International Journal for Parasitology 53 (2023) 197–205

Contents lists available at ScienceDirect

International Journal for Parasitology

journal homepage: www.elsevier.com/locate/ijpara

Mining various genomic resources to resolve old alpha-taxonomy questions: A test of the species hypothesis of the *Proteocephalus longicollis* species complex (Cestoda: Platyhelminthes) from salmonid fishes ☆

Jan Brabec [a,b,*], Eloïse C. Rochat [c], Rune Knudsen [c,*], Tomáš Scholz [b], Isabel Blasco-Costa [a,c]

[a] Department of Invertebrates, Natural History Museum of Geneva, Geneva, Switzerland
[b] Institute of Parasitology, Biology Centre of the Czech Academy of Sciences, České Budějovice, Czech Republic
[c] Department of Arctic Biology, The Arctic University of Norway, Tromsø, Norway

ARTICLE INFO

Article history:
Received 17 October 2022
Received in revised form 12 December 2022
Accepted 14 December 2022
Available online 24 January 2023

Keywords:
Eucestoda
Proteocephalus
Salmonids
Taxonomy
Mitochondrial genes
Ribosomal RNA

ABSTRACT

High-throughput sequencing strategies became commonly employed to study non-model parasites, but the corresponding genomes and transcriptomes were seldom mined following the original publication. Similar to the data generated with genome skimming techniques based on shallow-depth shotgun genomes, various genomic and transcriptomic resources can be screened for useful molecular phylogenetic markers traditionally characterised with Sanger sequencing. Here, we provide an example of a strategy using reduced-representation genomic as well as transcriptomic data to obtain broad insights into the molecular diversity of the cestode *Proteocephalus longicollis*, a common parasite of salmonids distributed throughout the Holarctic region. We extract popular mitochondrial and nuclear ribosomal markers from various genomic resources for hundreds of parasite specimens from multiple European whitefish populations and compare those with *Proteocephalus* representatives from other species of salmonids and various geographical regions. In contrast with the previous morphology-based assessments, molecular phylogeny reveals a high degree of genetic divergence between *Proteocephalus* isolates from different salmonids, contrastingly low genetic differentiation within the parasite's populations hosted by the European whitefish (*Coregonus lavaretus* species complex), and a sister species relationship of *Proteocephalus* from European whitefish and *Proteocephalus percae*, a parasite of European perch (*Perca fluviatilis*). *Proteocephalus* spp. from North American lake whitefish, brown trout and Arctic charr each formed clearly distinct lineages. These results advance our understanding of the interrelationships of the *Proteocephalus*-aggregate, a well-recognized clade of Holarctic freshwater fish proteocephalids, and support resurrection of some of the nominal species of *Proteocephalus*, including *Proteocephalus exiguus* La Rue, 1911 from North American coregonids and *Proteocephalus fallax* La Rue, 1911 from European *C. lavaretus*, reserving *Proteocephalus longicollis* (Zeder, 1800) exclusively for parasites of *Salmo trutta*.

1. Introduction

High-throughput sequencing (HTS) strategies became an economically accessible and commonplace way of generating data for non-model organisms, mainly thanks to the development of reduced-representation genome sequencing techniques (Davey et al., 2011). Nevertheless, those data are often generated to address rather specific biological problems and come in different kinds, depending on project goals and available budget, ranging from whole genome sequencing to targeting specific genome regions with restriction site-associated DNA sequencing (RADseq; Andrews et al., 2016). Despite becoming popular amongst population ecologists, evolutionary geneticists or conservationists, the breadth and technical specifics of the HTS data pose a barrier to many alpha taxonomists to adopt and merge with those in their toolboxes used to study organismal diversity. Consequently, HTS data are accumulating in public repositories but are seldom mined for goals other than those originally intended. Applicability of HTS data is wide and their adoption not only allows researchers to formulate and test a new generation of biological questions, but

☆ *Note*: Nucleotide sequence data reported in this paper are available in the GenBank under accession numbers **OP971515–21** and **OP972569–71** (*cox1*), and **OP972512–6** (IsrDNA).

* Corresponding author.
E-mail addresses: brabcak@paru.cas.cz (J. Brabec), rune.knudsen@uit.no (R. Knudsen).

https://doi.org/10.1016/j.ijpara.2022.12.005
0020-7519/

mainly to address old questions with unprecedentedly rich data rigour (Rokas, 2016). The field of cestode integrative taxonomy has already seen advances from estimating single gene phylogenies to broader genome-scale phylogenetic inferences, mostly mitogenomics, achieved using genome skimming techniques (Brabec et al., 2016; Trevisan et al., 2019, 2021), but the use of HTS data in cestode taxonomy remains uncommon.

Proteocephalus longicollis (Zeder, 1800) is a common and geographically widespread cestode found across the Holarctic region. Numerous nominal species of *Proteocephalus* were originally described from various genera of salmonid fishes including *Coregonus*, *Oncorhynchus*, *Salmo*, *Salvelinus* and *Thymallus*, foregrounding tapeworms' differences in size and shape of proglottids, and the number of testes. However, these differences were later considered a host-related intraspecific variation because all tapeworms invariably shared species-specific characteristics of presumed taxonomic value, i.e., a similar shape of the scolex and positions of suckers, the shape and size of the vestigial apical sucker, the presence of a large cirrus-sac of similar shape and its relative size (ratio of the cirrus-sac length to the proglottid width), and a well-developed ring-like vaginal sphincter. As a result, most of the species have been synonymised with *P. longicollis* based on the lack of distinctive morphological characters (Scholz and Hanzelová, 1998; Hanzelová and Scholz, 1999). *Proteocephalus longicollis* circulates between two aquatic hosts to complete its life cycle: copepod crustaceans as intermediate and salmonids as definitive hosts. Similar to other proteocephalids, *P. longicollis* does not possess a motile free-swimming ciliated larva (coracidium) and successful transmission thus relies on activity of the intermediate host, copepods, which ingest eggs passively floating in the water (Scholz, 1999).

Together with more than a dozen other species of *Proteocephalus* from Holarctic freshwater teleosts, *P. longicollis* forms a group known as the *Proteocephalus*-aggregate de Chambrier, Zehnder, Vaucher, and Mariaux, 2004, a clade of unresolved relationships amongst its members continuously recovered by multiple studies based on various fragments of nuclear ribosomal or mitochondrial (mt) genes (Zehnder and Mariaux, 1999; Škeříková et al., 2001; Scholz et al., 2007, 2019; de Chambrier et al., 2015). So far, all the analyses had *P. longicollis* represented with limited data, mostly a single representative, and rarely a few representatives from a limited geographical range. Parasites of European whitefish (*Coregonus lavaretus* L. species complex, *C. pollan* Thompson, 1835) and North American lake whitefish (*Coregonus clupeaformis* Mitchill, 1818) have never been directly compared, and specimens from the other genera of salmonid hosts remained molecularly uncharacterised. The species hypothesis, so far based exclusively on the morphology of *P. longicollis*, slender worms with elongate proglottids and euryxenous host specificity, has thus never been tested with molecular data.

