BIOINFORMATICS

The digenean complex life cycle: phylostratigraphy analysis of the molecular signatures

Maksim Nesterenko[1,2], Sergei Shchenkov[1], Sofia Denisova[1], and Viktor Starunov[1,2]

[1]Department of Invertebrate Zoology, Faculty of Biology, Saint Petersburg State University, Universitetskaya nab., 7–9, Saint Petersburg, 199034, Russian Federation
[2]Zoological Institute, Russian Academy of Sciences, Universitetskaya nab., 1, Saint Petersburg, 199034, Russian Federation

Address correspondence and requests for materials to Maksim Nesterenko, maxnest.research@gmail.com

Abstract

The parasitic flatworms from Digenea group have been the object of numerous in-depth studies for several centuries. The question of the evolutionary origin and transformation of the digenean complex life cycle remains relevant and open due to the biodiversity of these parasites and the absence of fossil records. However, modern technologies and analysis methods allow to get closer to understanding the molecular basis of both the realization of the cycle and its complication. In the present study, we have applied phylostratigraphy and evolutionary transcriptomics approaches to the available digenean genomic and transcriptomic data and built ancestral genomes models. The comparison results of Platyhelminthes and Digenea ancestor genome models made it possible to identify which genes were gained and duplicated in the possible genome of digenean ancestor. Based on the bioprocesses enrichment analysis results, we assumed that the change in the regulation of many processes, including embryogenesis, served as a basis for the complication of the ancestor life cycle. The evolutionary transcriptomics results obtained revealed the "youngest" and "oldest" life cycle stages of *Fasciola gigantica*, *F. hepatica*, *Psilotrema simillimum*, *Schistosoma mansoni*, *Trichobilharzia regenti*, and *T. szidati*. Our results can serve as a basis for a more in-depth study of the molecular signatures of life cycle stages and the evolution transformation of individual organ systems and stage-specific traits.

Keyword: flatworms, Digenea, complex life cycle, molecular signature, phylostratigraphy, evolutionary transcriptomics

Introduction

Digenea is one of the numerous and species-rich groups of parasitic flatworms. The study of its complex life cycle is of medical and fundamental importance. Although the first description of the life cycle stages (LCS) could be found even in a work dated 1379 (Reinhard, 1957), the first complete scheme of the digenean life cycle was obtained in the early 1880s (Reinhard, 1957). A sequential alternation of several contrasting LCS occurs during the complex life cycle of Digenea (Figure 1). The first one is free-living miracidium, the larva of the mother sporocyst. Mother sporocyst parasitizes in the first intermediate host, usually a mollusk. The next parasitic stage is daughter redia / sporocyst. After several acts of self-reproduction, these LCS produce cercariae, the larvae of the amphimictic generation. Cercariae leave the first intermediate host and encyst on a suitable substrate (to form adolescaria) or inside the second intermediate host (to form metacercaria). After infection of the definitive host, the parasite undergoes maturation, turning into an adult worm, which produces eggs with miracidia by amphimixis.

Citation: Nesterenko, M., Shchenkov, S., Denisova, S., and Starunov, V. 2022. The digenean complex life cycle: phylostratigraphy analysis of the molecular signatures. Bio. Comm. 67(2): 65–87. https://doi.org/10.21638/spbu03.2022.201

Authors' information: Maksim Nesterenko, PhD Student, orcid.org/0000-0002-8807-1115; Sergei Shchenkov, Senior Research Assistant, orcid.org/0000-0002-0579-1660; Sofia Denisova, PhD Student, orcid.org/0000-0002-5602-5894; Viktor Starunov, PhD, Senior Researcher, orcid.org/0000-0002-9001-2069

Manuscript Editor: Kirill Antonets, Department of Cytology and Histology, Faculty of Biology, Saint Petersburg State University, Saint Petersburg, Russia

Received: May 11, 2021;

Revised: February 2, 2022;

Accepted: February 15, 2022.

Funding: The research was funded by the Russian Foundation for Basic Research, project No. 19-34-90111 and by the State research project No. 1021051703357-3.

Ethics statement: This paper does not contain any studies involving human participants or animals performed by any of the authors.

Supplementary information: Supplemental material to the article is available at https://doi.org/10.21638/spbu03.2022.201. Supplementary files are published as submitted by the authors, and are not copyedited.

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

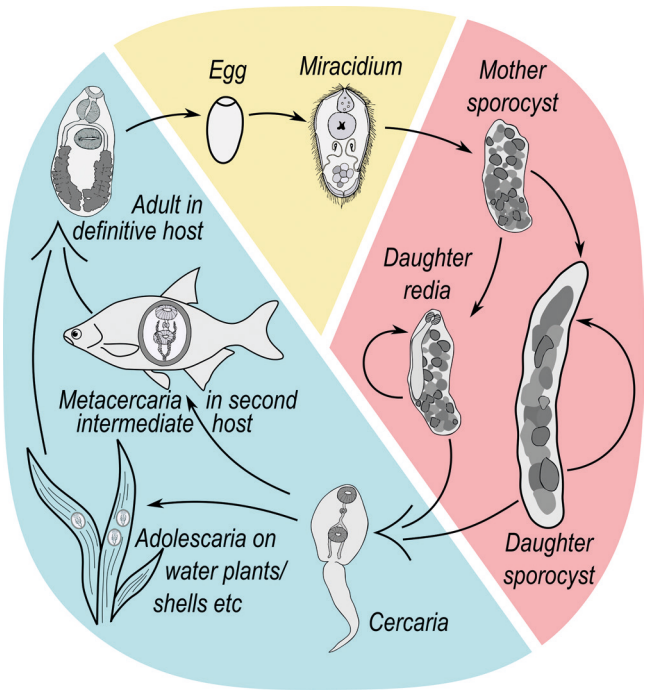

Fig. 1. Generalized scheme of the digenean complex life cycle. Color sectors indicate contrast generations: yellow and pale red ones cover parthenogenetic generations, whereas blue sector covers amphimictic generation. Miracidium and cercaria are free-swimming stages. Mother sporocyst, daughter redia / sporocyst parasitize inside the first intermediate host, usually Gastropoda. The reverse loop shows parthenogenetic reproduction of daughter rediae and sporocysts.

There are two general hypotheses on the origin of the digenean life cycle: "adult-first" and "parthenitae-first" (Gibson, 1987; Galaktionov and Dobrovolskij, 2003). These hypotheses are based on the existence of two fundamentally different generations — amphimictic (cercaria, metacercaria, adult) and parthenogenetic (miracidium, mother sporocyst, daughter sporocyst / redia) ones. The morphological and functional adaptations of these generations differ significantly. In general, these LCS possess shared features with various species of flatworms. This applies to the organization of tegument, body cavity, excretory, digestive, and nervous systems. The morphological diversity of digeneans, along with their broad ecological radiation, makes the determination of "ancestral" LCS even more difficult. Questions about the molecular basis of life cycle complication remain open.

The problem of the digenean life cycle origin is close to the unsolved question on the interpretation of the parthenogenetic LCS reproduction. The reproductive organ of sporocysts and rediae is represented by the so-called germinal mass. This is an organ of several undifferentiated and supporting cells. Germinal mass may be connected with the body wall or float in the body cavity. Young embryos at the initial stages of development could be found inside the germinal mass. The nature of undifferentiated cells is still unclear. Some researchers classify them as reproductive lineage cells, others call them stem cells. If undifferentiated cells are the reproductive lineage cells, then parthenitae multiply by apomictic parthenogenesis, i.e. a type of sexual reproduction, although without fertilization. If undifferentiated cells are stem cells, parthenitae use clonal reproduction. For more information see, for example, Galaktionov and Dobrovolskij, 2003.

The formation of digenean life cycles and interactions with their hosts took place a long time ago, probably in the Cambrian or Ordovician period (Littlewood, 2006). The absence of solid covers or skeletons in flatworms precludes accurate data from the fossil record. It made impossible to reconstruct the early stages of host-parasite interactions with paleontological methods. However, modern methods of bioinformatic analysis may serve as a basis for the reconstruction of early evolutionary events in Digenea.

According to the concept of genomic phylostratigraphy, the genome of every extant species retains parts of the picture of the evolutionary epochs (Domazet-Lošo, Brajković and Tautz, 2007). The method based on this concept provides a statistical approach for reconstructing macro evolutionary trends since it identifies the first emergence of the founder for new gene lineage or family and based on punctuated appearance of protein families. As a result, the whole gene set of the analyzed species can be distributed among different phylostrata — sets of genes that coalesce to founder genes having a common phylogenetic origin (Domazet-Lošo, Brajković and Tautz, 2007). The results of both phylostratigraphy and expression quantification allow describing transcriptome conservation in terms of expression dynamics of genes with different phylogenetic origins. For example, this approach has significantly expanded the knowledge about the sporulation in *Bacillus subtilis* (Shi et al., 2020) as well as the origin of insect wings (Almudi et al., 2020). Moreover, the phylostratigraphy approach was successfully applied to study even more complex and multicomponent traits like a biphasic life cycle and its emergence in evolution (Wang et al., 2020).

Here we studied LCS-specific molecular signatures within and between complex life cycles of six digenean species (*Fasciola gigantica*, *F. hepatica*, *Psilotrema simillimum*, *Schistosoma mansoni*, *Trichobilharzia regenti*, and *T. szidati*). In addition, we reconstructed the genome models of Platyhelminthes and Digenea ancestors. We suggest that genetic innovations acquired by the digenean ancestor allowed to complicate the molecular basis of numerous processes, including the regulation of gene expression and embryogenesis. Moreover, we applied phylostratigraphy to genomic and transcriptomic data of 14 flatworm species, both free-living and parasitic. Based on the obtained results, we identified the "oldest" and "youngest" LCS among considered for 6 digenean species.

Materials and methods

Data preparation

In the current research, we used publicly available data for flatworms both free-living — (*Macrostomum lignano* Ladurner, Schorer, Salvenmoser & Rieger, 2005, *Prostheceraeus vittatus* Montagu, 1815, and *Schmidtea mediterranea* Benazzi, Baguna, Ballester & del Papa, 1975) and parasitic (*Clonorchis sinensis* Cobbold, 1875, *Fasciola gigantica* Cobbold, 1855, *F. hepatica* Linnaeus, 1758, *Opisthorchis felineus* (Rivolta, 1884) Blanchard, 1895, *O. viverrini* (Poirier, 1886) Stiles & Hassal, 1896, *Psilotrema simillimum* Mühling, 1898, *Schistosoma haematobium* Bilharz, 1852, *S. japonicum* Katsurada, 1904, *S. mansoni* Sambon, 1907, *Trichobilharzia regenti* Horák, Kolářová & Dvořák, 1998, and *T. szidati* Neuhaus, 1952). The condition on which a species was included into analysis was the presence of a high-quality assembly of the genome (N50 > 175 Kb and more than 50 % of successfully assembled single-copy orthologues that are present in more than 90 % of animals according to WormBase Parasite statistics) or the presence of a reference transcriptome and paired-end read libraries from at least two LCS. For most of the species (*C. sinensis* (PRJNA386618; ASM360417v1, GCA_003604175.1), *F. gigantica* (PRJNA230515; F_gigantica_1.0.allpaths, GCA_006461475.1), *F. hepatica* (PRJEB25283; Fasciola_10x_pilon, GCA_900302435.1), *M. lignano* (PRJNA371498; Mlig_3_7, GCA_002269645.1), *O. felineus* (PRJNA413383; ICG_Ofel_1.0, GCA_004794785.1), *O. viverrini* (PRJNA222628; OpiViv1.0, GCF_000715545.1), *S. haematobium* (PRJNA78265; SchHae_2.0, GCA_000699445.2), *S. japonicum* (PRJNA520774; ASM636876v1, GCA_006368765.1), *S. mansoni* (PRJEA36577; Smansoni_v7, GCA_000237925.3), and *S. mediterranea* (PRJNA379262; ASM260089v1, GCA_002600895.1) latest genome assemblies [accessed 19 Aug 2021] from WormBase Parasites (Howe et al., 2017) database were analyzed. The reference transcriptomes of *P. simillimum* (Nesterenko et al., 2020), *P. vittatus* (Brandl et al., 2016; Martín-Durán et al., 2017) (PlanMine: bg_Pvit_v1), *T. regenti* (Leontovyč et al., 2016) (after direct request), and *T. szidati* (Leontovyč et al., 2019) (after direct request) were used. To reduce the redundancy of transcriptome data, the MMseqs2 (Mirdita, Steinegger and Söding, 2019) clustering algorithm was applied on sets of predicted amino acid sequences with the following parameters: mmseqs easy-cluster –cov-mode 0 –min-seq-id 0.9. In subsequent analyses, only genes encoding proteins with lengths ≥ 100 amino acids ("long proteins") were included for each of the 14 flatworm species.

