Protist, Vol. xx, xxx–xxx, xx 2017
http://www.elsevier.de/protis
Published online date xxx

Protist

ORIGINAL PAPER

Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical Records

Hunter N. Hines[a,c,1,2], Henning Onsbring[b,1,2], Thijs J.G. Ettema[b], and Genoveva F. Esteban[a]

[a]Bournemouth University, Faculty of Science and Technology, Department of Life and Environmental Sciences, Poole, Dorset BH12 5BB, UK
[b]Department of Cell and Molecular Biology, Science for Life Laboratory, Uppsala University, SE-75123 Uppsala, Sweden
[c]Harbor Branch Oceanographic Institute, Florida Atlantic University, Fort Pierce, FL 34946, USA

Submitted May 4, 2018; Accepted August 9, 2018
Monitoring Editor: Eric Meyer

The ciliate *Spirostomum semivirescens* is a large freshwater protist densely packed with endosymbiotic algae and capable of building a protective coating from surrounding particles. The species has been rarely recorded and it lacks any molecular investigations. We obtained such data from *S. semivirescens* isolated in the UK and Sweden. Using single-cell RNA sequencing of isolates from both countries, the transcriptome of *S. semivirescens* was generated. Phylogenetic analysis of the rRNA gene cluster revealed both isolates to be identical. Additionally, rRNA sequence analysis of the green algal endosymbiont revealed that it is closely related to *Chlorella vulgaris*. Along with the molecular species identification, an analysis of the ciliates' stop codons was carried out, which revealed a relationship where TGA stop codon frequency decreased with increasing gene expression levels. The observed codon bias suggests that *S. semivirescens* could be in an early stage of reassigning the TGA stop codon. Analysis of the transcriptome indicates that *S. semivirescens* potentially uses rhodoquinol-dependent fumarate reduction to respire in the oxygen-depleted habitats where it lives. The data also shows that despite large geographical distances (over 1,600 km) between the sampling sites investigated, a morphologically-identical species can share an exact molecular signature, suggesting that some ciliate species, even those over 1 mm in size, could have a global biogeographical distribution.

Key words: Protist; stop codon; RNA-seq; anaerobic respiration; symbiotic algae; Heterotrich; biogeography

Introduction

The genus *Spirostomum* Ehrenberg, 1834, currently comprises eight species of ciliates found globally in fresh and brackish water habitats (Boscaro et al. 2014). These single-celled eukary-

[1]Corresponding authors;
[2]These authors contributed equally.
e-mail hhines@bournemouth.ac.uk (H.N. Hines), henning.onsbring@icm.uu.se (H. Onsbring).

https://doi.org/10.1016/j.protis.2018.08.001
1434-4610/© 2018 Published by Elsevier GmbH.

otes can be found in high abundances, and some species can obtain body sizes that are visible to the naked eye, e.g. *S. ambiguum*. The ciliate *S. semivirescens* is a large (usually over 1 mm in length) protist with densely-packed endosymbiotic green algae that resemble *Chlorella* (Esteban et al. 2009a; Kreutz and Foissner 2006). Despite its large size and conspicuous bright green color, it is still largely absent from published global ciliate species lists, with only a few sparse records of the species. This makes the species an ideal candidate in which to investigate its biogeography. Neither the algal endosymbiont nor the ciliate host have benefitted from molecular examinations, even though this is an active area of research for other species of ciliates, especially *Paramecium bursaria* (Hoshina and Imamura 2008; Hoshina and Kusuoka 2016).

Different ways to adapt to anoxic environments have been described among ciliate species (Esteban et al. 2009b; Finlay et al. 1983; Yarlett et al. 1981). Ciliates are also known for their wide diversity of genetic codes, where stop codons are recoded to be translated into amino acids (Knight et al. 2001). To get insight into how such traits have evolved, large-scale data sets, which cover the whole genome content of the species of interest, are needed. In this study we generate such data by RNA sequencing at the single cell level. *S. semivirescens* was specifically targeted, as it has been missing from earlier examinations of this well-studied genus (Boscaro et al. 2014; Shazib et al. 2014). In the research presented here, *S. semivirescens* was isolated from freshwater habitats in the UK and Sweden. Transcriptome data was also generated from another *Spirostomum* species to complement our investigation into *S. semivirescens*. Data generated in this study is a necessary piece for improved understanding of the *Spirostomum* genus and the whole suborder Heterotricha. Molecular data for *S. semivirescens* is provided for the first time, along with the first molecular identification of the symbiotic algae associated with this species.

Results

The *Spirostomum semivirescens* found thriving in the UK's anoxic ditch sediments matched exactly the previously described records of occurrence and morphology from a fen pond ~100 meters away (Esteban et al. 2009a). Densities of up to 15 cells per mL were observed, with cells being maintained in natural samples for one week after collection. When left undisturbed for about one hour the ciliate builds an external case or coating; the ciliate is contractile, and retracts into the casing if disturbed. This could provide protection during a dispersal event (e.g. wind, birds). *S. semivirescens* was not observed to form cysts; however, there are records of other *Spirostomum* species being able to form cysts precursors (Ford 1986, and own observations) in cultures of *S. ambiguum*). Cells were always found to be densely packed with bright green endosymbiotic algae (Fig. 1).

S. semivirescens from Swedish study sites was immediately identified from the freshly-collected samples from both locations as being morphologically identical to the UK strain, and the additional diagnostic literature (Esteban et al. 2009a). *S. semivirescens* found was 800–1,500 μm in length and 25–45 μm in width with more than 50 cells being measured (Fig. 1). Densities of up to 30 cells per mL were observed, but more often were found to be 5 per mL from both locations, each showing productive ciliate concentrations, with green *Frontonia* reaching up to ~1,000 per mL, especially from an algal mat sampled in Stadsskogen. The *S. semivirescens* cells were observed to build a loose casing, contractile, and always densely-packed with endosymbiotic green algae. The casing observed in the Swedish specimens of *S. semivirescens* was larger (wider) and less densely packed than observed in the UK, perhaps due to different composition of available sediments and/or to the length of time that the ciliate samples were left undisturbed, allowing them to build a larger protective coat. The samples were collected during a warm period in August 2015, but *S. semivirescens* was later found to thrive during much colder periods in winter, even being regularly recovered from the habitat under a ~15 cm thick layer of ice.

Sequencing and Transcriptome Quality

For all seven transcriptomes (Table 1) a total of 9.3 Gb sequencing data was generated. Low levels of contamination were indicated by MEGAN that assigned less than 5% of the contigs as prokaryotic in each assembly. Less than 4% of the contigs were classified as Viridiplantae, despite the high number of algal endosymbionts in *S. semivirescens*. For 17% of the 23,933 transcripts in the co-assembly more than 10 reads from each of the six *S. semivirescens* mapped and for 49% of the transcripts 10 reads or more from at least three different replicates mapped. Based on this level of consistency between the transcriptomes and the similar relative expression level of transcripts between replicates (Supplementary Material Fig.

