Revisiting the phylogenetic position of *Caullerya mesnili* (Ichthyosporidia), a common *Daphnia* parasite, based on 22 protein-coding genes

Yameng Lu ^{1,2,*}, Eduard Ocaña-Pallarès ^{3,*}, David López-Escardó ^{3,4}, Stuart R. Dennis ²,
Michael T. Monaghan ^{1,5,6}, Iñaki Ruiz-Trillo ^{3,7,8}, Piet Spaak ², Justyna Wolinska ^{1,6}

*these authors contributed equally to this work

¹Leibniz Institute of Freshwater Ecology and Inland Fisheries (IGB), Berlin, Germany;

²Swiss Federal Institute of Aquatic Science and Technology (Eawag), Dübendorf,
Switzerland;

³Institut de Biologia Evolutiva (CSIC-Universitat Pompeu Fabra), Barcelona, Catalonia,
Spain;

⁴Institut de Ciències del Mar (ICM-CSIC), Barcelona, Catalonia, Spain;

⁵Berlin Center for Genomics in Biodiversity Research (BeGenDiv), Berlin, Germany;

⁶Institut für Biologie, Freie Universität Berlin (FU), Berlin, Germany;

⁷Departament de Genètica, Microbiologia i Estadística, Institut de Recerca de la Biodiversitat
(IRBio), Universitat de Barcelona (UB), Barcelona 08028, Catalonia, Spain;

⁸ICREA, Pg. Lluís Companys 23, 08010 Barcelona, Catalonia, Spain;

This document is the accepted manuscript version of the following article:
Lu, Y., Ocaña-Pallarès, E., López-Escardó, D., Dennis, S. R., Monaghan, M. T., Ruiz-Trillo, I., ... Wolinska, J. (2020). Revisiting the phylogenetic position of *Caullerya mesnili* (Ichthyosporidia), a common *Daphnia* parasite, based on 22 protein-coding genes. *Molecular Phylogenetics and Evolution*, 106891 (31 pp.). <https://doi.org/10.1016/j.ympev.2020.106891>

This manuscript version is made available under the CC-BY-NC-ND 4.0
license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

Abstract

Caullerya mesnili is a common and virulent parasite of the water flea, *Daphnia*. It was classified within the Haplosporidia (Rhizaria) for over a century. However, a recent molecular phylogeny based on the 18S rRNA gene suggested it belonged to the Ichthyosporea, a class of protists closely related to animals within the Opisthokonta clade. The exact phylogenetic position of *C. mesnili* remained uncertain because it appeared in the 18S rRNA tree with a very long branch and separated from all other taxa, suggesting that its position could be artifactual. A better understanding of its phylogenetic position has been constrained by a lack of molecular markers and the difficulty of obtaining a suitable quantity and quality of DNA from *in vitro* cultures, as this intracellular parasite cannot be cultured without its host. We isolated and collected spores of *C. mesnili* and sequenced genomic libraries. Phylogenetic analyses of a newly generated multi-protein data set (22 proteins, 4,998 amino acids) and of sequences from the 18S rRNA gene both placed *C. mesnili* within the Ichthyophonida sub-clade of Ichthyosporea, as sister-taxon to *Abeoforma whisleri* and *Pirum gemmata*. Our study highlights the utility of metagenomic approaches for obtaining genomic information from intracellular parasites and for more accurate phylogenetic placement in evolutionary studies.

Keywords

evolution; metagenomic; molecular systematics; multi-protein phylogeny; protists; phylogenomics; 18S rRNA gene

1. Introduction

Caullerya mesnili (Chatton, 1907) (hereafter *Caullerya*) is a parasite of the water flea, *Daphnia*, a crustacean zooplankter (Ebert, 2005; Stewart et al., 2018; Wolinska et al., 2007). It causes regular epidemics and can infect up to 40% of a given *Daphnia* population at any one time (Turko et al., 2018; Wolinska et al., 2007, 2011). *Caullerya* has become a model for evolutionary research (Schoebel et al., 2010; Turko et al., 2018; Wolinska et al., 2014) because of its common occurrence, amenability to experimental manipulation, and extreme virulence (Lohr et al., 2010a, b), the latter leading to strong selection induced on its host.

Caullerya was classified within the Haplosporidia (Rhizaria) for over a century (Chatton, 1907; Green, 1974). This classification was challenged based on morphological characters (Bittner et al., 2002; Ebert, 2005). A phylogenetic analysis based on the 18S rRNA gene (SSU) suggested that *Caullerya* belongs to the Ichthyosporea (Lohr et al., 2010a), a class of protists closely related to animals (Mendoza et al., 2002; Ruiz-Trillo et al., 2006, 2008; Torruella et al., 2012). The Ichthyosporea clade, also known as Mesomycetozoa (Mendoza et al., 2002), is an early-branching lineage among the Holozoa, the group that comprises metazoans and their closest unicellular relatives (Lang et al., 2002; Ruiz-Trillo et al., 2008; Torruella et al., 2012). Ichthyosporeans are used as a model for the study of the origin of Metazoa and multicellular evolution (de Mendoza et al., 2015; Suga and Ruiz-Trillo, 2013).

Most ichthyosporean species that have been described were isolated from guts of aquatic organisms such as fish, arthropods, and molluscs (Glockling et al., 2013; Mendoza et al., 2002; Ragan et al., 1996). While this suggests ichthyosporeans are pathogens, it is uncertain if they are obligate parasites because their life cycles in natural hosts have not been described. Moreover, many ichthyosporean species have been successfully transferred to *in vitro* cultures (Grau-Bove et al., 2017; Marshall et al., 2008), indicating that some species are not host-dependent. Attempts to isolate and cultivate *Caullerya* using similar techniques have failed consistently (JW, personal observation). The parasitic life cycle of *Caullerya* is well

documented: it infects the gut epithelium of its host and is transmitted horizontally, i.e. spores of *Caullerya* released from one *Daphnia* host are directly taken up by another *Daphnia*, resulting in a rapidly spreading infection (Bittner et al., 2002; Lohr et al., 2010a). *Caullerya* is thought to evade host immunity by undermining host defences that would otherwise prevent its development in the *Daphnia* gut (Lu et al., 2018). Infection results in early *Daphnia* mortality and severely reduces reproduction (Lohr et al., 2010a, b).