We used the paired-end libraries of raw reads for *F. gigantica* (eggs, miracidia, rediae, cercariae, metacercariae, 42-days-old and 70-days-old juveniles, adults) (Zhang et al., 2019), *F. hepatica* (eggs, metacercariae, NEJ at 1, 3 and 24 hours post excystment, 21-day-old juveniles, and adult) (Cwiklinski et al., 2015; McNulty et al., 2017), *P. simillimum* (rediae, cercariae, adults) (Nesterenko et al., 2020), *S. mansoni* (cercariae, 3 hours and 24 hours post-infection schistosomula, and 7-week-old mixed sex adults) (Protasio et al., 2012), *T. regenti* (cercariae, schistosomula) (Leontovyč et al., 2016), and *T. szidati* (cercariae, schistosomula) (Leontovyč et al., 2019). A complete list of library identifiers is provided in the Supplementary Table S1. Low quality and adapter sequences were removed from the libraries using fastP (Chen, Zhou, Chen and Gu, 2018) and the following parameters: –cut_right –cut_window_size 4 –cut_mean_quality 20 –qualified_quality_phred 20 –length_required 25. To avoid possible contamination, read libraries were compared with a database we created. The custom database included: (i) reference libraries of archaea, bacteria, fungi, plasmids, protozoa, vectors, viruses and *Homo sapiens* Linnaeus, 1758 from Kraken2 (Wood and Salzberg, 2014; Wood, Lu and Langmead, 2019) sources, (ii) genomes of 8 Gastropoda species (*Aplysia californica* Cooper, 1863, *Biomphalaria glabrata* Say, 1818, *Chrysomallon squamiferum* Chen, Linse, Copley & Rogers, 2015, *Elysia chlorotica* Gould 1870, *Gigantopelta aegis* Chen, Linse, Roterman, Copley & Rogers, 2015, *Haliotis discus hannai* Ino, 1953, *Lottia gigantea* Gray, 1834, *Pomacea canaliculate* Lamarck, 1828) from MolluscDB (Liu et al., 2021), (iii) the genomes of definitive hosts of the studied species, such as cow (*Bos taurus* Linnaeus, 1758; ARS-UCD1.2 (GCA_002263795.2)), chicken (*Gallus gallus* Linnaeus, 1758; GRCg6a (GCA_000002315.5)), duck (*Anas platyrhynchos platyrhynchos* Linnaeus, 1758; CAU_duck1.0 (GCA_002743455.1)), mouse (*Mus musculus* Linnaeus, 1758; GRCm39 (GCA_000001635.9)), and sheep (*Ovis aries* Linnaeus, 1758; Oar_rambouillet_v1.0 (GCA_002742125.1)) from Ensembl (Yates et al., 2020). The search for possible contamination was performed using Kraken2 (Wood, Lu and Langmead, 2019).

The orthogroups identification

For the orthogroups identification, we used the OMA standalone (Altenhoff et al., 2019) program (v2.5.0), and the analysis of long proteins sets was carried out in three steps. First, the program was run with default parameters with the "bottom-up" algorithm for inference of HOGs. The first launch was carried out without a phylogenetic tree, but with the indication of 3 free-living species (*M. lignano*, *P. vittatus*, *S. mediterranea*) as an outgroup. Second, we reconstructed the phylogenetic tree following the protocol of Dylus et al. (Dylus et al., 2020). In brief, using the filter_groups.py script provided, we selected OMA groups that included at least 13 of the 14 species

of flatworms. Then, using MAFFT (v7.487) (Katoh and Standley, 2013), multiple alignment of the sequences in each orthogroup was performed (–maxiterate 1000 –localpair). The alignments have been concatenated into a supermatrix using the concat_alignments.py script. The selection of suitable sites in the supermatrix was carried out using trimAl (-automated1) (Capella-Gutiérrez, Silla-Martínez, and Gabaldón, 2009). We used the ProtTest program (v3.4.2) (Guindon and Gascuel, 2003; Darriba, Taboada, Doallo and Posada, 2011) to determine the most appropriate sequence evolution model. The phylogenetic tree was reconstructed using the IQ-TREE (v2.1.4-beta) (Nguyen, Schmidt, Von Haeseler and Minh, 2015; Minh et al., 2020) with the following parameters: -m LG + I + G + F –seed 12345 -B 1000 –nmax 1000. Rooting by the outgroup of the consensus tree was performed using the "ape" (Paradis and Schliep, 2019) library for R. Third, the phylogenetic tree was used when OMA standalone was re-run with default settings. The construction of a heatmap with the number of common OMA groups between the studied species was performed in RStudio using the "ggplot2" (v3.3.5), "pheatmap" (v1.0.12), and "RColorBrewer" (v1.1-2) libraries.

Gene expression analysis

We used Salmon (v1.2) (Patro et al., 2017) to quantify the gene expression. The indices construction (–kmerLen 25) was performed on either mRNA transcripts from WormBase Parasite (*F. gigantica*, *F. hepatica*, *S. mansoni*) or available transcriptomes (*P. simillimum*, *T. regenti*, *T. szidati*). The following parameters were used to align previously prepared read libraries: -l A –seqBias –gcBias –minScoreFraction 0.50 –softclip –validateMappings. Tables with unaveraged and averaged between biological replicates TPM values were prepared.

Given the *F. hepatica* libraries were obtained from two different studies, we used ComBat-seq (Zhang, Parmigiani and Johnson, 2020) from the "sva" library (v3.36.0) for R to remove the batch effect. The analysis was performed using the NumReads calculated by Salmon and the adjusted counts were then converted to TPM. To separate *F. hepatica* metacercaria samples obtained in different studies, we renamed them as follows: "met0h" samples obtained by Cwiklinski et al. (Cwiklinski et al., 2015) and "meta" samples obtained by McNulty et al. (McNulty et al., 2017) were renamed to "early" and "late" metacercariae, respectively.

As a "molecular signature" of a LCS, we considered a set of genes with an expression level of ≥ 2TPM at the LCS under consideration. The expression threshold value was chosen in accordance with the results of studies by Wagner, Kin, and Lynch, according to which "genes with more than two transcripts per million transcripts (TPM) are highly likely from actively transcribed genes" (Wagner, Kin, and Lynch, 2013). If a gene had an expression of ≥ 2TPM at all LCS considered, we classified it as "common expressed".

The detection of significant variation of gene expression was performed with "RNentropy" (v1.2.2) (Zambelli et al., 2018) library for R. We used the corrected global sample specificity test $P < 0.01$ by the Benjamini-Hochberg method, and a local sample specificity test $P < 0.01$. The analysis was carried out on the tables with unaveraged TPM values for *F. gigantica*, *F. hepatica*, *P. simillimum*, *S. mansoni*, *T. regenti*, and *T. szidati*.

The co-expressed gene clusters searching was performed using Clust (v1.10.8) (Abu-Jamous and Kelly, 2018) on sets of genes with expression level ≥ 2TPM at least on 1 LCS. The analysis was carried out only for *F. gigantica*, *F. hepatica*, *P. simillimum*, and *S. mansoni* since the transcriptomes from three or more LCS were available. The analysis was carried out on tables with averaged TPM values.

Multidimensional scaling

For Multidimensional Scaling (MDS), the matrices of presence ("1") and absence ("0") of expression were prepared. We used 2TPM as the expression threshold for classifying presence, both for intraspecies and interspecies comparison. Interspecies analysis was performed on OMA groups which have the property that all members are orthologous to all other members of the same group. For interspecies comparison, additional matrices were prepared indicating whether the ortholog is overexpressed on the LCS ("1") or not ("0"). If the species was not included in this OMA group, then "NA" was indicated. Rows (genes) containing only the same symbols (invariant) or at least 1 "NA" were excluded from the tables. Interspecies analysis was carried out in two steps: (i) comparison of "redioid" (*F. gigantica*, *F. hepatica*, and *P. simillimum*) and "sporocystoid" (*S. mansoni*, *T. regenti*, and *T. szidati*) species separately, (ii) comparison of all species together.

The optimal number of clusters was determined using the "silhouette" method implemented in the "factoextra" library (v1.0.7) for R. We used the metaMDS function from the "vegan" library (v2.5-7) with the following parameters: distance = "manhattan", try = 100, trymax = 1000, autotransform = FALSE, binary = TRUE, k = the optimal number of clusters. Both in determining the optimal number of clusters and in MDS, we set the seed to be 1234. To visualize the results, we used ggscatter function from "ggpubr" (v0.4.0) library for R.

GSEA

The protein sets of *F. gigantica*, *F. hepatica*, *P. simillimum*, *S. mansoni*, *T. regenti*, and *T. szidati* flatworm species were annotated using eggNOG-mapper (v2) (Huerta-Cepas

et al., 2019; Cantalapiedra et al., 2021) web-resource with default parameters. The GSEA was performed using "topGO" library (v2.40.0) for R. The analysis was performed for LCS-specific sets of over-expressed genes for *F. gigantica*, *F. hepatica*, *P. simillimum*, *S. mansoni*, *T. regenti*, and *T. szidati*. Only the GO-terms describing biological processes were considered. We used Fisher's exact test, and among the results (GO-terms with P-value < 0.01) extracted only terms in which at least 10 significant genes were included. Redundancy reducing was carried out with "rrvgo" library (v1.0.2) for R. We used the minus log10-transformed p-values as scores and 0.7 as the threshold for reduceSimMatrix.

Ancestors' models reconstruction and analysis

The Platyhelminthes and Digenea ancestor's genome models construction, as well as the vertical comparison between them, were performed using pyHAM (v1.1.10) (Train, Pignatelli, Altenhoff and Dessimoz, 2019) for Python. As input, we used the rooted tree with the internal node's names, as well as the HierarchicalGroups.orthoxml obtained by restarting OMA standalone. Given the possible incompleteness of the protein sets of the studied species, we excluded HOG from the ancestor genome model if it included sequences of less than 75 % of the species considered. The "Platyhelminthes" model corresponded to the last common ancestor of free-living and parasitic digenean flatworm species, and the "Digenea" model corresponded to the last common ancestor of all studied digenean species. We used *S. mansoni* genome as a reference for extracting genes from ancestor's genome models since *S. mansoni* is one of the most studied digenean species and high-quality genomes are available. The complete python code is available at the link: https://github.com/maxnest/The_phylostratigraphy_analysis_of_the_digenean_molecular_signatures/blob/main/pyHAM_flatworm_ancestor_genomes.py.

The GSEA was performed for the lists of gained and duplicated genes from Digenea ancestor genome model. We used the same parameters for the analysis of molecular signatures, with one exception that only *S. mansoni* genes included in the digenean ancestor genome model were used as a background. The redundancy reduction and visualization were done using the "rrvgo" library.

ESP identification and analysis

The identification of the potential ESP for 11 digenean species was carried out according to the pipeline described in Garg and Ranganathan's (Garg and Ranganathan, 2011) manuscript. First, all long proteins of each species were analyzed with SignalP (v5.0b) (Almagro Armenteros et al., 2019). Based on the analysis results, the proteins were divided into potential "classical" (SP ≥ 0.5) and "non-classical" (SP < 0.5) ESP. Second, the SecretomeP (v1.0) (Bendtsen et al., 2004) was used to analyze sets of potential "non-classical" ESP. Only proteins with NN-scores ≥ 0.9 and which were predicted not to contain a signal peptide were considered as potential "non-classical" ESP. Third, all potential ESP were scanned for the presence of the mitochondrial transit peptide with TargetP (v2.0) (Emanuelsson, Nielsen, Brunak and Von Heijne, 2000). The proteins with such signals were excluded from the analysis. Fourth, TMHMM (v2.0c) (Krogh, Larsson, Von Heijne and Sonnhammer, 2001) was used to detect transmembrane domains in proteins, and proteins without them were considered as potential ESP.

The GSEA was performed for both potential "classical" and "non-classical" ESP for each of 11 digenean species using "topGO" in a similar way to molecular signature analysis.

Phylostratigraphy and TAI analysis

The phylostratigraphy for all 14 flatworm species considered was performed using "phylostratr" library (v0.2.1) (Arendsee et al., 2019) for R. For each flatworm species analyzed, we used data of all other species as well as the prebuilt dataset of prokaryotes (function "use_recommended_prokaryotes"), human (function "add_taxa(9606)"), and yeast (function "add_taxa(4932)"). Similarity searching between proteins of analyzed species and prepared dataset was carried out with BLASTp (v2.6.0+) (Camacho et al., 2009). Tables with BLAST results in "6" output format were used as input for "phylostratr" for protein distribution between phylostrata: 1) "Cellular organisms", 2) "Eukaryota", 3) "Opisthokonta", 4) "Metazoa", 5) "Eumetazoa", 6) "Bilateria", 7) "Protostomia", 8) "Spiralia", 9) "Lophotrochozoa", 10) "Platyhelminthes", 11) "Class" (Rhabditophora / Digenea), 12) "Order" (Plagiorchiida), 13) "Family" (Echinostomatoidea / Opisthorchiidae / Schistosomatidae), 14) "Genus" (Fasciola / Opisthorchis / Schistosoma / Trichobilharzia), 15) "Species". The results of phylostratigraphy were used to determine the composition of the following groups of genes: (i) genes included in the genome models of Platyhelminthes and Digenea ancestors, (ii) genes having an expression level ≥ 2TPM at all LCS considered, (iii) genes encoding potential "classical" and "non-classical" ESP. The composition visualization was carried out using "ggplot2", "viridis" (v0.6.1), and "reshape" (v0.8.8) libraries for R.

The TAI definition was performed for *F. gigantica*, *F. hepatica*, *P. simillimum*, *S. mansoni*, *T. regenti*, and *T. szidati* using phylostratigraphy results and tables with averaged TPM values. The analysis was carried out with "myTAI" library (v0.9.3) (Drost et al., 2018) for R. Genes with an expression level < 2 TPM at all compared LCS were excluded from the tables. The analysis was performed on the transformed TPM values (log2(x + 1)). The FlatLineTest

function was used to quantify the statistical significance of the global TAI pattern. For analysis using PlotRE and PlotBarRE functions, the phylostrata were divided into groups "before" (1-11 phylostrata), and "after" (from 12 to species-specific phylostrata) the division of Digenea. For *P. simillimum*, the entire analysis was performed twice: (i) with all genes included and (ii) without genes assigned to the species-specific phylostratum.

Using the pMatrix function from "myTAI", the contributions of genes to TAI of LCS were determined. For the LCS with the smallest and largest TAI within each of the 6 species among the genes with the GO annotation, 500 genes with the largest contribution were selected. Further, GSEA for selected genes was performed in a similar way to a molecular signature.

