PLOS NEGLECTED TROPICAL DISEASES

RESEARCH ARTICLE

An improved genome assembly of the fluke *Schistosoma japonicum*

Fang Luo[1]☯, Mingbo Yin[1,2]☯, Xiaojin Mo[2]☯, Chengsong Sun[1], Qunfeng Wu[1], Bingkuan Zhu[1], Manyu Xiang[1], Jipeng Wang[1], Yi Wang[1], Jian Li[1], Ting Zhang[2], Bin Xu[2], Huajun Zheng[3], Zheng Feng[2], Wei Hu[1,2]*

1 Department of infectious diseases, Huashan Hospital, State Key Laboratory of Genetic Engineering, Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, Ministry of Education Key Laboratory of Contemporary Anthropology, School of Life Sciences, Fudan University, Shanghai, China, **2** National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention, Key Laboratory of Parasite and Vector Biology of China Ministry of Health, WHO Collaborating Centre for Tropical Diseases, Joint Research Laboratory of Genetics and Ecology on Parasite-host Interaction, Chinese Center for Disease Control and Prevention & Fudan University, Shanghai, China, **3** Shanghai-MOST Key Laboratory of Health and Disease Genomics, Chinese National Human Genome Center at Shanghai, Shanghai, China

☯ These authors contributed equally to this work.
* huw@fudan.edu.cn

Citation: Luo F, Yin M, Mo X, Sun C, Wu Q, Zhu B, et al. (2019) An improved genome assembly of the fluke *Schistosoma japonicum*. PLoS Negl Trop Dis 13(8): e0007612. https://doi.org/10.1371/journal.pntd.0007612

Editor: Alessandra Morassutti, PUCRS, BRAZIL

Received: March 24, 2019

Accepted: July 8, 2019

Published: August 7, 2019

Data Availability Statement: The raw sequencing data are available via NCBI under SRA accessions PRJNA515567. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession SKCS00000000. The version described in this paper is version SKCS01000000.

Funding: This research was funded by the National Natural Science Foundation of China (No. 31725025 and 91431104) to Wei Hu. The website of the funder " the National Natural Science Foundation of China (NSFC)" is " https://isisn.nsfc.gov.cn". The funders had no role in study design,

Abstract

Background

Schistosoma japonicum is a parasitic flatworm that causes human schistosomiasis, which is a significant cause of morbidity in China and the Philippines. A single draft genome was available for *S. japonicum*, yet this assembly is very fragmented and only covers 90% of the genome, which make it difficult to be applied as a reference in functional genome analysis and genes discovery.

Findings

In this study, we present a high-quality assembly of the fluke *S. japonicum* genome by combining 20 G (~53X) long single molecule real time sequencing reads with 80 G (~ 213X) Illumina paired-end reads. This improved genome assembly is approximately 370.5 Mb, with contig and scaffold N50 length of 871.9 kb and 1.09 Mb, representing 142.4-fold and 6.2-fold improvement over the released WGS-based assembly, respectively. Additionally, our assembly captured 85.2% complete and 4.6% partial eukaryotic Benchmarking Universal Single-Copy Orthologs. Repetitive elements account for 46.80% of the genome, and 10,089 of the protein-coding genes were predicted from the improved genome, of which 96.5% have been functionally annotated. Lastly, using the improved assembly, we identified 20 significantly expanded gene families in *S. japonicum*, and those genes were primarily enriched in functions of proteolysis and protein glycosylation.

Conclusions

Using the combination of PacBio and Illumina Sequencing technologies, we provided an improved high-quality genome of *S. japonicum*. This improved genome assembly, as well

data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

as the annotation, will be useful for the comparative genomics of the flukes and more importantly facilitate the molecular studies of this important parasite in the future.

Author summary

Schistosomiasis is an acute and chronic disease that remains one of the most prevalent and serious of the parasitic diseases in the world. Three major *Schistosoma* species cause human schistosomiasis, including *Schistosoma japonicum*, *S. mansoni* and *S. haematobium*. However, the three schistosome references or draft genomes were released in the last decade, which greatly facilitate the progress in the whole research field of schistosome. However, limited by the sequencing technique and mixture samples at that time, only a genome draft was suppled to *S. japonicum*, which is fragmented and difficult to be a reference in functional genome analysis and gene discovery. Here, using the combination of PacBio and Illumina Sequencing technologies, we present a high-quality assembly of *S. japonicum* with contig and scaffold N50 length of 871.9 kb and 1.09 Mb, representing 142.4-fold and 6.2-fold improvement over the released WGS-based assembly, respectively. The assembly genome with high quality will certainly supply a new reference genome of *S. japonicum* and be beneficial to functional genomic and comparative genomics of schistosome, as well as other helminths.

Introduction

Schistosomiasis is an acute and chronic disease that remains one of the most prevalent and serious of the parasitic diseases [1]. Globally, an estimated 700 million people are at risk of infection and more than 250 million people are affected in 78 countries and territories [2, 3]. Three major *Schistosoma* species cause human schistosomiasis, including *Schistosoma japonicum*, *S. mansoni* and *S. haematobium*. *S. japonicum* is mainly epidemic in South China, Indonesia and the Philippines [4, 5], with 147,642 patients being treated in China in 2016 [6]. *S. japonicum* has a complex life-cycle that involves an aquatic snail (*Oncomelania hupensis*) as an intermediate host [7] and a wide definitive host range, infecting humans as well as more than 40 other mammals, and these make it difficult to control [8–10]. Moreover, *S. japonicum* is the most pathogenic among human *Schistosoma* due to a high fecundity. Each pairing of *S. japonicum* deposits up to 3,000 eggs per day, which is 10-fold greater than both *S. mansoni* or *S. haematobium* [11, 12]. Despite the remarkable success of schistosomiasis control over the past 60 years, this disease remains a major public health problem in South China, Indonesia and the Philippines [4, 13].

Three *Schistosoma* draft genomes (*S. mansoni*, *S. haematobium*, *S. japonicum*) have been published. The genome of *S. mansoni* was first published in 2009 with the whole genome shotgun (WGS) sequencing strategy, then improved with second generation DNA, Single-molecule real time (SMRT) sequencing technology and genetic markers. The latest genome V7 is 409.6 Mb with a N50 scaffold length of 50.5 Mb, possessing 95.9% of the genome assembled into chromosomes [14–16]. The genome of *S. haematobium* was assembled using Illumina-based technology, achieving an assembly of 385 Mb (29,834 scaffolds; N50 scaffold size of 317 kb) [17]. The genome draft of *S. japonicum* was generated by WGS sequencing strategy with mixed and outbred adult worms. This assembly was highly fragmented, including 25,048 scaffolds and 95,269 contigs with contig and scaffold N50 length of 6.12 kb and 176.9 kb [18]. Therefore, it is difficult to serve as the reference genome for curated gene prediction, as well as comparative and functional genomic analysis. An improved high-quality genome assembly of *S. japonicum* is urgently required.

SMRT sequencing technology from Pacific Biosciences (PacBio) can provide an opportunity to significantly improve genome assembly [19–21]. In the present study, we utilized PacBio and Illumina sequencing data from clonal worms to generate the genome of *S. japonicum*. Then, we re-annotated the improved genome assembly by additionally applying RNA sequencing data and performed comparative genomic analysis for *S. japonicum* and other six flatworms. Our improved assembly and annotation will provide valuable genomic resource for future studies, especially functional genomic analysis for *S. japonicum*.

Materials and methods

Ethics statement

All protocols involving animals were performed based on the guidelines of the Association for Assessment and Accreditation of Laboratory Animal Care International. The study procedures followed institutional ethical guidelines that were approved by the ethics committee at the National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention (NIPD, China CDC; Permit No: IPD2008-4).