Here, we demonstrate a strategy of mining useful data from various sources of HTS data to supplement alpha-taxonomy studies of otherwise opportunistically collected parasites. We reconstruct the interrelationships of the *Proteocephalus*-aggregate, a group of common Holarctic fish cestodes, by filtering novel sequence data from various genomic and transcriptomic projects representing hundreds of parasite individuals from European whitefish and other salmonid fish hosts.

2. Materials and methods

2.1. Sanger sequencing

Sequence data utilized within this study were either generated de novo or retrieved from various resources using bioinformatic strategies as detailed in the following paragraphs. Genomic DNA was extracted with Chelex and 0.1 mg/ml of Proteinase K from the following specimens: i) a newly collected *Proteocephalus* sp. from Arctic charr (*Salvelinus alpinus* L.) in Lake Luktvatn, Norway; ii) *Proteocephalus longicollis* from brown trout (*Salmo trutta* L.) in the Walenbrunnen tributary of Lake Luzern, Switzerland of Rochat et al. (2021); iii) *Proteocephalus fallax* from European whitefish in Lake Walen, Switzerland (specimen W024Pma of Brabec et al., unpublished data). Partial nuclear large subunit ribosomal RNA (lsrDNA) and complete mt cytochrome c oxidase subunit 1 (*cox1*) were amplified following the previously described PCR protocols in de Chambrier et al. (2019) and Sanger sequenced by Macrogen Europe (Amsterdam, Netherlands). Assembly of the contiguous sequences also followed the previously described strategy (e.g., de Chambrier et al., 2019).

2.2. Filtering lsrDNA/cox1 data from third-party genomic resources

The lsrDNA and *cox1* markers are broadly used to discriminate amongst parasitic flatworm species (Blasco-Costa et al., 2016) and represent the most densely sampled markers available for the remaining species of the *Proteocephalus*-aggregate.

2.2.1. Extracting data from reduced-representation genomic reads

Specimens of *Proteocephalus* from *Coregonus* spp. in Europe were collected under the scope of the Brabec et al. (unpublished data) study from several lakes in Switzerland and northern Norway, and molecularly characterised through a reduced-representation genome sequencing technique. More specifically, Brabec et al. (unpublished data) generated double digest RADseq (ddRAD) tags for hundreds of specimens using two relatively frequent restriction enzymes, *Mse*I and *Nla*III, and Illumina HiSeq2500 sequencing. For the current study, we filtered their Illumina raw data for reads corresponding to the *cox1* and lsrDNA loci as follows. The demultiplexed raw Illumina reads were downloaded and quality trimmed with Trimmomatic 0.39 (Bolger et al., 2014) to remove Illumina adapters flanking the restriction sites, as well as any low-quality reads and reads shorter than 60 nucleotides. After trimming, the surviving reads were mapped either on the *cox1* or the lsrDNA Sanger sequenced reference of *Proteocephalus* sp. from Lake Walen, using bowtie2 (Langmead and Salzberg, 2012). Following the assembly, SAMtools (Li et al., 2009) were used to convert the resulting SAM files to BAM files and to generate a consensual sequence for alignment positions of a minimal depth of 6 (scripts and settings utilized can be found through the links provided in Section 2.5). Consensual sequences corresponding to *cox1* and lsrDNA loci were imported in Geneious Prime 2020.0.5 (https://www.geneious.com) and aligned using the translational L-INS-i (*cox1*) and the E-INS-i (lsrDNA) algorithms of MAFFT (Katoh and Standley, 2013), respectively. Resulting alignments were visually inspected, manually corrected for erroneously incorporated missing data within the alignments and aligned to the reference sequence to evaluate the proportion of the sequence recovered from the ddRAD data.

2.2.2. Extracting data from genomes & transcriptome Illumina reads

In addition to the ddRAD tags, Brabec et al. (unpublished data) generated a reference genome sequence of *P. fallax* from *Coregonus* sp. of Lake Léman (Lake Geneva), consisting of 4060 scaffolds. To extract the scaffolds containing the *cox1* and lsrDNA loci, we imported the genome assembly into Geneious and mapped the contigs to the Sanger sequenced reference *cox1* and lsrDNA of the *P. fallax* from Lake Walen using Geneious' Map to Reference built-in tool. Contigs corresponding to a nearly complete mt genome and a nuclear ribosomal operon were extracted and their authenticity checked through: i) an alignment to annotated ortho-

logs available in GenBank; and ii) an open reading frame compatibility with the mt protein-coding genes.

Access to RNA-sequencing data corresponding to raw Illumina MiSeq reads of transcriptomes of further two representatives of the *Proteocephalus*-aggregate (Table 1) was used to complement the data analysed in this study. Transcriptomic data, typically available from the Sequence read archive (SRA) of GenBank, were downloaded and low quality bases and adapters removed with Trimmomatic analogously to the protocol used with the ddRAD data described in the Section 2.2.1 above. Surviving reads were assembled using rnaSPAdes (Bushmanova et al., 2019), resulting nucleotide transcripts imported in Geneious and mapped against the *cox1* and lsrDNA sequences of *P. fallax* from Lake Walen as described above.

2.3. Cox1 haplotype networks of Proteocephalus from European whitefish

The contiguous sequence fragments of the mt *cox1* of sufficient coverage (minimum of six) assembled from the ddRAD data showed variability across *Proteocephalus* individuals from the six lakes sampled by Brabec et al. (unpublished data). To select representatives of the most frequent haplotype per lake using as many data as possible, the *cox1* dataset consisting of 432 individuals was subdivided into six lake-specific datasets. Sites of 75 percent missing data or more were masked and removed from these alignments using Geneious prior to being imported in PopART (Leigh and Bryant, 2015) to identify identical sequences and the longest sequence from the most common haplotype was saved for phylogenetic inference. The intraspecific relationships within *Proteocephalus* from European *Coregonus* spp. were visualised through median-joining network analysis (Bandelt et al., 1999) of the 432 *cox1* sequences calculated in PopART with $\varepsilon = 0$.

