

RESEARCH

Open Access



# Phylogeny and species delimitation of ciliates in the genus *Spirostomum* (class Heterotrichea) using single-cell transcriptomes

Shahed Uddin Ahmed Shazib<sup>1,2,4,5\*</sup>, Auden Cote-L'Heureux<sup>2</sup>, Ragib Ahsan<sup>1,2,3</sup>, Sergio A. Muñoz-Gómez<sup>4,5</sup>, JunMo Lee<sup>6,7</sup>, Laura A. Katz<sup>2,3\*</sup> and Mann Kyoon Shin<sup>1\*</sup>

## Abstract

**Background** Ciliates are single-celled microbial eukaryotes that diverged from other eukaryotic lineages more than a billion years ago. The long evolutionary timespan of ciliates has led to enormous genetic and phenotypic changes, contributing significantly to their high level of diversity. Recent analyses based on molecular data have revealed numerous cases of cryptic species complexes in different ciliate lineages, demonstrating the need for a robust approach to delimit species boundaries and elucidate phylogenetic relationships. Species of the genus *Spirostomum* are difficult to identify due to the lack of distinctive morphological characters. Previous molecular studies have focused on only a few loci, namely the nuclear ribosomal RNA genes, alpha-tubulin, and mitochondrial CO1, suggesting the presence of several cryptic *Spirostomum* species. In this study, we increased taxon sampling and obtained single-cell transcriptomes of 25 *Spirostomum* specimens (representing six morphospecies) sampled from South Korea and the USA. We evaluated the utility of the transcriptomic data by constructing species trees using concatenation and coalescent-based methods. In addition, we used neighbor-net network analysis to visualize and quantify potential phylogenetic conflicts within the concatenated dataset. Furthermore, coalescent-based species delimitation was performed with transcriptomic data to define the species boundaries within the genus *Spirostomum*.

**Results** Phylogenomic analysis of 37 *Spirostomum* specimens (25 newly generated transcriptomes and 12 from GenBank) and 265 protein-coding genes provides robust insight into the evolutionary relationships among *Spirostomum* species. Our results confirm that species with moniliform and compact macronucleus each form a distinct monophyletic lineage, with the compact macronucleus likely representing the ancestral state, while the moniliform macronucleus being a derived trait. Furthermore, our analyses suggest that ancestral polymorphism and rapid radiation may have shaped the genetic diversity and evolutionary history of *Spirostomum*, and the *S. minus*-like

\*Correspondence:

Shahed Uddin Ahmed Shazib  
shahed30ahmed@gmail.com; sshazib@purdue.edu  
Laura A. Katz  
lkatz@smith.edu  
Mann Kyoon Shin  
mkshin@ulsan.ac.kr

Full list of author information is available at the end of the article



© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

appearance represents the ancestral state of the species with a moniliform macronucleus. Therefore, the *S. minus*-like species share ancestral morphological traits and cannot be morphologically delimited. The multispecies coalescent (MSC) model suggests that two cryptic species from each of *S. minus*, *S. ambiguum*, *S. subtilis*, *S. teres*, and *S. aff. minus* represent distinct lineages within the genus *Spirostomum*. We also provide a workflow for reconstructing nuclear ribosomal RNA gene sequences from transcriptome sequences using a read mapping approach, and compare different mapping methods to reconstruct reliable contigs.

**Conclusion** Our sampling of closely related *Spirostomum* populations and comprehensive single-cell RNA sequencing (scRNA-seq) data allowed us to reveal the hidden crypticity of species within the genus *Spirostomum* and to resolve and provide much stronger support than hitherto to the phylogeny of this model ciliate genus.

**Keywords** Cryptic species, Multispecies coalescent (MSC), Read mapping, scRNA-seq, Species tree

## Background

Ciliates are among the most diverse clades of eukaryotes and are an essential component in aquatic and terrestrial food webs [1–4]. They are some of the most complex single-celled eukaryotes in terms of their cell structure and have even more intricate genomic architectures than multicellular eukaryotes. Ciliates of the genus *Spirostomum* Ehrenberg, 1834 (Class: Heterotrichea) are large cells (150–4,000  $\mu\text{m}$  in length) that inhabit fresh and brackish water [5–8]. The genus *Spirostomum* comprises ten morphospecies characterized by a vermiform, cylindrical, and laterally flattened cell; ciliary rows that spiral during contraction; and a long collecting canal leading to a posterior contractile vacuole [9–15]. These “worm-like” ciliates are considered model organisms for studying ciliate ecology and physiology due to their relative abundance and sensitivity to abiotic factors [7, 16, 17]. It was recently reported that many *Spirostomum* species use a rhodoquinone-dependent pathway under low oxygen conditions for anaerobic fumarate reduction [18, 19]. *Spirostomum* species are suitable for endosymbiosis research as they can harbor either or both bacteria and eukaryotes in their cytoplasm [20]. Several Rhizobiales and Rickettsiales-like bacteria have been found in *Spirostomum* species; *S. semivirescens* is the only *Spirostomum* species reported to host *Chlorella*-like endosymbiotic green algae [18, 21–25]. A unique endosymbiotic system was also found in *Pseudoblepharisma*, which is phylogenetically sister and morphologically similar to *Spirostomum*. More specifically, *P. tenue* harbors contrasting photosynthetic symbionts in its cytoplasm, purple bacteria and green algae [26, 27]. Therefore, understanding the evolution of *Spirostomum* species is also key to shedding light on the evolution of the extraordinarily rare purple-green photosymbiosis of *P. tenue*.

Several authors have previously attempted to resolve the phylogenetic relationships in the genus *Spirostomum* using morphological characters and a single or a few genes [10–14]. These single- and multigene analyses suggested the presence of cryptic species within the genus *Spirostomum*. For example, *S. minus* exhibits

considerable diversity at the molecular level and may contain at least two morphologically indistinguishable cryptic species. *Spirostomum teres* also appears to show case species crypticity, and has further been claimed to be a polyphyletic taxon [11, 13]. Nuclear ribosomal RNA gene sequences may not have enough phylogenetic signals to reveal species boundaries, which were not unambiguously resolved at the molecular level using both the primary and secondary structures of the rRNA gene sequences [12]. In our recent work, we increased the taxon sampling and used two protein-coding genes, e.g., nuclear alpha-tubulin and mitochondrial cytochrome *c* oxidase subunit 1 (CO1), to delimit *Spirostomum* species [13]. We also used the Bayesian multispecies coalescent (MSC) model to estimate a species tree, which allowed us to resolve the phylogenetic relationships between *Spirostomum* species and delimit species based on molecular data. In the present study, we improved the taxon sampling, utilized transcriptomic data information from 37 *Spirostomum* specimens, and used the MSC model to further understand the species complex structure and the evolutionary relationships within the genus *Spirostomum*.

Here, we generated 25 new single-cell transcriptomes from *Spirostomum* specimens collected from freshwater and brackish water environments in South Korea and the USA. We used these, together with 12 additional publicly available transcriptomes, belonging to seven morphospecies (*S. ambiguum*, *S. caudatum*, *S. minus*, *S. teres*, *S. semivirescens*, *S. subtilis*, and *S. yagiui*), to infer a robust phylogenetic tree and investigate the species boundaries of the genus *Spirostomum*. To do this, we performed concatenation- and coalescent-based analyses on a dataset that encompasses 37 *Spirostomum* specimens and 265 protein-coding genes. We also assembled a pipeline to reconstruct the nuclear 18S rRNA gene, ITS1-5.8S-ITS2 region, and 28S rRNA gene from transcriptomic data, and constructed the rRNA operon gene tree with improved taxon sampling to confirm the broad topological framework. Our study represents the first attempt to understand the phylogeny and species boundaries of the genus *Spirostomum* using single-cell transcriptome data.

## Materials and methods

### Sample collection, isolation, and culturing

Samples were collected from various sites, including fresh and brackish waters (Table 1). After transportation to the laboratory, the collected samples were immediately stored in a cool box and examined directly under a dissecting microscope. Single cells of each ciliate morpho-species were isolated, washed in sterile distilled water, and used to establish new cultures. The cultures were maintained at 20–24 °C in Petri dishes, and wheat grains were periodically added to stimulate the growth of prey bacteria. Culture dishes contained filtered in situ water or commercial mineral water (Evian, France) for freshwater species, and sterile seawater for brackish water species. Specimens from each clonal culture were examined under a light microscope, using bright-field and differential interference contrast (DIC) optics (Axio Imager A1; Carl Zeiss, Germany). Species characterization and identification were performed according to the following guidelines [11–14, 28–30] (Additional file 1: Fig. S1).

### Single-cell preparation

Single cells were isolated using a capillary glass pipette under a dissecting microscope. The cells were washed three times in sterile distilled water through successive transfers in PYREX 6-well depression plates, thereby diluting the original medium and removing eukaryotic and bacterial contaminants. Ciliates from brackish water habitats were washed three times in sterile in situ water. In the final step, cells were washed with distilled water and transferred to 1.5 mL microcentrifuge tubes for DNA extraction and 0.2 mL PCR tubes for single-cell transcriptome amplification. Cells were frozen at –80 °C or processed immediately.

### Single-cell DNA extraction

DNA was extracted using the RED Extract-N-Amp Tissue PCR Kit (Sigma, St. Louis, MO) following the protocol of Shazib et al. [13]. The nuclear 18S rRNA gene, ITS1–5.8S–ITS2 region, and 28S rRNA gene were amplified by polymerase chain reaction (PCR) using the TaKaRa Ex Taq polymerase kit (TaKaRa Korea Biomedical Inc., Seoul, South Korea). Polymerase chain reaction (PCR) cycling parameters and primer information were described in Shazib et al. [13]. Products of PCR reactions were directly sequenced on an ABI 3730 automated sequencer (Macrogen Inc., Seoul, South Korea) using seven sequencing primers (Additional file 2: Table S1). Sequence fragments from individual samples were checked, trimmed, and assembled into contigs using Geneious ver. Prime 2019 (Biomatters, <http://www.geneious.com/>) [31].

### Single-cell transcriptomes

Single-cell whole transcriptome amplifications (scWTA) were performed from frozen cells (–80 °C) or cells freshly picked from the clonal cultures. Whole transcriptome amplifications (WTAs) for the individual ciliates were generated according to the Smart-seq2 protocol [32] and SMART-Seq v4 Ultra Low Input RNA Kit (TaKaRa Biomedicals, Seoul, South Korea) with 17 cycles of cDNA amplification for each cell. In principle, both methods are selective for polyadenylated RNA, which allows the generation of full-length cDNA. After transcriptome amplification, cDNA was further purified using the ProNex® Size-Selective Purification System (Promega, WI, USA) according to the manufacturer's protocol. Following ProNex® purification, samples were quantified using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, USA). Sequencing libraries were constructed using the Nextera XT DNA Library Preparation Kit (Illumina, FC-131-1024) and the Nextera XT Index Kit (Illumina, FC-131-1001). Sequencing was performed on HiSeq 2500 (100 bp PE) and NovaSeq 6000 (150 bp PE) instruments at Thera-gen Bio (Seoul, South Korea).

Whole transcriptome amplification for three *Spirosomum* specimens (Sm\_U6, Sm\_U25, Su\_U10) from the Katz lab was performed using the SMART-Seq® v4 Ultra® Low Input RNA Kit for Sequencing (TaKaRa Bio USA, Inc., Mountain View, CA, USA). Amplified cDNA was purified using the AMPure XP PCR Purification system (Beckman Coulter Life Sciences, Indianapolis, IN, USA) and quantified using a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). Illumina sequencing libraries were constructed using the Nextera XT DNA Library Preparation Kit (Illumina, FC-131-1096) and the Nextera XT Index Kit v2 Set A (Illumina, FC-131-2001). These samples were then sequenced on a HiSeq 4000 instrument (Illumina) at the Institute for Genome Sciences, University of Maryland (Baltimore, USA).