Sample collection and sequencing

Oncomelania hupensis snails were individually exposed to a single miracidia of *S. japonicum* in tissue culture plates for 2 hours. After group cultivating for 3 months in small plastic aquaria at 25˚C, each snail was checked individually for cercarial release under a strong light. Those releasing cercariae were housed subsequently in single centrifuge tube of pond water marked as "clonal". The "clonal" cercariae from each single miracidium infection of *O. hupensis* snail were then used to infect Kunming mice, using 100–150 cercariae per mouse. After 30 days, 100–120 adult male worms and 60–80 adult female worms were collected by perfusing the hepatic portal system and mesenteric veins for genome sequencing, respectively. Genomic DNA (gDNA) extraction were carried out using the DNeasy Blood& Tissue Kit (Qiagen, Germany) according to manufacturer instruction, which generate 2.0 ug gDNA of male worms and 1.0 ug gDNA of female worms. The gDNA was quantified by applying Qubit dsDNA HS Assay Kit (Invitrogen, Thermo Fisher Scientific, Waltham, USA) and then by gel electrophoresis on 1% agarose gel, NanoDrop 2000/c Spectrophotometers (Thermo Fisher Scientific, Waltham, USA) and Agilent Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA).

Three 20-kb SMRTbell libraries were constructed using BluPippin Size Selection System protocol and sequenced on PacBio RS II platform using P5-C3 chemistry. Two male libraries (two clones) were sequenced in 12 SMRT cells while one female library (one clone) was sequenced in 2 SMRT cells. An Illumina paired-end library with 350 base pair (bp) insert size (PE350) was also constructed using the male gDNA and sequenced on Illumina platform for genome size estimation, correction of genome assembly and assembly evaluation.

RNA was extracted from three developmental stages of *S. japonicum* (i.e. adult, sporocyst and cercariae) using RNAiso Plus, respectively. Three paired-end (PE) RNA-seq libraries were generated using TruSeq RNA Library Preparation Kit v2 with different insert sizes (150 bp for adult and 250 bp for sporocyst and cercariae) and then sequenced on Illumina platform.

Genome estimation and assembly

Pacbio sequencing reads were analyzed for sequence quality with FastQC v0.11.8 [22]. The adaptor contamination, PCR duplication, low-quality reads and mice contamination of all

Illumina sequencing data were filtered out using Trimmomatic v0.36 [23] before genome estimation and assembly. To estimate genome of *S. japonicum*, jellyfish v1.1.12 was used to construct a k-mer frequency spectrum (k = 21) with PE350 cleaned sequencing data, and GenomeScope [24] were then performed to estimate the genome based on 21-kmer frequency.

The *de novo* genome assembly was executed by wtdbg v1.1.006 [25]. The WTDBG assembled raw reads without error correction and built the consensus from intermediate assembly output. Therefore, wtdbg-cns, minimap [26] and map2dbgcns in WTDBG were applied for initial error-correction process. Further, we polished the consensus genome assembly using pbalign v0.3.1 and Arrow v2.2.1 (Pacific Biosciences) with SMRT long reads, followed by polishing with pilon v1.22 [27] using PE350 clean sequencing reads. Finally, P_RNA_scaffolder [28] was used for scaffolding genomes with 150-bp library RNA-seq reads. Assembly statistics of final assembled genome (V2) were assessed with QUAST v4.6.3 [29].

Genome evaluation

To evaluate the completeness and accuracy of the V2 assembly of *S. japonicum*, we mapped PE350 clean data to both the V2 assembly and the first genome (v1) of *S. japonicum* [18] using bwa v0.7.12 [30]. Average sequencing depth and mapping rates were calculated using SAMTOOLS v1.8 [31]. The completeness of the coding gene sets was also evaluated by benchmarking universal single-copy orthologous genes (BUSCO v3) [32]. Last, a synteny analysis was also performed using SyMAP v4.2 [33] to further assess the quality of V2 assembly, considering only scaffolds of at least 100 kb and ordering the *S. japonicum* scaffolds based on chromosome-level *Schistosoma mansoni* genome V7 [14, 15] form WormBase ParaSite database [34]. We determined synteny for *S. mansoni* against both the v1 and v2 assembly for comparison. The genomes of *S. mansoni*, *S. haematobium* and other organisms for comparison were retrieved in WormBase Parasite database (S1 Table) [34].

Protein coding gene prediction and ncRNA prediction

By applying RepeatMasker v4.0.7 [35] with Repbase [36]and a *de novo* repeat database constructed with RepeatModeler v1.0.11 [37], the repeats elements were identified in V2 assembly before protein-coding gene prediction. Gene prediction was conducted by combining *ab initio* prediction, homology-based prediction and transcriptome-based prediction. AUGUSTUS v2.5.5 [38], SNAP v2006-07-28 [39] and GeneMark-ES v4.33 [40] with default parameters were used for the *ab initio* gene prediction in the repeat-masked genome. GeMoMa v2.3 were applied for the homology-based gene prediction, and the protein repertoires of flatworms including *S. japonicum* (GCA_000151775.1) [18], *S. mansoni* (GCA_000237925.2) [14, 15], *S. haematobium* (GCA_000699445.1) [17], *Clonorchis sinensis* (GCA_000236345.1) [41], *Opisthorchis viverrini* (GCF_000715545.1) [42], *Hymenolepis microstom* (GCA_000469805.2) [43], *Echinococcus multilocularis* (GCA_000469725.3) [43], *Echinococcus granulosus* (GCA_000524195.1) [44] from GenBank were used as the references. For the transcriptome-based gene prediction, All RNA-seq data were assembled by Trinity v2.7.0 [45]. The Trinity assembly contain 270,293 transcripts with N50 length of 1,155 bp. These assembled sequences plus 106,621 expressed sequence tags (EST) and mRNA from GenBank were aligned against the V2 assembly by Program to Assemble Spliced Alignment (PASA) [46, 47]. Valid transcript alignments were clustered based on genome mapping location and assembled into gene models. Besides, RNA-seq reads were also directly mapped to the genome by HISAT2 v2.1.0 [48] and assemble into gene models by StringTie v1.3.4 [49]. Gene models generated from all the above methods were integrated by EvidenceModeler (EVM) [49]. The gene models were further updated to generate untranslated regions (UTRs) and alternative splicing variation by PASA

[50, 51]. Finally, limit manual refinement of genome annotations was perform using Apollo [52] to fix reading frames. Then, the predicted genes length were summarized with an R package "GenomicFeatures" [53].

In addition, the non-coding RNAs (ncRNAs) including miRNA, rRNA, snRNA, and tRNA were also predicted. tRNAs were annotated by tRNAscan-SE v1.3.1 with default parameters for eukaryotes [54], snRNAs were extracted by INFERNAL v1.1.2 [55] against the RFAM v14.0 database [56], rRNAs were predicted by RNAmmer v1.2 [57], miRNA were annotated with deep-sequencing data from Sequence Read Archive (SRA) (SRR2927289) using miRDeep2 [58].

Gene functions of protein-coding genes were annotated by a successive blastp v2.7.1 [59] analysis against Swiss-Prot [60], TrEMBL and non-redundant (NR) databases [61] with an *E*-value cut-off of 10^{-5}. Gene domains were annotated using InterProScan5 [62] and HMMER v3.2.1 [63] against the PFAM database [64]. Gene Ontology (GO) [65, 66] annotations were generated combining the results from InterProScan5 and blastp analysis against NR database using BLAST2GO_CLI v1.15 [67]. Additional functional information was also derived via pathway analysis based on homology to the characterized pathways in Kyoto Encyclopaedia of Genes and Genomes (KEGG) [68] with online KAAS [69].

Gene family identification and phylogenetic analysis

To investigate gene family evolution in the *S. japonicum* genome, nucleotide and protein sequence of *S. mansoni*, *S. haematobium*, *C. sinensis*, *O. viverrini*, *Fasciola hepatica* and *H. diminuta* were retrieved from WormBase ParaSite databases [34]. Only the proteins from the longest transcripts were retained for each gene locus with alternative splicing variants. Those proteins with length $\geq$ 20 aa were used to calculate pair-wise similarities all-against-all BLASTp with *E*-value cut-off of $1e^{-10}$, and low-quality hits (coverage < 50%) were removed. Orthologous groups were further constructed by OrthoMCL v2.0.9 [70] with an inflation parameter of 1.5.