A phylogenetic reconstruction using the 18S rRNA gene placed *Caullerya* within the Ichthyophonida sub-clade of the Ichthyosporea together with three uncultured environmental isolates, although its position remained uncertain because of the low nodal support for the position of the *Caullerya* sequence (Lohr et al., 2010a). In addition, the 18S rRNA gene sequence of *Caullerya* occurs as a remarkably long branch in all existing phylogenetic reconstructions of Ichthyosporea diversity (Glockling et al., 2013; Reynolds et al., 2017). The internal transcribed spacer region (ITS) nucleotide sequence of *Caullerya* is also considerably different from any published ITS sequence within Ichthyosporea (see Lohr et al., 2010a). This lack of resolution in the 18S rRNA-based phylogenies demands a more complete phylogenetic analysis, including as many genes as possible to further clarify the phylogenetic position of *Caullerya*.

To increase the data available for phylogenetic reconstruction, we performed whole-genome sequencing of *Caullerya* spores free of *Daphnia* host DNA contamination, used a metagenomic approach to assemble the genome and identify *bona fide* *Caullerya* proteins, and carried out a multi-protein phylogenetic analysis. We also updated the 18S rRNA phylogeny of ichthyosporeans using newly available data. Our results support the placement of *Caullerya* within the Ichthyophonida sub-clade of Ichthyosporea, as a sister-taxon to *Abeoforma whisleri* and *Pirum gemmata*.

2. Materials and Methods

2.1. Cultures

The *Caullerya mesnili* isolate originated from lake Skulska Wies in Poland (52°29'32.2"N 18°18'58.5"E) and was sampled in September 2016. Ten heavily infected *Daphnia* were selected under a dissecting microscope from a mixed zooplankton sample and placed together into an Eppendorf tube. Infected *Daphnia* were homogenized with a pestle and the resulting spore-cocktail was added to a jar containing three uninfected *Daphnia galeata* Sars juveniles (standard laboratory clone G100; inbred once), in ca. 30 mL of synthetic *Daphnia* medium (Saebelfeld et al., 2017). The *Caullerya* culture was multiplied by adding juvenile *Daphnia* from uninfected stock cultures at approximately 2-week intervals (Wolinska et al., 2006). Both uninfected and infected *D. galeata* cultures were kept in *Daphnia* medium at a constant temperature of 20 ± 1 °C, on a light: dark cycle of 12h:12h, and with an unlimited food supply of the unicellular green algae *Scenedesmus obliquus* (> 1 mg C L⁻¹, added three times per week).

2.2. Spore harvesting

Six months after the isolation of the original strain, 50 heavily infected *Daphnia* were selected from the parasite cultures. They were washed three times in autoclaved *Daphnia* medium and then transferred into ten wells of a 24-well cell culture plate (5 individuals per well) containing 1.5 ml autoclaved *Daphnia* medium in each well. After 5 hours, the faecal material was examined under a microscope for the presence of spores. *Caullerya* spores (~10,000 spores) were obtained from *Daphnia* and medium by sequential filtration through a 20-µm pore size mesh, followed by a 3-µm pore size mesh. The harvested spores retained on the 3-µm filter and were re-suspended with autoclaved *Daphnia* medium, then centrifuged at 6,000 g for 10 min at 15°C to pellet them. A subset of spores were examined under a microscope and photographed with a Nikon Digital Camera. Spores of *Caullerya* were oval-

shaped with a mean length of 10.4 μm (SD = 0.74 μm) and mean width of 8.4 μm (SD = 0.39 μm) (n = 50, Supplementary Fig. S1, and Bittner et al., 2002).

2.3. DNA extraction

The supernatant was decanted and the spore pellet was ground to a fine powder with liquid nitrogen in a precooled mortar and pestle. The ground tissue was transferred immediately to a 2-ml tube with 180 μl of Buffer ATL (Qiagen, Hilden, Germany) and 20 μl of Protease K (Qiagen) stock solution according to the manufacturer's protocol (Qiagen Genomic DNA Handbook). DNA extraction was performed using the DNeasy Blood & Tissue Kit (Qiagen) with elution volume of 100 μl Buffer AE (Qiagen). The quality and quantity of the DNA were determined using a NanoDrop ND-1000 spectrometer (PeqLab, Erlangen, Germany) and a Genomic DNA ScreenTape on an Agilent TapeStation 2200 instrument (Agilent Technologies, Böblingen, Germany).

2.4. *Caullerya* DNA confirmation

The presence of *Caullerya* or *Daphnia* DNA was tested using polymerase chain reaction (PCR) amplifications. Primers specific to *Caullerya* or to host *Daphnia* were applied to samples individually. Primers and PCR conditions used to amplify a 621-bp fragment of ITS in *Caullerya* and the microsatellite locus SwiD10 in *Daphnia* are described in Lohr et al. (2010a) and in Yin et al. (2010), respectively. A single band was obtained from the genomic DNA isolated from purified spores after PCR amplification with *Caullerya*-specific ITS primers, whereas no PCR product was generated using *Daphnia* specific microsatellite primers (Supplementary Fig. S2).

2.5. Library preparation and sequencing

Because there was a small amount of template DNA (0.39 μg extracted from *Caullerya* spores harvested from 50 infected *Daphnia*: Supplementary Fig. S3), two library preparation

methods were used that are optimised for ultra-low input DNA: Nextera (Illumina Nextera XT library prep kit) and NEBNext (New England BioLabs next ultra II DNA library prep kit). Approximately 28 ng DNA was used for the preparation of Nextera libraries and 100 ng was used to prepare NEBNext libraries. Both were carried out according to the manufacturers' instructions. The quality and quantity of libraries were determined using a DNA Nano Chip assay on a 2100 Bioanalyzer device (Agilent Technologies) with a broad size distribution (300 - 2000 bp; Supplementary Fig. S4). The Nextera library was sequenced three times and the NEBNext library was sequenced once on an Illumina MiSeq, in 150 bp paired-end (PE) mode (Supplementary Table S1).

2.6. Genome assembly and binning

A workflow of the *Caullerya* genomic data analysis is shown in Fig. 1. Raw reads from four sequencing runs were quality-filtered prior to assembly. Illumina adapters were removed using Cutadapt v 1.15 (Martin, 2011) and sequence quality was assessed using FastQC v 0.11.4 (Andrews, 2010). After removing leading bases below a Phred threshold of 30 or N bases (leading: 30), low quality bases were scanned using a 12-base wide sliding window and cut when the average quality per base dropped below 30 (sliding window:12:30). An additional 10 bases were removed from the start of each read (headcrop:10) using trimmomatic v 0.36 (Bolger et al., 2014). Trimmed reads were corrected with Spades v 3.12.0 (-only-error-correction option) (Bankevich et al., 2012). A final FastQC run was performed to ensure that the previous quality trimming and correction steps successfully conserved high quality reads without being too stringent or introducing new technical biases.