Please cite this article in press as: Hines HN, et al. Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical Records. Protist (2017), https://doi.org/10.1016/j.protis.2018.08.001

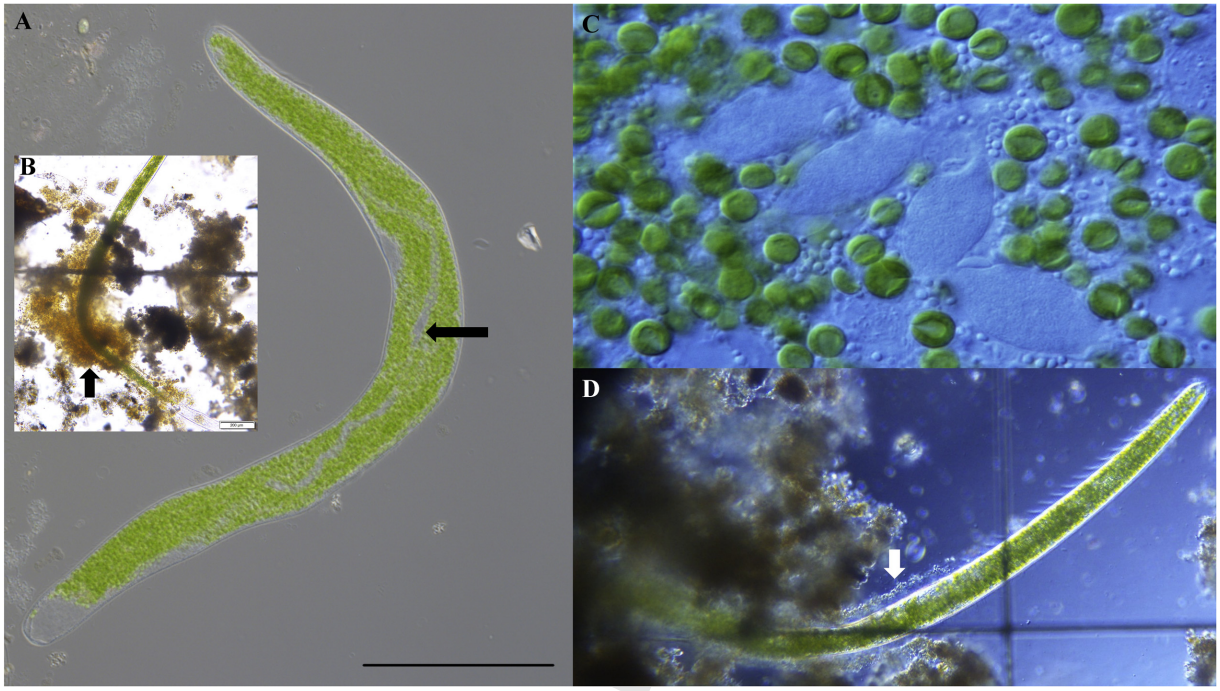

Figure 1. *Spirostomum semivirescens* in vivo micrographs of the specimens collected from UK and Sweden. **A**: *S. semivirescens* collected from Sweden. Note the long moniliform macronucleus (arrow) running along the center of the ciliate. The cell is packed with endosymbiotic green algae, a diagnostic characteristic. Scale bar 200 μm. **B**: *S. semivirescens* collected from UK. The cell is shown here after leaving it undisturbed for a few hours on a counting chamber, as evident by the thick casing (arrow) it has produced. **C**: *S. semivirescens*' nuclear apparatus from a UK cell. Note the nodes of the macronucleus, and the small micronucleus at top of the oval shapes. The densely-packed endosymbiotic green algae are clearly in view in this cell. **D**: *S. semivirescens* collected from the UK. This specimen has begun to build its coating, which is the thin layer (arrow) around the center of the ciliate.

Table 1. Transcriptome data generated. The sum of the pro- and eukaryotic fraction of contigs will not be 100%. This is due to the high number of contigs where diamond could not find any hit in the nr database and therefore no MEGAN assignment could be done.

Species	Sampling site	Contigs*	Prokaryotic contigs (%)**	Eukaryotic contigs (%)**
S. semivirescens	Stadsskogen	22396	2.6	44
S. semivirescens	Stadsskogen	34101	2.3	49
S. semivirescens	Oxhagen	25329	2.5	48
S. semivirescens	Oxhagen	31377	2.9	51
S. semivirescens	Dorset	17644	4.3	46
S. semivirescens	Dorset	27952	2.8	53
Spirostomum sp.	Stadsskogen	35543	3.2	47

*Number of contigs in assembly after 95% CD-HIT reduction.
**Fraction of contigs in assembly assigned by MEGAN as pro- or eukaryotic.

S1) we decided to use the co-assembly (Table 2) in downstream analysis.

Phylogeny

The phylogenetic analysis of the *Spirostomum* genus was based on a concatenation of 18S rRNA, 28S rRNA and the internal transcribed spacer region in between the two rRNA sequences. The tree topology showed that members of major *Spirostomum* clades grouped together in the same way as observed in earlier studies (Boscaro et al. 2014; Shazib et al. 2014). However, relationships between these clades changed and *S. teres*