### High-throughput sequencing data processing

We assessed the quality of the sequenced raw reads using FastQC v.0.11.5 [33], and trimmed adapters and removed low-quality sequences with BBTools v.38.86 [34], and used q24 as the minimum quality and minimum length, 125 bp for 150 bp and 300 bp PE reads, and 70 bp for 100 bp PE reads. Transcriptome assembly was carried out using rnaSPAdes, included in the SPAdes package v.3.14.1 [35] and subsequently processed by a series of custom Python scripts from PhyloToL v.4.3.1 (PhyloToL part 1- Post assembly Pipeline) [36], available at <https://github.com/Katzlab>. The post-assembly pipeline removes contaminants (rRNA and non-eukaryotic sequences) and translates the nucleotide sequences with >200 bp after assigning amino acids to gene families (GFs) with an appropriate genetic code. Further, we assessed the

**Table 1** Summary of *Spirostomum* species and transcriptomes included in the present work

MAC pattern	Species	Popula- tion ID	sc RNA-seq protocol	Raw reads (M)	Eukaryotic Contigs*	BUSCO Analy- sis (%)	GenBank accession #	Country	Habitat	Refer- ence
Moniliform	<i>S. ambiguum</i>	Sa_C1	Total RNA extraction	61	16,523	87.7	SRR12125180	China	Fresh water	Mukhtar et al., [19]
Moniliform	<i>S. ambiguum</i>	Sa_U1	SMART-Seq v4	48	9,567	90.7	SRR10512970	USA	Fresh water	Yan et al., [36]
Moniliform	<i>S. ambiguum</i>	Sa_K1	SMART-Seq v4	48	7,421	80.1	SRR29149931	South Korea	Fresh water	This study
Moniliform	<i>S. ambiguum</i>	Sa_K2	SMART-Seq v4	53	8,531	86.5	SRR29149930	South Korea	Fresh water	This study
Moniliform	<i>S. ambiguum</i>	Sa_K3	SMART-Seq v4	106	8,510	88.9	SRR29149919	South Korea	Fresh water	This study
Moniliform	<i>S. ambiguum</i>	Sa_K4	Smart-seq2	128	9,515	88.3	SRR29149913	South Korea	Fresh water	This study
Moniliform	<i>S. ambiguum</i>	Sa_K5	Smart-seq2	124	7,070	81.3	SRR29149912	South Korea	Fresh water	This study
Moniliform	<i>S. ambiguum</i>	Sa_K6	Smart-seq2	114	8,140	87.2	SRR29149911	South Korea	Fresh water	This study
Moniliform	<i>S. ambiguum</i>	Sa_K7	Smart-seq2	48	8,580	84.1	SRR29149910	South Korea	Fresh water	This study
Ellipsoid	<i>S. caudatum</i>	Sc_K1	Smart-seq2	70	2,078	29.3	SRR29149909	South Korea	Brackish water	This study
Moniliform	<i>S. minus</i>	Sm_U1	SMART-Seq v4	59	3,789	63.7	SRR10513010	USA	Fresh water	Yan et al., [36]
Moniliform	<i>S. minus</i>	Sm_U6	SMART-Seq v4	13	7,085	82.5	SRR29149908	USA	Fresh water	This study
Moniliform	<i>S. minus</i>	Sm_U25	SMART-Seq v4	36	2,091	45.1	SRR29149926	USA	Fresh water	This study
Moniliform	<i>S. minus</i>	Sm_K1	Smart-seq2	54	4,969	72.6	SRR29149907	South Korea	Fresh water	This study
Moniliform	<i>S. aff. minus</i>	Sm_K2	Smart-seq2	73	11,040	93	SRR29149929	South Korea	Fresh water	This study
Moniliform	<i>S. minus</i>	Sm_K3	Smart-seq2	71	8,456	83	SRR29149928	South Korea	Fresh water	This study
Moniliform	<i>S. minus</i>	Sm_K4	Smart-seq2	85	10,188	88.3	SRR29149927	South Korea	Fresh water	This study
Moniliform	<i>S. sp.</i>	Ss_E1	Smart-seq2	12	7,799	85.5	SRR7141210	Sweden	Fresh water	Hines et al., [18]
Moniliform	<i>S. semivirescens</i>	Sv_E1	Smart-seq2	2	2,743	55.6	SRR7141204	England	Fresh water	Hines et al., [18]
Moniliform	<i>S. semivirescens</i>	Sv_E2	Smart-seq2	4	6,454	77.8	SRR7141205	England	Fresh water	Hines et al., [18]
Moniliform	<i>S. semivirescens</i>	Sv_E3	Smart-seq2	4	5,473	74.9	SRR7141206	Sweden	Fresh water	Hines et al., [18]
Moniliform	<i>S. semivirescens</i>	Sv_E4	Smart-seq2	3	5,102	76	SRR7141207	Sweden	Fresh water	Hines et al., [18]
Moniliform	<i>S. semivirescens</i>	Sv_E5	Smart-seq2	1	3,052	62	SRR7141208	Sweden	Fresh water	Hines et al., [18]
Moniliform	<i>S. semivirescens</i>	Sv_E6	Smart-seq2	3	7,285	84.7	SRR7141209	Sweden	Fresh water	Hines et al., [18]
Moniliform	<i>S. subtilis</i>	Su_U10	SMART-Seq v4	12	9,684	91.8	SRR29149925	USA	Fresh water	This study
Moniliform	<i>S. subtilis</i>	Su_C1	Total RNA extraction	51	17,617	91.2	SRR12125179	China	Fresh water	Mukhtar et al., [19]
Moniliform	<i>S. subtilis</i>	Su_K1	Smart-seq2	56	9,110	86	SRR29149924	South Korea	Fresh water	This study
Moniliform	<i>S. subtilis</i>	Su_K2	Smart-seq2	80	8,591	85.4	SRR29149923	South Korea	Fresh water	This study
Moniliform	<i>S. subtilis</i>	Su_K3	Smart-seq2	51	10,216	93	SRR29149922	South Korea	Fresh water	This study
Ellipsoid	<i>S. teres</i>	St_C1	Total RNA extraction	53	14,595	94.2	SRR12125178	China	Fresh water	Mukhtar et al., [19]
Ellipsoid	<i>S. teres</i>	St_K1	Smart-seq2	54	5,336	72.5	SRR29149921	South Korea	Fresh water	This study
Ellipsoid	<i>S. teres</i>	St_K3	SMART-Seq v4	53	7,929	84.9	SRR29149920	South Korea	Fresh water	This study
Elongated	<i>S. yaguii</i>	Sy_K1	Smart-seq2	54	6,037	78.4	SRR29149918	South Korea	Fresh water	This study
Elongated	<i>S. yaguii</i>	Sy_K2	SMART-Seq v4	58	6,396	80.7	SRR29149917	South Korea	Fresh water	This study
Elongated	<i>S. yaguii</i>	Sy_K3	Smart-seq2	165	7,594	84.8	SRR29149916	South Korea	Fresh water	This study
Elongated	<i>S. yaguii</i>	Sy_K4	Smart-seq2	109	7,558	81.3	SRR29149915	South Korea	Fresh water	This study
Elongated	<i>S. yaguii</i>	Sy_K5	Smart-seq2	113	7,118	87.8	SRR29149914	South Korea	Fresh water	This study
Moniliform	<i>Anigsteinia</i> sp.	As_U1	SMART-Seq v4	183	2,877	33.9	SRR10512978	USA	Marine water	Yan et al., [36]

M=million.; SMART-Seq v4=scRNA-seq kit, Smart-seq2=In-house protocol for scRNA-seq; BUSCO analysis=Completeness+Partial, \* = number of eukaryotic contigs after PhyloToL pipeline part 1 (PLT1) (see materials and methods)



completeness of the translated sequences using BUSCO ver. 5.4.7 (Alveolata dataset -l alveolata\_odb10, and proteome mode -m protein) [37] (Table 1; Additional file 2: Table S2).

#### **Nuclear rRNA gene reconstruction, taxonomic assignment, and tree construction**

We sequenced the nuclear rRNA gene operon (18S, ITS1-5.8S-ITS2, 28S) of 19 specimens using the Sanger method (Additional file 2: Tables S3-S4). In addition, we reconstructed the rRNA gene sequences from other *Spirostomum* and *Anigsteinia* specimens by mapping reads from their respective raw transcriptomes against a set of reference sequences downloaded from NCBI (87 reliable rRNA gene sequences of the genera *Spirostomum* and *Anigsteinia*) (see Shazib et al. [13]) (Fig. 1C). We tested various mapping software, including Bowtie 2 [38], BMAP [39], BWA [40], and HISAT2 [41], to assess their influences on the accuracy of the reconstructed contigs. After read mapping to the rRNA genes, three different read sets with differing quality thresholds (all reads, q20, and q40) were used for read assembly with rnaSPAdes [35, 42] (Additional file 2: Table S5). We aligned all the contigs reconstructed using various read mappers and reference sequences to identify errors and artifacts (Additional file 1: Fig. S2). To construct the rRNA gene sequence data matrix, all reliable *Spirostomum* sequences were aligned with the newly obtained sequences using the MUSCLE algorithm in Geneious ver. Prime 2019 [43]. Ambiguous sequences and poorly aligned columns were removed from the data matrix, resulting in a final multiple sequence alignment (MSA) of 2,459 bp which was then analyzed with IQ-TREE ver. 2.2.0.3, and 1,000 replicates to construct a maximum likelihood (ML) tree [44]. The TIM2e + I + G4 model was identified as the best substitution model based on the Bayesian information criterion (BIC) calculated in IQ-TREE ver. 2.2.0.3. Phylogenetic trees were visualized and edited using FigTree ver. 1.4 (<http://tree.bio.ed.ac.uk/software/figtree/>). The reliability of branch support was assessed using ultrafast bootstrap approximation (UFBoot), and the Shimodaira Hasegawa-like approximate likelihood ratio test (SH-aLRT) with 1,000 replicates in both methods [45, 46]. A clade's reliability was considered significant only if the UFboot support and SH-aLRT values were  $\geq 95\%$  and  $\geq 80\%$ , respectively [45, 47]. However, a clade with  $< 70\%$  UFBoot was considered to have low support, whereas a clade with 70–95% UFBoot was considered to have moderate support [46].