Genome assembly was produced using MEGAHIT (Liu et al., 2015). This assembler was chosen for two reasons. On the one hand, MEGAHIT has been shown to produce adequate contiguity metrics with low computational time. On the other hand, it can perform with better sensitivity than other assemblers in recovering also poorly abundant genomic fragments in the

sequenced sample, that can result in a higher recovery of potentially valid phylogenetic markers (Vollmers et al., 2017). For the assembly, multiple k-mers were used and otherwise the default parameters. Assembly quality (including genome size, number of contigs, N50, L50) was assessed using QUAST v 4.6.3 (Gurevich et al., 2013).

The contigs from the metagenomic assembler MEGAHIT were binned to form genome drafts, using MetaBAT2 v 2.12.1 (Kang et al., 2019) with a default minimum contig length (2500 bp). Binning quality was assessed using QUAST v 4.6.3 (Gurevich et al., 2013). The completeness of each bin was evaluated using BUSCO v 3.0.2 with default settings based on predicted transcripts from the eukaryote orthodb v 9 (Simão et al., 2015). BUSCO proteins that were found complete and in a single copy were extracted and blasted against taxonomy proteomes of different organisms including Ichthyosporidia, fungi, green algae, and human (Supplementary Table S2) in order to estimate the level of contamination. BUSCO proteins that were duplicated, fragmented, or missing were not further considered.

2.7. Multi-protein concatenated phylogeny

Typical phylogenomic analyses use highly conserved and low-copy gene markers, which are individually aligned and concatenated into a super matrix from which the phylogeny is inferred (Grau-Bové et al., 2017). Because of the levels of incompleteness and fragmentation in our *Caullerya* genome, accurate gene models could not be produced. We therefore used BUSCO proteins, as they are presumably well-suited phylogenetic markers because they are, by definition, highly conserved and low-copy genes (Simão et al., 2015).

Single-copy BUSCO-predicted sequences (27 BUSCO sequences, queries) from bin 6 were aligned to a *Caullerya* project database (targets) with an E-value threshold of 0.001 using BLASTP v 2.3.0+. The *Caullerya* project database was specifically designed for this study and included mainly Opisthokonta species (4 Metazoa, 6 Fungi, and 17 unicellular lineages phylogenetically related to Metazoa and Fungi, including 7 ichthyosporidians). Three

non-Opisthokonta species (2 Amoebozoa and 1 Apusozoa) were also included as outgroups for tree-rooting purposes (Supplementary Table S3). Preliminary analyses revealed potential contamination from the ichthyosporean *Pirum gemmata* into the protein data of *Abeoforma whisleri* (hereafter referred to as “*Pirum*” and “*Abeoforma*”, respectively). Contamination was removed by aligning the *Abeoforma* and *Pirum* targets all-against-all and discarding those *Abeoforma* targets that aligned to *Pirum* targets with > 95% identity.

For every BUSCO protein, the single best-scoring target sequence per species was kept. Every BUSCO protein and its corresponding target (maximum 1 per species) were considered as potential phylogenetic markers. Individual phylogenies were constructed to validate the inclusion of each potential marker into the phylogenetic dataset. For that, sequences were aligned with MAFFT v 7.123b using the -einsi option (Katoh et al., 2002), alignments were trimmed with trimAl v 1.4.rev15 using the -gappypout parameter (Capella-Gutiérrez et al., 2009), and maximum-likelihood (ML) phylogenies were inferred with IQ-TREE v 1.6.7 (Nguyen et al., 2014), using the model suggested by ModelFinder (see below, Kalyaanamoorthy et al., 2017) according to BIC criteria. Five of the 27 potential markers were discarded because the *Caullerya* sequence was highly divergent and located in an unexpected position in the phylogeny, most likely due to non-orthologous relationships or contamination. From the remaining 22 potential markers, a total of 15 highly divergent targets were removed for the same reasons. The final phylogenetic dataset consisted of 22 curated phylogenetic markers that were separately aligned and trimmed (same methods as described above) before being concatenated into a supermatrix. For the ML phylogenetic inference, the model suggested by ModelFinder was: LG matrix for substitution rates and the profile mixture model C60 to deal with compositional heterogeneity among sites (Si Quang et al., 2008), with the +F option to empirically count amino acid frequencies from the data and optimize mixture weights (-mwopt parameter is automatically enabled with +F), as well as the FreeRate model (Soubrier et al., 2012) with 6 categories to deal with substitution rate

heterogeneity among sites (LG+C60+F+R6). However, C40 was used instead of C60 in order to avoid overfitting, as some mixture weights were estimated close to zero with C60 and C50. IQ-TREE ultrafast bootstrap values (UFBoot2) (Hoang et al., 2017) were optimized with the -bnni option. A total of 110 non-parametric standard bootstraps were also computed.

2.8. Additional phylogenetic analyses with the multi-protein concatenated dataset

The fastest evolving sites of the supermatrix were determined using the *Posterior mean site rate weighted by posterior probability*, computed with IQ-TREE using the -wsr option. From this, 10 additional supermatrices were generated by iteratively removing the 5% fastest evolving sites, such that the first supermatrix excluded the 5% fastest evolving sites of the original one, the second supermatrix excluded the 10% fastest evolving sites of the original one, and the supermatrix from the last iteration excluded the 50% fastest evolving sites of the original one. Phylogenetic inferences on each supermatrix were also carried out using IQ-TREE with the LG+C40+F+R6 model and the -bnni parameter, as above.

The recoding of the amino acid supermatrix was done according to the 4-bin scheme specified by Susko *et al.* (2007). The phylogenetic inference of the 4-bin recoded supermatrix was done with the GTR+F+R5 model, in agreement with ModelFinder results (BIC criteria), and also refining the UFBoot values with -bnni.

The approximately-unbiased test analyses (AU-test) (Shimodaira, 2002), including the inference of the constrained phylogenies, were also done with IQ-TREE. The inferences of the constrained phylogenies were done under the LG+C40+F+R6 model for the non-recoded supermatrix (the original one), and under the GTR+F+R4 model for the 4-bin recoded supermatrix. See the list of constrained topologies explored and AU-test results in Supplementary Table S4.