A total of 2,322 Single copy genes were retrieved for divergence time estimation. Proteins from each single-copy families were aligned using MAFFT v7.407 [71] with the parameter of "—localpair—maxiterate 1000". The corresponding CDS alignments were back-translated from the corresponding protein alignments with pal2nal, followed by the removal of poorly aligned regions using trimAl v1.2 with automated parameter [72]. The curated alignments of each family were concatenated into a super alignment matrix using phyutility [73]. Divergence times were estimated from a refined concatenated CDS alignment using BEAST2 v2.5.1 with a strict clock model. Priors used calibrated Yule model, time calibration constrains with two previously estimated dates between Trematoda and Cestoda (~106 million years ago (Mya)) and between Opistorchiida and Schistosomatoidea (~70 Mya) [74], and default setting for other priors. Samples from the posterior were drawn every 1,000 steps over a total of 10,000,000 steps per MCMC run.

Gene family evolution

Gene family expansion and contraction analyses was performed with Computational Analysis of gene Family Evolution software (CAFÉ v4.0.1) [75] according to divergence times. For each gene family, CAFÉ generated a family-wide *P* value and A branch/node-specific Viterbi *P* value indicating a possible gene-family expansion or contraction event. In this study, family-wide *P*-value less than 0.05 and a branch/node 'Viterbi *P*-value' less than 0.001 was considered as a signature of significant expansion or contraction for a specific gene family and specific species, respectively. Enrichment of GO terms for unique and expanded gene families of *S.*

japonicum were identified using clusterProfiler package [76] using the Benjamini-Hochberg FDR correction. Significantly enriched GO terms were identified with corrected *P* value of < 0.05. KEGG pathway enrichment were performed with KOBAS v3.0 software [77, 78].

Accession numbers

The raw sequencing data are available via NCBI under SRA accessions PRJNA515567. This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession SKCS00000000. The version described in this paper is version SKCS01000000.

Results

Genome assembly

A total of 20 G PacBio sequences, which has an average length of 7.4 kb and average sequence quality (Phred score) of 10 (S1 Fig), and 80 G Illumina paired-end (PE) clean sequences were generated from the whole genome sequencing. Additionally, a total of 33 G PE clean sequences was also generated for RNA samples on the Illumina platforms (S2 Table). The estimated genome size was approximately 375,706,643 bp based on 21-kmer analysis from the PE350 library, with the main peak at a depth of approximately 157× (S2 Fig). In the 21-kmer frequency distribution, the second peak at approximately half the coverage value of the main peak, indicates the high heterozygosity (1.05%) of the *S*. *japonicum* genome. The final assembled *S*. *japonicum* genome had a total length of 369.9 Mb (98.5% of the estimated genome) with 1,789 scaffolds and 33.76% GC ratio (S3 Fig). The overall assembly statistics of the improved genome version (*S*. *japonicum* V2) were dramatically improved when compared with the previously released version (*S*. *japonicum* V1; S1 Table) [18]: contig N50 increased from 6.12 kb to 871.9 kb, the scaffold N50 increased from 176 kb to 1.09 Mb and the number of gaps decreased from 70,219 to 319 (Table 1, Fig 1A).

Genome quality evaluation

More than 90.3% of the high-quality PE350 reads could be mapped concordantly to the V2 assembly, which is an improvement when compared with those of the V1 assembly (S3 Table). Besides, approximately 95.7% of the V2 assembly had a sequencing depth > 10-fold (S4 Fig) indicating high accuracy at the nucleotide level of V2 assembly. Assessment of genome completeness using BUSCO analysis confirmed that 85.2% of the 303 core eukaryotic genes and 72.6% of the 978 core metazoan genes were completely presented in the V2 assembly (Fig 1B). The BUSCO results were similar with those of recent studies for other flatworms and represent an improvement over the V1 assembly (S4 Table). Synteny were computed between *S*. *japonicum* and the chromosome-level *S*. *mansoni* genomes V7 and both the V2 and V1 assemblies were used for comparison. The increased contiguity of V2 assembly allowed us to compute synteny between *S*. *japonicum* and *S*. *mansoni* genome, more than doubling the percentage of the genome covered by synteny blocks from 29% to 67% and increasing the size of synteny blocks, with 97 of 277 synteny blocks > 1 Mb in length (Fig 1C and 1D).

Gene annotation

Approximately 46.87% of V2 assembly was identified as repeat elements, which is similar with the V1 assembly (44.56%) and the close relatives *S*. *mansoni* (49.23%) and *S*. *haematobium* (42.83%). Long interspersed elements were the most predominant elements, which account for 19.89% in V2 assembly (S5 Table). By combining *de novo* prediction, homology-based prediction and transcriptome-based prediction, a total of 10,089 protein-coding genes and 16,936

Table 1. Genome assembly statistics for the improved genome of *S. japonicum* in comparison with three published *Schistosoma* genome. V1 indicated conventional capillary sequenced genome and V2 indicated our improved genome.

	S. japonicum V2	*S. japonicum* V1	*S. mansoni*	*S. haematobium*
Genome size (bp)	369,900,518	402,743,189	409,579,008	375,894,156
Number of scaffolds	1,789	25,048	320	29,834
Number of contigs	2,108	95,267	602	59,195
Longest scaffold (bp)	6,264,197	1,730,213	88,881,357	1,826,302
Average scaffold length (bp)	210,145	16,078	1,279,934	12,560
Number of scaffolds: >10 kb	1,052	4,707	318	2,384
Number of Gaps	319	70,219	282	29,361
Scaffold N50 (bp)	1,093,989	176,869	50,458,499	317,484
Contig N50 (bp)	871,911	6,121	5,339,380	22,446
GC content (%)	33.76	34.08	35.47	34.22
Repeat content (%)	46.87	44.56	49.23	42.83

https://doi.org/10.1371/journal.pntd.0007612.t001

transcripts were predicted in the V2 assembly. Of the 10,089 protein-coding genes in the V2 assembly, 9,387 (93.0%) were supported by RNA-Seq clean data and 1,207 were newly detected genes. The number of predicted genes in V2 assembly is significantly lower (79.2%) than those in V1 assembly. Furthermore, BUSCO analysis with the metazoan_odb9 database [32] showed that the proportion of complete genes increased from 67.4% in the V1 assembly to 81.8% in our V2 assembly, while the ratio of the fragmented genes decreased from 15% to 4.3% (Fig 1B, Table 2). These results indicated that the large number of contigs in V1 assembly comprised a substantial number of fragmented or misassembled sequences, resulting in an overestimate of the number of unique protein-coding genes. The average gene length and average coding DNA sequence (CDS) size were 18,370 bp and 1,537 bp, which were longer than those in V1 gene annotation (Table 2). Additionally, the gene number, gene length distribution, CDS length distribution, exon length distribution and intron length distribution were similar with those in other Trematoda species (Fig 2). Among the 10,089 predicted genes, 9,291 (92.1%) genes had matches in the NR database, 6,642 (65.8%) generated hits to Swiss-Prot database, 8,689 (86.1%) were identified in InterPro, 8,368 (82.9%) and 4,547 (45.1%) were assigned GO terms and KEGG pathways (Table 3, S6 Table). Besides, four types of non-coding RNAs were also identified, including 172 miRNAs, 1,263 tRNAs, 10 rRNAs and 54 snRNAs (Table 3).

Genome family evolution

Comparative analysis between *S. japonicum* and other six Platyhelminthes species was conducted. In brief, 12,001 gene families were constructed in all 7 species using OrthoMCL [70], 2,322 gene families were identified as single-copy orthologous gene families, and 3,798 gene families were common to all 7 species. Furthermore, 8,278 orthologous genes were detected in *S. japonicum* genome, and 103 gene families corresponding 351 genes that were specific to *S. japonicum* (Fig 3B, S7 Table). These species-specific genes were significantly enriched in molecular functional categories related to RNA-directed DNA polymerase activity, endonuclease activity, nucleic acid binding, ribonuclease T2 activity and metalloendopeptidase activity (S8 Table).