2.4. Phylogenetic analyses

The most completely characterised *cox1* sequences retrieved from the ddRAD data representing the most frequent *cox1* haplotype of *Proteocephalus* found in each of the six lakes of Brabec et al. (unpublished data) were aligned together with the complete *cox1* sequences obtained from the screens of the reference genome and transcriptome assemblies, the newly generated sequences with Sanger sequencing and data downloaded from GenBank. To build the lsrDNA dataset, we used the data from the same ddRAD representatives selected according to the most frequent *cox1* haplotype, even though those did not always represent the most complete lsrDNA assembly from the lake. However, as we did not find any nucleotide variation between the partial lsrDNA assemblies from the ddRAD and the complete lsrDNA sequences recovered from the genome/transcriptome data, we included the only two non-identical and complete representative lsrDNA sequences in the final dataset.

Cox1 and lsrDNA datasets were aligned with the translational L-INS-i and E-INS-i algorithms of MAFFT, respectively, trimmed to exclude the PCR primers, and end gaps were coded as missing data. Phylogenetic analyses were carried out under the maximum likelihood criterion in IQ-TREE 2.0.5 (Minh et al., 2020). The corrected Akaike information criterion was used to select the following best-fitting models of nucleotide evolution for the two genes: TIM2 + F + I + G4 (*cox1*) and GTR + F + R2 (lsrDNA) (Kalyaanamoorthy et al., 2017). Nodal supports were estimated through running 1000 non-parametric standard bootstrap replicates.

2.5. Data accessibility

The genomic data used within this study were retrieved from the NCBI BioProject no. PRJNA910576 (ddRAD accessions SAMN32132446–959, reference genome accession SAMN32134074). The resulting gene assemblies, alignments, tree files and scripts were made accessible through Zenodo (10.5281/zenodo.7404234). The newly characterised *cox1* and lsrDNA sequences were deposited in GenBank under accession numbers **OP971515–21**, **OP972512–6**, and **OP972569–71**. Molecular voucher specimens of the newly characterised species of *Proteocephalus* are deposited in the collection of the Natural History Museum of Geneva, Switzerland (Table 1).

3. Results

3.1. Genomic data provide classical molecular markers for alpha-taxonomy

Consensual sequences of complete *cox1* and lsrDNA assembled from the ddRAD data of *Proteocephalus* from European whitefish yielded 1626 bp and 4540 bp long gene sequences, respectively, containing variable proportion of missing data. Average completeness of *cox1* was lower (52.8% missing data in 431 sequences) than lsrDNA (20.5% missing data in 437 sequences) and varied among lakes (see Table 2). Missing data of the selected European whitefish *Proteocephalus* representatives ranged from 3.0 to 41.7% (*cox1*) and 0.1 to 38.4% (lsrDNA) between lakes. Both molecular markers recovered from the reference genome of *P. fallax* were complete and the transcriptomic data allowed assembly of complete lsrDNA and nearly complete *cox1* (0.7 and 3.2% missing data, respectively). A median-joining haplotype network based on the *cox1* data (the lsrDNA sequences assembled from the ddRAD data were found to be identical after exclusion of the sites containing more than 5% of missing data) found three major, relatively distant clusters of sequences, corresponding to the Norwegian lakes Suohpatjávri and Langfjordvatn, and the Swiss lakes (Fig. 1). Haplotypes of specimens from the Norwegian lakes were fully segregated by lake. In the four Swiss lakes, haplotypes were separated by fewer (one to two) mutational steps forming relatively less distant groups, with haplotypes not entirely segregated spatially. Nonetheless, only the three most common haplotypes in the Swiss lakes were present in up to three lakes, either Lakes Bienne, Brienz, and Thun or Lakes Bienne, Thun and Walen, whereas Lakes Brienz and Thun shared up to five haplotypes.

3.2. Multiple species of Proteocephalus parasitise different salmonid hosts

Phylogenetic analysis based on *cox1* data (Fig. 2) showed that *Proteocephalus* from European whitefish forms a well-defined monophyletic lineage of closely related populations, despite the wide latitudinal spread of the specimens collected. Within this group, representatives from the Norwegian Lake Langfjordvatn formed the earliest branching lineage and the Norwegian Lake Suohpatjávri was sister to the lineage composed of representatives of the Swiss lakes. The lsrDNA (Fig. 3) data were found much more conserved, resulting in a phylogram with notably shorter internal branches and overall, less-resolved topology. *Proteocephalus percae* (Müller, 1780) formed a sister lineage to the *Proteocephalus* from European whitefish (although well-supported only by the *cox1* data), while the representatives of the *Proteocephalus* from North American lake whitefish *C. clupeaformis* and other North American fishes formed an independent lineage. *Proteocephalus* representatives from Arctic charr and brown trout formed sister lineages sup-

Table 1
List of specimens evaluated phylogenetically within this study, with sampling details and GenBank accession numbers.

