Received: 16 August 2022 | Revised: 14 November 2022 | Accepted: 30 November 2022

DOI: 10.1111/jeu.12959

ORIGINAL ARTICLE

18S rDNA sequence–structure phylogeny of the *Euglenophyceae* (Euglenozoa, Euglenida)

Antonia S. Rackevei[1] | **Anna Karnkowska**[2] | **Matthias Wolf**[1]

[1]Department of Bioinformatics, Biocenter, University of Würzburg, Würzburg, Germany

[2]Institute of Evolutionary Biology, Faculty of Biology, Biological and Chemical Research Centre, University of Warsaw, Warsaw, Poland

Correspondence
Matthias Wolf, Department of Bioinformatics, Biocenter, University of Würzburg, Am Hubland, 97074 Würzburg, Germany.
Email: matthias.wolf@biozentrum.uni-wuerzburg.de

Funding information
European Molecular Biology Organization, Grant/Award Number: 4150; Ministry of Education and Science, Poland

Abstract

The phylogeny of *Euglenophyceae* (*Euglenozoa*, *Euglenida*) has been discussed for decades with new genera being described in the last few years. In this study, we reconstruct a phylogeny using 18S rDNA sequence and structural data simultaneously. Using homology modeling, individual secondary structures were predicted. Sequence–structure data are encoded and automatically aligned. Here, we present a sequence–structure neighbor-joining tree of more than 300 taxa classified as *Euglenophyceae*. Profile neighbor-joining was used to resolve the basal branching pattern. Neighbor-joining, maximum parsimony, and maximum likelihood analyses were performed using sequence–structure information for manually chosen subsets. All analyses supported the monophyly of *Eutreptiella*, *Discoplastis*, *Lepocinclis*, *Strombomonas*, *Cryptoglena*, *Monomorphina*, *Euglenaria*, and *Colacium*. Well-supported topologies were generally consistent with previous studies using a combined dataset of genetic markers. Our study supports the simultaneous use of sequence and structural data to reconstruct more accurate and robust trees. The average bootstrap value is significantly higher than the average bootstrap value obtained from sequence-only analyses, which is promising for resolving relationships between more closely related taxa.

KEYWORDS
euglena, euglenids, phylogenetics, secondary structure

INTRODUCTION

ORGANISMS classified within *Euglenophyceae* Schoenich 1925 (*Euglenozoa*, *Euglenida*) are unicellular marine or freshwater algae. Most of the genera were described before the era of molecular phylogenetics and taxonomic changes have been made since the group was first revised using molecular data by Marin et al. (2003). *Euglenophyceae* are represented by the three orders *Rapazida*, *Eutreptiales*, and *Euglenales* (Kostygov et al., 2021). The genus *Rapaza*, representing the order *Rapazida*, is monotypic and sister to all remaining *Euglenophyceae* (Yamaguchi et al., 2012). *Eutreptiales* branch off prior to *Euglenales* and encompasses two genera, *Eutreptia* and *Eutreptiella* (Kostygov et al., 2021). A multigene phylogenetic analysis divided the *Euglenales* into the monophyletic families *Euglenaceae* and *Phacaceae* (Kim et al., 2010). Subsequent studies have supported the monophyly of *Euglenaceae* and *Phacaceae* (Karnkowska et al., 2015; Kim et al., 2015).

Currently, the family *Euglenaceae* includes eight genera supported by molecular data (*Colacium*, *Cryptoglena*, *Euglena*, *Euglenaformis*, *Euglenaria*, *Monomorphina*, *Strombomonas*, and *Trachelomonas*) (Kostygov et al., 2021). Several other genera were described, but due to a lack of molecular data, *Ascoglena*, *Euglenamorpha*, *Euglenopsis*, *Hegneria*, and *Klebsina* are not included in molecular studies. The family *Phacaceae* consists of four genera *Discoplastis*, *Lepocinclis*, and *Phacus* (Kim et al., 2010; Kostygov et al., 2021), and the more

recently added genus *Flexiglena* (Łukomska-Kowalczyk et al., 2021). These genera are monophyletic on phylogenetic trees; however, the genus *Euglena* tends to be polyphyletic since the first molecular phylogeny of the group in 2003 (Marin et al., 2003). Since then, several new genera, such as *Discoplastis* Triemer et al. (2006), *Euglenaria* Linton et al. (2010), *Euglenaformis* Bennett et al. (2014), and *Flexiglena* Łukomska-Kowalczyk et al. (2021), have been erected and representatives of *Euglena* were transferred to them based on the molecular analyses. However, the genus *Euglena* remains paraphyletic or polyphyletic (Karnkowska et al., 2015; Kim et al., 2015) due to several species branching out off the main *Euglena* clade.

The phylogeny of *Euglenophyceae* has been thoroughly studied throughout the years using the 18S ribosomal RNA (rRNA) gene, often combined with other molecular and/or morphological data (Bennett & Triemer, 2012; Karnkowska et al., 2015; Karnkowska-Ishikawa et al., 2012; Kim et al., 2010, 2015; Linton et al., 2010; Lukešová et al., 2020; Marin et al., 2003; Milanowski et al., 2006; Nudelman et al., 2003; Triemer et al., 2006; Wang et al., 2021). In this study, we use 18S ribosomal DNA (rDNA) sequence and structural information simultaneously for inferring phylogenetic relationships. This approach was shown to increase the accuracy and robustness of inferred phylogenetic trees and was recently reviewed by Keller et al. (2010) and Wolf et al. (2014).

Keller et al. (2010) used ITS2 rRNA gene sequences and 600,000 different alignments testing the methodical influence concerning different tree topologies, branch lengths, and ancestral sequences in a complex simulation that integrated the coevolution process of sequences and their individual secondary structures. The simulation showed a negative correlation of accuracy and robustness in neighbor-joining tree reconstruction with an increase in the number of taxa. Keller et al. (2010) showed that including individual secondary structure information broadens the range of the optimal marker performance and a higher level of divergence results in better performances. Whereas marker elongation increases robustness, that is, bootstrap support, the inclusion of individual secondary structures additionally improves accuracy. Keller et al. (2010) suggested that this approach can be applied to other ribosomal genes like 18S rDNA and to other tree reconstruction methods like parsimony and likelihood, thereby combining variable sequences with conserved secondary structures is the most beneficial and promising approach for phylogenetic analyses.

As mentioned above, the phylogeny of the *Euglenophyceae* is quite well studied, but it was shown that multigene phylogenies are usually needed to resolve the tree. To come as close as possible, with our approach, we want to get more out of the 18S rDNA data, in particular, because still most of the phylogenetic data available come from 18S rDNA; in other words, for many species multiple genes used in concatenated analyses are still not available.

MATERIALS AND METHODS

Taxon sampling

For a flowchart of methods used in this study, see Figure 1. Using different search strings, all available 18S rDNA sequences from organisms classified within *Euglenophyceae* (*Euglenozoa*, *Euglenida*) ranging from >2000 to <3000 nucleotides, to ensure complete secondary structures, were obtained from the nucleotide database (GenBank) from the National Center of Biotechnology Information (NCBI) (retrieved on 04/21/2022) (Benson et al., 2013). Only strains that were identified by NCBI down to the species level were further processed in this study. Using default settings, 18S rDNA sequences were aligned using ClustalX 2.1 (Larkin et al., 2007). Introns were removed in ALIGN 07/04 (Hepperle, 2004).

