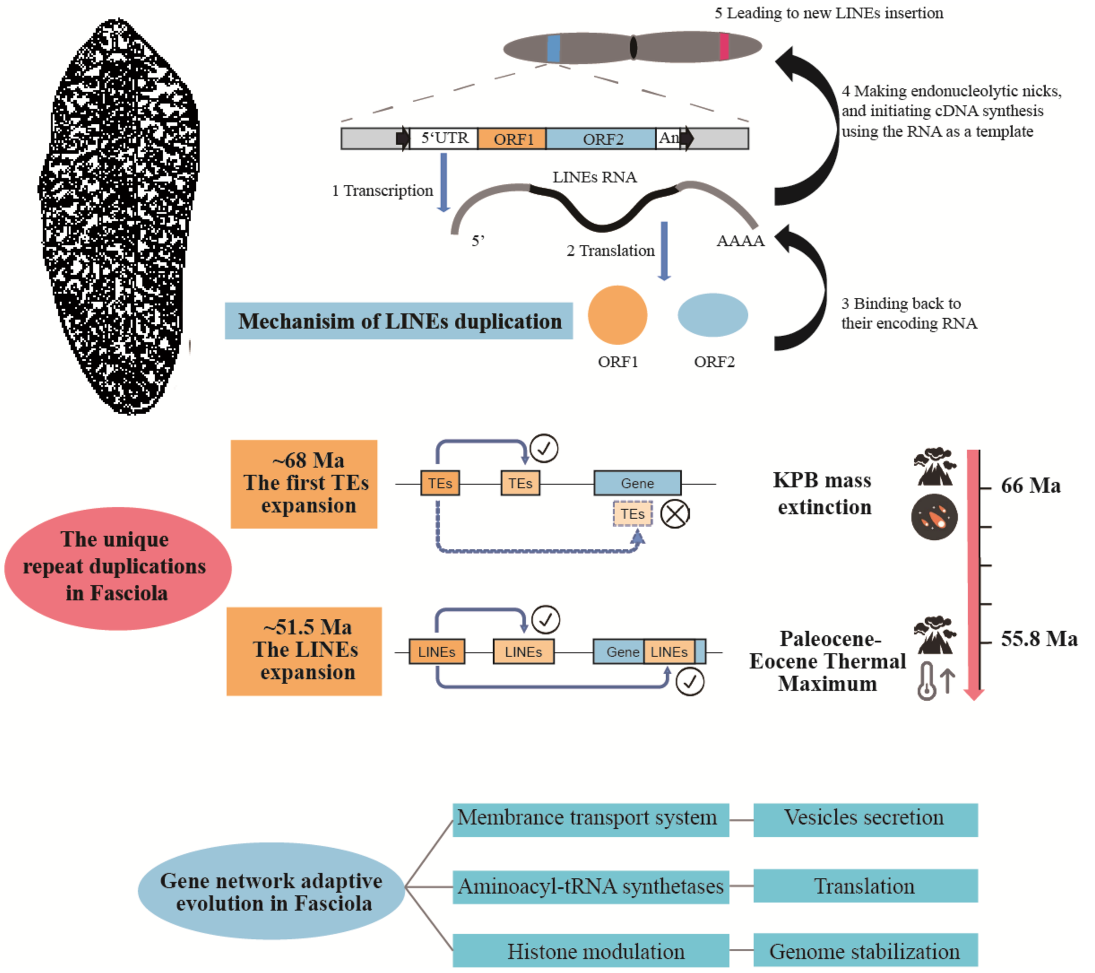

5 Leading to new LINEs insertion
4 Making endonucleolytic nicks, and initiating cDNA synthesis using the RNA as a template
5'UTR
ORF1
ORF2
An
1 Transcription
LINEs RNA
5'
2 Translation
AAAA
3 Binding back to their encoding RNA
Mechanisim of LINEs duplication
ORF1
ORF2
~68 Ma The first TEs expansion
TEs
TEs
Gene
TEs
The unique repeat duplications in Fasciola
KPB mass extinction
66 Ma
~51.5 Ma The LINEs expansion
LINEs
LINEs
Gene
LINEs
Paleocene-Eocene Thermal Maximum
55.8 Ma
Gene network adaptive evolution in Fasciola
Membrance transport system
Vesicles secretion
Aminoacyl-tRNA synthetases
Translation
Histone modulation
Genome stabilization

Table 1. Summary statistics for the genome sequences and annotation.

	F. gigantica	
Genome	Total Genome Size (Mb)	1,348
	Chromosome Number	10
	Scaffold Number [a]	10+24
	Scaffold N50 (Mb)	133
	Scaffold L50	4
	Contig Number	1,022
	Contig N50 (Mb)	4.89
	Heterozygosity Rate (%)	1.9×10^{-3}
Annotation	Total Gene Number	12,503
	Average CDS Length (bp)	1552.7
	Average Gene Length (kb)	28.8
	Percentage of Genome Covered by CDSs (%)	1.5%
	BUSCO Assessment	90.4%
	Repeat Content	70.0%

[a] number of chromosome level scaffolds and unplaced scaffolds.

CDS, coding sequence.

Fig. S1. Genome-wide all-by-all chromosome conformation capture sequencing (Hi-C) interaction in *F. gigantica* (Bins = 500 K).