Species	Host	Country, Locality, Voucher, Specimen no.[a]	lsrDNA	*cox1*	Original Identification, Reference
Glanitaenia osculata	*Silurus glanis*	Switzerland	KX768937	KX768943	
Proteocephalus exiguus	*Micropterus dolomieu*	Canada	MN061862	MN061850	*P. longicollis* (Scholz et al., 2019)
Proteocephalus exiguus	*Coregonus clupeaformis*	USA	MN061863	MN061851	*P. longicollis* (Scholz et al., 2019)
Proteocephalus exiguus	*Sander vitreus*	USA	MN061864	MN061852	*P. longicollis* (Scholz et al., 2019)
Proteocephalus fallax	*Coregonus lavaretus*	Norway, Lake Langfjordvatn, MHNG-PLAT- 0144231, L071Pm4	Zenodo[d]	Zenodo[d]	
Proteocephalus fallax	*Coregonus lavaretus*	Norway, Lake Suohpatjávri, MHNG-PLAT- 0144232, S060Pmb	Zenodo[d]	Zenodo[d]	
Proteocephalus fallax	*Coregonus confusus*	Switzerland, Lake Bienne, MHNG-PLAT-0144234, N007Pxa	Zenodo[d]	Zenodo[d]	
Proteocephalus fallax	*Coregonus alpinus*	Switzerland, Lake Brienz, MHNG-PLAT-0144235, B046Pxc	Zenodo[d]	Zenodo[d]	
Proteocephalus fallax	*Coregonus* sp.	Switzerland, Lake Léman, MHNG-PLAT-0144236, Ex67-4	**OP972514**	**OP972569**	
Proteocephalus fallax	*Coregonus* sp.	Switzerland, Lake Neuchâtel, MHNG-PLAT-0144237, NEU003	**OP972513**[c]	**OP972571**[c]	
Proteocephalus fallax	*Coregonus acrinasus*	Switzerland, Lake Thun, MHNG-PLAT-0144238, T002Pxd	Zenodo[d]	Zenodo[d]	
Proteocephalus fallax	*Coregonus duplex*	Switzerland, Lake Walen, MHNG-PLAT-0144240, W014Pxe	Zenodo[d]	Zenodo[d]	
Proteocephalus fallax	*Coregonus heglingus*	Switzerland, Lake Walen, MHNG-PLAT-0144241, W024Pma	**OP972515**[b]	**OP971515**[b]	
Proteocephalus longicollis	*Salmo trutta*	Switzerland, Walenbrunnen Stream, MHNG-PLAT-0144243, Eex05-03	MT738715	**OP971516**[b]	*Proteocephalus* sp. (Rochat et al., 2021)
Proteocephalus sp.	*Salvelinus alpinus*	Norway, Lake Luktvatn, MHNG-PLAT-0144245, Eex28-06	**OP972516**[b]	**OP971521**[b]	
Proteocephalus demshini	*Barbatula toni*	Russia	KX768942	KX768950	
Proteocephalus filicollis	*Gasterosteus aculeatus*	United Kingdom	AJ388636	n/a	
Proteocephalus fluviatilis	*Micropterus dolomieu*	Japan	KP729390	KX768945	
Proteocephalus gobiorum	*Neogobius fluviatilis*	Ukraine	KP729393	KX768944	
Proteocephalus lucipercae	*Sander vitreus*	USA	MN061853	MN061841	
Proteocephalus lucipercae	*Sander vitreus*	Canada	MN061855	MN061843	
Proteocephalus lucipercae	*Sander vitreus*	Canada	MN061856	MN061844	
Proteocephalus macrocephalus	*Anguilla anguilla*	Czech Republic	AJ388609	n/a	
Proteocephalus macrocephalus	*Anguilla anguilla*	United Kingdom	EF095261	JQ268552	
Proteocephalus midoriensis	*Lefua echigonia*	Japan	AJ388610	n/a	
Proteocephalus misgurni	*Misgurnus anguillicaudatus*	Russia	KX768941	KX768949	
Proteocephalus percae	*Perca fluviatilis*	Switzerland, Lake Neuchâtel, MHNG-PLAT-36744, Ex79-1	n/a	**OP971517**[b]	
Proteocephalus percae	*Perca fluviatilis*	Switzerland, Lake Morat, MHNG-PLAT-36745, Ex79-2	n/a	**OP971518**[b]	
Proteocephalus percae	*Perca fluviatilis*	Switzerland, Lake Léman, MHNG-PLAT-54160, Ex79-3	n/a	**OP971519**[b]	
Proteocephalus percae	*Perca fluviatilis*	Switzerland, Lake Léman, MHNG-PLAT-63395, Ex79-4	n/a	**OP971520**[b]	
Proteocephalus percae	*Perca fluviatilis*	Switzerland	AJ388594	KX768947	
Proteocephalus pearsei	*Perca flavescens*	USA	MN061857	MN061845	
Proteocephalus pearsei	*Esox niger*	USA	MN061858	MN061846	
Proteocephalus pinguis	*Esox lucius*	USA	MN061859	MN061847	
Proteocephalus pinguis	*Esox lucius*	USA	MN061860	MN061848	
Proteocephalus pinguis	*Esox lucius*	USA	MN061861	MN061849	
Proteocephalus plecoglossi	*Plecoglossus altivelis*	Japan	AJ388606	KX768946	
Proteocephalus sagittus	*Barbatula barbatula*	Czech Republic	KP729391	KX768948	
Proteocephalus tetrastomus	*Hypomesus nipponensis*	Japan	AJ388635	n/a	
Proteocephalus tetrastomus	*Hypomesus nipponensis*	Japan, JP244a	**OP972512**[c]	**OP972570**[c]	

[a] Detailed locality, molecular voucher and specimen codes are provided only for newly sequenced specimens. MHNG, Muséum d'histoire naturelle de Genève, Switzerland.
[b] Sequences characterised using Sanger sequencing (see Section 2 for details).
[c] Sequences obtained from transcriptomic data.
[d] Gene assemblies derived from ddRAD data are available from Zenodo (10.5281/zenodo.7404234).

Table 2
Average completeness of the *cox1* (1626 bp) and lsrDNA (4540 bp) markers recovered from the ddRAD data shown as percentage of mean missing data in representatives from Norwegian (Langfjordvatn, Suohpatjárvi) and Swiss (Bienne, Brienz, Thun, Walen) lake populations of *Proteocephalus* from European whitefish. Numbers in brackets show numbers of individuals.

Lake	Bienne	Brienz	Langfjordvatn	Suohpatjárvi	Thun	Walen
cox1 (431)	48.8 (30)	51.6 (83)	63.5 (82)	54.4 (90)	46.3 (94)	49.0 (52)
lsrDNA (437)	19.8 (33)	18.3 (75)	23.6 (94)	27.5 (95)	10.4 (87)	22.4 (53)