Structure prediction

Using the "model" option as implemented in the ITS2 database (Ankenbrand et al., 2015), homology modeling (Selig et al., 2008; Wolf et al., 2005) was used to predict individual 18S rRNA secondary structures for all sequences used in this study. The 18S rRNA sequence–structure information of *Euglena gracilis* (GenBank Accession: M12677), recently published by Matzov et al. (2020), was used as template (Figure S3 and Data S1). Sequences with a structural homology of less than 70% were discarded.

Sequence–structure alignment

An automatic sequence–structure alignment, using a sequence–structure specific scoring matrix, was generated in ClustalW 1.83 (Larkin et al., 2007) as implemented in 4SALE 1.7.1 (Seibel et al., 2006, 2008), that is, 4SALE simultaneously and automatically aligns sequences and their individual secondary structures using a 12×12 scoring matrix (Wolf et al., 2014). The scoring matrix (Seibel et al., 2006) was derived after translating the four nucleotides and their three structural states (unpaired, paired right, and paired left) into a 12-letter alphabet (one-letter encoded sequence–structure data), as shown in Figure 2. Sequences that could not be properly aligned were removed from the dataset.

Sequence–structure tree reconstruction

With ProfDistS 0.9.9 (Friedrich et al., 2005; Wolf et al., 2008), using the alignment in .xfasta format (saved by 4SALE as alignment including secondary structures), an overall sequence–structure neighbor-joining (NJ) (Saitou & Nei, 1987) tree (Figure 3) was calculated using

a sequence–structure specific Jukes-Cantor (JC) correction (cf. Jukes & Cantor, 1969).

According to Müller et al. (2004) and Rahmann et al. (2006) in phylogenetics, we often do have access to additional information that allows the definition of subclades. We can keep the whole set of taxa but restrict the set of allowable tree topologies to those that

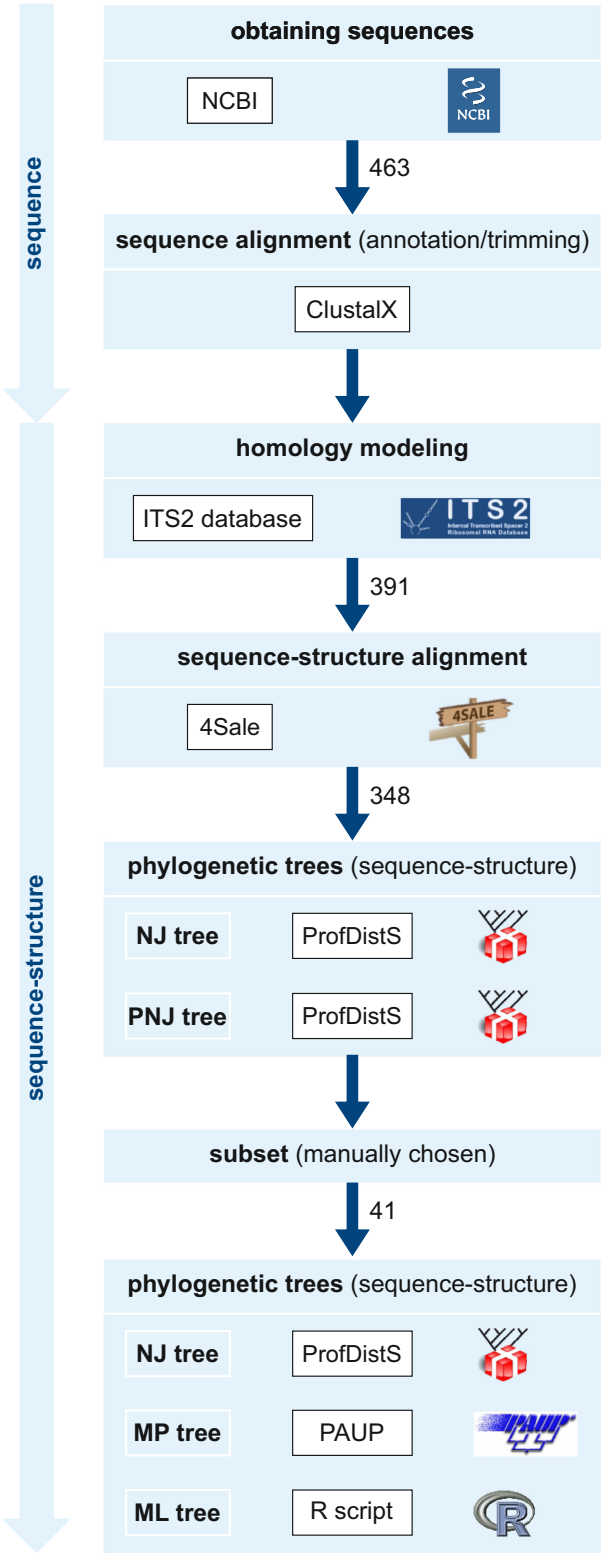

FIGURE 1 Flowchart of materials and methods. 18S rDNA sequences of *Euglenophyceae* were obtained from GenBank (Benson et al., 2013) and aligned using ClustalX (Larkin et al., 2007). Strains classified as "sp." and sequences that could not be properly aligned were removed from the dataset. Secondary structures were predicted using homology modeling (Selig et al., 2008; Wolf et al., 2005). Sequence and structural data were simultaneously aligned in 4SALE (Seibel et al., 2006, 2008). With ProfDistS (Friedrich et al., 2005; Wolf et al., 2008), a sequence–structure NJ tree and a sequence–structure PNJ tree were reconstructed. A subset was manually chosen for sequence–structure NJ, MP, and ML analyses. For comparison, sequence-only trees for the overall NJ analysis and the subset analysis as obtained by default settings using ProfDistS (Friedrich et al., 2005; Wolf et al., 2008) and RAxML (Stamatakis, 2014) can be found in the supplement (Figures S1 and S2).

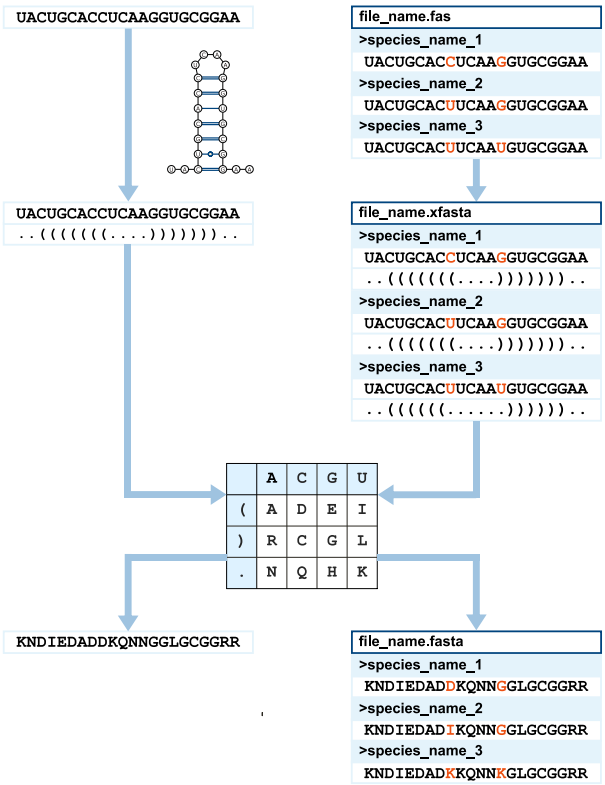

FIGURE 2 Translation of sequence–structure information into one-letter encoded files. An RNA sequence with its individual secondary structure in bracket dot-bracket notation and a 2D structure is shown. Using the 12-letter translation table, for sequence–structure alignments 4SALE (Seibel et al., 2006, 2008; Wolf et al., 2014) encodes the sequence–structure information into a pseudoprotein sequence. For tree reconstructions, ProfDistS (Friedrich et al., 2005; Wolf et al., 2008) uses and encodes .xfasta files, whereas PAUP* (Swofford, 2002) and R (R Core Team, 2018) directly use one-letter encoded .fasta files (pseudoprotein).

