GigaScience

Comparative genomics and transcriptomics of four Paragonimus species provide insights into lung fluke parasitism and pathogenesis

--Manuscript Draft--

Manuscript Number:	GIGA-D-19-00411
Full Title:	Comparative genomics and transcriptomics of four Paragonimus species provide insights into lung fluke parasitism and pathogenesis
Article Type:	Research
Funding Information:	National Institutes of Health - National Institute of Allergy and Infectious Diseases (AI081803) — Dr. Makedonka Mitreva National Institutes of Health - National Human Genome Research Institute (U54HG003079) — Dr. Makedonka Mitreva National Institutes of Health - National Institute of General Medical Sciences (GM097435) — Dr. Makedonka Mitreva Thailand Research Fund (TH) - Distinguished Research Professor Grant (DPG6280002) — Dr. Wanchai Maleewong
Abstract:	Background Paragonimus spp. (lung flukes) are among the most injurious food-borne helminths, infecting ~23 million people, (~293 million with infection risk). Paragonimiasis is acquired from infected undercooked crustaceans and primarily affects the lungs, but often causes lesions elsewhere including the brain. The disease is easily mistaken for tuberculosis due to similar pulmonary symptoms, so diagnostics are in demand. Results We assembled, annotated and compared draft genomes of four prevalent and distinct Paragonimus species: P. miyazakii, P. westermani, P. kellicotti and P. heterotremus. Genomes ranged from 697 to 923 Mb, included 12,072 to 12,853 genes, and were 87% to 96% complete. Orthologous group (OG) analysis spanning 21 species (lung, liver and blood flukes, additional platyhelminths and hosts) provided insights into lung fluke biology, including identifying 256 lung fluke-specific and conserved OGs enriched for iron acquisition, immune modulation and other parasitic functions, and eight “host mimic” OGs including a STOX1 ortholog (proliferation of pulmonary artery cells). Transcriptome analysis identified consistent adult-stage Paragonimus expression profiles, and 388 genes differentially expressed between host body cavities and tissues, enriched for functions including proteolysis, nutrient transport and iron acquisition. Previously identified Paragonimus diagnostic antigens were matched to genes, providing an opportunity to optimize and ensure consistent cross-reactivity for diagnostic assays. Conclusions We anticipate that these novel genomic and transcriptomic resources will be invaluable for future lung fluke research. This report represents a major contribution to ongoing trematode genome sequencing efforts and bootstraps future studies into the biology, evolution and pathogenesis of Paragonimus and related food-borne flukes.
Corresponding Author:	Makedonka Mitreva UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Bruce A Rosa
First Author Secondary Information:	

Order of Authors:	Bruce A Rosa
	Young-Jun Choi
	Samantha N McNulty
	Hyeim Jung
	John Martin
	Takeshi Agatsuma
	Hiromu Sugiyama
	Thanh Le Hoa
	Pham Ngoc Doanh
	Wanchai Maleewong
	David Blair
	Paul J. Brindley
	Peter U. Fischer
	Makedonka Mitreva
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes
Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible.	Yes

Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Comparative genomics and transcriptomics of four *Paragonimus* species provide insights into lung fluke parasitism and pathogenesis

Bruce A. Rosa[1], Young-Jun Choi[2], Samantha N. McNulty[2], Hyeim Jung[2], John Martin[2], Takeshi Agatsuma[3], Hiromu Sugiyama[4], Thanh Le Hoa[5], Pham Ngoc Doanh[6,7], Wanchai Maleewong[8], David Blair[9], Paul J. Brindley[10], Peter U. Fischer[1], Makedonka Mitreva[1,2*]

[1]Department of Internal Medicine, Washington University School of Medicine, St. Louis, MO 63110, USA

[2]The McDonnell Genome Institute at Washington University, School of Medicine, St. Louis, MO 63108, USA

[3]Department of Environmental Health Sciences, Kochi Medical School, Oko, Nankoku City, Kochi 783-8505, Japan

[4]Laboratory of Helminthology, Department of Parasitology, National Institute of Infectious Diseases, Tokyo 162-8640, Japan

[5]Department of Immunology, Institute of Biotechnology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

[6]Institute of Ecology and Biological Resources, Vietnam Academy of Science and Technology, Hanoi, Vietnam

[7]Graduate University of Science and Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam

[8]Research and Diagnostic Center for Emerging Infectious Diseases, Khon Kaen University, Khon Kaen, Thailand, Department of Parasitology, Faculty of Medicine, Khon Kaen University, Khon Kaen, Thailand

[9]College of Marine and Environmental Sciences, James Cook University, Townsville, Queensland 4811, Australia

[10]Departments of Microbiology, Immunology and Tropical Medicine, and Research Center for Neglected Diseases of Poverty, and Pathology School of Medicine & Health Sciences, George Washington University, Washington, DC 20037, USA

***** Correspondence should be addressed to Makedonka Mitreva. Tel. +1-314-285-2005, Fax +1-314-286-1800, Email: mmitreva@wustl.edu

Emails:

Bruce A. Rosa: barosa@wustl.edu

Young-Jun Choi: choi.y@wustl.edu

Samantha N. McNulty: samantha.n.mcnulty@gmail.com

Hyeim Jung: jungh@wustl.edu

John Martin: jmartin@wustl.edu

Takeshi Agatsuma: agatsuma@kochi-u.ac.jp

Hiromu Sugiyama: hsugi@niid.go.jp

Thanh Le Hoa: imibtvn@gmail.com

Pham Ngoc Doanh: pndoanh@yahoo.com

Wanchai Maleewong: wanch_ma@kku.ac.th

David Blair: david.blair@jcu.edu.au

Paul J. Brindley: pbrindley@gwu.edu

Peter U. Fischer: pufischer@wustl.edu

Makedonka Mitreva: mmitreva@wustl.edu

Keywords

Lung flukes, genomics, transcriptomics, paragonimiasis, infectious disease, trematodes

Abstract

Background

Paragonimus spp. (lung flukes) are among the most injurious food-borne helminths, infecting ~23 million people, (~293 million with infection risk). Paragonimiasis is acquired from infected undercooked crustaceans and primarily affects the lungs, but often causes lesions elsewhere including the brain. The disease is easily mistaken for tuberculosis due to similar pulmonary symptoms, so diagnostics are in demand.

Results

We assembled, annotated and compared draft genomes of four prevalent and distinct *Paragonimus* species: *P. miyazakii*, *P. westermani*, *P. kellicotti* and *P. heterotremus*. Genomes ranged from 697 to 923 Mb, included 12,072 to 12,853 genes, and were 87% to 96% complete. Orthologous group (OG) analysis spanning 21 species (lung, liver and blood flukes, additional platyhelminths and hosts) provided insights into lung fluke biology, including identifying 256 lung fluke-specific and conserved OGs enriched for iron acquisition, immune modulation and other parasitic functions, and eight "host mimic" OGs including a STOX1 ortholog (proliferation of pulmonary artery cells). Transcriptome analysis identified consistent adult-stage *Paragonimus* expression profiles, and 388 genes differentially expressed between host body cavities and tissues, enriched for functions including proteolysis, nutrient transport and iron acquisition. Previously identified *Paragonimus* diagnostic antigens were matched to genes, providing an opportunity to optimize and ensure consistent cross-reactivity for diagnostic assays.

Conclusions

We anticipate that these novel genomic and transcriptomic resources will be invaluable for future lung fluke research. This report represents a major contribution to ongoing trematode genome sequencing efforts and bootstraps future studies into the biology, evolution and pathogenesis of *Paragonimus* and related food-borne flukes.

Background

The trematode genus *Paragonimus*, the lung flukes, is among the most injurious taxon of food-borne helminths. About 23 million people are infected with lung flukes [1], an estimated 292 million people are at-risk, mainly in eastern Asia [2] , and billions of people live in areas where *Paragonimus* infections of animals are endemic. The life-cycle of *Paragonimus* species involves freshwater snails, crustacean intermediate hosts and mammals in Asia, parts of West Africa, and the Americas [3]. Human paragonimiasis is acquired by consuming raw or undercooked shrimp and crabs containing the metacercaria, which is the infective stage. Although primarily affecting the lungs, lesions can occur at other sites, including the brain [4], and pulmonary paragonimiasis is frequently mistaken for tuberculosis due to similar respiratory symptoms [4].

Pathogenesis ensues because of the migration of the newly invading juveniles from the gut to the lungs and through not-infrequent ectopic migration to the brain, reproductive organs, and subcutaneous sites at the extremities, and because of toxins and other mediators released by the parasites during the larval migration [4, 5]. The presence of the flukes in the lung causes hemorrhage, inflammation with leukocytic infiltration and necrosis of lung parenchyma that gradually proceeds to the development of fibrotic encapsulation except for a fistula from the evolving lesion to the respiratory tract. Eggs of the lung fluke exit the encapsulated lesion through the fistula to reach the sputum and/or feces of the host, where they pass to the external environment, accomplishing transmission of the parasite [6]. There are signs and symptoms that allow characterization of acute and chronic stages of paragonimiasis. In pulmonary paragonimiasis, for example, the most noticeable clinical symptom of an infected individual is a chronic cough with gelatinous, rusty brown, pneumonia-like, blood-streaked sputum [6]. Heavy work commonly induces hemoptysis. Pneumothorax, empyema from secondary bacterial infection and pleural effusion might also be presented. When symptoms include only a chronic cough, the disease may be misinterpreted as chronic bronchitis and bronchiectasis or bronchial asthma. Pulmonary paragonimiasis is frequently confused with pulmonary tuberculosis [7]. The symptoms of extra-pulmonary paragonimiasis vary depending on the location of the fluke, including cerebral [5] and abdominal paragonimiasis [6].

Paragonimus is a large genus that includes more than 50 nominal species [8]. Seven of these species or species complexes of *Paragonimus* are known to infect humans [3]. This is also an ancient genus, thought to have originated before the breakup of Gondwana [9], but possibly also distributing as colonists from the original East

Asian clade, based on the distribution of host species [10]. To improve our understanding of pathogens across this genus at the molecular level, we have assembled, annotated and compared draft genomes of four of these, three from Asia (*P. westermani* from Japan, *P. heterotremus, P. miyazakii*) and one from North America (*P. kellicotti*). Among them, P. westermani is the best-known species causing pulmonary paragonimiasis. This name has been applied to a genetically and geographically diverse complex of lung fluke populations differing widely in biological features including infectivity to humans [11]. The complex extends from India and Sri Lanka eastwards to Siberia, Korea and Japan, and southwards into Vietnam, Indonesia and the Philippines. However, human infections are reported primarily from China, Korea, Japan and the Philippines. Until this study, an Indian member of the *P. westermani* complex was the only lung fluke species for which a genome sequence was available [12]. *Paragonimus heterotremus* is the most common cause of pulmonary paragonimiasis in southern China, Lao PDR, Vietnam, northeastern India and Thailand [6, 8]. *Paragonimus miyazakii* is a member of the *P. skrjabini* complex, to which Blair and co-workers accorded sub-specific status [13]. Flukes of this complex tend not to mature in humans but frequently cause ectopic disease at diverse sites, including the brain. In North America, infection with *P. kellicotti* is primarily a disease of native, crayfish-eating mammals including the otter and mink. The occasional human infections can be severe, and thoracic involvement is typical [14, 15].

These four species represent a broad sampling of the phylogenetic diversity of the genus. Most of the known diversity, as revealed by DNA sequences from portions of the mitochondrial genome and the nuclear ribosomal genes, resides in Asia [16]. Analysis of the ITS2 marker by Blair et al [16] indicates that each of the species sequenced occupies a distinct clade within the phylogenetic tree.

In addition to a greater understanding of the genome contents of this group of food-borne trematodes, the findings presented here provide new information to assist development of diagnostic tools and recognition of potential drug targets. The findings will facilitate evolutionary, zoogeographical and phylogenetic investigation of the genus *Paragonimus* and its host-parasite relationships through the comparative analysis of gene content relative to other sequenced platyhelminth and host species, as well as through a comparative transcriptomic analysis.

Data Description

Genome assemblies and annotations

DNA and RNA samples were collected from adult-stage parasites of four distinct *Paragonimus* species: *P. miyazakii* (Japan), *P. heterotremus* (LC strain, Vietnam), *P. kellicotti* (Missouri, USA) and *Paragonimus westermani* (Japan). Illumina DNA sequencing produced fragments, 3kb- and 8kb-insert whole-genome shotgun libraries, as well as PacBio reads being generated for *P. kellicotti*, which were used to produce genomic assemblies using assemblers [17, 18] and scaffolders [19]. Transcriptome datasets utilizing RNA-Seq from the adult stages of each species were mapped and assembled using Stringtie [20], and included: *P. miyazakii* samples collected from stages in the liver, peritoneal cavity (2 replicates), lung (adult) and pleural cavity; *P. heterotremus* samples from adults and young adults (2 replicates), and previously published adult-stage transcriptomic samples for *P. westermani* [21] and *P. kellicotti* [22]. These transcript assemblies, in addition to available trematode protein sequence information [23, 24] were used as evidence to perform genome annotation using BRAKER [25] and MAKER [26], after masking for repetitive elements [27], transposable elements [26] ribosomal RNA and tRNA [28, 29]. Gene models were refined using transcript and protein evidence, including selection based on Annotation Edit Distance (AED) [30], InterProScan functional annotations [31], and by balancing sensitivity and specificity [26, 32]. Gene products were named using PANNZER2 [33] and sma3s v2 [34]. Genome assemblies, annotations and raw reads are available for download from the NCBI Sequence Read Archive (SRA [35]), with accession numbers provided in **Supplementary Table S1**.

