Gene duplication and gain in the trematode *Atriophallophorus winterbourni* contributes to adaptation to parasitism

Natalia Zajac*[,1,2], Stefan Zoller[2], Katri Seppälä[1,4], David Moi[5,6,7], Christophe Dessimoz[5,6,7,8,9], Jukka Jokela[1,2], Hanna Hartikainen[1,2,3], Natasha Glover[5,6,7]

1. Eawag, Swiss Federal Institute of Aquatic Science and Technology, CH-8600 Dübendorf, Switzerland
2. ETH Zurich, Department of Environmental Systems Science, Institute of Integrative Biology, CH-8092 Zurich, Switzerland
3. School of Life Sciences, University of Nottingham, University Park, NG7 2RD, Nottingham, UK
4. Research Department for Limnology, University of Innsbruck, 5310 Mondsee, Austria
5. Department of Computational Biology, University of Lausanne 1015 Lausanne, Switzerland
6. Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland
7. Center for Integrative Genomics, 1015 Lausanne, Switzerland
8. Centre for Life's Origins and Evolution, Department of Genetics Evolution and Environment, University College London, Gower St, London WC1E 6BT, UK
9. Department of Computer Science, University College London, Gower St, London WC1E 6BT, UK

*Author for Correspondence: Natalia Zajac, ETH Zurich, Department of Environmental Systems Science, Institute of Integrative Biology, Zurich, Switzerland, +41 58 765 1122, natalia.zajac@usys.ethz.ch

Abstract

Gene duplications and novel genes have been shown to play a major role in helminth adaptation to a parasitic lifestyle because they provide the novelty necessary for adaptation to a changing environment, such as living in multiple hosts. Here we present the *de novo* sequenced and annotated genome of the parasitic trematode *Atriophallophorus winterbourni* and its comparative genomic analysis to other major parasitic trematodes. First, we reconstructed the species phylogeny, and dated the split of *A. winterbourni* from the Opisthorchiata suborder to approximately 237.4 MYA (± 120.4 MY). We then addressed the question of which expanded gene families and gained genes are potentially involved in adaptation to parasitism. To do this, we used Hierarchical Orthologous Groups to reconstruct three ancestral genomes on the phylogeny leading to *A. winterbourni* and performed a GO enrichment analysis of the gene composition of each ancestral genome, allowing us to characterize the subsequent genomic changes. Out of the 11,499 genes in the *A. winterbourni* genome, as much as 24% have arisen through duplication events since the speciation of *A. winterbourni* from the Opisthorchiata, and as much as 31.9% appear to be novel, *i.e.* newly acquired. We found 13 gene families in *A. winterbourni* to have had more than 10 genes arising through these recent duplications; all of which have functions potentially relating to host behavioural manipulation, host tissue penetration, and hiding from host immunity through antigen presentation. We identified several families with genes evolving under positive selection. Our results provide a valuable resource for future studies on the genomic basis of adaptation to parasitism and point to specific candidate genes putatively involved in antagonistic host-parasite adaptation.

Keywords: comparative genomics, evolution, phylogeny, selection

Significance statement

Transition to parasitism has been associated with gene duplication and gain of novel genes for host exploitation, invasion, and escape from host immunity. In our study, we trace gene duplications and gains across a phylogeny from an ancestral trematode genome to our focal species, the newly sequenced trematode *Atriophallophorus winterbourni*. We characterize gene duplications and gains in 3 ancestral genomes leading to *A. winterbourni* and outline candidate gene families that have recently undergone duplication and are potentially involved in parasitism.

Introduction

The adoption of a parasitic lifestyle represents a major niche shift that has occurred multiple times across the tree of life (Poulin & Randhawa, 2015; Weinstein & Kuris, 2016). The similar selective pressures involved in exploiting hosts have resulted in convergent macroevolutionary features, such as a tendency for morphological simplification (O'Malley et al., 2016) and the associated genome compaction, reduction and streamlining across many parasite lineages (Chang et al., 2015; Lu et al., 2019; Peyretaillade et al., 2011; Slyusarev et al., 2020). At the same time, parts of the parasite genome involved in e.g. host exploitation and life-cycle complexity may have experienced expansions. Comparative genomic analyses have implied that gene duplications can drive innovation in gene function during radiations of parasitic lineages (Zarowiecki & Berriman, 2015).

Novel gene functions involved in the response to host immunity may be particularly important for the evolution of parasitism. For example, mucins, a family of heavily glycosylated surface epithelial proteins, have undergone multiple rounds of duplication in the blood fluke, *Schistosoma mansoni.* Mucins frequently recombine, generating antigenic variation through splice variants (Roger et al., 2008). Increased life-cycle complexity, especially within the parasitic flatworms (Poulin & Randhawa, 2015), may have also driven the evolution of functional novelty involved in host exploitation strategies. For instance in *S. mansoni,* multiple duplication events in the gene superfamily SCP/TAPS (sperm-coating protein/TPx/antigen 5/pathogenesis-related protein 1) have led to an array of proteins that are now associated with an active role in penetration of the snail host tissues (Cantacessi & Gasser, 2012). Duplicated genes, which evolve beyond sequence recognition, can also give rise to lineage-specific genes ("gained" genes), which can confer specific, novel traits, important in adaptation of that lineage to its particular niche (David et al., 2008; Takeuchi et al., 2016).

With the whole genome sequences of over 30 nematodes (roundworms) and 25 platyhelminth (flatworms, including trematodes) species, it has been possible to characterize the births and expansions of new gene families arising by duplication at key taxonomic levels (Rödelsperger, 2018). Nematodes and platyhelminths are two invertebrate animal phyla consisting of parasitic and free living organisms with the parasitic ones causing major animal, crop and human diseases, as well as being a major economic burden (Disease and Injury Incidence and Prevalence Collaborators, GBD, 2016; International Helminth Genomes, Consortium, 2019).

The microphallid *Atriophallophorus winterbourni* (syn. *Microphallus sp.* or *Microphallus livelyi*) is a digenean trematode parasite native to the lakes of New Zealand (Blasco-Costa et al., 2020). It alternates between two hosts in its life cycle; the intermediate host is *Potamopyrgus antipodarum,* a prosobranch dioecious mud snail (Warwick, 1952; Winterbourn, 1970) and the final hosts are waterfowl, mainly dabbling ducks (Lively & McKenzie, 1991). Multi-host life cycle is a general characteristic of all digenean trematodes, and always includes a molluscan species as an intermediate host and a vertebrate as the final host (Galaktionov & Dobrovolskij, 2003) (Supplementary Box 1). The metacercarial asexual stage of *A. winterbourni* develops in the gonad of the snail, which is consequently castrated. The adult worm stage occurs in the gut of waterfowl, where the worms reproduce sexually, producing eggs released with the waterfowl faeces (Lively & McKenzie, 1991). *A. winterbourni* notably lacks several life cycle stages known to occur in other digenean trematodes, including sporocyst, redia, cercaria, and possibly miracidia stages (Figure 1). Unlike some other well-studied digenean trematodes (see Figure 1 and Supplementary Box 1), *A. winterbourni* is not known to infect humans and has low virulence in its final bird host. The *Potamopyrgus-Atriophallophorus* system has been studied intensively because the parasite seems to be in a tight coevolutionary relationship with its host in natural populations (Lively et al., 2004). The host-parasite interaction has been used to test alternative explanations for the

maintenance of sex in *Potamopyrgus* snails (Lively, 1987, 1989). Previous field and laboratory studies suggest that *A. winterbourni* adaptation to local host populations is genotype-specific to a degree that the parasite population can adapt to specifically infect the most common host genotypes, which creates negative frequency-dependent dynamics between the two (Dybdahl & Lively, 1996; Jokela et al., 2009; Lively et al., 2004). Additionally, recent experimental evidence has indicated that the parasite alters the behaviour of the snail, causing it to migrate to the shallow parts of the lake where the final host resides (Feijen et al. in prep).

In this study, we assembled *de novo* the *A. winterbourni* reference genome, annotated protein-coding genes, and assigned putative functions using Gene Ontology. With the knowledge from previous studies of pathways and gene families potentially important in trematode adaptation to parasitism, we used comparative genomics to contrast *A. winterbourni* with other trematodes. We studied the evolution of homologous gene families across the phylogeny of platyhelminths using Hierarchical Orthologous Groups (HOGs), or sets of orthologs/paralogs which all originate from a single gene in the last common ancestor of a clade of interest (Altenhoff et al., 2013). By tracing HOGs along the species tree, it is possible to infer the evolutionary history of gene loss, gain, and duplications since the ancestral gene. Using HOGs, we reconstructed the ancestral digenean trematode genome, the Plagiorchiida ancestral genome, and the ancestral genome before the split of Xiphidiata and Opisthorchiata suborders. Using these ancestral genomes, we identified the evolutionary events (duplications, gains, and retention of 1:1 orthologs) that shaped each gene family in the lineage leading to *A. winterbourni.* We characterized the duplicated, gained, and 1:1 orthologs (i.e. conserved/retained) genes shared between all trematodes, as well as those specific to *A. winterbourni.* We discuss the relevance and function of these gene families in *A. winterbourni* and search for signatures of positive selection in two of the largest gene families. We use the inferred changes in the gene content to better understand the genetic novelty necessary for adaptation to parasitic lifestyles in the lineage leading to *A. winterbourni.* Through

outlining candidate genes for parasitism, we provide a basis for future studies on the genomics of parasite-host coevolution and we broaden the knowledge on trematode evolutionary history.

Methods

Parasite collection and DNA extraction

P. antipodarum snails infected with *A. winterbourni* were collected from Lake Alexandrina (New Zealand, South Island) in January 2017 from several shallow localities (< 1.5 m) by pushing a kicknet through the vegetation. The snails were transported to the Swiss Federal Institute of Aquatic Science (Eawag, Dübendorf, Switzerland) within two weeks of collections and were kept in boxes of 500 snails in a flow-through system that filtered the water every 12 h. Snails were fed spirulina *ad libitum* (*Arthospira platensis,* Spirulina California, Earthrise) once a day.

Infected snails were individually dissected and 200-1000 *A. winterbourni* metacercariae were isolated under 10x-20x magnification. The metacercariae were hatched into adult worms (see Supplementary Methods 1 for details). Obtaining adult worms was necessary to separate the parasite from the double-walled metacercarial cyst that contained both the parasite and the snail DNA (Galaktionov & Dobrovolskij, 2003). The worms were lysed using a CTAB buffer and Proteinase K (2mg/ml) with overnight incubation at 55°C (Yap & Thompson, 1987). DNA was isolated using a chloroform: isoamyl alcohol solution (24:1) and precipitated with sodium acetate (3M). The resulting pellet was washed twice with 70% ethanol. DNA was stored in RNase/DNase-free water (Sigma-Aldrich, Missouri, United States) at -20°C until sequencing library preparation.