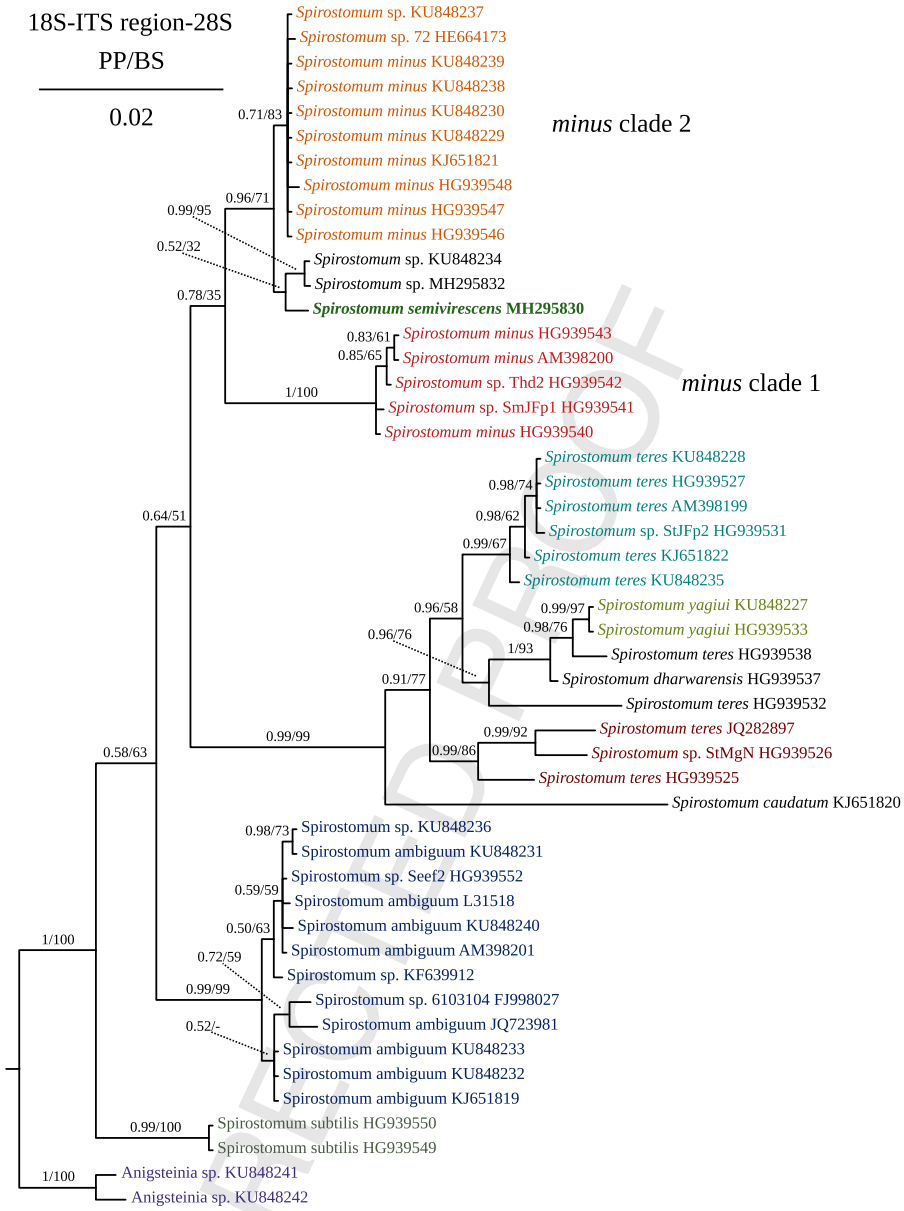

Figure 2. Phylogenetic relationships for the *Spirostomum* genus. Topology is inferred from a concatenation of 18S rRNA, 28S rRNA and the ITS region with PhyloBayes using the CAT + GTR model, bootstrap values from the maximum likelihood analysis is mapped on the tree as well. The branch length of the outgroup is reduced by ten times for visual purposes. For taxa where the different rRNA sequences have different accession numbers, i.e. are not available as one contig, the accession number for the 18S rRNA gene is displayed. The concatenated rRNA sequences assembled from each of the *S. semivirescens* replicates were identical or within the expected error from polymerase, therefore only one *S. semivirescens* taxon was placed in the tree (see text). *minus* clade 1 and 2 is marked on the tree based on how these two clades are referred to in earlier literature (Boscaro et al. 2014; Shazib et al. 2014).

together with *S. yahiui*, *S. dharwarensis* and *S. caudatum* was placed as a sister clade next to *S. minus*. The regions used to infer the phylogeny shared only a total of one mismatch in the 18S rRNA region, no mismatches in the ITS region and three in the 28S rRNA region between the six replicates. Based on the number of PCR cycles used prior to sequencing this is in line with what could be expected from polymerase errors. Therefore, only one *S. semivirescens* taxon is placed in the tree

Please cite this article in press as: Hines HN, et al. Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical Records. Protist (2017), https://doi.org/10.1016/j.protis.2018.08.001

Table 2. Assembly statistics for the co-assembly of six *S. semivirescens* replicates.

Total number of contigs	23933
Transcriptome size (Mb)	31.5
Average contig length (nucleotides)	1317
Longest contig (nucleotides)	12455
Shortest contig (nucleotides)	210
Average GC content (%)	51
Total amount of data assembled (Gb)	6.3

(Fig. 2). The phylogenetic analysis indicates that *S. semivirescens* is most closely related to the members of the clade earlier referred to as "*minus* clade 2" (Boscaro et al. 2014). This clade consists of *Spirostomum minus* and an unnamed species first discovered by Shazib et al. (2014). The unnamed species in "*minus* clade 2" were placed as sister taxon with high support together with the colorless *Spirostomum* species found during this study which was not identified prior to sequencing.

No algal 18S rRNA could be found in any of the transcriptome assemblies, despite the high number of algal endosymbionts in *S. semivirescens*. It is possibly that lysis of the algae was inefficient leading to poor transcriptome coverage of the endosymbiont. However, in five of the six *S. semivirescens* transcriptome assemblies a 28S rRNA gene could be found with high identity to *Chlorella vulgaris*. Transcriptome data of another ciliate that harbors similar endosymbiotic algae, *Stentor polymorphus*, has been generated for species from the same pond in Stadsskogen sampled in this study (Onsbring et al. 2018). If the algae observed in the *S. semivirescens* transcriptomes are contamination, the same contamination could potentially be observed in the *S. polymorphus* transcriptome. No 28S rRNA gene identical to the assumed *S. semivirescens* endosymbiont could be found in the *S. polymorphus* data. Instead another 28S rRNA gene with high identity to *Chlorella vulgaris* was found. Except for the 28S rRNA with high identity to *Chlorella vulgaris*, no other algae related rRNA sequence was detected more than once in each transcriptome. Both the *Spirostomum* and *Stentor* algal endosymbiont sequences branched together with *Chlorella vulgaris* in the tree (Fig. 3).

Identification of Anaerobic Metabolism

Members of the *Spirostomum* genus are often encountered in the oxygen-depleted sediment layers of water bodies. Thriving in these habitats would require the ciliate to be able to respire under anoxic conditions. Therefore, proteins involved in previously described anaerobic respiration pathways (Müller et al. 2012; Stairs et al. 2015) were searched for in the *S. semivirescens* transcriptome to get better insight into its anaerobic lifestyle. A tblastn search could identify a match for the rhodoquinone biosynthesis protein RquA of the bacterium *Rhodospirillum rubrum* (WP_011390975). The match had an 86% query coverage and 47% sequence identity. The putative RquA sequence from *S. semivirescens* carried a 21 amino acid mitochondrial targeting sequence at the N terminus with a predicted probability to be targeted to the mitochondrion of 86% (Claros and Vincens 1996). This is consistent with previous reports that have shown eukaryotic RquA has predicted mitochondrial localization (Stairs et al. 2014). The assumption that the identified potential RquA sequence in the *S. semivirescens* transcriptome is indeed a true RquA is further supported by the presence of a 9 amino acid motif (Lonjers et al. 2012). This motif contains glutamine and valine in RquA instead of aspartate and glycine used by UbiE and UbiG at corresponding positions. UbiE and UbiG are methyl transferases involved in the ubiquinone biosynthesis, which has a high sequence similarity to RquA. Based on the tblastn search we could not find any evidence for the presence of hydrogenosomes, pyruvate formate lyase activity or dissimilatory nitrate reduction, which have been found in other microbial eukaryotes (Müller et al. 2012; Stairs et al. 2015).