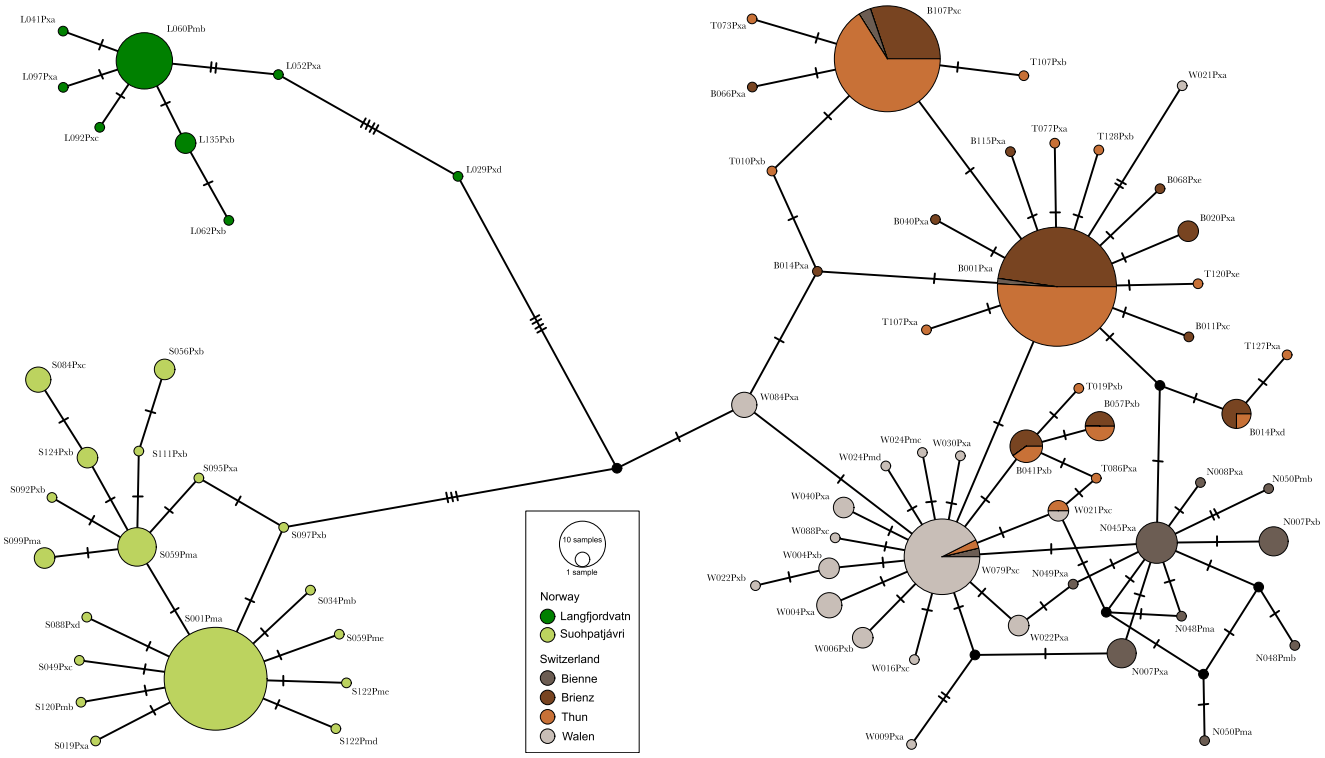

Fig. 1. Intraspecific relationships within *Proteocephalus* from European whitefish (*Coregonus* spp.) visualised through a median-joining network in PopART. Hash marks represent mutational steps, the size of each circle is proportional to the *cox1* haplotype frequency. Single haplotype code initials represent the following lakes: B, Brienz; L, Langfjordvatn; N, Bienne; S, Suohpatjárvi; T, Thun; W, Walen.

ported by a bootstrap of 91 in the *cox1* and 57 in the lsrDNA analysis, independent and relatively basal to the two lineages from whitefish and *P. percae*. The phylogenetic position of *Proteocephalus tetrastomus* (Rudolphi, 1810) was inferred inconsistently between the two datasets: *cox1* data resolved it (although without statistical support) as sister to *Proteocephalus* from North American lake whitefish, while the lsrDNA grouped it with *Proteocephalus plecoglossi* Yamaguti, 1934, in the phylogenetic position sister to all the proteocephalids of salmonids plus *P. percae*. The *cox1* tree found only *P. plecoglossi* to occupy this position. The remaining representatives of the *Proteocephalus*-aggregate all represented by data downloaded from GenBank were placed basally relative to the proteocephalids of salmonids, *P. percae*, *P. plecoglossi* and *P. tetrastomus* and the topology of this part of the tree was inconsistent between the *cox1* and lsrDNA trees (see Figs. 2 and 3).

4. Discussion

By gathering molecular data from an unprecedently large sample of proteocephalids parasitizing European whitefish, as well as representatives of other salmonids, we were able to gain a more detailed understanding of the species diversity and interrelationships of the *Proteocephalus*-aggregate group, common parasites of freshwater fishes of the Holarctic region. The history of species descriptions of proteocephalids from Holarctic freshwater fishes is convoluted and partially speculative given the fact that type specimens of most species described before 1911 do not exist, the early species descriptions provided insufficient morphological details and some species descriptions are erroneous (overviewed by Scholz and Hanzelová, 1998). For example, several authors have based their descriptions of *P. longicollis* on proteocephalid specimens isolated from smelt (*Osmerus eparlanus* L.; Osmeriformes), a known host of another *Proteocephalus*-aggregate representative, *P. tetrastomus*. While the rare presence of *P. longicollis* in the smelt host was confirmed by morphological observations of Scholz and Hanzelová (1998), who studied several specimens from fish collected in Russia and found worms morphologically identical with *P. longicollis* from coregonids, molecular data representative of *P. longicollis* from smelt remain to be generated.

Commonly used mitochondrial and nuclear ribosomal markers represented here with *cox1* and lsrDNA, respectively, support the existence of multiple species of proteocephalids parasitising different salmonid hosts. Adding to this elevated species complexity, the European and North American lake whitefishes each host their own phylogenetically close but unrelated species of *Proteocephalus*. This finding contrasts significantly with the conclusions of the morphological treatments of these parasites, most notably with the detailed works of Scholz and Hanzelová (1998) and Hanzelová and Scholz (1999), who synonymised most of the nom-