Estimation of genome size

To guide the *de novo* genome assembly, genome size was estimated using flow cytometry with Propidium Iodide staining (CyFlow Space, Sysmex). *A. winterbourni* worms were hatched according to the above described protocol. A pool of 15 worms was stained for 1 h with Propidium Iodide (according to the Partec protocol of CyStain PI Absolute T kit) and treated with DNase-free RNase. Three batches of 15 worms were measured, each taken from a different snail host. The DNA content of 2C nuclei was calculated using heads of isoline *Drosophila melanogaster* males and a laboratory clone of *Daphnia galeata* as two independent standards. For the haploid DNA content of *Drosophila melanogaster*, a value of 175 Mb (Bennett et al., 2003) was used and for *Daphnia galeata* a value of 158 Mb (S. Dennis, personal communication, December 12, 2019). Each standard was run separately with each batch of worms.

Sequencing

The DNA of *A. winterbourni* was sequenced using Illumina and Pacific Biosciences technologies (Ambardar et al., 2016). For Illumina sequencing, two infected snails were selected from a shallow water habitat from one site sampled at Lake Alexandrina. A total of 200 ng DNA was extracted from approximately 800-1000 worms and was sent to the Functional Genomics Center Zurich (University of Zurich, Zurich) for library preparation and paired end sequencing using the Illumina HiSeq4000 sequencing platform. A single TruSeq library was constructed from the DNA using the TruSeq Nano DNA library prep kit according to Illumina protocols, obtaining an average of 500 bp insert size. The library was sequenced without indexing on a single Illumina lane. For Pacific Biosciences sequencing, we selected 33 infected snails from two different sites from a shallow water habitat with a high infection prevalence within Lake Alexandrina. We assumed no distinct or significant population structure for the parasite from different sites within

the same habitat zone, as previously shown for the snail host (Paczesniak et al., 2013). Genomic material was isolated from a pool of approximately 13,000-30,000 worms. The high molecular weight DNA with an average length of 45,000 bp (assessed with a Bioanalyser) was sent for sequencing to the Functional Genomics Center Zurich (University of Zurich, Zurich), where it was sequenced with the Pacific Biosciences RSII sequencing platform. A 10 kb SMRT-bell library was constructed from a total of 10 µg of DNA. The library was sequenced using 3 SMRT-cells using P6/C4 chemistry. Primary filtering was performed by Functional Genomics using the SMRT Link software from Pacific Biosciences. We performed secondary filtering, choosing only reads of at least 1000 bp in length and with read quality > 80%. No error correction was performed on the PacBio data at this stage, as it was corrected later with the Illumina data during the hybrid assembly.

Illumina data correction

A quality trimming step was performed with Trimmomatic 0.35 on the raw Illumina HiSeq data before proceeding with the assembly. Adapter sequences were removed and bases with a phred quality score below 5 were removed from the start and the end of the reads. Reads were scanned with a sliding window of 4 and were clipped if the average quality per base dropped below 15. Reads shorter than 50 bp were discarded. The reads were then submitted to PRINSEQ (Schmieder & Edwards, 2011) for filtering for ambiguous bases (Ns), characters different than A, C, G, T or N, and for removal of exact duplicates. For assessment of contamination, we used taxonomic interrogation of the paired reads with Kraken v2, standard database (Wood & Salzberg, 2014).

Hybrid assembly and annotation

Paired reads from Illumina were used together with long reads from Pacific Biosciences for a hybrid assembly with the MaSuRCA 3.2.3 assembler using default parameters (Zimin et al., 2013). Redundans 0.13c (Pryszcz & Gabaldón, 2016) and AGOUTI (Zhang et al., 2016) were used for improvements. Redundans improves the quality of the assembly by reduction, scaffolding and gap closing (Pryszcz & Gabaldón, 2016). The reduction steps consist of identification and removal of heterozygous contigs, based on pairwise sequence similarity searches. Heterozygous contigs are expected to have high sequence identity (Pryszcz & Gabaldón, 2016). The quality was assessed using the N50 statistic, BUSCO 3.0.2 (Benchmarking Universal Single Copy Orthologs) (Waterhouse et al., 2017), and Blobtools 0.9.19.5 (Laetsch & Blaxter, 2017). BUSCO 3.0.2 assesses the completeness of single copy orthologs based on evolutionary-informed expectations about gene content using the lineage dataset metazoa_odb9. Blobtools 0.9.19.5 was used for taxonomic partitioning of the assembly. All scaffolds >50,000 bp (2718 scaffolds) plus a random sample of scaffolds <50,000 bp from the assembly (2661 scaffolds) were submitted to BLAST 2.3.0 using the NCBI nr database for taxonomic annotation. Taxonomic assessment of those scaffolds was used as input for Blobtools**.** The paired and filtered Illumina reads and PacBio reads of at least 1000 bp in length and with read quality > 80% were mapped back to the final assembly with BWA-MEM 0.7.17, yielding an average of 143x coverage per base (125x from the Illumina reads and 18x from the PacBio reads).

Genome annotation was performed using the Maker 2.31.9 annotation pipeline (for details see Supplementary Methods 2) (Cantarel et al., 2008). The completeness and quality of the annotation was assessed with BUSCO and with full-length transcript analysis using BLAST+ (see Supplementary Methods 3). Gene Ontology annotation of the coding sequences was performed with Pannzer2 (Törönen, Medlar, & Holm, 2018), EggNOG (Diamond mapping mode) (Huerta-

Cepas et al., 2016) and OMA ("Orthologous MAtrix") (Altenhoff et al., 2017) web browsers with each dataset used separately for GO enrichment analyses (http://ekhidna2.biocenter.helsinki.fi/sanspanz/ (last accessed: 10.2019), http://eggnogdb.embl.de/#/app/home (last accessed: 01.2020), https://omabrowser.org/oma/functions/ (last accessed: 12.2019)). We also assessed the percentage of all GO terms annotated in *A. winterbourni* with experimental evidence in nematode or trematode (Supplementary Methods 4).

Comparative genomics and ancestral genome reconstruction

We selected 20 species of platyhelminthes and nematodes for comparative genomic analysis. The choice of both nematodes and trematodes was based on their comparisons in other helminth genomic analyses (International Helminth Genomes, Consortium, 2019; Zarowiecki & Berriman, 2015) and will allow for future comparison of trematodes to model species of nematodes. The species consisted of: 14 digenean trematodes (including our focal species), 3 species of parasitic cestodes, 1 species of parasitic monogeneans, and 2 species of free living nematodes (see Supplementary Box 1, Figure 2). We chose these species on the basis of close relatedness to *A. winterbourni* and quality of the genome assembly and annotation (species also used in (International Helminth Genomes, Consortium, 2019)). The proteomic, genomic and transcriptomic sequences for analysis were obtained from the NCBI database of invertebrate genomes (ftp.ncbi.nlm.nih.gov) and from the EBI database (ftp://ftp.ebi.ac.uk/). For the analysis, we used the most recent genomes from those databases with available transcriptomic data (CDS_genomic) and protein annotation (see Supplementary Table 1).

The OMA standalone (Orthologous Matrix) software was used for inference of Hierarchical Orthologous Groups (HOGs) of genes shared between species (Altenhoff et al., 2019). This software conducts an all-against-all comparison to identify the evolutionary relationships between

all pairs of proteins included in the custom-made database of the 20 genomes. The program was run with default parameters and with the “bottom-up” algorithm for inference of HOGs. *C. elegans* and *P. pacificus* were specified as outgroup species. After obtaining the phylogenetic species tree (see next section), OMA was rerun with the precise species tree specified.

The data obtained from OMA was then analyzed with the python library pyHam (Train et al., 2018). With pyHam we reconstructed a model of the ancestral genomes at each stage of the phylogeny leading to *A. winterbourni* and carried out all comparisons between ancestral and extant genomes to obtain classes of duplicated, gained, retained, or lost genes (see Jupyter notebook Supplementary Material 6). We also used pyHam to visualise genomic changes along each branch of the phylogenetic tree.

Phylogenetic species tree

OMA Groups, i.e. Orthologous Groups, from the OMA output were used for phylogenetic tree construction, as they are stringent groups of orthologs and do not contain paralogs (Zahn-Zabal, Dessimoz, Glover, 2020). The phylogenetic tree was constructed following the protocol of Dylus et al. 2020 (Dylus et al., 2020). Briefly, Orthologous Groups containing at least 15 species of monogeneans, cestodes and trematodes were extracted using the custom script filter_groups.py from the git repository https://github.com/DessimozLab/f1000_PhylogeneticTree. Nematodes were excluded from precise phylogenetic and time tree reconstruction, as they are too evolutionarily distant. Within each Orthologous Group, sequences were aligned using MAFFT (mafft 7.273, 1000 cycles of iterative refinement) (Katoh et al., 2009). The separate alignments were concatenated into one supermatrix using a custom script concat_alignment.py from the git repository https://github.com/DessimozLab/f1000_PhylogeneticTree. The final size of the supermatrix was 145,802 sites for all 18 species. No columns from the supermatrix were excluded. The supermatrix was used as input for IQ-TREE maximum likelihood phylogenetic tree

construction (Hoang et al., 2017; Kalyaanamoorthy et al., 2017; Trifinopoulos et al., 2016) using the ModelFinder Plus option for finding the best fitting model. Branch support was calculated with 1000 Ultrafast bootstrap alignments and 1000 iterations. The maximum likelihood tree was confirmed with ASTRAL III (Zhang et al. 2018) by constructing a species tree from gene trees of the 238 Orthologous Groups. Each gene tree was first constructed with IQ-TREE using ModelFinder Plus for choosing an appropriate model; branch support was calculated with 1000 bootstrap alignments and 1000 iterations. The IQ-TREE tree, together with the supermatrix, were used in Mega-X 6.06 for time tree reconstruction using the Maximum Likelihood RelTime method (Tamura et al., 2013). We used two pieces of evidence for time calibration, discussed in the Results.

Gene Ontology Enrichment Analysis

We performed Gene Ontology (GO) annotation for each species using Pannzer2, EggNOG (Diamond mapping mode) and OMA (Törönen et al., 2018; Huerta-Cepas et al., 2016; Altenhoff et al., 2017). Each extant species genome was functionally annotated with orthology-informed putative functions using OMA, Pannzer2 and EggNOG reaching between 26% to 96% of genes annotated for each species (Supplementary Table 2). We then performed GO enrichment analysis using GOATOOLS (Klopfenstein et al., 2018), which finds statistically over- and under-represented GO terms in the set of genes of interest compared to all the GO terms in the background population. For analyses that were species-specific, the background set was all the genes in the genome. For analyses of ancestral genomes, the background population was all the ancestral genes, i.e. the set of HOGs at that taxonomic level. To get the GO terms for any particular ancestral gene/HOG, we took the union of all the GO terms in the extant “children” species. Fisher’s exact test was used for computing uncorrected p-values. The p-values were then corrected using the Bonferroni method and retained if the corrected p-value was <0.05.