#### **Mitochondrial CO1 gene reconstruction and genetic divergence analyses**

We reconstructed mitochondrial cytochrome *c* oxidase subunit 1 (CO1) gene sequences from the WTA data. Out

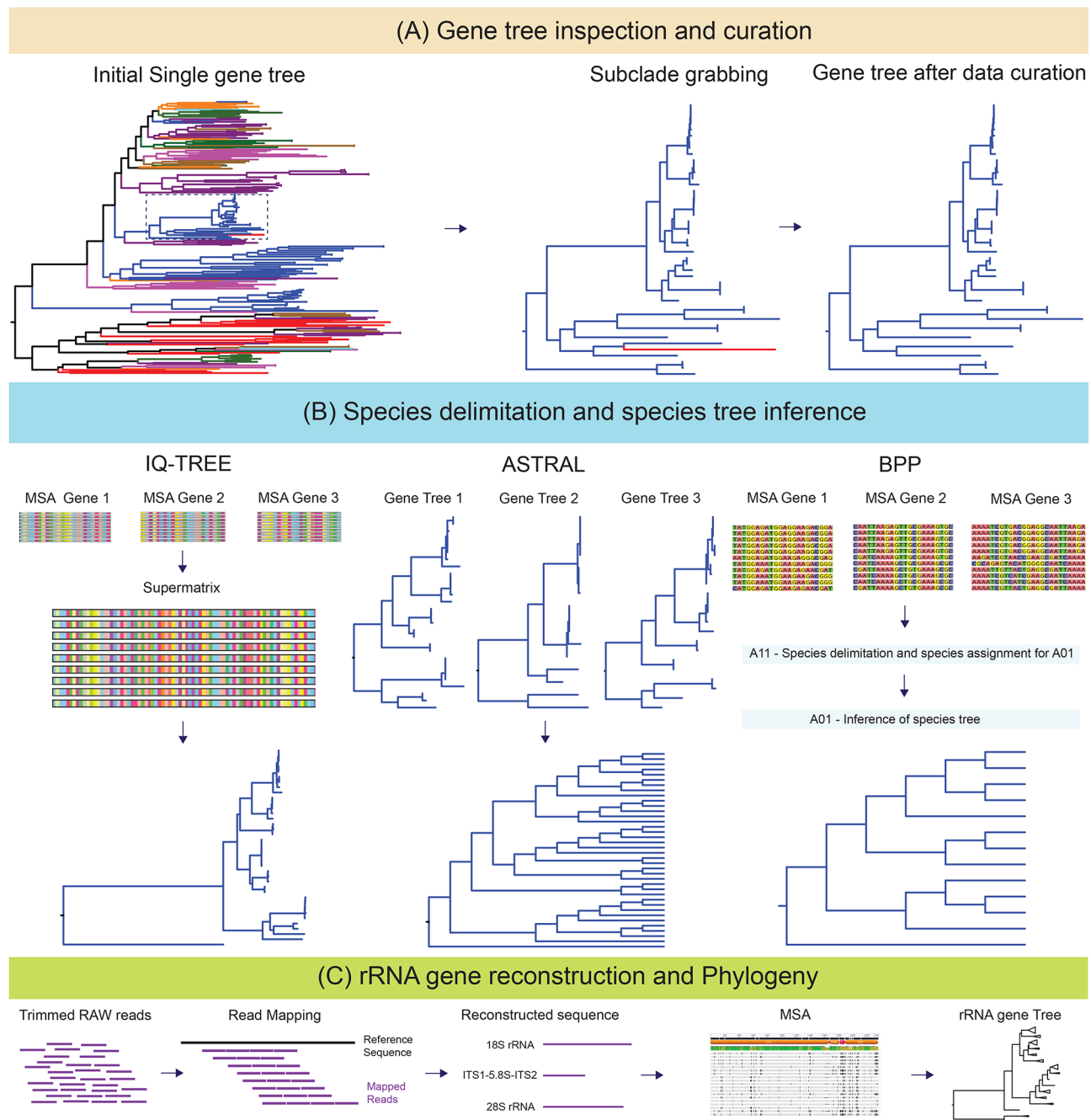
of 37 specimens, we were able to construct CO1 contigs for 19 specimens due to the low abundance of organelle (e.g., mitochondria and chloroplast) reads in the single-cell RNA sequencing (scRNA-seq) data [48]. To construct a phylogenetic tree, we included 49 *Spirostomum* specimens and one *Anigsteinia* sp. as an outgroup. The CO1 gene sequence dataset consisted of 38 sequences from NCBI GenBank (MK721498–MK721535, Additional file 2: Table S4) and 12 reconstructed CO1 gene sequences (sequences  $> 400$  bp only) obtained using read mapping and transcriptomic data (Additional file 2: Tables S3, S6). We aligned the CO1 gene sequences using the MUSCLE algorithm and the protozoan mitochondrial genetic code in Geneious ver. Prime 2019. Maximum likelihood (ML) analyses were performed using a final data matrix with 699 nucleotides and IQ-TREE ver. 2.2.0.3 using the TIM2e + F + I + G4 best fitted-model [49]. We used the ultrafast bootstrap approximation (UFBoot) with 1,000 replicates [46, 47] and the SH-like approximate likelihood ratio test (SH-aLRT), also using 1,000 replicates [45]. Intra- and interspecific pairwise *p*-distances, and the number of nucleotide differences, were calculated using MEGA X with the pairwise deletion option (Additional file 2: Tables S7-S8) [50].

#### **Phylogenomic dataset construction**

All publicly available *Spirostomum* transcriptomes were downloaded from NCBI GenBank as of April 2021, processed, and added to the PhyloToL database (Table 1; Additional file 2: Table S4). Initially, we generated 1,245 gene trees using PhyloToL and assessed transcript diversity in each family for each species (Fig. 1A). Bacterial and eukaryotic contamination was identified by analyzing single-gene trees, which were summarized by a custom Python script (ContaminationBySisters.py). Next, we inspected paralogous families in these initial single-gene trees and selected the 1,141 best clades (subtrees) containing only target sequences using the custom Python script (GetBestSequences.py) (Fig. 1A). Sequences of short lengths and low coverage ( $\text{cov} \leq 10$ ) were removed prior to alignment and tree reconstruction. After manual hand curation, we selected 265 subtrees with sequences that include single-copy genes for each specimen (gene tree inspection and curation steps in Fig. 1; Additional file 2: Table S9). The PhyloToL scripts used above are available at <https://github.com/Katzlab/PhyloToL-6/tree/main/Utilities>.

#### **Concatenation-based phylogenomic approach**

The well-curated datasets of 265 orthologous genes from 38 specimens were aligned using MAFFT [51]. All MSAs were inspected, and gaps were removed using trimAl (-gt 0.9) [52]. The concatenation of the 265 MSAs into a supermatrix was performed using Geneious ver. Prime



**Fig. 1** Schematic workflow for phylogenomic analysis and reconstruction of nuclear rRNA genes from single-cell transcriptomic data. **(A)** Initial gene trees were constructed using the taxon/gene-rich pipeline PhyloToL. Each tree was inspected, and subclades containing the maximum number of target taxa were extracted. Subclades were manually checked for non-target taxa. Curated gene trees with single-copy genes from all taxa were used to prepare the data matrix. **(B)** For concatenation analysis, each dataset with well-curated sequences were aligned and 265 datasets were concatenated into an amino acid supermatrix for maximum likelihood (ML) phylogeny. The ASTRAL method used 265 protein-coding gene trees to infer a species tree. Species delimitation and species tree inference were performed using BPP with 265 well-aligned datasets of nucleotides **(C)** Nuclear rRNA gene sequences were reconstructed from raw single-cell RNA sequencing (scRNA-seq) data using read mapping software (HISAT2 and Bowtie 2). Reconstructed sequences were aligned, and rRNA gene phylogeny was inferred using the ML approach. Detailed methods are provided in the Materials and methods section

2019 [31]. The final supermatrix comprises 62,043 unambiguously aligned amino acid residues (Fig. 1B). Performing phylogenetic analysis using the site-heterogeneous CAT-GTR model in PhyloBayes is often computationally

demanding for a dataset of 265 protein-coding sequences. Therefore, only maximum likelihood (ML) analyses were performed for the large dataset. The ML phylogenetic analyses were conducted using IQ-TREE ver. 2.2.0.3 with

the LG + C60 + F + G4 profile mixture model [49]. Branch support was estimated using the ultrafast bootstrap approximation (UFBoot) with 1,000 replicates [46, 47] and the SH-like approximate likelihood ratio test (SH-aLRT) also with 1,000 replicates [45].

To visualize potential phylogenetic conflicts or noise in the concatenated dataset of 265 protein-coding genes, a neighbor-net network analysis was performed with uncorrected distances in SplitsTree CE [53, 54]. The reliability of the splits was assessed using 1,000 replicates [55].

### Coalescent-based species delimitation and species tree analyses

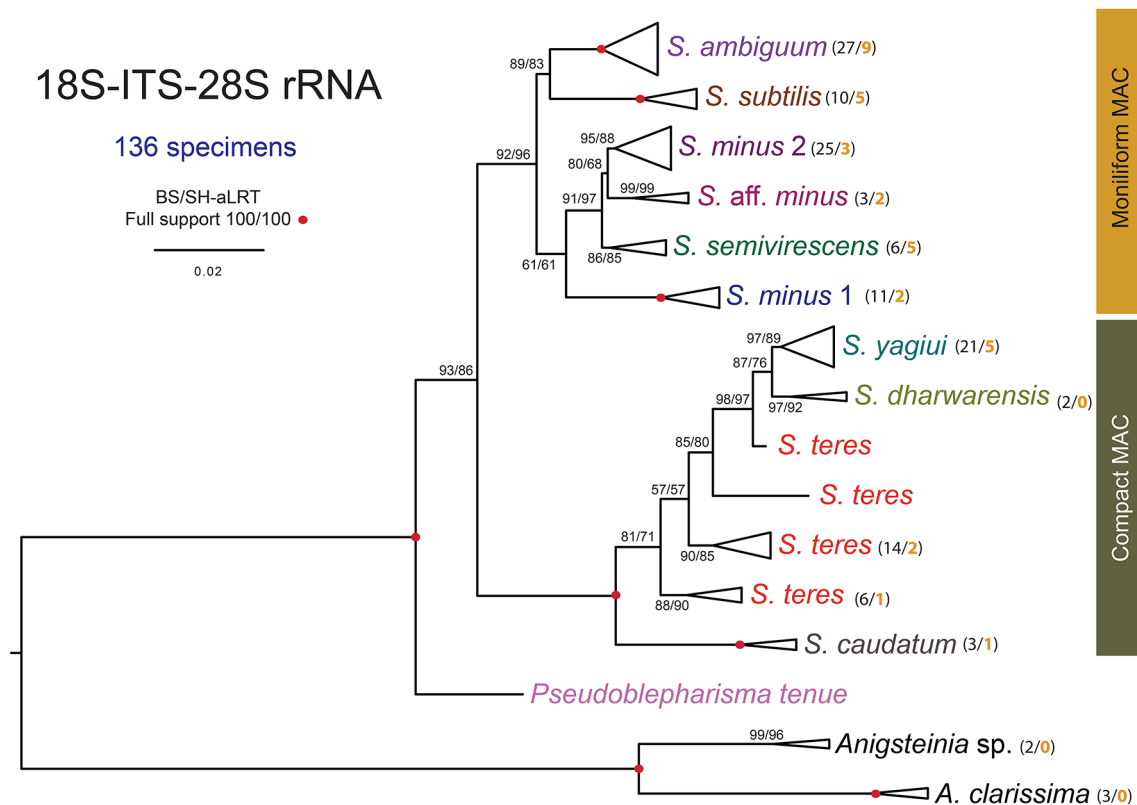
Discordance between gene and species trees occurs due to several biological factors, such as incomplete lineage sorting (ILS) and introgression. Such discordance can be minimized by using coalescent-based methods [56]. We used two different approaches to estimate the species tree using the multispecies coalescent (MSC) model. First, we used the Accurate Species TRee ALgorithm (ASTRAL) III ver. 5.7.1 software, which uses a multispecies coalescent (MSC) model to account for incomplete lineage sorting (ILS) [57] and inferred the species tree from the protein-coding gene trees. This method is widely used, and the accuracy of the species tree relies on the given error-free input gene trees [58]. We constructed individual gene trees from 265 protein-coding gene alignments using the maximum likelihood (ML) approach in IQ-TREE ver. 2.2.0.3. These single-gene trees were used by ASTRAL to infer a species tree, and the certainty of the tree nodes was quantified using local posterior probability (LPP) [59] (Fig. 1B).

Second, we used the Bayesian Phylogenetics and Phylogeography (BPP) ver. 4.6.2 software [60, 61] which uses the multispecies coalescent (MSC) model under the Bayesian framework and estimates the posterior probabilities (PP) for the alternative species tree and a species delimitation hypothesis [60, 62]. Using the reversible-jump Markov chain Monte Carlo (rjMCMC) algorithm and a set of priors, BPP can combine information from many loci and estimate a reliable species tree even when each locus has a weak phylogenetic signal [63, 64].

For species delimitation, we performed the A11 analysis (species delimitation = 1 and species tree = 1) in BPP, which accounts for joint species delimitation and species-tree estimation. The software uses nucleotide sequences, as synonymous codon changes in the nucleotide sequences provide more information when working with closely related species [65]. Therefore, we used MSAs of DNA sequences for BPP analyses, and each alignment file was used as a single-locus input file for BPP (Fig. 1B).