are consistent with the known monophyletic groups, or we solve many small problems, each with a reduced set of taxa, and assemble the resulting subtrees into a consistent supertree. Here, we replace the set of taxa forming a known subclade by a single supertaxon, which we represent by a sequence profile. To estimate the evolutionary distances between supertaxa, we generalize

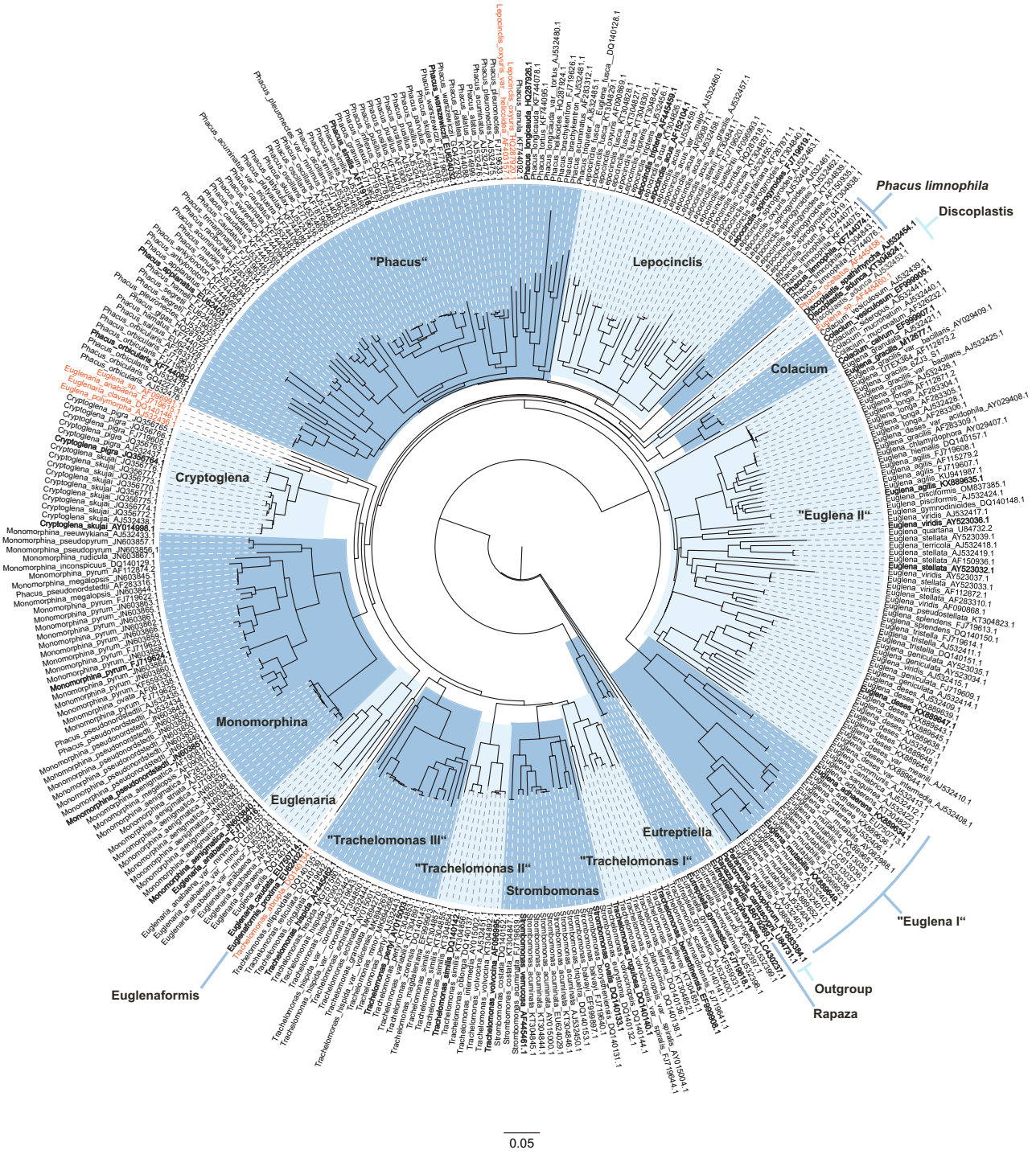

FIGURE 3 18S rDNA overall sequence–structure neighbor-joining (NJ) tree. The tree was based on 348 sequence–structure pairs and was reconstructed using ProfDistS (Friedrich et al., 2005; Wolf et al., 2008). 4SALE (Seibel et al., 2006, 2008) was used for the global multiple sequence–structure alignment. *Peranema trichophorum* and *Petalomonas cantuscygni* were used as outgroup taxa. Each taxon name is accompanied by the GenBank accession number. *Euglena viridis* (GenBank accession: AF445460) and *Euglena agilis* (GenBank accession: AF096991) are currently discussed as *Euglena* sp. according to the EukRef database. Clades are alternating in light and dark blue. Clades that are not monophyletic are indicated by quotation marks. Taxa marked in orange were not assigned to a specific clade. Sequence–structure pairs which were chosen for the subset are marked bold. The scale bar indicates evolutionary distances.

the maximum likelihood distance estimator of two sequences to evolutionary distances of profiles. The derived distance matrix is used to reconstruct a tree by NJ, referred to as profile neighbor-joining (PNJ).

Evaluations indicated that PNJ is both more robust and accurate than NJ (Müller et al., 2004). According to Müller et al. (2004), we might simply use one representative taxon from each subclade or estimate the most likely

sequence in the subclade root, but the profile-based approach appears preferable because it integrates information from all sequences. A sequence profile is a stochastic model of a sequence family. A profile is also a sequence, but it is composed of probability distribution vectors instead of characters. Because we are more interested in the "center of gravity" of the sequences in a known subclade, we simply take the position-specific relative nucleotide frequencies over all sequences within the subclade. This results in a robust estimate that is independent of estimated subclade topologies. With ProfDistS, a sequence–structure PNJ (Müller et al., 2004) tree was reconstructed in five iterations (Figure 4A), whereby sequences that could not be unambiguously assigned to a monophyletic subclade (cf. Figure 3) were not included in predefined profiles. However, additionally, a subset from the overall sequence–structure NJ tree was manually chosen with each genus represented proportionally.

For subset trees (Figure 5), using a sequence–structure specific JC correction, a sequence–structure NJ tree was reconstructed with ProfDistS (Figure S4). Using the one-letter encoded sequence–structure alignment in .fasta format (saved by 4SALE), and translated to NEXUS (Maddison et al., 1997) with ALIGN, a sequence–structure maximum parsimony (Camin & Sokal, 1965) (MP) tree was calculated with default settings in PAUP* 4.0a (Swofford, 2002) (Figure S5). With phangorn (Schliep, 2011) as implemented in R 4.2.1 (R Core Team, 2018), using the one-letter encoded sequence–structure alignment in .fasta format, a sequence–structure maximum likelihood (Felsenstein, 1981) (ML) tree was reconstructed with a GTR+I+G model as estimated from the data. The R-script is available at http://4sale.bioapps.biozentrum.uni-wuerzburg.de (Wolf et al., 2014). For the PNJ tree and for all trees of the manually chosen subset (NJ, MP, and ML), bootstrap support (Felsenstein, 1985) was estimated (due to the complexity of the 12×12 approach) based on "only" 100 pseudo-replicates. All trees were rooted with *Peranema trichophorum* (*Euglenozoa*, *Euglenida*, *Heteronematina*) and *Petalomonas cantuscygni* (*Euglenozoa*, *Euglenida*, and *Scytomonadidae*) as outgroup.