Subsequently, all enriched GO terms were categorized into GO slim categories using the AGR subset (Alliance of Genome Resources, http://geneontology.org/docs/download-ontology/, last accessed: 7.05.2020) and unique genes within each enriched GO slim category were counted. For each GO term, the IC (Information Content) score was calculated as: IC(t)=−log(p(t)) with p(t) being estimated as the empirical frequency of the term in the UniProt-GOA database (Barrell et al., 2009). The average IC was calculated for each GO slim term using the IC values of all enriched GO terms in each category (Mazandu & Mulder, 2014; Mistry & Pavlidis, 2008). GO slim terms were used in summarizing the data.

Estimation of dN/dS in gene families in *Atriophallophorus winterbourni*

HOGs 25969 and 36190 with over 30 *A. winterbourni* genes were investigated for signatures of positive selection. All proteins within the two families were submitted to NCBI BLASTP to find their best hit against the nr database and obtain putative functions. We then applied the protocol from Jeffares et al. (2015) to estimate the non-synonymous to synonymous substitution rate ratio within each HOG and to investigate whether selection models explain the data better than null models (Yang, 1997; Kohlhase, 2006). Protein sequences were aligned using Clustal Omega (Madeira et al., 2019), then converted to codon alignment in Phylip format with PAL2NAL (Suyama et al., 2006). Positive selection analyses are sensitive to alignment errors; thus the gap-ridden alignment of HOG 36190 was subjected to a more stringent alignment filtering, guided by the approach proposed by (Moretti et al., 2014) (for details see Supplementary Methods 5). Branch site models in codeml were used to estimate dN, dS and ω (dN/dS) (model=2, NSsites=2). The likelihood ratio test (LRT) was used to determine significance. Gene trees were constructed with protein sequence alignments using IQ-TREE (Hoang et al., 2017; Kalyaanamoorthy et al., 2017; Trifinopoulos et al., 2016). First an initial parsimony tree was

created by a phylogenetic likelihood library; 168 protein models were then tested for best fit with the data according to the Bayesian Information Criterion. Branch support was calculated with 1000 bootstrap alignments (ultrafast bootstrap) and 1000 iterations. The models chosen were JTT+F+G4 for HOG 25969 (General matrix with empirical amino acid frequencies from the data and discrete Gamma model with 4 categories) and WAG+G4 for HOG 36190 (General matrix with discrete Gamma model with 4 categories).

Results and Discussion

Genome of *A. winterbourni*

The *de novo* sequenced genome of *A. winterbourni*, resulted in a final assembly of 601.7 Mb in size, consisting of 26,114 scaffolds with an N50 of 40,108 (see Table 1 and Supplementary Results 1 for details). The assembly size was similar to the flow cytometry-based genome size estimate of 550-600 Mb (Supplementary Figure 1). The annotation yielded 11,499 predicted protein-coding loci spanning 163.7 Mb, with a mean of 5.8 exons and a median of 4 exons per gene (Table 1). The final BUSCO gene set completeness for the annotation was 72% of complete single copy conserved orthologs (see Supplementary Results 3 for protein coding sequence length analysis using BLAST+). Relative to other published trematode genomes, the *A. winterbourni* genome showed good protein sequence length distribution and a comparable BUSCO complete single copy conserved orthologs (Supplementary Figure 4, Table 1). Functional annotation via Gene Ontology (GO) was successful for 84% of genes using OMA, Pannzer2, and EggNOG (9674 genes, see Supplementary Figure 5 and Supplementary Table 2), with 45.3% of the OMA GO terms and 32% of Pannzer2 GO terms assigned to *A. winterbourni* having experimental evidence in nematodes or trematodes (see Supplementary Table 2 and Supplementary Results 2). In comparison to other Plagiorchiida genomes, the *A. winterbourni*

assembly was of similar size and showed similar percentages of non-coding regions, suggesting that no significant genome reduction has occurred in this species (Table 1). Transposable elements, interspersed repeats and low complexity DNA comprised 51.7% of the genome (Supplementary Table 11). This elevated level of TE content in comparison to closely related Opisthorchiata species (33% *C. sinensis*, 30.3% *O. felineus*, 30.9% *O. viverrini*) (Esch et al., 2002) might be an indication of increased importance of transposable elements in *A. winterbourni* genome evolution.

Species phylogeny and molecular clock

To reconstruct a robust maximum likelihood phylogenetic tree, 238 Orthologous Groups (groups containing only orthologs, with a maximum one gene per species) shared between at least 15 out of 18 species of Platyhelminths were used. The phylogenetic estimate was well resolved and congruent with previous publications based on genetic markers or whole genomes (Figure 2) (Blasco-Costa et al., 2020; Galaktionov & Dobrovolskij, 2003; International Helminth Genomes, Consortium, 2019; Lee et al., 2013). *A. winterbourni* was placed as sister to the Ophisthorchiata clade with 100% bootstrap support. The time of speciation of *A. winterbourni* from the Opisthorchiata species was estimated to have been 237.4 MYA (± 120.4 MY), *i.e.* during the Carboniferous through the Cretaceous period (Figure 2). The divergence time estimates across the phylogeny were inferred using several independent pieces of evidence, used as calibration points for Time Tree: the existence time of the proto-trematode first associated with a molluscan host around 400 MYA, and the origin of Schistosoma species in the Creataceous period (66-145MYA) (Blair et al., 2001; Gibson, 1987; Hausdorf, 2000; Parfrey et al., 2011; Peterson et al., 2004).

Evolutionary patterns of gene 1:1 orthology, gain, loss and duplication across Trematoda

The OMA analysis identified 38,144 HOGs among all the species included (2 Nematodes and 18 Platyhelminthes). Specifically, in *A. winterbourni* 5,815 gene families were found (comprising 7,828 out of a total of 11,499 genes, 68.1%) with the rest being identified by OMA as singletons not belonging to any family (3671 genes, 31.9%). Comparisons of three ancestral genomes among the trematode phylogeny (the ancestral Trematoda, the ancestral Plagiorchiida and the Opisthorchiata/Xiphidiata ancestor) revealed many duplicated and gained gene families (Figure 3, Supplementary Figure 7). A particularly high proportion of genomic novelty was inferred during the initial speciation of Trematoda from the Trematoda/Cestoda common ancestor (37.2% of newly acquired genes), and again during the divergence of *A. winterbourni* from the most recent Opisthorchiata/Xiphidiata ancestor (31.9% of newly acquired genes, Figure 3B). The proportion of duplicated genes in the *A. winterbourni* genome was also high (24%) when compared to the Trematoda/Cestoda split (10.9%) (Figure 3B). In *A. winterbourni,* many of the duplicated genes were found in expanded gene families (503 genes comprising 66 HOGs with a minimum of 5 duplications per HOG) and 13 of these HOGs were massively expanded, with over 10 duplicated genes since the Opisthorchiata/Xiphidiata speciation (Supplementary Table 4).

We found only 660 genes lost in the ancestral Trematode from the previous ancestor. We observed a progressive increase in the number of lost genes to the Opisthorchiata/Xiphidiata ancestor (Figure 3A). The Plagiorchiida ancestor exhibited comparable gene loss to gene gain and duplication whereas in the Opisthorchiata/Xiphidiata ancestor, gene loss exceeded the number of duplications or gains (Figure 3A).

1:1 orthologs in trematodes

Based on previous studies, we assumed that many of the genes that remain conserved throughout speciation are housekeeping genes, the building blocks of the organism, and necessary for life, growth, and reproduction (Wu et al. 2006, Duarte et al. 2010). The prediction was confirmed through the GO annotations associated with the genes retained at a 1:1 orthologous gene ratio for each of the ancestral genomes (Supplementary Table 5). The enriched GO terms for retained genes over all ancestors and *A. winterbourni* can be summarized as: RNA processing, the establishment of protein localization, organelle organization, embryo development, cellular catabolism, developmental process, reproduction, and response to stress and stimulus. What is more, since the ancestral trematode species 400 MYA, the number of genes retained at a 1:1 ratio remained relatively constant for each of the 14 extant trematodes, between 2966-5203 genes (Supplementary Table 6).

Additionally, we found 28 single-copy orthologs present in all species, which have been maintained since the trematode ancestor. Examination of their functions through the annotations of best studied trematodes (*Fasciola hepatica* (NCBI, 2017), *Schistosoma mansoni*, (Protasio et al., 2012; Wang et al., 2016) revealed that the 28 retained gene families shared between them all were largely involved in cell functioning and growth, division, and cell-to-cell or protein-to-protein interactions (Supplementary Results 4, Supplementary Table 7).

Genes duplicated and gained in trematodes

We hypothesized that the duplicated genes are more likely to be adaptive than the single-copy orthologs due to the redundant second copy being functionally maintained through positive selection to play a new or same role within the organism (Ohno, 2013; Yang et al., 2015). Multiple duplications within gene families would further suggest an adaptive importance of these key

HOGs. The novel (gained) genes may similarly indicate areas of genetic innovation that were crucial in adoption of new hosts, expansion/streamlining of life cycles and adaptation to changing environments. The origins of the gained genes may stem from neofunctionalization or high divergence of duplicated genes, therefore also potentially involved in adaptive functions as suggested by the gene duplication model of Ohno (2013).

Examination of the enriched functions from the trematode ancestor to the most recent ancestor of *A. winterbourni* are presented in Figure 3C and appear to indicate a progressive gain and duplication of potentially adaptive genes. An “ancestral GO enrichment” analysis of the ancestral genomes was used to retrieve the putative functions of all gained genes (shared between at least 70% of the extant species in Trematoda, 66.7% of Plagiorchiida, or 50% of Xiphidiata/Opisthorchiata) and duplicated genes (minimum 5 duplicated genes per family) (Supplementary Methods 6). Here, we concentrate on functional analysis of ancestral genomes because the inferred gene duplications, gains, and losses are based on evidence present in all of the extant genomes. For example, a gene is inferred to be gained at a particular ancestral level if it is present in at least two species only in that clade. Therefore, ancestral genomes (i.e. internal nodes in the species tree) are more robust than extant genomes in terms of inferred evolutionary duplications, gains, and losses. Additionally, by only considering gained genes present in the majority of the extant species of a given clade, or duplicated genes present with at least five copies, we have more confidence that we are looking at *bona fide* gains and duplications. The ancestral genome annotation was based on combining the GO terms assigned to the extant genomes. We further categorized the enriched GO terms into GO slim categories to give a

broader overview of the functions, and counted unique genes within each of those categories (summarised in Figure 3C). Although there was a similar number of enriched functions for the duplicated and gained genes in the Trematoda and Plagiorchiida ancestors, we found more functions enriched in duplicated than in gained genes in the Xiphidiata/Opisthorchiata ancestor and in *A. winterbourni*. Considering only the duplicated genes, from the trematode ancestral genome to the *A. winterbourni* genome, there was a progressive increase in the number of enriched GO slim functions over time, and an overall increase in the number of unique genes contributing to each function. The increase in the number of unique genes could possibly reflect the increasing importance of this function over time or increased duplication rate of certain families.