The concatenated phylogenomic tree of 265 protein-coding genes was used as a guide tree for species delimitation (A11). However, species tree estimation and species delimitation should not be affected by the starting tree if there are no convergence and mixing problems [66, 67]. Large datasets, such as the 265 loci one used here, can cause problems with MCMC convergence and mixing, leading to inconsistent results [60, 61, 68] (Additional file 2: Table S9). To ensure consistency and proper mixing across different analyses initially, we first used a subset dataset of 100 loci, each with more than 70% taxon occupancy, and the length ranged from 200 bp to 1,000 bp after removing any sites with missing data (Additional file 2: Table S10). Using the 100 loci dataset, we set up six sets of priors to explore the sensitivity of the BPP analysis using  $\alpha = 3$  in the inverse-gamma (InvG) as a diffuse prior and  $\alpha = 21$  as an informative prior [60] (Additional file 2: Table S11). After an initial assessment using six different prior settings, we assigned the priors  $\theta \sim \text{InvG}(3, 0.002)$ , with a mean of 0.001, and  $\tau \sim \text{InvG}(3, 0.4)$ , with a mean of 0.2. We used reversible-jump Markov Chain Monte Carlo (rjMCMC) analyses and algorithm 0 and the fine-tuning parameter was set to  $e = 10$  to ensure good mixing in the reverse jump (rj) algorithm. The rjMCMC was set to 100,000 generations with burnin = 10,000 and a sample frequency of 5. A posterior probability of 0.95 or higher was interpreted as strong support, indicating that the branch or clade is well-supported by the data within the chosen model. Each analysis was run at least twice, and convergence was assessed by examining consistency between runs. We also ran BPP on 265 loci using the same parameters for the 100 loci dataset. Both datasets will generate consistent results if the MCMC chains are properly mixed and the phylogenetic signals in both datasets are consistent.

We performed A01 analysis (species delimitation = 0 and species tree = 1) with BPP [61, 67] for species phylogeny and partitioned all 38 specimens into 13 “species” following the results of the species delimitation analyses from A11. Species tree inferences using the A01 method under the MSC model are more efficient at estimating the correct species tree than concatenation and summary-based methods when large datasets are used [60]. We used the uniform rooted tree prior (species-modelprior = 1) and excluded all ambiguous sites (clean-data = 1). The ML tree from the concatenated 265-gene dataset was used as the guide tree for this analysis, and the same prior settings were used for species delimitation (A11). To check the reliability of the results, we ran each analysis at least twice. Each analysis was run for 100,000 generations, sampling every 5th generation with 10,000 burnin generations. Posterior probabilities greater than 0.95 were considered highly supported lineage.



**Fig. 2** Maximum likelihood (ML) phylogenetic tree of the genus *Spirostomum*. The topology is based on 136 18S-ITS-28S rRNA gene operon sequences (2,459 bp) and the TIM2e + I + G4 model. Monophyletic clades were collapsed for clarity (see also Additional file 1: Fig. S3). Numbers in parentheses indicate the total number of specimens (in black font) in a lineage and the number of specimens with transcriptomic data used for phylogenomic analyses (in bold orange font). Numbers at nodes are ML bootstrap values and SH-aLRT from 1,000 replicates. Red solid circle on nodes indicates maximum support. The scale bar corresponds to the number of nucleotide substitutions

We used DensiTree to visualize the agreements in the trees generated during the MCMC run of the A01 BPP analysis [69]. All trees generated in the A01 analysis (200,000 MCMC trees, from two runs) were exported as Newick format files and visualized in DensiTree.

## Results

### Ciliate transcriptome amplification using scRNA-seq

We amplified transcriptomes from 25 specimens using the SMART-Seq V4 kit and Smart-seq2 protocols (see Methods, Table 1). Both methods rely on the amplification of polyadenylated mRNA and have previously been successfully employed on single-cell eukaryotes [70–72]. Sequencing results from 35 specimens using the Smart-seq technique showed that both methods performed similarly, and data on contig number plus BUSCO analyses varied across specimens and sequencing depths (Table 1; Additional file 2: Table S2).

### Construction of nuclear rRNA genes from HTS data

We reconstructed rRNA genes by recruiting reads mapped to references and then reconstructing a sequence for each specimen (Fig. 1C; Additional file 2: Table S6).

We found that the contigs generated using BWA from *S. subtilis* (Su\_K1) WTA data had poor quality, with over 50% sequence mismatch nucleotides against the reference sequences. HISAT2 produced the longest contig, with over 99% concordance with the reference sequences; Bowtie 2 and BMap generated the second and third longest contigs, respectively (Additional file 1: Fig. S2). HISAT2 and Bowtie 2 showed comparable results (only two nucleotide mismatches, and over 99.9% similarity) to the references. BMap showed some mismatched or erroneous sequences in the assembled contig. Meanwhile, q40 retrieved the most reliable reads except BWA, and reads from the other three mappers (BMap, Bowtie 2 and HISAT2) generated reliable and single assembled contigs (Additional file 2: Table S5). Our results indicated that HISAT2 and Bowtie 2, with quality scores of 40, were the best for constructing rRNA genes from ciliate WTA data.

### Phylogenetic tree of nuclear rRNA genes

In the rRNA operon gene tree, *Spirostomum* was divided into two major clades that correspond with the macronuclear morphology (Fig. 2; Additional file 1: Fig. S3). The



rRNA gene tree generally supported the monophyly of all *Spirostomum* morphospecies except *S. minus* and *S. teres*. However, most of the phylogenetic relationships among *Spirostomum* morphospecies were either poorly or moderately supported.

The first clade, all of which shared a moniliform macronucleus (MAC), united *S. ambiguum*, *S. minus*, *S. aff. minus* (Latin word *affinis*, abbreviation: aff. = having affinity with, but no identical to), *S. semivirens*, and *S. subtilis* with moderate support (92% UFBoot and 96% SH-aLRT) (Fig. 2). *Spirostomum minus* split into two clades, and the monophyly of *S. minus* clade 1 was fully supported, whereas *S. minus* clade 2 formed a sister relationship with *S. aff. minus* clade, with moderate statistical support (80% UFBoot and 68% SH-aLRT). *Spirostomum aff. minus* clade comprised three specimens (SKS255, Sm\_K2, Ss\_E1) and formed a highly supported clade. *Spirostomum aff. minus* (Sm\_K2), which morphologically resembles *S. minus*, and the rRNA gene tree suggested that all three specimens were close to *S. minus* clade 2 (Fig. 2; Additional file 1: Fig. S3). *Spirostomum subtilis* formed a sister clade to *S. ambiguum* (89% UFBoot and 83% SH-aLRT), and *S. semivirens* fell sister to the *S. minus* 2 + *S. aff. minus* clade with moderate support (91% UFBoot and 97% SH-aLRT).

The second clade consisted of lineages with a compact macronucleus (MAC) and included *S. teres*, *S. yagiui*, *S. dharwarensis*, and *S. caudatum*. The monophyletic origin of species with a compact macronucleus was fully supported, with *S. caudatum* as an early divergent taxon within the second clade (Fig. 2). All the taxa in the second major clade were monophyletic, but *S. teres* appeared to be non-monophyletic and split into four clades. Furthermore, *Pseudoblepharisma tenue* (Pt\_MG) with a compact MAC, appeared to have diverged first among all *Spirostomum* species, although the topology was only moderately supported (93% UFBoot and 86% SH-aLRT). All *Anigsteinia* species formed a fully supported clade, consistent with the classification as the sister to all *Spirostomum* spp. in all our previous phylogenetic analyses [73–75].

#### Genetic divergence analyses and phylogeny of CO1 gene sequences

A total of 49 mitochondrial cytochrome *c* oxidase subunit 1 (CO1) gene sequences were used for the genetic distance analyses. The initial alignment of the CO1 sequences included 753 nucleotides; therefore, only the sequences over 400 bp were retained for the pairwise distance analyses. The mean intraspecific genetic diversity in the morphospecies *S. ambiguum* and *S. teres* was comparatively high in the CO1 region (11.30%, and 11.20%, respectively), and a significant genetic diversity was also observed in *S. yagiui* (7.48% mean intraspecific distance) (Additional file 2: Table S8). In contrast, *S. minus* and *S.*

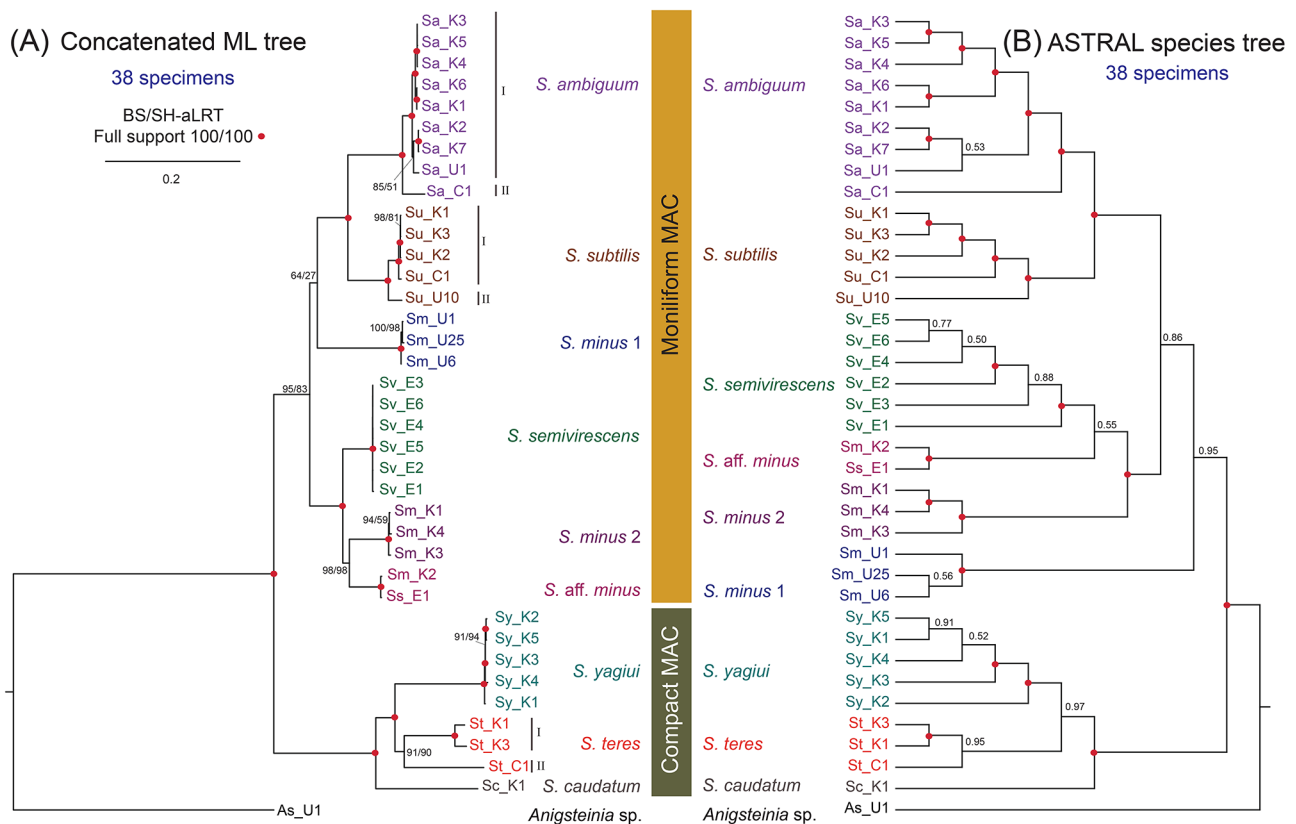
*subtilis* had low nucleotide variability (0.41% and 1.72% mean intra-specific distance, respectively). Nevertheless, the CO1 gene sequences were only available from *S. minus* clade 2, and the genetic divergence among 17 specimens of *S. minus* clade 2 was up to 2.29%. Furthermore, the intraspecific genetic divergence in the CO1 gene sequences was significantly high in *S. ambiguum* (up to 21.83%), suggesting the presence of a cryptic species complex. The CO1 gene sequence variability within *S. subtilis* was up to 2.30%, and only CO1 gene sequences were available for *S. subtilis* lineage I specimens (Additional file 1: Fig. S4, Additional file 2: Table S8). However, the genetic variability in the *S. yagiui* CO1 gene was up to 15.22% (six specimens, Additional file 1: Fig. S4) and 2.91% (two specimens, Sy\_K3 and Sy\_K4, Additional file 2: Table S7).