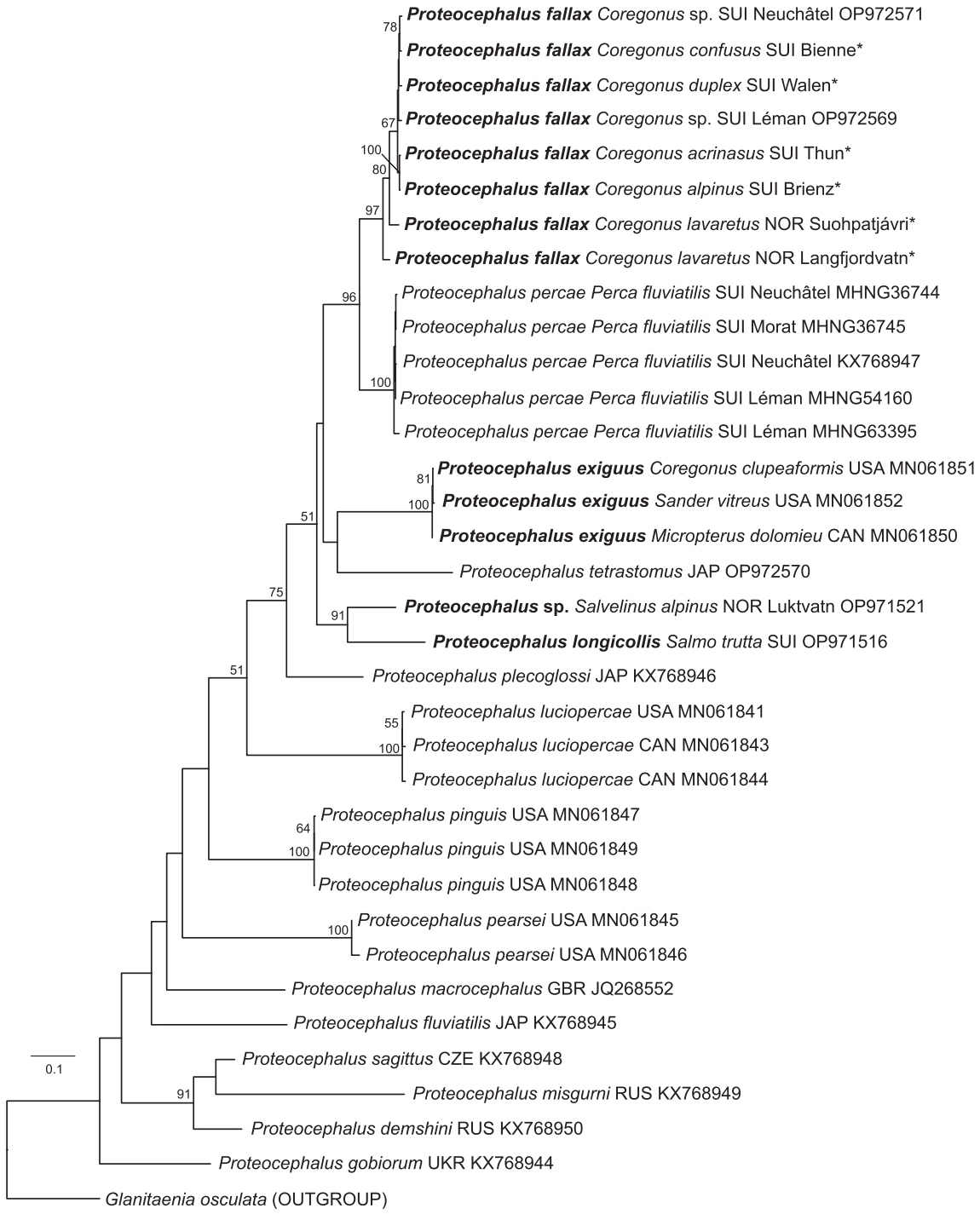

Fig. 2. Phylogenetic interrelationships of species of *Proteocephalus* parasitising salmonids and their position within the *Proteocephalus*-aggregate group. Maximum likelihood estimate based on *cox1* data analysed as a single partition. Parasites of salmonids are highlighted in bold. Note that the operational taxonomic units marked with asterisks represent multiple specimens of identical sequence. Branch length scale bar represents the number of substitutions per site. Nodal values show non-parametric standard bootstrap support (1000 replicates) above 50.

inal species of *Proteocephalus* from various salmonid hosts, including species of *Coregonus*, *Oncorhynchus*, *Salmo* and *Salvelinus*, with *P. longicollis*. Based on our molecular data, we argue that some of the nominal species of *Proteocephalus* should be resurrected.

Following the first original species description of a proteocephalid from European whitefish, *P. fallax* collected by (La Rue, 1911) from *Coregonus fera* Jurine, 1825 in Lake Luzern, we suggest that the European whitefish isolates from Norwegian and Swiss lakes (Table 1) represent *P. fallax*. The specimen of *Proteocephalus* from *Salmo trutta* sampled by Rochat et al. (2021) in a tributary of Lake Luzern most likely represents a conspecific species that Zeder (1800) sampled from the same host and we propose to refer to it as *P. longicollis* (sensu stricto). The proteocephalids found in North American lake whitefish as well as other freshwater fishes in North America also represent a lineage genetically distinct from the European whitefish or other salmonid fishes of the Palaearctic region (Figs. 2 and 3) and we consider those to represent *P. exiguus*, the species originally described from *Coregonus nigripinnis* (Milner,