Results

Low-quality sequences and contamination were successfully removed from available data

The available high-quality genomic and transcriptomic data of three free-living flatworms (*Macrostomum lignano*, *Prostheceraeus vittatus*, and *Schmidtea mediterranea*) and 11 digenean (*F. gigantica*, *F. hepatica*, *Clonorchis sinensis, Opisthorchis felineus*, *O. viverrini*, *P. simillimum*, *S. haematobium*, *S. japonicum*, *S. mansoni*, *T. regenti*, and *T. szidati*) species were used. We focused only on genes encoding proteins of at least 100 amino acids in length. In each of the species, the share of such long proteins from the total number considered was at least 81 %.

Comparative transcriptomic analysis between contrasting LCS of the digenean life cycle plays a key role in understanding its genome activity. We used the transcriptomes available for six species (*F. gigantica*, *F. hepatica, P. simillimum, S. mansoni, T. regenti*, and *T. szidati*) and prepared the data for further analysis. The summary of data processing is presented in Table 1. The low-quality sequences, as well as adapters and short sequences were removed from libraries. At least 79 % of the raw reads successfully passed all filters for each library. Given that most of the libraries were obtained from parasitic LCS living in various definitive hosts, a possible contamination was removed. We prepared a database including a standard collection of sequences (archaea, bacteria, fungi, plasmids, protozoa, vectors, viruses and *Homo sapiens*) and the genomes of eight Gastropoda species as well as the definitive hosts of the studied species, such as cow (*Bos taurus*), chicken (*Gallus gallus*), duck (*Anas platyrhynchos platyrhynchos*), mouse (*Mus musculus*), and sheep (*Ovis aries*). The percentage of contamination varied widely between samples but did not exceed 8 % of the total number of reads in each species (Table 1). Additional information about complex life cycles of digenean species considered is available in the Supplementary Table S2.

More than 23 and 36 thousand of HOG and OMA groups were discovered, respectively

We used the OMA standalone program (Altenhoff et al., 2019) for the orthogroups identification. The analysis of long proteins sets was carried out in three steps. The orthologs searching was carried out without the phylogenetic tree, but with the indication of three free-living species as an outgroup. As a result, 36486 OMA groups were identified. Among them the proteins of all 14 flatworm species under consideration were presented simultaneously in 175 OMA groups only, as well as all digenean species were presented in 438 OMA groups.

Then, the phylogenetic tree of the considered species had been reconstructed based on the results obtained during the previous step. Among OMA groups we selected 614 OMA groups that included proteins of at least 13 out of 14 species. Then we built a multiple amino acid sequences alignment in each selected orthogroup separately and concatenated the alignments into a supermatrix. After filtering, the supermatrix contained 345931 sites. The topology of the reconstructed tree had full support for all main branches (Figure 2A) and corresponded to modern ideas about the phylogenetic relationship of the studied species (Pérez-Ponce de León and Hernández-Mena, 2019). The reconstructed tree was used when OMA standalone was relaunched.

According to the results obtained, 36486 OMA group and 23852 hierarchical orthologous groups (HOG) were identified. The number of OMA groups common for species pairs are shown in Figure 2B. Close-

Table 1. Summary of paired-end reads libraries processing

Metrics	*F. gigantica*	*F. hepatica*	*P. simillimum*	*S. mansoni*	*T. regenti*	*T. szidati*
Total # of raw reads	1159738312	2299488146	315007934	427288262	226839830	170841194
Total # of reads passed filters	1097332390 (94.61 %)	2227302396 (96.86 %)	297575468 (94.46 %)	389356068 (91.12 %)	204942936 (90.34 %)	147496064 (86.33 %)
Total # of reads after decontamination	1079253166 (98.35 %)	2186776391 (98.18 %)	291340169 (97.9 %)	381044468 (97.86 %)	189863940 (92.64 %)	136050134 (92.23 %)

Total # of reads passed filters — the number of reads that met the quality requirement; Total # of reads after decontamination — the number of reads that remained in the libraries after comparing with the prepared database with potential sources of contamination. In parentheses are the percentages of the number of reads relative to the number of reads at the previous step of the analysis.

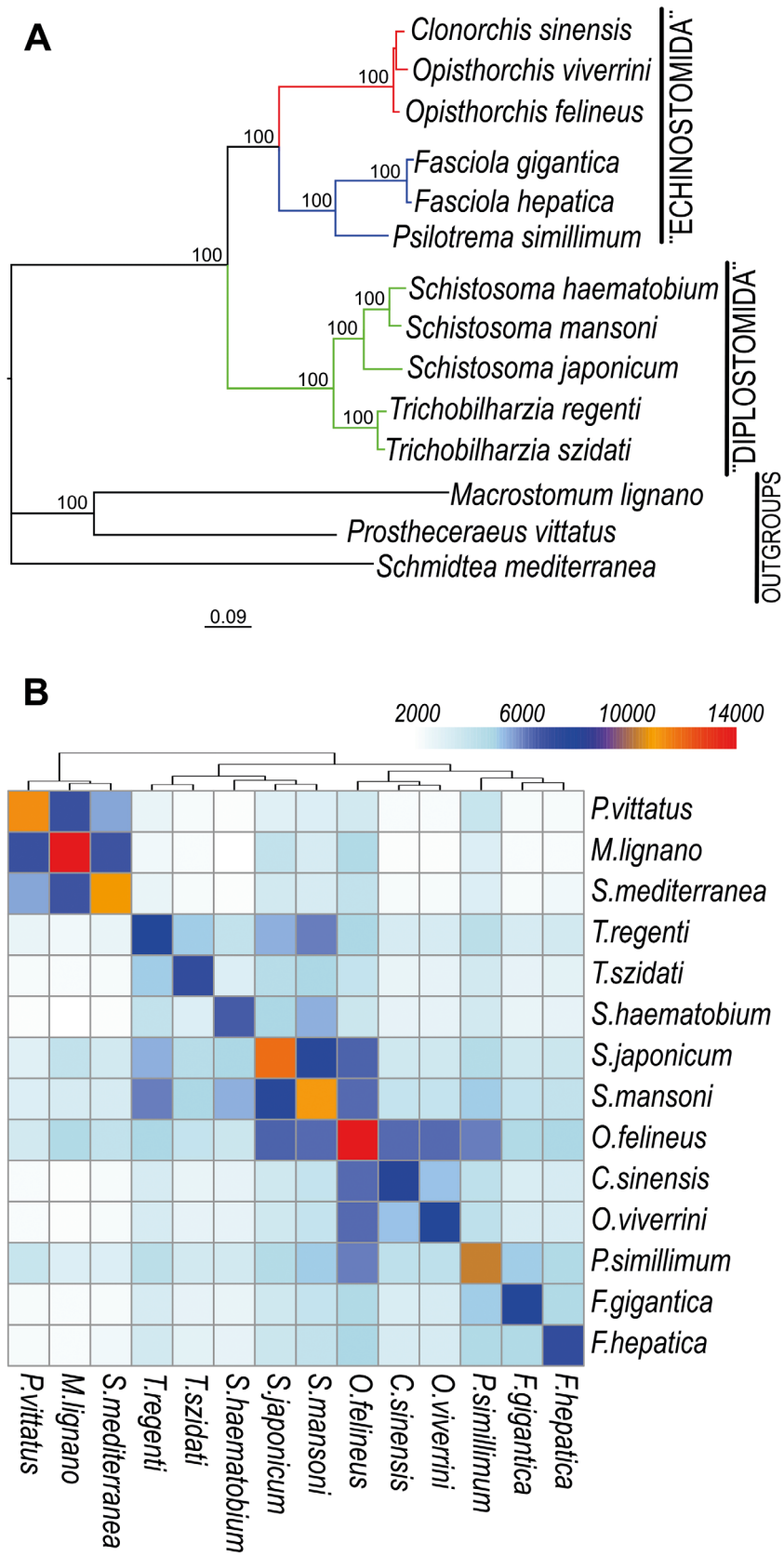

Fig. 2. Relationships between the studied flatworm species based on sequence similarities (A) and the number of common OMA groups (B). On the phylogenetic tree (A), colors indicate different taxa: Opisthorchiidae (*C. sinensis*, *O. viverrini*, *O. felineus*), Echinostomatoidea (*F. gigantica*, *F. hepatica*, *P. simillimum*), Schistosomatidae (*S. haematobium*, *S. mansoni*, *S. japonicum*, *T. regenti*, *T. szidati*). Names of clades are given according to Odening, 1974. Data of "Plagiorchiida" s. str. are absent in public databases. The color key on heatmap (B) shows the number of shared OMA groups among species.

Table 2. Summary of genes expression analysis

Metrics	*F. gigantica*	*F. hepatica*	*P. simillimum*	*S. mansoni*	*T. regenti*	*T. szidati*
# of genes with "common" expression	5104	5698	8116	8212	6766	6837
Min mol.signature size (LCS)	6349 (egg)	6421 (egg)	9938 (adult)	9480 (adult)	6948 (cercaria)	6958 (cercaria)
Max mol.signature size (LCS)	9273 (metacercaria)	7999 (juvenile)	11481 (redia)	10353 (somule 3hr)	9197 (somule)	9393 (somule)
Total # of overexpr. genes	7538	5757	7171	4539	6578	6104
Min # of overexpr. genes (LCS)	352 (juvenile 70d)	446 (adult)	1137 (adult)	585 (cercaria)	535 (cercaria)	449 (cercaria)
Max # of overexpr. genes (LCS)	3466 (metacercaria)	2648 (juvenile)	4415 (redia)	1944 (adult)	6043 (somule)	5655 (somule)
Number of clusters	16	9	12	16	X	X
Total # of genes in clusters	5125 (49.09 %)	3201 (37.03 %)	9789 (70.02 %)	8039 (69.46 %)	X	X
# of genes not included in any cluster	5314	5443	4192	3534	X	X
Min cluster size	13	106	426	36	X	X
Max cluster size	998	897	1898	1471	X	X
Average cluster size	320.3125	355.6667	815.75	502.4375	X	X

# of genes with "common" expression — the number of genes with an expression level ≥ 2TPM at all stages of the cycle considered; Min / Max mol.signature size — the minimum and maximum size of the molecular signature in genes, respectively; Total # of overexpr. genes — the total number of genes with a statistically significant increase in expression at least at 1 stage of the life cycle; Min / Max # of overexpr. genes — the minimum and maximum number of genes with a statistically significant increase in expression, respectively; Number of clusters — the number of co-expression clusters; Total # of genes in clusters — the total number of genes included in the co-expression clusters; # of genes not included in any cluster — the number of genes not included in any co-expression cluster; Min / Max cluster size — minimum and maximum size of co-expression clusters in genes, respectively; LCS — life cycle stages for which the value was obtained; somule — schistosomulum; X — the analysis was not performed.

ly related species have more common OMA groups than distant relatives. All digenean species were divided into two distinct clusters, such as "Plagiorchiida" (*C. sinensis, F. gigantica, F. hepatica, P. simillimum, O. felineus, O. viverrini*) and "Schistosomatidae" (*S. haematobium, S. japonicum, S. mansoni, T. regenti,* and *T. szidati*). The three considered free-living species form an outgroup.

During the digenean complex life cycle, numerous genes change their expression

The living abilities of LCS require complex regulation of genome activity. Previously, we analyzed gene expression in two Psilostomatidae species during their complex life cycles (Nesterenko et al., 2020). In this study, we updated the set of analyzed data to clarify and expand our notions of the genome activity: 1) what is the molecular signature of different LCS, 2) how many genes have a noticeable expression throughout the life cycle, 3) how many genes are differentially expressed between the LCS, and 4) how many clusters of co-expressed genes can be found in different species.

We define the "LCS molecular signature" as the set of all genes with an expression level ≥ 2 Transcripts-Per-Million (TPM) at the LCS under consideration. According to the results obtained, each molecular signature in the six analyzed species included at least 60 % of the total number of genes encoding long proteins. In each case the signature included several thousand of genes, and the minimum and maximum sizes of signatures are presented in Table 2.

Among all species under consideration, the number of protein-coding genes with noticeable expression throughout the life cycle was 48.89 % in *F. gigantica* only. In all other species it was above 50 %. At the same time, the share of differentially expressed genes from the molecular signature size varied widely, but in most cases (18 / 27 LCS) did not exceed 20 %.

In order to determine whether some groups of genes alter their expression in a similar way throughout the life cycle, the *in silico* identification of the co-expression clusters was carried out. As it can be seen from Table 2, in two *Fasciola* species analyzed less than 50 % of genes were included into clusters, whereas in *P. simillimum* and *S. mansoni* approximately 70% of genes considered were co-expressed.

The results of LCS molecular signature definition and differential expression analysis are available in the Supplementary Table S3.