We present the average Information Content (IC) per GO slim category, which can be used as a proxy to estimate the specificity of a particular GO term (see methods). The higher the IC, the more specific a term. For the gained genes, we found a progressive increase in IC value of the different GO slim categories but we did not find an increase in the number of enriched GO slim functions or the number of unique genes within them (Figure 3C). The increase in average IC values of GO slim categories enriched for gained genes could suggest an increase in specificity of functions over time (Figure 3C). These observations are best illustrated with enriched GO slim functions such as catalytic activity (GO:0003824), including microtubule motor activity, but also cellular component organization (GO:0005634), including actin bundle filament organization and response to stimulus. A literature review relates them to the importance of the microtubule-based and actin-based cytoskeletal system building the outer body layering (tegument), through which

the parasite interacts with the host environment. Microtubule associated proteins in the tegument, including tubulin, paramyosin, actin, dynein light chains and various antiporters, participate in absorption and secretion (e.g. nitrogen utilization), transport of vesicles from sub-tegumental cells to the tegument cytoplasm, and cell motility (Githui et al., 2009; Young et al., 2010). Molecular characterization and immunostaining studies have also shown dynein light chains to function as tegument associated antigens (Hoffmann & Strand, 1997; Jones et al., 2004; Yang et al., 1999), important in hiding from host immunity. The tegument has been shown to be an essential structure for adaptation to the external environment (Kim et al., 2012) including the pH of the digestive system of the hosts. Indeed, our results show dynein light chain, tegument-associated antigen, and a tubulin-beta chain to be the functions of 3 of the 12 HOGs duplicated since the Trematode ancestor and with at least 3 copies in 75% of the extant species (Supplementary Table 8). We also found dynein light chain to be the putative function of one of the most duplicated HOGs in *A. winterbourni* (Supplementary Table 4, see next section), as well as a HOG duplicated in all 14 trematode species (Supplementary Table 9). Thus, we speculate the functions related to the tegument to be also of great importance in our focal species.

The results might indicate acquisition of more complex and specific adaptations to hosts and environments over time. More experimentally-validated GO annotations in our species of interest could shed light on this hypothesis in the future.

Gene loss in trematodes

Gene loss is known to be common for intracellular parasites (Sakharkar et al. 2004; Corradi 2015) and it is much rarer in parasites with complex life cycles and multiple hosts

(Zarowiecki and Berriman 2015). However, in several helminths there has been a loss of a mitochondrial gene *atp8* (Egger et al. 2017) or cytochrome P450 redox enzymes (Tsai et al. 2013) as well as other functional losses and gene family contractions (International Helminth Genomes, Consortium, 2019). Here, we again focused on ancestral genomes because they are inferred by the accumulation of gene presence and absence information from the extant genomes, *i.e.* if a gene is not found in all the extant genomes of a clade, we can assume it was lost in the last common ancestor of that clade. Thus, ancestral genome analysis is less prone to being undermined by poorer quality genomes (Deutekom et al., 2019). In our study, the robustness was exhibited by the number of losses being always much lower in ancestral than extant genomes (Supplementary Fig 7). We also performed a GO enrichment of the lost genes for *A. winterbourni* as well as the ancestors leading to it. For the ancestral genomes, the background population for GO enrichment was the union of all the GO terms in the extant children species constituting the previous ancestor to the ancestor of interest.

Although there was a progressive increase in the number of genes lost from Trematoda to Opisthorchiata/Xiphidiata ancestor, a GO enrichment analysis of lost genes did not reveal any functions to be enriched in the Trematoda or Opisthorchiata/Xiphidiata ancestor. In the Plagiorchiida ancestor we found loss of genes related to intrinsic components of membrane (GO:0016021) and wide pore channel activity (GO:0022829). We did not find any enrichment of GO terms for the lost genes in *A. winterbourni*. Since the functions of the lost genes appear to not be related to any specific biological processes, we speculate that there is a greater importance of gene gains and duplications in adaptation to parasitism.

Role of gene duplication and gain in driving adaptation of *A. winterbourni*

The Opisthorchiata species exhibit a high similarity in life cycle traits and set of hosts. The *A. winterbourni* genome exhibited comparable proportions of gained, retained and duplicated genes since the Opisthorchiata/Xiphidiata ancestor (31.9%, 44.1%, 24% respectively) as *Opisthorchis viverrini* (41%, 50.5%, 8.5% respectively), *i.e.* in both species the highest proportion of genes was retained and the smallest proportion of genes was duplicated. On the other hand, *Opisthorchis felineus* exhibited a much higher proportion of genes originating through duplication since Opisthorchiata/Xipihidiata ancestor (52.4%) and *Clonorchis sinensis* had the most genes originating through gain since the Opisthorchiata/Xiphidiata ancestor (54.3%). Thus, across the four species, sometimes gene duplication and sometimes gene gain seems to play a greater role in gene family evolution. However, it is important to note that inferences regarding gene duplications, gains, and losses in extant species rather than ancestral species are impacted to a greater extent by fragmentation in genome assemblies, likely inflating the numbers in these categories of genes.

The *A. winterbourni* genome revealed a massive expansion of 13 HOGs that occurred after the speciation from Opisthorchiata/Xiphidiata ancestor (over 10 duplicated genes/HOG, comprising 221 genes, Supplementary Table 4). Comparing *A. winterbourni* to the Opisthorchiata/Xiphidiata ancestor, two gene families stood out due to the presence of more than 30 *A. winterbourni* genes: HOG 25969, with 31 genes in *A. winterbourni* out of 56 genes in all trematodes, and HOG 36190, with 36 genes in *A. winterbourni* out of 72 genes. In these two families, 29 and 31 genes originated through duplication since the Opisthorchiata/Xiphidiata ancestor for HOG 25969 and HOG 36190, respectively. In any other trematode, only 1-5 copies were found. These genes were investigated for being artificially duplicated due to a high

proportion of BUSCO duplicated genes found within the assembly. Genes could be considered artificial duplications due to being fragmented by breaks between scaffolds (Alkan et al., 2011). We looked at the positions of the duplicated genes of HOG 25969 and 36190 on their scaffolds, and we did not find this to be the case (Supplementary Table 10). We thus concluded our genes are likely real duplications rather than artificial duplications due to assembly fragmentation.

Functions of massively expanded gene families in *A. winterbourni*

Examination of GO annotations of the 13 HOGs with over 10 recently duplicated genes (Supplementary Table 4) led us to speculate that the genes are likely involved in host tissue invasion and exploitation (metallohydrolases, Baskaran et al. 2017), escape from host immunity (serpins, Bao et al. 2018), and host behavioural manipulation (glutamine synthase, Helluy et al. 2010) (Supplementary Table 4).

Specifically, we examined the two most highly duplicated gene families in depth. We determined HOG 36190 (36 genes in *A. winterbourni*) to be a gene family of putative glutamine synthases (Supplementary Table 4). Already from the Plagiorchiida ancestor to the Opisthorchiata/Xiphidiata ancestor there was a significant enrichment in biological processes and cellular components related to glutamine family amino acid metabolic processes, including glutamate ammonia ligase activity (GO:0004356), positive regulation of synaptic transmission, glutamatergic (GO:0051968), glutamate binding (GO:0016595), glutamate catabolic process (GO:0006538), and glial cell projection (GO:0097386) (Supplementary Table 5). The glutamine biosynthesis pathway is a pathway in which one of the end products is proline, a non-essential amino acid. An extremely active proline pathway has already been observed in most helminths infecting humans (*Fasciola hepatica*, Schistosomes), with host derived arginine used as a substrate (Ertel & Isseroff, 1976; Mehlhorn, 2016; Toledo & Fried, 2010). These excessive proline levels have been implicated in the pathogenesis of trematode infections. Proline alters antioxidant

defenses, activating secondary metabolite virulence factors, but also provides an energy source for a metabolic shift appropriate for adaptation to the host environment (Ertel & Isseroff, 1976; Mehlhorn, 2016; Toledo & Fried, 2010). Glutamine synthase has also been found to be a marker for glial cells, immunity cells of the central nervous system. A study on *Microphallus papillorobustus*, a trematode parasite of *Gammarus* crustaceans, found disruption of the glutamine metabolism in the brain of the gammarids due to astrocyte-like glia and nitric oxide production by the parasite metacercariae, resulting in altered neuromodulation and behaviour of the host (Helluy & Thomas, 2010). The gene family is thus especially interesting and a potential candidate in parasite-host interactions as previous research has shown *A. winterbourni* to be affecting the behaviour of its snail host (Feijen et al. in prep; Levri & Lively, 1996).

The second highest-duplicated gene family in *A. winterbourni* was HOG 25969, with 31 genes. It consists of proteins putatively encoding for O-sialoglycoprotein endopeptidase, tRNA N6-adenosine threonylcarbamoyltransferase, metallohydrolase and/or glycoprotease/Kae1, all related to DNA repair, protein binding and metal ion binding (Supplementary Table 4). The GO annotation indicates the family to be potentially involved in DNA repair, nuclease activity and nucleic acid phosphodiester bond hydrolysis. Metalloproteases have been found to be duplicated and under positive selection in other parasitic worms (*Strongyloides papillosus*), showing them to be involved in host tissue penetration at final larval stage (Baskaran et al., 2017).

Signatures of selection in two expanded gene families of *A. winterbourni*

We next investigated the two highly duplicated (> 30 genes) HOGs described above for signatures of positive selection. Signatures of positive selection were detected by comparing the dN/dS ratio at branches leading to the radiation of *A. winterbourni* genes, indicated with a #1 on the gene tree, with the dN/dS ratio of background branches (Figure 4, Supplementary Figure 9).

Selection is generally considered negative/purifying if ω (or dN/dS) is less than one, neutral if ω is one, and positive if ω is greater than one.