Transcriptome datasets and gene functional annotations

RNA-seq datasets were trimmed for adapters [36] and aligned [37] to their respective genome assemblies, and gene expression levels (FPKM) were quantified per gene per sample in each of the four species [38]. For *P. miyazakii*, differential gene expression analysis [39] identified genes significantly differentially expressed between the cavity and tissue samples. Interpro domains and Gene Ontology (GO) terms [31], KEGG enzymes [40], and protease [41] annotations of the genes were used to identify putative functions of genes of interest and perform pathway enrichment [42]. All raw RNA-Seq fastq files were uploaded to the NCBI Sequence Read Archive (SRA [35]), and complete sample metadata and accession information are provided in **Supplementary Table S1**. **Supplementary Table S2** provides, for each of the species, complete gene lists and

gene expression levels for each of the RNA-Seq samples. Complete functional annotations for every gene and the differential gene expression dataset are also provided for *P. miyazakii* in this table.

Gene family conservation dataset

Orthologous groups (OG) of genes from 21 species (including the four new *Paragonimus* gene sets) were inferred with OrthoFinder v1.1.4 [43] (Worm gene sets retrieved from WormBase ParaSite [24]; Outgroup species gene sets retrieved from Ensembl [44]). For each orthologous group identified in the analysis, **Supplementary Table S3** provides complete gene lists, counts of genes per species, and average gene expression levels from each the *Paragonimus* transcriptome datasets described above. As an additional gene-level functional annotation, OG membership information is provided for each of the *P. miyazakii* genes **in Supplementary Table S2** (for example, identifying if a gene is *Paragonimus*-specific, or conserved across all species).

The methods section provides detailed descriptions of the data generation described here, as well as the downstream analyses providing insights into *Paragonimus* biology. All raw and processed data are available in the supplementary tables, and represent an invaluable resource for future *Paragonimus* research, providing previously unavailable genome assemblies, gene identifications, functional annotations and expression levels, as well as conservation information spanning trematode and host species.

Analyses

Genome features

The sizes of the four novel *Paragonimus* genomes range from 697 to 923 Mb, containing between 12,072 and 12,853 genes. These draft genomes are estimated to be between 87% and 96% complete, with the new *P. westermani* genome produced from a sample collected from Japan being slightly more complete than the previously-sequenced genome produced from a sample collected from India [12] (96.4% vs 94.1%, respectively; **Table 1**). Here, statements about *P. westermani* apply to the new Japanese genome, unless otherwise stated. The total genome lengths of the *Paragonimus* spp. are larger than those of the Schistosomatidae and Opisthorchiidae, but smaller than those of Fasciolidae. However, the total numbers of protein-coding genes are comparable (**Figure 1A**). Repetitive sequences occupy between 49% and 54% of the *Paragonimus* genomes. The repeat landscapes, depicting the relative abundance of repeat classes in the genome, versus the Kimura

194 divergence from the consensus, revealed that *P. kellicotti* in particular has a significant number of copies of
195 transposable elements (TE) with high similarity to consensus (Kimura substitution level: 0-5), indicating recent
196 and current TE activity (**Figure 1B**). In a recent study [45], TE activity in the Fasciolidae was found to be low.
197 TEs are potent sources of mutation that can rapidly create genetic variance, especially following genetic
198 bottlenecks and environmental changes, providing bursts of allelic and phenotypic diversity upon which selection
199 can act [46, 47]. Therefore, changes in TE activity, modulated by environmentally induced physiological or
200 genomic stress, may have a major effect on adaptation of populations and species facing novel habitats and
201 large environmental perturbations [48].

202 Focusing on the gene content, *P. kellicotti* had the shortest average total gene length among the species,
203 and the lung flukes overall had similar gene lengths compared to other flukes, while platyhelminth species other
204 than trematodes had shorter genes overall (**Figure 2A**). The variability in gene lengths observed between
205 species is explained by differences in both average intron lengths (**Figure 2B**) and the average number of exons
206 per gene (**Figure 2C**) while the average coding sequence (CDS) lengths of the exons across all the platyhelminth
207 species were similar to each other (**Figure 2D**). Whereas there was species-to-species variability in gene lengths
208 and exon counts, consistent patterns among the types of flukes were not apparent.

209 Mitochondrial whole genome-based clustering was performed for the four *Paragonimus* species plus
210 some additional existing mitochondrial genome assemblies for *P. ohirai* and four for *P. westermani*, including
211 previously-sequenced mitochondrial genomes of *Paragonimus* (**Figure 3A**). This clustering indicated that our *P.*
212 *westermani* sample clustered with the existing known *P. westermani* samples from eastern Asia, and that all the
213 other three newly sequenced species were distinct from *P. ohirai*.

214 We generated a PacBio long-read based mitochondrial assembly for *P. kellicotti*. The fully circularized
215 complete genome was 17.3 kb in length, including a 3.7 kb non-coding repeat region between $tRNA^{Gly}$ and *cox3*
216 (**Supplementary Figure S1**). There are seven copies of long repeats (378 bp) and 9.5 copies of short repeats
217 (111 bp). The long repeats overlap with 6 copies of $tRNA^{Glu}$. This structural organization of repeat sequences
218 does not resemble those found in *Paragonimus ohirai* [12] and *P. westermani* [12] where the non-coding region
219 is partitioned by $tRNA^{Glu}$ into two parts.

220 Clustering of the four new lung fluke genomes, four liver fluke genomes, three blood fluke genomes, five
221 other platyhelminth species, four host species and a yeast outgroup was performed based on the shared

phylogeny among orthologous protein groups (using OrthoFinder [43]). These findings mirrored the mitochondrial clustering results for the lung fluke species (**Figure 3B**), indicating that *P. westermani* is the earlier-diverging taxon, as previously suggested based on ribosomal RNA [49]. p

Although our *P. westermani* reference genome was assembled using samples collected from Japan (Amakusa, Kyusyu), we also generated additional sequence data using samples from China (Hunan province). We compared the genomic sequences of our East Asian *P. westermani* to the recently published *P. westermani* genome from India (Changlang, Arunachal Pradesh) [12] to estimate the genetic divergence between geographically diverse samples. The average sequence identity between the genomic samples from Japan and China was 98.3%. The isolate from India was genetically more distant from the East Asian *P. westermani* with the average sequence identity of 87.6% and 87.9% for the Japanese vs. Indian and the Chinese vs. Indian comparisons, respectively. Overall, the average degree of identity between strains was 91.3%.

Gene-family dynamics identify expanded functions distinguishing lung fluke species

We investigated large-scale differences in gene complements among families of digenetic trematodes (**Figure 4A**) and modeled gene gain and loss while accounting for the phylogenetic history of species [50]. Gene families of interest that displayed pronounced differential expansion or contraction (**Figure 4B**) included the papain-family cysteine proteases, cathepsins L, B and F, dynein heavy chain, spectrin/dystrophin, heat shock 70 kDa protein, major vault protein, and multidrug resistance protein. Total protease counts are shown in **Figure 4C**.

Lineage-specific expansion was observed in cathepsin F genes in *Paragonimus* spp. *Paragonimus miyazakii* RNA-seq reads showed that nine cathepsin F genes (out of 24 total) were differentially expressed, with expression levels in (peritoneal and pleural) cavity stage parasites significantly higher than in the tissue (lung and liver) developmental stages. This suggested that (1) these enzymes are highly expressed during parasite penetration of the intestinal wall and invasion and migration through the abdominal and thoracic cavities (1- to 7-week-old immature stages), and (2) they might participate in metacercarial excystment, tissue invasion/migration and immune evasion. The remaining 15 constitutively expressed cathepsin F genes might be involved in nutrient digestion and remodeling of other physiologically active molecules. Ahn et al. [51] also reported differential expression of cathepsin F genes during development of *P. westermani*, and showed that

most are highly immunogenic. This flagged them as prospective diagnostic targets. The importance of cathepsin F for *Paragonimus* contrasts with its function in the fasciolids, where cathepsin L genes are expanded and are thought to play a more critical role in host invasion [45, 52].

Differential expansion of cytoskeletal molecules is of interest in the context of tegument physiology [53]. Dynein is a microtubule motor protein, which transports intracellular cargo. Spectrin is an actin-binding protein, with a key role in maintenance of integrity of the plasma membrane. Dystrophin links microfilaments with extracellular matrix. The syncytial tegument is a complex structure on the surface of flatworms, is a major adaptation to parasitism, and plays critical roles in nutrient uptake, immune response modulation and evasion, and other processes [53].

In *Paragonimus* spp., expanded gene families included heat shock proteins, major vault proteins, and multidrug resistance proteins that play roles in maintaining cellular homeostasis under stress conditions. Flatworm heat shock proteins (HSPs) play a key role as molecular chaperones in the maintenance of protein homeostasis. They also are immunogenic and immunomodulatory. HSP is the most abundant family of proteins in the immature and mature egg of *Schistosoma mansoni*, and in the miracidium [54] and is highly abundant in the tegument of the adult schistosome [55]. In addition, HSP is abundant in the excretory/secretory products of the adult *Schistosoma japonicum* blood fluke [56]. HSP stimulates diverse immune cells, eliciting release of pro- and anti-inflammatory cytokines [57], binds human LDL (the purpose of which is unknown but may be associated with transport of apoprotein B or in lipid trafficking [58]), and given these properties represents a promising vaccine and diagnostic candidate [59]. Vaults, ribonucleoprotein complexes, are highly conserved in eukaryotes. Although their exact function remains unclear, it may be associated with multidrug resistance phenotypes and with signal transduction. In *S. mansoni*, up-regulation of major vault protein has been observed during the transition from cercaria to schistosomulum and in praziquantel-resistant adult worms [60]. ATP-binding cassette transporters (ABC transporters) are essential components of cellular physiological machinery, and some ABC transporters, including P-glycoproteins, pump toxins and xenobiotics out of the cell. Overexpression of P-glycoprotein has been reported in a praziquantel-resistant *S. mansoni* [61].

Tetraspanin sequence evolution in *P. kellicotti*

We searched for genes that evolved under positive selection in the four *Paragonimus* spp. based on the non-synonymous to synonymous substitution rate ratio (d_N/d_S). We conducted the branch-site test of positive selection to identify adaptive gene variants that became fixed in each species [62] (**Supplementary Table S4**). A tetraspanin from *P. kellicotti* (PKEL_00573) reached statistical significance after correction for multiple testing (d_N/d_S = 9.9, FDR = 0.018). Tetraspanins are small integral proteins bearing four transmembrane domains which form two extracellular loops [63]. In trematodes, they are major components of the tegument at the host-parasite interface [64], are highly immunogenic vaccine antigens [65, 66], and may play a role in immune evasion [67]. In the tetraspanin sequence of *P. kellicotti*, we detected six amino acid sites under positive selection (**Supplementary Figure S2**). Five of the six sites were predicted to be located within the extracellular loops believed to interact with the immune system of the host. A similar pattern of positive selection within regions that code for extracellular loops has been reported in tetraspanin-23 from African *Schistosoma* species [68].

Gene phylogeny analysis identifies functions conserved and specific to fluke groups

OrthoFinder [43] was used to determine the conservation of orthologous groups (OGs) containing one or more orthologs from each of the 21 species indicated in **Figure 3B**. Complete gene counts and lists per species and per OG are provided in **Supplementary Table S3**. These results were parsed to identify the OGs containing members among the platyhelminth species, and those that were conserved across all members of each group (lung, liver, and blood flukes, and other platyhelminth species (**Figure 5A**). This analysis identified 256 OGs that were conserved among, and exclusive to, the lung flukes (**Figures 5A and 5B**). The lung fluke-conserved and -specific genes were significantly enriched for several gene ontology (GO) terms (**Table 2**; using *P. miyazakii* genes to test significance), most of which were related to peptidase activity (including serine proteases which are involved in host tissue invasion, anticoagulation, and immune evasion [69]), as well as “iron binding” (which may be related to novel iron acquisition mechanisms from host tissue, which is not well understood in most metazoan parasites, but has been described in schistosomes [70]).

Expansion of unique aspartic proteases (including those predicted to be retropepsins) and other peptidases in the lung flukes may be associated with digestion of ingested blood, given the key role of this category of hydrolases and their inhibitors in nutrition and digestion of hemoglobin by schistosomes, and indeed

other blood-feeding worms including hookworms [71, 72]. Given that pulmonary hemorrhage and hemoptysis are cardinal signs of lung fluke infection, it can be anticipated that the lung flukes ingest host blood when localized at the ulcerous lesion induced in the pulmonary parenchyma by infection. Overall, protease counts across species were similar (**Figure 4C**) although *P. kellicotti* had substantially fewer protease inhibitors compared to the other *Paragonimus* species (34 vs 57, 62 and 66), *F. hepatica* (61) and *S. mansoni* (55). Protease inhibitors in flukes are thought to be important for creating a safe environment for the parasite inside the host by inhibiting and regulating protease activity and immunomodulation [91], so this may suggest a novel host interaction strategy by *P. kellicotti*.

Analysis of the adult-stage gene expression levels of the various protease classes (**Supplementary Figure S3**) did not identify substantial differences among the *Paragonimus* species, except for a lower expression of threonine proteases in *P. kellicotti.* During the adult stage, cysteine proteases in all *Paragonimus* species exhibited significantly higher expression overall compared to *F. hepatica*, but similar expression levels to *S. mansoni.* A previous study identified immunodominant excretory-secretory cysteine proteases of adult *Paragonimus westermani* involved in immune evasion [73] and another study identified critical roles for excretory-secretory cysteine proteases during tissue invasion by newly excysted metacercariae of *P. westermani* [74]. The rapid diversification and critical host-interaction functions of the proteases highlights their importance, both in terms of understanding *Paragonimus* biology and in terms of identifying targets for control.