Molecular signatures of similar LCS were clustered together in both intraspecific and interspecific comparisons

To determine how similar molecular signatures are within a single life cycle as well as between species, we used multidimensional scaling (MDS). Figure 3 shows the results of the distribution of samples within a life cycle on a reduced 2-dimensional space. In most cases (11 / 18 clusters) clusters contained only biological replicates of one LCS. In *F. gigantica*, cluster #3 contained replicates of juveniles of different ages (42- and 70-days post infection) and adults, and cluster #4 contained replicates of cercariae and metarcercariae. Cluster #1 in *F. hepatica* included replicates of an adult and a juvenile, while cluster #2 combined replicates of the early and late metacercariae, as well as all newly emerged juveniles (NEJs). In *P. simillimum*, redia and cercariae were combined into one cluster. Two

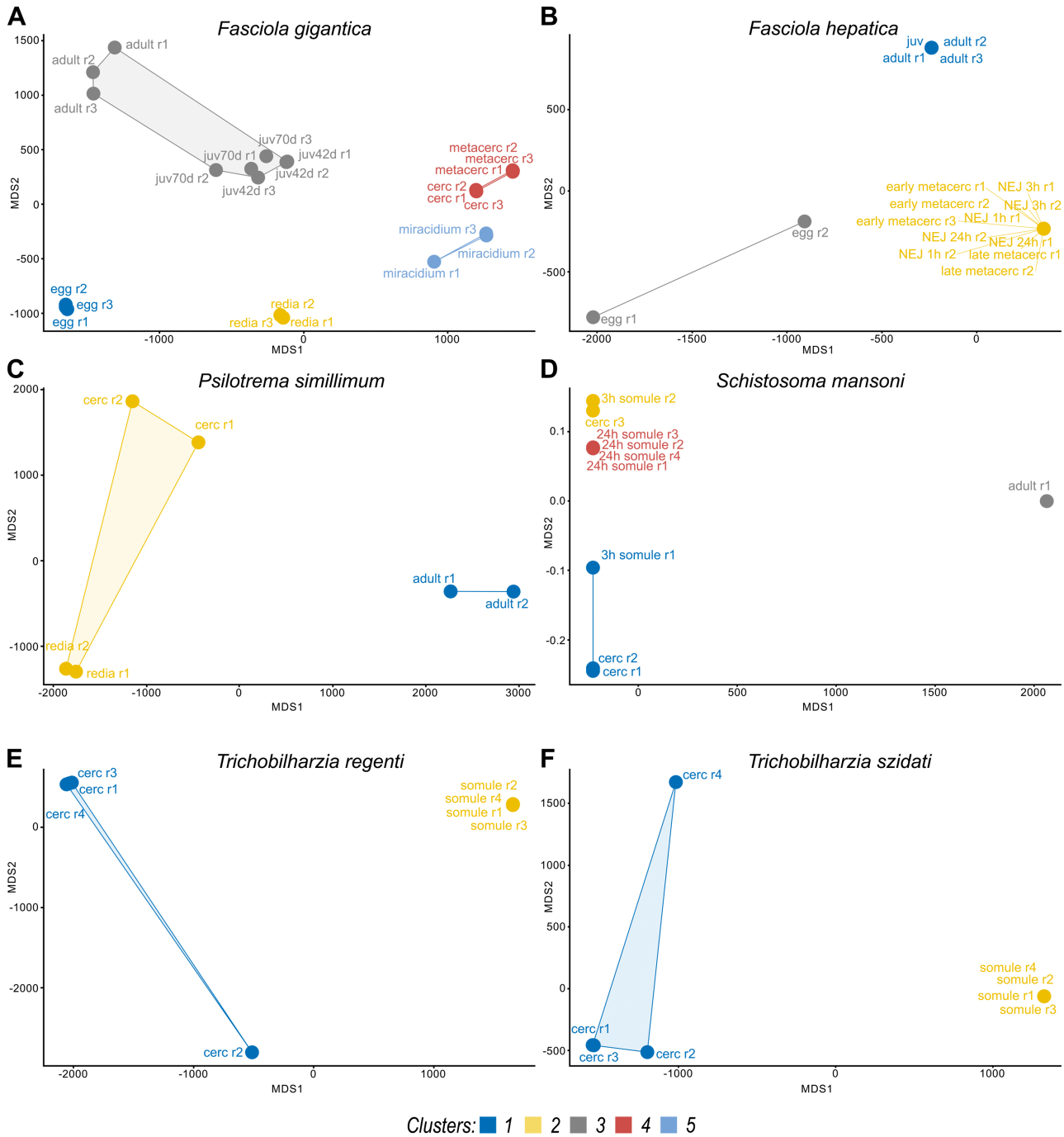

Fig. 3. Multidimensional Scaling (MDS) plots for stages within complex life cycles.
Abbreviations: cerc — cercaria, metacerc — metacercaria, juv — juvenile, juv42/70d — 42- and 70-days-old juveniles, respectively; NEJ 1/3/24h — newly excysted juveniles at 1, 3 and 24 h post excystment, respectively; somule — schistosomula, 3/24 h somule — 3- and 24-hours post-infection schistosomula, respectively; r1/2/3/4 — biological replication identifier. Different clusters are marked with colors.

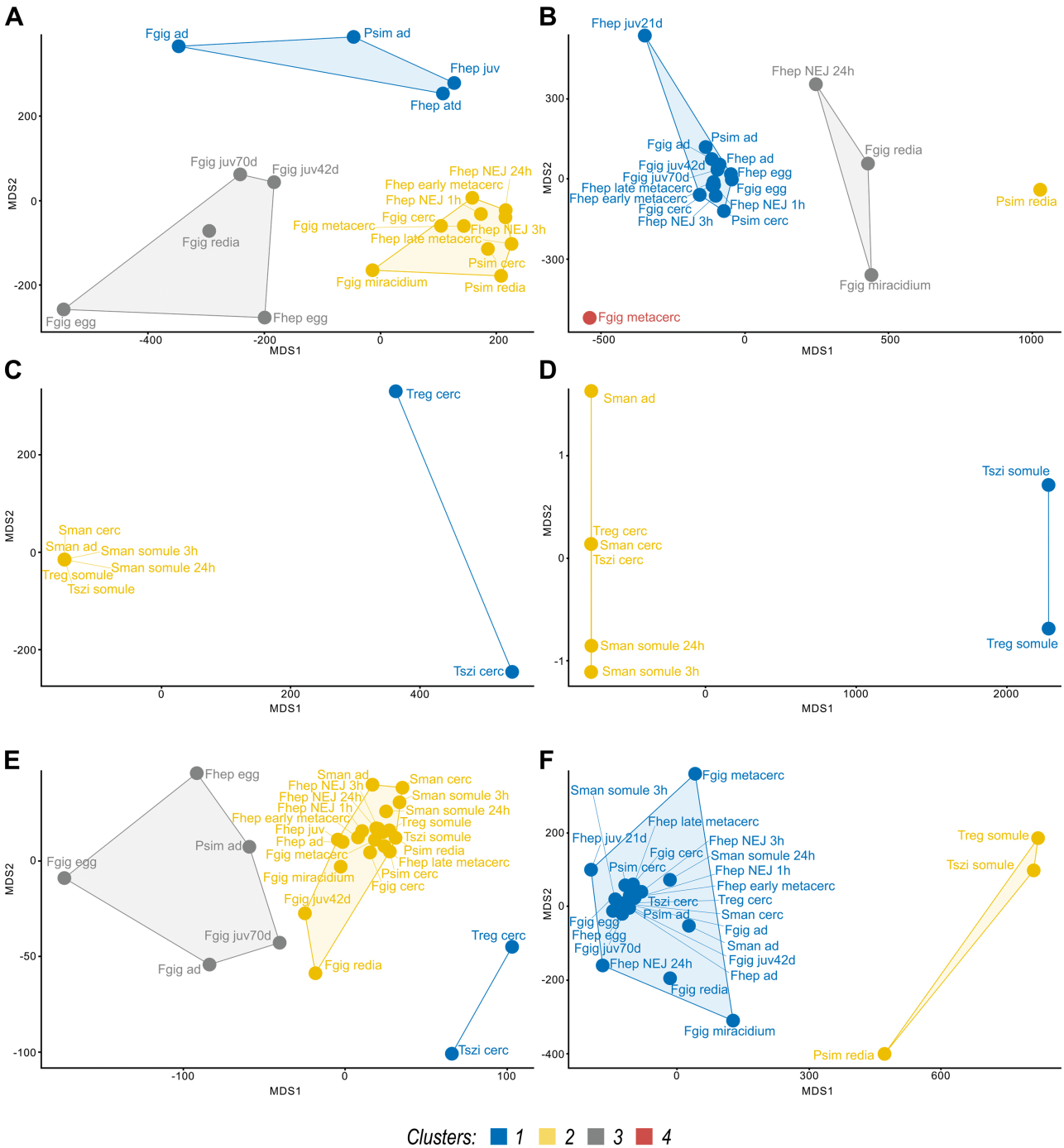

Fig. 4. Multidimensional Scaling (MDS) plots for interspecies comparison based on OMA group expression analysis. The "redioid" (A, B), "sporocystoid" (C, D) species, as well as all species together (E, F) were analyzed. Interspecies comparison was carried out based on the molecular signature (A, C, E) and sets of overexpressed orthologs (B, D, F) analysis results. Abbreviations: cerc — cercaria, metacerc — metacercaria, juv — juvenile, juv42/70d — 42- and 70-days-old juveniles, respectively; NEJ 1/3/24h — newly excysted juveniles at 1, 3 and 24 hours post excystment, respectively; somule — schistosomula, 3/24 h somule — 3- and 24-hours post-infection schistosomula, respectively; ad — adult worm. Different clusters are marked with colors.

clusters (#1 and #2) of *S. mansoni* had a similar composition and included one replicate of three hours post-infection schistosomula and one or two replicates of cercariae.

We carried out the interspecies analysis two times separately for different comparison matrices. In the first one, we divided the considered digenean species into two groups according to their phylogenetic relationship and the life cycles traits. The first group included the "redioid" species *F. gigantica*, *F. hepatica*, and *P. simillimum*. In the second group only the "sporocystoid" species were included (*S. mansoni*, *T. regenti*, and *T. szidati*). In the second analysis, we performed MDS of six species together. For all interspecies analyses, we built presence/absence matrices based on the expression of genes from OMA

groups. A gene was assigned "1" if it was included in LCS-specific molecular signature or was over-expressed at LCS, respectively. Otherwise, the status "0" (expression absent) or "NA" (a species was not included in the given OMA group) was assigned. With an increase in the number of compared species (from 3 to 6), the number of OMA groups that passed the filter and were included in the analysis decreased twofold approximately. All the results of interspecies MDS are shown in Figure 4.

When comparing entire molecular signatures among "redioid" species, we observed three clusters (see Figure 4A). One of them (cluster #1) included similar LCS of different species. In the other two clusters the composition was more diverse, but also included similar LCS of related species with rare exceptions. When considering differential expression (Figure 4B), four clusters were identified, two of which contained only one LCS: cluster #2 with *P. simillimum* redia and cluster #4 with *F. gigantica* metacercaria.

We found two clusters in "sporocystoid" species when analyzing both the entire molecular signatures (Figure 4C) and only overexpressed genes (Figure 4D). In the first case, the cercariae of *T. regenti* and *T. szidati* were united in a cluster separate from other LCS, while the schistosomula of these species were combined together with the cercaria, schistosomula, and the adult of *S. mansoni*. On the contrary, schistosomula of *Trichobilharzia* were united in a separate cluster, and their cercariae were included into a single cluster with all LCS of *S. mansoni* (Figure 4D).

The presence of a large cluster, which included most of the LCS considered, was a common result when comparing entire molecular signatures of LCS (Figure 4E) and only the sets of over-expressed genes (Figure 4F) of six species. Clustering of *Trichobilharzia* samples in both cases resembles those obtained when analyzing only "sporocystoid" species: either cercariae or schistosomula clustered separately from other LCS. However, in the second case, *P. simillimum* redia joined the schistosomula of *Trichobilharzia* (Figure 4F).

Interspecies comparative analysis revealed key bioprocesses

A Gene Set Enrichment Analysis (GSEA) with genes included in molecular signatures or over-expressed genes only makes it possible to distinguish the processes which occur at a particular LCS. According to the results of GSEA with over-expressed genes, the number of "enriched" bioprocesses was not less than 80 at each of the LCS considered. We reduced the diversity by keeping only the parental bioprocesses (PB) and compared the results between similar LCS of different species. The whole presence / absence matrix for parental bioprocesses is available in the Supplementary Table S4.

Among the 61 PB that were enriched in at least one of the *Fasciola* egg, only 14 PB are common for *F. gigantica* and *F. hepatica*, such as "regulation of cell death" and "developmental pigmentation". Data on miracidiae were available for *F. gigantica* only. In its set of PB were "cellular respiration", "toxin transport", "mitotic spindle organization", "liver regeneration", and the ones associated with metabolism and RNA processing.

Rediae of *F. gigantica* and *P. simillimum* share 7 PB only: "ribonucleoprotein complex biogenesis", "cellular nitrogen compound metabolic process", "RNA processing", "ribonucleoprotein complex assembly", "nuclear-transcribed mRNA catabolic process", "mitochondrial translation", "DNA metabolic process". We observed processes that may be associated with the development only in *P. simillimum*, including "DNA-dependent DNA replication", "G1/S transition of mitotic cell cycle", "sex determination", "maintenance of cell number", "Endodermal cell fate commitment", "midbrain dopaminergic neuron differentiation".

Data on cercariae were available for five digenean species: *F. gigantica*, *P. simillimum*, *S. mansoni*, *T. regenti*, *T. szidati*. A PB was considered as "common" if it was enriched for at least three species. In most cases (10 /15 PB) common PB were associated with metabolism, including ATP, NADH, NAD metabolic processes. Bioprocesses connected with muscle movement such as "muscle contraction" (*F. gigantica*), "regulation of muscle contraction" (*P. simillimum, S. mansoni*), and "skeletal muscle contraction" (*P. simillimum*) were also found.

The metacercariae of *F. gigantica* and *F. hepatica* were found to have 20 common PB. None of them were common to all three samples, and most (18 / 20 PB) were common to the metacercaria of *F. gigantica* and late metacercaria of *F. hepatica*. Among the common PB were such bioprocesses as "response to abiotic stimulus", "signaling", "locomotion", "cell-cell signaling", "tissue development", "epithelium migration", "behavior", "response to external stimulus", "defense response to other organism", "post-embryonic animal organ development". The only common bioprocess for the early metacercaria of *F. hepatica* and *F. gigantica* metacercaria was the "interspecies interaction between organisms".