Codon Usage

The investigation of the codon usage showed that TAA, TAG and TGA are used by *S. semivirescens* as stop codons. A relationship between gene expression levels and TGA frequency could be observed where TGA was more common among the genes with low expression (Fig. 4). Only 3% of the 500 genes with the highest expression had TGA as a stop codon while 19% of the 500 genes with the lowest expression were terminated with TGA. A similar relationship between TGA stop codon frequency and gene expression level was observed when analyzing the *Spirostomum* transcriptome generated in this study and the *Stentor polymorphus* transcriptome previously generated (Onsbring et al. 2018; data not shown).

Discussion

Specimens of *S. semivirescens* have been recorded from oxygen-depleted, freshwater habi-

Please cite this article in press as: Hines HN, et al. Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical Records. Protist (2017), https://doi.org/10.1016/j.protis.2018.08.001

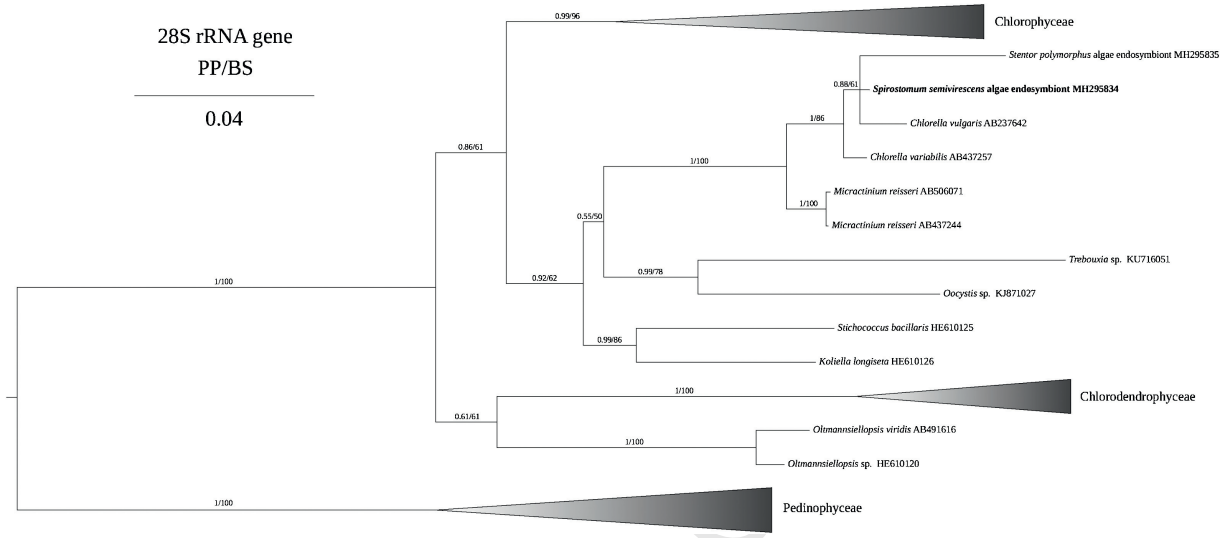

Figure 3. Phylogenetic relationships for green algae species of the class Trebouxiophyceae and other close relatives. The endosymbiont obtained from within the *Spirostomum semivirescens* sequences is in bold. Topology is inferred from the 28S rRNA gene with PhyloBayes using the CAT + GTR model, bootstrap values from the maximum likelihood analysis is mapped on the tree as well. A closely related, but not identical, 28S rRNA gene was found in the algal endosymbiont of *Stentor polymorphus* sampled from the same pond as *S. semivirescens* (Stadsskogen, Sweden).

tats in the UK before (Esteban et al. 2009a); the isolates used in this study represents a further habitat in the UK, and new records for Sweden from two Sweden separated by ~30 km. All strains were observed to be morphologically identical. The molecular analysis revealed identical molecular sequence between strains at the highly variable 18S rRNA level, confirming the match between the two sample groups of this large >1 mm ciliate. By investigating this species at a wider global resolution, the distribution of these micro-organisms has been expanded; at a distance of over 1,600 km between the sampling sites investigated in the UK and Sweden, the discovery of a strain with matching molecular sequences supports previous findings that microbial species thrive wherever the right conditions for their growth are found globally (Finlay 2002; Hines et al. 2016). This has wider implications for global microbial dispersal, particularly ciliate biogeography and biodiversity, with this species being a good target to investigate from other world regions (i.e. the tropics) for comparisons at a global level.

S. semivirescens has thus far been recorded from Germany (Kreutz and Foissner 2006), UK (Bradley et al. 2010; Esteban et al. 2009a, and this study), and now Sweden. The *Spirostomum minus viride* investigated by Foissner and Gschwind (1998) in Germany fits the morphological features of *S. semivirescens* and both are probably conspecific. Records from Russia (Boscaro et al. 2014) and Japan have also been reported (Esteban et al. 2009a), which comes to demonstrate that species' biogeography expands as sampling efforts increase.

The phylogenetic relationships found in this study show that the *S. ambiguum* clade was placed differently compared to previously published phylogenies (Boscaro et al. 2014; Shazib et al. 2014). When using RAxML (Stamatakis 2014) instead of IQ-TREE to calculate the phylogeny the same topology as (Shazib et al. 2014) was achieved with 66 in bootstrap support for *S. ambiguum* together with *S. subtilis* branch as sister clade to both groups of *S. minus*. The bootstrap support for *S. subtilis* as sister clade to *S. ambiguum* was 56. Since the IQ-TREE package contains a wider selection of evolutionary models to choose from and is reported to often find topologies with higher likelihoods (Nguyen et al. 2015) than RAxML, the bootstrap values from IQ-TREE were maped on the bayesian tree (Fig. 2). *S. subtilis* was placed as the deepest branching taxa in the *Spirostomum* genus as seen before in Boscaro et al. (2014) but not in Shazib et al. (2014). *S. semivirescens* could be placed with high support in the Bayesian tree as a close relative to *S. minus*. This is consistent with the similar morphology of the nuclear apparatus, where *S. semivirescens and S. minus* share the moniliform macronucleus shape (Shazib et al.