Focusing on the more specific functional data provided by InterPro domain analysis, functional enrichment among the lung, liver and blood fluke conserved-and-exclusive OGs (**Figure 5C**) indicated that each class of fluke has evolved a distinct set of aspartic peptidases, trematode eggshell synthesis genes and saposin-like genes (which interact with lipids and are strongly immunogenic during fascioliasis [75]). The lung flukes, meanwhile, have uniquely expanded sets of serine proteases, as well as other gene functions including FAR1 DNA binding (a class of proteins which are important secreted host-interacting proteins in some parasitic nematodes [76]), fatty-acid binding, and ferritin-like functions (intracellular proteins involved in iron metabolism, localized in vitelline follicles and eggs [77]).

We also identified OGs that are exclusive only to one or more of the lung flukes and one or more of the mollusk and mammal host species (**Figure 5D**; full species list in **Figure 3B**). Since these OGs are not found in any of the other platyhelminth species but are conserved with the host of the lung fluke, these are informative

candidates for potential host-interacting proteins that may have evolved uniquely in lung flukes to mimic host factors specific to their environmental niche. The four "host mimic" proteins conserved among all *Paragonimus* species and both mammalian and molluskan species (**Table 3**) include orthologs of human STOX1 (which induces the proliferation of artery smooth-muscle cells [78]), and a glutaredoxin-like protein (C5orf63; a class of proteins that impair bacterial clearance from the lungs in pneumonia [79]), both of which have functional relevance to the survival of adult-stage *Paragonimus* in the lungs. Similarly, host-mimic proteins only conserved among mammalian species include Zip67 (a transcriptional repressor of genes including PPARG, which is involved in inflammation reduction and is highly expressed in alveolar macrophages [80]) and a PAC domain-containing protein (may function in oxygen detection [81]), both of which have potential function in survival in the mammalian lungs. Host-mimic proteins conserved only with the mollusk host species included a Zinc-finger transcription factor of unknown function, and a protein containing Ankyrin repeats, which have previously been shown to be involved in host-parasite interactions by intracellular bacteria [82]. Overall, this analysis of broad and specific orthologous group conservation among platyhelminth species and their hosts identifies gene sets that have uniquely evolved in lung flukes, primarily for survival in their distinct niche in host tissues including functions such as iron acquisition, host environment oxygen level detection and immune cell modulation.

Gene expression analysis identifies stage-specific lung fluke functions

Lung (adult) stage RNA-Seq datasets were collected for each of the four lung fluke species (accessions in **Supplementary Table S1**), and reads were mapped to each of their respective genomes. Based on the 1:1 gene orthologs (as defined by the previously described OG dataset), the orthologous genes across the lung flukes had consistent adult-stage gene expression levels, with Pearson correlations ranging from 0.72 to 0.85 (**Figure 6A**, **6B**). Worms from additional life cycle stages were collected for *P. miyazakii*, with samples sequenced from cavities (peritoneal and pleural cavities) and tissues (lung and liver). DESeq2 [39] differential expression analysis comparing tissue and cavity stages identified 216 genes significantly overexpressed in the cavities relative to the tissues, and 172 genes significantly overexpressed in the tissues relative to the cavities (**Figure 6C**). Functional enrichment among these gene sets (**Table 4**) indicates that within the tissues, *P. miyazakii* overexpresses genes related to cytoskeleton and microtubules, lyases and phosphatases, carbon-oxygen lyase and ribonucleotide binding. Within the cavities, *P. miyazakii* overexpresses genes related to

cysteine peptidase activity (critical for larval migration through host tissues [83]), iron ion binding (related to oxygen scavenging), and sulfotransferase (responsible for anthelmintic resistance in *S. mansoni* [84]).

The *P. miyazakii* genes belonging to the lung fluke-conserved and -exclusive OGs (described above) on average had significantly higher expression levels in the liver stage compared to the pleural cavity and lung stages, and significantly lower expression in the lung stage compared to all of the other stages. These results suggest that most of these OGs contain genes that are actively expressed during the transit through the host mammal, en route to the lungs (**Figure 6D**). However, some were more highly expressed in the lung stage relative to the other stages (**Table 5**), and these genes had annotated functions including serine and aspartic peptidases and an MFS transporter gene (transports nutrients and ions between cells and the environment [85]). Gene expression levels and orthologous group identifiers for each gene in each of the four species are provided in **Supplementary Table S2**, along with detailed functional annotations for each of the *P. miyazakii* genes.

This stage-specific gene expression offers insight into known and novel biological functions of lung flukes at different developmental stages and within different organs and tissues of the mammalian host, and represents a sophisticated new resource for study of specific genes of the lung fluke.

Treatments, vaccine targets and diagnostics

The World Health Organization (WHO) currently recommends the use of praziquantel or triclabendazole (considered a backup) for the treatment of paragonimiasis, both of which are highly effective for curing infections [86]. However, there are concerns about the development of resistance to these drugs; triclabendazole resistance of *P. westermani* was reported in a human case from Korea [87]. Furthermore, there is widespread resistance to triclabendazole in liver flukes in cattle in Australia and South America [88], and praziquantel resistance is anticipated in the future due to its widespread use as a single treatment for schistosomiasis, a problematic situation which has encouraged the search for novel drugs [89]. The comparative analysis presented here identifies valuable protein targets for drug treatments, including *Paragonimus*-specific proteins and trematode-conserved proteins which do not share orthology to human proteins. These can provide a starting point for future bioinformatic prioritization and drug testing (**Supplementary Tables S2** and **S3**).

Vaccination to prevent future infections would offer an attractive alternative to treatment, but vaccine protection against trematode infection has so far been unsuccessful and is unlikely to be practical for

paragonimiasis in the near future [90]. However, the complete genome sequences and comparative analysis of the gene sets presented here provides a valuable resource for potential future vaccine target development.

Pulmonary paragonimiasis is frequently mistaken for tuberculosis or pneumonia, and often patients do not shed eggs, which leads to false positive diagnoses of other conditions such as malaria or pneumonia [4, 91, 92]. This highlights a pressing need for accurate, rapid and affordable diagnostic approaches for paragonimiasis, a topic which has been the focus of many previous studies. We performed BLAST sequence similarity searches of previously identified *Paragonimus* diagnostic antigen targets among the four species (**Supplementary Figure S4**). These included: (i) *P. westermani* and *P. pseudoheterotremus* cysteine proteases identified in two previous studies [93, 94] (matching to the same protein targets from both studies in *P. heterotremus* and *P. kellicotti*), one of which had high adult-stage expression levels in all four species [93]; (ii) three different tyrosine kinases (one of which was identified in two different studies, in *Clonorchis sinensis* and in *P. westermani* [95, 96]), all of which had relatively low gene expression levels in adult stages; (iii) a previously unannotated *P. heterotremus* ELISA antigen [97] with low expression across life cycle stages, which we now annotate as a saposin protein (which we found to rapidly evolve among flukes [**Figure 5C**], and which is strongly immunogenic in fascioliasis [75]); (iv) eggshell proteins of *P. westermani* [98], for which we now provide full-length sequences. We observed that this gene was conserved across and specific to the lung flukes, with lower gene expression in the young adult stage (*P. heterotremus*), but higher expression in the adult stages of all species; (v) among serodiagnostic *P. kellicotti* antigens based on a transcriptome assembly and proteomic evidence [21], we identified the top 10 of the 25 prioritized transcripts that best matched between the transcript sequence and the newly annotated draft genome of *P. kellicotti.* Thereafter, the full-length gene sequence in *P. kellicotti* was employed to query the other species. Several of these were highly expressed in the adult stage of all four species, including one that is fluke specific (PKEL_05597). However, not all of these had high sequence conservation across all species, with two only having weak hits in *P. heterotremus* (PKEL_00171 and PKEL_01872).

As a result of this newly developed genomic resource for the lung flukes, previously identified diagnostic targets were identified with full gene sequences across all four species. The complete gene sequences, conservation information and transcriptomic gene expression data for these target proteins can allow for optimization of the targets for diagnostic testing that is effective on species spanning the genera

(**Supplementary Figure S4**). This is noteworthy given the absence of a standardized, commercially-available test for serodiagnosis for human paragonimiasis.

Discussion

To substantially improve our understanding of the lung flukes at the molecular level, we assembled, annotated and compared draft genomes of four species of *Paragonimus*, three from Asia (*P. miyazakii*, *P. westermani* from Japan, *P. heterotremus*) and one from North America (*P. kellicotti*), thereby providing novel and hopefully valuable genomic resources across these important parasites for the first time. We have utilized these new resources to compare and analyze phylogenies, to identify gene sets and biological functions associated with parasitism in lung flukes, and to provide a valuable resource for future investigation into host-parasite interactions for these poorly-understood agents of neglected tropical disease. Our identification of previously prioritized *Paragonimus* diagnostic markers in each of the four lung fluke species revealed that the same protein targets were identified in multiple studies, and hence the availability of full gene sequences now should enable the optimization of diagnostic assays aiming for reactivity across all species of lung fluke. Overall, we hope that the novel genomic and transcriptomic resources developed here will be invaluable for future research on paragonimiasis, and to underpin sequence information to guide experimental design and generate novel hypotheses.

Methods

Parasite specimens

Samples of DNA and RNA of *Paragonimus westermani* were sourced in Japan. *Paragonimus heterotremus* (LC strain, Vietnam) were recovered from a cat experimentally infected with metacercariae from Lai Chau province, northern Vietnam (70% ethanol preserved; whole worm). *P. miyazakii* Metacercariae were recovered from freshwater crabs (*Geothelphusa dehaani*), collected in Shizuoka Prefecture, central Japan [15], and were raised to adulthood in rats. DNA and RNA samples were prepared for each of the (pre-)adult flukes recovered from the lungs and from the pleural and peritoneal cavities of experimentally infected rats. *Paragonimus kellicotti* adult worms for genome sequencing were recovered from the lungs of Mongolian gerbils infected in the laboratory with metacercariae recovered from Missouri crayfish [99].

Genome sequencing, assembly and annotation

The assemblies consist of fragments, 3kb- and 8kb-insert whole-genome shotgun libraries. The sequences were generated on the Illumina platform and assembled using Allpaths_LG [17]. Scaffolding was improved using an in-house tool called Pygap (gap closure tool), the Pyramid assembler with Illumina paired reads to close gaps and extend contigs, and L_RNA_scaffolder [19] which uses transcript alignments to improve contiguity. For *P. kellicotti*, PacBio reads were assembled using PBJelly [18], utilizing the Illumina allpaths assembly as the reference. Nanocorr was used to perform error correction on the PacBio data. The nuclear genomes were annotated using the MAKER pipeline v2.31.8 [26]. Repetitive elements were softmasked with RepeatMasker v4.0.6 using a species-specific repeat library created by RepeatModeler v1.0.8, RepBase repeat libraries [27], and a list of known transposable elements provided by MAKER [26]. RNA-seq reads were aligned to their respective genome assemblies and assembled using StringTie v1.2.4 [20] (*P. westermani* [21] and *P. kellicotti* [22] transcriptomic reads were retrieved from published reports). The resulting alignments and transcript assembly were used by BRAKER [25] and MAKER pipelines, respectively, as extrinsic evidence. In addition, for each species, mRNA and EST sequences were retrieved from NCBI and passed to MAKER as transcript evidence. Protein sequences from UniRef100 [23] (Trematoda-specific, n=205,161) and WormBase ParaSite WBPS7 [24] were provided to MAKER as protein homology evidence. *Ab initio* gene predictions from BRAKER v2 [25] and AUGUSTUS v3.2.2 (trained by BRAKER and run within MAKER) were refined using the transcript and protein evidence. Previously unpredicted exons and UTRs were added, and split models were merged. The best-supported gene models were chosen based on Annotation Edit Distance (AED) [30]. To reduce false positives, gene predictions without supporting evidence were excluded in the final annotation build, with the exception of those encoding Pfam domains, as detected by InterProScan v5.19 [31]. These Pfam encoding domains were rescued in order to improve the annotation accuracy overall by balancing sensitivity and specificity [26, 32]. Gene products were named using PANNZER2 [33] and sma3s v2 [34]. **Supplementary Table S1** provides details of database accessions for the genomes. The completeness of annotated gene sets was assessed using BUSCO v3.0 [100]. Gene Ontology (GO), KEGG and protease annotations were performed using InterProScan v5.19 [31], GhostKOALA [40], and MEROPS [41], respectively. Functional enrichment testing was performed using GOSTATS [42] for GO enrichment and negative binomial distribution tests for InterPro

domain enrichment (minimum 3 annotated genes required for significant enrichment). Ribosomal RNA and tRNA were annotated using RNAmmer v1.2.1 [28] and tRNAscan-SE v1.23 [29], respectively. Genome charateristics and statistics including CDS, numbers and lengths of genes, exons, introns were defined using the longest complete mRNA (with start and stop codon) for each gene. Across the four species of *Paragonimus*, complete mRNAs were found for an average of 86.2% of all annotated genes.

Assembly of the mitochondrial genome of *P. kellicotti* was achieved using CANU [101] to align PacBio long-reads, followed by error-correction using Pilon [102].

Repeat analysis

RepeatModeler v1.0.8 (with WU-BLAST as its search engine) was used to build, refine and classify consensus models of putative interspersed repeats for each species. With the resulting repeat libraries, genomic sequences were screened using RepeatMasker v4.0.6 in “slow search” mode to generate a detailed annotation of the interspersed and simple repeats. Per-copy distances to consensus were calculated (Kimura 2-parameter model, excluding CpG sites) and were plotted as repeat landscapes where divergence distribution reflected the activity of transposable elements (TE) on a relative time scale per genome using the calcDivergenceFromAlign.pl and createRepeatLandscape.pl scripts included in the RepeatMasker package.