Our current phylogenetic tree, inferred from 49 *Spirostomum* specimens, was congruent with our previous results based on 37 *Spirostomum* specimens [13]. To summarize, the monophyletic origin of the *Spirostomum* species with a moniliform macronucleus, as well as the monophyly of those with a compact macronucleus, was not supported in the CO1 gene tree (Additional file 1: Fig. S4). The monophyletic origin of almost all *Spirostomum* morphospecies was recognized and fully supported in the CO1 gene trees except for *S. teres*. In the *S. ambiguum* clade, two highly diverged subclades were recognized. Meanwhile, the *S. teres* individuals were divided into three subclades, and individual SKS787 diverged from the other *S. yagiui* individuals.

#### Concatenated analyses

Maximum likelihood (ML) analyses of the concatenated supermatrix of 265 protein-coding genes yielded a topology with strong support for most branches (Fig. 3A). The majority of the currently defined morphospecies were monophyletic, except for *S. minus*. Both major clades identified in the rRNA gene trees were fully supported (Figs. 2 and 3A).

The concatenated tree fully supported a moniliform clade of *Spirostomum*, with five taxa (*S. ambiguum*, *S. minus*, *S. aff. minus*, *S. semivirens*, and *S. subtilis*) (Fig. 3A). *Spirostomum ambiguum* and *S. subtilis* specimens formed a clade with full support. However, genetically divergent individuals were identified from both the *S. ambiguum* and *S. subtilis* clades, suggesting the existence of cryptic species in both lineages. *Spirostomum minus* formed two well-supported clades: the *S. minus* clade 1 received high support and was sister to the *S. ambiguum* + *S. subtilis* clade with low support (64% UFBoot and 27% SH-aLRT). Meanwhile, the *S. minus* clade 2 formed a monophyletic relationship with *S. aff. minus*, which received high statistical support (98% UFBoot and 98% SH-aLRT), and clustered together with



**Fig. 3** Phylogenomic relationships of *Spirostomum* species from 38 specimens and 265 protein-coding genes. **(A)** Maximum likelihood (ML) tree topology of a phylogenomic matrix comprising 62,043 unambiguously aligned amino acid residues was inferred using IQ-TREE under the LG + C60 + F + G4 model. **(B)** ASTRAL species tree topology inferred from 265 gene trees. One *Anigsteinia* sp. was used as an outgroup. Numbers at nodes are ML bootstrap values and SH-aLRT from 1,000 replicates (IQ-TREE) and local posterior probability (LPP) for ASTRAL are given at the nodes. Red solid circle on nodes means maximum support. The scale bar corresponds to the number of nucleotide substitutions. (C=China; E=Europe; K=South Korea; U=USA)

*S. semivirens* with full support. In the second major clade (compact macronucleus), *S. yagiui* was monophyletic and well supported. Individuals assigned to *S. teres* formed two distinct intraspecific lineages, with *S. teres* (St\_C1, isolates from China) forming a clade with two other *S. teres* (St\_K1 and St\_K3, isolates from South Korea) that received moderate statistical support (91% UFBoot and 90% SH-aLRT). *Spirostomum caudatum* was the deepest branching taxon in the second major clade and formed a fully supported sister relationship with a clade comprising *S. yagiui* and *S. teres*.

Based on the concatenated dataset of 265 protein-coding genes, the neighbor-net network (Fig. 4) indicated the existence of non-vertical signals, which may have been caused by deep incomplete lineage sorting, recombination, hybridization, or convergent evolution. Specifically, two conflicting noisy splits were identified. One was observed in the *S. minus* clade 1, which was linked to *S. minus 2* + *S. aff. minus* + *S. semivirens* and *Anigsteinia* through small net-like splits. The other split was in the *S. teres* clade, where two *S. teres* lineages were connected

with *S. yagiui* in a mesh-like manner, further complicating the relationships.

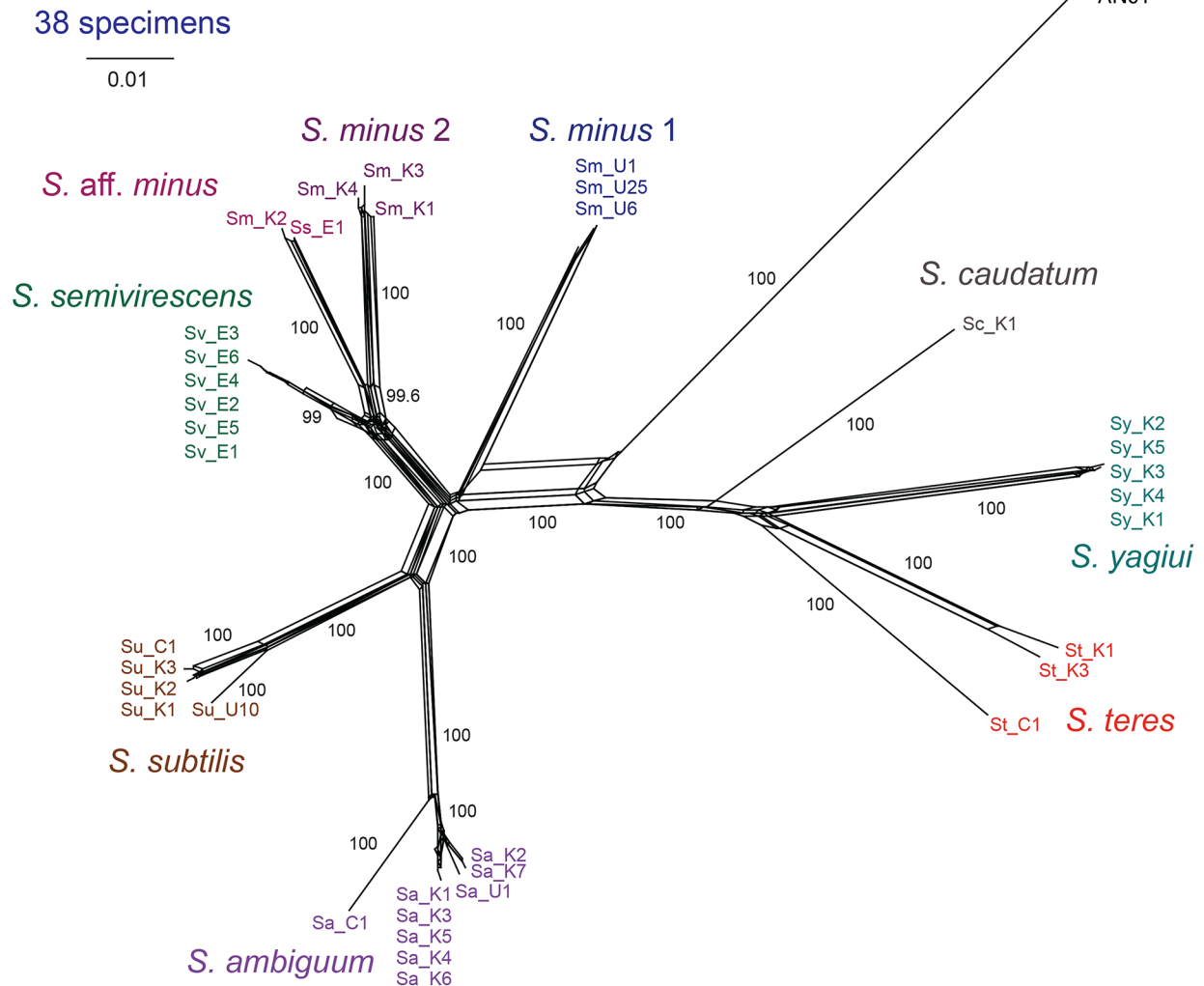
#### Coalescent-based species tree using ASTRAL

To estimate the species tree using ASTRAL, we used the same number of genes as in the concatenated IQ-TREE. Thus, ASTRAL estimated a species tree from the 265 gene trees (Fig. 3B). The resulting coalescent tree revealed relationships similar to those of the concatenated ML tree, and the discordant results received moderate/poor statistical support. However, in the ASTRAL species tree, *S. minus* clade 1 was sister to the other taxa of the moniliform MAC clade and received a higher local posterior probability (LPP = 0.95) (Fig. 3B). Notably, *S. aff. minus* was more closely related to *S. semivirens*, with very poor statistical support (LPP = 0.55).

#### Coalescent-based species delimitation and species tree analysis using BPP

To determine whether the genetically divergent lineages found in the IQ-TREE and ASTRAL analyses represented distinct species, we utilized the MSC model in BPP. Two

## Neighbor-net network



**Fig. 4** Phylogenetic network inferred from the concatenated alignment of 265 protein-coding genes using the Neighbor-Net algorithm and uncorrected distances. Numbers along edges are bootstrap support values coming from 1,000 replicates. Values < 50% are not shown

different datasets (100 and 265 loci) were used for species delimitation analysis using the A11 algorithm (Table 2, Additional file 2: Tables S9–S10). The species delimitation approach using BPP supported a species delimitation scheme in which 13 lineages (including outgroup taxa) were identified as distinct species (lineages), with full statistical support (posterior probability = 1.00) in both the 100 loci and 265 loci datasets (Fig. 5A). Furthermore, we treated all these *Spirostomum* specimens as 12 distinct lineages and used them for species tree analysis using BPP.

The species trees inferred from the 100 and 265 loci using BPP (A01 analysis) were generally concordant with the concatenated alignment tree (Figs. 3A and 5A). In the MSC approach, the tree topology received full statistical support for all nodes, except for the relationships between two *S. teres* lineages, which received poor

statistical support in both BPP species trees (100 and 265 loci dataset; Table 2). The 265 loci dataset received higher support than the 100 loci dataset (posterior probability 0.79 vs. 0.54), which might indicate that data size influenced the node support (Fig. 5A, Additional file 2: Table S11).

The DensiTree (Fig. 5B) of all 200,000 MCMC trees from 265 loci revealed conflicting relationships between *S. teres* and *S. yagiui*. The results showed that the *S. teres* lineages (I and II) were closely related and formed a monophyletic relationship, as shown in the blue- and red-colored majority gene trees. On the other hand, the fewer (green-colored) gene trees suggested that the *S. teres* lineages also formed a close relationship with *S. yagiui*, which may have been caused by high levels of incomplete lineage sorting during rapid speciation events, leading to incongruence between gene trees and species trees.

**Table 2** List of analyses performed in the present study

Analysis	No. of specimens	Description
18S-ITS-28S rRNA tree	136	130 <i>Spirostomum</i> + 1 <i>Pseudoblepharisma</i> + 5 <i>Anigsteinia</i> specimens
CO1 gene tree	50	49 <i>Spirostomum</i> + 1 <i>Anigsteinia</i> specimens
Concatenated tree	38	Concatenated data matrix of 265 protein-coding genes. 37 <i>Spirostomum</i> + 1 <i>Anigsteinia</i> specimens
Neighbor-net network	38	Concatenated data matrix of 265 protein-coding genes. 37 <i>Spirostomum</i> + 1 <i>Anigsteinia</i> specimens
ASTRAL species tree	38	Gene trees of 265 protein-coding genes were used to estimate the species tree under the MSC model. 37 <i>Spirostomum</i> + 1 <i>Anigsteinia</i> specimens
BPP 265 loci	38	Multiple sequence alignments (MSA) of 265 protein-coding genes were used to estimate the species tree and species delimitation under the MSC model. 37 <i>Spirostomum</i> + 1 <i>Anigsteinia</i> specimens
BPP 100 loci	38	Multiple sequence alignments (MSA) of 100 protein-coding genes were used to estimate the species tree and species delimitation under the MSC model. 37 <i>Spirostomum</i> + 1 <i>Anigsteinia</i> specimens

Discussion

To further understand the evolutionary history and species complex structure of *Spirostomum*, we used transcriptomic data and a supermatrix-based phylogenetic approach to estimate a tree with IQ-TREE and a network with SplitsTree. We then used the MSC model for species tree inference using ASTRAL and performed both species tree and species delimitation analyses using BPP. Our analyses of the transcriptomic data provided insights into the phylogeny of *Spirostomum* and defined the boundaries of its constituent species.