2.9. 18S rRNA phylogeny

Sequences of 18S rRNA were collected from other described Ichthyosporea species (Grau-Bove et al., 2017; Marshall and Berbee, 2011), uncultured lineages of freshwater ichthyosporeans (FRESHIP1, FRESHIP2), marine ichthyosporeans (MAIP1), environmental sequences associated to a genus (del Campo and Ruiz-Trillo, 2013), and other unicellular Opisthokonta lineages as outgroups (Hehenberger et al., 2017; López-escardó et al., 2018). In addition, a BLAST search on GenBank (Benson et al., 2012) was performed to update the dataset of ichthyosporean sequences that can be environmental or were miss-annotated in GenBank. It ended up with 583 Ichthyosporea-like sequences, and, after a preliminary phylogenetic inference (performed as further explained), 133 sequences were selected representing Ichthyosporea diversity. We added an extensive outgroup of other Opisthokonta sequences, and ended up with 203 sequences to perform the phylogenetic analysis of the 18S ribosomal gene (Supplementary File S1). Sequences were aligned with MAFFT v 7.123b using the `-einsi` option. After trimming sequence ends, long insertions, and spuriously aligned sites with Gblocks v 0.91b (Castresana, 2000), a total of 1,492 sites remained. Phylogenetic trees were inferred from this alignment, using a maximum likelihood inference. The best substitution model for phylogenetic inference was selected using IQ-TREE, with the TESTNEW model selection procedure, following the BIC criterion. In all four cases (the best tree and three constrained topologies, see Supplementary Table S4), the GTR substitution matrix with a 5-category free-rate distribution, a modification of the standard Γ distribution (Yang, 1995), was selected as the best-fitting model. Maximum likelihood inferences were performed with IQ-TREE, and statistical supports were drawn from 1,000 ultrafast bootstrap values with a 0.99 minimum correlation as convergence criterion (Minh et al., 2013) and 1,000 replicates of the SHlike approximate likelihood ratio test (Guindon et al., 2010) (alignments before and after trimming, as well as the phylogenetic tree in nexus format are available in Supplementary File S2 and S3).

3. Results

3.1. Genome sequencing and assembly

Illumina MiSeq sequencing yielded 48,480,041 reads from two DNA libraries and four sequencing runs. After quality filtration and error correction, a total of 44,932,272 reads (93%) were used for genome assembly (Supplementary Table S1). This assembly from MEGAHIT contained 133,382 contigs (58.8 Mb, N50 = 1,714 bp, L50 = 6,066) (Supplementary Table S5). This genome size was smaller than the size estimated by k-mer assessment (155 Mb) and was incomplete (24.1% of BUSCO). A total of 3,835 MEGAHIT contigs (≥ 2.5 kb) were binned using MetaBAT2, yielding six bins (2,924 contigs). Statistics of binning results are summarized in Supplementary Table S6. A total of 75 single-copy BUSCO proteins were extracted from the six bins (Supplementary Table S7) and were blasted against taxonomy proteomes of Ichthyosporea, fungi, green algae, and human (Supplementary Table S2). There were no BUSCO proteins identified in bin 1 and 2. The contaminants (i.e. matching known proteomes) were identified in bin 3, 4 and 5, whereas most of the Ichthyosporea proteins were found in bin 6 (Supplementary Table S8). BUSCO proteins identified as contaminants, or that were duplicated, fragmented, or missing were not further considered.

3.2. Taxonomic partitioning of putative *Caullerya* contigs

Bin 6 contained 1,537 contigs, spanning 11.34 Mb, and included 27 BUSCO proteins, more than any other bin. In order to investigate whether bin 6 corresponded to *Caullerya*, the 27 BUSCO proteins from bin 6 were aligned to a comprehensive eukaryotic database (Ocaña-Pallarès et al., 2019). The most similar proteins to 21 of the 27 BUSCO proteins were sequences from the ichthyosporeans *Pirum* and *Abeoforma*. This suggests that bin 6 corresponded to the genome of *Caullerya*, and that *Caullerya* was within Ichthyosporea, in agreement with Lohr et al. (2010a), and could be closely related to *Pirum* and *Abeoforma*. A

multi-protein concatenated phylogeny was performed to validate this hypothetical phylogenetic association (see the following results section).

3.3. Multi-protein concatenated phylogeny

The phylogenetic dataset consisted of 22 markers (see Material and Methods) and the resulting alignment supermatrix included 4,998 amino acid sites from *Caullerya* and 30 other species. Of these, 3,549 sites were parsimony-informative, and only ~5% were missing in *Caullerya*. The maximum likelihood phylogeny recovered *Caullerya* within the Ichthyosporea as sister taxon to *Abeoforma whisleri* and *Pirum gemmata* with node support of 100% (IQ-TREE refined ultrafast bootstrap) (Fig. 2A).

3.4. Additional phylogenetic analyses with the multi-protein concatenated dataset

A series of additional phylogenetic analyses were performed to explore the robustness of this topology (see Material and Methods). First, three alternative phylogenies were reconstructed from the same supermatrix used in Fig. 2A, but constraining the nodes (i) *Caullerya* + *Abeoforma*; (ii) *Caullerya* + *Pirum*; and (iii) *Caullerya* + (*Abeoforma* + *Pirum* + *Creolimax* + *Sphaeroforma* + *Amoebidium* + *Ichthyophonus*). AU-tests indicated that these three alternative topologies had significantly less support ($p < 0.01$, Fig. 2C) than the ML tree with *Caullerya* + (*Abeoforma* + *Pirum*) (Fig. 2A). Second, because the long-branch of *Caullerya* (Fig. 2A) is indicative of high substitution rates, we tested whether the topology could have been influenced by potential saturation in some alignment positions by reconstructing the phylogenies of a subset of supermatrices that were generated from iteratively excluding the fastest evolving sites of the original one. Nodal support of 100% was found for the *Caullerya* + *Abeoforma* + *Pirum* and for the *Abeoforma* + *Pirum* clades independent of the proportion of sites removed (yellow and blue topologies in Fig. 2B). In contrast, nodal support for the three other topologies evaluated was consistently 0% (green, red, and black topologies in Fig. 2B). Third, to discard that the position of *Caullerya* in Fig.

2A could have been influenced by compositional heterogeneity at the amino-acid level (Jermiin et al., 2004), the alignment supermatrix was recoded into a 4-bin scheme (SR-4) (Susko and Roger, 2007), from which a ML unconstrained phylogeny was inferred (Supplementary Fig. S5). For topology testing purposes, phylogenies of the three constrained topologies that were explored with the non-recoded supermatrix were also reconstructed from the SR-4 recoded matrix (see above). The *Caullerya* + *Abeoforma* topology was rejected with the AU-test ($p < 0.01$, Fig. 2C), but not the *Caullerya* + *Pirum* topology ($p = 0.304$). Notwithstanding, *Caullerya* + (*Abeoforma* + *Pirum*) was the most supported topology with the SR-4 dataset ($p = 0.7021$). Overall, the phylogenomic analysis strongly supports a phylogenetic position of *Caullerya* as sister-taxon to *Pirum* and *Abeoforma*.