RESULTS

Taxon sampling

From NCBI, 572 sequences ranging in their sequence length from >2000 to <3000 nucleotides were obtained. In NCBI, 111 sequences were unclassified as "sp." and were discarded. Four hundred sixty-one 18S rDNA sequences from organisms classified within *Euglenophyceae* and two outgroup sequences were obtained (Tables S1 and S2) using different search strings and further aligned in ClustalX (Larkin et al., 2007). An intron with the length of 434 nucleotides was removed from the sequence

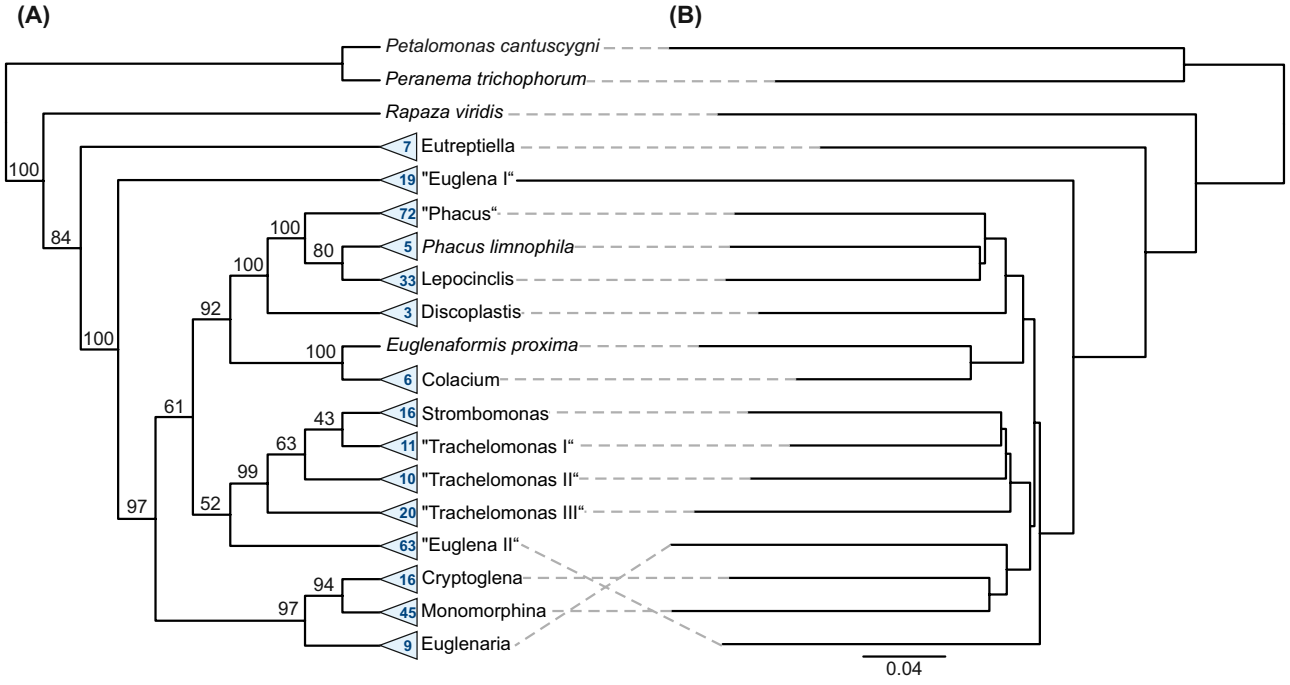

FIGURE 4 18S rDNA sequence–structure profile neighbor-joining (PNJ) tree based on 339 taxa. The tree was reconstructed using ProfDistS (Friedrich et al., 2005; Wolf et al., 2008) and was rooted with *Peranema trichophorum* and *Petalomonas cantuscygni*. (A) Five-times iterated 18S rDNA sequence–structure PNJ tree. Based on bootstrap values (>75), in each iteration, super-profiles of profiles have been built. Bootstrap values from 100 pseudo-replicates are mapped at internodes. Numbers in blue represent taxa included in a profile. (B) Original PNJ tree with branch lengths reconstructed with ProfDistS (Friedrich et al., 2005; Wolf et al., 2008). The scale bar indicates evolutionary distances.