The phylogenetic analysis based on 2,322 single-copy orthologs showed that the divergence time of *S. japonicum*-*S. mansoni* and *S. mansoni*-*S. haematobium* occurred at ~14 Mya and ~4 Mya (Fig 3A), which are consistent with previous estimates [79, 80]. 20 significantly expanded and 5 significantly contracted gene families were identified in *S. japonicum* (S9 Table). The genes from these significantly expanded families were mainly enriched in nucleic acid binding,

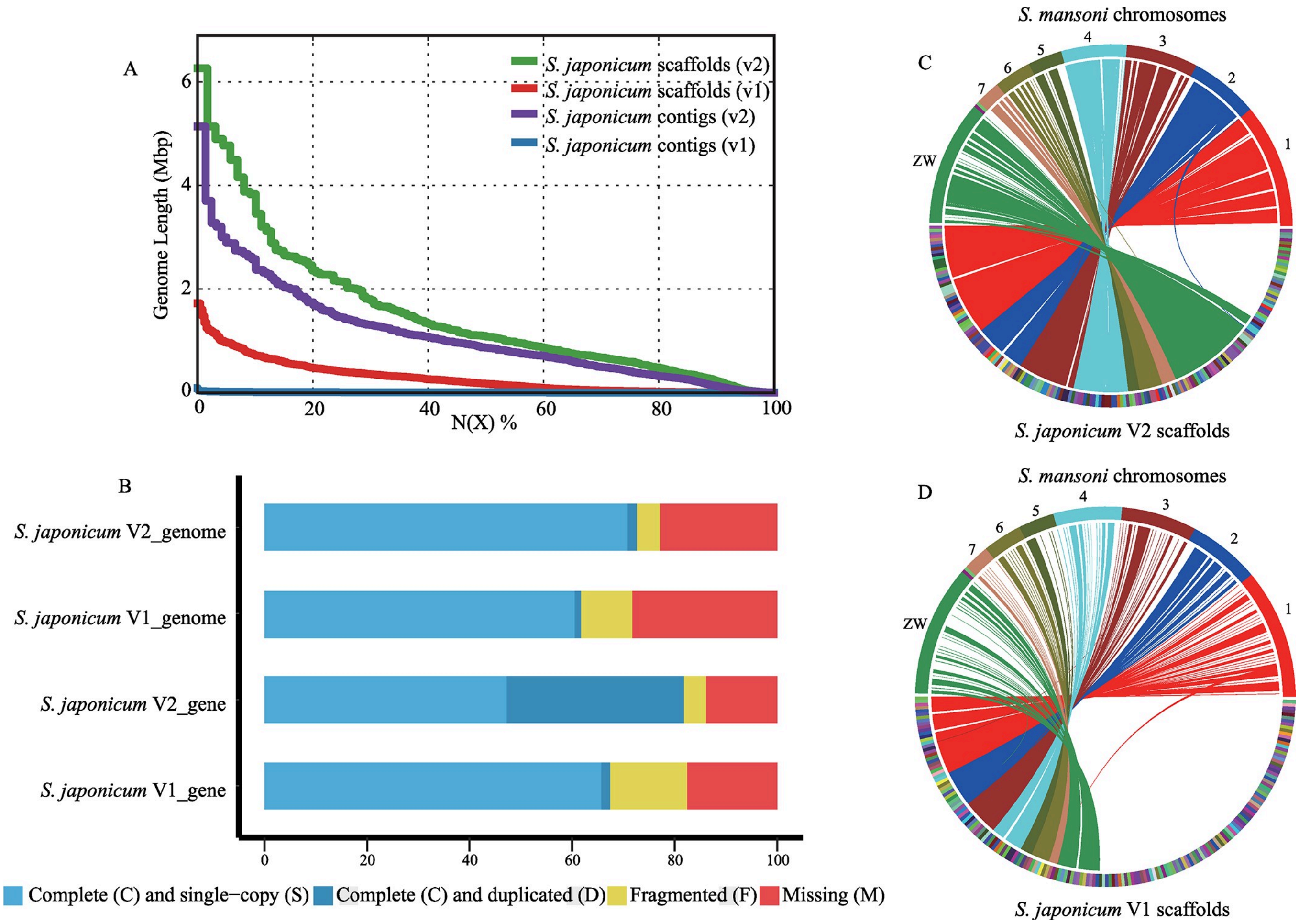

Fig 1. Genome assembly of *Schistosoma japonicum*. (A) Comparison of contiguity between the two versions of *S. japonicum* genome assembly. N(x)% graph shows the contig and scaffold sizes (y-axis), where x% of the genome assembly consists of contigs and scaffolds of at least that size. (B) comparison between two version of *S. japonicum* genome assembly, showing the portions of the genomes that are complete (blue), fragmented (yellow) or missing (red), as determined by benchmarking universal single-copy orthologs (BUSCO) analysis with metazoan_odb9 database. (C) Circle plot of synteny between the second version of *S. japonicum* genome and *S. mansoni* genome V7 made using SyMAP. It shows a high degree of synteny, with many long *S. japonicum* scaffolds covering significant portions of *S. mansoni* chromosome. (D) Circle plot of synteny between the first version of *S. japonicum* genome and *S. mansoni* genome. V1 indicated conventional capillary sequenced genome and V2 indicated our improved genome.

https://doi.org/10.1371/journal.pntd.0007612.g001

Table 2. Comparison of predicted genes of the two version of *S. japonicum* genome assembly. V1 indicated conventional capillary sequenced genome and V2 indicated our improved genome.

	S. japonicum (V2)	*S. japonicum* (V1)
Gene number	10,089	12,738
Average gene length (bp)	18,370	9,960
Average CDS length (bp)	1,537	1,172
Average exons per gene	8.3	5.3
Average exon length (bp)	370	223
Average intron length (bp)	2,521	2,058
BUSCO analysis		
Complete	81.8%	67.4%
partial	4.3%	15.0%
Missing	13.9%	17.6%