HOG 36190 was the most massively expanded HOG and selection was found to be acting on some but not all genes within this family. In the dN/dS ratio analysis, the null model (allowing ω ≤ 1) explained the data better than the alternative model (allowing ω > 1) for 2 out of 3 of the investigated branches, indicating neutral evolution (Table 2, Supplementary Figure 9). The signature of selection was detected only on one branch, a long branch leading to a subset of 13 *A. winterbourni* genes within this family (Supplementary Figure 9, branch #1.3). Eleven sites were identified as >50% probability to be under positive selection with one having a probability >90%. From this we conclude that selection might be acting on some, but not all, genes within this family potentially indicating a certain structure evolving under positive selection. However, considering we do not find selection on any other branches in the gene tree, it also has to be taken into account that genes in this family might be highly proliferating due to being in genomic locations prone to duplication events. Their increasing number can be causing redundancy, which can ultimately be deleterious to the organism (Schiffer et al., 2016).

For the gene family HOG 25969 the alternative model (allowing ω > 1) explained the data better than the null model (allowing ω <=1) for all the investigated branches, indicating a signature of positive selection on all of the investigated foreground branches (Table 2, Figure 4). With this result we followed up with the post hoc BEB (Bayes Empirical Bayes) analysis implemented in the alternative model (Ziheng Yang et al., 2005). For the branch leading to all 31 *A. winterbourni* genes, the BEB analysis identified 39 amino acids residues to be under positive selection in the alignment with 4 sites having an over 95% probability of being selected. For sites under positive selection among different subsets of foreground lineages see Table 3. Analysis of positive selection on the structures of the enzyme showed the active, DNA or mental binding site to be under highest probability of selection suggesting an important role (Supplementary Figure 8,

Supplementary Results 5). However, without experimental characterization it is difficult to say what role the family might be playing in *A. winterbourni*.

Conclusions

In our study we report a new *de novo* sequenced genome of a digenean trematode parasite, *Atriophallophorus winterbourni*, its phylogenetic position among other digenean trematodes, and the time of speciation of its ancestor from Opisthorchiata suborder. Using 14 other currently available and well-studied parasitic digenean trematodes we reconstruct the ancestral trematode genome and investigate which genes have originated through duplication, which were gained and which have remained conserved (retained) through each speciation point until the extant genome of *A. winterbourni*. The comparative genomic approach is a powerful tool for identifying candidate duplicated gene families involved in adaptation. We find 13 gene families expanded recently in *A. winterbourni,* and for two we infer signatures of positive selection. Our description of candidate gene families putatively involved in parasite infectivity will facilitate the identification of genomic regions directly involved in the host-parasite coevolutionary arms race and will facilitate studying coadaptation in the laboratory. Gene expression studies in diverse life-cycle stages and functional confirmation via e.g. RNAi knock-out studies will be required to provide a direct link between the genes and phenotypes involved. By focusing on gene duplications and retention across the digenean trematodes our work informs on the genomic basis of adaptation to parasitic lifestyles and paves the way for future adaptation genomics focusing on antagonistic relationships between host and parasites.

Data Availability Statement

Data deposition: The reference genome of *Atriophallophorus winterbourni* has been deposited on NCBI: https://www.ncbi.nlm.nih.gov/nuccore/JACCGJ000000000. It will soon be released with annotation at WormBase Parasite.

Acknowledgments

We thank Frida Feijen for helpful comments on the manuscript and Kirsten Klappert for help in field work and sample collection. We thank Alex Warwick Vesztrocy for data of IC scores calculated for the GO terms in the OMA database. Finally, we thank the reviewers for their constructive comments on this work. The research was funded by an ETH grant ETH-36 15-2 obtained by Jukka Jokela and Hanna Hartikainen. Christophe Dessimoz further acknowledges Swiss National Science Foundation grant #183723. Data produced and analyzed in this paper were generated in collaboration with the Genetic Diversity Centre (GDC), ETH Zurich.

References

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. *Nature Methods*, *8*(1), 61–65.

Altenhoff, A. M., Gil, M., Gonnet, G. H., & Dessimoz, C. (2013). Inferring Hierarchical Orthologous Groups from Orthologous Gene Pairs. *PloS One*, *8*(1), e53786.

Altenhoff, A. M., Glover, N. M., Train, C.-M., Kaleb, K., Warwick Vesztrocy, A., Dylus, D., de Farias, T. M., Zile, K., Stevenson, C., Long, J., Redestig, H., Gonnet, G. H., & Dessimoz, C. (2017). The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Research*, *46*(D1), D477–D485.

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Warwick Vesztrocy, A., Dalquen, D. A., Müller, S., Telford, M. J., Glover, N. M., Dylus, D., & Dessimoz, C. (2019). OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Research*, *29*(7), 1152–1163.

Ambardar, S., Gupta, R., Trakroo, D., Lal, R., & Vakhlu, J. (2016). High Throughput Sequencing: An Overview of Sequencing Chemistry. *Indian Journal of Microbiology*, *56*(4), 394–404.

Bao, J., Pan, G., Poncz, M., Wei, J., Ran, M., & Zhou, Z. (2018). Serpin functions in host-pathogen interactions. *PeerJ*, *6*, e4557.

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., & Apweiler, R. (2009). The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Research*, *37*(Database issue), D396–D403.

Baskaran, P., Jaleta, T. G., Streit, A., & Rödelsperger, C. (2017). Duplications and Positive Selection Drive the Evolution of Parasitism-Associated Gene Families in the Nematode Strongyloides papillosus. *Genome Biology and Evolution*, *9*(3), 790–801.

Bennett, M. D., Leitch, I. J., Price, H. J., & Johnston, J. S. (2003). Comparisons with Caenorhabditis (~100 Mb) and Drosophila (~175 Mb) Using Flow Cytometry Show Genome Size in Arabidopsis to be ~157 Mb and thus ~25 % Larger than the Arabidopsis Genome Initiative Estimate of ~125 Mb. *Annals of Botany*, *91*(5), 547–557.

Blair, D., Davis, G. M., & Wu, B. (2001). Evolutionary relationships between trematodes and snails emphasizing schistosomes and paragonimids. *Parasitology*, *123 Suppl*, S229–S243.

Blasco-Costa, I., Seppälä, K., Feijen, F., Zajac, N., Klappert, K., & Jokela, J. (2020). A new species of Atriophallophorus Deblock & Rosé, 1964 (Trematoda: Microphallidae) described from in vitro-grown adults and metacercariae from Potamopyrgus antipodarum (Gray, 1843) (Mollusca: Tateidae). *Journal of Helminthology*, *94*, e108.

Cantacessi, C., & Gasser, R. B. (2012). SCP/TAPS proteins in helminths – Where to from now? *Molecular and Cellular Probes*, *26*(1), 54–59.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196.

Chang, E. S., Neuhof, M., Rubinstein, N. D., Diamant, A., Philippe, H., Huchon, D., & Cartwright, P. (2015). Genomic insights into the evolutionary origin of Myxozoa within Cnidaria. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(48), 14912–14917.

Corradi, N. Microsporidia: Eukaryotic Intracellular Parasites Shaped by Gene Loss and Horizontal Gene Transfers. *Annu. Rev. Microbiol.* **69**, 167–183 (2015).

David, C. N., Ozbek, S., Adamczyk, P., Meier, S., Pauly, B., Chapman, J., Hwang, J. S., Gojobori, T., & Holstein, T. W. (2008). Evolution of complex structures: minicollagens shape the cnidarian nematocyst. *Trends in Genetics: TIG*, *24*(9), 431–438.

Dennis, S. (2019, December 12). [Personal communication].

Deutekom, E. S., Vosseberg, J., van Dam, T. J. P., & Snel, B. (2019). Measuring the impact of gene prediction on gene loss estimates in Eukaryotes by quantifying falsely inferred absences. PLoS Computational Biology, 15(8), e1007301.

Disease and Injury Incidence and Prevalence Collaborators, GBD. (2016). Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet*, *388*(10053), 1545–1602.

Duarte, J. M., Wall, P. K., Edger, P. P., Landherr, L. L., Ma, H., Pires, P. K., . . . Claude, W. d. (2010). Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryzaand their phylogenetic utility across various taxonomic levels. *BMC Evolutionary Biology, 10*(1), 61.

Dybdahl, M. F., & Lively, C. M. (1996). THE GEOGRAPHY OF COEVOLUTION: COMPARATIVE POPULATION STRUCTURES FOR A SNAIL AND ITS TREMATODE PARASITE. *Evolution; International Journal of Organic Evolution*, *50*(6), 2264–2275.

Dylus D, Nevers Y, Altenhoff AM *et al.* (2020) How to build phylogenetic species trees with OMA [version 1; peer review: awaiting peer review]. *F1000Research* 2020, **9**:511

Egger, B., Bachmann, L. & Fromm, B. Atp8 is in the ground pattern of flatworm mitochondrial genomes. *BMC Genomics* **18**, 414 (2017).

Ertel, J. C., & Isseroff, H. (1976). Proline in fascioliasis: II. Characteristics of partially purified ornithine-δ-transaminase from Fasciola. *Rice Institute Pamphlet-Rice University Studies*, *62*(4).

Esch, G. W., Barger, M. A., & Fellis, K. J. (2002). The Transmission of Digenetic Trematodes: Style, Elegance, Complexity1. *Integrative and Comparative Biology*, *42*(2), 304–312.

Feijen, F., Buser, C. C., Klappert, K., Kopp, K., Lively, C.M., Zajac, N.H., Jokela, J. Hotspots for parasite transmission emerge from large infection source habitats.*in prep.*

Galaktionov, K., & Dobrovolskij, A. (2003). *The Biology and Evolution of Trematodes*.

Gibson, D. I. (1987). Questions in digenean systematics and evolution. *Parasitology*, *95 (Pt 2)*, 429–460.

Githui, E. K., Damian, R. T., Aman, R. A., Ali, M. A., & Kamau, J. M. (2009). Schistosoma spp.: Isolation of microtubule associated proteins in the tegument and the definition of dynein light chains components. *Experimental Parasitology*, *121*(1), 96–104.

Hausdorf, B. (2000). Early evolution of the bilateria. *Systematic Biology*, *49*(1), 130–142.

Helluy, S., & Thomas, F. (2010). Parasitic manipulation and neuroinflammation: Evidence from the system Microphallus papillorobustus (Trematoda) - Gammarus (Crustacea). *Parasites & Vectors*, *3*, 38.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2017). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522.

Hoffmann, K. F., & Strand, M. (1997). Molecular Characterization of a 20.8-kDa Schistosoma mansoni Antigen: Sequence similarity to tegumental associated antigens and dynein light chains. *The Journal of Biological Chemistry*, *272*(23), 14509–14515.

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C., & Bork, P. (2016). eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research*, *44*(D1), D286–D293.

International Helminth Genomes, Consortium. (2019). Comparative genomics of the major parasitic worms. *Nature Genetics*, *51*(1), 163–174.

Jeffares, D. C., Tomiczek, B., Sojo, V., & dos Reis, M. (2015). A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome. In C. Peacock (Ed.), *Parasite Genomics Protocols* (pp. 65–90). Springer New York.