Gene family evolution

Orthologous groups (OG) of genes of 21 species were inferred with OrthoFinder v1.1.4 [43] using the longest isoform for each gene (*Paragonimus* genome source information in **Supplementary Table S1**; Worm gene sets were retrieved from WormBase ParaSite in June 2017 [24]; Outgroup species gene sets were retrieved from Ensembl in June 2017 [44]). CAFE method [50] was employed to model gene gain and loss while accounting for the species’ phylogenetic history based on an ultrametric species tree and the number of gene copies found in each species for each gene family. Birth-death (λ) parameters were estimated and the statistical significance of the observed family size differences among taxa were assessed. Results from OrthoFinder [43] were parsed to identify the OGs of interest based on conservation, including the lung fluke-conserved, liver fluke-conserved and blood fluke-conserved OGs and gene sets per species. **Supplementary Table S3** provides details of full OG counts per species and gene membership.

We used PosiGene [103] to search genome-wide for genes that evolved under positive selection based on the non-synonymous to synonymous substitution ratio. TMMOD [104] and Protter [105] were used for transmembrane helical topology prediction and visualization, respectively. We searched for genes that evolved under positive selection in the four *Paragonimus* spp. based on the non-synonymous to synonymous substitution rate ratio (d_N/d_S). We conducted the branch-site test of positive selection to identify adaptive gene variants that became fixed in each species [62].

Previously identified *Paragonimus* diagnostic antigen search

Nucleotide sequences (or, if unavailable, amino acid sequences) were retrieved from each of the cited publications (**Supplementary Figure S4**). Diamond blastx (nucleotides; v0.9.9.110) or Diamond blastp (amino acids; v0.9.9.110) were used to identify the top hit gene in each *Paragonimus* genome annotation (default settings). The best BLAST E-value was used to identify the top match, followed by top bitscore, length and % ID in the case of ties. For the top 25 *P. kellicotti* immunodominant antigen transcripts identified in McNulty et al, 2014 [22], matches were identified between the assembled transcript and the annotated gene. For the other 3 species, the BLAST searches are performed against the identified *P. kellicotti* gene, and not the original transcript sequence.

RNAseq-based gene expression profiling

After adapter trimming using Trimmomatic v0.36 [36], RNA-seq reads were aligned to their respective genome assemblies using the STAR aligner [37] (2-pass mode, basic). All raw RNA-Seq fastq files were uploaded to the NCBI Sequence Read Archive (SRA [35]), and complete sample metadata and accession information are provided in **Supplementary Table S1**. Read fragments (read pairs or single reads) were quantified per gene per sample using featureCounts (version 1.5.1) [38]. FPKM (fragments per kilobase of gene length per million reads mapped) normalization was also performed. For *P. miyazakii*, significantly differentially expressed genes between the cavity and tissue sample sets were identified using DESeq2 (version 1.4.5) [39] with default settings, and a minimum *P*-value significance threshold of 0.05 (after False Discovery Rate [FDR [106]] correction for the number of tests). Pearson correlation-based RNA-Seq sample clustering was performed in R (using the hclust package, complete linkage).

Statistics

To compare genome statistics and protease expression between species (**Figure 2**, **Supplementary Figure S3**), ANOVA analysis followed by Tukey's HSD post-hoc testing was performed. Because there was a very large number of values being compared for the genome statistics, which can lead to false positive significance in a *t*-test, a random selection of 100 values from each species was used (excluding the upper and lower 1% of data to exclude outliers). Letter labels above the species indicate statistical groups, i.e., if two species share the same letter then they are not statistically significant from each other.

Availability of supporting data and materials

Genomic raw reads, genome assemblies, genome annotations, and raw transcriptomic (RNA-Seq) fastq files were uploaded and are available for download from the NCBI Sequence Read Archive (SRA [35]), with all accession numbers and relevant metadata provided in **Supplementary Table S1**. **Supplementary Table S2** provides, for each of the species, complete gene lists and gene expression levels for each of the RNA-Seq samples. Complete functional annotations for every gene and the differential gene expression dataset are also provided for *P. miyazakii* in this table. For each orthologous group identified, **Supplementary Table S3** provides complete gene lists, counts of genes per species, and average gene expression levels from each the *Paragonimus* transcriptome datasets described above. All results of the genome-wide selection scan are provided in **Supplementary Table S4**.

Declarations

List of Abbreviations

FPKM - Fragments Per Kilobase of gene length per Million reads mapped (gene expression level)

OG - Orthologous Group

TE – Transposable Elements

Consent for Publication

Not Applicable.

Competing Interests

The authors declare that they have no competing interests.

Funding

Sequencing of the genomes was supported by the 'Sequencing the etiological agents of the Food-Borne Trematodiases' project (National Institutes of Health - National Human Genome Research Institute award number U54HG003079). Comparative genome analysis was funded by grants National Institutes of Health - National Institute of Allergy and Infectious Diseases AI081803 and National Institutes of Health - National Institute of General Medical Sciences GM097435 to M.M. Parasite material from Thailand was supported by Distinguished Research Professor Grant (WM), Thailand Research Fund (Grant no. DPG6280002).

Author's Contributions

1. **Conceptualization:** MM PJB.
2. **Formal analysis:** BAR YJC SNM HJ JM.
3. **Funding acquisition:** PJB MM.
4. **Methodology:** PJB PUF DB MM.
5. **Resources:** MM TA HS TLH PND WM DB PUF.
6. **Visualization:** BAR YJC.
7. **Writing – original draft:** BAR YJC MM.
8. **Writing – review & editing:** DB PJB PUF MM.

Acknowledgements

We gratefully acknowledge assistance provided by Xu Zhang and Kymberlie Pepin with genome assembly and annotation and by Rahul Tyagi for figure graphics. We thank Kurt Curtis for his help generating *P. kellicotti* parasite material.

References

1. Furst T, Keiser J and Utzinger J. Global burden of human food-borne trematodiasis: a systematic review and meta-analysis. Lancet Infect Dis. 2012;12 3:210-21. doi:10.1016/S1473-3099(11)70294-8.
2. Utzinger J, Becker SL, Knopp S, Blum J, Neumayr AL, Keiser J, et al. Neglected tropical diseases: diagnosis, clinical management, treatment and control. Swiss Med Wkly. 2012;142:w13727. doi:10.4414/smw.2012.13727.
3. Blair D. Paragonimiasis. Adv Exp Med Biol. 2014;766:115-52. doi:10.1007/978-1-4939-0915-5_5.
4. Furst T, Sayasone S, Odermatt P, Keiser J and Utzinger J. Manifestation, diagnosis, and management of foodborne trematodiasis. BMJ. 2012;344:e4093. doi:10.1136/bmj.e4093.
5. Lv S, Zhang Y, Steinmann P, Zhou XN and Utzinger J. Helminth infections of the central nervous system occurring in Southeast Asia and the Far East. Adv Parasitol. 2010;72:351-408. doi:S0065-308X(10)72012-1 [pii] 10.1016/S0065-308X(10)72012-1.
6. Sripa B, Kaewkes S, Intapan PM, Maleewong W and Brindley PJ. Food-borne trematodiases in Southeast Asia epidemiology, pathology, clinical manifestation and control. Adv Parasitol. 2010;72:305-50. doi:S0065-308X(10)72011-X [pii] 10.1016/S0065-308X(10)72011-X.
7. Liu Q, Wei F, Liu W, Yang S and Zhang X. Paragonimiasis: an important food-borne zoonosis in China. Trends Parasitol. 2008;24 7:318-23. doi:S1471-4922(08)00137-2 [pii] 10.1016/j.pt.2008.03.014.
8. Blair D, Xu ZB and Agatsuma T. Paragonimiasis and the genus Paragonimus. Adv Parasitol. 1999;42:113-222.
9. Blair D, Davis GM and Wu B. Evolutionary relationships between trematodes and snails emphasizing schistosomes and paragonimids. Parasitology. 2001;123:S229-S43. doi:Doi 10.1017/S003118200100837x.
10. Attwood SW, Upatham ES, Meng XH, Qiu DC and Southgate VR. The phylogeography of Asian Schistosoma (Trematoda: Schistosomatidae). Parasitology. 2002;125 Pt 2:99-112. doi:10.1017/s0031182002001981.

11. Doanh NP, Tu AL, Bui TD, Loan TH, Nonaka N, Horii Y, et al. Molecular and morphological variation of Paragonimus westermani in Vietnam with records of new second intermediate crab hosts and a new locality in a northern province. Parasitology. 2016;143 12:1639-46. doi:10.1017/S0031182016001219.
12. 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. 2019;8 1 doi:10.1093/gigascience/giy146.
13. Blair D, Chang Z, Chen M, Cui A, Wu B, Agatsuma T, et al. Paragonimus skrjabini Chen, 1959 (Digenea: Paragonimidae) and related species in eastern Asia: a combined molecular and morphological approach to identification and taxonomy. Syst Parasitol. 2005;60 1:1-21. doi:10.1007/s11230-004-1378-5.
14. Lane MA, Marcos LA, Onen NF, Demertzis LM, Hayes EV, Davila SZ, et al. Paragonimus kellicotti flukes in Missouri, USA. Emerg Infect Dis. 2012;18 8:1263-7. doi:10.3201/eid1808.120335.
15. Fischer PU and Weil GJ. North American paragonimiasis: epidemiology and diagnostic strategies. Expert Rev Anti-Infe. 2015;13 6:779-86. doi:10.1586/14787210.2015.1031745.
16. Blair D, Nawa Y, Mitreva M and Doanh PN. Gene diversity and genetic variation in lung flukes (genus Paragonimus). Trans R Soc Trop Med Hyg. 2016;110 1:6-12. doi:10.1093/trstmh/trv101.
17. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011;108 4:1513-8. doi:10.1073/pnas.1017351108.
18. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS One. 2012;7 11:e47768. doi:10.1371/journal.pone.0047768.
19. Xue W, Li JT, Zhu YP, Hou GY, Kong XF, Kuang YY, et al. L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics. 2013;14:604. doi:10.1186/1471-2164-14-604.
20. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT and Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33 3:290-5. doi:10.1038/nbt.3122.

21. Li BW, McNulty SN, Rosa BA, Tyagi R, Zeng QR, Gu KZ, et al. Conservation and diversification of the transcriptomes of adult Paragonimus westermani and P. skrjabini. Parasit Vectors. 2016;9:497. doi:10.1186/s13071-016-1785-x.

22. McNulty SN, Fischer PU, Townsend RR, Curtis KC, Weil GJ and Mitreva M. Systems biology studies of adult paragonimus lung flukes facilitate the identification of immunodominant parasite antigens. PLoS Negl Trop Dis. 2014;8 10:e3242. doi:10.1371/journal.pntd.0003242.

23. The UniProt C. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45 D1:D158-D69. doi:10.1093/nar/gkw1099.

24. Howe KL, Bolt BJ, Shafie M, Kersey P and Berriman M. WormBase ParaSite - a comprehensive resource for helminth genomics. Mol Biochem Parasitol. 2017;215:2-10. doi:10.1016/j.molbiopara.2016.11.005.

25. Hoff KJ, Lange S, Lomsadze A, Borodovsky M and Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32 5:767-9. doi:10.1093/bioinformatics/btv661.

26. Holt C and 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.

27. Bao W, Kojima KK and Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11. doi:10.1186/s13100-015-0041-9.

28. Lagesen K, Hallin P, Rodland EA, Staerfeldt HH, Rognes T and Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35 9:3100-8. doi:10.1093/nar/gkm160.

29. Lowe TM and Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25 5:955-64.

30. Eilbeck K, Moore B, Holt C and Yandell M. Quantitative measures for the management and comparison of annotated genomes. BMC Bioinformatics. 2009;10:67. doi:10.1186/1471-2105-10-67.

31. Jones P, Binns D, Chang HY, 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.

32. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 2014;164 2:513-24. doi:10.1104/pp.113.230144.
33. Koskinen P, Toronen P, Nokso-Koivisto J and Holm L. PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. Bioinformatics. 2015;31 10:1544-52. doi:10.1093/bioinformatics/btu851.
34. Casimiro-Soriguer CS, Munoz-Merida A and Perez-Pulido AJ. Sma3s: A universal tool for easy functional annotation of proteomes and transcriptomes. Proteomics. 2017;17 12 doi:10.1002/pmic.201700071.
35. Leinonen R, Sugawara H, Shumway M and on behalf of the International Nucleotide Sequence Database C. The Sequence Read Archive. Nucleic Acids Res. 2011;39 Database issue:D19-D21. doi:10.1093/nar/gkq1019.
36. Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30 15:2114-20. doi:10.1093/bioinformatics/btu170.
37. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29 1:15-21. doi:10.1093/bioinformatics/bts635.
38. Liao Y, Smyth GK and Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30 7:923-30. doi:10.1093/bioinformatics/btt656.
39. Anders S and Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11 10:R106. doi:10.1186/gb-2010-11-10-r106.
40. Kanehisa M, Sato Y and Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol. 2016;428 4:726-31. doi:10.1016/j.jmb.2015.11.006.
41. Rawlings ND, Barrett AJ and Finn R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2016;44 D1:D343-50. doi:10.1093/nar/gkv1118.
42. Falcon S and Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23 2:257-8. doi:10.1093/bioinformatics/btl567.