3.5. 18S rRNA phylogeny

The 18S rRNA phylogeny shows that together with *Caullerya*, *Pirum*, and *Abeoforma*, there is an important diversity of uncultured sequences related to these three taxa that form a highly-supported monophyletic clade (100% ultrafast bootstrap and 100% replicates of the SH-like approximate likelihood ratio test, Fig. 3), that we here name the CABEPI clade (*Caullerya*-*ABEoforma*-*PIrum* clade). In particular, two new monophyletic environmental groups were defined within CABEPI diversity that are highly supported: the MACABEPI (MARine CABEPI)1 and the MACABEPI2 groups. Both groups correspond to sequences isolated from marine environments. In addition, there were uncultured marine sequences related to *Abeoforma* with high nodal support. The position of the remaining uncultured marine sequences was not highly supported, as well as the internal relationships among CABEPI groups. For that a multi-protein approach would be needed. Currently, only the phylogenetic relationship between *Caullerya*, *Abeoforma* and *Pirum* can be clearly defined; *Caullerya* is the earliest-branching of the CABEPI clade. Further genomic information of other taxa of CABEPI clade will better solve the internal relationships within the group.

Regarding *Caullerya*, a closely related sequence was found (accession number JN619316) that is annotated in GenBank as the 18S ribosomal gene of the aquatic predaceous diving beetle *Sanfilipoodytes rossi* (Coleoptera, Dytiscidae). Based on the close relationship between *Caullerya* and the 18S sequence JN619316 in our phylogenetic tree, it is likely that the latter is instead an ichthyosporean symbiont or parasite of aquatic beetles. Sister to *Caullerya* and the JN61931 sequence was another sequence related to a beetle, JN699060 which belongs to a symbiont of the mealworm beetle *Tenebrio molitor* (Coleoptera, Tenebrionidae) (Lord et al., 2012). The monophyly of the three sequences is well-supported (>90% ultrafast bootstrap and >80% replicates of the SH-like approximate likelihood ratio test, Fig. 3).

Finally, the 18S analysis defined two new environmental Ichthyosporean groups outside of the CABEPI clade, namely the FRESHIP 3 (FRESHwater IchthysPoreans, following the nomenclature in (del Campo and Ruiz-Trillo, 2013), and the MAIPH1 (MARine IchthysPHonida 1) groups). Interestingly, the sequences previously defined as the MAIP1 clade (del Campo and Ruiz-Trillo, 2013) were paraphyletic in our 18S phylogeny, and are located within MACABEPI1, MACABEPI2, and MAIPH1 groups.

4. Discussion

Both the multi-protein and 18S rRNA gene phylogenetic analyses strongly supported the placement of *Caullerya* within the sub-clade Ichthyophonida of the Ichthyosporea. This phylogenetic placement was suggested by earlier studies using 18S rRNA sequences (Lohr et al., 2010a), but the multi-protein phylogeny here provided higher resolution than the phylogeny based solely on 18S rRNA sequences, clearly placing *Caullerya* as a sister-taxon to *Abeoforma* and *Pirum*. The analysis of 18S rRNA sequences positioned *Caullerya* within a monophyletic clade containing *Abeoforma* and *Pirum* and other environmental sequences that

we named CABEPI clade, but the exact relationship among the internal groups of the clade remain unresolved.

Most ichthyosporeans described to date were proposed to be parasites or commensals of multiple host species (Mendoza et al., 2002). For example, *Sphaeroforma* sp. was reported from several phyla of marine invertebrates (Glockling et al., 2013). Allele frequency analyses of *Sphaeroforma tapetis* indicate geographical differentiation of populations, but not host specificity (Marshall and Berbee, 2013). Similar results were found for *Ichthyophonus hoferi*, described as a ubiquitous parasite of marine and freshwater fishes including herring and chinook salmon (Hershberger et al., 2002; Rahimian and Thulin, 1996). These examples and others suggest that many ichthyosporeans might be sapronotic agents, facultative parasites, or obligate parasites, but with a broad host spectrum. In contrast, *Caullerya* is strictly dependant on its host, and has thus far been reported from only one host, the crustacean zooplankter *Daphnia* (Table 1). *Pirum* was isolated from the gut content of the detritivorous peanut worm *Phascolosoma agassizii* (phylum Sipuncula) whereas *Abeoforma* was isolated from material from the digestive tract of the filter-feeding mussel *Mytilus* sp. (phylum Mollusca) (Marshall and Berbee, 2011). Their presence in the digestive tract, however, does not prove their parasitic nature. There are no shared morphological characters currently recognized between *Caullerya* and *Pirum* or *Abeoforma* (Marshall and Berbee, 2011). This may be the result of having studied the *Caullerya* life cycle *in vivo*, in comparison to *in vitro* observations of *Pirum* and *Abeoforma*. Conversely, *Caullerya* might be distinct from these and other ichthyosporeans because of its obligatory parasitic life cycle. The long branch of *Caullerya* in the multi-gene phylogeny is consistent with its parasitic lifestyle, as high divergence is characteristic of many parasitic lineages (Giovannoni et al., 2014; Janouskovec and Keeling, 2016). The 18S rRNA phylogeny suggests that *Caullerya* might have had a less aggressive ancestor, a symbiont of beetles (Coleoptera). In this sense, the study of the Coleoptera symbionts may provide clues in the parasitic transition of *Caullerya*.

While the present study describes the first attempt to sequence the *Caullerya* genome, it was not possible to assemble a full genome for several reasons. First, *Caullerya* DNA was extracted from multiple spores. Although these spores originated from a parasite isolate hosted by a single *Daphnia* genotype for a period of six months, it is not certain that all spores were genetically identical (i.e. that they represented a single parasite “clone”). Second, only very small amounts of DNA could be obtained from the isolated spores, and library preparation kits specifically designed for ultra-low input DNA were needed which resulted in uneven coverage. Consequently, the obtained *Caullerya* genome was fragmented and incomplete, which hindered the gene prediction and comparative analyses, e.g. gene content evolution or synteny conservation with other Ichthyosporeans. Similar challenges were encountered while sequencing the genomes of *Pirum* and *Abeoforma*. Compared to other Ichthyosporean species, these two genomes were large and difficult to assemble, with only 11.9% and 17.0% of BUSCO orthologs detected, respectively (Grau-Bove et al., 2017). The DNA of *Pirum* and *Abeoforma* was extracted from axenic cell cultures grown in marine broth medium without host, excluding DNA quantity or contamination being the reason for sequencing difficulties. Additional sequencing efforts using long-read sequencing technologies could help to produce satisfactory genomic data for these ichthyosporean taxa.