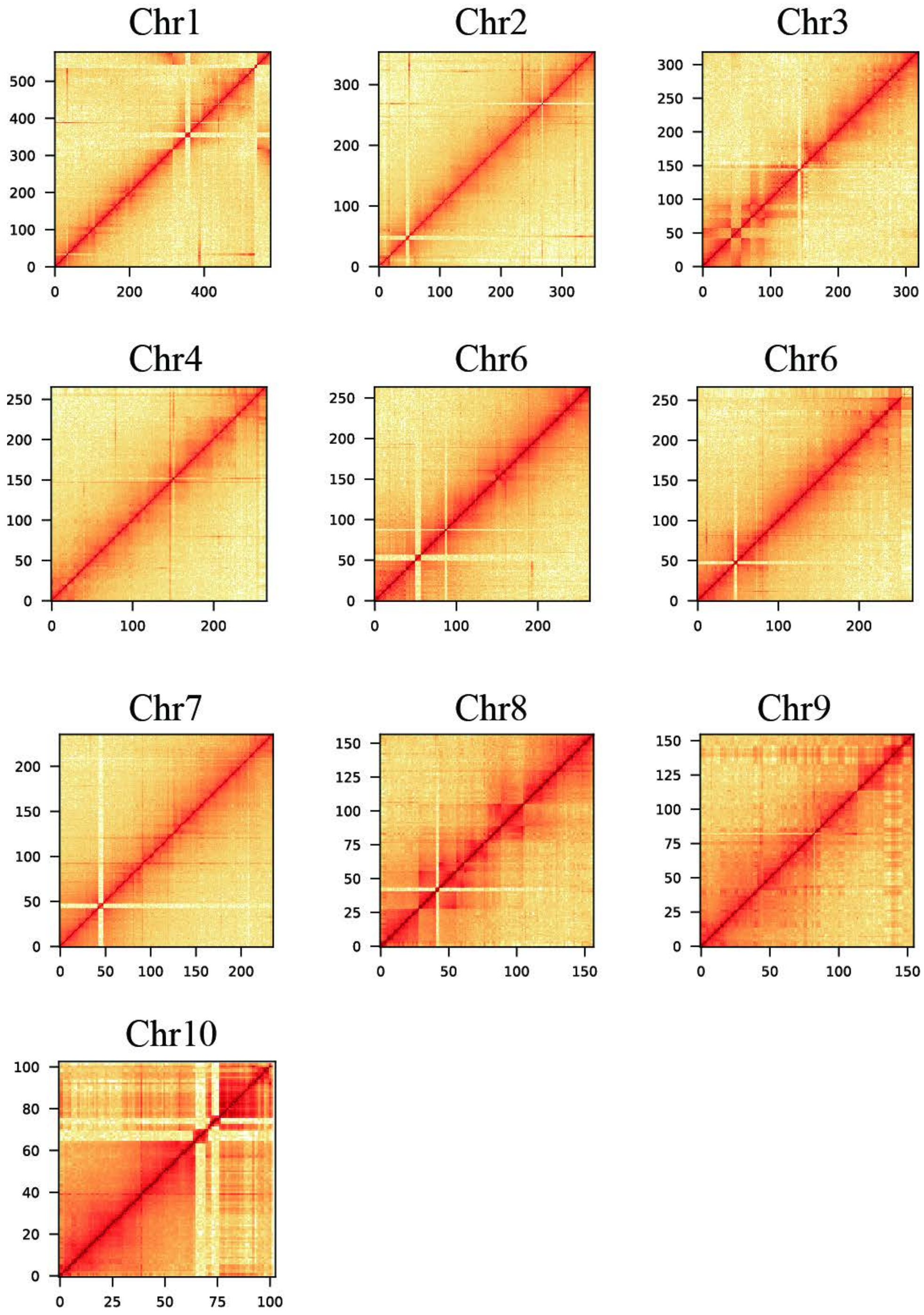


Fig. S2. Comparison of chromosome length between the chromosome conformation capture sequencing (Hi-C) assembly and estimates from published karyotype data by Jae Ku Rhee.

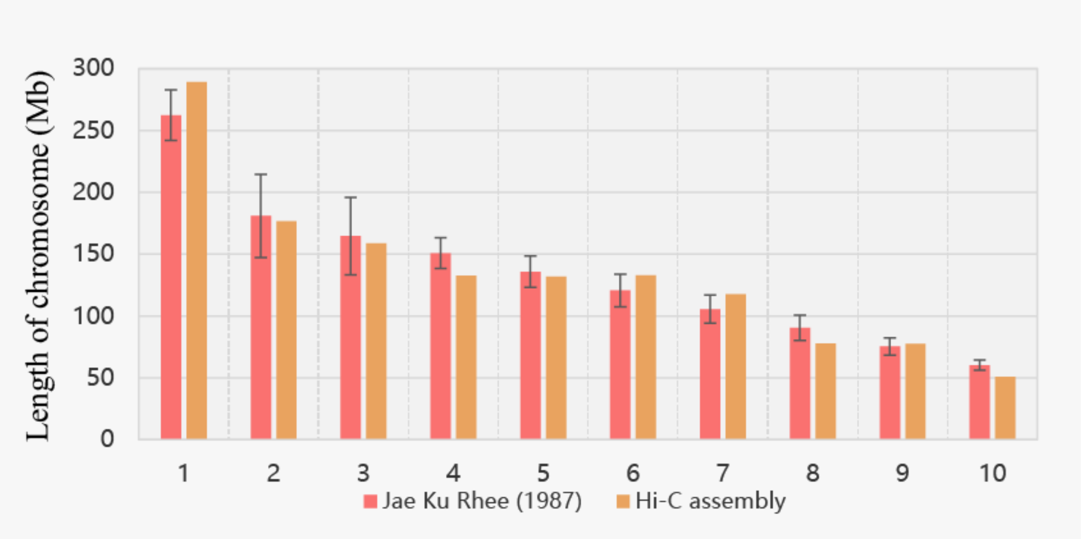


Fig. S3. Boxplot of average gene length.

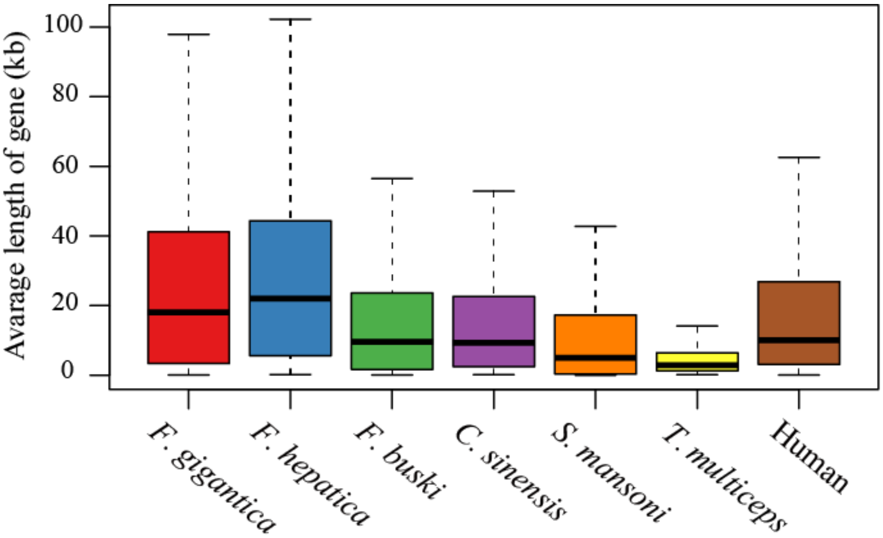


Fig. S4. Boxplot of average coding sequence (CDS) length per gene.