https://doi.org/10.1371/journal.pntd.0007612.t002

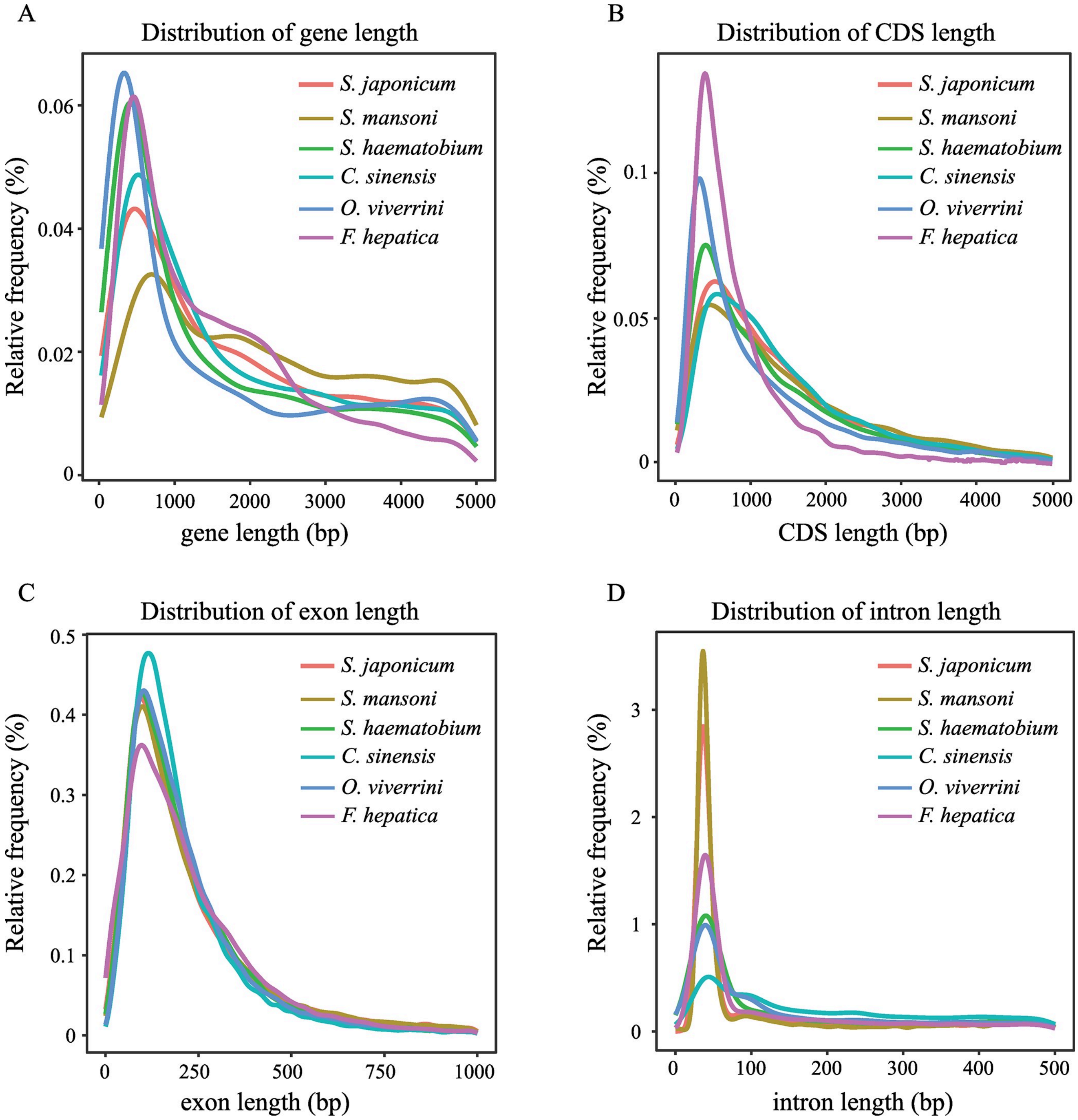

Fig 2. Length distribution comparison on total gene, CDS, exon, and intron of annotated gene models of the *S. japonicum* with other closely related Trematoda species. Length distribution of total genes (A), CDS (B), exon (C), and intron (D) were compared to those of *S. mansoni*, *S. haematobium*, *C. sinensis*, *O. viverrini*, and *F. hepatica*.

https://doi.org/10.1371/journal.pntd.0007612.g002

polypeptide N-acetylgalactosaminyltransferese activity, 5'-nucleotidase activity, cysteine-type endopeptidase activity, galactosyltransferase activity, adenylate cyclase activity and ribonuclease T2 activity (Fig 4, S10 Table). In addition, these expanded genes were also significantly

Table 3. Annotation of protein-coding genes and noncoding RNA elements in the improved *S. japonicum* genome assembly.

	Number (%)
Protein annotations	
SWISSPROT	6,642 (65.8)
TrEMBL	6,137 (60.8)
NCBI nr database	9,291 (92.1)
KEGG database	4,547 (45.1)
InterProScan	8,689 (86.1)
Gene ontology annotation	8,368 (82.9)
Conserved noncoding RNA elements	
Small nuclear RNA (snRNA)	54
Transfer RNA (tRNA)	1,263
Micro RNA (miRNA)	172
Ribosomal RNA (rRNA)	10

https://doi.org/10.1371/journal.pntd.0007612.t003

enriched in the KEGG pathway related to mucin type O-Glycan biosynthesis (Corrected *P* value = 2.77e-06).

Discussion

A draft genome of *S. japonicum* was released in 2009 and provided a resource for gene discovery and data mining. However, the genome was generated using WGS sequencing strategy with mixed and outbred adult worms, resulting its highly fragments in the assembly. Here, by applying combination of the long-reads PacBio and short-reads Illumina sequencing data

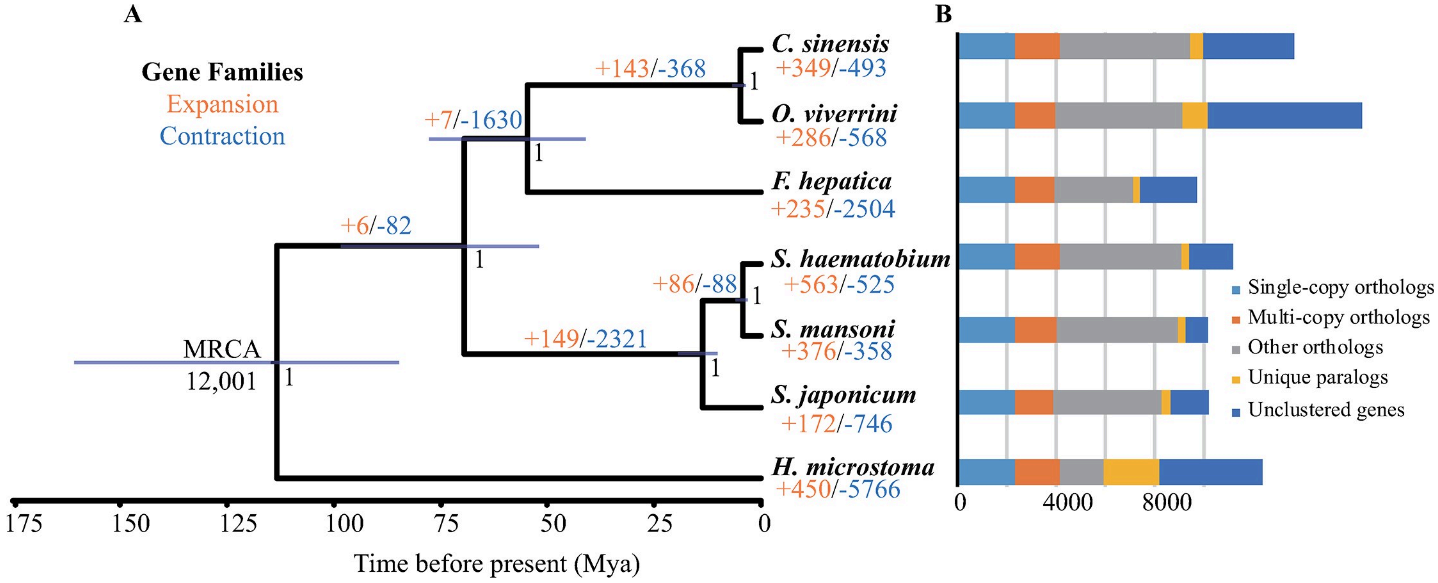

Fig 3. Comparative genome analysis between *S. japonicum* and other six flatworms. (A) Phylogenetic tree and expansion and contraction of gene families. The phylogenetic tree and divergence time were generated from 2,322 single-copy orthologous genes using BEAST2. The branch lengths of the phylogenetic tree are scaled to estimated divergence time. Tree topology is supported by posterior probability of 1.0 for all nodes. The blue bars on the nodes indicate the 95% credibility intervals of the estimated posterior distributions of the divergence times. The number of expanded (orange) and contracted (blue) gene families is designated on each branch. Bar charts indicates the orthologous and paralogous gene families in *S. japonicum* and other six flatworm species. (B) Comparison of the number of gene families in 7 Platyhelminthes species.