Genome assembly is challenging even for single genomes (Nagarajan and Pop, 2013), and simultaneous assembly of multiple genomes at varying abundances in a mixed sample, i.e. metagenomics, is even more complicated (Wooley and Ye, 2010). Despite all the purification steps undertaken in the present study, it was not possible to obtain contaminant-free DNA from *in vivo* cultures. Fortunately, a large part of the *Caullerya* genome was successfully placed in a single bin by metagenomic analysis, indicating the usefulness of a metagenome assembler in dealing with contaminated genomes. We suggest that a similar approach might be helpful in obtaining data from contaminated DNA of non-model organisms, particularly from intracellular parasites.

Acknowledgements

We thank Marcin Dziuba for providing us with a zooplankton sample from lake Skulska Wies (from which the *Caullerya* strain was isolated), and Ursula Newen for the maintenance of *Daphnia* cultures. Thanks to Susan Mbedi and Sarah Sparmann for assistance during the Illumina sequencing at the Berlin Center for Genomics in Biodiversity Research (BeGenDiv). We are grateful to Camila Mazzoni, Paul Johnston, Felix Heeger, Marcela Uliano, Maximilian Driller, Calvinna Caswara, and the entire BeGenDiv team for their technical support and excellent collaboration. This work was supported by a joint “lead agency” grant from the German Science Foundation (WO 1587/6-1 to JW) and Swiss National Science Foundation (310030 L 166628 to PS). We also acknowledge support from a grant (BFU2014-57779-P and BFU2017-90114-P) from the Ministerio de Economía y Competitividad (MINECO), Agencia Estatal de Investigación (AEI), and Fondo Europeo de Desarrollo Regional (FEDER) to IRT. EOP was supported by a pre-doctoral FPI grant from MINECO.

Availability of data and materials

The genome data have been submitted to SRA at NCBI under the Accession no. PRJNA631182. The primary sequences have been submitted to NCBI under the Accession no. SRX8386394 - SRX8386397.

References

- Andreou, D., Arkush, K.D., Guégan, J.-F., Gozlan, R.E., 2012. Introduced pathogens and native freshwater biodiversity: a case study of *Sphaerothecum destruens*. PLoS One 7, e36998.
- Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data.
- Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., Pyshkin, A.V., Sirotkin, A.V., Vyahhi, N., Tesler, G., Alekseyev, M.A., Pevzner, P.A., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19, 455-477.

- Benny, G.L., O'Donnell, K., 2000. *Amoebidium parasiticum* is a protozoan, not a Trichomycete. *Mycologia*, 1133-1137.
- Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Sayers, E.W., 2012. GenBank. *Nucleic Acids Research* 41, D36-D42.
- Bittner, K., Rothhaupt, K.-O., Ebert, D., 2002. Ecological interactions of the microparasite *Caullerya mesnili* and its host *Daphnia galeata*. *Limnology and Oceanography* 47, 300-305.
- Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25, 1972-1973.
- Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17, 540-552.
- Chatton, E., 1907. *Caullerya mesnili* ngn sp. Haplosporidie parasite des Daphnies. *Société du Biologie* 62, 529-531.
- de Mendoza, A., Suga, H., Permanyer, J., Irimia, M., Ruiz-Trillo, I., 2015. Complex transcriptional regulation and independent evolution of fungal-like traits in a relative of animals. *eLife* 4.
- del Campo, J., Ruiz-Trillo, I., 2013. Environmental survey meta-analysis reveals hidden diversity among unicellular opisthokonts. *Molecular Biology and Evolution* 30, 802-805.
- Ebert, D., 2005. Ecology, epidemiology, and evolution of parasitism in *Daphnia*. National Library of Medicine.
- Giovannoni, S.J., Thrash, J.C., Temperton, B., 2014. Implications of streamlining theory for microbial ecology. *The ISME Journal* 8, 1553.
- Glockling, S.L., Marshall, W.L., Gleason, F.H., 2013. Phylogenetic interpretations and ecological potentials of the Mesomycetozoa (Ichthyosporea). *Fungal Ecology* 6, 237-247.
- Gozlan, R.E., St-Hilaire, S., Feist, S.W., Martin, P., Kent, M.L., 2005. Biodiversity: disease threat to European fish. *Nature* 435, 1046.
- Gozlan, R.E., Whipps, C.M., Andreou, D., Arkush, K.D., 2009. Identification of a rosette-like agent as *Sphaerothecum destruens*, a multi-host fish pathogen. *International Journal for Parasitology* 39, 1055-1058.
- Grau-Bove, X., Torruella, G., Donachie, S., Suga, H., Leonard, G., Richards, T.A., Ruiz-Trillo, I., 2017. Dynamics of genomic innovation in the unicellular ancestry of animals. *eLife* 6, e26036.
- Green, J., 1974. Parasites and epibionts of Cladocera. *Philosophical Transactions of the Royal Society of London. B* 47, 219-247.
- Grigg, R.D., Williams, M.C., 1989. Distribution of *Amoebidium* and *Smittium* species (Trichomycetes) in mosquito larvae on the Platte river floodplain of Central Nebraska.
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* 59, 307-321.

- Gurevich, A., Saveliev, V., Vyahhi, N., Tesler, G., 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29, 1072-1075.
- Hehenberger, E., Tikhonenkov, D.V., Kolisko, M., del Campo, J., Esaulov, A.S., Mylnikov, A.P., Keeling, P.J., 2017. Novel predators reshape holozoan phylogeny and reveal the presence of a two-component signaling system in the ancestor of animals. *Current Biology* 27, 2043-2050.
- Hershberger, P.K., Stick, K., Bui, B., Carroll, C., Fall, B., Mork, C., Perry, J.A., Sweeney, E., Wittouck, J., Winton, J., 2002. Incidence of *Ichthyophonus hoferi* in Puget Sound fishes and its increase with age of Pacific herring. *Journal of Aquatic Animal Health* 14, 50-56.
- Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., Vinh, L.S., 2017. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* 35, 518-522.
- Janouskovec, J., Keeling, P.J., 2016. Evolution: Causality and the origin of parasitism. *Current Biology* 26, R174-R177.
- Jermiin, L.S., Ho, S.Y.W., Ababneh, F., Robinson, J., Larkum, A.W.D., 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic Biology* 53, 638-643.
- Jøstensen, J.-P., Sperstad, S., Johansen, S., Landfald, B., 2002. Molecular-phylogenetic, structural and biochemical features of a cold-adapted, marine ichthyosporean near the animal-fungal divergence, described from *in vitro* cultures. *European Journal of Protistology* 38, 93-104.
- Kang, D., Li, F., Kirton, E.S., Thomas, A., Egan, R.S., An, H., Wang, Z., 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ Preprints* 7, e27522v27521.
- Katoh, K., Misawa, K., Kuma, K.i., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30, 3059-3066.
- Lang, B.F., O'Kelly, C., Nerad, T., Gray, M.W., Burger, G., 2002. The closest unicellular relatives of animals. *Current Biology* 12, 1773-1778.
- Liu, C.-M., Li, D., Sadakane, K., Luo, R., Lam, T.-W., 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct *de Bruijn* graph. *Bioinformatics* 31, 1674-1676.
- Lohr, J.N., Laforsch, C., Koerner, H., Wolinska, J., 2010a. A *Daphnia* parasite (*Caullerya mesnili*) constitutes a new member of the Ichthyosporea, a group of protists near the animal-fungi divergence. *Journal of Eukaryotic Microbiology* 57, 328-336.
- Lohr, J.N., Yin, M., Wolinska, J., 2010b. Prior residency does not always pay off-co-infections in *Daphnia*. *Parasitology* 137, 1493-1500.
- López - escardó, D., López - garcía, P., Moreira, D., Ruiz - Trillo, I., Torruella, G., 2018. *Parvularia atlantis* gen. et sp. nov., a Nucleariid Filose Amoebe (Holomycota, Opisthokonta). *Journal of Eukaryotic Microbiology* 65, 170-179.