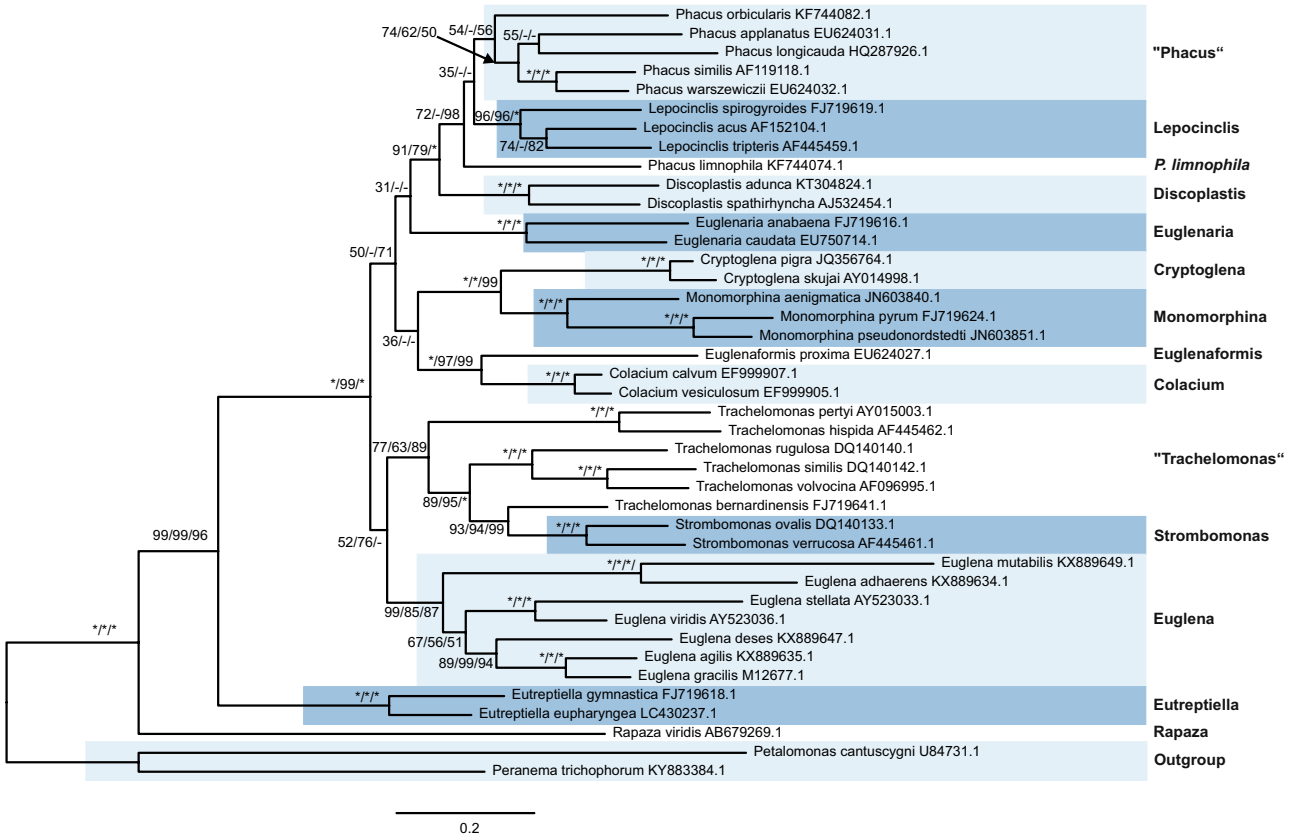

FIGURE 5 18S rDNA sequence–structure maximum likelihood (ML) tree. The tree was reconstructed with R (R Core Team, 2018). A subset of 41 sequence–structure pairs was manually chosen. Bootstrap values from 100 pseudo-replicates are from ML, MP, and NJ analyses. Bootstrap values of 100 are marked with "*." Different tree topologies are indicated with "-." The MP tree was reconstructed using PAUP* (Swofford, 2002). The NJ tree was reconstructed with ProfDistS (Friedrich et al., 2005; Wolf et al., 2008). *Peranema trichophorum* and *Petalomonas cantuscygni* were used as outgroup taxa. Monophyletic clades are highlighted alternating in light and dark blue and are additionally named alongside the tree. Clades that are not monophyletic are indicated by quotation marks. Each taxon name is accompanied by the GenBank accession number. The scale bar indicates evolutionary distances.

of *Rapaza viridis* (GenBank Accession: AB679269). Individual secondary structures of all 463 18S rDNA sequences were predicted using homology modeling (Selig et al., 2008; Wolf et al., 2005). Seventy-two sequences with structural homology of less than 70% were discarded.

Overall sequence–structure neighbor-joining tree

In 4SALE, a sequence–structure alignment of 391 sequence–structure pairs was generated. Four sequences could not be properly aligned but were represented by other taxa of the same species and were therefore discarded. The aligned sequences were trimmed uniformly at the ends (Figure S3) and 39 sequences that were too short after aligning were removed from the dataset. The final alignment consists of 348 taxa with a length of 6151 characters (Data S2). Based on 348 sequence–structure pairs, an overall sequence–structure NJ tree (Figure 3, Data S3) was reconstructed. Most genera were recovered as monophyletic. *Euglena* was paraphyletic with most members grouped in one clade (*Euglena* II). Organisms classified as *E. mutabilis*, *E. adhaerens*, and *E. carterae* formed a separate clade (*Euglena* I). Three members of *Euglena* (*E.* sp, [AF096991], *E.* sp. [AF445460] and *E. polymorpha* [AJ532436]) could not be assigned to a clade. Most members of *Trachelomonas* were positioned in three groups and the *Strombomonas* clade was positioned within the *Trachelomonas* clade. *T. abrupta* formed a separate lineage. Most members of *Phacus* were grouped in a single clade, with two organisms classified as *Lepocinclis* positioned within the clade. Organisms classified as *P. limnophila* formed a separate clade and *P. ocellatus* formed a separate lineage. Two organisms classified as *Euglenaria* could not be assigned to a clade.

Sequence–structure profile neighbor-joining tree

Nine taxa could not be unambiguously assigned to a subclade and were not included in the predefined profiles for PNJ analysis. The final alignment consists of 339 sequence–structure pairs with a length of 6016 characters (Data S4). Profiles were defined from subclades that could be inferred from the overall NJ tree (Figure 3).

In the iterated (five times) PNJ tree (Figure 4A and Data S5), *Rapaza viridis* was positioned at the base of the tree. The genus *Eutreptiella* was represented by seven taxa and formed a sister clade to *Euglenales*. *Euglenales* formed a well-supported (100 = bootstrap support) monophyletic clade. The genus *Euglena* was recovered to be polyphyletic. The *Euglena* I clade, including *E. mutabilis*, *E. adhaerens*, and *E. carterae*, was located at the base of *Euglenales*. The clade including the genera *Cryptoglena*, *Monomorphina*, and *Euglenaria* was well supported (97) and diverged after the *Euglena* I clade. *Euglenaria* was represented by nine taxa and was positioned at the base of this clade. The genus *Cryptoglena* was represented by 16 taxa and formed a sister clade with high support (94) to the genus *Monomorphina*, which was represented by 45 taxa.

Other genera of *Euglenales* split into two clades. One clade with low support (52) included the *Euglena* II clade and the genera *Strombomonas* and *Trachelomonas*. The *Euglena* II clade included 63 taxa and was positioned at the base of this clade. The *Trachelomonas*/*Strombomonas* clade was well-supported (99). *Trachelomonas* was found to be paraphyletic with three clades. The genus *Strombomonas* was represented by 16 taxa and formed a sister clade to one of the *Trachelomonas* clades with no support (43). The other clade included *Colacium*, *Euglenaformis*, and the family *Phacaceae* (*Phacus*, *Lepocinclis*, and *Discoplastis*) and was well-supported (92). *Colacium* was represented by six taxa and formed a fully supported (100) sister clade to a single taxon representing *Euglenaformis*. The *Euglenaformis*/*Colacium* clade was sister to *Phacaceae*.

Discoplastis was represented by three taxa and was positioned at the base of *Phacaceae*. The *Phacus*/*Lepocinclis* clade was fully supported (100). *Phacus* was paraphyletic with *P. limnophila* forming a separate clade. The genus *Lepocinclis* was represented by 33 taxa and was sister to the *P. limnophila* clade with moderate support (80). The PNJ tree showed some divergence to the overall NJ tree. In the overall NJ tree, both *Euglena* clades diverged at the base of *Euglenales* and *Euglenaformis* was positioned within the *Trachelomonas*/*Strombomonas* clade. Also, *P. limnophila* was sister to the main *Phacus*/*Lepocinclis* clade. The clade including *Euglenaria*, *Monomorphina*, and *Cryptoglena* was sister to *Phacaceae*. The genus *Colacium* was positioned at the base of the clade including all genera of the family *Phacaceae* and the genera *Cryptoglena*, Monomorphina, and *Euglenaria*.

The original PNJ tree (Figure 4B, Data S6) showed a similar topology as the iterated PNJ tree. However, in the original PNJ tree, both *Euglena* clades diverged at the base of *Euglenales*. The clade including *Colacium*, *Euglenaformis*, and the family *Phacaceae* diverged after the *Euglena* clades. Also, the *Monomorphina*/*Cryptoglena*/*Euglenaria* clade was sister to the *Trachelomonas*/*Strombomonas* clade.

Average bootstrap support

The average bootstrap value within subclades (data not shown) in our large-scale NJ analysis is much higher than the average bootstrap value in Kolisko et al. (2020), the largest currently available sequence-only analysis (based on ML) with a comparable number of taxa. This was exemplarily tested for the main *Phacus* clade (average bootstrap values are 89 vs. 65, respectively), and for the *Cryptoglena*/*Monomorphina* clade (91 vs. 63).

Sequence–structure subset trees

A subset of 41 taxa was manually chosen from the overall tree (Figure 3) to represent each genus proportionally. Whenever possible, the type species was chosen. The final alignment of the subset consists of 41 taxa with 3836 characters (Data S7). NJ, MP, and ML trees were reconstructed based on the sequence–structure alignment of the subset and bootstrap support was estimated based on 100 pseudo-replicates. The ML tree is shown in Figure 5 (Data S8–S10) with bootstrap values from ML, MP, and NJ analyses.