43. Emms DM and Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157. doi:10.1186/s13059-015-0721-2.
44. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018;46 D1:D754-D61. doi:10.1093/nar/gkx1098.
45. Choi Y-JC, 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. Molecular Biology and Evolution. 2019;In Press.
46. Stapley J, Santure AW and Dennis SR. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Mol Ecol. 2015;24 9:2241-52. doi:10.1111/mec.13089.
47. Schrader L and Schmitz J. The impact of transposable elements in adaptive evolution. Mol Ecol. 2018; doi:10.1111/mec.14794.
48. Chenais B, Caruso A, Hiard S and Casse N. The impact of transposable elements on eukaryotic genomes: from genome size increase to genetic adaptation to stressful environments. Gene. 2012;509 1:7-15. doi:10.1016/j.gene.2012.07.042.
49. Prasad PK, Tandon V, Biswal DK, Goswami LM and Chatterjee A. Phylogenetic reconstruction using secondary structures and sequence motifs of ITS2 rDNA of Paragonimus westermani (Kerbert, 1878) Braun, 1899 (Digenea: Paragonimidae) and related species. BMC Genomics. 2009;10 Suppl 3:S25. doi:10.1186/1471-2164-10-S3-S25.
50. Han MV, Thomas GW, Lugo-Martinez J and 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.
51. Ahn CS, Na BK, Chung DL, Kim JG, Kim JT and Kong Y. Expression characteristics and specific antibody reactivity of diverse cathepsin F members of Paragonimus westermani. Parasitol Int. 2015;64 1:37-42. doi:10.1016/j.parint.2014.09.012.
52. 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

723 Potomac Horse and Human Sennetsu Fevers. PLoS Genet. 2017;13 1:e1006537.
724 doi:10.1371/journal.pgen.1006537.
725 53. Jones MK, Gobert GN, Zhang L, Sunderland P and McManus DP. The cytoskeleton and motor proteins
726 of human schistosomes and their roles in surface maintenance and host-parasite interactions.
727 Bioessays. 2004;26 7:752-65. doi:10.1002/bies.20058.
728 54. Mathieson W and Wilson RA. A comparative proteomic study of the undeveloped and developed
729 Schistosoma mansoni egg and its contents: the miracidium, hatch fluid and secretions. Int J Parasitol.
730 2010;40 5:617-28. doi:10.1016/j.ijpara.2009.10.014.
731 55. Sotillo J, Pearson M, Becker L, Mulvenna J and Loukas A. A quantitative proteomic analysis of the
732 tegumental proteins from Schistosoma mansoni schistosomula reveals novel potential therapeutic
733 targets. Int J Parasitol. 2015;45 8:505-16. doi:10.1016/j.ijpara.2015.03.004.
734 56. Liu F, Cui SJ, Hu W, Feng Z, Wang ZQ and Han ZG. Excretory/secretory proteome of the adult
735 developmental stage of human blood fluke, Schistosoma japonicum. Mol Cell Proteomics. 2009;8
736 6:1236-51. doi:10.1074/mcp.M800538-MCP200.
737 57. Kolinski T, Marek-Trzonkowska N, Trzonkowski P and Siebert J. Heat shock proteins (HSPs) in the
738 homeostasis of regulatory T cells (Tregs). Cent Eur J Immunol. 2016;41 3:317-23.
739 doi:10.5114/ceji.2016.63133.
740 58. Pereira AS, Cavalcanti MG, Zingali RB, Lima-Filho JL and Chaves ME. Isoforms of Hsp70-binding
741 human LDL in adult Schistosoma mansoni worms. Parasitol Res. 2015;114 3:1145-52.
742 doi:10.1007/s00436-014-4292-z.
743 59. He S, Yang L, Lv Z, Hu W, Cao J, Wei J, et al. Molecular and functional characterization of a mortalin-
744 like protein from Schistosoma japonicum (SjMLP/hsp70) as a member of the HSP70 family. Parasitol
745 Res. 2010;107 4:955-66. doi:10.1007/s00436-010-1960-5.
746 60. Reis EV, Pereira RV, Gomes M, Jannotti-Passos LK, Baba EH, Coelho PM, et al. Characterisation of
747 major vault protein during the life cycle of the human parasite Schistosoma mansoni. Parasitol Int.
748 2014;63 1:120-6. doi:10.1016/j.parint.2013.10.005.
749 61. Messerli SM, Kasinathan RS, Morgan W, Spranger S and Greenberg RM. Schistosoma mansoni P-
750 glycoprotein levels increase in response to praziquantel exposure and correlate with reduced

praziquantel susceptibility. Mol Biochem Parasitol. 2009;167 1:54-9. doi:10.1016/j.molbiopara.2009.04.007.

62. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24 8:1586-91. doi:10.1093/molbev/msm088.

63. Huang S, Yuan S, Dong M, Su J, Yu C, Shen Y, et al. The phylogenetic analysis of tetraspanins projects the evolution of cell-cell interactions from unicellular to multicellular organisms. Genomics. 2005;86 6:674-84. doi:10.1016/j.ygeno.2005.08.004.

64. Chaiyadet S, Krueajampa W, Hipkaeo W, Plosan Y, Piratae S, Sotillo J, et al. Suppression of mRNAs encoding CD63 family tetraspanins from the carcinogenic liver fluke Opisthorchis viverrini results in distinct tegument phenotypes. Sci Rep. 2017;7 1:14342. doi:10.1038/s41598-017-13527-5.

65. Krautz-Peterson G, Debatis M, Tremblay JM, Oliveira SC, Da'dara AA, Skelly PJ, et al. Schistosoma mansoni Infection of Mice, Rats and Humans Elicits a Strong Antibody Response to a Limited Number of Reduction-Sensitive Epitopes on Five Major Tegumental Membrane Proteins. PLoS Negl Trop Dis. 2017;11 1:e0005306. doi:10.1371/journal.pntd.0005306.

66. Tran MH, Pearson MS, Bethony JM, Smyth DJ, Jones MK, Duke M, et al. Tetraspanins on the surface of Schistosoma mansoni are protective antigens against schistosomiasis. Nat Med. 2006;12 7:835-40. doi:10.1038/nm1430.

67. Wu C, Cai P, Chang Q, Hao L, Peng S, Sun X, et al. Mapping the binding between the tetraspanin molecule (Sjc23) of Schistosoma japonicum and human non-immune IgG. PLoS One. 2011;6 4:e19112. doi:10.1371/journal.pone.0019112.

68. Sealey KL, Kirk RS, Walker AJ, Rollinson D and Lawton SP. Adaptive radiation within the vaccine target tetraspanin-23 across nine Schistosoma species from Africa. Int J Parasitol. 2013;43 1:95-103. doi:10.1016/j.ijpara.2012.11.007.

69. Yang Y, Wen Y, Cai YN, Vallee I, Boireau P, Liu MY, et al. Serine proteases of parasitic helminths. Korean J Parasitol. 2015;53 1:1-11. doi:10.3347/kjp.2015.53.1.1.

70. Glanfield A, McManus DP, Anderson GJ and Jones MK. Pumping iron: a potential target for novel therapeutics against schistosomes. Trends Parasitol. 2007;23 12:583-8. doi:10.1016/j.pt.2007.08.018.

71. Brindley PJ, Kalinna BH, Wong JY, Bogitsh BJ, King LT, Smyth DJ, et al. Proteolysis of human hemoglobin by schistosome cathepsin D. Mol Biochem Parasitol. 2001;112 1:103-12.
72. Williamson AL, Brindley PJ, Abbenante G, Prociv P, Berry C, Girdwood K, et al. Cleavage of hemoglobin by hookworm cathepsin D aspartic proteases and its potential contribution to host specificity. FASEB J. 2002;16 11:1458-60. doi:10.1096/fj.02-0181fje.
73. Lee EG, Na BK, Bae YA, Kim SH, Je EY, Ju JW, et al. Identification of immunodominant excretory-secretory cysteine proteases of adult Paragonimus westermani by proteome analysis. Proteomics. 2006;6 4:1290-300. doi:10.1002/pmic.200500399.
74. Na BK, Kim SH, Lee EG, Kim TS, Bae YA, Kang I, et al. Critical roles for excretory-secretory cysteine proteases during tissue invasion of Paragonimus westermani newly excysted metacercariae. Cell Microbiol. 2006;8 6:1034-46. doi:10.1111/j.1462-5822.2006.00685.x.
75. Caban-Hernandez K and Espino AM. Differential expression and localization of saposin-like protein 2 of Fasciola hepatica. Acta Trop. 2013;128 3:591-7. doi:10.1016/j.actatropica.2013.08.012.
76. Basavaraju SV, Zhan B, Kennedy MW, Liu Y, Hawdon J and Hotez PJ. Ac-FAR-1, a 20 kDa fatty acid- and retinol-binding protein secreted by adult Ancylostoma caninum hookworms: gene transcription pattern, ligand binding properties and structural characterisation. Mol Biochem Parasitol. 2003;126 1:63-71.
77. Jones MK, McManus DP, Sivadorai P, Glanfield A, Moertel L, Belli SI, et al. Tracking the fate of iron in early development of human blood flukes. Int J Biochem Cell Biol. 2007;39 9:1646-58. doi:10.1016/j.biocel.2007.04.017.
78. Xu Y, Sun Z, Wang Q, Wang T, Liu Y and Yu F. Stox1 induced the proliferation and cell cycle arrest in pulmonary artery smooth muscle cells via AKT signaling pathway. Vascul Pharmacol. 2019;120:106568. doi:10.1016/j.vph.2019.106568.
79. Janssen-Heininger Y. P3 - Glutathione, Glutaredoxin And S-Glutathionylation In Lung Disease. Free Radical Biology and Medicine. 2017;112:3. doi:https://doi.org/10.1016/j.freeradbiomed.2017.10.350.
80. Børud B, Mellgren G, Lund J and Bakke M. Cloning and Characterization of a Novel Zinc Finger Protein that Modulates the Transcriptional Activity of Nuclear Receptors. Mol Endocrinol. 2003;17 11:2303-19. doi:10.1210/me.2003-0158 %J Molecular Endocrinology.

81. Taylor BL and Zhulin IB. PAS domains: internal sensors of oxygen, redox potential, and light. Microbiology and molecular biology reviews : MMBR. 1999;63 2:479-506.
82. Jernigan KK and Bordenstein SR. Ankyrin domains across the Tree of Life. PeerJ. 2014;2:e264. doi:10.7717/peerj.264.
83. Grote A, Caffrey CR, Rebello KM, Smith D, Dalton JP and Lustigman S. Cysteine proteases during larval migration and development of helminths in their final host. PLoS Negl Trop Dis. 2018;12 8:e0005919. doi:10.1371/journal.pntd.0005919.
84. Taylor AB, Roberts KM, Cao X, Clark NE, Holloway SP, Donati E, et al. Structural and Enzymatic Insights into Species-specific Resistance to Schistosome Parasite Drug Therapy. 2017; doi:10.1074/jbc.M116.766527.
85. Pao SS, Paulsen IT and Saier MH, Jr. Major facilitator superfamily. Microbiol Mol Biol Rev. 1998;62 1:1-34.
86. World Health Organization. 2019. Accessed August 25, 2019.
87. Kyung SY, Cho YK, Kim YJ, Park JW, Jeong SH, Lee JI, et al. A paragonimiasis patient with allergic reaction to praziquantel and resistance to triclabendazole: successful treatment after desensitization to praziquantel. Korean J Parasitol. 2011;49 1:73-7. doi:10.3347/kjp.2011.49.1.73.
88. Kelley JM, Elliott TP, Beddoe T, Anderson G, Skuce P and Spithill TW. Current Threat of Triclabendazole Resistance in Fasciola hepatica. Trends Parasitol. 2016; doi:10.1016/j.pt.2016.03.002.
89. Mader P, Rennar GA, Ventura AMP, Grevelding CG and Schlitzer M. Chemotherapy for Fighting Schistosomiasis: Past, Present and Future. ChemMedChem. 2018;13 22:2374-89. doi:10.1002/cmdc.201800572.
90. Stutzer C, Richards SA, Ferreira M, Baron S and Maritz-Olivier C. Metazoan Parasite Vaccines: Present Status and Future Prospects. Front Cell Infect Microbiol. 2018;8:67. doi:10.3389/fcimb.2018.00067.
91. Radzikowska E, Chabowski M and Bestry I. Tuberculosis mimicry. Eur Respir J. 2006;27 3:652; author reply doi:10.1183/09031936.06.00121205.
92. Eapen S, Espinal E and Firstenberg M. Delayed diagnosis of paragonimiasis in Southeast Asian immigrants: A need for global awareness. 2018;4 2:173-7. doi:10.4103/ijam.Ijam_2_18.

93. Yang SH, Park JO, Lee JH, Jeon BH, Kim WS, Kim SI, et al. Cloning and characterization of a new cysteine proteinase secreted by Paragonimus westermani adult worms. Am J Trop Med Hyg. 2004;71 1:87-92.

94. Yoonuan T, Nuamtanong S, Dekumyoy P, Phuphisut O and Adisakwattana P. Molecular and immunological characterization of cathepsin L-like cysteine protease of Paragonimus pseudoheterotremus. Parasitol Res. 2016;115 12:4457-70. doi:10.1007/s00436-016-5232-x.

95. Kim SH and Bae YA. Lineage-specific expansion and loss of tyrosinase genes across platyhelminths and their induction profiles in the carcinogenic oriental liver fluke, Clonorchis sinensis. Parasitology. 2017;144 10:1316-27. doi:10.1017/S003118201700083X.

96. Bae YA, Kim SH, Ahn CS, Kim JG and Kong Y. Molecular and biochemical characterization of Paragonimus westermani tyrosinase. Parasitology. 2015;142 6:807-15. doi:10.1017/S0031182014001942.

97. Pothong K, Komalamisra C, Kalambaheti T, Watthanakulpanich D, Yoshino TP and Dekumyoy P. ELISA based on a recombinant Paragonimus heterotremus protein for serodiagnosis of human paragonimiasis in Thailand. Parasit Vectors. 2018;11 1:322. doi:10.1186/s13071-018-2878-5.

98. Bae YA, Kim SH, Cai GB, Lee EG, Kim TS, Agatsuma T, et al. Differential expression of Paragonimus westermani eggshell proteins during the developmental stages. Int J Parasitol. 2007;37 3-4:295-305. doi:10.1016/j.ijpara.2006.10.006.

99. Fischer PU, Curtis KC, Marcos LA and Weil GJ. Molecular characterization of the North American lung fluke Paragonimus kellicotti in Missouri and its development in Mongolian gerbils. Am J Trop Med Hyg. 2011;84 6:1005-11. doi:10.4269/ajtmh.2011.11-0027.