- Lord, J.C., Hartzer, K.L., Kambhampati, S., 2012. A nuptially transmitted ichthyosporean symbiont of *Tenebrio molitor* (Coleoptera: Tenebrionidae). *Journal of Eukaryotic Microbiology* 59, 246-250.
- Lu, Y., Johnston, P.R., Dennis, S.R., Monaghan, M.T., John, U., Spaak, P., Wolinska, J., 2018. *Daphnia galeata* responds to the exposure to an ichthyosporean gut parasite by down-regulation of immunity and lipid metabolism. *BMC Genomics* 19, 932.
- Marshall, W.L., Berbee, M.L., 2011. Facing unknowns: living cultures (*Pirum gemmata* gen. nov., sp. nov., and *Abeoforma whisleri*, gen. nov., sp. nov.) from invertebrate digestive tracts represent an undescribed clade within the unicellular Opisthokont lineage Ichthyosporea (Mesomycetozoea). *Protist* 162, 33-57.
- Marshall, W.L., Berbee, M.L., 2013. Comparative morphology and genealogical delimitation of cryptic species of sympatric isolates of *Sphaeroforma* (Ichthyosporea, Opisthokonta). *Protist* 164, 287-311.
- Marshall, W.L., Celio, G., McLaughlin, D.J., Berbee, M.L., 2008. Multiple isolations of a culturable, motile Ichthyosporean (Mesomycetozoa, Opisthokonta), *Creolimax fragrantissima* n. gen., n. sp., from marine invertebrate digestive tracts. *Protist* 159, 415-433.
- Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal* 17, pp-10.
- Mendoza, L., Taylor, J.W., Ajello, L., 2002. The class Mesomycetozoea: a heterogeneous group of microorganisms at the animal-fungal boundary. *Annual Reviews in Microbiology* 56, 315-344.
- Minh, B.Q., Nguyen, M.A.T., von Haeseler, A., 2013. Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and Evolution* 30, 1188-1195.
- Nagarajan, N., Pop, M., 2013. Sequence assembly demystified. *Nature Reviews Genetics* 14, 157.
- Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2014. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 32, 268-274.
- Ocaña-Pallarès, E., Najle, S.R., Scazzocchio, C., Ruiz-Trillo, I., 2019. Reticulate evolution in eukaryotes: Origin and evolution of the nitrate assimilation pathway. *PLoS Genetics* 15, e1007986.
- Ragan, M.A., Goggin, C.L., Cawthorn, R.J., Cerenius, L., Jamieson, A.V., Plourde, S.M., Rand, T.G., Söderhäll, K., Gutell, R.R., 1996. A novel clade of protistan parasites near the animal-fungal divergence. *Proceedings of the National Academy of Sciences* 93, 11907-11912.
- Rahimian, H., Thulin, J., 1996. Epizootiology of *Ichthyophonus hoferi* in herring populations off the Swedish west coast. *Diseases of Aquatic Organisms* 27, 187-195.
- Reynolds, N.K., Smith, M.E., Tretter, E.D., Gause, J., Heeney, D., Cafaro, M.J., Smith, J.F., Novak, S.J., Bourland, W.A., White, M.M., 2017. Resolving relationships at the animal-fungal divergence:

- A molecular phylogenetic study of the protist trichomycetes (Ichthyosporidia, Eccrinidae). *Molecular Phylogenetics and Evolution* 109, 447-464.
- Ruiz-Trillo, I., Lane, C.E., Archibald, J.M., Roger, A.J., 2006. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. *Journal of Eukaryotic Microbiology* 53, 379-384.
- Ruiz-Trillo, I., Roger, A.J., Burger, G., Gray, M.W., Lang, B.F., 2008. A phylogenomic investigation into the origin of metazoa. *Molecular Biology and Evolution* 25, 664-672.
- Saebelfeld, M., Minguez, L., Griebel, J., Gessner, M.O., Wolinska, J., 2017. Humic dissolved organic carbon drives oxidative stress and severe fitness impairments in *Daphnia*. *Aquatic Toxicology* 182, 31-38.
- Schoebel, C.N., Wolinska, J., Spaak, P., 2010. Higher parasite resistance in *Daphnia* populations with recent epidemics. *Journal of Evolutionary Biology* 23, 2370-2376.
- Si Quang, L., Gascuel, O., Lartillot, N., 2008. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics* 24, 2317-2323.
- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210-3212.
- Soubrier, J., Steel, M., Lee, M.S.Y., Der Sarkissian, C., Guindon, S., Ho, S.Y.W., Cooper, A., 2012. The influence of rate heterogeneity among sites on the time dependence of molecular rates. *Molecular Biology and Evolution* 29, 3345-3358.
- Stewart, T.E., Torchin, M.E., Cáceres, C.E., 2018. Invisible parasites and their implications for coexisting water fleas. *Journal of Parasitology* 104, 101-106.
- Suga, H., Ruiz-Trillo, I., 2013. Development of ichthyosporidia sheds light on the origin of metazoan multicellularity. *Developmental Biology* 377, 284-292.
- Susko, E., Roger, A.J., 2007. On reduced amino acid alphabets for phylogenetic inference. *Molecular Biology and Evolution* 24, 2139-2150.
- Torruella, G., Derelle, R., Paps, J., Lang, B.F., Roger, A.J., Shalchian-Tabrizi, K., Ruiz-Trillo, I., 2012. Phylogenetic relationships within the Opisthokonta based on phylogenomic analyses of conserved single-copy protein domains. *Molecular Biology and Evolution* 29, 531-544.
- Turko, P., Tellenbach, C., Keller, E., Tardent, N., Keller, B., Spaak, P., Wolinska, J., 2018. Parasites driving host diversity: incidence of disease correlated with *Daphnia* clonal turnover. *Evolution* 72, 619-629.
- Vollmers, J., Wiegand, S., Kaster, A.K., 2017. Comparing and evaluating metagenome assembly tools from a microbiologist's perspective-not only size matters. *Plos One* 12, e0169662.
- Wolinska, J., Bittner, K., Ebert, D., Spaak, P., 2006. The coexistence of hybrid and parental *Daphnia*: the role of parasites. *Proceedings of the Royal Society B: Biological Sciences* 273, 1977.