Most genera were found to be monophyletic, except for *Phacus* and *Trachelomonas*. *Rapaza viridis* was positioned at the base of the tree. *Euglenales* form a well-supported clade (100/99/100 = bootstrap support from ML/MP/NJ analyses) and are a sister group to the *Eutreptiella* clade. The *Euglenales* were split into two clades, one clade including all members of *Trachelomonas*, *Strombomonas*, and *Euglena* with low support (52/76/−). The *Strombomonas* clade was positioned within the *Trachelomonas* clade. The *Trachelomonas*/*Strombomonas* clade showed moderate bootstrap support (77/63/89), and the sister relationship of *Strombomonas* and *Trachelomonas bernardinensis* was well-supported (93/94/99). The *Trachelomonas*/*Strombomonas* clade formed a sister clade to the *Euglena* clade with low support (52/76/−). With high support (99/85/87), *Euglena* appeared as monophyletic. Other members of *Euglenales* formed a clade with low support (50/−/71). This clade was split into two clades. One of those clades included members of *Colacium*, *Monomorphina*, *Euglenaformis*, and *Cryptoglena* with no support (36/−/−). The sister relationship of *Euglenaformis* and *Colacium* was well-supported (100/97/99). *Monomorphina* formed a sister clade to *Cryptoglena* with high support (100/100/99).

The other clade included the genus *Euglenaria* and all members of *Phacaceae* (*Phacus*, *P. limnophila*, *Discoplastis*, and *Lepocinclis*) with *Euglenaria* as a sister clade to *Phacaceae*. *Phacaceae* formed a well-supported clade (91/79/100) with *Discoplastis* as the basal lineage. The *Phacus*/*Lepocinclis* clade showed moderate support (72/−/98). The genus *Phacus* was paraphyletic with *Lepocinclis* positioned within the

Phacus clade. All members of *Phacus* except for *P. limnophila* formed a single clade with low bootstrap support (54/−/56). This clade was sister to the *Lepocinclis* clade with no support (35/−/−) in maximum likelihood analysis.

DISCUSSION

Several studies have used a secondary structure to guide the alignment (Bennett & Triemer, 2012; Ciugulea et al., 2008; Karnkowska et al., 2015; Karnkowska-Ishikawa et al., 2012; Kim et al., 2010, 2015; Kim & Shin, 2008, 2014; Linton et al., 2010; Marin et al., 2003; Milanowski et al., 2006; Nudelman et al., 2003; Triemer et al., 2006). In this study, we infer the alignment based on sequence and structural information simultaneously. According to Keller et al. (2010), using sequence–structure information simultaneously improves the accuracy and robustness of reconstructed trees. This was shown only for NJ trees based on internal transcribed spacer 2 (ITS2) rRNA gene sequence–structure data. However, Keller et al. (2010) suggested that other ribosomal genes might benefit from the inclusion of individual secondary structures, whereby markers with a conserved structure and a variable sequence benefit the most. Therefore, subsequent case studies applied the approach as suggested by Keller et al. (2010) for ITS2 and 18S sequence–structure information using NJ, MP, and ML (Borges et al., 2021; Buchheim et al., 2017; Czech & Wolf, 2020; Heeg & Wolf, 2015; Lim et al., 2016; Markert et al., 2012; Plieger & Wolf, 2022).

The scoring matrices and substitution models that have been used (cf. Seibel et al., 2006) assume that the RNA secondary structures are conserved. Which is supported by the simulation study from Keller et al. (2010). According to Wolf et al. (2014), on a plain RNA substitution model the replacement e.g. U to C is negatively scored, whereas a more complex RNA sequence–structure model yields large positive and negative scores for the 3×3 U–C mutations on the sequence–structure level. In particular, the mutation "U)" to "C)" gets a highly positive score. This relies on the fact that often these mutations will maintain the underlying RNA secondary structure (because of the possible GU pairs), while the structure-destroying mutations "U(" to "C)" and "C(" to "U)" are scored highly negative, and therefore, such mutations will only very rarely be aligned. However, it is apparent that the higher the impact of the used scoring matrix and the used substitution model on the tree inferences, the more important is the correctness of the underlying data (e.g. doubtful secondary structures and/or weak areas in the multiple sequence alignment). Therefore, the increased statistical and computational complexity must necessarily go hand in hand with the quality of the underlying data. Taxa with a doubtful secondary structure and taxa that could not be unambiguously aligned need to be removed. Homology-modeled structures are not perfect. In the future, such structures could be optimized and fine-tuned with constantly evolving structure prediction methods.

In this study, we applied the approach from Keller et al. (2010) for 18S rDNA sequence–structure data using PNJ, NJ, MP, and ML. The PNJ tree was well-supported with only one branch with a support <50. Some basal branches of the ML tree differed from the PNJ tree due to a lack of bootstrap support. In agreement with previous studies, the reconstructed sequence–structure trees support the monophyly of *Eutreptiella*, *Discoplastis*, *Lepocinclis*, *Strombomonas*, *Cryptoglena*, *Monomorphina*, *Euglenaria*, and *Colacium*. In addition, the position of the genera *Rapaza*, *Eutreptiella*, *Discoplastis*, *Lepocinclis*, *Cryptoglena*, and *Monomorphina* were generally consistent with previous studies based on multiple markers (Karnkowska et al., 2015; Kim et al., 2010, 2015; Linton et al., 2010). The individual groups are discussed below.

Rapaza/Eutreptiella

Rapaza viridis was positioned at the base of *Euglenophyceae*, which is consistent with previous single-gene studies using 18S rDNA (Cavalier-Smith, 2016; Kolisko et al., 2020; Lukešová et al., 2020; Yamaguchi et al., 2012). In previous studies using 18S rDNA (Cavalier-Smith, 2016; Kolisko et al., 2020; Lukešová et al., 2020; Marin et al., 2003; Yamaguchi et al., 2012) or multiple chloroplast genes (Dabbagh & Preisfeld, 2018), *Eutreptiella* branched off prior to *Euglenales*. This study supports this position.

Discoplastis/Lepocinclis/Phacus

In this study, *Phacaceae* formed a well-supported group, which is consistent with previous studies combining multiple markers (Karnkowska et al., 2015; Kim et al., 2010, 2015). Kim et al. (2010) established the family *Phacaceae* in a multigene study using a combined data set of nuclear and chloroplast genes. *Phacaceae* included the genera *Discoplastis*, *Phacus*, and *Lepocinclis* and was sister to the family *Euglenaceae*. Subsequent studies based on multiple genetic markers supported the monophyly and sister relationship of both groups (Karnkowska et al., 2015; Kim et al., 2015). In this study, *Phacaceae* was positioned within *Euglenaceae*. This position differs from multigene studies (Karnkowska et al., 2015; Kim et al., 2010, 2015) and was inconsistent in the PNJ and subset analyses. Most studies using only 18S rDNA as a marker could not recover either families as monophyletic (Cavalier-Smith, 2016; Kolisko et al., 2020; Marin et al., 2003; Wang et al., 2021). A recent single-gene study using 18S rDNA recovered both families but with no support (Lukešová et al., 2020).