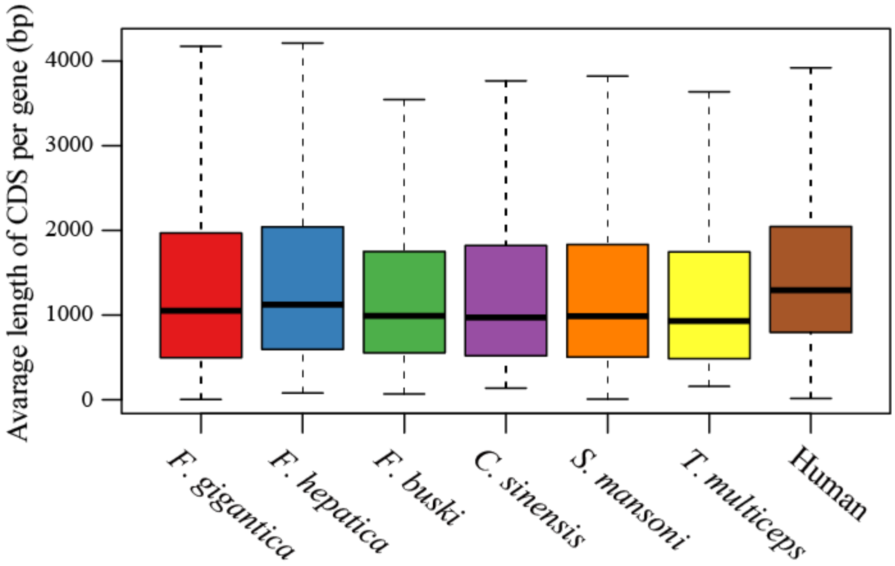


Fig. S5. Divergence distribution of classified families of transposable elements. The classified transposon families in *F. gigantica*.

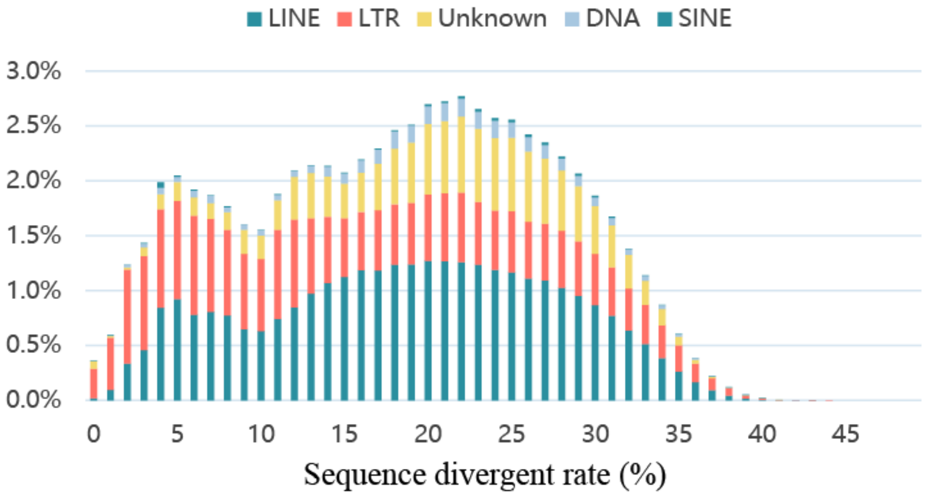


Fig. S6. Expansion time of long terminal repeats (LTRs) and long interspersed elements (LINEs). The mutation rate was 1.73×10^{-9}.

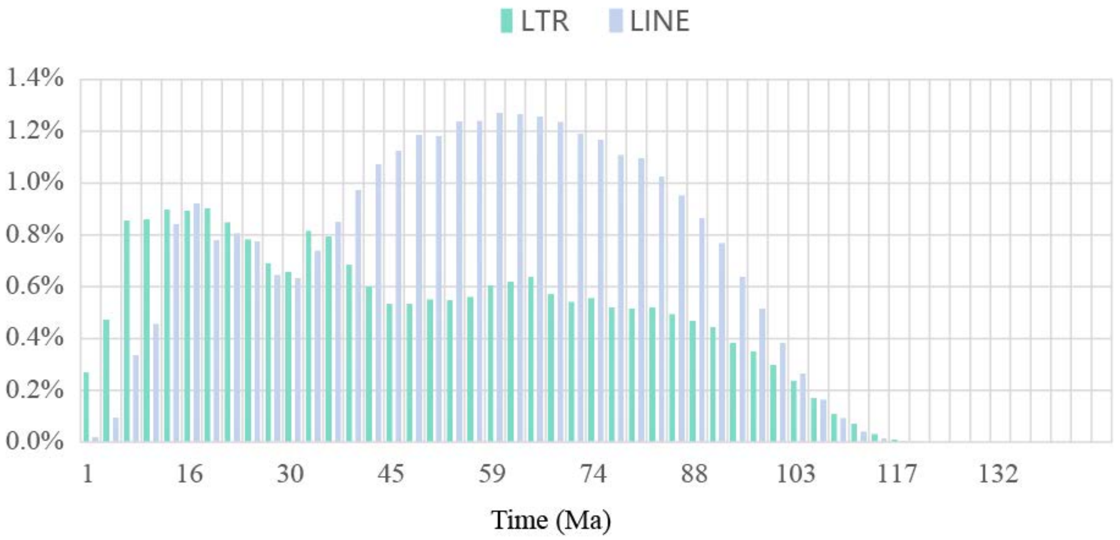
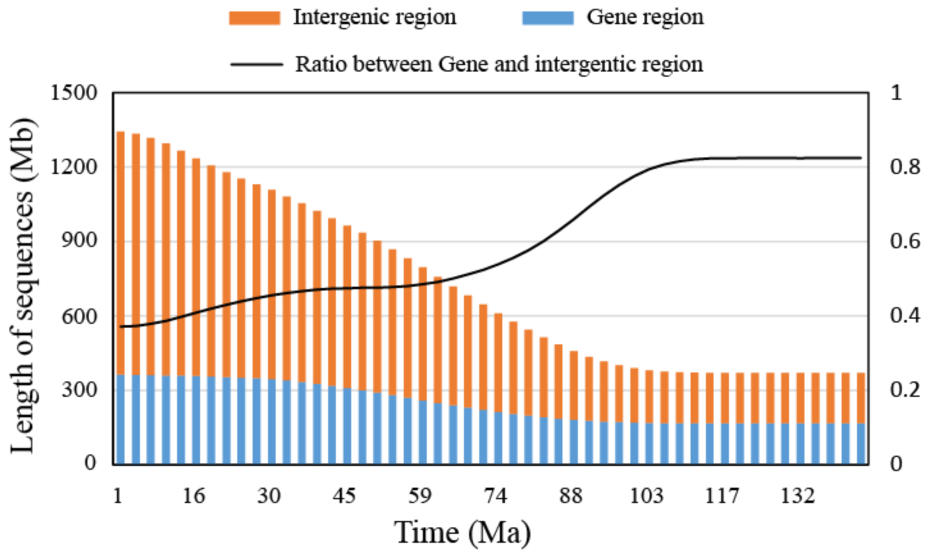