F. hepatica NEJ transcriptomes were collected from 3 time points, which allows us to trace how exactly the activity of a young juvenile changes after leaving the metacercaria cyst (Cwiklinski et al., 2015). We identified 86 active PB, of which only 11 are common for at least two time points: "oxidation-reduction process", "regulation of cell death", "locomotion", "negative regulation of biological process", "glycosylation", and "cell surface receptor signaling pathway", etc. The PB "developmental process" (NEJ at 1 hour post excystment), "interspecies interaction between organisms" (1), "response to external stimulus" (1), "biological adhesion" (1), "neutrophil mediated immunity" (1), "regulation of hormone levels"

(NEJ at 3 hours post excystment), "negative regulation of secretion" (3), "regulation of cell differentiation" (3), "sarcomere organization" (NEJ at 24 hours post excystment), and "response to nutrient" (24) were found among the unique biological processes.

Samples of *F. gigantica* (42- and 70-days-old) and *F. hepatica* (21-days-old) juveniles were combined in the analysis. A total of 52 PB were found, most of which were "enriched" in only one of the LCS compared. For example, "cilium organization" and "cilium or flagellum-dependent cell motility" in *F. hepatica*, "maternal process involved in female pregnancy" or "female pregnancy" in 42-days-old or 70-days-old juvenile of *F. gigantica* were among the enriched PB. After excluding various metabolic processes, only six processes remained among 14 common PB, including "response to glucocorticoid", "epithelial cell proliferation", and "extracellular matrix organization".

Comparative analysis between sets of enriched bioprocesses in the schistosomula of *S. mansoni, T. regenti*, and *T. szidati* revealed an overlap in 40 PB. Most of them (33 / 40 PB) were common for *Trichobilharzia* species only: "cell communication", "cell migration", "cell fate commitment", "cell population proliferation", "embryo development", "immune system process", "reproduction", and "neurogenesis", etc. At the same time, the remaining processes were either common for one species of *Trichobilharzia* and *S. mansoni* ("post-embryonic animal organ development", "chaperone-mediated protein folding", "RNA polyadenylation"), or for two species of *Trichobilharzia* and one sample of *S. mansoni* schistosomula ("regulation of cellular process", "negative regulation of cellular process", "cellular component morphogenesis", and "regulation of response to stimulus"). There were no overlaps between 3- and 24-hours post-infection schistosomula of *S. mansoni.* Among the 11 common PB for adults "cilium organization" (*F. gigantica*, *P. simillimum*), "cilium movement" (*F. gigantica*, *P. simillimum*), "sperm motility" (*F. gigantica*, *P. simillimum*), and "tissue remodeling" (*F. gigantica*, *F. hepatica*) PB were found. On the contrary, 77 PB were characterized as "enriched" in only one of the considered samples: "skin development" (*F. gigantica*), "cellular response to thyroid hormone stimulus" (*F. gigantica*), "development of primary male sexual characteristics" (*F. hepatica*), "determination of left/right symmetry" (*P. simillimum*), "movement of cell or subcellular component" (*P. simillimum*), "vitellogenesis" (*S. mansoni*), and "response to estrogen" (*S. mansoni*), etc.

The number of gained genes in Digenea ancestor exceeds the number of duplicated ones according to the results of comparison with Platyhelminthes ancestor genome model

The digenean complex life cycle with a sequential alternation of contrast generations is one of the key traits of this group of parasitic flatworms. We can assume that the transition from a simple to a complex life cycle required significant changes both in the genome itself and in its regulation. Ancestral genome models reconstruction and comparative analysis between Platyhelminthes and Digenea might shed light on possible evolutionary transformations of their genomes and the establishment of molecular basis for the increase in life cycles complexity.

We analyzed the results of HOG identification using the pyHAM library (Train, Pignatelli, Altenhoff and Dessimoz, 2019) to build the ancestral genome models. According to the results obtained, the genome model of the Platyhelminthes ancestor included 5952 genes, while the Digenea ancestor genome model included 10372 genes. To clarify the models, we introduced an additional filter by including in the analysis only the HOGs that contained proteins of at least 75 % of the analyzed species (8 / 11 and 11 / 14 for the digenean and Platyhelminthes models, respectively). As a result, the updated models contained 2579 (Platyhelminthes) and 4622 (Digenea) genes. Here, 2258 genes were retained, 64 duplicated, and 1850 gained in the Digenea model compared to Platyhelminthes.

Genes that have been duplicated and gained in the digenean ancestor genome were of particular interest. Given high-quality genome assembly and numerous results of molecular biology research available for *S. mansoni,* we chose the genome of this species as a reference and extracted 2633 and 116 genes that were gained and duplicated, respectively. The total number of *S. mansoni* genes in the models were 3687 for the Platyhelminthes ancestor and 6431 for the ancestor of Digenea. Next, the gene set enrichment of bioprocesses was carried out. Figure 5A shows all the variety of processes in which the gained genes take part. Among them are genes, associated with signaling, cell cycle, development, reproduction, and regulation of gene expression. Duplicated genes are also involved in several biological processes, including those related to animal organ development, female genitalia development, cellular response to different stimulus, phagocytosis, estrogen and fatty acid metabolic process, regulation of various processes and epithelial cell migration (see Figure 5B). All GSEA results for gained and duplicated genes from digenean ancestor genome model are available in the Supplementary Table S5.

Hundreds of potential excretory/secretory proteins were identified

Given digeneans usually exploit several hosts during their complex life cycle, a study of the molecular basis of a host-parasite interaction is important. In our study, we applied the pipeline developed to identify *in silico* potential excretory/secretory proteins (ESP) (Garg and Ranganathan, 2011). As a result, the "classical" and

Fig. 5. The scatter plots for parental bioprocesses enriched in sets of genes gained (A) and duplicated (B) in digenean ancestor genome model. Distances between points represent the similarity between terms, and axes are the first 2 components of applying a PCoA to the similarity matrix. Size of the points represents the scores equal to minus log10 (Fisher's Test p-values).

"non-classical" ESP were found, the main difference between them is that the former have classical N-terminal signal peptides.

Hundreds of ESP were found in the sets of long proteins of all 11 studied trematode species: from 327 (239 "classical" + 88 "non-classical" ESP) in *T. szidati* to 1899 (974 "classical" + 925 "non-classical" ESP) in *P. simillimum*. According to the results of the GSEA analysis, potential "classical" ESP-encoding genes are involved in various processes such as "extracellular structure organization" (*F. gigantica, O. felineus, S. mansoni*), "tissue remodeling" (*F. gigantica, F. hepatica, P. simillimum*), "regulation of chemotaxis" (*F. hepatica, P. simillimum, S. japonicum, S. mansoni*), etc. However only four bioprocesses were enriched in more than half considered digenean species: "proteolysis", "response to endoplasmic reticulum stress", "response to stimulus", and "leukocyte degranulation". In contrast, there was no overlap between different species in the sets of bioprocesses, enriched by genes encoding potential "non-classical" ESP. The lists of potential ESP and enriched bioprocesses are presented in Supplementary Table S6.

The LCS with the "oldest" and "youngest" molecular signatures were identified

In our study, we applied this approach to analyze prepared sets of long proteins from 14 flatworm species. With rare exceptions (one and two proteins in *F. hepatica* and *P. vittatus*, respectively), almost all proteins were successfully distributed across different groups of genes with a common phylogenetic origin, called phylostrata. During the analysis the following phylostrata were identified: 1) "Cellular organisms", 2) "Eukaryota", 3) "Opisthokonta", 4) "Metazoa", 5) "Eumetazoa", 6) "Bilateria", 7) "Protostomia", 8) "Spiralia", 9) "Lophotrochozoa", 10) "Platyhelminthes", 11) "Class", 12) "Order", 13) "Family", 14) "Genus", 15) "Species".

According to the results obtained, in most species (12 / 14), the largest phylostratum is the "Cellular organisms", and in all species it includes at least 20 % of proteins from the considered sets. *Psilotrema simillimum* and *P. vittatus* were two exceptions. In the former, the species-specific phylostratum included 22.37 % of proteins, while "Cellular organisms" included 21.44 %. In *P. vittatus*, the difference between the sizes of the species-specific (38.11 %) and the "Cellular organisms" (20.68 %) phylostrata is almost twofold. The smallest phylostratum in all species was "Spiralia", except for the cases when a particular phylostratum was not distinguished in the species (for example, "Order" or "Genus") due to the limitation of the available data.

The phylostratigraphy results can be used to reveal how the molecular signatures change during the life cycle in terms of phylostrata contribution. One metric to quantify transcriptome conservation on a global scale is the Transcriptome Age Index (TAI) (Domazet-Lošo and Tautz, 2010), which denotes the average transcriptome age throughout the biological process of interest (Drost et al., 2018). In general, a lower TAI value describes an older transcriptome age, whereas a higher TAI denotes a younger one.

Figure 6 shows the variation of TAI through the life cycle of *F. gigantica* (Figure 6A1), *F. hepatica* (Figure 6B1), *P. simillimum* (Figure 6C1), *S. mansoni* (Figure 6D1), *T. regenti* (Figure 6E1), and *T. szidati* (Figure 6F1). For all species, except for *P. simillimum*, significant differences were revealed between the LCS when considering whole molecular signatures. Given the large size of *P. simillimum*-specific phylostratum we decided to rerun the analysis for this species without the latter phylostratum. In this case, the differences between the LCS of the *P. simillimum* became significant.

According to the results obtained, the smallest TAI belong to the eggs of *F. gigantica* (3.63) and *F. hepatica* (3.24), redia of *P. simillimum* (3.59), adult of *S. mansoni* (3.28), and cercariae of *T. regenti* (3.66) and *T. szidati* (3.82). On the contrary, the highest TAI were obtained for cercaria of *F. gigantica* (4.08), 21-days-old juvenile *F. hepatica* (3.59), adult of *P. simillimum* (3.75), 3 hours post-infection schistosomula of *S. mansoni* (3.37), schistosomulum of *T. regenti* (3.74) and *T. szidati* (4.04).

Figure 6 shows the patterns of phylostrata relative expression levels. All phylostrata were divided into two groups: the first one includes phylostrata from "Cellular organisms" to "Digenea", and the second group — from "Order" — to species-specific ones. Significant differences between groups were found only NEJ of *F. hepatica* at 24-hour post excystment, redia of *P. simillimum*, and 24 hours post-infection schistosomula of *S. mansoni*. Most of phylostrata have the lowest relative expression levels on the *F. gigantica* egg (15 / 15 phylostrata), *F. hepatica* egg (14 / 15), adults of *P. simillimum* (13 / 13) and *S. mansoni* (10 / 14), and cercariae of two *Trichobilharzia* species analyzed (14 / 14). The highest relative expression level for most phylostrata was found at metacercaria of *F. gigantica* (14 / 15 phylostrata), juvenile of *F. hepatica* (11 / 15), redia of *P. simillimum* (8 / 13), 3 hours post-infection schistosomulum of *S. mansoni* (11 / 14), and schistosomula of *T. regenti* (14 / 14) and *T. szidati* (14 / 14).

Given the differences between TAI of LCS, we tried to find the biological processes containing the genes with the greatest contribution to TAI. Since the greatest contribution may come from unannotated genes, we focused only on the genes with GO-annotation and selected the top 500 genes with the greatest contribution to the highest and lowest TAI within the life cycle. In all 12 cases, some of the genes are involved in processes associated with mitochondria. At the same time, the lists

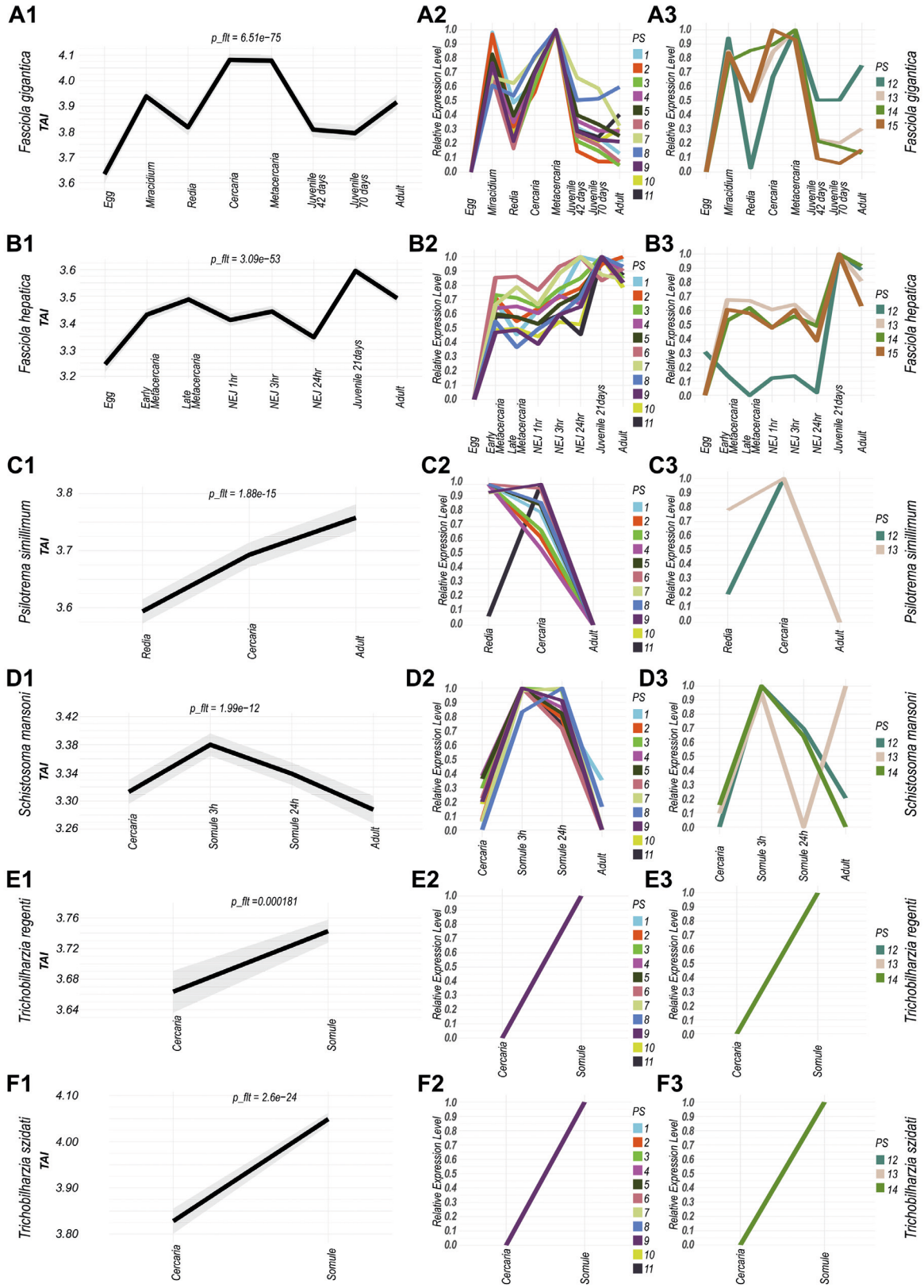

Fig. 6. The Transcriptome Age Indices (TAI) variation and phylostrata relative expression throughout the digenean life cycles. Results are presented for *F. gigantica* (A1–3), *F. hepatica* (B1–3), *P. simillimum* (C1–3), *S. mansoni* (D1–3), *T. regenti* (E1–3), and *T. szidati* (F1–3). A lower TAI value describes an "older" transcriptome, whereas a higher TAI denotes a "younger" one. Phylostrata were divided into two groups: the first one includes phylostrata from "Cellular organisms" to "Digenea" (A2, B2, C2, D2, E2, F2), and the second group — from "Order" — to species-specific ones (A3, B3, C3, D3, E3, F3). For *P. simillimum*, the results are shown without the species-specific phylostratum. Abbreviations: NEJ 1/3/24hr — newly emerged juveniles at 1, 3 and 24 hours post excystment; Somule — schistosomula; Somule 3/24hr — schistosomula at 3 and 24 hours post infection.

contained numerous bioprocesses related to development.