Please cite this article in press as: Hines HN, et al. Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical Records. Protist (2017), https://doi.org/10.1016/j.protis.2018.08.001

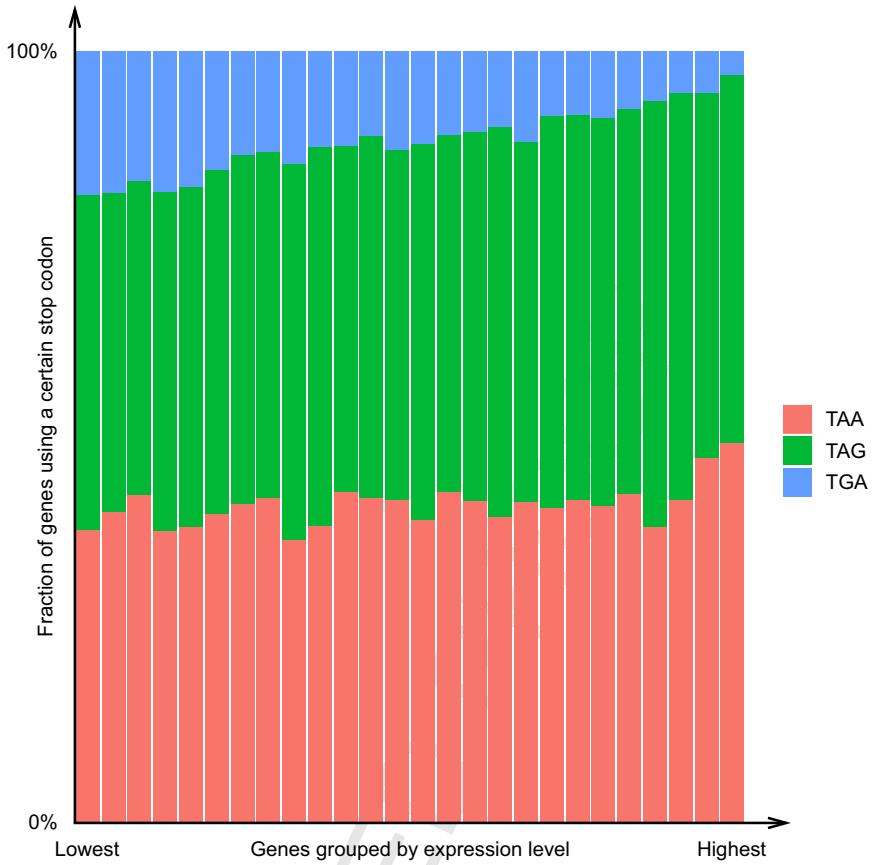

Figure 4. Relationship between stop codon frequency and gene expression levels. *Spirostomum semivirescens* transcripts ranked by expression level and then separated by groups of 500. Each bar displays the fraction of 500 genes using a specific stop codon, except for the lowest expressed group which had a total number of transcripts not evenly divided by 500. A trend is shown such that the fraction of transcripts terminating the translation with TGA is decreasing as the expression level is increasing.

2014). The closest relatives found for the endosymbiotic algae were *C. vulgaris* and *C. variabilis*, both reported as endosymbionts in other ciliate species (Zagata et al. 2016).

The TGA frequency was estimated to 11% of the stop codons in *S. semivirescens*, based on the genes used to investigate the relationship between expression level and stop codon frequency. In another heterotrich, *Stentor coeruleus*, the TGA frequency is 9%, based on the CDS file available from the online database StentorDB. Swart et al. (2016) report 5% and 1% TGA stop codon frequency for *Climacostomum virens* and *Fabrea salina* respectively. However these estimations for *C. virens* and *F. salina* were only based on 285 and 96 proteins respectively. Only 38 species out of 283 had a TGA stop codon frequency below 12% (Swart et al. 2016), several of these species could have already had their TGA reassigned since Swart et al. (2016) predicted less than 10 TGA stop codons for 11 of these cases. The relatively low TGA frequency among these heterotrichs indicates that TGA termination could have a higher fitness cost compared to the other stop codons. Therefore, there could potentially be a higher fitness gain in replacing the TGA codon for genes with a high expression level compared to genes with a lower expression level. Such selection pressure could cause the observed bias with fewer TGA in highly expressed genes as in *S. semivirescens*. Since it has been suggested that codon frequency is reduced prior to reassignment of codons (Mühlhausen et al. 2016), this leads to the question if *S. semivirescens* could be in an early stage of codon reassignment? Close relatives such as *Blepharisma* has already reassigned the TGA stop codon (Lozupone et al. 2001), *Condylostoma magnum* can use all three stop codons, uncluding TGA, as both stop and sense codon (Swart et al. 2016) and *Climacostomum virens* has been suggested to be in a transitory state of stop

Please cite this article in press as: Hines HN, et al. Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical Records. Protist (2017), https://doi.org/10.1016/j.protis.2018.08.001

codon reassignment (Heaphy et al. 2016). Given these observations in other heterotrichs, the connection between stop codon reassignment and stop codon bias with expression level could be worth further investigations. Gene expression levels for different stop codons have been investigated before in model organisms, but have turned out to have no relationship (Sun et al. 2005). Gene expression levels for different codons have mainly been investigated for sense codons before and have been observed in eukaryotes, e.g. *Saccharomyces pombe*, which stop codon frequency also correlate with expression level (Hiraoka et al. 2009). However, in *S. semivirescens* for most sense codons the frequency is changed slightly for the top 1000 highest or lowest expressed genes and for some sense codons the frequency is not affected by expression level at all (Supplementary Material Figs. S2–S5). Interestingly, the TAA frequency, which seems to be affected by mutational biases (Swart et al. 2016), is rather constant for *S. semivirescens* when comparing different expression levels (Fig. 4). In *S. semivirescens* the decrease in TGA frequency with higher expression levels is leading to an increase in TAG frequency, a change that requires the change of two nucleotides, instead of one.

We suggest that *S. semivirescens* uses rhodoquinol-dependent fumarate reduction to respire under anaerobic conditions. This is based on the high sequence identity to RquA found in *Rhodospirillum rubrum*, and the presence of the expected motif and mitochondrial targeting tag. A potential RquA sequence could also be found as well in the *Spirostomum* sp. data generated in this study. In both cases the sequence identity to *R. rubrum* RquA was above 40% and query coverage above 85%, the RquA motif was found and probability of export to mitochondria were over 70%. Since a putative rquA gene was found in the two *Spirostomum* species and this gene has also been reported in several other ciliates from the class Heterotrichea (Stairs et al. 2018), the whole *Spirostomum* genus might use this pathway for anaerobic respiration. The heterotrichs formed a monophyletic group in a phylogenetic analysis of the RquA protein from both prokaryotic and eukaryotic species. The relationship between the heterotrichs in the RquA phylogeny mirrored the topology of a phylogenetic analysis for their respective 18S rRNA genes (Supplementary Material Fig. S6). Additionally, Stairs et al. (2018) located a potential rquA sequence in the *Stentor coeruleus* genome, generated by Slabodnick et al. (2017), which gives further support to that heterotrichs code for rquA within their genome. Therefore we suggest that the identified rquA genes in this study are highly unlikely to be a contamination.