Fig. S7. Estimation of *F. gigantica* genome size based on the expansion time of repeat sequences during evolution. The mutation rate was 1.73×10^{-9}.



Supplementary Table 1. Genome sequencing strategy for buffaloes

Supplementary Table 2. Summary of the *Fasciola gigantica* genome assembly

Supplementary Table 3. Summary of different assemblies in *Fasciola* species

Supplementary Table 4. Summary of chromosome conformation capture sequencing (Hi-C) assembly of the chromosome length in *Fasciola gigantica*

Supplementary Table 5. Assessment of the completeness and accuracy of the genome

Supplementary Table 6. BUSCO assessment of the genome

Supplementary Table 7. Number of genes with functional classification gained using various methods

Supplementary Table 8. Transposable element content of *Fasciola gigantica* genome

Supplementary Table 9. The list of genes with more than 10 kb of long interspersed element (LINE) insertion between 41 Ma and 62 Ma

Supplementary Table 10. Gene ontology (GO) term category enrichment for genes with more than 10 kb of long interspersed element (LINE) insertion between 41 Ma and 62 Ma

Supplementary Table 11. Kyoto Encyclopedia of Genes and Genomes (KEGG pathway enrichment for genes with more than 10 kb of long interspersed element (LINE) insertion between 41 Ma and 62 Ma

Supplementary Table 12. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment for genes with more than 10 kb of long interspersed element (LINE) insertion between 41 Ma and 62 Ma

Supplementary Table 13. Protein inhibitors in the *Fasciola gigantica* genome

Supplementary Table 14. Excretory/secretory (E/S) proteins in the *Fasciola gigantica* genome

Supplementary Table 15. Gene ontology (GO) term category enrichment for excretory/secretory (E/S) proteins

Supplementary Table 16. Gene ontology (GO) term category enrichment for rapidly evolving families specific to *F. gigantica*.