https://doi.org/10.1371/journal.pntd.0007612.g003

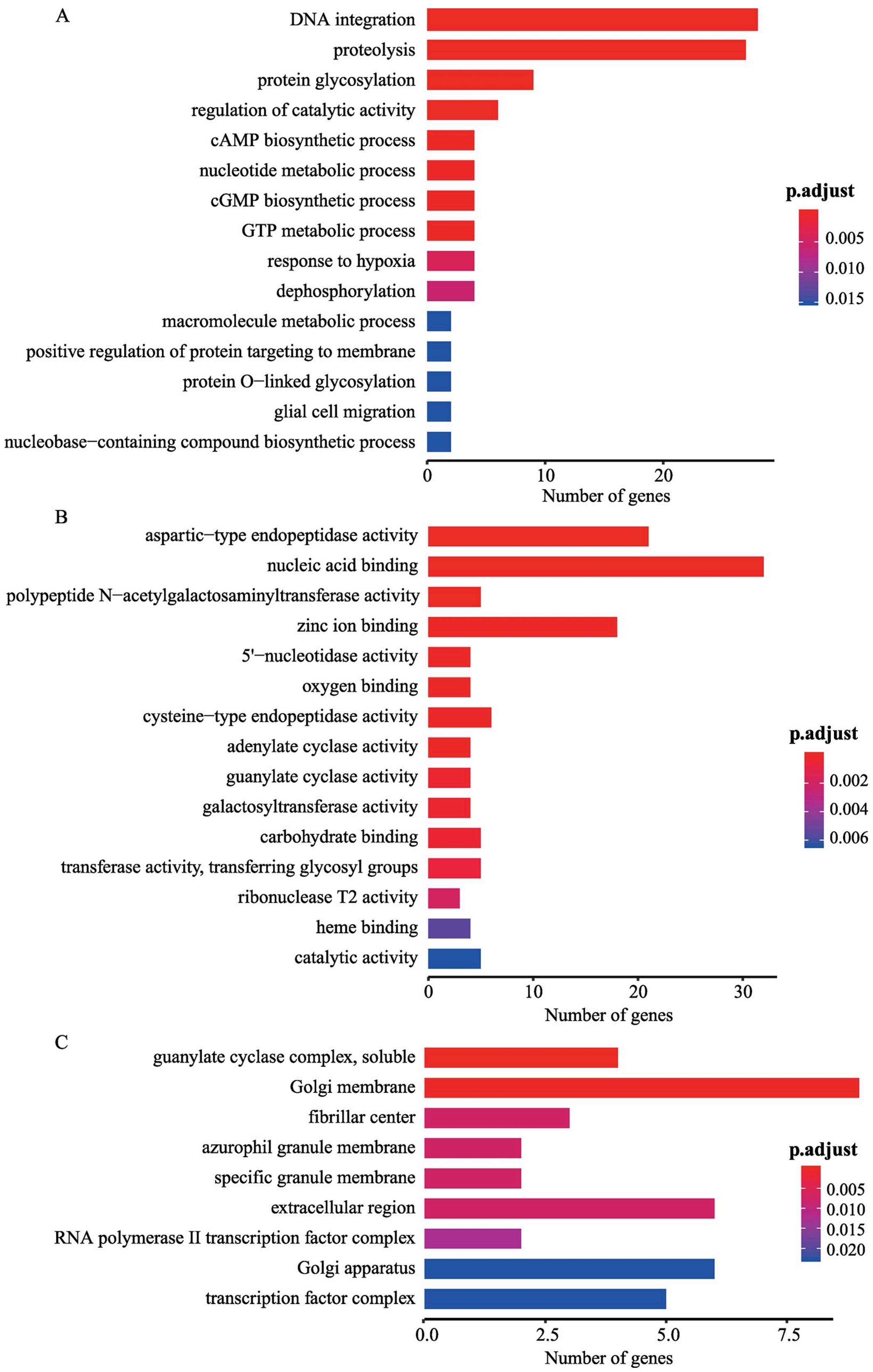

Fig 4. Gene Ontology enrichment analysis of significantly expanded gene families. (A) biological processes, (B) molecular function and (C) cellular component.

https://doi.org/10.1371/journal.pntd.0007612.g004

from the clonal worms, we provided an improved high-quality genome assembly (V2) of *S. japonicum*. When compared to the released WGS genome (V1) of *S. japonicum* [18], our V2 assembly showed significant improvements in assembly contiguity, accuracy and completeness, with 142.4-fold increase in contig N50, 6.2-fold increase in scaffold N50, 5% improvement in the mapping rate of PE reads and 10% improvement in the proportion of complete genes. Genome annotation has also upgraded by using deep coverage transcriptome data. Over 93.0% of the 10,089 protein-coding genes are supported by RNA-Seq data and 1,207 new genes were predicted in our improved assembly. Also, the length of protein-coding genes and completeness were increased significantly, when compared with those of the V1 assembly. Here, we found that genes typically had large introns (the average length of 2,521 bp) and much smaller exons (the average length of 370 bp). This pattern was also detected in *S. mansoni* [14] and in many other eukaryotes [81]. The large introns and much smaller introns might be attributable to the high activity of transposable elements [81].

Applying our improved V2 assembly, 20 significantly expanded gene families in *S. japonicum* were detected. Specifically, expansion of cathepsin B-like cysteine proteinase genes in *S. japonicum* could be associated with host invasion, hemoglobin degradation and immune invasion [82]. Another expanded gene family encoding Trematoda eggshell synthesis protein could compose the surface of the eggs, and thus protect the embryo from environmental challenges [83]. Previous study has shown that this family is essential for vitellarium development and egg production in *S. japonicum* [84]. Additionally, we detected expansions in the gene families that encode ribonuclease Oy, polypeptide N-acetylgalactosaminyltransferase (ppGalNAcTs) and N-acetylglucosaminyltransferase (UDP-GlcNAc). Ribonuclease Oy is an egg antigen molecule and induce a polarized Th2 type host immune response, which may enable eggs to escape from host tissues and initiate granuloma formation [85]. ppGalNAcTs and UDP-GlcNAc are important in the biosynthesis of glycan and glycoconjugates that interact with both the innate and adaptive arms of immunity in human and animal hosts [86, 87]. We detected 103 gene families (351 genes) that were specific to *S. japonicum*. Interestingly, they mainly related to the central metabolism, such as RNA-directed DNA polymerase activity and proteolysis. We assumed that the function of "proteolysis" was related with invasion and hemoglobin degradation of *S. japonicum* [88, 89]. Therefore, this function might lead to the wide host range of *S. japonicum*, when compared with other *Schistosoma* species [8]. The pao retrotransposon and reverse transcriptase, which was related with RNA-directed DNA polymerase activity, were regarded as principal forces driving the evolution of eukaryotic genomes [90]. Moreover, the species-specific duplication of the retrotransposon and reverse transcriptase were also observed in *S. haematobium* and *S. mansoni*, which imply that these genes could drive the genome divergence. Additionally, we found 32 egg proteins that were specific to *S. japonicum*; these proteins might be related with the egg formation and thus lead to a higher fertility, when compared to *S. mansoni* or *S. haematobium* [12]. These expanded and unique genes could be potential targets to investigate the molecular mechanisms of adaptations to diverse definitive host and high egg production of *S. japonicum*, and thus will provide the candidates for vaccine and drug targets.

Overall, our improved genome assembly of *S. japonicum*, together with its newly annotation, will serves as a framework for the functional analysis of *S. japonicum*, which may facilitate the development of new disease interventions for its control and eventual elimination.

Supporting information

S1 Fig. Distribution of quality score over all PacBio sequences.
(EPS)

S2 Fig. Graph of K-mer distribution (K = 21) generated from the PE350 library using GenomeScope. The big peak at the coverage of 157 in the graph is the homozygous portion of the genome, which accounts for the strands of the DNA having identical 21-kmers. The smaller shoulder to the left of the peak corresponds to the heterozygous portion of the genome, which accounts for the strands of the DNA having different 21-kmers. The *S*. *japonicum* genome size was estimated to be 375.7 Mb.
(EPS)

S3 Fig. The GC distribution of the new version of *S*. *japonicum* genome assembly.
(EPS)

S4 Fig. Sequencing depth distribution of the new version of *S*. *japonicum* genome assembly.
(EPS)

S5 Fig. Comparison of genome completeness between the genome of *S*. *japonicum* and other Trematoda species based on BUSCO evaluation, using either the eukaryote_odb9 database (A) and metazoa_odb9 database (B). The portions of the genomes that are complete (blue), fragmented (yellow), or missed (red), are determined by benchmarking universal single-copy orthologs (BUSCO) analysis for eukaryotic-conserved (left) and metazoan-conserved (right) genes.
(EPS)

S1 Table. Species used for comparison of genome parameters and genomic comparative analysis.
(DOCX)

S2 Table. Summary of improved genome sequencing data generated using PacBio and Illumina platform.
(DOCX)

S3 Table. Comparison of mapping rates for PE350 between the two versions of the *S*. *japonicum* genome assembly. V1 indicated conventional capillary sequenced genome and V2 indicated our improved genome.
(DOCX)

S4 Table. BUSCO results for completeness assessment for the first, second version assemblies of *S*. *japonicum* and genome of other Trematoda species, using either the eukaryote or metazoan databases.
(XLSX)

S5 Table. Composition of repetitive elements in the two version of *S*. *japonicum* genome assembly. V1 indicated conventional capillary sequenced genome and V2 indicated our improved genome.
(DOCX)

S6 Table. Annotations of the predicted gene models.
(XLSX)

S7 Table. Unique gene families in *Schistosoma japonicum* based on OrthoMCL analysis.
(XLSX)

S8 Table. Gene ontology (GO) enrichment analysis for unique gene families of *Schistosoma japonicum*. (MF: molecular function; CC: cell component; BP: biological process).
(DOCX)

S9 Table. Gene families that showed significant expansions in *Schistosoma japonicum* based on CAFE4 analysis. (sja: *S. japonicum*; smm: *S. mansoni*; shx: *S. haematobium*; ovi: *O. viverrini*; csi: *C. sinensis*; fhe: *F. hepatica*; hmi: *H. microstoma*).
(XLSX)

S10 Table. Gene ontology (GO) enrichment analysis for expanded gene families of *S. japonicum*.
(DOCX)

Acknowledgments

We thank Dave Dyer for linguistic modification and, Leandro de Mattos Pereira and two anonymous reviewers for useful comments on the earlier version of this manuscript.