- Wolinska, J., Keller, B., Manca, M., Spaak, P., 2007. Parasite survey of a *Daphnia* hybrid complex: host - specificity and environment determine infection. *Journal of Animal Ecology* 76, 191-200.
- Wolinska, J., Petrusek, A., Yin, M., Koerner, H., Seda, J., Giessler, S., 2014. Population structure of a microparasite infecting *Daphnia*: spatio-temporal dynamics. *BMC Evolutionary Biology* 14, 247.
- Wolinska, J., Seda, J., Koerner, H., Smilauer, P., Petrusek, A., 2011. Spatial variation of *Daphnia* parasite load within individual water bodies. *Journal of Plankton Research* 33, 1284-1294.
- Wooley, J.C., Ye, Y., 2010. Metagenomics: facts and artifacts, and computational challenges. *Journal of Computer Science and Technology* 25, 71-81.
- Yang, Z., 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139, 993-1005.
- Yin, M., Wolinska, J., Gießler, S., 2010. Clonal diversity, clonal persistence and rapid taxon replacement in natural populations of species and hybrids of the *Daphnia longispina* complex. *Molecular Ecology* 19, 4168-4178.

Table 1. Ichthyosporean species from which the genome information (Grau-Bove et al., 2017) was used for the construction of the multi-protein phylogeny in the present study.

Parasite	Presumably host	Reference
<i>Caullerya mesnili</i>	Several <i>Daphnia</i> species	Bittner et al. (2002); Ebert (2005); Stewart et al. (2018); Wolinska et al. (2007)
<i>Abeoforma whisleri</i>	Peanut worm <i>Phascolosoma agassizii</i>	Marshall and Berbee (2011)
<i>Pirum gemmata</i>	Mussel <i>Mytilus</i> sp.	Marshall and Berbee (2011)
<i>Creolimax fragrantissima</i>	Different marine invertebrates (e.g. <i>Phascolosoma agassizii</i> , <i>Corella</i> sp., <i>Leptosynapta clarki</i> , <i>Katharina tunicata</i>)	Marshall et al. (2008)
<i>Sphaeroforma arctica</i>	Marine and freshwater fishes	Jøstensen et al. (2002); Mendoza et al. (2002)
<i>Ichthyophonus hoferi</i>	Marine and freshwater fishes	Hershberger et al. (2002); Rahimian and Thulin (1996)
<i>Amoebidium parasiticum</i>	Freshwater larval insects and small crustaceans	Benny and O'Donnell (2000); Grigg and Williams (1989)
<i>Sphaerothecum destruens</i>	Marine and freshwater fishes (e.g. Chinook salmon, <i>Oncorhynchus tshawytscha</i> , and Atlantic salmon, <i>Salmo salar</i>)	Andreou et al. (2012); Gozlan et al. (2005); Gozlan et al. (2009)

Fig. 1. Workflow of *Caullerya* genomic data analysis.

Fig. 2. (A) Maximum likelihood phylogeny of a concatenated alignment supermatrix of 4,998 amino acid sites, done with IQ-TREE under the LG+C60+F+R6 model. Statistical nodal supports were assessed with (i) 1,000 optimized UFBoot2 bootstrap replicates and (ii) 110 non-parametric (standard) bootstrap replicates. Nodal supports are only represented (in %) in the nodes where either (i) or (ii) support values are lower than 100%. (B) Nodal support variation for 5 topologies evaluated in relation to the iterative removal of the fastest-evolving sites from the original supermatrix (5% sites removed per iteration). (C) Results from the Approximately Unbiased (AU) tests comparing the ML supports between the best tree found (ML tree), represented in (A), and three alternative topologies under two distinct alignment supermatrix datasets: the default amino-acid supermatrix, and a supermatrix in which amino-acids were recoded to four distinct categories according to Susko and Rogers 4-bin scheme (see references). logL: ML score in log-10 base. ΔL : increments in logL respect to the ML tree. p AU-test: AU-test p -value. '*' Indicates values below a significance level of 0.05.

Fig. 3. Phylogenetic tree based on 203 sequences of 18S rRNA representing all known molecular diversity of Ichthyosporea and other Opisthokonts, including environmental lineages. The tree was inferred by maximum likelihood under the GTR+F+R5 bootstraps of single branch test (SH-aLRT, right) and ultrafast bootstraps (UFBS, left) calculated with IQ-TREE.

Figure 1

Figure2
A UFBoot2 (optimized) / non-parametric bootstrap
Tree scale: 0.1

C

Default alignment supermatrix

Topologies tested	logL	ΔL	p AU-test
ML tree (panel A)	-175210.24	0	0.996
Green (panel B)	-175528.52	318.28	0*
Red (panel B)	-175258.58	48.34	0.001*
Black (panel B)	-175245.53	35.29	0.004*

SR4-recoded alignment supermatrix

Topologies tested	logL	ΔL	p AU-test
ML tree (panel A)	-70819.75	0	0.702
Green (panel B)	-70994.14	174.39	0*
Red (panel B)	-70844.14	24.38	0.003*
Black (panel B)	-70828.45	8.70	0.304

Figure3

Highlights

- We obtained the first genomic data from *Caullerya mesnili*, gut parasite of *Daphnia*
- Phylogenetic analyses of multiple proteins placed *C. mesnili* within Ichthyosporaea
- *Abeoforma* and *Pirum* were well supported as sister taxa
- Our metagenomic approach enabled the analysis of an intracellular parasite

Graphical Abstract

Author Statement

JW and PS obtained the funding. YL, SRD, MTM, PS and JW designed the study. YL obtained parasite DNA and performed all molecular work and sequencing. YL analysed genomic data with input from SRD and IRT. EOP made the multi-protein phylogeny whereas DLE made the 18S phylogeny. YL wrote most of the manuscript with support of JW and input of PS. EOP and DLE wrote specifically the parts corresponding to multi-protein phylogeny and the 18S tree, respectively, with support of IRT. All authors commented on, and approved, the final version of the manuscript.