Species delimitation using molecular data analyses

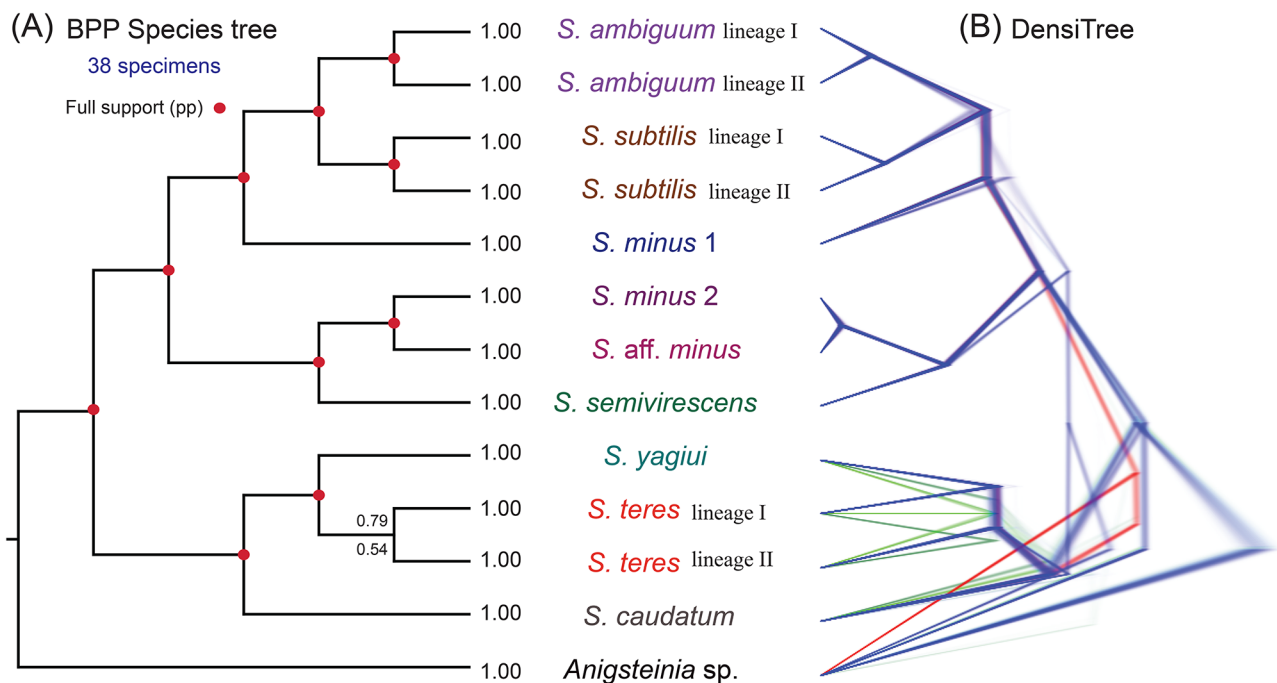
Identifying ciliate species using molecular criteria has consistently posed a challenge [76, 77]. Our results, based on CO1 gene sequences, suggest the presence of cryptic species within several *Spirostomum* morphospecies, which is congruent with our recent findings [13]. For example, the CO1 gene sequences of *S. ambiguum*, *S. teres*, and *S. yagiui* exhibited more than 15% variability, whereas *S. minus* and *S. subtilis* showed just over 2% variability (Additional file 2: Table S8). Genetic divergence thresholds of 4–5% in the CO1 gene are used to distinguish ciliate species, whereas intraspecies divergence typically ranges from 0 to 1% in most ciliate species [76, 78–82]. Based on these thresholds, *S. ambiguum*, *S. teres*, and *S. yagiui* likely comprise cryptic species complexes.

Our rDNA operon tree and phylogenomic analyses did not indicate cryptic species within *S. ambiguum* and *S. subtilis* until we included *S. ambiguum* Sa\_C1 and *S. subtilis* Su\_U10 isolates in our analyses. Furthermore, the coalescent-based species delimitation approach using BPP confirmed Sa\_C1 (*S. ambiguum* lineage II) and Su\_U10 (*S. subtilis* lineage II) isolates as distinct species, while CO1 gene sequences were available only for the *S. ambiguum* lineage I and *S. subtilis* lineage I isolates (Additional file 1: Fig. S4). The high variability in nucleotide CO1 gene sequences likely reflects synonymous mutations within the mitochondrial CO1 insert region, as suggested in our previous study [13]. However, these results need to be interpreted cautiously due to differences in specimen numbers across datasets.

Using the MSC-based analyses, we identified 12 putative species (Fig. 5). *Spirostomum minus* was divided into at least two distinct species, which contrasts with the poor support for this clade in previous analyses that relied on rRNA gene sequences [10–12, 14, 15]. Conserved rRNA gene regions could lead to ambiguous results in taxonomic studies, and no reliable morphological key characters have been identified to differentiate *S. minus* clade species. However, the MSC-based species delimitation approach using transcriptomic data and BPP confirmed two species within *S. minus*. This finding is consistent with our previous study, which was based on five molecular markers [13]. However, we do not know which one corresponds to the originally described *S. minus* species, as data on type specimens are lacking. Therefore, molecular data from type specimens (which do not exist and, therefore, from specimens from the type locality) are needed to resolve this taxonomic uncertainty.

Coalescent-based species delimitation analysis supported the existence of two genetically distinct species within *S. teres*, a taxon found in fresh and brackish water environments and characterized by a compact ellipsoidal macronucleus [11]. Only *S. teres* and *S. caudatum* have an ellipsoidal macronucleus, but these species can be easily distinguished from each other by their body shape (posterior end of the body narrowly rounded vs. tail-like). In previous studies, multiple *S. teres* individuals appeared as non-monophyletic in rRNA gene trees and were split into several clades [10–14]. Even analyses of the secondary structure of the ITS2 molecule, which is extremely short (79–80 bp) within the genus *Spirostomum*, could not clearly differentiate *S. teres* isolates [12]. In this study, only three *S. teres* specimens (St\_K1, St\_K3, St\_C1) with rRNA gene and transcriptomic data were available, and *S. teres* (St\_C1) had both CO1 gene sequence and transcriptomic data (Additional file 1: Figs. S3–S4, Additional file 2: Table S4). However, species delimitation approaches (A11 analysis) using transcriptomic data defined *S. teres*





**Fig. 5** Multispecies coalescent-based species trees inferred using BPP. **(A)** Species tree using A01 approach with datasets of 265 loci and 100 loci from 38 specimens assigned into 13 lineages. Posterior probability (PP) values are given at each node, the upper value at the node indicates the PP of 265 loci and the lower value at the node indicates the PP of 100 loci. Fully supported branches are marked with solid red circles. **(B)** DensiTree plot for 200,000 MCMC trees of 265 loci. Blue, red, and green colors represent the first, second, and third most common topologies, respectively, whereas gray represents other topologies

as two distinct lineages. As with *S. minus*, molecular data from type specimens (type locality) are required to determine the “true” *S. teres*.

#### Phylogenetic relationships among *Spirostomum* species

Our current analyses, based on the phylogenomic approach, enhance our understanding of the evolutionary patterns of *Spirostomum* species. For example, phylogenomic approach supports the deep division of *Spirostomum* into two major clades: clade 1 includes members with moniliform macronucleus (*S. ambiguum*, *S. minus*, *S. aff. minus*, *S. semivirescens*, and *S. subtilis*), while clade 2 members (*S. caudatum*, *S. dharwarensis*, *S. teres*, and *S. yagiui*) possess compact (ellipsoidal or elongated) macronucleus (Figs. 3 and 5). This supports the two major clade species hypothesis [11–14].

The macronuclear pattern within the class Heterotricha is variable. The morphological evolution of the macronucleus is also complex at the genus level [83–85]. According to our nuclear rDNA operon tree (Fig. 2), *Pseudoblepharisma tenue* is closely related to the genus *Spirostomum*, which is consistent with previous studies using 18S rRNA gene as well as multiple genes (rDNA operon + CO1) [26, 86]. Morphologically, *P. tenue* shares similarities with *Spirostomum* ciliates, as the cell spirally contracts under stress and hence can be misidentified as a *Spirostomum* species. Moreover, *P. tenue* is close to *S.*

*teres* with respect to its cell size and ellipsoidal macronucleus [26]. To date, only two species of *Pseudoblepharisma* have been documented, both of which possess an ellipsoidal macronucleus [26, 86]. We hypothesize that a compact macronucleus represents an ancestral state, whereas a moniliform macronucleus might be derived, as observed in other heterotrich genera (*Blepharisma*, *Stentor*) [83–85].

Phylogenomic analyses using both concatenation and coalescent approaches suggest the presence of high ancestral genetic polymorphism within the first major clade (species with moniliform macronucleus) (Fig. 3–5). In contrast to previous hypotheses, the phylogenomic data indicated that the *S. minus* clade 2 is more closely related to *S. semivirescens* than to the *S. minus* clade 1. The concatenated and BPP species trees depicted *S. minus* clade 1 as a sister to the *S. ambiguum*+*S. subtilis* clade (Figs. 3A and 5). Thus, the phylogenomic data, together with improved taxon sampling, provide further insights into understanding the phylogenetic relationships among the species of the first major clade. In previous molecular studies, *S. minus* appeared to be two morphologically cryptic species, and their monophyly and/or separation (*minus* 1 and 2 clades) received poor support based on the rDNA operon phylogeny [13]. However, the coalescent-based approach received moderate support when *S. minus* was split into two distinct



clades (see Fig. 5 in Shazib et al. [13]). Phylogenetic analyses based on five markers were limited to the *S. minus* clade 2; as no data were available for species in *S. minus* clade 1 or *S. semivirescens*. In our phylogenomic analyses, *S. aff. minus*-like species were closely related to *S. minus* clade 2 species. Both the concatenated analysis and coalescent-based species tree supported their monophyletic relationships (Figs. 3A and 5). However, the coalescent-based species delimitation analysis suggested that the *S. minus* clade 2 and *S. aff. minus* represent distinct species. The evolutionary relationship within the clade with a moniliform macronucleus seems to be complex, and our study confirms that the *minus*-like appearance is likely the ancestral trait of the major clade 1 species, and *S. minus*-like species share the ancestral gross body architecture, making morphological delimitation of this group challenging.

In the present study, concatenation and coalescent-based analyses indicated that *S. caudatum* is an early branching lineage within the second clade of *Spirostomum* (with ellipsoidal macronuclei; Figs. 3 and 5). *Spirostomum yagiui* is a monophyletic taxon and forms a sister relationship with *S. teres*. Individuals identified as *S. teres* were non-monophyletic in the rDNA operon tree and split into several clades but were recovered as monophyletic taxa in all phylogenomic analyses with moderate support (Fig. 2–5, Additional file 1: Fig. S3). This may be due to the limited taxon sampling or the influence of the increased data size. Our results are consistent with that of Shazib et al. [13] and support the hypothesis of Boscaro et al. [11] that *S. yagiui* and *S. dharwarensis* (with an elongated macronucleus) originated during the rapid radiation of a *S. teres*-like ancestor with an ellipsoidal macronucleus. Taken together, our data suggest that *S. teres* represents at least two distinct lineages. Transcriptomic or genomic data from other *S. teres* specimens are needed to verify the monophyly of *S. teres* at the phylogenomic level.