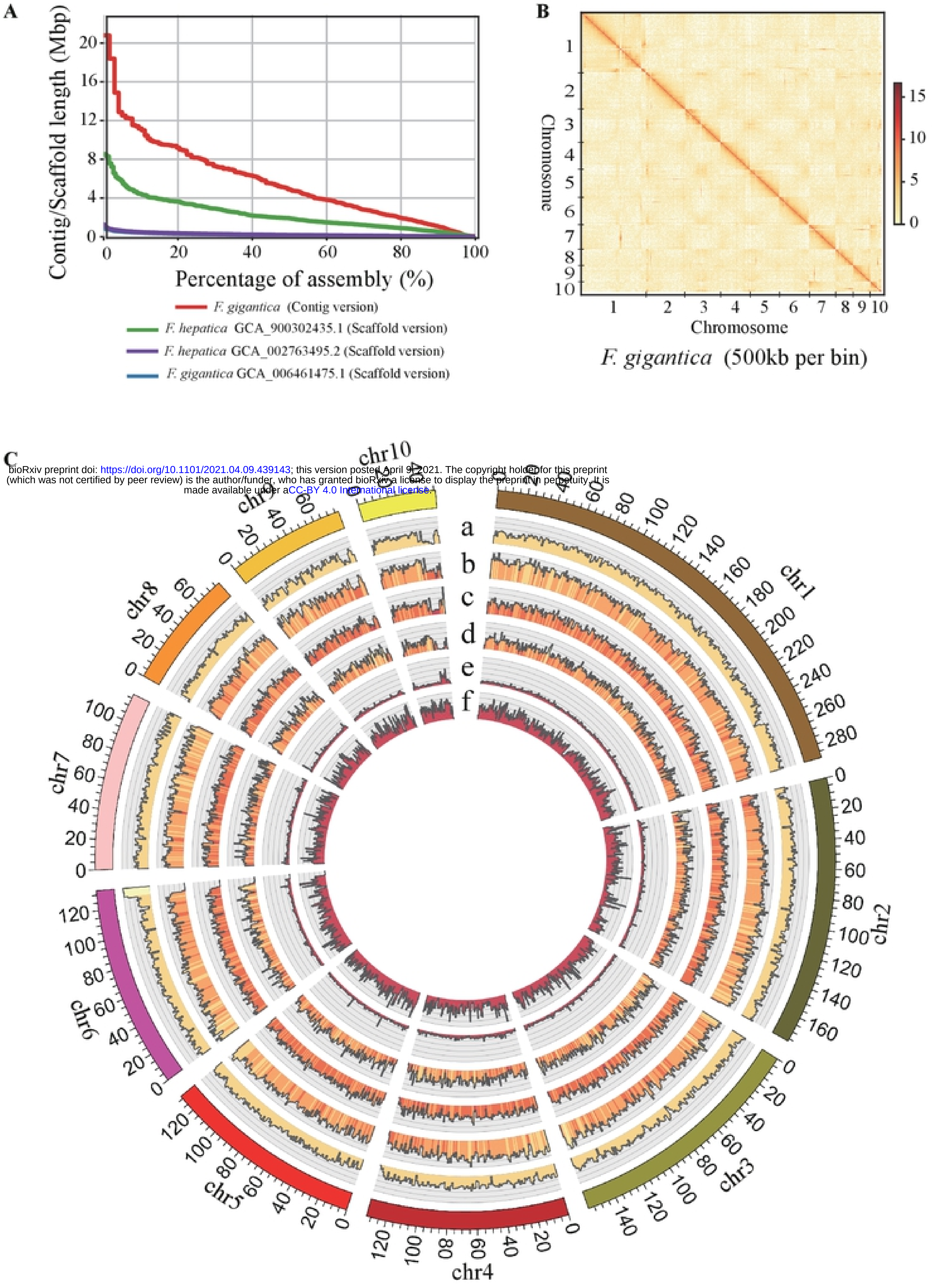

A
Contig/Scaffold length (Mbp)
0
4
8
12
16
20
Percentage of assembly (%)
0
20
40
60
80
100
F. gigantica (Contig version)
F. hepatica GCA_900302435.1 (Scaffold version)
F. hepatica GCA_002763495.2 (Scaffold version)
F. gigantica GCA_006461475.1 (Scaffold version)
B
Chromosome
1
2
3
4
5
6
7
8
9
10
Chromosome
15
10
5
0
F. gigantica (500kb per bin)
C
a
b
c
d
e
f
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10