Jokela, J., Dybdahl, M. F., & Lively, C. M. (2009). The maintenance of sex, clonal dynamics, and host-parasite coevolution in a mixed population of sexual and asexual snails. *The American Naturalist*, *174 Suppl 1*(s1), S43–S53.

Jones, M. K., Gobert, G. N., Zhang, L., Sunderland, P., & McManus, D. P. (2004). The cytoskeleton and motor proteins of human schistosomes and their roles in surface maintenance and host-parasite interactions. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *26*(7), 752–765.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589.

Katoh, K., Asimenos, G., & Toh, H. (2009). Multiple Alignment of DNA Sequences with MAFFT. In D. Posada (Ed.), *Bioinformatics for DNA Sequence Analysis* (pp. 39–64). Humana Press.

Kim, Y.-J., Yoo, W. G., Lee, M.-R., Kim, D.-W., Lee, W.-J., Kang, J.-M., Na, B.-K., & Ju, J.-W. (2012). Identification and characterization of a novel 21.6-kDa tegumental protein from Clonorchis sinensis. *Parasitology Research*, *110*(5), 2061–2066.

Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., & Tang, H. (2018). GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports*, *8*(1), 10872.

Kohlhase, M. (2006). CodeML: an open markup format the content and presentatation of program code. Computer Science, Carnegie Mellon University Pittsburgh.

Laetsch, D. R., & Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. *F1000Research*, *6*(1287). https://doi.org/10.12688/f1000research.12232.1

Lee, D., Choe, S., Park, H., Jeon, H.-K., Chai, J.-Y., Sohn, W.-M., Yong, T.-S., Min, D.-Y., Rim, H.-J., & Eom, K. S. (2013). Complete mitochondrial genome of Haplorchis taichui and comparative analysis with other trematodes. *The Korean Journal of Parasitology*, *51*(6), 719.

Levri, E. P., & Lively, C. M. (1996). The effects of size, reproductive condition, and parasitism on foraging behaviour in a freshwater snail, Potamopyrgus antipodarum. *Animal Behaviour*, *51*(4), 891–901.

Lively, C. M. (1987). Evidence from a New-Zealand Snail for the Maintenance of Sex by Parasitism. *Nature*, *328*(6130), 519–521.

Lively, C. M. (1989). Adaptation by a Parasitic Trematode to Local-Populations of Its Snail Host. *Evolution; International Journal of Organic Evolution*, *43*(8), 1663–1671.

Lively, C. M., Dybdahl, M. F., Jokela, J., Osnas, E. E., & Delph, L. F. (2004). Host Sex and Local Adaptation by Parasites in a Snail-Trematode Interaction. *The American Naturalist*, *164*(S5), S6–S18.

Lively, C. M., & McKenzie, J. C. (1991). Experimental infection of a freshwater snail, *Potamopyrgus antipodarum*, with a digenetic trematode, *Microphallus* sp. *New Zealand Natural Sciences*, *18*, 59–62.

Lu, T.-M., Kanda, M., Furuya, H., & Satoh, N. (2019). Dicyemid Mesozoans: A Unique Parasitic Lifestyle and a Reduced Genome. *Genome Biology and Evolution*, *11*(8), 2232–2243.

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D., & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research*, *47*(W1), W636–W641.

Mazandu, G. K., & Mulder, N. J. (2014). Information content-based Gene Ontology functional similarity measures: which one to use for a given biological data type? *PloS One*, *9*(12), e113859.

Mehlhorn, H. (2016). Amino Acids. In H. Mehlhorn (Ed.), *Encyclopedia of Parasitology* (pp. 104–107). Springer Berlin Heidelberg.

Mistry, M., & Pavlidis, P. (2008). Gene Ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, *9*, 327.

Moretti, S., Laurenczy, B., Gharib, W. H., Castella, B., Kuzniar, A., Schabauer, H., Studer, R. A., Valle, M., Salamin, N., Stockinger, H., & Robinson-Rechavi, M. (2014). Selectome update: quality control and computational improvements to a database of positive selection. *Nucleic Acids Research*, *42*(Database issue), D917–D921.

NCBI. (2017, November 9). *Fasciola hepatica genome assembly, unpublished*. NCBI F_hepatica_1.0.allpaths.pg. https://www.ncbi.nlm.nih.gov/assembly/GCA_002763495.1

Ohno, S. (2013). *Evolution by Gene Duplication*. Springer Science & Business Media.

O'Malley, M. A., Wideman, J. G., & Ruiz-Trillo, I. (2016). Losing Complexity: The Role of Simplification in Macroevolution. *Trends in Ecology & Evolution*, *31*(8), 608–621.

Paczesniak, D., Jokela, J., Larkin, K., & Neiman, M. (2013). Discordance between nuclear and mitochondrial genomes in sexual and asexual lineages of the freshwater snail Potamopyrgus antipodarum. *Molecular Ecology*, *22*(18), 4695–4710.

Parfrey, L. W., Lahr, D. J. G., Knoll, A. H., & Katz, L. A. (2011). Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33), 13624–13629.

Peterson, K. J., Lyons, J. B., Nowak, K. S., Takacs, C. M., Wargo, M. J., & McPeek, M. A. (2004). Estimating metazoan divergence times with a molecular clock. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(17), 6536–6541.

Peyretaillade, E., El Alaoui, H., Diogon, M., Polonais, V., Parisot, N., Biron, D. G., Peyret, P., & Delbac, F. (2011). Extreme reduction and compaction of microsporidian genomes. *Research in Microbiology*, *162*(6), 598–606.

Poulin, R., & Randhawa, H. S. (2015). Evolution of parasitism along convergent lines: from ecology to genomics. *Parasitology*, *142 Suppl 1*, S6–S15.

Protasio, A. V., Tsai, I. J., Babbage, A., Nichol, S., Hunt, M., Aslett, M. A., De Silva, N., Velarde, G. S., Anderson, T. J. C., Clark, R. C., Davidson, C., Dillon, G. P., Holroyd, N. E., LoVerde, P. T., Lloyd, C., McQuillan, J., Oliveira, G., Otto, T. D., Parker-Manuel, S. J., … Berriman, M. (2012). A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. *PLoS Neglected Tropical Diseases*, *6*(1), e1455.

Pryszcz, L. P., & Gabaldón, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, *44*(12), e113–e113.

Rödelsperger, C. (2018). Comparative Genomics of Gene Loss and Gain in Caenorhabditis and Other Nematodes. In J. C. Setubal, J. Stoye, & P. F. Stadler (Eds.), *Comparative Genomics: Methods and Protocols* (pp. 419–432). Springer New York.

Roger, E., Mitta, G., Moné, Y., Bouchut, A., Rognon, A., Grunau, C., Boissier, J., Théron, A., & Gourbal, B. E. F. (2008). Molecular determinants of compatibility polymorphism in the Biomphalaria glabrata/Schistosoma mansoni model: New candidates identified by a global comparative proteomics approach. *Molecular and Biochemical Parasitology*, *157*(2), 205–216.

Sakharkar, K. R., Dhar, P. K. & Chow, V. T. K. Genome reduction in prokaryotic obligatory intracellular parasites of humans: a comparative analysis. *Int. J. Syst. Evol. Microbiol.* **54**, 1937–1941 (2004).

Schiffer, P. H., Gravemeyer, J., Rauscher, M., & Wiehe, T. (2016). Ultra Large Gene Families: A Matter of Adaptation or Genomic Parasites? *Life* , *6*(3). https://doi.org/10.3390/life6030032

Schmieder, R., & Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. *Bioinformatics* , *27*(6), 863–864.

Slyusarev, G. S., Starunov, V. V., Bondarenko, A. S., Zorina, N. A., & Bondarenko, N. I. (2020). Extreme Genome and Nervous System Streamlining in the Invertebrate Parasite Intoshia variabili. *Current Biology: CB*, *30*(7), 1292–1298.e3.

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, *34*(suppl_2), W609–W612.

Takeuchi, T., Koyanagi, R., Gyoja, F., Kanda, M., Hisata, K., Fujie, M., Goto, H., Yamasaki, S., Nagai, K., Morino, Y., Miyamoto, H., Endo, K., Endo, H., Nagasawa, H., Kinoshita, S., Asakawa, S., Watabe, S., Satoh, N., & Kawashima, T. (2016). Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. *Zoological Letters*, *2*(1), 3.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., & Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution*, *30*(12), 2725–2729.

Toledo, R., & Fried, B. (2010). *Biomphalaria snails and larval trematodes*. Springer Science & Business Media.

Törönen, P., Medlar, A., & Holm, L. (2018). PANNZER2: a rapid functional annotation web server. *Nucleic Acids Research*, *46*(W1), W84–W88.

Train, C.-M., Pignatelli, M., Altenhoff, A., & Dessimoz, C. (2018). iHam and pyHam: visualizing and processing hierarchical orthologous groups. *Bioinformatics* , *35*(14), 2504–2506.

Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., & Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, *44*(W1), W232–W235.

Wang, T., Zhao, M., Rotgans, B. A., Strong, A., Liang, D., Ni, G., Limpanont, Y., Ramasoota, P., McManus, D. P., & Cummins, S. F. (2016). Proteomic Analysis of the Schistosoma mansoni Miracidium. *PloS One*, *11*(1), e0147247.

Warwick, T. (1952). Strains in the mollusc Potamopyrgus jenkinsi (Smith). *Nature*, *169*, 551–552.

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2017). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548.

Weinstein, S. B., & Kuris, A. M. (2016). Independent origins of parasitism in Animalia. *Biology Letters*, *12*(7). https://doi.org/10.1098/rsbl.2016.0324

Winterbourn, M. J. (1970). Population studies on the New Zealand freshwater gastropod *Potamopyrgus antipodarum* (Gray). *Proceedings of the Malacological Society of London*, *39*, 139–149.

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46.

Wu, F., Mueller, L. A., Crouzillat, D., Pétiard, V., & Tanksley, S. D. (2006). Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics, 174(3), 1407-1420.

Yang, W., Jones, M. K., Fan, J., Hughes-Stamm, S. R., & McManus, D. P. (1999). Characterisation of a family of Schistosoma japonicum proteins related to dynein light chains1The nucleotide sequences reported in this paper have been submitted to the GenBank/EMBL Data Bank with accession numbers AF072327–AF072332.1. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, *1432*(1), 13–26.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences: CABIOS*, *13*(5), 555–556.

Yang, Z., Wafula, E. K., Honaas, L. A., Zhang, H., Das, M., Fernandez-Aparicio, M., Huang, K., Bandaranayake, P. C. G., Wu, B., Der, J. P., Clarke, C. R., Ralph, P. E., Landherr, L., Altman, N. S., Timko, M. P., Yoder, J. I., Westwood, J. H., & dePamphilis, C. W. (2015). Comparative transcriptome analyses reveal core parasitism genes and suggest gene duplication and repurposing as sources of structural novelty. *Molecular Biology and Evolution*, *32*(3), 767–790.