The phylostratigraphy and TAI analysis results are available in the Supplementary Tables S7 and S8, respectively.

A complex phylostratigraphic composition of different groups of genes was identified

Figure 7 shows the phylostratigraphic composition analysis results for genes from the constructed models of ancestral genomes (A), genes with noticeable expression throughout complex life cycles (B), as well as potential "classical" (C) and "non-classical" (D) ESP.

In the digenean ancestor genome model the percentage of genes from the phylostrata "Cellular organisms" was significantly low according to the results obtained: 53.19 % for Platyhelminthes *vs* 42.51 % for Digenea. Larger percentages of genes in the digenean ancestral genome compared to Platyhelminthes were also accounted for phylostrata "Opisthokonta" (5.8 *vs* 7.54 % in Platyhelminthes and Digenea, respectively), "Metazoa" (3.55 *vs* 4.45 %), "Eumetazoa" (2.17 *vs* 4.49 %), "Bilateria" (0.62 *vs* 2.1 %), "Protostomia" (0.46 *vs* 0.75 %), "Spiralia" (0.03 *vs* 0.06 %), "Lophotrochozoa" (0.24 *vs* 0.75 %), "Platyhelminthes" (0.62 *vs* 4.37 %). The proportion of genes attributable to the "Digenea" phylostratum in the reconstructed digenean ancestor genome was 0.9 %. Moreover, the size of each phylostratum in the digenean ancestral model genome is much larger than the corresponding ones in the Platyhelminthes ancestral model genome.

In each of the six analyzed species, more than 55 % of genes with noticeable expression levels at all considered LCS were in the "Cellular organism" and "Eukaryota" (Figure 7B), whereas the proportion of species-specific genes in most species did not exceed 2 %. The only exception was *P. simillimum*, which had approximately 6.5 % of such genes.

We also analyzed the phylostratigraphic composition of the co-expression clusters. All clusters contain genes from several phylostrata. The phylostratigraphic composition of the clusters was various. In most cases (48 / 53), the largest number of co-expressed genes belonged to "Cellular organisms". The contribution of species-specific phylostratum varied widely: from 0 to 15.38 % in *F. gigantica*, from 0.38 to 2.31 % in *F. hepatica*, from 7.88 to 18.17 % in *P. simillimum*, and from 0 to 1.21 % in *S. mansoni*.

The phylostratigraphic composition of the sets of ESP, both "classical" (Figure 7C) and "non-classical" (Figure 7D), differed in all the considered digenean species. We divided all phylostrata into two groups: 1) from "Cellular organism" to "Platyhelminthes", 2) from "Digenea" to "Species". In most species, the first group included most of the "classical" (11 / 11 species) and "non-classical" (8 / 11 species) ESP. The exceptions were *C. sinensis*, *O. viverrini*, and *P. simillimum* species, in which the majority of "nonclassical" ESP belong to the second group of phylostrata. In general, the percentage of "classical" and "nonclassical" ESP in the second group varied from 17.58 (*T. szidati*) to 49.18 % (*P. simillimum*) and from 27.27 (*T. szidati*) to 73.13 % (*C. sinensis*), respectively.

Discussion

According to the results obtained, i) all LCS use one shared genome, and we see significant overlap between the molecular signatures of contrasting LCS in terms of active genes, ii) the sizes of molecular signatures of the complex life cycle stages vary greatly, iii) over-expressed genes often make up a relatively small portion of the molecular signature, iv) numerous genes are co-expressed during the life cycle. However, for a better understanding of the nature of molecular signatures, several key points should be noted. Firstly, the LCS-specific molecular signature is formed by transcriptomes of various cell types. Therefore, the analyzed gene expression is the derivative of this gene expression in different cell types / states. Secondly, a molecular signature is a dynamic system that changes depending on the impact of different conditions. For example, the over-expression of a particular gene can be caused by a response to influences. Therefore, we conclude that the molecular signature of a LCS is a multicomponent system, where the component is not so much the gene itself as its expression under certain conditions. The basis of molecular signature is the expression landscape created by the co-expression of the genes.

We can expect that the closer is the cellular composition of the compared LCS or the more similar are the habitat conditions of LCS, the more similar the molecular signatures should be. Joint clustering of similar LCS in both intraspecific and interspecific comparison corroborates this suggestion. The results obtained for *F. gigantica*, *F. hepatica*, and *P. simillimum* are striking examples. In *F. gigantica*, cercariae and metacercariae are merged into a single cluster, that is, two LCS following each other during the life cycle. Similarly, in *F. hepatica* the early and late metacercariae are grouped with different juveniles that had just left the cysts. In both liver fluke species, the union of juveniles and adults (immature and mature individuals of the amphimictic generation) was distinguished. The clustering of cercariae and redia in *P. simillimum* may be a result of the presence of developing cercariae embryos inside the rediae.

The results of interspecific comparison between similar species also confirm the similarity of its molecular signatures: i) clustering of *F. gigantica*, *F. hepatica*,

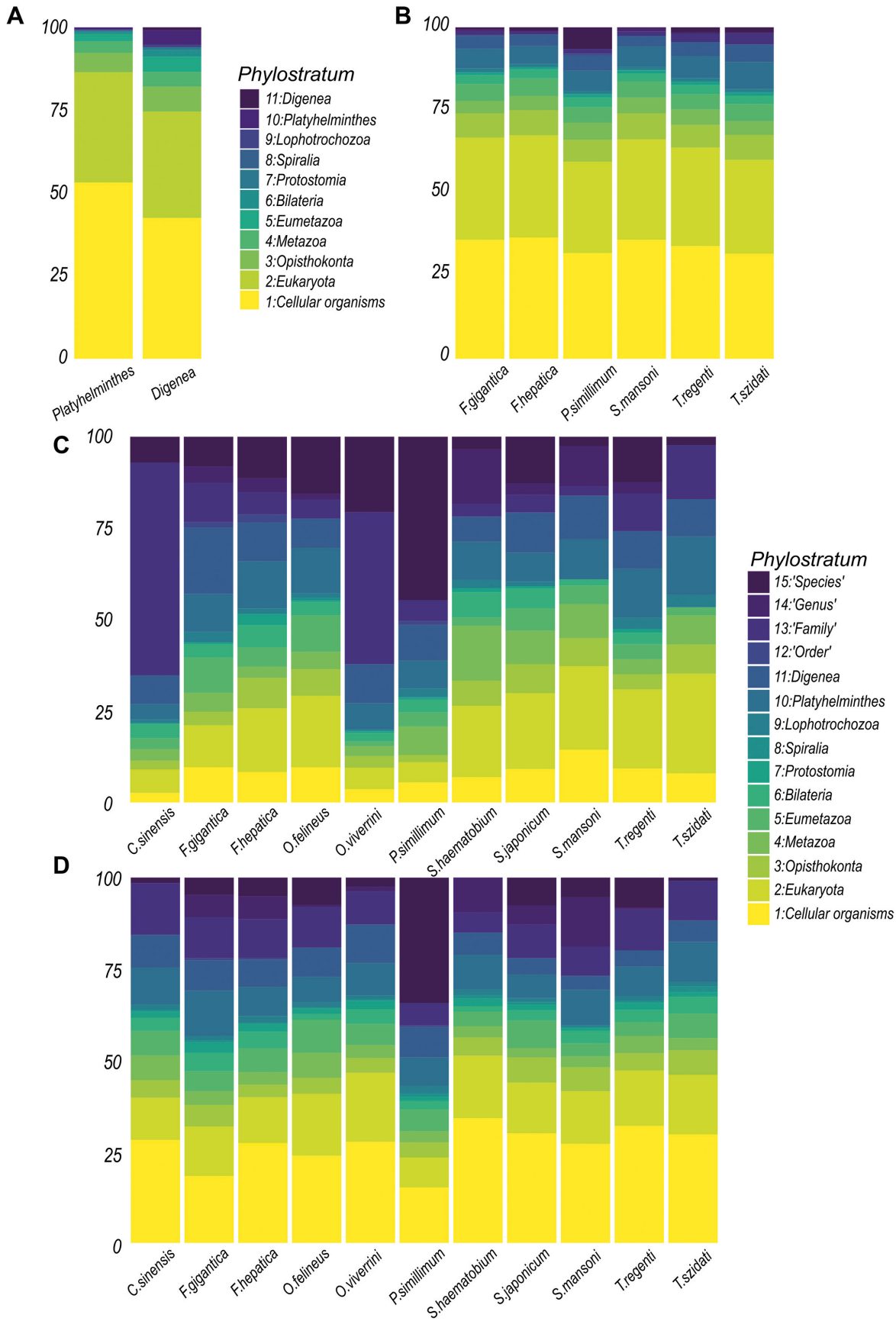

Fig. 7. The phylostratigraphic composition of different gene sets.
The analysis results for the gene sets of the constructed models of ancestral genomes (A), genes with noticeable (≥ 2 TPM) expression throughout complex life cycles (B), as well as genes encoding potential "classical" (C) and "non-classical" (D) excretory/secretory proteins are presented. Numerical values correspond to percentages, and each phylostratum has its own color.

P. simillimum adults and juveniles of *F. hepatica*; ii) miracidia of *F. gigantica*, cercariae of *F. gigantica* and *P. simillimum*, metacercariae of *F. gigantica* and *F. hepatica*, and NEJ of *F. hepatica*; iii) schistosomula of different schistomatids, and iv) cercariae of *T. regenti* and *T. szidati*. However, with the simultaneous analysis of all the species, it seems that there is not enough data to resolve massive clusters combined of numerous LCS. At the same time, clustering of the adults of *F. gigantica* and *P. simillimum*, together with the 70-days-old juvenile of *F. gigantica* and flukes' eggs, may be a result of the egg developmental processes in adults. Similarly, the activity of developmental processes can explain the association of the schistosomula of *Trichobilharzia* and *P. simillimum* redia.

Clustering of cercariae and three hours post-infection schistosomula of *S. mansoni* deserves special attention. All analyzed data on *S. mansoni* were obtained in the same study and belong to the same Puerto Rican strain (Protasio et al., 2012), but each sample contained mixed specimens of both sexes. The significant molecular differences between *Schistosoma* sexes appear already at the cercaria, although phenotypic dimorphism becomes visible only after infection of the vertebrate host (Picard et al., 2016). Thus, the separation of cercarial biological replicates and their clustering with schistosomula may be influenced by molecular differences between sexes of the morphologically similar cercariae.

Comparative analyses between the signature-specific sets of active bioprocesses were also performed. One of the key challenges is to deal with the LCS-specific abundance of active bioprocesses, some of which are interconnected. Given the hierarchical structure of the GO database, we focused on parental bioprocesses, which combine several more private processes at once. According to the results obtained, a relatively small number of bioprocesses are common for similar LCS of different species. However, based on the distinguished common bioprocesses, it is easy to establish the key traits of the LCS biology: the development of miracidia in eggs, energy-consuming muscle-based movement in cercariae, transformation of the metacercariae and schistosomula, tegument changes of juveniles, the formation of male gametes and eggs with miracidia inside the adults. So, for example, the "developmental pigmentation" in an egg may be associated with the development of miracidium eyespot. The activity of "epithelial cell proliferation" and "extracellular matrix organization" in juveniles may be related to tegument changes during their migration in the definitive host's body.