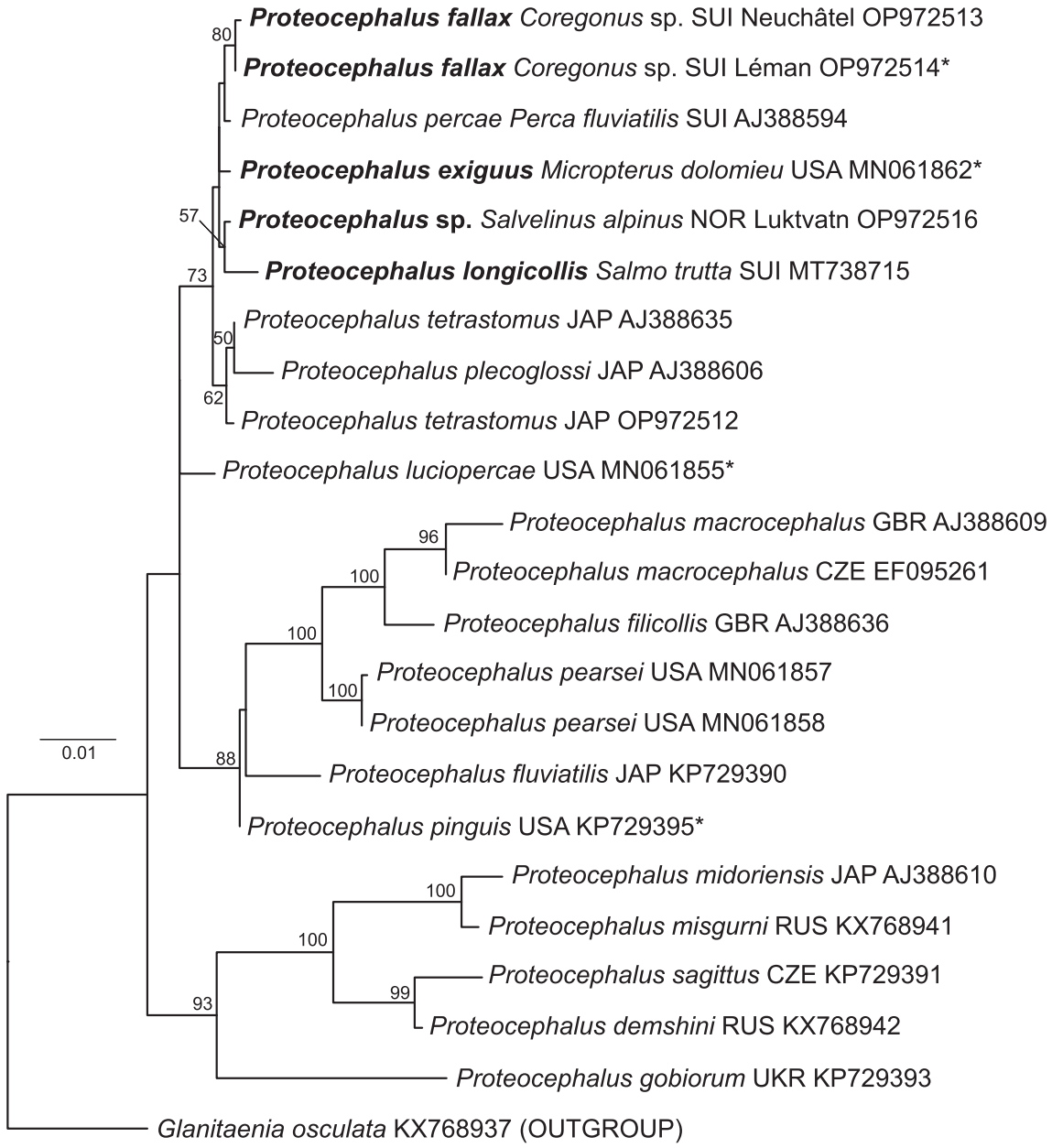

Fig. 3. Phylogenetic interrelationships of species of *Proteocephalus* parasitising salmonids and their position within the *Proteocephalus*-aggregate group. Maximum likelihood estimate based on lsrDNA data. Parasites of salmonids are highlighted in bold. Note that the operational taxonomic units marked with asterisks represent multiple specimens of identical sequence. Branch length scale bar represents the number of substitutions per site. Nodal values show non-parametric standard bootstrap support (1000 replicates) above 50.

1874) in Lake Michigan, USA, by La Rue in 1911. Dubinina (1952) was able to differentiate *P. exiguus* as a common parasite of salmonids and *P. longicollis* as a parasite of smelt (*O. eperlanus*) and grayling (*Thymallus* spp.). Both *P. fallax* and *P. exiguus* were morphologically characterised in detail by Scholz and Hanzelová (1994) and Hanzelová et al. (1995). Our data also suggest that at least one further species-level lineage of *Proteocephalus* hosted by Arctic charr exists in Europe, represented by a single specimen molecularly characterised within this study. Based on the lack of material suitable for morphological characterisation of this taxon, however, we refrain from suggesting a species name for the *Proteocephalus* isolate from *S. alpinus* and retain the specimen as undetermined *Proteocephalus* sp. Given the high degree of polymorphism of *Proteocephalus* spp. reported in previous studies (Scholz and Hanzelová, 1998; Hanzelová and Scholz, 1999), an accurate delineation of species would require a thorough morphological evaluation of proteocephalids based on multiple specimens from each of its salmonid hosts together with its molecular data. The current study did not span such a wealth of specimens from the individual salmonid hosts to revise the tapeworms' species delineations at present.

Within the *Proteocephalus*-aggregate, two further nominal species currently considered valid, *P. percae* and *Proteocephalus thymalli* (Annenkova-Chlopina, 1923), parasites of perch and grayling, respectively, differ only inconspicuously in their morphology from *P. longicollis*. The three species share some prominent morphological characteristics including a well-developed vaginal sphincter and a thick, relatively long cirrus sac, rendering the shape of the scolex (along the parasite's host spectrum) the only morphological trait differentiating it from *P. longicollis* (Scholz

et al., 2007). While the lsrDNA phylogeny presented here reflects well the morphological similarity of *P. longicollis* and *P. percae* seen by the former studies (noting the fact that the shared morphological traits are diagnostic characters of the group uniting proteocephalids of salmonids plus *P. percae*, instead of a single species), the lack of homologous molecular data for *P. thymalli* prevented us from evaluating the validity of this species. Scholz et al. (2007) were the only ones to obtain some molecular data of *P. thymalli* and retained it provisionally as valid. The similarity of the internal transcribed spacer 2 (ITS2) sequence of *P. thymalli* to *Proteocephalus ambiguus* (Dujardin, 1845), a morphologically distinct parasite of nine-spined sticklebacks and the type species of the genus, was also treated with caution by Scholz et al. (2007) until more molecular data of the two species are available, a problem that waits to be addressed. Future applications of HTS data such as ddRAD or other '-omics' approaches will permit both the use of classical markers such as *cox1*, but also of the genome-wide data, to delimit species of proteocephalids in salmonids within the *Proteocephalus*-aggregate group. Together with a re-evaluation of the morphological evidence, these efforts will streamline resolution of the true diversity and host specificity of this taxonomically complex group of tapeworms and set the stage for other helminths.

The application of molecular approaches to both model and non-model parasitic species often lay decades behind those of free-living species (Selbach et al., 2019) and we feel we need to keep encouraging parasite taxonomists to adopt genomic and other HTS approaches to advance the knowledge in our field. This study showed that previously generated reduced-representation genomic data allow retrieval of popular mitochondrial and nuclear ribosomal markers typically used to delimit species boundaries in helminths (Blasco-Costa et al., 2016), thus permitting comparison of molecular data generated across taxonomical and geographical scales, and time periods. More importantly, we revealed that it is possible to benefit from specimens processed through a range of HTS approaches and compare them with data available in GenBank to address fundamental questions at the intersection between species and populations with unprecedented genome-wide breadth (Brabec et al., unpublished data), as well as to delimit species and address long-standing taxonomic questions.

In summary, we describe a strategy using various HTS data to inform long-standing taxonomical problems. We show how to extract a wealth of data from previously generated reduced-representation genomic datasets, as well as other genomic resources, and integrate those with classical molecular markers to advance our understanding of organismal taxonomy, taking a broadly distributed group of cestodes of salmonids as an example. Contrary to morphology-based interpretations (Scholz and Hanzelová, 1998; Scholz et al., 2007), our results show that specificity of *P. longicollis* s.l. for the definitive host is high in the Palaearctic region. Each of the salmonid fish genus included in this study hosts a genetically distinct species-level lineage of *Proteocephalus*, all of which belong to the *Proteocephalus*-aggregate, group formed by parasites of Holarctic freshwater fishes. We propose to resurrect *P. exiguus* from Lake whitefish in North America and *P. fallax* from European whitefish. *Proteocephalus longicollis* s. s. should be used for the specimens from *Salmo trutta*, the original type host. According to the *cox1* and lsrDNA data, proteocephalids of salmonids, together with *P. percae*, parasite of perch, form the most derived lineage of the *Proteocephalus*-aggregate group.

Acknowledgements

This work received funding from the Swiss National Science Foundation (SNSF grant 31003A_169211 to I.B.-C. and J.M.). We would like to express our gratitude to Mia Delacombaz, Pierre Rizzolo and Janik Pralong from the Natural History Museum of Geneva, Switzerland (MHNG) for their help with fish dissections and to Alain de Chambrier and Jean Mariaux (both MHNG) for informal discussions and comments.