Author Contributions

Conceptualization: Wei Hu.

Data curation: Fang Luo, Mingbo Yin, Xiaojin Mo, Chengsong Sun, Qunfeng Wu, Bingkuan Zhu, Manyu Xiang, Jipeng Wang, Huajun Zheng.

Formal analysis: Fang Luo, Mingbo Yin, Xiaojin Mo, Chengsong Sun, Qunfeng Wu, Bingkuan Zhu, Manyu Xiang, Jipeng Wang, Yi Wang, Ting Zhang, Bin Xu, Huajun Zheng, Wei Hu.

Funding acquisition: Wei Hu.

Methodology: Mingbo Yin, Zheng Feng, Wei Hu.

Project administration: Wei Hu.

Resources: Xiaojin Mo, Jian Li, Ting Zhang, Bin Xu.

Software: Fang Luo.

Supervision: Mingbo Yin, Zheng Feng, Wei Hu.

Visualization: Fang Luo.

Writing – original draft: Fang Luo.

Writing – review & editing: Mingbo Yin, Wei Hu.

References

1. Vos T, Abajobir AA, Abate KH, Abbafati C, Abbas KM, Abd-Allah F, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet. 2017; 390 (10100):1211–59. https://doi.org/10.1016/S0140-6736(17)32154-2 PMID: 28919117
2. Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. Lancet Infect Dis. 2006; 6 (7):411–25. https://doi.org/10.1016/S1473-3099(06)70521-7 PMID: 16790382
3. McManus DP, Dunne DW, Sacko M, Utzinger J, Vennervald BJ, Zhou X-N. Schistosomiasis. Nat Rev Dis Primers. 2018; 4(1):13. https://doi.org/10.1038/s41572-018-0013-8 PMID: 30093684
4. Zhou X-N, Wang L-Y, Chen M-G, Wu X-H, Jiang Q-W, Chen X-Y, et al. The public health significance and control of schistosomiasis in China—then and now. Acta Trop. 2005; 96(2):97–105.
5. Colley DG, Bustinduy AL, Secor WE, King CH. Human schistosomiasis. Lancet. 2014; 383 (9936):2253–64. https://doi.org/10.1016/S0140-6736(13)61949-2 PMID: 24698483
6. Li-Juan Z, Zhi-Min X, Ying-Jun Q, Hui D, Shan L, Jing X, et al. Endemic status of schistosomiasis in People's Republic of China in 2016. Chin J Schisto Control. 2017; 29(6):669–77.

7. Weerakoon KGAD, Gobert GN, Cai P, McManus DP. Advances in the diagnosis of human Schistosomiasis. Clin Microbiol Rev. 2015; 28(4):939. https://doi.org/10.1128/CMR.00137-14 PMID: 26224883
8. He Y-X, Salafsky B, Ramaswamy K. Host–parasite relationships of *Schistosoma japonicum* in mammalian hosts. Trends Parasitol. 2001; 17(7):320–4. PMID: 11423374
9. McGarvey ST, Zhou XN, Willingham Iii AL, Feng Z, Olveda R. The epidemiology and host–parasite relationships of *Schistosoma japonicum* in definitive hosts. Parasitol Today. 1999; 15(6):214–5. PMID: 10366824
10. Liu S, Zhou X, Piao X, Wu C, Hou N, Chen Q. Comparative analysis of transcriptional profiles of adult *Schistosoma japonicum* from different laboratory animals and the natural host, water buffalo. PLoS Negl Trop Dis. 2015; 9(8):e0003993. https://doi.org/10.1371/journal.pntd.0003993 PMID: 26285138
11. Shen J, Lai D-H, Wilson RA, Chen Y-F, Wang L-F, Yu Z-L, et al. Nitric oxide blocks the development of the human parasite *Schistosoma japonicum*. Proc Natl Acad Sci USA. 2017; 114(38):10214. https://doi.org/10.1073/pnas.1708578114 PMID: 28874579
12. Roberts L, Schmidt GD Jr. Janovy J. Foundations of Parasitology: McGraw-Hill Education; 2008.
13. Sun L-P, Wang W, Hong Q-B, Li S-Z, Liang Y-S, Yang H-T, et al. Approaches being used in the national schistosomiasis elimination programme in China: a review. Infect Dis Poverty. 2017; 6(1):55. https://doi.org/10.1186/s40249-017-0271-9 PMID: 28292327
14. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, et al. The genome of the blood fluke *Schistosoma mansoni*. Nature. 2009; 460:352. https://doi.org/10.1038/nature08160 PMID: 19606141
15. Protasio AV, Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, et al. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. PLoS Negl Trop Dis. 2012; 6(1):e1455. https://doi.org/10.1371/journal.pntd.0001455 PMID: 22253936
16. Holroyd N, Sanchez-Flores A. Producing parasitic helminth reference and draft genomes at the Wellcome Trust Sanger Institute. Parasite Immunol. 2012; 34(2-3):100–7. https://doi.org/10.1111/j.1365-3024.2011.01311.x PMID: 21707658
17. Young ND, Jex AR, Li B, Liu S, Yang L, Xiong Z, et al. Whole-genome sequence of *Schistosoma haematobium*. Nat Genet. 2012; 44:221. https://doi.org/10.1038/ng.1065 PMID: 22246508
18. *Schistosoma japonicum* genome sequencing functional analysis consortium. The *Schistosoma japonicum* genome reveals features of host-parasite interplay. Nature. 2009; 460(7253):345–51. https://doi.org/10.1038/nature08140 PMID: 19606140
19. Xia M, Han X, He H, Yu R, Zhen G, Jia X, et al. Improved *de novo* genome assembly and analysis of the Chinese cucurbit *Siraitia grosvenorii*, also known as monk fruit or luo-han-guo. GigaScience. 2018; 7 (6):giy067.
20. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017; 546:524. https://doi.org/10.1038/nature22971 PMID: 28605751
21. Jiao W-B, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, et al. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res. 2017; 27(5):778–86. https://doi.org/10.1101/gr.213652.116 PMID: 28159771
22. Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010.
23. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404
24. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017; 33(14):2202–4. https://doi.org/10.1093/bioinformatics/btx153 PMID: 28369201
25. Ruan J. wtdbg 2017. Available from: https://github.com/ruanjue/wtdbg.
26. Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. Bioinformatics. 2016; 32(14):2103–10. https://doi.org/10.1093/bioinformatics/btw152 PMID: 27153593
27. 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):e112963. https://doi.org/10.1371/journal.pone.0112963 PMID: 25409509
28. Zhu B-H, Xiao J, Xue W, Xu G-C, Sun M-Y, Li J-T. P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. BMC Genomics. 2018; 19(1):175. https://doi.org/10.1186/s12864-018-4567-3 PMID: 29499650
29. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013; 29(8):1072–5. https://doi.org/10.1093/bioinformatics/btt086 PMID: 23422339

30. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009; 25(14):1754–60. https://doi.org/10.1093/bioinformatics/btp324 PMID: 19451168

31. 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. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

32. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015; 31(19):3210–2. https://doi.org/10.1093/bioinformatics/btv351 PMID: 26059717

33. Soderlund C, Bomhoff M, Nelson WM. SyMAP v3.4: a turnkey synteny system with application to plant genomes. Nucleic Acids Res. 2011; 39(10):e68. https://doi.org/10.1093/nar/gkr123 PMID: 21398631

34. Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. WormBase ParaSite − a comprehensive resource for helminth genomics. Mol Biochem Parasitol. 2017; 215:2–10. https://doi.org/10.1016/j.molbiopara.2016.11.005 PMID: 27899279

35. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009; 25(1):4.10.1–4..4.

36. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 2015; 6:11. https://doi.org/10.1186/s13100-015-0041-9 PMID: 26045719

37. Smit A, Hubley R. RepeatModeler-1.0. 5 2012. Available from: http://www.repeatmasker.org/RepeatModeler/.

38. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006; 34(suppl_2):W435–W9.

39. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004; 5(1):59.