Currently, there is no doubt that the digenean ancestor was a free-living flatworm. But it is still unclear which of the genome changes have led to the rise of such a numerous contrasting LCS. Moreover, the evolution of the digenean life cycles is associated with the evolution of the host-parasite interactions. Nevertheless, modern methods of analysis, such as the analysis of hierarchical groups of orthologs and phylostratigraphy, bring us closer to the understanding of the evolution of this amazing group of parasites. In the recent publication of Zajac et al. (Zajac et al., 2021), the detailed analysis of digenean genomes was carried out and the putative molecular basis of the transition mechanisms to parasitism was revealed. For example, gene functions associated with most duplicated HOGs were host tissue penetration, host behavioral manipulation, and hiding from host immunity through antigen presentation. One of the ways to implement such function is the excreting ESP through the tegument or specialized glands. The results of our analysis indicate that such ESP were involved in bioprocesses, probably associated with the transformation of tegument ("extracellular structure organization", "tissue remodeling"). Moreover, since parasite receives nutrients from the host, the participation of ESP in processes such as "regulation of chemotaxis" and "proteolysis" may be related to nutrients searching and digestion. As one of the examples, digeneans use a diverse array of cathepsin peptidases, which are promising drug targets according to numerous research (Young et al., 2011; Cantacessi et al., 2012; Choudhary et al., 2015; Leontovyč et al., 2016; Li et al., 2016; Zhang et al., 2019).

Following Zajac et al. (Zajac et al., 2021) we conducted our own research to identify the possible molecular basis for the complication of the life cycle. Assuming that accessible genomic and transcriptomic data may be incomplete, the additional condition for ancestor genome models' reconstruction was introduced: a gene should be present in at least 75 % of the considered modern species to be included in the model. We regard this additional condition as sufficient to consider the constructed models reliable. Comparison of digenean and all-Platyhelminthes ancestor genome models indicates that gene appearances and/or duplications are associated with regulation of gene expression, signaling, estrogen and fatty acid metabolic process, and embryonic development.

The digenean life cycle consists of extremely phenotypically different LCS. What mechanism can be used to create several contrasting phenotypes based on the single genome? Polyphenism is one of the possible answers. Minelli and Fusco (Minelli and Fusco, 2010) note that "the hypothesis that a former polyphenism may have been 'internalized' in the course of evolution, resulting in the coexistence of alternative genetically controlled phenotypes as parts of a system of predictable complexity was first suggested long ago (Zakhvatkin, 1949)". In line with this idea, Minelli and Fusco make several key assumptions. First, authors suggest that "from the same beginnings (a polyphenism) a temporally (rather than spatially) consistent, predictable array of phenotypes in

the form of structurally distinct stages within a complex life cycle may have evolved" (Minelli and Fusco, 2010). Second, "life cycle complexification is likely to originate, preferentially, in a less predictable and less stable, perhaps seasonally changing environment" (Minelli and Fusco, 2010). In general, the GSEA results obtained for gained and duplicated genes are consistent with these suggestions. Thus, for example, the complication of the perception of external and internal signals, as well as the subsequent regulation of gene expression via epigenetics and/or long non-coding RNAs, especially in embryogenesis and metamorphosis, could serve as the basis for the creation and consolidation of several contrasting LCS.

One of the main topics in the discussion of the life cycle evolution is a transformation of the reproduction mechanisms. The key question is the cell source of new individuals during the development inside the generations of mother sporocyst and daughter sporocyst/redia. Do they come from primary oocytes or from stem cells? According to Galaktionov and Dobrovolskij (Galaktionov and Dobrovolskij, 2003), i) typical germinal masses morphologically and functionally correspond to the ovaries of mature individuals of hermaphroditic generation; ii) in both germinal masses and ovaries proliferation of non-differentiated cells occur, followed by their physiological maturation which is not accompanied by meiosis; iii) in both ovaries and in germinal masses primary oocytes are formed; iv) formation and differentiation of these cells in parthenitae and hermaphroditic individuals are similar, but their subsequent fate is different; v) early embryogenetic stages of the fertilized ovum of adult worms and germinal cell of parthenitae are almost identical. At the same time, Wang et al. (Wang et al., 2018) conducted a single-cell RNA-seq, which revealed the presence of 4 major stem cell classes in sporocysts. It is also important that the authors note the similarity between schistosome stem cells and the neoblasts that drive regeneration in free-living planarians (Wang et al., 2018). Moreover, the detected heterogeny in schistosome stem cells is also reminiscent of that observed in the planarian neoblasts, and the striking overlap in a group of genes co-regulated between stem cell classes from both organisms was observed (Wang et al., 2018). Based on the results obtained, the authors suggest that sporocyst undergoes asexual clonal expansion to produce new individuals. According to our results, among bioprocesses enriched by genes originated or duplicated in the digenean ancestor, those associated with germ cells, stem cells, reproduction, reproductive system development were found. Given GSEA results, we can assume that genetic innovations in the ancestor of Digenea played a role in both reproduction and differentiation of stem-like cells. Moreover, it seems that the regulation of stem cell division and differentiation really contributed to the implementation of an ancestor life cycle. Nevertheless, it is currently difficult to determine the role of stem-like cells in the formation and activity of the germinal mass of sporocyst / redia-ancestor.

Phylostratigraphy provides a statistical approach for reconstructing macro evolutionary trends. In our study, the phylostratigraphy affiliation to one of the 15 phylostrata from "Cellular organisms" to species-specific ones for almost all long-protein-encoding genes of 14 flatworm species was identified. However, the detection of phylostrata occurs considering the phylogenetic tree and taxonomic affiliation of the analyzed species. Therefore, if the phylogenetic tree used is incomplete and there are no data for the identification of more specific phylostratum, then some of the sequences can be attributed to an incorrect phylostrata. The following two cases can serve as examples. The first one is three fasciolid species (*F. gigantica, F. hepatica* and *P. simillimum*) which were isolated into phylostratum called "Plagiorchiida". The Plagiorchiida La Rue, 1957 was suggested as an order that combined all plagiorchiids and echinostomatids (Olson et al., 2003), which are totally different in terms of morphology and their life cycles. We prefer to consider these flukes belonging to different evolutionary branches and separated orders as was suggested earlier: Diplostomida, Echinostomida, and Plagiorchiida (Odening, 1974). So, the fasciolid species incorporated into our analysis should belong to another phylostratum, which is lacking owing to clearly formal reasons. Regarding the *P. simillimum,* due to a lack of data on other Psilostomatidae species, genes that are common on the family- and genus level were also included in the species-specific phylostratum. The inclusion of a larger number of species in the future analysis will allow refining the results obtained in both the first and second case. Nevertheless, the ability to divide sequences into non-overlapping groups corresponding to the steps on the phylogenetic tree is one of the strengths of phylostratigraphy.

One of the applications of the phylostratigraphy results obtained is the study of the composition of different gene sets. So, for example, we found that the reconstructed ancestors genome models differ in the percentage ratios of all phylostrata. We can conclude that the genes that appeared or duplicated in the ancestor of Digenea originated from genes that appeared at different steps of evolution. A complex phylostratigraphic composition also had gene sets with noticeable expression throughout the life cycle, co-expressed clusters of genes, and sets of genes encoding potential "classical" and "non-classical" ESP. Thus, the complication of the digenean ancestor life cycle (the emergence of new LCS and the transformation of existing ones, the transition to parasitism, and the inclusion of several hosts into the life cycle) required a change in the expression regulation of

different groups of genes, especially those already present in the ancestor of Platyhelminthes.

Nevertheless, the obtained phylostratigraphy results have great potential for studying molecular signatures of LCS. They give the opportunity to study the expression patterns of different phylostrata, as well as to calculate the TAI to compare the LCS and determine the evolutionary "youngest" and "oldest" ones. In our study, statistically significant differences, considering all phylostrata, were found for the LCS of *F. gigantica, F. hepatica, S. mansoni, T. regenti*, and *T. szidati*. With the exclusion of the species-specific phylostratum in *P. simillimum*, which includes genes that are also common for the whole family and genus it belongs to, we also obtained significant differences between LCS. In both *Fasciola* species analyzed, the evolutionary "oldest" LCS was an egg, in *P. simillimum* such a LCS was a redia, and the evolutionary "youngest" in all three Schistosomatidae considered was the strictly specific schistosomulum LCS. However, the results obtained cannot be used to reconstruct the sequence of LCS appearance in the life cycle for the following reasons. First, the molecular signature of LCS in a modern species may significantly differ from those of the corresponding LCS in an ancestral life cycle. Probably the ancestral signature would have some degree of similarity at once with several molecular signatures of the modern different LCS. Second, transcriptomes of important LCS (for example, mother sporocyst) are currently absent, and the sets of available transcriptomes between species differ greatly. Third, the molecular signatures are dynamic systems, which was reflected in TAI variation between the same LCS when the transcriptomes were taken at different time points. The striking examples of such "signature plasticity" are NEJ of *F. hepatica* and *S. mansoni* schistosomula. Fourth, the LCS are characterized by numerous traits. Therefore, we suggest that considering the appearance of a LCS in a life cycle as a single and one-time event is incorrect.

In general, the molecular signature is a "snapshot" of cell types / states transcriptomes that form the LCS considered. If a certain tissue or an organ system has been strongly transformed during evolution, it may affect the relative age of the whole transcriptome. The confirmation to our assumption is the fact that for each of the six LCS with the highest TAI among the genes with the greatest contribution are those that take part in bioprocesses involved in the nervous system development. At the same time, the following should be considered: i) a gene can take part in several processes, ii) the lists of parental bioprocesses were obtained using *H. sapiens* database, iii) the taxon- and LCS-specific biological traits should be considered. However, further investigation of signals from different organ systems in molecular signatures of LCS is of particular interest, especially in view of recent studies of gene gain and loss across the metazoan tree of life. Fernández and Gabaldón (Fernández and Gabaldón, 2020) conclude that ancient gene duplications related to neural activity could be co-opted in a convergent manner to generate a growing neural complexity in animal phyla, while subsequent lineage-specific duplications potentially enabled expanding structural plasticity, neuronal morphology, and connectivity.

In conclusion, we suggest that genetic innovations acquired by the digenean ancestor have allowed complicating the molecular basis of numerous processes, including the regulation of gene expression and embryogenesis. During the complex life cycle, a sequential change of contrasting LCS occurs, and a specific molecular signature can be distinguished for each of them. It is currently difficult to determine the sequence of the LCS appearance in the digenean life cycle, due to the evolution of the LCS themselves. Nevertheless, modern methods of analysis make it possible to delve deeper into the evolution of individual processes and traits that are important for the LCS-specific ontogenies.

Acknowledgements

Data analysis was performed at the Bioinformatics Shared Access Center of the Institute of Cytology and Genetics of the Siberian Branch of the Russian Academy of Sciences. The authors thank Roman Leontovyč for the access to transcriptomic data for *T. regenti* and *T. szidati,* Krystyna Cwiklinski and Rui-Si Hu for the access to *F. gigantica* data. The authors also thank two anonymous reviewers for their constructive comments and valuable suggestions. MAN thanks his relatives and friends for their support and encouragement.

References

Abu-Jamous, B. and Kelly, S. 2018. Clust: automatic extraction of optimal co-expressed gene clusters from gene expression data. *Genome Biology* 19(1):1–11. https://doi.org/10.1186/s13059-018-1536-8

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., von Heijne, G., and Nielsen, H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology* 37(4):420–423. https://doi.org/10.1038/s41587-019-0036-z

Almudi, I., Vizueta, J., Wyatt, C. D. R., de Mendoza, A., Marlétaz, F., Firbas, P. N., Feuda, R., Masiero, G., Medina, P., Alcaina-Caro, A., Cruz, F., Gómez-Garrido, J., Gut, M., Alioto, T. S., Vargas-Chavez, C., Davie, K., Misof, B., González, J., Aerts, S., Lister, R., Paps, J., Rozas, J., Sánchez-Gracia, A., Irimia, M., Maeso, I., and Casares, F. 2020. Genomic adaptations to aquatic and aerial life in mayflies and the origin of insect wings. *Nature Communications* 11(1):1–11. https://doi.org/10.1038/s41467-020-16284-8

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Vesztrocy, A. W., Dalquen, D. A., Müller, S., Telford, M. J., Glover, N. M., Dylus, D., and Dessimoz, C. 2019. OMA standalone: Orthology inference among public and custom genomes and transcriptomes. *Genome Research* 29(7):1152–1163. https://doi.org/10.1101/gr.243212.118

Arendsee, Z., Li, J., Singh, U., Seetharam, A., Dorman, K., and Wurtele, E. S. 2019. Phylostratr: A framework for phylos-

tratigraphy. *Bioinformatics* 35(19):3617–3627. https://doi.org/10.1093/bioinformatics/btz171

Bendtsen, J. D., Jensen, L. J., Blom, N., Von Heijne, G., and Brunak, S. 2004. Feature-based prediction of non-classical and leaderless protein secretion. *Protein Engineering, Design and Selection* 17(4):349–356. https://doi.org/10.1093/protein/gzh037

Brandl, H., Moon, H. K., Vila-Farré, M., Liu, S. Y., Henry, I., and Rink, J. C. 2016. PlanMine — A mineable resource of planarian biology and biodiversity. *Nucleic Acids Research* 44(D1):D764–D773. https://doi.org/10.1093/nar/gkv1148

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. L. 2009. BLAST+: Architecture and applications. *BMC Bioinformatics* 10(1):1–9. https://doi.org/10.1186/1471-2105-10-421

Cantacessi, C., Mulvenna, J., Young, N. D., Kasny, M., Horak, P., Aziz, A., Hofmann, A., Loukas, A., and Gasser, R. B. 2012. A deep exploration of the transcriptome and " excretory / secretory " proteome of adult *Fascioloides magna*. *Molecular & Cellular Proteomics* 11(11):1340–1353. https://doi.org/10.1074/mcp.M112.019844

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., and Huerta-Cepas, J. 2021. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular Biology and Evolution* msab293. https://doi.org/10.1093/molbev/msab293

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Chen, S., Zhou, Y., Chen, Y., and Gu, J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34(17):i884–i890. https://doi.org/10.1093/bioinformatics/bty560

Choudhary, V., Garg, S., Chourasia, R., Hasnani, J. J., Patel, P. V, Shah, T. M., Bhatt, V. D., Mohapatra, A., Blake, D. P., and Joshi, C. G. 2015. Transcriptome analysis of the adult rumen fluke *Paramphistomum cervi* following next generation sequencing. *Gene* 570(1):64–70. https://doi.org/10.1016/j.gene.2015.06.002

Cwiklinski, K., Dalton, J. P., Dufresne, P. J., Course, J. L., Williams, D. J. L., Hodgkinson, J., and Paterson, S. 2015. The *Fasciola hepatica* genome : gene duplication and polymorphism reveals adaptation to the host environment and the capacity for rapid evolution. *Genome Biology* 16(1):1–13. https://doi.org/10.1186/s13059-015-0632-2

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. 2011. ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics* 27(8):1164–1165. https://doi.org/10.1093/bioinformatics/btr088

Domazet-Lošo, T., Brajković, J., and Tautz, D. 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. *Trends in Genetics* 23(11):531–533. https://doi.org/10.1016/j.tig.2007.07.007

Domazet-Lošo, T. and Tautz, D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 468(7325):815–819. https://doi.org/10.1038/nature09632

Drost, H.-G., Gabel, A., Liu, J., Quint, M., and Grosse, I. 2018. MyTAI: Evolutionary transcriptomics with R. *Bioinformatics* 34(9):1589–1590. https://doi.org/10.1093/bioinformatics/btx835

Dylus, D., Nevers, Y., Altenhoff, A. M., Gürtler, A., Dessimoz, C., and Glover, N. M. 2020. How to build phylogenetic species trees with OMA. *F1000Research* 9:511. https://doi.org/10.12688/f1000research.23790.1

Emanuelsson, O., Nielsen, H., Brunak, S., and Von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology* 300(4):1005–1016. https://doi.org/10.1006/jmbi.2000.3903

Fernández, R. and Gabaldón, T. 2020. Gene gain and loss across the metazoan tree of life. *Nature Ecology and Evolution* 4(4):524–533. https://doi.org/10.1038/s41559-019-1069-x

Galaktionov, K. V. and Dobrovolskij, A. A. 2003. The biology and evolution of trematodes. Kluwer Academic Publ., St Petersbug.