Conclusion

As more data are generated at the genomic level for different species in the *Spirostomum* genus, the relationship between major clades can be resolved. With the rRNA data that is currently available, *S. semivirescens* can be assigned as closest relative to *S. minus* and the endosymbiotic algae was identified as a member of the *Chlorella* genus. Insights into the transcriptome suggest that *S. semivirescens* use rhodoquinol-dependent fumarate reduction for respiration under anoxic conditions, which is likely also used by the other members of the genus since it has been observed in other species from the class Heterotrichea (Stairs et al. 2018) that also thrive in anoxic habitats similar to those where *S. semivirescens* is found. Our observations indicate that *S. semivirescens* could be in an early stage of codon reassignment. Therefore *S. semivirescens* could potentially be a relevant species to study for a better understanding of the evolution of the genetic code. Our results also indicate that it is possible for ciliates with identical morphologies, but from distant geographical areas, to have also identical molecular signatures.

Methods

Study sites: *UK study site.* Ciliates were sampled during June 2015 in Dorset, South England from a fen pond (50°40′44.5″N 2°11′29.5″W) and from a freshwater ditch, both located on the flood plain of the River Frome (50°40′52.0″N 2°11′21.8″W). *Spirostomum semivirescens* had previously been shown to thrive within this area (Esteban et al. 2009a) and this site is known to be a hotspot of ciliate biodiversity, with sampling efforts often revealing the *S. semivirescens* species. The fen habitat is densely wooded and dimly-lit with temporary ponds rich in organic sediment. The ditch had similar parameters, and was about 100 meters away from the fen. Oxygen levels were very low (<5%). The sediment water interface was sampled using a corked 500 mL caged sample bottle on a line. The corked line was pulled once the apparatus had sunk, to allow water and sediment within the desired oxygen-depleted depths to be collected. The area sampled in the fen pond and the ditch had a depth of less than 30 cm. 1 mL subsamples were observed in a Sedgewick Rafter chamber. Many cells were encountered and examined, with densities of up to 15 cells per mL of sediment subsample.

S. semivirescens cells collected from this location were hand-picked under a dissecting microscope using a micropipette, and were stored in RNAlater (Thermo Fisher Scientific) for transport to Uppsala University, Sweden for transcriptome analyses. cDNA synthesis (see below) was

Please cite this article in press as: Hines HN, et al. Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical Records. Protist (2017), https://doi.org/10.1016/j.protis.2018.08.001

performed within three days of removal from the UK sampling site and storage.

Sweden study sites. Samples were collected from two freshwater locations (see below) during August 2015. Air temperature of 25 °C in full sun, with water temperature of 18 °C recorded during a normal productive sample day. The first location investigated was Stadsskogen "city forest" – an ancient, densely-wooded and dimly-lit forest area (59°50′19.0″N 17°37′21.4″E). Within this habitat, a small pond location was chosen, and a pH of 6.0 was recorded, with a conductivity of 47 μS/cm. Samples collected ranged from a ~30 cm depth to shallow ~4 cm samples obtained by hand along the shoreline and on submerged algal mats. The second location sampled was a shallow eutrophic farmland pond with dense organic sediment, at "Oxhagen" in full sunlight with some aquatic plant coverage (60°5′47.6″N 17°46′4.7″E). Within this location, a pH of 6.6 was recorded, with a conductivity of 292 μS/cm. Samples were taken from 30 cm deep zones from various areas along the middle and edge of the pond.

Sampling methods were identical to the technique used in the UK (see above). Samples were taken back to the laboratory at Uppsala University, with subsamples being analyzed on a 1 mL Sedgewick Rafter chamber. 1 liter of water was taken from the sampling locations for laboratory analysis of the pH using a handheld PW9420 pH meter (Philips). To determine conductivity a Crison conductimeter 522 was also utilized on the removed samples within 2 hours of collection.

Samples were examined within 3 hours of removal, and the ciliates were found to thrive naturally for at least one week in the 500 mL bottles. Both sites were extremely productive for ciliates, with many harboring endosymbiotic algae such as *Stentor polymorphus*, *Frontonia* sp. and *Loxodes rostrum*. Anaerobic ciliates of the genus *Plagiopyla*, *Metopus* and *Caenomorpha* were present, as the sediment layer was largely oxygen depleted.

cDNA generation and sequencing: Both preserved and fresh ciliates were washed twice in double distilled water before single cells were picked in a 0.4 μL volume into a 0.2 mL PCR tube (VWR). cDNA synthesis was done according to the Smart-seq2 protocol (Picelli et al. 2014). Aliquots were diluted to 0.2 ng/μL based on dsDNA concentration measured with a Quant-iT PicoGreen dsDNA Assay Kit (Invitrogen). The diluted aliquots were prepared for sequencing using Nextera XT DNA Library Preparation Kit (Illumina). Two *S. semivirescens* replicates were sampled from each sampling site, i.e. the fen in Dorset, the pond in Stadsskogen and the pond in Oxhagen. The *Spirostomum* species lacking algae were sampled in Stadsskogen. For all six *S. semivirescens* replicates sequencing was done on Illumina MiSeq, 300 base pair, pair-end reads using v3 chemistry. The unidentified *Spirostomum* species were sequenced on Illumina HiSeq, 250 base pair, pair-end reads.

Transcriptome assembly: Raw reads were trimmed with Trimmomatic v0.35 (Bolger et al. 2014) by first removing primer sequences and DNA library preparation related sequences with the settings ILLUMINACLIP:2:30:10. Also in the following order LEADING:5, TRAILING:5 SLIDINGWINDOW:5:16 and MINLEN:80 were applied. Artificial reads were identified and removed using BLAST v2.2.30+ (Altschul et al. 1990) by a blastn search against the NCBI UniVec database. Transcriptome assembly was both carried out with Trinity v2.2.0 (Grabherr et al. 2011) and SPAdes v3.9.0 (Bankevich et al. 2012). The SPAdes assembly was done with a k-mer size of 99 and only used for the phylogenetic analysis since the rRNA contigs assembled by SPAdes were larger than in the Trinity assembly. In all other analysis, the Trinity assemblies were used. Full transcriptome alignment to NCBI nr database was done with DIAMOND v0.8.37 on sensitive blastx mode (Buchfink et al. 2015). The alignment results were analyzed with MEGAN v5.8.3 (Huson et al. 2007), which contig assignments were used to estimate the fraction of the data originating from the host, algae or prokaryotes.