References

Andrews, K.R., Good, J.M., Miller, M.R., Luikart, G., Hohenlohe, P.A., 2016. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat. Rev. Genet. 17, 81–92.

Bandelt, H., Forster, P., Röhl, A., 1999. Median-joining networks for inferring intraspecific phylogenies. Mol. Biol. Evol. 16, 37–48.

Blasco-Costa, I., Cutmore, S.C., Miller, T.L., Nolan, M.J., 2016. Molecular approaches to trematode systematics: 'best practice' and implications for future study. Syst. Parasitol. 93, 295–306.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Brabec, J., Kuchta, R., Scholz, T., Littlewood, D.T.J., 2016. Paralogues of nuclear ribosomal genes conceal phylogenetic signals within the invasive Asian fish tapeworm lineage: evidence from next generation sequencing data. Int. J. Parasitol. 46, 555–562.

Bushmanova, E., Antipov, D., Lapidus, A., Prjibelski, A.D., 2019. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. GigaScience 8, giz100.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., Blaxter, M.L., 2011. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat. Rev. Genet. 12, 499–510.

de Chambrier, A., Zehnder, M., Vaucher, C., Mariaux, J., 2004. The evolution of the Proteocephalidea (Platyhelminthes, Eucestoda) based on an enlarged molecular phylogeny, with comments on their uterine development. Syst. Parasitol. 57, 159–171.

de Chambrier, A., Waeschenbach, A., Fisseha, M., Scholz, T., Mariaux, J., 2015. A large 28S rDNA-based phylogeny confirms the limitations of established morphological characters for classification of proteocephalidean tapeworms (Platyhelminthes, Cestoda). Zookeys 500, 25–59.

de Chambrier, A., Brabec, J., Tran, B.T., Scholz, T., 2019. Revision of *Acanthotaenia* von Linstow, 1903 (Cestoda: Proteocephalidae), parasites of monitors (*Varanus* spp.), based on morphological and molecular data. Parasitol. Res. 118, 1761–1783.

Dubinina, M.N., 1952. Some remarks on the classification of tapeworm family Proteocephalidae La Rue and their distribution in the USSR. Parazitologicheskii Sbornik Zoologicheskogo Instituta AN SSSR 14, 281–302.

Hanzelová, V., Scholz, T., Fagerholm, H.-P., 1995. Synonymy of *Proteocephalus neglectus* La Rue, 1911, with *P. exiguus* La Rue, 1911, two fish cestodes from the Holarctic Region. Syst. Parasitol. 30, 173–185.

Hanzelová, V., Scholz, T., 1999. Species of *Proteocephalus* Weinland, 1858 (Cestoda: Proteocephalidae), Parasites of coregonid and salmonid fishes from North America: taxonomic reappraisal. J. Parasitol. 85, 94–101.

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

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. Mol. Biol. Evol. 30, 772–780.

La Rue, G.R., 1911. A revision of the cestode family Proteocephalidae. Zool. Anz. 38, 473–482.

Langmead, B., Salzberg, S., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359.

Leigh, J.W., Bryant, D., 2015. PopART: Full-feature software for haplotype network construction. Methods Ecol. Evol. 6, 1110–1116.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Subgroup, 1000 Genome Project Data Processing, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37, 1530–1534.

Rochat, E.C., Brodersen, J., Blasco-Costa, I., 2021. Conspecific migration and environmental setting determine parasite infracommunities of non-migratory individual fish. Parasitology 148, 1057–1066.

Rokas, A., 2016. Perspective: Systematics in the age of genomics. In: Olson, P., Hughes, J., Cotton, J. (Eds.), Next Generation Systematics (Systematics Association Special Volume 85). Cambridge University Press, Cambridge, pp. 219–228.

Scholz, T., 1999. Life cycles of species of *Proteocephalus*, parasites of fishes in the Palearctic Region: a review. J. Helminthol. 73, 1–19.

Scholz, T., Hanzelová, V., 1998. Tapeworms of the genus *Proteocephalus* Weinland, 1858 (Cestoda: Proteocephalidae), parasites of fishes in Europe. Studie AV ČR, No. 2/98. Academia, Prague, Czech Republic.

Scholz, T., Hanzelová, V., 1994. Taxonomic study of two *Proteocephalus* species (Cestoda: Proteocephalidae) parasitizing coregonid fish: synonymization of *P. fallax* with *P. exiguus*. Syst. Parasitol. 27, 1–12.

Scholz, T., Hanzelová, V., Škeříková, A., Shimazu, T., Rolbiecki, L., 2007. An annotated list of species of the *Proteocephalus* Weinland, 1858 aggregate sensu de Chambrier et al. (2004) (Cestoda: Proteocephalidea), parasites of fishes in the

Palaearctic Region, their phylogenetic relationships and a key to their identification. Syst. Parasitol. 67, 139–156.
Scholz, T., Choudhury, A., Uhrová, L., Brabec, J., 2019. The *Proteocephalus* species-aggregate in freshwater centrarchid and percid fishes of the Nearctic Region (North America). J. Parasitol. 105, 798–812.
Selbach, C., Jorge, F., Dowle, E., Bennett, J., Chai, X., Doherty, J.-F., Eriksson, A., Filion, A., Hay, E., Herbison, R., Lindner, J., Park, E., Presswell, B., Ruehle, B., Sobrinho, P. M., Wainwright, E., Poulin, R., 2019. Parasitological research in the molecular age. Parasitology 146, 1361–1370.
Škeříková, A., Hypša, V., Scholz, T., 2001. Phylogenetic analysis of European species of *Proteocephalus* (Cestoda: Proteocephalidea): compatibility of molecular and morphological data, and parasite-host coevolution. Int. J. Parasitol. 31, 1121–1128.
Trevisan, B., Alcantara, D.M.C., Machado, D.J., Marques, F.P.L., Lahr, D.J.G., 2019. Genome skimming is a low-cost and robust strategy to assemble complete mitochondrial genomes from ethanol preserved specimens in biodiversity studies. PeerJ 7, e7543.
Trevisan, B., Machado, D.J., Lahr, D.J.G., Marques, F.P.L., 2021. Comparative characterization of mitogenomes from five orders of cestodes (Eucestoda: tapeworms). Front. Gene. 12, 788871.
Zeder, J.G.H., 1800. Erster Nachtrag zur Naturgeschichte der Eingeweidewürmer von Johann August Ephraim Goeze. Leipzig.
Zehnder, M.P., Mariaux, J., 1999. Molecular systematic analysis of the order Proteocephalidea (Eucestoda) based on mitochondrial and nuclear rDNA sequences. Int. J. Parasitol. 29, 1841–1852.