## Conclusion

Our phylogenomic species tree and species delimitation approach, using transcriptomic data, allows us to robustly understand the evolutionary relationships between *Spirostomum* species and supports our previous hypothesis [13]. Our results confirm that *Spirostomum* species with a moniliform and compact macronucleus form distinct clades, in agreement with previous studies [11–13]. We also tested the MSC model implemented in BPP on scRNA-seq data from 37 *Spirostomum* specimens for the first time. This confirmed that three distinct *S. minus*-like species and *S. teres* contain multiple distinct lineages. Furthermore, improved taxon sampling revealed cryptic species in *S. ambiguum* and *S. subtilis*. Our exploratory network analyses also suggest the occurrence of

incomplete lineage sorting in the evolutionary history of the genus *Spirostomum*. Overall, our study demonstrated that transcriptomic data together with the MSC-based approach are valuable tools for species tree inference and delimitation in single-celled eukaryotes.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12862-025-02353-3>.

Supplementary Material 1: Additional file 1: figure S1. Six morphospecies of the genus *Spirostomum* isolated from different habitats in South Korea. Morphology of living cells, (A–C) species with moniliform macronucleus, and (D–F) species with compact macronucleus. (A) *Spirostomum ambiguum* has a moniliform macronucleus, the peristome occupying about 50% of the body length, and the posterior body end is truncated. (B) *S. subtilis* has a moniliform macronucleus, the peristome occupies about 35% of the body length, and the posterior body end is truncated. (C) *S. minus* has a moniliform macronucleus, the peristome occupying about 40% of the body length, and the posterior body end is truncated. (D) *S. yagiui* has an elongated macronucleus, the peristome occupying about 42% of the body length, and the posterior body end is truncated. (E) *S. teres* has an ellipsoid macronucleus, the peristome occupies about 38% of the body length, and the posterior body end is truncated. (F) *S. caudatum* has an ellipsoid macronucleus, the peristome occupying about 35% of the body length, and a tail-like posterior body end. MAC, macronucleus; CV, contractile vacuole; CY, cytostome. Scale bars: 500 µm (A, B), 100 µm (C, D, F), 50 µm (E). Additional file 1: figure S2 Reconstruction of the nuclear rRNA gene from *S. subtilis* Su\_K1 transcriptome data using different read mapping software. PCR Sanger sequencing of Su\_K1 was used as a reference sequence for read mapping and contig assembling. BWA read mappers generated small and highly diverged contigs and were not included in this comparison analysis. Only contigs from BMap, Bowtie 2, HISAT2 and q40 map quality were used here. Nucleotides differing from the reference sequence (Su\_K1 Sanger sequencing contig) are highlighted with red box. Additional file 1: figure S3 Maximum likelihood (ML) analysis using IQ-TREE of the genus *Spirostomum*. The topology is based on 136 18S-ITS-28S rRNA gene sequences and 2,459 bp data matrix using the TIM2e+I+G4 model. Numbers at nodes are ML bootstrap values and SH-aLRT from 1,000 replicates. Blue bars indicate the availability of RNA-seq data, while green bars represent the availability of the CO1 gene sequence of a specimen. The scale bar corresponds to the number of nucleotide substitutions. Additional file 1: figure S4 Maximum likelihood (ML) analysis using IQ-TREE of the genus *Spirostomum*. The topology is based on 50 Spirostomidae CO1 gene sequences (49 *Spirostomum* + 1 *Anigsteinia*) and 699 bp data matrix using the TIM2e+I+G4 model. Numbers at nodes are ML bootstrap values and SH-aLRT from 1,000 replicates. Blue bars indicate the availability of RNA-seq data of a specimen. Red solid circle on nodes indicates maximum support. The scale bar corresponds to the number of nucleotide substitutions.

Supplementary Material 2: Additional file 2: table S1 List of primers used for PCR amplification and sequencing of the nuclear rRNA gene. Additional file 2: table S2 BUSCO genome completeness analysis of 38 Spirostomidae ciliates used in this study. Additional file 2: table S3 Characteristics of the nuclear rRNA and mitochondrial CO1 gene sequences of 38 Spirostomidae ciliates used in this study (RM, contig reconstructed using read mapping). Additional file 2: table S4 Dataset and sequence information of 139 Spirostomidae ciliates used for rRNA and phylogenomic analyses (RM, contig reconstructed using read mapping). Additional file 2: table S5 Nuclear rRNA gene reconstruction from WTA data using different mapping algorithm. For the comparisons here, we used *Spirostomum subtilis* (Su\_K1) as a reference. Additional file 2: table S6 Reconstructed nuclear rRNA and mitochondrial CO1 gene sequences using read mapping approach and reference sequences. Additional file 2: table S7 Numbers of nucleotide differences (above diagonal) and pairwise similarities (% below diagonal) of the mitochondrial CO1 gene sequences among 49 *Spirostomum* specimens. Additional file 2: table S8 Intra- and interspecific variability of mitochondrial CO1 gene sequences in 49 *Spirostomum* specimens.

Additional file 2: table S9 List of 265 protein-coding genes and species distribution across gene families (GFs). These 265 genes were utilized for phylogenomic and BPP analyses. Gene Family defined by OrthoMCL number followed by subclade number (e.g., OG5\_126891\_1, where 126891 is the ortholog group, and subclade number 1 contains the maximum number of target taxa). Additional file 2: table S10 List of 100 protein-coding genes and species distribution across gene families (GFs). These 100 genes were utilized for species tree analyses and species delimitation using BPP. Gene Family defined by OrthoMCL number followed by subclade number (e.g., OG5\_126891\_1, where 126891 is the ortholog group, and subclade number 1 contains the maximum number of target taxa). Additional file 2: table S11 Species delimitation (A11) and species tree (A01) analyses were conducted using different prior settings on 100 loci and 265 loci datasets to determine the best prior settings fit the dataset.

## Acknowledgements

We thank Ying Yan and Xyrus Maurer-Alcalá for collecting ciliates, and Angela Jiang, and Emma Schumacher for helping in writing additional scripts for the pipeline in Katz's laboratory, Smith College.

## Author contributions

S.U.A.S. planned the research, collected samples, performed laboratory work, dataset preparation, curation and analysis, writing – original draft, review & editing, and funding acquisition. A.C.L. dataset preparation, writing - review & editing. R.A. laboratory work, writing – review & editing. S.A.M. methodology, writing – review & editing. J.M.L. data analysis, writing - review & editing. L.A.K. planned the research, methodology, writing – original draft, review & editing, and funding acquisition. M.K.S. planned the research, collected samples, methodology, writing – original draft, review & editing, and funding acquisition.

## Funding

This work was supported by the grants from the NRF Korea funded by the Ministries of Science and ICT, and Education of the Republic of Korea under Grants 2017R1D1A2B03032963 (awarded to S.U.A.S) and RS-2021-NR065806 (awarded to M.K.S). L.A.K. was supported by grants from the NIH (R15HG010409) and NSF (OCE-1924570).

## Data availability

Raw transcriptome reads are deposited in GenBank under the BioProject number PRJNA1113686, along with the rRNA gene sequences of species (PQ164953 – PQ164971). The workflow to reconstruct rRNA gene from transcriptomic data is available at <https://github.com/shahed30/SeqToDNA>.

## Declarations

### Human ethics and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare no competing interests.

### Author details

<sup>1</sup>Department of Biological Sciences, University of Ulsan, Ulsan 44610, South Korea

<sup>2</sup>Department of Biological Sciences, Smith College, 44 College Lane, Northampton, MA 01063, USA

<sup>3</sup>University of Massachusetts Amherst, Program in Organismic and Evolutionary Biology, Amherst, MA, USA

<sup>4</sup>Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA

<sup>5</sup>Center for Plant Biology, Purdue University, West Lafayette, IN 47907, USA

<sup>6</sup>Department of Oceanography, Kyungpook National University, Daegu 41566, South Korea

<sup>7</sup>Kyungpook Institute of Oceanography, Kyungpook National University, Daegu 41566, South Korea

Received: 2 August 2024 / Accepted: 7 February 2025

Published online: 27 February 2025

## References

- Clamp JC, Lynn DH. Investigating the biodiversity of ciliates in the 'Age of integration'. *Eur J Protistol*. 2017;61:314–22.
- Dopheide A, Lear G, Stott R, Lewis G. Molecular characterization of ciliate diversity in stream biofilms. *Appl Environ Microbiol*. 2008;74:1740–7.
- Hakenkamp CC, Morin A. The importance of meiofauna to lotic ecosystem functioning. *Freshw Biol*. 2000;44:165–75.
- Lynn DH. The Ciliated Protozoa. Characterization, Classification, and Guide to the Literature. third edition. Springer, Dordrecht; 2008.
- Bradley MW, Esteban GF, Finlay BJ. Ciliates in chalk-stream habitats congregate in biodiversity hot spots. *Res Microbiol*. 2010;161:619–25.
- Finlay BJ. Global dispersal of Free-Living Microbial Eukaryote species. *Science*. 2002;296:1061–3.
- Nalecz-Jawecki G, Sawicki J. Spirotox—a new tool for testing the toxicity of volatile compounds. *Chemosphere*. 1999;38:3211–8.
- Specht H. Aerobic respiration in *Spirostomum ambiguum* and the production of ammonia. *J Cell Comp Physiol*. 1934;5:319–33.
- Jang SW, Kwon C-B, Shin MK. First Records of Two *Spirostomum* Ciliates (Heterotricha: Heterotrichida: Spirostomidae) from Korea. *Anim Syst Evol Divers*. 2012;28.
- Fernandes NM, Silva Neto IDD. Morphology and 18S rDNA gene sequence of *Spirostomum minus* and *Spirostomum teres* (Ciliophora: Heterotricha) from Rio De Janeiro, Brazil. *Zoologia (Curitiba)*. 2013;30:72–9.
- Boscaro V, Carducci D, Barbieri G, Senra MVX, Andreoli I, Erra F, et al. Focusing on Genera to Improve species Identification: revised systematics of the Ciliate *Spirostomum*. *Protist*. 2014;165:527–41.
- Shazib SUA, Vďačný P, Kim JH, Jang SW, Shin MK. Molecular phylogeny and species delimitation within the ciliate genus *Spirostomum* (Ciliophora, Postciliodesmatophora, Heterotricha), using the internal transcribed spacer region. *Mol Phylogenet Evol*. 2016;102:158–77.
- Shazib SUA, Vďačný P, Slovák M, Gentekaki E, Shin MK. Deciphering phylogenetic relationships and delimiting species boundaries using a bayesian coalescent approach in protists: a case study of the ciliate genus *Spirostomum* (Ciliophora, Heterotricha). *Sci Rep*. 2019;9:16360.
- Chi Y, Duan L, Luo X, Cheng T, Warren A, Huang J, et al. A new contribution to the taxonomy and molecular phylogeny of three, well-known freshwater species of the ciliate genus *Spirostomum* (Protozoa: Ciliophora: Heterotricha). *Zool J Linn Soc*. 2020;189:158–77.
- Chi Y, Wang Z, Ye T, Wang Y, Zhao J, Song W, et al. A new contribution to the taxonomy and phylogeny of the ciliate genus *Spirostomum* (Alveolata, Ciliophora, Heterotricha), with comprehensive descriptions of two species from wetlands in China. *Water Biol Secur*. 2022;1:100031.
- Berger H, Foissner W. Illustrated guide and ecological notes to ciliate indicator species (Protozoa, Ciliophora) in running waters, lakes, and sewage plants. In: Steinberg, C., Calmano, W., Klapper, H., Wilken, R.-D, editors, *Handbuch Angewandte Limnologie*. Ecomed Verlag, Landsberg am Lech, 17. Erg.-Lieferung 10/03, pp. 1–160. 2003. pp. 1–160.
- Nalecz-Jawecki G, Wawryniuk M, Giebułtowicz J, Olkowski A, Drobnińska A. Influence of selected antidepressants on the Ciliated Protozoan *Spirostomum ambiguum*: toxicity, Bioaccumulation, and Biotransformation products. *Molecules*. 2020;25:1476.
- Hines HN, Onsbring H, Ettema TJG, Esteban GF. Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical records. *Protist*. 2018;169:875–86.
- Mukhtar I, Wu S, Wei S, Chen R, Cheng Y, Liang C, et al. Transcriptome profiling revealed multiple rquA genes in the species of *Spirostomum* (Protozoa: Ciliophora: Heterotricha). *Front Microbiol*. 2020;11:574285.
- Fokin SI, Schweikert M, Brümmer F, Görtz H-D. *Spirostomum* spp. (Ciliophora, Protista), a suitable system for endocytobiosis research. *Protoplasma*. 2005;225:93–102.
- Akter S, Shazib SUA, Shin MK. *SegnochrobaSpirostomiostomi* gen. nov., sp. nov., isolated from the ciliate *Spirostomum yagiui* and description of a novel family, Segnochrobactraceae fam. nov. within the order Rhizobiales of the class Alphaproteobacteria. *Int J Syst Evol Microbiol*. 2020;70:1250–8.
- Esteban GF, Bradley MW, Finlay BJ. A case-building *Spirostomum* (Ciliophora, Heterotrichida) with zoochlorellae. *Eur J Protistol*. 2009;45:156–8.