Identification of anaerobic respiration pathway: Anaerobic respiration proteins previously found in other eukaryotes (Müller et al. 2012; Stairs et al. 2015) were searched for in the transcriptomes via tblastn search. To search for the presence of hydrogenosomes queries with [FeFe] hydrogenase, pyruvate:ferredoxin oxidoreductase and the maturase proteins HydE, HydF and HydG were used. Both pyruvate formate lyase and the enzyme to activate this protein were search for to detect pyruvate formate lyase activity. Also nitrate reductase, fumarase and RquA were used as queries to detect other anaerobic pathways.

Phylogenetic analysis: The rRNA sequences used in the phylogenies were identified with Barrnap (Seemann 2013). The ciliate sequences used to infer the phylogeny (Supplementary Material Table S1) were gathered by downloading all *Spirostomum* sequences available in the SILVA database and all sequences generated by Shazib et al. (2014). The algae sequences were gathered by using the identified 28S rRNA gene from *S. semivirescens* as a seed in a blastn search against NCBI nt database. CD-HIT V4.6.6 (Li and Godzik 2006) was used to remove identical sequences. Multiple sequence alignments were produced by MAFFT X-INS-i (Katoh 2002) where the CONTRAfold algorithm (Do et al. 2006) was used for pairwise structural alignment. The multiple sequence alignments were manually curated. BMGE was used to trim the curated alignment (Criscuolo and Gribaldo 2010). Bayesian inference tree topology was calculated with PhyloBayes v1.5a (Lartillot and Philippe 2004) using the CAT + GTR model. Four chains were used and both trees ran until maxdiff calculated by the PhyloBayes bpcomp-command were below 0.1. Burn-in was selected by monitoring – log likelihood plotted against generation of trees. For the ciliate Tree 13000 generations was generated and the burn-in was set to 1000. For the algae Tree 37000 generations was generated and the burn-in was set to 1000. Maximum likelihood trees were calculated with IQ-TREE (Nguyen et al. 2015) using the TIM + R2 model for the ciliate and TN + R3 model for the algae. The model tester in the IQ-TREE package selected the models in the maximum likelihood tree according to the Bayesian Information Criterion. Two long branches were removed in both the ciliate and the algae phylogeny that could potentially produce artifacts in the tree topology. To rule out that that the identified rquA sequences from the tblastn search were contamination we repeated the phylogenetic analysis by Stairs et al. (2018). Additional sequences added in this phylogeny were the potential rquA sequences identified in this study and a potential rquA sequence from the transcriptome of *Stentor polymorphus* (Onsbring et al. 2018). Multiple sequence alignment done by MAFFT L-INS-i (Katoh 2002) was trimmed with trimal (Capella-Gutiérrez et al. 2009) and the tree topology was calculated with IQ-TREE (Nguyen et al. 2015), using ultrafast bootstrap approximation with the LG + C50 model that was selected by the Bayesian Information Criterion.

Stop codon usage analysis: To analyze codon usage all six *S. semivirescens* replicates were assembled with Trinity v2.2.0 to a single assembly. To this assembly raw reads were mapped using Bowtie 2 (Langmead and Salzberg 2012) with the settings "–end-to-end –k 20 –D 20 –R 3 –N 1 –L 20 –i S,1,0.50 –X 1000". Because of the redundancy, often caused when assembling transcriptomes de novo (Duan et al. 2012), the contigs were clustered to transcripts using Corset v1.06 (Davidson Q3

Please cite this article in press as: Hines HN, et al. Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical Records. Protist (2017), https://doi.org/10.1016/j.protis.2018.08.001

and Oshlack 2014). The longest open reading frame from the longest contig in each cluster where then extracted. A blastp search against NCBI nr database using DIAMOND v0.8.37 (Buchfink et al. 2015) was then used to select all contigs with hits to *Stentor coeruleus*, *Paramecium tetraurelia*, *Oxytricha trifallax*, *Stylonychia lemnae*, *Tetrahymena thermophila*, *Pseudocohnilembus persalinus* and *Ichthyophthirius multifiliis* to discard contamination for downstream analysis. The species used to select contigs for further analysis represented the seven ciliates with the most blast hits. Selecting contigs based on more species would not have changed the outcome of the analysis since potential additional species had few hits and would in most cases have a hit as well to any of the seven mentioned species. The count matrix calculated with Corset during the clustering step was then used to rank the extracted open reading frames based on their expression level. To take different sequencing depth for each library into consideration, the total number of mapped reads for each species were used to normalize the number of mapped reads to each transcript. These values were then added together for each transcript to rank all the transcripts based on their normalized sum of mapped reads. The statistics for stop codon usage and the relationship to expression level was finally collected based on the transcripts selected by the blast search and the ranking of the normalized sum of mapped reads. Additionally, the transcripts were ranked based on expression level for each of the individual replicates, for comparison with the average, to assess the feasibility of averaging out the noise and the consistency between replicates (Supplementary Material Fig. S1). The redundancy-reduced co-assembly used in the codon analysis has been deposited in GenBank under the accession GGNT00000000. An assembly for the unidentified *Spirostomum* species was generated in the same way and deposited in GenBank under the accession GGNU00000000. The first versions for both transcriptome assemblies are described in this paper. The accession number for the raw reads reported in this paper is SRA: SRP145156.

Conflict of interest

None.

Uncited reference

Sun et al. (2016).

Acknowledgements

The *Spirostomum* transcriptomes were sequenced at the National Genomics Infrastructure sequencing platforms at the Science for Life Laboratory at Uppsala University (Sweden), a national infrastructure supported by the Swedish Research Council (VR-RFI) and the Knut and Alice Wallenberg Foundation. This work was supported by grants of the European Research Council (ERC Starting grant 310039-PUZZLE_CELL), the Swedish Foundation for Strategic Research (SSF-FFL5) and the Swedish Research Council (VR grant 2015-04959) to TJGE; GFE received financial support from the Alice Ellen Cooper Dean Trust (UK).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.protis.2018.08.001.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215**:403–410

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA (2012) SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comp Biol **19**:455–477

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

Boscaro V, Carducci D, Barbieri G, Senra MVX, Andreoli I, Erra F, Petroni G, Verni F, Fokin SI (2014) Focusing on genera to improve species identification: Revised systematics of the ciliate *Spirostomum*. Protist **165**:527–541

Bradley MW, Esteban GF, Finlay BJ (2010) Ciliates in chalkstream habitats congregate in biodiversity hot spots. Res Microbiol **161**:619–625

Buchfink B, Xie C, Huson DH (2015) Fast and sensitive protein alignment using DIAMOND. Nat Method **12**:59–60

Capella-Gutiérrez S, Silla-Martínez M, Gabaldón T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics **25**:1972–1973

Claros MG, Vincens P (1996) Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem **241**:779–786

Criscuolo A, Gribaldo S (2010) BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol **10**:210

Davidson NM, Oshlack A (2014) Corset: enabling differential gene expression analysis for de novo assembled transcriptomes. Genome Biol **15**:410