40. Besemer J, Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res. 2005; 33(suppl_2):W451–W4.

41. 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 Biol. 2011; 12(10):R107. https://doi.org/10.1186/gb-2011-12-10-r107 PMID: 22023798

42. Young ND, Nagarajan N, Lin SJ, Korhonen PK, Jex AR, Hall RS, et al. The *Opisthorchis viverrini* genome provides insights into life in the bile duct. Nat Commun. 2014; 5:4378. https://doi.org/10.1038/ncomms5378 PMID: 25007141

43. Tsai IJ, Zarowiecki M, Holroyd N, Garciarrubio A, Sanchez-Flores A, Brooks KL, et al. The genomes of four tapeworm species reveal adaptations to parasitism. Nature. 2013; 496:57. https://doi.org/10.1038/nature12031 PMID: 23485966

44. Zheng H, Zhang W, Zhang L, Zhang Z, Li J, Lu G, et al. The genome of the hydatid tapeworm *Echinococcus granulosus*. Nat Genet. 2013; 45:1168. https://doi.org/10.1038/ng.2757 PMID: 24013640

45. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29:644. https://doi.org/10.1038/nbt.1883 PMID: 21572440

46. 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 Res. 2003; 31(19):5654–66. https://doi.org/10.1093/nar/gkg770 PMID: 14500829

47. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008; 9(1):R7–R. https://doi.org/10.1186/gb-2008-9-1-r7 PMID: 18190707

48. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015; 12:357. https://doi.org/10.1038/nmeth.3317 PMID: 25751142

49. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015; 33:290. https://doi.org/10.1038/nbt.3122 PMID: 25690850

50. Haas BJ. Analysis of Alternative Splicing in Plants with Bioinformatics Tools. In: Reddy ASN, Golovkin M, editors. Nuclear pre-mRNA Processing in Plants. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 17–37.

51. Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics. 2006; 7(1):327.

52. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, et al. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 2013; 14(8):R93. https://doi.org/10.1186/gb-2013-14-8-r93 PMID: 24000942

53. Lawrence M, Huber W, Pagès H, Aboyoun P, Carlson M, Gentleman R, et al. Software for computing and annotating genomic ranges. PLoS Comp Biol. 2013; 9(8):e1003118.

54. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997; 25(5):955–64. https://doi.org/10.1093/nar/25.5.955 PMID: 9023104

55. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013; 29 (22):2933–5. https://doi.org/10.1093/bioinformatics/btt509 PMID: 24008419

56. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, et al. Non-coding RNA analysis using the Rfam database. Curr Protoc Bioinformatics. 2018; 62(1):e51. https://doi.org/10.1002/cpbi.51 PMID: 29927072

57. Lagesen K, Hallin P, Rødland EA, Stærfeldt H-H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007; 35(9):3100–8. https://doi.org/10.1093/nar/gkm160 PMID: 17452365

58. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012; 40(1):37–52. https://doi.org/10.1093/nar/gkr688 PMID: 21911355

59. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10(1):421.

60. Boeckmann B, Bairoch A, Apweiler R, Blatter M-C, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003; 31(1):365–70. https://doi.org/10.1093/nar/gkg095 PMID: 12520024

61. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007; 35(suppl_1):D61–D5.

62. 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. https://doi.org/10.1093/bioinformatics/btu031 PMID: 24451626

63. Eddy SR. HMMER: Profile hidden Markov models for biological sequence analysis. 2001.

64. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014; 42(D1):D222–D30.

65. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000; 25:25. https://doi.org/10.1038/75556 PMID: 10802651

66. The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Res. 2017; 45(Database issue):D331–D8. https://doi.org/10.1093/nar/gkw1108 PMID: 27899567

67. Conesa A, Götz S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. Int J Plant Genomics. 2008; 2008:12.

68. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017; 45(D1):D353–D61. https://doi.org/10.1093/nar/gkw1092 PMID: 27899662

69. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007; 35:W182–W5. https://doi.org/10.1093/nar/gkm321 PMID: 17526522

70. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. 2003; 13(9):2178–89. https://doi.org/10.1101/gr.1224503 PMID: 12952885

71. Katoh K, Standley DM. MAFFT multiple sequence alignment software Version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80. https://doi.org/10.1093/molbev/mst010 PMID: 23329690

72. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25(15):1972–3. https://doi.org/10.1093/bioinformatics/btp348 PMID: 19505945

73. Smith SA, Dunn CW. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics. 2008; 24(5):715–6. https://doi.org/10.1093/bioinformatics/btm619 PMID: 18227120

74. Oey H, Zakrzewski M, Narain K, Devi KR, Agatsuma T, Nawaratna S, et al. Whole-genome sequence of the oriental lung fluke *Paragonimus westermani*. GigaScience. 2018:giy146.

75. 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. https://doi.org/10.1093/molbev/mst100 PMID: 23709260

76. Yu G, Wang L-G, Han Y, He Q-Y. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS: J Integrative Biol. 2012; 16(5):284–7.

77. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011; 39(suppl_2):W316–W22.

78. Wu J, Mao X, Cai T, Luo J, Wei L. KOBAS server: a web-based platform for automated annotation and pathway identification. Nucleic Acids Res. 2006; 34(suppl_2):W720–W4.

79. Lawton SP, Hirai H, Ironside JE, Johnston DA, Rollinson D. Genomes and geography: genomic insights into the evolution and phylogeography of the genus *Schistosoma*. Parasite Vector. 2011; 4(1):131.

80. Attwood SW, Fatih FA, Mondal MMH, Alim MA, Fadjar S, Rajapakse RPVJ, et al. A DNA sequence-based study of the *Schistosoma indicum* (Trematoda: Digenea) group: population phylogeny, taxonomy and historical biogeography. Parasitology. 2007; 134(14):2009–20.

81. Koonin EV, Csuros M, Rogozin IB. Whence genes in pieces: reconstruction of the exon–intron gene structures of the last eukaryotic common ancestor and other ancestral eukaryotes. WIRES RNA. 2013; 4(1):93–105. https://doi.org/10.1002/wrna.1143 PMID: 23139082

82. Buathong S, Leelayoova S, Mungthin M, Tan-ariya P. Role of Cathepsin B in *Schistosoma japonicum* infection. J Trop Med Parasitol. 2014; 37:43–53.

83. deWalick S, Bexkens ML, van Balkom BWM, Wu Y-P, Smit CH, Hokke CH, et al. The proteome of the insoluble *Schistosoma mansoni* eggshell skeleton. Int J Parasitol. 2011; 41(5):523–32. https://doi.org/10.1016/j.ijpara.2010.12.005 PMID: 21236260

84. Liu F, Ding H, Tian J, Zhou C, Yang F, Shao W, et al. Differential gene expression, including Sjfs800, in *Schistosoma japonicum* females before, during, and after male-female pairing. bioRxiv. 2018:452458.

85. Steinfelder S, Andersen JF, Cannons JL, Feng CG, Joshi M, Dwyer D, et al. The major component in schistosome eggs responsible for conditioning dendritic cells for Th2 polarization is a T2 ribonuclease (omega-1). J Exp Med. 2009; 206(8):1681. https://doi.org/10.1084/jem.20082462 PMID: 19635859

86. Prasanphanich N, Mickum M, Heimburg-Molinaro J, Cummings R. Glycoconjugates in host-helminth interactions. Front Immunol. 2013; 4(240).

87. Mickum ML, Prasanphanich NS, Heimburg-Molinaro J, Leon KE, Cummings RD. Deciphering the glycogenome of schistosomes. Front Genet. 2014; 5(262).

88. Ranjit N, Zhan B, Hamilton B, Stenzel D, Lowther J, Pearson M, et al. Proteolytic degradation of hemoglobin in the intestine of the human hookworm *Necator americanus*. J Infect Dis. 2009; 199(6):904–12. https://doi.org/10.1086/597048 PMID: 19434933

89. Mora Huertas AC, Schmelzer CEH, Luise C, Sippl W, Pietzsch M, Hoehenwarter W, et al. Degradation of tropoelastin and skin elastin by neprilysin. Biochimie. 2018; 146:73–8. https://doi.org/10.1016/j.biochi.2017.11.018 PMID: 29196110

90. Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 1994; 371(6494):215–20. https://doi.org/10.1038/371215a0 PMID: 8078581