The genus *Discoplastis* was found to be monophyletic with high support and was positioned at the base of *Phacaceae*, which is consistent with previous studies using a combined dataset (Karnkowska et al., 2015; Kim et al., 2015; Linton et al., 2010; Łukomska-Kowalczyk et al., 2021). In other single-gene studies using 18S rDNA, the position of *Discoplastis* is often discrepant (Cavalier-Smith, 2016; Kolisko et al., 2020; Marin et al., 2003; Milanowski et al., 2006; Wang et al., 2021). The sister relationship of *Phacus* and *Lepocinclis* was well-supported in previous studies using two or more markers (Bennett & Triemer, 2012; Kim et al., 2010, 2015; Linton et al., 2010; Łukomska-Kowalczyk et al., 2021; Triemer et al., 2006). Marin et al. (2003) emended the genus *Phacus*, making it monophyletic. The monophyly was supported by subsequent molecular studies (Kim et al., 2010; Kim & Shin, 2008; Linton et al., 2010; Milanowski et al., 2006; Triemer et al., 2006). Linton et al. (2010) added *Euglena limnophila* [*Phacus limnophila*] and *Lepocinclis salina* [*Phacus salina*] to the genus *Phacus*. In a study using five different genes, *P. limnophila* and *P. warszewiczii* were positioned basal to *Phacus* sensu Marin *et* Melkonian and *Lepocinclis* (Karnkowska et al., 2015). Kim and Shin (2014) reconstructed the phylogeny of *Phacus* using combined nuclear and plastid genes and showed that *P. limnophila* formed a clade separate from the main *Phacus* clade. Kim et al. (2015) used the same genetic markers but increased the taxon sampling and confirmed the monophyly of the genus *Phacus*. The monophyly was supported in Bayesian analysis but had low bootstrap support in ML analysis. In this study, *Phacus* was paraphyletic and *P. limnophila* formed a separate clade. In most single-gene studies using 18S rDNA, *P. limnophila* was also not included in the main *Phacus* clade (Cavalier-Smith, 2016; Kolisko et al., 2020; Marin et al., 2003; Milanowski et al., 2006; Wang et al., 2021). In the PNJ analysis, *P. limnophila* was sister to *Lepocinclis*. The ML tree and the overall NJ tree showed a similar topology as Karnkowska et al. (2015) and Kim and Shin (2014) with *P. limnophila* forming a separate lineage. The ML tree was reconstructed on a small dataset, whereas the overall NJ tree and the PNJ tree were reconstructed on a large dataset. This suggests that an increased taxon sampling alone is not enough to resolve the taxonomy of *Phacus*, and more molecular markers might be necessary.

Euglena

Recent studies found that *Euglena* is not monophyletic. *E. archaeoplastidiata* was not included in the main *Euglena* clade (Dabbagh & Preisfeld, 2018; Karnkowska et al., 2015; Kim et al., 2015; Kolisko et al., 2020) and several strains of *E. velata* were sister to *Colacium* (Kim et al., 2015). Due to a lack of appropriate molecular data in NCBI, *E. archaeoplastidiata* and *E. velata* were not included in this study and *Euglena* formed a well-supported monophyletic clade in the subset analysis. This is consistent with single-gene studies using 18S rDNA (Cavalier-Smith, 2016), chloroplast large subunit (cpLSU) (Kim & Shin, 2008), and multigene studies using a combined dataset (Bennett & Triemer, 2012; Linton et al., 2010; Triemer et al., 2006).

In the PNJ analysis and the overall NJ tree, *Euglena* was found to be polyphyletic. *E. mutabilis*, *E. adhaerens*, and *E. carterae* formed a separate clade that was positioned at the base of *Euglenales*. This position is in accordance with Marin et al. (2003), where *E. mutabilis* and *E. carterae* diverged at the base of *Euglenales* and *E. adhaerens* was not included in this study. According to Marin et al. (2003), long-branch attraction (LBA) may have caused this divergence. In a subsequent study, Marin et al. (2003) changed the taxon sampling as well as the outgroup and used a more complex model of evolution (GTR+I+G) for ML analysis. There, *E. mutabilis* and *E. carterae* were included in the main Euglena clade. In our study, in the subset analysis *Euglena* was monophyletic with *E. mutabilis* and *E. adhaerens* positioned at the base of the *Euglena* clade. LBA could be the cause for the divergence in the PNJ tree and overall NJ tree. In most studies that combine two or more genetic markers, *Euglena* diverges after *Euglenaformis* at the base of *Euglenales* (Bennett & Triemer, 2012; Dabbagh & Preisfeld, 2018; Karnkowska et al., 2015; Linton et al., 2010; Milanowski et al., 2006). In this study, *Euglena* was sister to the *Trachelomonas*/*Strombomonas* clade, which differs from previous studies using multiple markers. However, the position of *Euglena* is not well-supported in this study. Other single-gene studies using 18S rDNA were discrepant in the position of the *Euglena* clade and bootstrap support was low (Cavalier-Smith, 2016; Kolisko et al., 2020; Lukešová et al., 2020; Marin et al., 2003; Milanowski et al., 2006; Wang et al., 2021).

Trachelomonas/Strombomonas

Trachelomonas and *Strombomonas* were found to be monophyletic sister groups in studies using multiple nuclear genes (Ciugulea et al., 2008; Triemer et al., 2006) or combined nuclear and chloroplast genes (Karnkowska et al., 2015; Kim et al., 2010, 2015; Linton et al., 2010). In this study, *Strombomonas* was found to be monophyletic and positioned within *Trachelomonas*, making *Trachelomonas* paraphyletic. This position differs from studies using multiple genetic markers but was supported only in the subset analysis. Early single-gene studies using 18S rDNA could not confirm the monophyly of *Trachelomonas* (Brosnan et al., 2003; Müllner et al., 2001; Nudelman et al., 2003). Marin et al. (2003) included *Strombomonas* within the genus *Trachelomonas*. More recent studies using 18S rDNA found *Trachelomonas* to be monophyletic but with low

support (Lukešová et al., 2020) or para- or polyphyletic in Kolisko et al. (2020) and Wang et al. (2021). This suggests that additional genetic markers may be necessary to confirm the monophyly of *Trachelomonas*.

Cryptoglena/Monomorphina/Euglenaria

Cryptoglena and *Monomorphina* were well-supported monophyletic sister groups in studies using different nuclear and/or plastid markers (Dabbagh & Preisfeld, 2018; Karnkowska et al., 2015; Kim et al., 2010, 2015; Linton et al., 2010; Triemer et al., 2006). The sister relationship was also recovered in single-gene studies using 18S rDNA (Karnkowska-Ishikawa et al., 2012; Kolisko et al., 2020; Lukešová et al., 2020; Marin et al., 2003; Wang et al., 2021) and cpLSU (Kim & Shin, 2008). However, bootstrap support was low (Karnkowska-Ishikawa et al., 2012; Kim & Shin, 2008; Kolisko et al., 2020; Lukešová et al., 2020; Marin et al., 2003). This study supports the monophyly and the sister relationship of *Cryptoglena* and *Monomorphina* with high support. In most studies using two or more markers, *Euglenaria* was sister to *Monomorphina* and *Cryptoglena* (Dabbagh & Preisfeld, 2018; Karnkowska et al., 2015; Kim et al., 2010; Linton et al., 2010). This sister relationship was recovered in PNJ analysis with high bootstrap support. In the ML tree, the position differs but was not supported. In other single-gene studies using 18S rDNA, the position of *Euglenaria* is often discrepant (Cavalier-Smith, 2016; Linton et al., 2000; Marin et al., 2003; Milanowski et al., 2006; Müllner et al., 2001; Wang et al., 2021).