100. 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. 2017; doi:10.1093/molbev/msx319.

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

102. 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. doi:10.1371/journal.pone.0112963.

103. Sahm A, Bens M, Platzer M and Szafranski K. PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes. Nucleic Acids Res. 2017;45 11:e100. doi:10.1093/nar/gkx179.

104. Kahsay RY, Gao G and Liao L. An improved hidden Markov model for transmembrane protein detection and topology prediction and its applications to complete genomes. Bioinformatics. 2005;21 9:1853-8. doi:10.1093/bioinformatics/bti303.

105. Omasits U, Ahrens CH, Muller S and Wollscheid B. Protter: interactive protein feature visualization and integration with experimental proteomic data. Bioinformatics. 2014;30 6:884-6. doi:10.1093/bioinformatics/btt607.

106. Benjamini Y and Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society Series B (Methodological). 1995;57 1:289-300. doi:10.2307/2346101.

107. Shield AJ, Murray TP, Cappello JY, Coggan M and Board PG. Polymorphisms in the human glutathione transferase Kappa (GSTK1) promoter alter gene expression. Genomics. 2010;95 5:299-305. doi:https://doi.org/10.1016/j.ygeno.2010.02.007.

889 **Figure Captions**

890

891 **Figure 1.** Comparisons of the overall content of the assembled *Paragonimus* genome assemblies. Comparisons
892 are based on (**A**) length (including statistics for other sequenced trematode genomes) and (**B**) Repeat
893 landscapes, measured using the Kimura substitution level, which indicates how much a repeat sequence has
894 degenerated since its incorporation into the genome (i.e., how recently the repeat sequence was added). The
895 high peak at the far left of *P. kellicotti* indicates a recent incorporation or active transposable element activity.

896

897 **Figure 2**: Comparison of genome annotation characteristics and attributes among several species of flatworms.
898 Attributes characterized included (**A**) Full gene lengths, including coding and noncoding sequences, (**B**) Average
899 intron lengths per gene, (**C**) Number of exons per gene, and (**D**) Coding sequence (CDS) length per exon. *P*
900 values and letter groupings indicating significant differences among species, as calculated using ANOVA with
901 Tukey's HSD post-hoc test.

902

903 **Figure 3.** Clustering of *Paragonimus* species. (**A**) Mitochondrial whole genome-based phylogeny, including
904 previously-sequenced *Paragonimus* mitochondrial genomes (with accessions indicated). (**B**) Species clustering
905 based on single-member OPF sequences. 262,720 genes (85% of all genes across the species) were assigned
906 to 17,953 OPFs; 2,493 genes are in 326 species-specific OPFs.

907

908 **Figure 4.** Gene-family dynamics among platyhelminth species. (**A**) Rapidly evolving families of interest are
909 quantified at each stage of the phylogeny, including genes gained (blue) and lost (red) relative to other species.
910 The number of rapidly evolving genes are indicated in parentheses. (**B**) Functionally annotated gene families of
911 interest that displayed most pronounced differential expansions or contractions. (**C**) Overall protease and
912 protease inhibitor abundance per species.

913

914 **Figure 5.** Orthologous Group (OG) distribution analysis. (**A**) OGs identified among groups of flukes. The OGs
915 conserved in at least one of the species from each group are indicated in black, and the OGs conserved among
916 all the species in the overlapping groups are indicated in red. (**B**) Counts of OGs among the four *Paragonimus*

species, with *Paragonimus*-specific gene sets indicated in red text. The 256 *Paragonimus* conserved-and-specific genes are indicated with highlight (Table 5). (**C**) Significant functional enrichment (Interpro domains) among the gene sets conserved among, and specific to each major group of flukes (256, 758 and 270 OPFs in lung, liver and blood flukes, respectively), relative to the functions in the complete gene sets. (**D**) Counts of OGs only shared between the *Paragonimus* species and one or more of the host mammal and mollusk species, but not the other platyhelminth species (potential host-interacting OGs). The potential host-interacting OGs conserved among all four *Paragonimus* species are highlighted.

Figure 6: Analysis of gene expression data for species of lung flukes of the genus *Paragonimus*. (**A**) Comparison of adult-stage gene expression levels among 1:1 orthologs shared by *P. westermani* and *P. miyazakii*. Pearson correlation = 0.79. (**B**) Pearson correlation values between all lung fluke species for the adult-stage expression levels of all 1:1 orthologous genes. (**C**) Differential gene expression between cavities (blue) and tissues (orange) in *P. miyazakii*. Clustering based on FPKM value across all genes is indicated in the bottom right (Pearson clustering, complete linkage). (**D**) A comparison of the average relative expression of the lung fluke-specific and -conserved genes in each *P. miyazakii* tissue type. ** $P < 0.01$, *** $P < 0.001$, according to an ANOVA test of all Z-score values.

Tables

Table 1: The draft genome of *Paragonimus*: assembly, size and annotation characteristics

Statistic	*Paragonimus miyazakii*	*Paragonimus heterotremus*	*Paragonimus kellicotti*	*Paragonimus westermani* (Japan)	*Paragonimus westermani* (India) [12]
Total genome length (Mb)	915.8	841.2	696.5	923.3	922.8
Number of contigs	22,318	27,557	29,377	22,477	30,455
Mean contig size (kb)	41	30.5	23.7	41.1	30.3
Median contig size (kb)	15.1	9.3	10.2	17.2	4.8
Max. contig size (kb)	919.8	715.6	826	829	809.4
N50 length (kb)	108.8	92.5	56.0	100.8	135.2
N50 number	2,320	2,506	3,316	2,664	1,943
Number of genes	12,652	12,490	12,853	12,072	12,771
Avg gene length (kb)	25.9	22.6	17.6	24.1	18.0
Avg CDS length (kb)	1.5	1.4	1.1	1.4	1.4
Avg intron length (kb)	4.2	4	3.6	4.2	4.0
Avg # exons per gene	6.7	6.2	5.3	6.3	5.2
BUSCO completeness	93.4%	93.4%	86.8%	96.4%	94.1%
% annotated InterPro	82%	85%	81%	87%	82%
% annotated KEGG	40%	41%	34%	43%	43%

Table 2. "Molecular Function" Gene Ontology terms enriched among *P. miyazakii* genes that are conserved among and exclusive to lung flukes.

GO ID	GO term name	P value	# Conserved and Specific	Total # in genome
GO:0004175	endopeptidase activity	5.2E-05	8	132
GO:0008236	serine-type peptidase activity	5.6E-05	6	67
GO:0017171	serine hydrolase activity	5.6E-05	6	67
GO:0004252	serine-type endopeptidase activity	1.6E-04	5	51
GO:0070011	peptidase activity, acting on L-amino acid peptides	6.1E-04	9	237
GO:0008233	peptidase activity	8.7E-04	9	249
GO:0004568	chitinase activity	2.1E-03	2	7
GO:0004190	aspartic-type endopeptidase activity	1.1E-02	2	16
GO:0070001	aspartic-type peptidase activity	1.1E-02	2	16
GO:0008199	ferric iron binding	1.1E-02	2	16

Table 3: Annotations of "host-mimic" orthologous groups that are conserved among the four *Paragonimus* species and one or more of their host taxa, but no other platyhelminth species.

Host conservation	OPF	Annotated genes in OPF		Known function
		Species	Gene	
Both mammal and mollusk	OG0008088	Human	STOX1	Induces proliferation of pulmonary artery smooth muscle cells [78]
	OG0007415	Human	C5orf63 (Glutaredoxin-like protein)	Another Glutaredoxin gene (Grx1) impairs bacterial clearance from the lungs and increases severity of pneumonia [79]
	OG0007990	Human	GSTK1	Associated with insulin secretion and fat deposition [107]
	OG0008725	Human	LRRC58	Leucine rich repeats - No annotated function specific to this protein
Mammal	OG0005049	Human	Zip67	Transcriptional repressor of genes including PPARG - involved in inflammation reduction, highly expressed in alveolar macrophages [80]
	OG0006588	Human	PAC domain-containing protein	Voltage-gated potassium channel; Possibly involved in oxygen detection [81]
Mollusk	OG0008337	*P. miyazakii*	IPR domain: Zinc finger C2H2-type	Common transcription factor domain
	OG0008914	*P. miyazakii*	KEGG: euchromatic histone-lysine N-methyltransferase (EHMT1) IPR: Ankyrin repeat domains	EHMT1: Histone modification Ankyrin repeat: host-parasite interactions by intracellular bacteria [82]

Table 4. "Molecular Function" Gene Ontology terms enriched among the 216 *P. miyazakii* genes that are overexpressed in cavities (peritoneal, pleural) relative to tissues (lung, liver), and among the 172 overexpressed in tissues relative to cavities.

Sample group	GO term	Term	FDR-corrected P value	Number of genes over-expressed
Cavities	GO:0008234	cysteine-type peptidase activity	1.0E-05	8
	GO:0005509	calcium ion binding	1.1E-05	13
	GO:0008233	peptidase activity	3.0E-04	11
	GO:0046872	metal ion binding	6.6E-04	17
	GO:0043169	cation binding	7.6E-04	17
	GO:0070011	peptidase activity, acting on L-amino acid peptides	8.3E-04	10
	GO:0008375	acetylglucosaminyltransferase activity	8.4E-04	3
	GO:0008194	UDP-glycosyltransferase activity	1.5E-03	3
	GO:0008146	sulfotransferase activity	2.4E-03	2
	GO:0016787	hydrolase activity	2.4E-03	21
	GO:0005544	calcium-dependent phospholipid binding	5.6E-03	2
	GO:0016782	transferase activity, transferring sulfur-containing groups	6.9E-03	2
	GO:0005506	iron ion binding	7.2E-03	3
Tissues	GO:0005200	structural constituent of cytoskeleton	1.1E-07	8
	GO:0016829	lyase activity	2.7E-04	6
	GO:0017111	nucleoside-triphosphatase activity	5.5E-04	14
	GO:0016462	pyrophosphatase activity	6.9E-04	14
	GO:0016818	hydrolase activity, acting on acid anhydrides…	7.6E-04	14
	GO:0004634	phosphopyruvate hydratase activity	8.2E-04	2
	GO:0016817	hydrolase activity, acting on acid anhydrides	8.3E-04	14
	GO:0003924	GTPase activity	1.6E-03	8
	GO:0003824	catalytic activity	2.0E-03	48
	GO:0016836	hydro-lyase activity	2.1E-03	3
	GO:0016835	carbon-oxygen lyase activity	2.6E-03	3
	GO:0003777	microtubule motor activity	3.1E-03	5
	GO:0016830	carbon-carbon lyase activity	3.6E-03	3
	GO:0016491	oxidoreductase activity	3.6E-03	10
	GO:0003774	motor activity	4.6E-03	5
	GO:0005198	structural molecule activity	6.6E-03	8
	GO:0032561	guanyl ribonucleotide binding	6.8E-03	8
	GO:0032550	purine ribonucleoside binding	6.8E-03	8
	GO:0001883	purine nucleoside binding	6.8E-03	8
	GO:0005525	GTP binding	6.8E-03	8
	GO:0032549	ribonucleoside binding	7.3E-03	8
	GO:0001882	nucleoside binding	7.6E-03	8
	GO:0019001	guanyl nucleotide binding	8.1E-03	8
	GO:0008017	microtubule binding	9.9E-03	4

977 **Table 5**: *Paragonimus*-conserved and specific genes with relatively high expression levels in the *P. miyazakii*
978 lung (adult) stage relative to other stages (minimum 1.5-fold expression difference compared to all other stages).
979

Gene	Gene function (InterPro)	Expression level (FPKM)			
		Peritoneal	Pleural	Liver	Lung
PMIY_10706	-	0	0	0	1.00
PMIY_04932	-	0	0	0	0.50
PMIY_04199	-	0	0	0	0.23
PMIY_05623	IPR009003: Peptidase S1, PA clan (Serine protease)	0	0	0	0.03
PMIY_05645	-	0.03	0	0	1.08
PMIY_10421	-	0	0	0	0.03
PMIY_10154	-	0.03	0.17	0.25	2.10
PMIY_10315	-	0.30	0.26	0	1.67
PMIY_07404	IPR028089: Domain of unknown function DUF4455	0.05	0	0	0.22
PMIY_06071	IPR035914: Spermadhesin, CUB domain superfamily	0.16	0.11	0.20	0.80
PMIY_09494	-	1.61	1.76	0.08	5.88
PMIY_01683	-	0.24	0.18	0.05	0.79
PMIY_12118	IPR021109: Aspartic peptidase domain superfamily	0.02	0	0	0.04
PMIY_06957	-	0.87	1.68	1.65	3.55
PMIY_00756	-	14.75	20.66	16.12	43.07
PMIY_01507	-	0.23	0.38	0.39	0.79
PMIY_11874	-	0.05	0.05	0	0.10
PMIY_05793	-	0.03	0.06	0.26	0.47
PMIY_05272	-	5.44	4.20	4.22	9.67
PMIY_12189	IPR036259: MFS transporter superfamily	1.84	1.72	1.73	3.25
PMIY_12491	-	0.63	0	0.18	1.08
PMIY_08946	-	0	0.06	0.00	0.10
PMIY_12111	-	1.23	0.35	1.04	2.04
PMIY_02371	IPR016024: Armadillo-type fold	10.30	9.68	12.40	20.06
PMIY_11682	-	4.36	6.64	5.66	10.58
PMIY_12247	IPR036236: Zinc finger C2H2 superfamily	2.84	3.22	4.16	6.60
PMIY_06566	-	0.63	1.02	1.04	1.64
PMIY_02942	-	5.32	5.71	8.45	13.11
PMIY_11404	-	2.40	2.82	2.58	4.36
PMIY_07606	IPR009060: UBA-like superfamily	1.33	0.12	0.49	2.03