Do CB, Woods DA, Batzoglou S (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics **22**:e90–e98

Duan J, Xia C, Zhao G, Jia J, Kong X (2012) Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data. BMC Genom **13**:392

Esteban GF, Bradley MW, Finlay BJ (2009a) A case-building *Spirostomum* (Ciliophora, Heterotrichida) with zoochlorellae. Eur J Protistol **45**:156–158

Esteban GF, Finlay BJ, Clarke KJ (2009b) Sequestered organelles sustain aerobic microbial life in anoxic environments. Environ Microbiol **11**:544–550

Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. Science **296**:1061–1063

Finlay BJ, Span ASW, Harman JMP (1983) Nitrate respiration in primitive eukaryotes. Nature **303**:333–336

Foissner W, Gschwind K (1998) Taxonomy of some freshwater ciliates (Protozoa: Ciliophora) from Germany. Ber nat-med Ver Salzburg **12**:S25–S76

Ford BJ (1986) Hibernatory behaviour of the heterorichidan protozoan *Spirostomum* Ehrenberg. Microscopy **35**:362–364

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol **29**:644–652

Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov PV (2016) Novel ciliate genetic code variants including the reassignment of all three stop codons to sense codons in *Condylostoma magnum*. Mol Biol Evol **33**:2885–2889

Hines HN, McCarthy PJ, Esteban GF (2016) The first record for the Americas of *Loxodes rex*, a flagship ciliate with an alleged restricted biogeography. Microb Ecol **71**:5–8

Hiraoka Y, Kawamata K, Haraguchi T, Chikashige Y (2009) Codon usage bias is correlated with gene expression levels in the fission yeast *Schizosaccharomyces pombe*. Genes to Cells **14**:499–509

Hoshina R, Imamura N (2008) Multiple origins of the symbioses in *Paramecium bursaria*. Protist **159**:53–63

Hoshina R, Kusuoka Y (2016) DNA analysis of algal endosymbionts of ciliates reveals the state of algal integration and the surprising specificity of the symbiosis. Protist **167**:174–184

Huson DH, Auch AF, Qi J, Schuster SC (2007) MEGAN analysis of metagenomic data. Genome Res **17**:377–386

Katoh K (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res **30**:3059–3066

Knight RD, Freeland SJ, Landweber LF (2001) Rewiring the keyboard: evolvability of the genetic code. Nature Rev Genet **2**:49–58

Kreutz M, Foissner W (2006) The *Sphagnum* Ponds of Simmelreid in Germany: A Biodiversity Hot-Spot for Microscopic Organisms. Shaker Verlag Aachen, 267 p

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Method **9**:357–359

Lartillot N, Philippe H (2004) A bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol **21**:1095–1109

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**:1658–1659

Lonjers ZT, Dickson EL, Chu TPT, Kreutz JE, Neacsu FA, Anders KR, Shepherd JN (2012) Identification of a new gene required for the biosynthesis of rhodoquinone in *Rhodospirillum rubrum*. J Bacteriol **194**:965–971

Lozupone CA, Knight RD, Landweber LF (2001) The molecular basis of nuclear genetic code change in ciliates. Curr Biol **11**:65–74

Mühlhausen S, Findeisen P, Plessmann U, Urlaub H, Kollmar M (2016) A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. Genome Res **26**:945–955

Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu RY, van der Giezen M, Tielens AGM, Martin WF (2012) Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. Microbiol Mol Biol Rev **76**:444–495

Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol **32**:268–274

Onsbring H, Jamy M, Ettema TJG (2018) RNA sequencing of *Stentor* cell fragments reveals transcriptional changes during cellular regeneration. Curr Biol **28**:1281–1288

Picelli S, Faridani OR, Björklund ÅK, Winberg G, Sagasser S, Sandberg R (2014) Full-length RNA-seq from single cells using Smart-seq2. Nat Protocols **9**:171

Seemann T (2013) barrnap 0.5: rapid ribosomal RNA prediction. http://www.vicbioinformatics.com/

Shazib SUA, Vd'ačný P, Kim JH, Jang SW, Shin MK (2014) Phylogenetic relationships of the ciliate class Heterotrichea (Protista, Ciliophora, Postciliodesmatophora) inferred from multiple molecular markers and multifaceted analysis strategy. Mol Phylogenet Evol **78**:118–135

Slabodnick M, Ruby G, Reiff B, Swart C, Gosai S, Prabakaran S, Witkowska E, Larue E, Fisher S, Freeman M Jr, Gunawardena J (2017) The macronuclear genome of *Stentor coeruleus* reveals tiny introns in a giant cell. Curr Biol **27**:569–575

Stairs CW, Leger MM, Roger AJ (2015) Diversity and origins of anaerobic metabolism in mitochondria and related organelles. Philos Trans R Soc B Biol Sci **370**:20140326

Stairs CW, Eme L, Muñoz-Gómez S, Cohen A, Dellaire G, Shepherd JN, Fawcett JP, Roger AJ (2018) Microbial eukaryotes have adapted to hypoxia by horizontal acquisitions of a gene involved in rhodoquinone biosynthesis. eLife **7**:e34292

Stairs CW, Eme L, Brown MW, Mutsaers C, Susko E, Dellaire G, Soanes DM, van der Giezen M, Roger AJ (2014) A SUF Fe-S cluster biogenesis system in the mitochondrion-related organelles of the anaerobic protist *Pygsuia*. Curr Biol **24**:1176–1186

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30**:1312–1313

Sun J, Chen M, Xu J, Luo J (2005) Relationships among stop codon usage bias, its context, isochores, and gene expression level in various eukaryotes. J Mol Evol **61**:437–444

Sun L, Fang L, Zhang Z, Chang X, Penny D, Zhong B (2016) Chloroplast phylogenomic inference of green algae relationships. Sci Rep **6**:20528

Please cite this article in press as: Hines HN, et al. Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical Records. Protist (2017), https://doi.org/10.1016/j.protis.2018.08.001

Swart C, Serra V, Petroni G, Nowacki M (2016) Genetic codes with no dedicated stop codon: context-dependent translation termination. Cell **166**:691–702

Yarlett N, Hann AC, Lloyd D, Williams A (1981) Hydrogenosomes in the rumen protozoon *Dasytricha ruminantium* Schuberg. Biochem J **200**:365–372

Zagata P, Greczek-Stachura M, Tarcz S, Rautian M (2016) The evolutionary relationships between endosymbiotic green algae of *Paramecium bursaria* syngens originating from different geographical locations. Folia Biol **64**: 47–54

Please cite this article in press as: Hines HN, et al. Molecular Investigation of the Ciliate *Spirostomum semivirescens*, with First Transcriptome and New Geographical Records. Protist (2017), https://doi.org/10.1016/j.protis.2018.08.001