Yang, Z., Wong, W. S. W., & Nielsen, R. (2005). Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, *22*(4), 1107–1118.

Yap, K. W., & Thompson, R. C. (1987). CTAB precipitation of cestode DNA. *Parasitology Today , 3*(7), 220–222.

Young, N. D., Hall, R. S., Jex, A. R., Cantacessi, C., & Gasser, R. B. (2010). Elucidating the transcriptome of Fasciola hepatica—a key to fundamental and biotechnological discoveries for a neglected parasite. *Biotechnology Advances*, *28*(2), 222–231.

Zahn-Zabal, M., Dessimoz, C., & Glover, N. M. (2020). Identifying orthologs with OMA: A primer. F1000Research, 9, 27. https://doi.org/10.12688/f1000research.21508.1

Zarowiecki, M., & Berriman, M. (2015). What helminth genomes have taught us about parasite evolution. *Parasitology*, *142 Suppl 1*, S85–S97.

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(Suppl 6), 153.

Zhang, S. V., Zhuo, L., & Hahn, M. W. (2016). AGOUTI: improving genome assembly and annotation using transcriptome data. *GigaScience*, *5*(1), 31.

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics* , *29*(21), 2669–2677.

Figures

Legends

Figure 1. A summary table showing several shared life cycle characteristics of the trematodes used in the study. The first seven columns indicate the presence (blue) or absence (grey) of developmental stages in each parasite's life cycle. "Host number" indicates the number of hosts in a parasite's life cycle, "Type of adult worm" indicates whether the adult worms in the final host are hermaphroditic or dioecious (both males and females present). Species within the genera *Schistosoma* and *Opisthorchis* are grouped due to identical characteristics. The photographs below represent the metacercaria and adult stage of *A. winterbourni* and the intermediate host of *A. winterbourni* (*P. antipodarum* snail) (photographs taken by N.Zajac and K.Seppälä).

Table 1. Information on the genome assemblies used in the analysis. For more information, see Supplementary Table 1. The BUSCO results refer to the protein annotation. The results on exon/intron number and length were calculated from the gff files with genestats script (available at: https://gist.github.com/darencard/fcb32168c243b92734e85c5f8b59a1c3, date accessed 14.07.2020) or obtained from (Tsai et al. 2013).

Figure 2. Phylogenetic tree and classification of species used in the analysis and classification into the different classes. The data used for the tree were all Orthologous Groups from the OMA analysis with genes from at least 15 species present (238 groups of orthologs). The combined data was used in IQ-TREE to create a robust consensus species tree. The tree and the combined alignment of 238 groups of orthologs was used in Mega-X 6.06 for reconstruction of the time tree. The scale below indicates divergence times in Million Years (MY). Each node has a divergence time with the confidence interval indicated in brackets in million years and a bootstrap support indicated after a slash.

Figure 3.A. Number of duplicated, retained (1:1 orthologs) and gained genes resulting after each point of speciation obtained from the analysis of Hierarchical Orthologous Groups in pyHam, mapped onto a phylogenetic tree of trematodes (for original see Supplementary Figure 6). The total number of genes at each point is indicated on the left-hand side of the bar and the total number of retained (pink), duplicated (green) and gained (yellow) genes are indicated on the right-hand side of the bar. The bars indicate the proportions of genes in each category. The lost genes are indicated only for the three ancestral genomes: the Trematoda ancestor, the Plagiorchiia ancestor and the

Opisthorchiata/Xiphidiata ancestor. **B.** The proportions (on the bars) and the total numbers (next to the bars) of retained (pink), duplicated (green) and gained (yellow) genes in each reconstructed ancestral genome leading to *A. winterbourni*. The oldest ancestral genome is on the left-hand side and the extant *A. winterbourni* genome on the right-hand side. The total number of genes per genome is above each bar beneath the name. **C.** Heatmaps summarising the GO enrichment analysis of the duplicated and gained genes in the 3 reconstructed ancestral genomes and the extant genome of *A. winterbourni.* All enriched GO terms were categorized into GO slims, listed on the y-axis of each heatmap. The colours indicate the mean IC value of each GO slim category and the number printed on top is the number of unique genes within that GO slim category (see Methods).

Figure 4. Gene tree of gene family HOG 25969 created with IQ-TREE. The tree is unrooted. Each name is a species name followed by the original gene name (protein name). *A. winterbourni* gene names are shortened version of gene names in Supplementary Table 10. The numbers above branches indicate ultrafast bootstrap support, for the #1 branches the bootstrap support is after a backslash. The branches labelled with #1.X indicate the separation between the foreground branches and the background branches (distinction used in codeml for investigation of selection). The test for selection compares the dN/dS between the foreground branch and the background branches. The total number of genes in this HOG per trematode species is given next to each species name.

Table 2. Results of studying positive selection in two majorly expanded gene families in *A. winterbourni*. The HOG indicates the ID of the gene family. The node relates to the nodes indicated in the gene trees of each HOG. LRT - results of likelihood ratio test, p-value is the result of chi^2 test of the LRT. Positively selected sites are the result of BEB (Bayes empirical Bayes) test implemented in codeml. The starred values indicate sites under significantly high probability of selection (>95%).

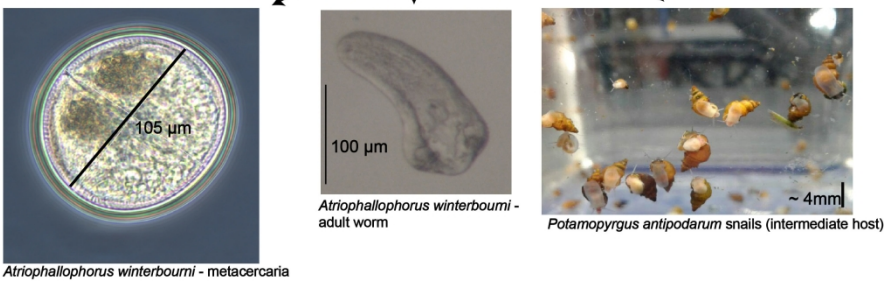
Table 1.

Species	Genome size (Mb)	NB. genes	scaff. count	N50	GC content (%)	Busco complete single (%)	Busco duplicated (%)	Busco fragmented (%)	Busco missing (%)	Total exon number	Average exon length (bp)	Total intron number	Average intron length (bp)	Total coding sequence (Mb)
Atriophallophorus winterbourni	601.7	11499	26114	40108	40.73	56.2	15.7	6.1	22	66672	233	54987	1732	163.7
Taenia solium	122	12467	11237	68000	42.9	77.9	1.9	6.7	13.5	69770	223	57289	574	48.3
Echinococcus granulosus	110.8	11319	957	712683	41.7	76.2	2.1	6.5	15.2	75264	211	63945	722	62
Gyrodactylus salaris	67.4	15436	6075	18400	33.9	67.7	1.5	8.9	21.9	61693	229	46257	584	41.1
Fasciola hepatica	1138	14642	23604	161103	44.1	71	1.2	9.5	18.3	83777	488	72560	4168	343.3
Echinostoma caproni	834.6	18607	86083	27000	42.5	50.3	1.2	26.9	21.6	65273	267	46666	2451	131.8
Opisthorchis felienus	679.25	11427	13306	621022	44.1	53.4	33	4.1	9.5	180879	261	160011	3527	291.9
Opisthorchis viverrini	472.26	13555	16038	79767	44	67	0.7	11.6	20.7	59112	242	48358	2734	146.5
Clonorchis sinensis	562.7	14538	2776	1628761	42.6	72.6	1	6.9	19.5	89304	234	74766	2745	226.2
Trichobilharzia regenti	701.76	22185	188369	7696	37.4	36.4	0.7	35.4	27.5	54402	277	32217	1829	74
Schistosoma japonicum	369.9	11416	1789	1093989	33.8	47.2	34.6	4.3	13.9	130068	336	113132	2372	185.4
Schistosoma mansoni	364.5	10772	885	32115376	35.5	71.5	8.6	6.9	13	70430	204	57138	2475	148.8
Schistosoma margrebowiei	367.4	26189	23355	35236	34.3	65.7	1.6	17.4	15.3	79991	262	53802	1925	122.3
Schistosoma haematobium	375.89	11140	29834	317484	34.2	71.4	1.6	11.7	15.3	64235	246	53148	2488	148.8
Schistosoma bovis	373.4	11576	4774	202989	34.4	68.3	4.3	11.9	15.5	65265	259	53689	2406	146.1
Schistosoma mattheei	340.82	22997	62061	12303	34.1	51.3	1.5	24.1	23.1	65852	263	43672	1569	84.1
Schistosoma curassoni	344.2	23546	60140	13861	34.2	54.6	1.4	19.6	24.4	69606	259	46060	1576	88.5
Caeorhabditis elegans	102.3	20184	7	17493829	35.4	98	0.6	0.8	0.6	285984	239	250855	438	63.3
Echinococcus multilocularis	115	10663	1217	13800000	42.2	79.8	2.8	5	12.4	71022	205	60677	663	49
Pristionchus pacificus	158.5	25991	47	23900000	42.8	91.6	1.3	3.7	22.4	312244	106	287319	275	112.2

Table 2.