Figure1

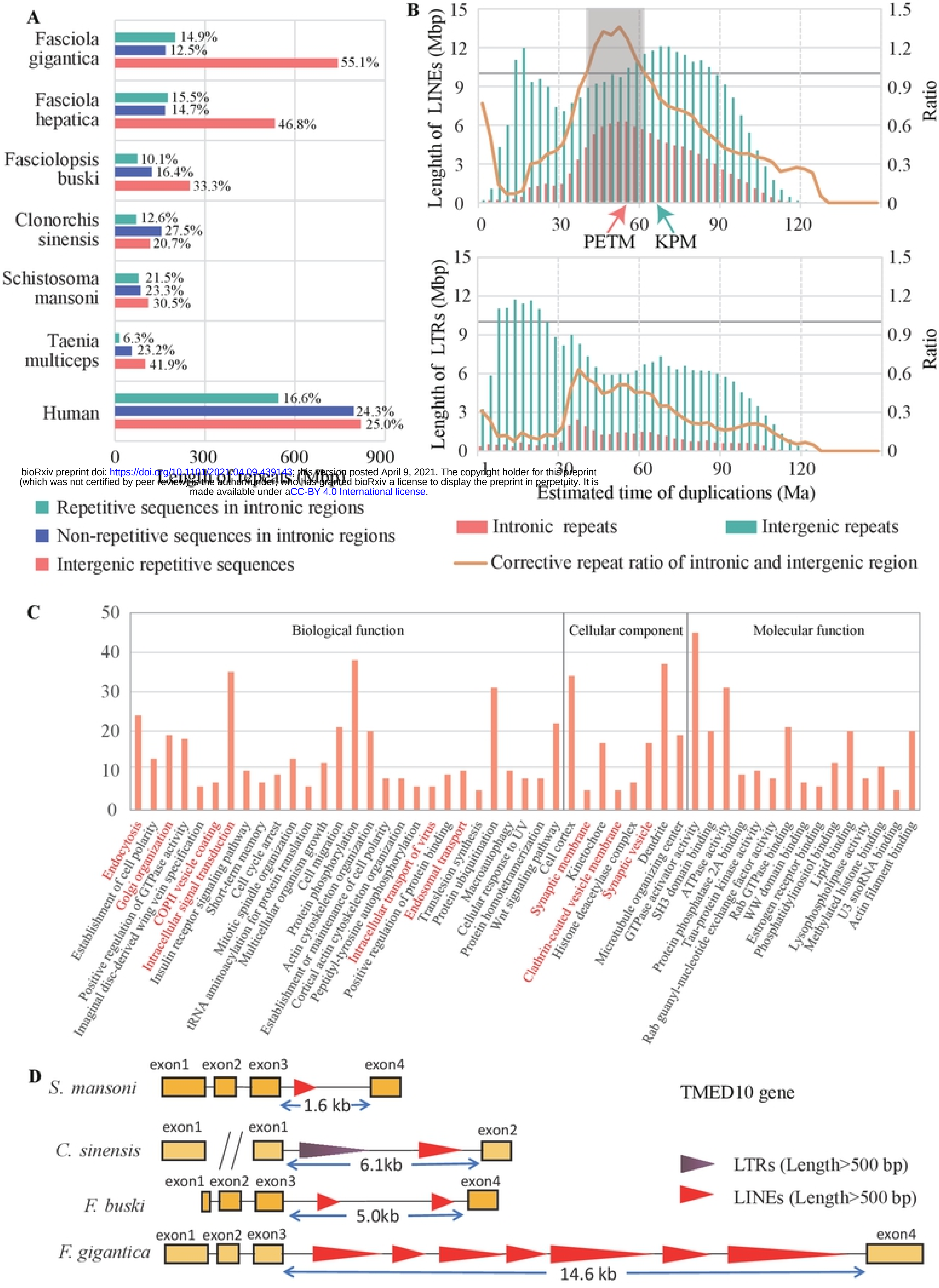

A
Fasciola gigantica
14.9%
12.5%
55.1%
Fasciola hepatica
15.5%
14.7%
46.8%
Fasciolopsis buski
10.1%
16.4%
33.3%
Clonorchis sinensis
12.6%
27.5%
20.7%
Schistosoma mansoni
21.5%
23.3%
30.5%
Taenia multiceps
6.3%
23.2%
41.9%
Human
16.6%
24.3%
25.0%
0
300
600
900
Length of repeats (Mbp)
Repetitive sequences in intronic regions
Non-repetitive sequences in intronic regions
Intergenic repetitive sequences
B
Lenghth of LINEs (Mbp)
Ratio
15
12
9
6
3
0
1.5
1.2
0.9
0.6
0.3
0
0
30
60
90
120
PETM
KPM
Lenghth of LTRs (Mbp)
Ratio
Estimated time of duplications (Ma)
Intronic repeats
Intergenic repeats
Corrective repeat ratio of intronic and intergenic region
C
50
40
30
20
10
0
Biological function
Cellular component
Molecular function
Endocytosis
Establishment of cell polarity
Golgi organization
Positive regulation of GTPase activity
Imaginal disc-derived wing vein specification
COPII vesicle coating
Intracellular signal transduction
Insulin receptor signaling pathway
Short-term memory
Cell cycle arrest
Mitotic spindle organization
tRNA aminoacylation for protein translation
Multicellular organism growth
Cell migration
Protein phosphorylation
Actin cytoskeleton organization
Establishment or maintenance of cell polarity
Cortical actin cytoskeleton organization
Peptidyl-tyrosine autophosphorylation
Intracellular transport of virus
Positive regulation of protein binding
Endosomal transport
Translesion synthesis
Protein ubiquitination
Macroautophagy
Cellular response to UV
Protein homotetramerization
Wnt signaling pathway
Cell cortex
Synaptic membrane
Clathrin-coated vesicle membrane
Kinetochore
Histone deacetylase complex
Synaptic vesicle
Dendrite
Microtubule organizing center
GTPase activator activity
SH3 domain binding
Protein phosphatase 2A binding
ATPase activity
Tau-protein kinase activity
Rab guanyl-nucleotide exchange factor activity
Rab GTPase binding
WW domain binding
Estrogen receptor binding
Phosphatidylinositol binding
Lipid binding
Lysophospholipase activity
Methylated histone binding
U3 snoRNA binding
Actin filament binding
D
S. mansoni
exon1 exon2 exon3 exon4
1.6 kb
C. sinensis
exon1 exon1 exon2
6.1kb
F. buski
exon1 exon2 exon3 exon4
5.0kb
F. gigantica
exon1 exon2 exon3 exon4
14.6 kb
TMED10 gene
LTRs (Length>500 bp)
LINEs (Length>500 bp)