Garg, G. and Ranganathan, S. 2011. *In silico* secretome analysis approach for next generation sequencing transcriptomic data. *BMC Genomics* 12(3):1–10. https://doi.org/10.1186/1471-2164-12-S3-S14

Gibson, D. I. 1987. Questions in digenean systematics and evolution. *Parasitology* 95(2):429–460. https://doi.org/10.1007/978-1-4614-3265-4_10

Guindon, S. and Gascuel, O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52(5):696–704. https://doi.org/10.1080/10635150390235520

Howe, K. L., Bolt, B. J., Shafie, M., Kersey, P., and Berriman, M. 2017. WormBase ParaSite — a comprehensive resource for helminth genomics. *Molecular and Biochemical Parasitology* 215:2–10. https://doi.org/10.1016/j.molbiopara.2016.11.005

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., Von Mering, C., and Bork, P. 2019. eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research* 47(D1):D309–D314. https://doi.org/10.1093/nar/gky1085

Katoh, K. and Standley, D. M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30(4):772–780. https://doi.org/10.1093/molbev/mst010

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. L. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology* 305(3):567–580. https://doi.org/10.1006/jmbi.2000.4315

Leontovyč, R., Young, N. D., Korhonen, P. K., Hall, R. S., Bulantová, J., Jeřábková, V., Kašný, M., Gasser, R. B., and Horák, P. 2019. Molecular evidence for distinct modes of nutrient acquisition between visceral and neurotropic schistosomes of birds. *Scientific Reports* 9(1):1–12. https://doi.org/10.1038/s41598-018-37669-2

Leontovyč, R., Young, N. D., Korhonen, P. K., Hall, R. S., Tan, P., Mikeš, L., Kašný, M., Horák, P., and Gasser, R. B. 2016. Comparative transcriptomic exploration reveals unique molecular adaptations of neuropathogenic trichobilharzia to invade and parasitize its avian definitive host. *PLoS Neglected Tropical Diseases* 10(2):e0004406. https://doi.org/10.1371/journal.pntd.0004406

Li, B., McNulty, S. N., Rosa, B. A., Tyagi, R., Zeng, Q. R., Gu, K., Weil, G. J., and Mitreva, M. 2016. Conservation and diversification of the transcriptomes of adult *Paragonimus westermani* and *P . skrjabini*. *Parasites & Vectors* 9(1):497. https://doi.org/10.1186/s13071-016-1785-x

Littlewood, D. T. J. 2006. Parasitic flatworms: molecular biology, biochemistry, immunology and physiology. Cabi Publishing-C a B Int. 480 pp.

Liu, F., Li, Y., Yu, H., Zhang, L., Hu, J., Bao, Z., and Wang, S. 2021. MolluscDB: An integrated functional and evolutionary genomics database for the hyper-diverse animal phylum Mollusca. *Nucleic Acids Research* 49(D1):D988–D997. https://doi.org/10.1093/nar/gkaa918

Martín-Durán, J. M., Ryan, J. F., Vellutini, B. C., Pang, K., and Hejnol, A. 2017. Increased taxon sampling reveals thousands of hidden orthologs in flatworms. *Genome Research* 27(7):1263–1272. https://doi.org/10.1101/gr.216226.116

McNulty, S. N., Tort, J. F., Rinaldi, G., Fischer, K., Rosa, B. A., Smircich, P., Fontenla, S., Choi, Y. J., Tyagi, R., Hallsworth-Pepin, K., Mann, V. H., Kammili, L., Latham, P. S., Dell'Oca, N., Dominguez, F., Carmona, C., Fischer, P. U., Brindley, P. J., and Mitreva, M. 2017. Genomes of *Fasciola hepatica* from the Americas reveal colonization with *Neorickettsia* endobacteria related to the agents of potomac horse and human sennetsu fevers. *PLoS Genetics* 13(1):e1006537. https://doi.org/10.1371/journal.pgen.1006537

Minelli, A. and Fusco, G. 2010. Developmental plasticity and the evolution of animal complex life cycles. *Philosophical Transactions of the Royal Society B* 365(1540):631–640. https://doi.org/10.1098/rstb.2009.0268

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., and Lanfear, R. 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution* 37(5):1530–1534. https://doi.org/10.1093/molbev/msaa015

Mirdita, M., Steinegger, M., and Söding, J. 2019. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* 35(16):2856–2858. https://doi.org/10.1093/bioinformatics/bty1057

Nesterenko, M., Starunov, V., Shchenkov, S., Maslova, A., Denisova, S., Granovich, A., Dobrovolskij, A., and Khalturin, K. 2020. Molecular signatures of the rediae, cercariae and adult worm stages in the complex life cycles of parasitic flatworms (Psilostomatidae, Trematoda). *Parasites & Vectors* 13(1):1–21. https://doi.org/10.1101/580225

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32(1):268–274. https://doi.org/10.1093/molbev/msu300

Odening, K. 1974. Verwandtschaft, System und zyklo-ontogenetische Besonderheiten der Trematoden. *Zoologischer Jahrbucher. Systematik* 101(3):345–396.

Olson, P. D., Cribb, T. H., Tkach, V. V, Bray, R. A., and Littlewood, D. T. J. 2003. Phylogeny and classification of the Digenea (Platyhelminthes: Trematoda). *International Journal for Parasitology* 33(7):733–755. https://doi.org/10.1016/S0020-7519(03)00049-3

Paradis, E. and Schliep, K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* 35(3):526–528. https://doi.org/10.1093/bioinformatics/bty633

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., and Kingsford, C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods* 14(4):417–419. https://doi.org/10.1038/nmeth.4197

Pérez-Ponce de León, G. and Hernández-Mena, D. I. 2019. Testing the higher-level phylogenetic classification of Digenea (Platyhelminthes, Trematoda) based on nuclear rDNA sequences before entering the age of the "next-generation" Tree of Life. *Journal of Helminthology* 93(3):260–276. https://doi.org/10.1017/S0022149X19000191

Picard, M. A. L., Boissier, J., Roquis, D., Grunau, C., Allienne, J.-F., Duval, D., Toulza, E., Arancibia, N., Caffrey, C., Long, T., Nidelet, S., Rohmer, M., and Cosseau, C. 2016. Sex-biased transcriptome of *Schistosoma mansoni*: host-parasite interaction, genetic determinants and epigenetic regulators are associated with sexual differentiation. *PLoS Neglected Tropical Diseases* 10(9):e0004930. https://doi.org/10.1371/journal.pntd.0004930

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., Quail, M. A., Wilson, R. A., Zerlotini, A., Dunne, D. W., and 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. https://doi.org/10.1371/journal.pntd.0001455

Reinhard, E. G. 1957. Landmarks of parasitology I. The discovery of the life cycle of the liver fluke. *Experimental Parasitology* 6(2):208–232. https://doi.org/10.1016/0014-4894(57)90017-6

Shi, L., Derouiche, A., Pandit, S., Rahimi, S., Kalantari, A., Futo, M., Ravikumar, V., Jers, C., Mokkapati, V. R. S. S., Vlahovicek, K., and Mijakovic, I. 2020. Evolutionary analysis of the *Bacillus subtilis* genome reveals new genes involved in sporulation. *Molecular Biology and Evolution* 37(6):1667–1678. https://doi.org/10.1093/molbev/msaa035

Train, C. M., Pignatelli, M., Altenhoff, A., and Dessimoz, C. 2019. IHam and pyHam: Visualizing and processing hierarchical orthologous groups. *Bioinformatics* 35(14):2504–2506. https://doi.org/10.1093/bioinformatics/bty994

Wagner, G. P., Kin, K., and Lynch, V. J. 2013. A model based criterion for gene expression calls using RNA-seq data. *Theory in Biosciences* 132(3):159–164. https://doi.org/10.1007/s12064-013-0178-3

Wang, B., Lee, J., Li, P., Saberi, A., Yang, H., Liu, C., Zhao, M., and Newmark, P. A. 2018. Stem cell heterogeneity drives the parasitic life cycle of *Schistosoma mansoni*. *eLife* 7:e35449. https://doi.org/10.7554/eLife.35449.001

Wang, J., Zhang, L., Lian, S., Qin, Z., Zhu, X., Dai, X., Huang, Z., Ke, C., Zhou, Z., Wei, J., Liu, P., Hu, N., Zeng, Q., Dong, B., Dong, Y., Kong, D., Zhang, Z., Liu, S., Xia, Y., Li, Y., Zhao, L., Xing, Q., Huang, X., Hu, X., Bao, Z., and Wang, S. 2020. Evolutionary transcriptomics of metazoan biphasic life cycle supports a single intercalation origin of metazoan larvae. *Nature Ecology and Evolution* 4(5):725–736. https://doi.org/10.1038/s41559-020-1138-1

Wood, D. E., Lu, J., and Langmead, B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biology* 20(1):1–13. https://doi.org/10.1101/762302

Wood, D. E. and Salzberg, S. L. 2014. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 15(3):1–12. https://doi.org/10.1186/gb-2014-15-3-r46

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., Giron, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J. G., Maurel, T., McDowall, M., McMahon, A., Mohanan, S., Moore, B., Nuhn, M., Oheh, D. N., Parker, A., Parton, A., Patricio, M., Sakthivel, M. P., Abdul Salam, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Sycheva, M., Szuba, M., Taylor, K., Thormann, A., Thread-

gold, G., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Flint, B., Frankish, A., Hunt, S. E., Iisley, G., Kostadima, M., Langridge, N., Loveland, J. E., Martin, F. J., Morales, J., Mudge, J. M., Muffato, M., Perry, E., Ruffier, M., Trevanion, S. J., Cunningham, F., Howe, K. L., Zerbino, D. R., and Flicek, P. 2020. Ensembl 2020. *Nucleic Acids Research* 48(D1):D682–D688. https://doi.org/10.1093/nar/gkz966

Young, N. D., Jex, A. R., Cantacessi, C., Hall, R. S., Campbell, B. E., Spithill, T. W., Tangkawattana, S., Tangkawattana, P., Laha, T., and Gasser, R. B. 2011. A portrait of the transcriptome of the neglected trematode, *Fasciola gigantica* — biological and biotechnological implications. *PLoS Neglected Tropical Diseases* 5(2):e1004. https://doi.org/10.1371/journal.pntd.0001004

Zajac, N., Zoller, S., Seppälä, K., Moi, D., Dessimoz, C., Jokela, J., Hartikainen, H., and Glover, N. 2021. Gene duplication and gain in the trematode *Atriophallophorus winterbourni* contributes to adaptation to parasitism. *Genome Biology and Evolution* 13(3):evab010. https://doi.org/10.1093/gbe/evab010

Zakhvatkin, A. A. 1949. The comparative embryology of the low invertebrates. Sources and method of the origin of metazoan development. Moscow, Soviet Science. (In Russian)

Zambelli, F., Mastropasqua, F., Picardi, E., D'Erchia, A. M., Pesole, G., and Pavesi, G. 2018. RNentropy: An entropy-based tool for the detection of significant variation of gene expression across multiple RNA-Seq experiments. *Nucleic Acids Research* 46(8):e46. https://doi.org/10.1093/nar/gky055

Zhang, X.-X., Cwiklinski, K., Hu, R.-S., Zheng, W.-B., Sheng, Z.-A., Zhang, F.-K., Elsheikha, H. M., Dalton, J. P., and Zhu, X.-Q. 2019. Complex and dynamic transcriptional changes allow the helminth *Fasciola gigantica* to adjust to its intermediate snail and definitive mammalian hosts. *BMC Genomics* 20(1):1–18. https://doi.org/10.1186/s12864-019-6103-5

Zhang, Y., Parmigiani, G., and Johnson, W. E. 2020. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genomics and Bioinformatics* 2(3):lqaa078. https://doi.org/10.1093/nargab/lqaa078