23. Fokin SI. Frequency and biodiversity of symbionts in representatives of the main classes of Ciliophora. *Eur J Protistol.* 2012;48:138–48.
24. Kreutz M, Foissner W. The Sphagnum ponds of Simmelried in Germany: a biodiversity hot-spot for microscopic organisms. Shaker Verlag GmbH; 2006. p. 1.
25. Schrällhammer M, Ferrantini F, Vannini C, Galati S, Schweikert M, Görtz H-D, et al. *Candidatus Megaira polyxenophila* gen. nov., sp. considerations of evolutionary history, Host Rangshift Shiearly divergent eukaryotes. *PLoS ONE.* 2013;8:e72581.
26. Muñoz-Gómez SA, Kreutz M, Hess S. A microbial eukaryote with a unique combination of purple bacteria and green algae as endosymbionts. *Sci Adv.* 2021;7:eabg4102.
27. Muñoz-Gómez SA, Hess S. Purple photosymbioses. *Curr Biol.* 2023;33:R167–70.
28. Repak A, Isquith I. The systematics of the genus *Spirostomum* Ehrenberg, 1838. *Acta. Protozool.* 1974;12:325–33.
29. Foissner W, Berger H, Kohmann F. Taxonomische Und ökologische Revision Der Ciliaten Des Saprobiesystems—Band II: Peritrichia, Heterotrichida, Odontostomatida. Informationsberichte Des Bayer. Landesamtes für Wasserwirtschaft. 1992;5/92:1–502.
30. Berger H, Foissner W, Kohmann F. Bestimmung Und Ökologie Der Mikrosaprobien Nach DIN 38 410. Stuttgart, Jena, Lübeck, Ulm: Gustav Fischer; 1997.
31. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics.* 2012;28:1647–9.
32. Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. *Nat Protoc.* 2014;9:171–81.
33. Andrews S, FastQC. A Quality Control Tool for High Throughput Sequence Data. 2010. Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc>
34. Bushnell B, Rood J, Singer E. BBMerge—accurate paired shotgun read merging via overlap. *PLoS ONE.* 2017;12:e0185056.
35. Bushmanova E, Antipov D, Lapidus A, Pribelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *Gigascience.* 2019;8:giz100.
36. Cerón-Romero MA, Maurer-Alcalá XX, Grattepanche J-D, Yan Y, Fonseca MM, Katz LA. PhyloTol: a Taxon/Gene-Rich Phylogenomic Pipeline to explore genome evolution of diverse eukaryotes. *Mol Biol Evol.* 2019;36:1831–42.
37. Manni M, Berkeley MR, Seppely M, Zdobnov EM. BUSCO: assessing genomic data Quality and Beyond. *Curr Protoc.* 2021;1:e323.
38. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* 2012;9:357–9.
39. Bushnell B, BMap. A Fast, Accurate, Splice-Aware Aligner. 2014.
40. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics.* 2009;25:1754–60.
41. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol.* 2019;37:907–15.
42. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol.* 2012;19:455–77.
43. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32:1792–7.
44. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol.* 2020;37:1530–4.
45. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to Estimate Maximum-Likelihood Phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 2010;59:307–21.
46. Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast approximation for Phylogenetic Bootstrap. *Mol Biol Evol.* 2013;30:1188–95.
47. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the Ultrafast bootstrap approximation. *Mol Biol Evol.* 2018;35:518–22.
48. Smith DR, Sanità Lima M. Unraveling chloroplast transcriptomes with ChloSeq, an organelle RNA-Seq bioinformatics pipeline. *Briefings Bioinf.* 2017;18:1012–6.
49. Si Quang L, Gascuel O, Lartillot N. Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics.* 2008;24:2317–23.
50. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing platforms. *Mol Biol Evol.* 2018;35:1547.
51. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013;30:772–80.
52. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 2009;25:1972–3.
53. Huson DH. SplitsTree: analyzing and visualizing evolutionary data. *Bioinformatics.* 1998;14:68–73.
54. Huson DH, Bryant D. Application of phylogenetic networks in Evolutionary studies. *Mol Biol Evol.* 2006;23:254–67.
55. Bryant D. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol.* 2003;21:255–65.
56. Huang J, Flouri T, Yang Z. A Simulation Study to examine the Information Content in Phylogenomic Data Sets under the multispecies Coalescent Model. *Mol Biol Evol.* 2020;37:3211–24.
57. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinf.* 2018;19:153.
58. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics.* 2014;30:i541–8.
59. Sayyari E, Mirarab S. Fast coalescent-based computation of local Branch support from quartet frequencies. *Mol Biol Evol.* 2016;33:1654–68.
60. Flouri T, Jiao X, Rannala B, Yang Z. Species Tree inference with BPP using genomic sequences and the multispecies Coalescent. *Mol Biol Evol.* 2018;35:2585–93.
61. Yang Z. The BPP program for species tree estimation and species delimitation. *Curr Zool.* 2015;61:854–65.
62. Flouri T, Jiao X, Rannala B, Yang Z. A bayesian implementation of the multispecies Coalescent Model with Introgression for Phylogenomic Analysis. *Mol Biol Evol.* 2020;37:1211–23.
63. Shi C-M, Yang Z. Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among Major groups of Gibbons. *Mol Biol Evol.* 2018;35:159–79.
64. Xu B, Yang Z. Challenges in Species Tree Estimation under the multispecies Coalescent Model. *Genetics.* 2016;204:1353–68.
65. Flouri T, Rannala B, Yang Z. A Tutorial on the Use of BPP for Species Tree Estimation and Species Delimitation. In: Scornavacca C, Delsuc F, Galtier N, editors. *Phylogenetics in the Genomic Era. No commercial publisher!* Authors open access book; 2020. p. 5.6:1–5.6:16.
66. Robert CP, Casella G. Monte Carlo Statistical methods. New York, NY: Springer New York; 2004.
67. Yang Z, Rannala B. Unguided species delimitation using DNA sequence data from multiple loci. *Mol Biol Evol.* 2014;31:3125–35.
68. Thawornwattana Y, Seixas FA, Yang Z, Mallet J. Full-likelihood genomic analysis clarifies a Complex History of species Divergence and Introgression: the Example of the erato-sara group of Heliconius butterflies. *Syst Biol.* 2022;71:1159–77.
69. Bouckaert RR. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics.* 2010;26:1372–3.
70. Kolisko M, Boscaro V, Burki F, Lynn DH, Keeling PJ. Single-cell transcriptomics for microbial eukaryotes. *Curr Biol.* 2014;24:R1081–2.
71. Liu Z, Hu SK, Campbell V, Tatters AO, Heidelberg KB, Caron DA. Single-cell transcriptomics of small microbial eukaryotes: limitations and potential. *ISME J.* 2017;11:1282–5.
72. Onsborg H, Tice AK, Barton BT, Brown MW, Ettema TJG. An efficient single-cell transcriptomics workflow for microbial eukaryotes benchmarked on Giardia Intestinalis cells. *BMC Genomics.* 2020;21:448.
73. Shazib SUA, Vd'ačny P, Kim JH, Jang SW, Shin MK. Phylogenetic relationships of the ciliate class Heterotricha (Protista, Ciliophora, Postciliodesmatophora) inferred from multiple molecular markers and multifaceted analysis strategy. *Mol Phylogenet Evol.* 2014;78:118–35.
74. Chen X, Kim JH, Shazib SUA, Kwon CB, Shin MK. Morphology and molecular phylogeny of three heterotrichid species (Ciliophora, Heterotricha), including a new species of *Anigsteinia*. *Eur J Protistol.* 2017;61:278–93. Pt A.
75. Chen X, Shazib SUA, Kim JH, Kim MS, Shin MK. New contributions to *Gruberia lanceolata* (Gruber, 1884) Kahl, 1932 based on analyses of multiple populations and genes (Ciliophora, Heterotricha, Gruberiidae). *Eur J Protistol.* 2018;65:16–30.
76. Doerder FP. Abandoning sex: multiple origins of asexuality in the ciliate *Tetrahymena*. *BMC Evol Biol.* 2014;14:112.
77. Nannery DL, McCoy JW. Characterization of the species of the *Tetrahymena pyriformis* complex. *Trans Am Microsc Soc.* 1976;95:664–82.

78. Kher CP, Doerder FP, Cooper J, Ikonomi P, Achilles-Day U, Kuepper FC, et al. Barcoding *Tetrahymena*: discriminating species and identifying unknowns using the cytochrome c oxidase subunit I (cox-1) Barcode. *Protist*. 2011;162:2–13.
79. Zhao Y, Yi Z, Gentekaki E, Zhan A, Al-Farraj SA, Song W. Utility of combining morphological characters, nuclear and mitochondrial genes: an attempt to resolve the conflicts of species identification for ciliated protists. *Mol Phylogenet Evol*. 2016;94:718–29.
80. Strueder-Kypke MC, Lynn DH. Comparative analysis of the mitochondrial cytochrome c oxidase subunit I (COI) gene in ciliates (Alveolata, Ciliophora) and evaluation of its suitability as a biodiversity marker. *Syst Biodivers*. 2010;8:131–48.
81. Chantangsi C, Lynn DH, Brandl MT, Cole JC, Hetrick N, Ikonomi P. Barcoding ciliates: a comprehensive study of 75 isolates of the genus *Tetrahymena*. *Int J Syst Evol Microbiol*. 2007;57:2412–25.
82. Obert T, Zhang T, Rurik I, Vďačný P. First molecular evidence of hybridization in endosymbiotic ciliates (Protista, Ciliophora). *Front Microbiol*. 2022;13.
83. Schmidt SL, Foissner W, Schlegel M, Bernhard D. Molecular phylogeny of the Heterotrichea (Ciliophora, Postciliodesmatophora) based on small subunit rRNA gene sequences. *J Eukaryot Microbiol*. 2007;54:358–63.
84. Taher MA, Kabir AS, Shazib SUA, Kim MS, Shin MK. Morphological Redescriptions and Molecular Phylogeny of Three *Stentor* Species (Ciliophora: Heterotrichea: Stentoridae) from Korea. *Zootaxa*. 2020;4732:zootaxa.4732.3.6.
85. Thamm M, Schmidt SL, Bernhard D. Insights into the phylogeny of the Genus *Stentor* (Heterotrichea, Ciliophora) with special emphasis on the evolution of the Macronucleus based on SSU rDNA data. *Acta Protozool*. 2010;2010:149–57.
86. Hines HN, McCarthy PJ, Esteban GF. A Case Building Ciliate in the Genus *Pseudoblepharisma* found in Subtropical Fresh Water. *Diversity*. 2022;14:174.

## Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.