980

Figure 1

Click here to download Figure Figure 1 (resized).tif

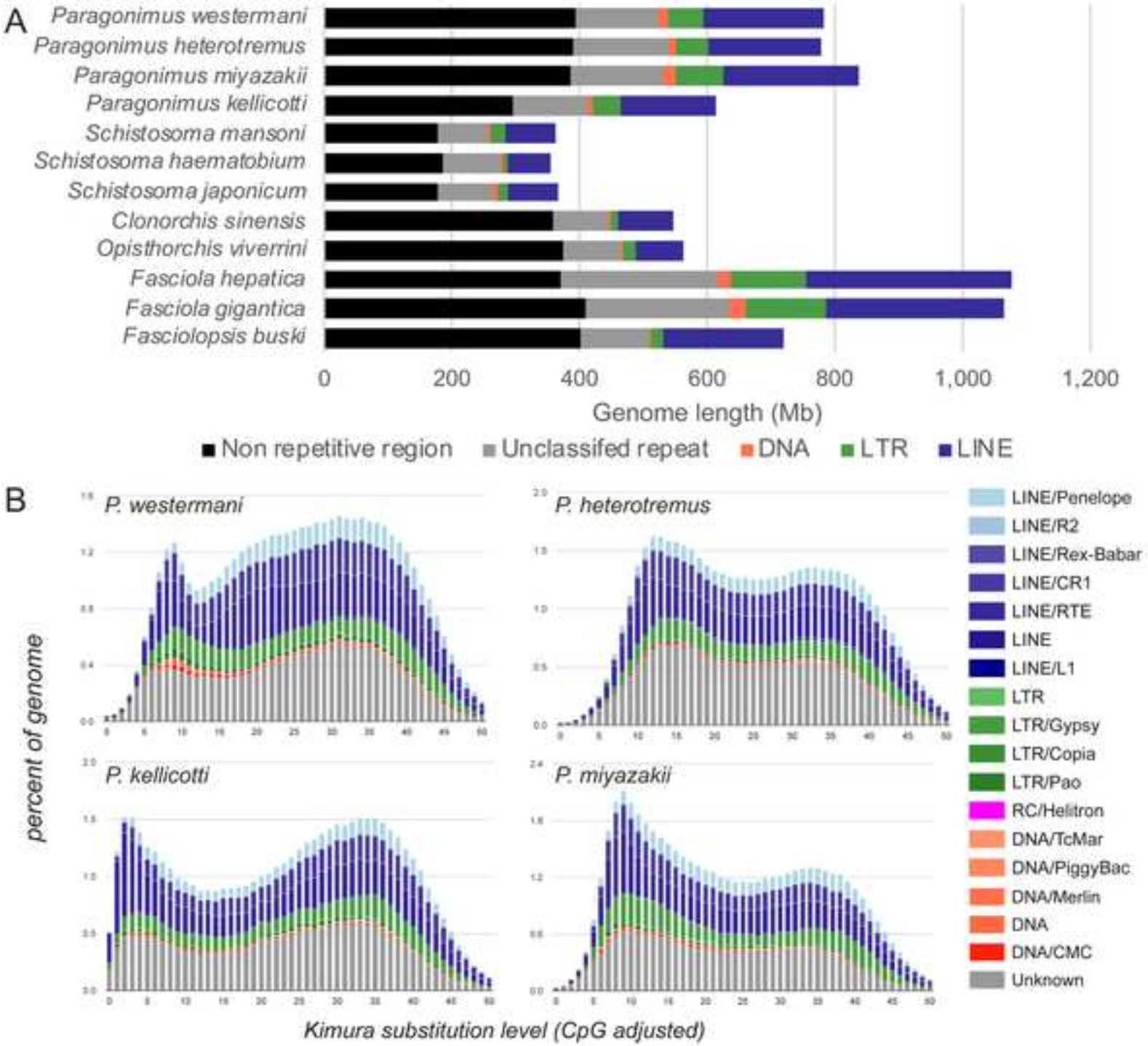


Figure 2

Click here to download Figure Figure 2 (resized).tif

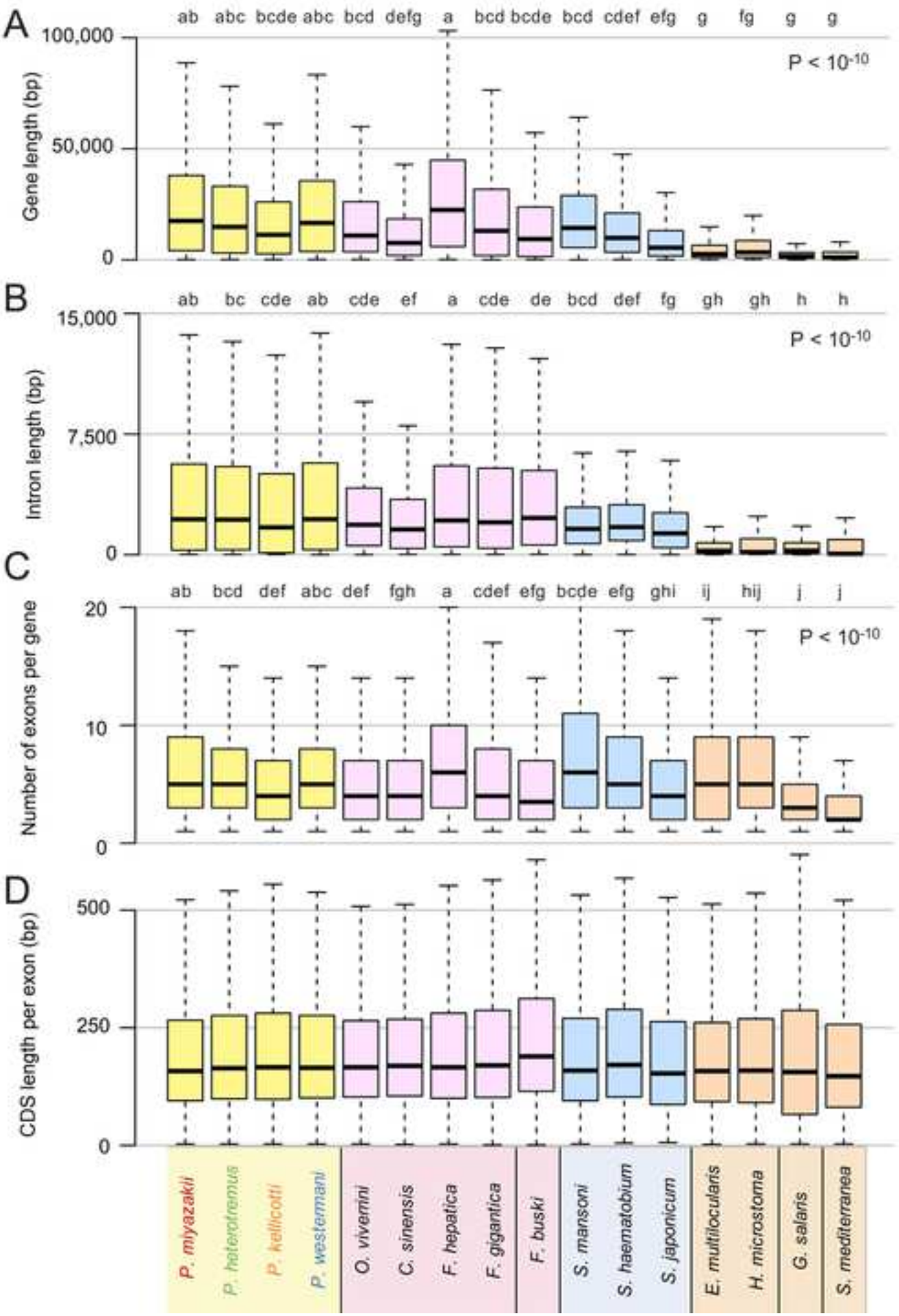


Figure 3

Click here to download Figure Figure 3 (resized).tif

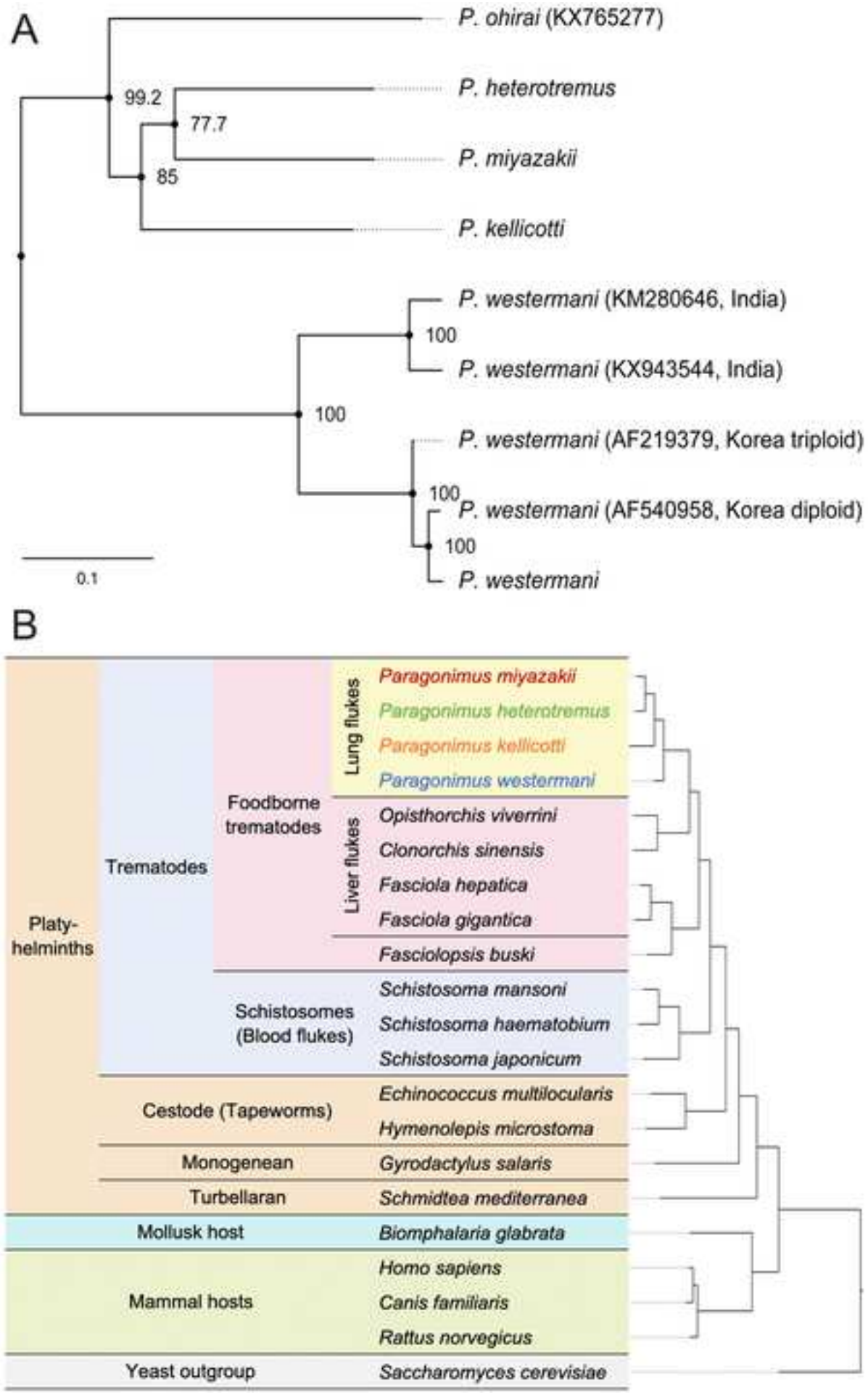


Figure 4

Click here to download Figure Figure 4 (resized).tif

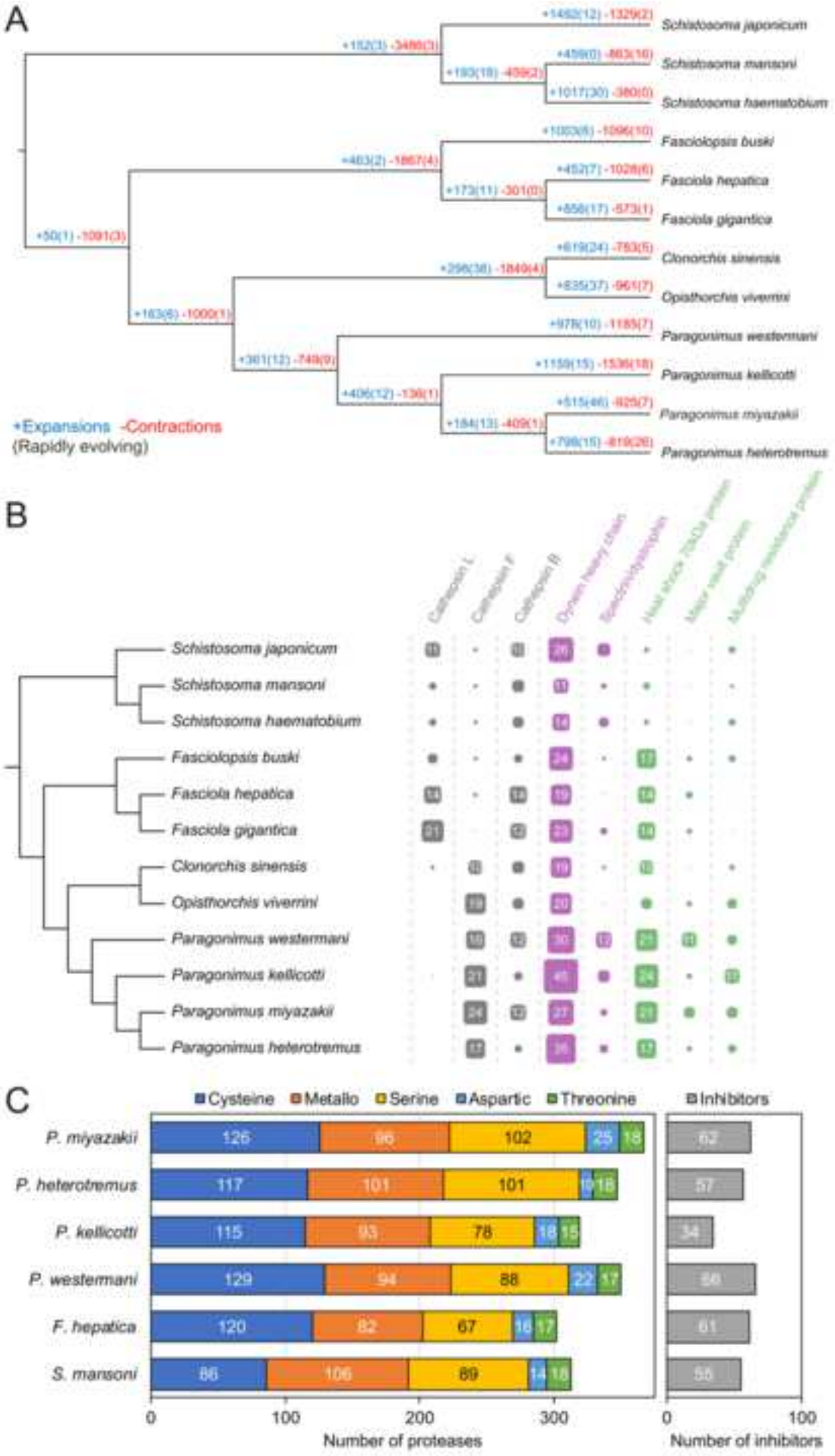


Figure 5

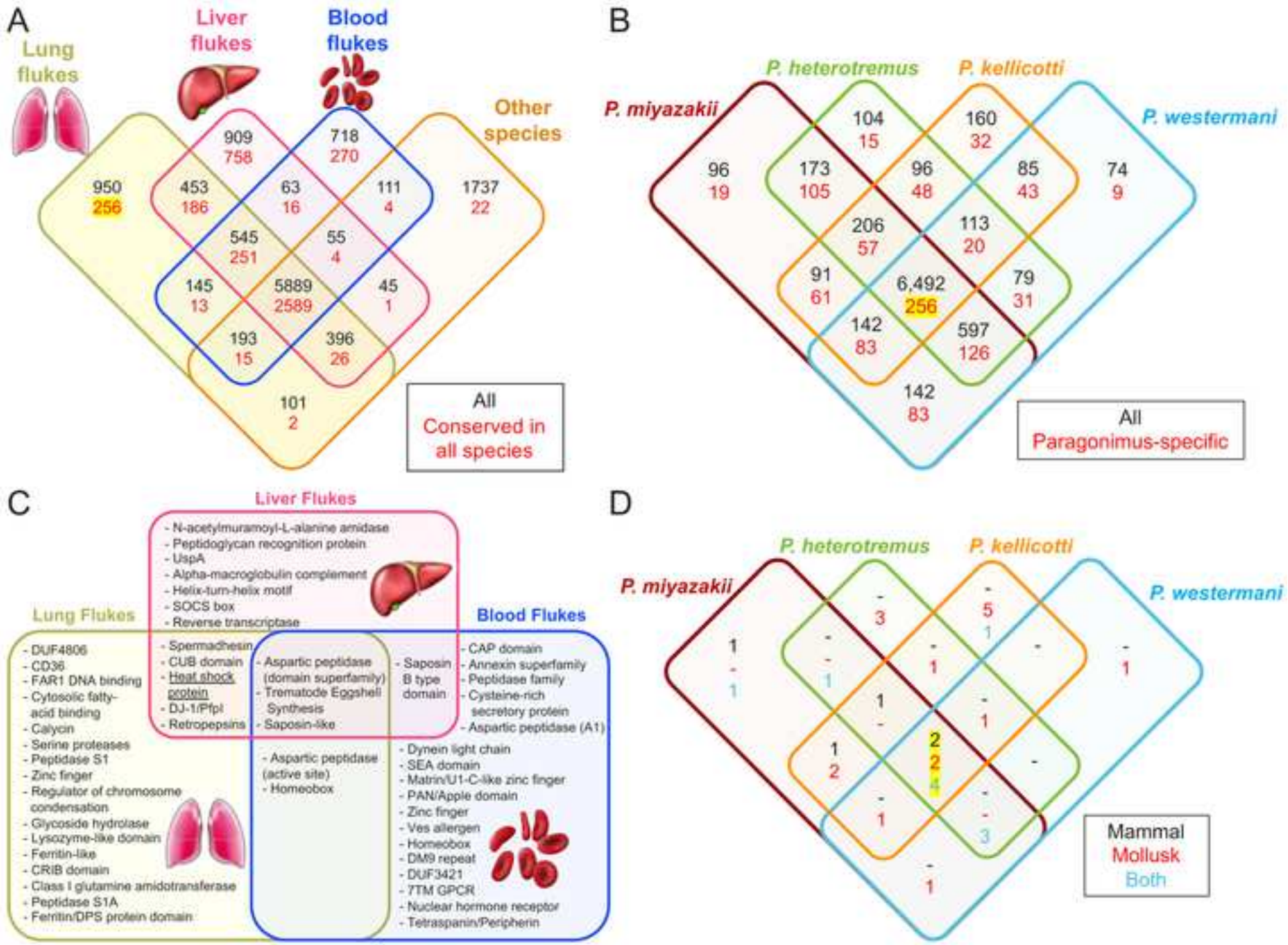
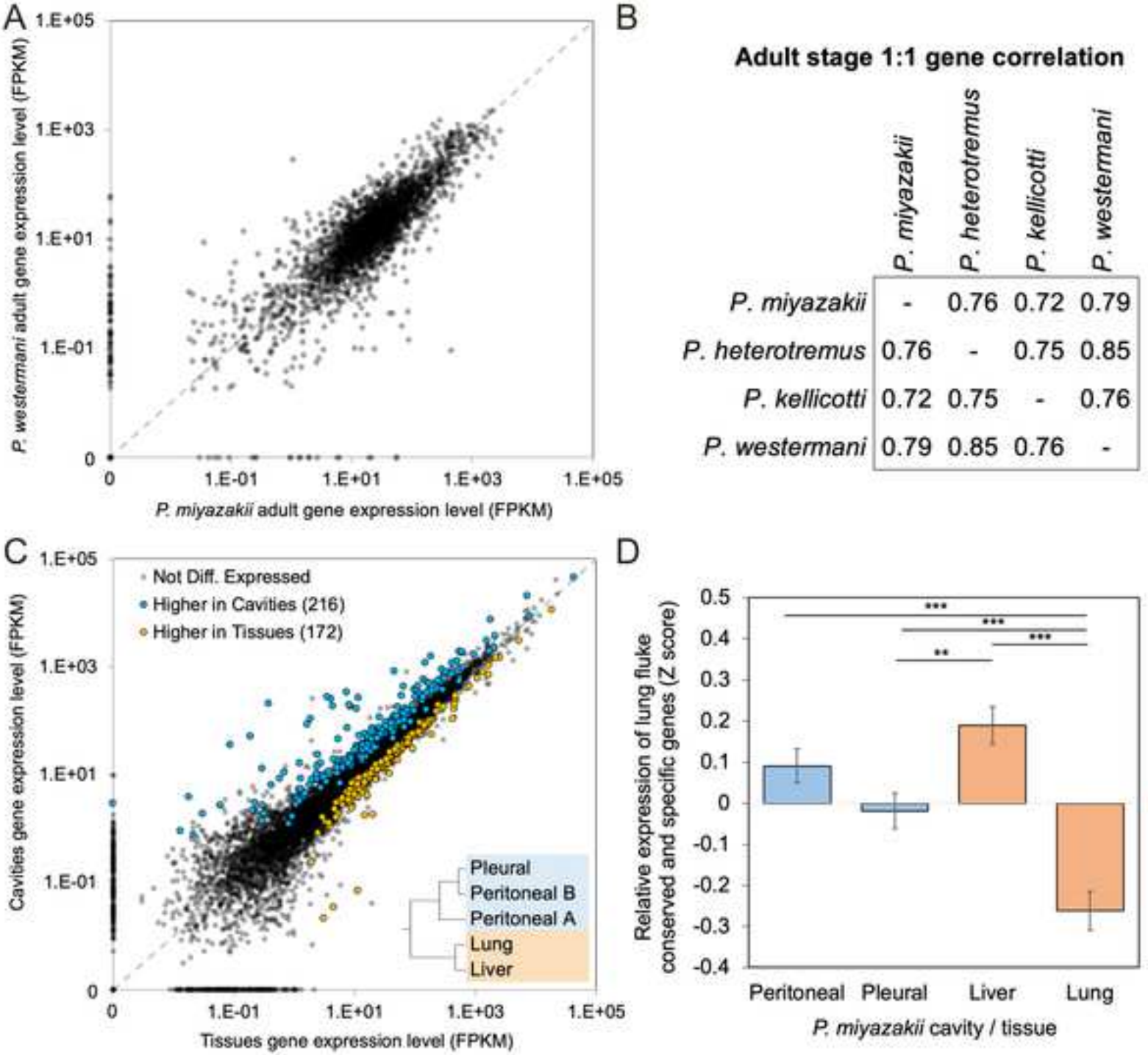


Figure 6

Click here to download Figure Figure 6 (resized).tif



Click here to access/download

Supplementary Material

Supp Table S2 - Paragonimus expression data per gene.xlsx

Washington University in St. Louis
School of Medicine
McDonnell Genome Institute

To
Dr. Laurie Goodman
Editor-in-Chief
Gigascience

11/26/2019

Dear Dr. Goodman,

Please accept our manuscript entitled "*Comparative genomics and transcriptomics of four* Paragonimus *species provide insights into lung fluke parasitism and pathogenesis*" as a candidate for publication in Gigascience.

In this manuscript, we present novel draft genomes, transcriptomes and detailed comparative analyses for four species of *Paragonimus* (lung flukes), which are among the most injurious food-borne helminths, infecting ~23 million people, with ~293 million people at risk for infection. The disease is easily mistaken for tuberculosis due to similar pulmonary symptoms, so there is an urgent need for the development of effective diagnostics, but currently the genomic resources for developing these are lacking.

In addition to characterizing and comparing the four genomes, we also performed orthologous group (OG) analysis spanning 21 species (lung, liver and blood flukes, additional platyhelminths and hosts), to provide insights into lung fluke biology. Among other findings, we identified 256 lung fluke-specific and conserved OGs that were enriched for many parasite-related functions including iron acquisition and immune modulation, and eight "host mimic" OGs including a STOX1 ortholog (proliferation of pulmonary artery cells). Our lung fluke transcriptome analysis identified consistent adult-stage *Paragonimus* expression profiles, and 388 genes differentially expressed between host body cavities and tissues that were enriched for functions including proteolysis, nutrient transport and iron acquisition. In order to accelerate diagnostic antigen research, we matched the new gene sets to previously identified *Paragonimus* diagnostic antigens, providing an opportunity to optimize and ensure consistent cross-reactivity for diagnostic assays.