HOG	node	LRT	df	p-value	positively selected sites (position in the alignment, amino acid, probability of being under positive selection)			
36190	#1.1	0.00017	1	0.98	-			
36190	#1.2	1.9	1	0.17	-			
36190	#1.3	16.9	1	3.9E-05	291 A 0.725	874 L 0.731	1030 - 0.681	
					367 E 0.910	875 S 0.800		
					370 K 0.745	878 Y 0.564		
					371 K 0.827	879 V 0.518		
					873 K 0.626	880 P 0.707		
25969	#1.1	25.6	1	4.13E-07	603 A 0.767	794 A 0.575	922 I 0.548	998 V 0.504
					607 S 0.664	795 K 0.694	932 N 0.516	1038 Q 0.542
					638 N 0.707	798 I 0.662	951 K 0.683	1073 S 0.605
					649 V 0.935	803 S 0.762	955 H 0.878	1090 S 0.875
					675 S 0.669	804 G 0.556	970 T 0.513	1095 Y 0.951*
					680 C 0.912	812 R 0.893	976 Q 0.846	1121 S 0.508
					701 I 0.684	833 S 0.514	986 N 0.906	1184 R 0.931
					705 K 0.541	871 A 0.695	989 F 0.544	1188 I 0.521
					716 Y 0.852	879 Q 0.889	994 S 0.624	1198 H 0.700
					719 C 0.966*	921 N 0.983*	996 F 0.957*	
25969	#1.2	5.5	1	1.80E-02	366 K 0.537			
					394 W 0.534			
					464 K 0.661			
					473 N 0.832			
					637 T 0.623			
					657 H 0.830			
					662 D 0.692			
					665 S 0.707			
25969	#1.3	4.3	1	3.70E-02	402 R 0.867			
					873 F 0.624			
25969	#1.4	26.7	1	2.30E-07	370 T 0.756	831 N 0.534	1004 N 0.974*	
					394 W 0.559	857 E 0.984*	1029 F 0.676	
					480 R 0.868	928 S 0.766	1044 D 0.846	
					482 L 0.590	929 G 0.986*	1080 E 0.767	
					611 C 0.632	931 N 0.687	1157 N 0.674	
					710 N 0.697	932 N 0.546	1162 Y 0.635	
					714 S 0.733	967 H 0.955*	1184 R 0.922	
					722 M 0.740	970 T 0.811	1189 L 0.886	
					749 P 0.791	988 M 0.940		
					814 E 0.637	994 S 0.729		

	Egg	Miracidium	Sporocyst	Redia	Cercaria	Meta-cercaria	Adult	Free-swimming cercaria	Host number	First intermediate host	Final host	Type of adult worm
Fasciola hepatica	YES	YES	YES	YES	YES	YES	YES	YES	2	Snail sp.	cattle, humans	hermaphrodite
Schistosoma species	YES	YES	YES	YES	YES	YES	YES	YES	2	Snail sp.	humans	dioecious
Trichobilharzia regenti	YES	YES	YES	YES	YES	YES	YES	YES	2	Snail sp.	aquatic birds	dioecious
Echinostoma caproni	YES	YES	YES	YES	YES	YES	YES	YES	3	Snail sp.	aquatic birds	hermaphrodite
Clonorchis sinensis	YES	YES	YES	YES	YES	YES	YES	NO	3	Snail sp.	humans	hermaphrodite
Opisthrochis species	YES	YES	YES	YES	YES	YES	YES	YES	3	Snail sp.	humans	hermaphrodite
Atriophallophorus winterbourni	YES	NO	NO	NO	NO	YES	YES	NA	2	Snail sp.	aquatic birds	hermaphrodite



Atriophallophorus winterbourni - metacercaria

Atriophallophorus winterbourni - adult worm

Potamopyrgus antipodarum snails (intermediate host)

Figure 1. A summary table showing several shared life cycle characteristics of the trematodes used in the study. The first seven columns indicate the presence (blue) or absence (grey) of developmental stages in each parasite's life cycle. "Host number" indicates the number of hosts in a parasite's life cycle, "Type of adult worm" indicates whether the adult worms in the final host are hermaphroditic or dioecious (both males and females present). Species within the genera Schistosoma and Opisthorchis are grouped due to identical characteristics. The photographs below represent the metacercaria and adult stage of *A. winterbourni* and the intermediate host of *A. winterbourni* (*P. antipodarum* snail) (photographs taken by N.Zajac and K.Seppälä).

247x132mm (300 x 300 DPI)

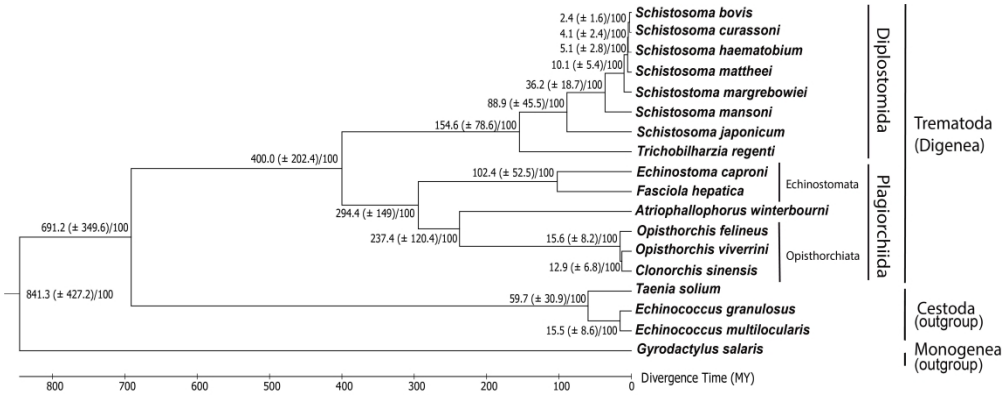
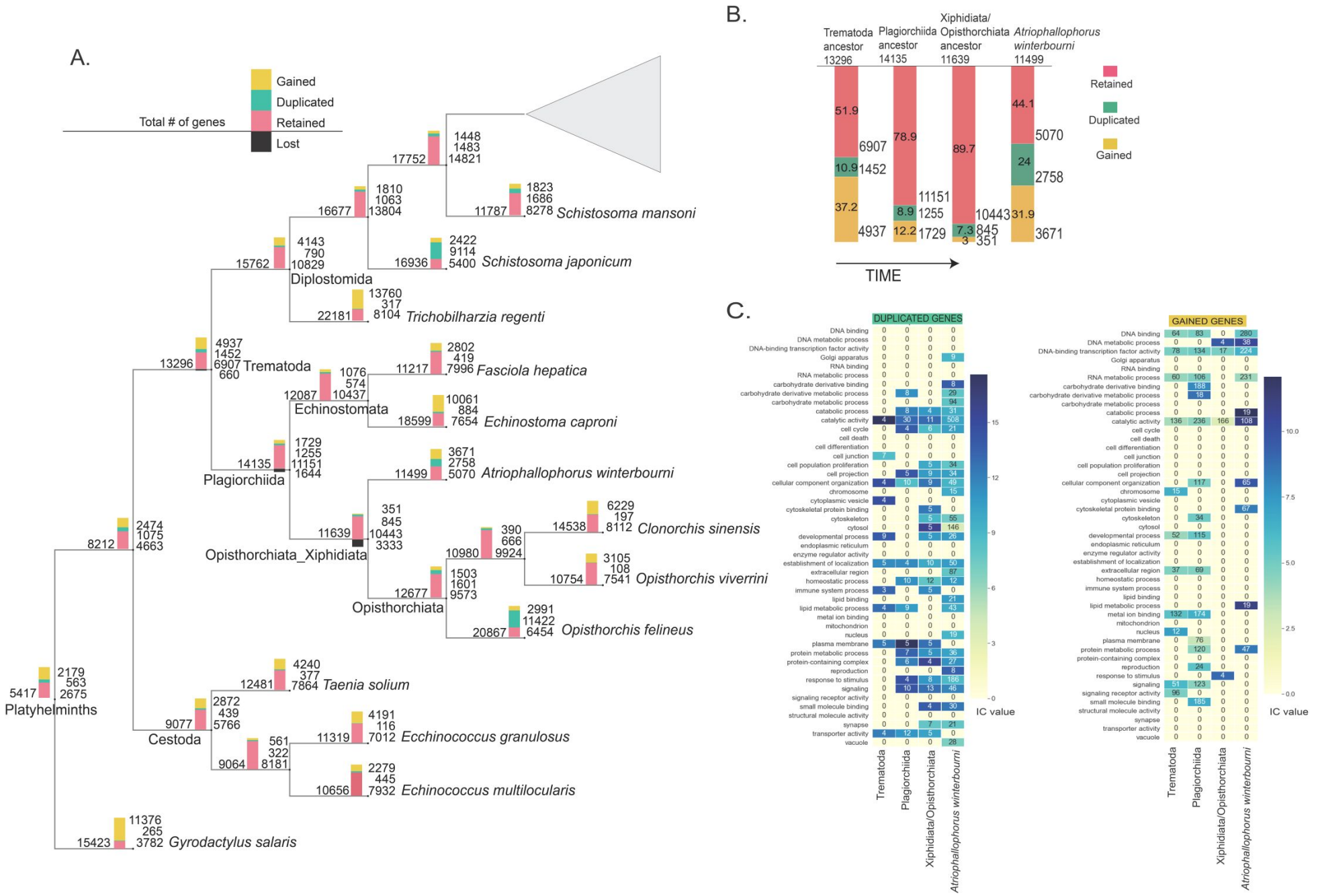


Figure 2. Phylogenetic tree and classification of species used in the analysis and classification into the different classes. The data used for the tree were all Orthologous Groups from the OMA analysis with genes from at least 15 species present (238 groups of orthologs). The combined data was used in IQ-TREE to create a robust consensus species tree. The tree and the combined alignment of 238 groups of orthologs was used in Mega-X 6.06 for reconstruction of the time tree. The scale below indicates divergence times in Million Years (MY). Each node has a divergence time with the confidence interval indicated in brackets in million years and a bootstrap support indicated after a slash.

433x169mm (300 x 300 DPI)



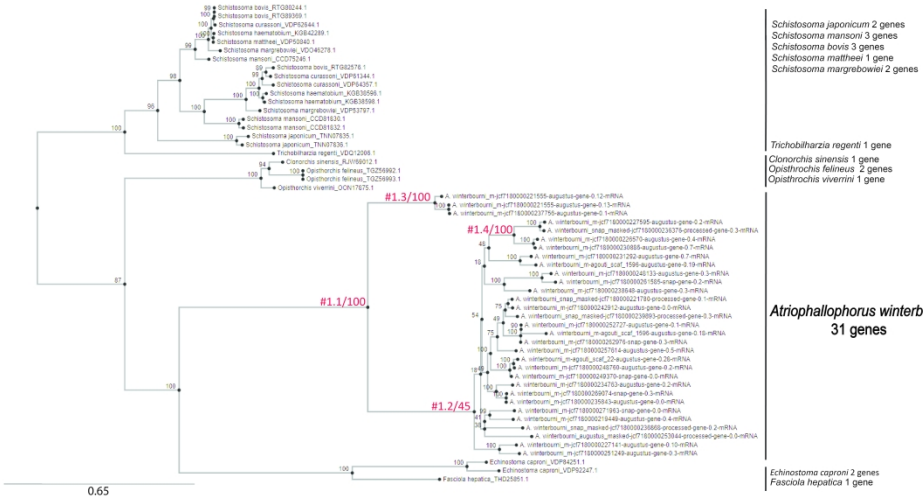


Figure 4. Gene tree of gene family HOG 25969 created with IQ-TREE. The tree is unrooted. Each name is a species name followed by the original gene name (protein name). *A. winterbourni* gene names are shortened version of gene names in Supplementary Table 10. The numbers above branches indicate ultrafast bootstrap support, for the #1 branches the bootstrap support is after a backslash. The branches labelled with #1.X indicate the separation between the foreground branches and the background branches (distinction used in codeml for investigation of selection). The test for selection compares the dN/dS between the foreground branch and the background branches. The total number of genes in this HOG per trematode species is given next to each species name.

685x348mm (300 x 300 DPI)