Figure2

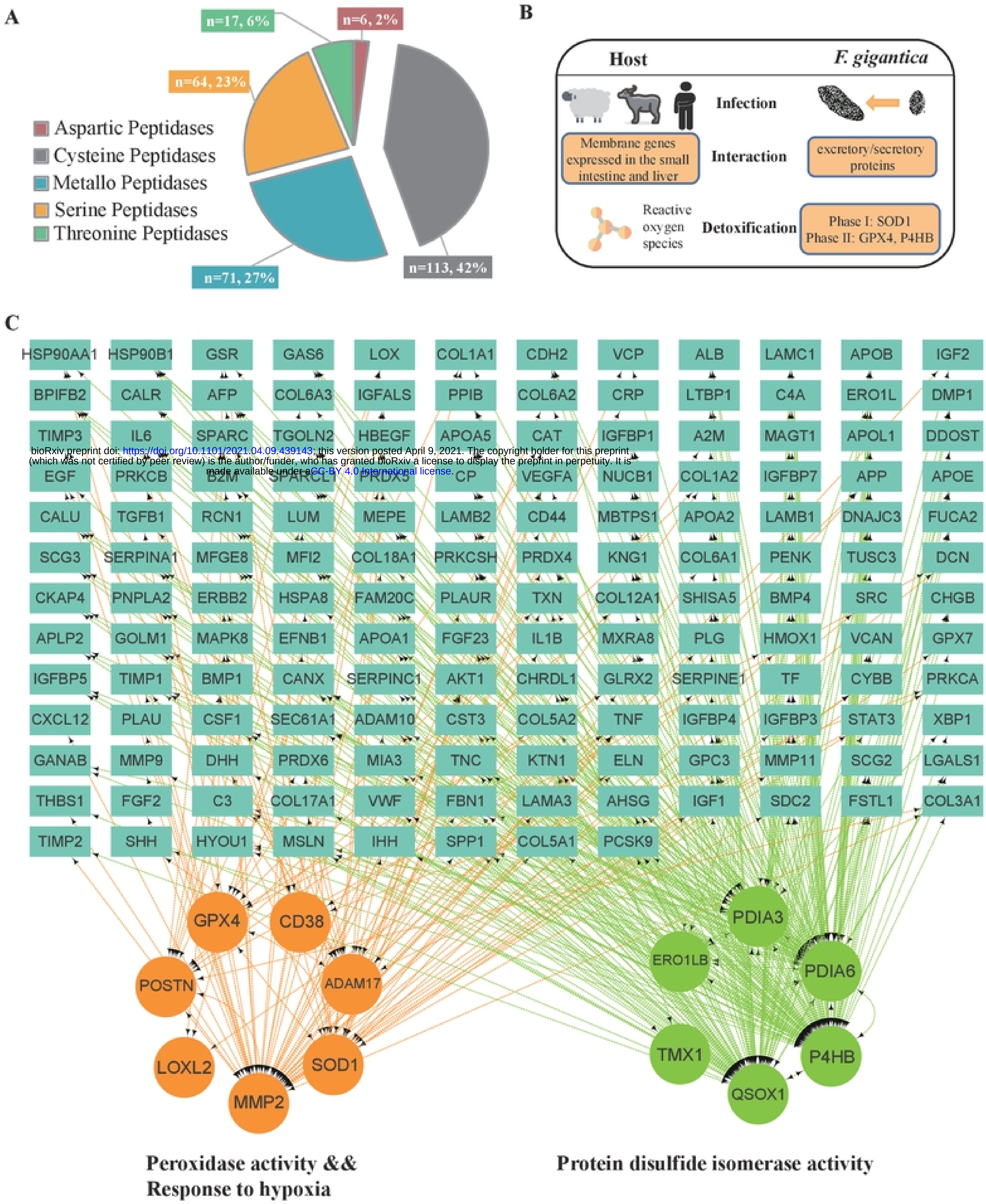

A
n=17, 6%
n=6, 2%
n=64, 23%
Aspartic Peptidases
Cysteine Peptidases
Metallo Peptidases
Serine Peptidases
Threonine Peptidases
n=71, 27%
n=113, 42%
B
Host
F. gigantica
Infection
Membrane genes expressed in the small intestine and liver
Interaction
excretory/secretory proteins
Reactive oxygen species
Detoxification
Phase I: SOD1
Phase II: GPX4, P4HB
C
HSP90AA1
HSP90B1
GSR
GAS6
LOX
COL1A1
CDH2
VCP
ALB
LAMC1
APOB
IGF2
BPIFB2
CALR
AFP
COL6A3
IGFALS
PPIB
COL6A2
CRP
LTBP1
C4A
ERO1L
DMP1
TIMP3
IL6
SPARC
TGOLN2
HBEGF
APOA5
CAT
IGFBP1
A2M
MAGT1
APOL1
DDOST
EGF
PRKCB
B2M
SPARCL1
PRDX5
CP
VEGFA
NUCB1
COL1A2
IGFBP7
APP
APOE
CALU
TGFB1
RCN1
LUM
MEPE
LAMB2
CD44
MBTPS1
APOA2
LAMB1
DNAJC3
FUCA2
SCG3
SERPINA1
MFGE8
MFI2
COL18A1
PRKCSH
PRDX4
KNG1
COL6A1
PENK
TUSC3
DCN
CKAP4
PNPLA2
ERBB2
HSPA8
FAM20C
PLAUR
TXN
COL12A1
SHISA5
BMP4
SRC
CHGB
APLP2
GOLM1
MAPK8
EFNB1
APOA1
FGF23
IL1B
MXRA8
PLG
HMOX1
VCAN
GPX7
IGFBP5
TIMP1
BMP1
CANX
SERPINC1
AKT1
CHRDL1
GLRX2
SERPINE1
TF
CYBB
PRKCA
CXCL12
PLAU
CSF1
SEC61A1
ADAM10
CST3
COL5A2
TNF
IGFBP4
IGFBP3
STAT3
XBP1
GANAB
MMP9
DHH
PRDX6
MIA3
TNC
KTN1
ELN
GPC3
MMP11
SCG2
LGALS1
THBS1
FGF2
C3
COL17A1
VWF
FBN1
LAMA3
AHSG
IGF1
SDC2
FSTL1
COL3A1
TIMP2
SHH
HYOU1
MSLN
IHH
SPP1
COL5A1
PCSK9
GPX4
CD38
POSTN
ADAM17
LOXL2
SOD1
MMP2
PDIA3
ERO1LB
PDIA6
TMX1
P4HB
QSOX1
Peroxidase activity && Response to hypoxia
Protein disulfide isomerase activity