We anticipate that these novel genomic and transcriptomic resources will be invaluable for future lung fluke research, and flatworm research in general. This report represents a major contribution to ongoing trematode genome sequencing efforts and bootstraps future studies into the biology, evolution and pathogenesis of *Paragonimus* and related food-borne flukes. This study was possible due to the interdisciplinary collaboration among teams across the world (USA, Australia, Japan, Vietnam and Thailand) with expertise in genomics / systems biology, evolutionary genomics and experimental parasitology and pathology.

We believe that our paper is suitable for this journal based on the following:

- The manuscript contains detailed comprehensive analyses of previously unavailable but critically important omics resources for pathogens of importance to human health (the lung flukes).
- We provide accessions for all raw data used in the data production, disseminating all available data to the community.

- We additionally provide supplementary tables formatted as comprehensive Excel databases, that include all gene expression data for every sample in every species, complete functional annotation data for all *P. miyazakii* genes, all calculated differential expression data for all genes, and complete orthologous protein group (OG) data including all gene members spanning 21 species, and the average expression data across the four lung flukes.
- These interactive datasets provide simple open access to all interested researchers, regardless of bioinformatic skill level, and provide all relevant data to reproduce analyses presented in the manuscript.

The manuscript has not been previously published, nor is submitted for publication elsewhere. The authors declare no potential competing interests, and all authors have approved the manuscript for submission. It has been formatted according to the author instructions on the journal's website.

We thank you for your kind consideration.

Yours sincerely,

Makedonka Mitreva, PhD

Professor, Department of Medicine and of Genetics,
Assistant Director, McDonnell Genome Institute,
Director, Center for Clinical Genomics of Microbial Systems,
Washington University School of Medicine

We suggest the following reviewers to evaluate the manuscript, based on their research backgrounds:

James Wasmuth
jwasmuth@ucalgary.ca
Associate Professor
Faculty of Veterinary Medicine
University of Calgary, Calgary, AB, Canada

Neil Young
nyoung@unimelb.edu.au
NHMRC Career Development Fellow
Faculty of Veterinary and Agricultural Sciences
University of Melbourne, Melbourne, Australia

Aaron Jex
ajex@unimelb.edu.au
Associate Professor
Faculty of Veterinary and Agricultural Sciences
University of Melbourne, Melbourne, Australia

Philip Loverde
loverde@uthscsa.edu
Professor
Graduate School of Biomedical Sciences
UT Health San Antonio, San Antonio, TX, USA

Tim Anderson
tanderso@txbiomed.org
Professor, Program Co-lead
Disease Intervention & Prevention
Texas Biomedical Research Institute, San Antonio, TX, USA