Figure3

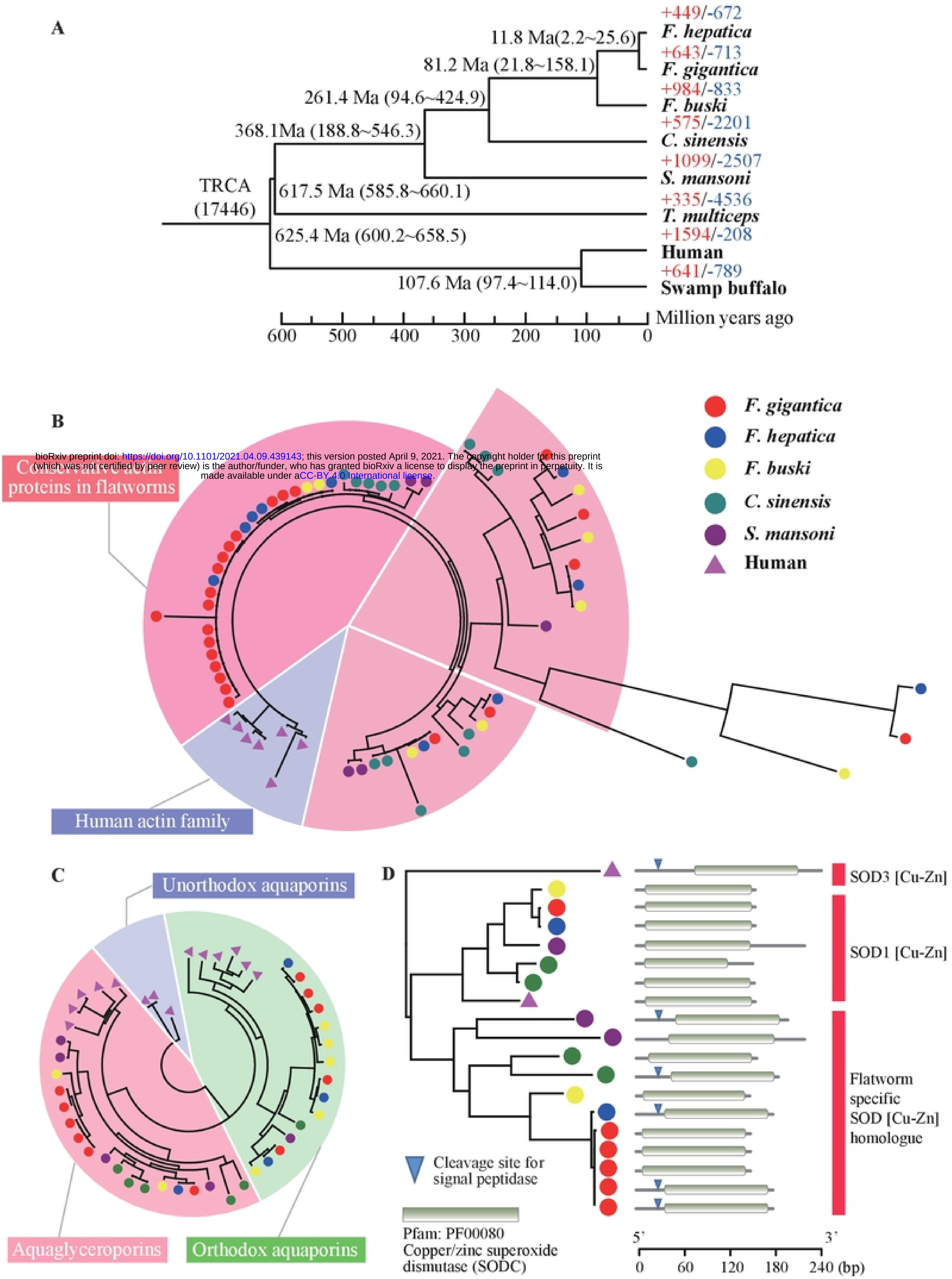

A
+449/-672
F. hepatica
11.8 Ma(2.2~25.6)
+643/-713
F. gigantica
81.2 Ma (21.8~158.1)
+984/-833
F. buski
261.4 Ma (94.6~424.9)
+575/-2201
C. sinensis
368.1Ma (188.8~546.3)
+1099/-2507
S. mansoni
TRCA
(17446)
617.5 Ma (585.8~660.1)
+335/-4536
T. multiceps
625.4 Ma (600.2~658.5)
+1594/-208
Human
+641/-789
Swamp buffalo
107.6 Ma (97.4~114.0)
600
500
400
300
200
100
0
Million years ago
B
Conservative actin proteins in flatworms
Human actin family
F. gigantica
F. hepatica
F. buski
C. sinensis
S. mansoni
Human
C
Unorthodox aquaporins
Aquaglyceroporins
Orthodox aquaporins
D
SOD3 [Cu-Zn]
SOD1 [Cu-Zn]
Flatworm specific SOD [Cu-Zn] homologue
Cleavage site for signal peptidase
Pfam: PF00080 Copper/zinc superoxide dismutase (SODC)
5'
3'
0
60
120
180
240 (bp)

Figure4

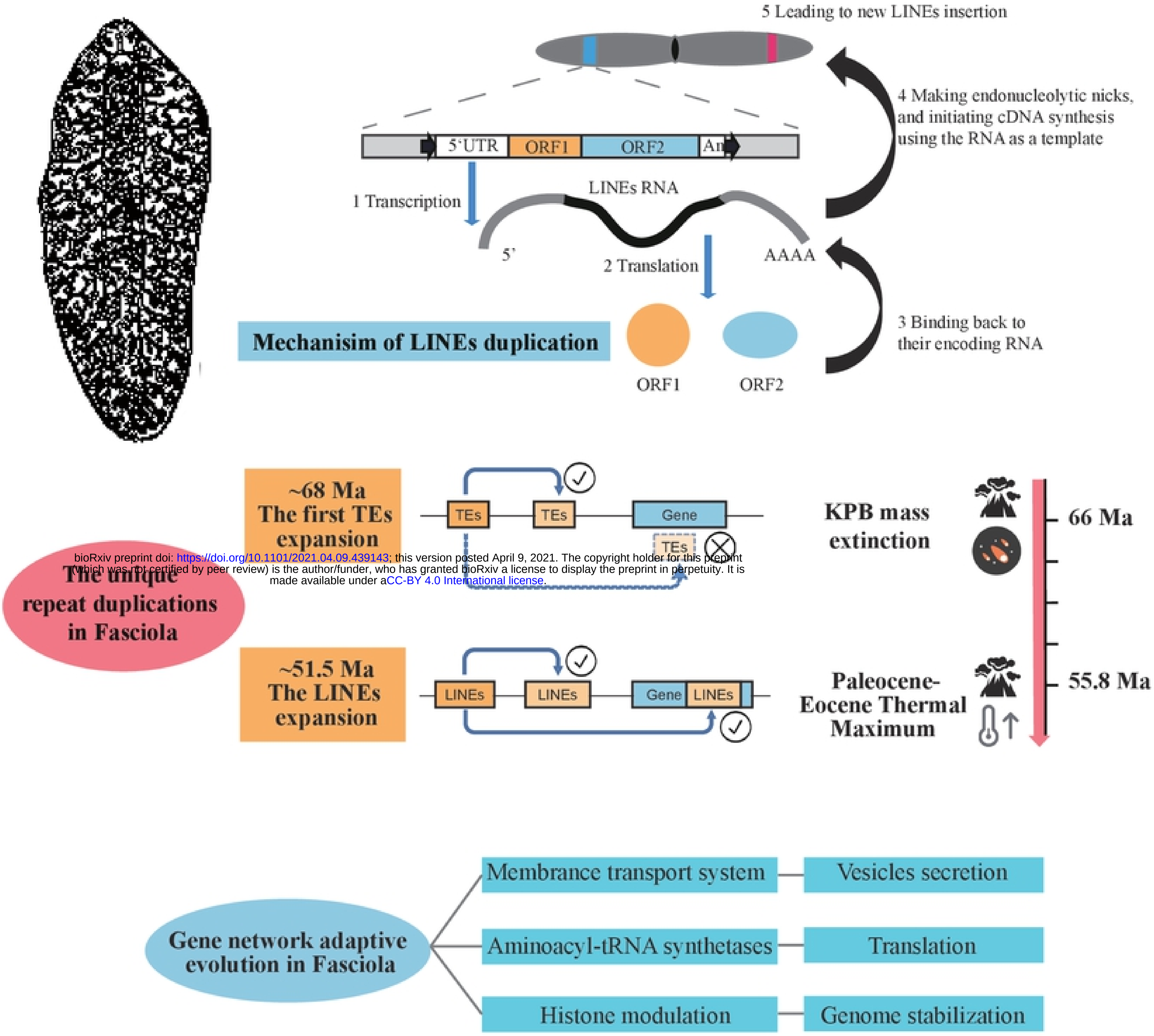

5 Leading to new LINEs insertion
5'UTR
ORF1
ORF2
An
4 Making endonucleolytic nicks, and initiating cDNA synthesis using the RNA as a template
1 Transcription
LINEs RNA
5'
AAAA
2 Translation
Mechanisim of LINEs duplication
ORF1
ORF2
3 Binding back to their encoding RNA
~68 Ma The first TEs expansion
TEs
TEs
Gene
TEs
KPB mass extinction
66 Ma
The unique repeat duplications in Fasciola
~51.5 Ma The LINEs expansion
LINEs
LINEs
Gene
LINEs
Paleocene-Eocene Thermal Maximum
55.8 Ma
Gene network adaptive evolution in Fasciola
Membrance transport system
Vesicles secretion
Aminoacyl-tRNA synthetases
Translation
Histone modulation
Genome stabilization

Figure5