Euglenaformis/Colacium

Previous studies using combined nuclear and/or chloroplast genes have placed *Euglenaformis* at the base of *Euglenaceae* (Bennett et al., 2014; Bennett & Triemer, 2012; Dabbagh & Preisfeld, 2018; Karnkowska et al., 2015; Kim et al., 2010, 2015; Linton et al., 2010; Milanowski et al., 2006). In other single-gene studies using 18S rDNA (Cavalier-Smith, 2016; Karnkowska-Ishikawa et al., 2012; Kolisko et al., 2020; Lukešová et al., 2020; Milanowski et al., 2006) or cpLSU (Kim & Shin, 2008), the position of *Euglenaformis* varied and bootstrap support was generally low. In this study, *Euglenaformis* and *Colacium* were sisters which differs from previous studies using multiple markers but was consistent in PNJ and subset analyses. *Colacium* formed a monophyletic sister clade to *Trachelomonas* and *Strombomonas* in previous studies using different nuclear markers (Ciugulea et al., 2008; Triemer et al., 2006) or combined nuclear and chloroplast genes (Karnkowska et al., 2015; Kim et al., 2010, 2015; Linton et al., 2010). This sister relationship was well-supported in Bayesian analyses but had low bootstrap support in ML analyses (Ciugulea et al., 2008; Karnkowska et al., 2015; Kim et al., 2010, 2015; Linton et al., 2010). The position of *Colacium* is often discrepant in single-gene studies using 18S rDNA (Karnkowska-Ishikawa et al., 2012; Kolisko et al., 2020; Marin et al., 2003; Milanowski et al., 2006; Nudelman et al., 2003; Wang et al., 2021) or cpLSU (Kim & Shin, 2008) as well as in some multigene studies using different nuclear and/or plastid markers (Brosnan et al., 2003; Dabbagh & Preisfeld, 2018; Milanowski et al., 2006). Whereas in most previous studies, the position of *Colacium* and *Euglenaformis* was not well-supported in bootstrap analysis; in this study, the sister relationship had high support. This suggests that more detailed molecular studies are necessary to unambiguously resolve the position of *Colacium* and *Euglenaformis*.

CONCLUSION

In this study, we used sequence–structure information simultaneously which was shown to improve robustness and accuracy by Keller et al. (2010). As Keller et al. (2010) suggested, this approach can be applied in NJ, MP, and ML analyses. Sequence–structure profile neighbor-joining yielded a well-supported tree. The ML tree showed a similar topology to the PNJ tree but differed in basal branches due to a lack of bootstrap support. Where bootstrap support was high, the reconstructed phylogenetic trees were generally consistent with studies using a combined set of genetic markers. Topologies that differed from previous multigene studies were generally not well-supported or inconsistent in our analysis and also often discrepant in sequence-only studies using 18S rDNA. The position of *Colacium* and *Euglenaformis* differed from most studies, but their sister relationship showed high support and was consistent in PNJ and subset analyses. Our study supports the simultaneous use of sequence and structural data to reconstruct more accurate and robust trees in comparison with sequence-only analyses and to come as close as possible to multigene marker phylogenies. The average bootstrap value obtained from sequence–structure analyses is significantly higher than the average bootstrap value obtained from sequence-only analyses, which is promising for resolving relationships between more closely related taxa.

ACKNOWLEDGMENTS

We would like to thank Marlyn Weimer (University of Würzburg, Germany) and Selina Brözel (University of Würzburg, Germany) for fruitful discussions. AK thanks for financial support from the EMBO Installation Grant 4150 and from the Ministry of Education and Science, Poland. Open Access funding enabled and organized by Projekt DEAL.

ORCID

Matthias Wolf https://orcid.org/0000-0003-2004-1806

REFERENCES

Ankenbrand, M.J., Keller, A., Wolf, M., Schultz, J. & Förster, F. (2015) ITS2 database V: twice as much. *Molecular Biology and Evolution*, 32, 3030–3032. Available from: https://doi.org/10.1093/molbev/msv174

Bennett, M.S. & Triemer, R.E. (2012) A new method for obtaining nuclear gene sequences from field samples and taxonomic revisions of the photosynthetic Euglenoids Lepocinclis (euglena) helicoideus and Lepocinclis (Phacus) horridus (Euglenophyta). *Journal of Phycology*, 48, 254–260. Available from: https://doi.org/10.1111/j.1529-8817.2011.01101.x

Bennett, M.S., Wiegert, K.E. & Triemer, R.E. (2014) Characterization of Euglenaformis gen. nov. and the chloroplast genome of Euglenaformis [euglena] proxima (Euglenophyta). *Phycologia*, 53, 66–73. Available from: https://doi.org/10.2216/13-198.1

Benson, D.A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. et al. (2013) GenBank. *Nucleic Acids Research*, 41, D36–D42. Available from: https://doi.org/10.1093/nar/gks1195

Borges, A.R., Engstler, M. & Wolf, M. (2021) 18S rRNA gene sequence-structure phylogeny of the Trypanosomatida (Kinetoplastea, Euglenozoa) with special reference to Trypanosoma. *European Journal of Protistology*, 81, 125824. Available from: https://doi.org/10.1016/j.ejop.2021.125824

Brosnan, S., Shin, W., Kjer, K.M. & Triemer, R.E. (2003) Phylogeny of the photosynthetic euglenophytes inferred from the nuclear SSU and partial LSU rDNA. *International Journal of Systematic and Evolutionary Microbiology*, 53, 1175–1186. Available from: https://doi.org/10.1099/ijs.0.02518-0

Buchheim, M.A., Müller, T. & Wolf, M. (2017) 18S rDNA sequence-structure phylogeny of the Chlorophyceae with special emphasis on the Sphaeropleales. *Plant Gene*, 10, 45–50. Available from: https://doi.org/10.1016/j.plgene.2017.05.005

Camin, J.H. & Sokal, R.R. (1965) A method for deducing branching sequences in phylogeny. *Evolution*, 19, 311–326. Available from: https://doi.org/10.2307/2406441

Cavalier-Smith, T. (2016) Higher classification and phylogeny of Euglenozoa. *European Journal of Protistology*, 56, 250–276. Available from: https://doi.org/10.1016/j.ejop.2016.09.003

Ciugulea, I., Nudelman, M.A., Brosnan, S. & Triemer, R.E. (2008) Phylogeny of the Euglenoid loricate genera Trachelomonas and Strombomonas (Euglenophyta) inferred from nuclear SSU and LSU rDNA. *Journal of Phycology*, 44, 406–418. Available from: https://doi.org/10.1111/j.1529–8817.2008.00472.x

Czech, V. & Wolf, M. (2020) RNA consensus structures for inferring green algal phylogeny: a three–taxon analysis for Golenkinia/Jenufa, Sphaeropleales and Volvocales (Chlorophyta, Chlorophyceae). *Fottea*, 20, 68–74. Available from: https://doi.org/10.5507/fot.2019.016

Dabbagh, N. & Preisfeld, A. (2018) Intrageneric variability between the chloroplast genomes of *Trachelomonas grandis* and *Trachelomonas volvocina* and phylogenomic analysis of phototrophic Euglenoids. *The Journal of Eukaryotic Microbiology*, 65, 648–660. Available from: https://doi.org/10.1111/jeu.12510

Felsenstein, J. (1981) Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution*, 35, 1229–1242. Available from: https://doi.org/10.1111/j.1558–5646.1981.tb04991.x

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using bootstrap. *Evolution*, 39, 783–791. Available from: https://doi.org/10.2307/2408678

Friedrich, J., Dandekar, T., Wolf, M. & Müller, T. (2005) ProfDist: a tool for the construction of large phylogenetic trees based on profile distances. *Bioinformatics*, 21, 2108–2109. Available from: https://doi.org/10.1093/bioinformatics/bti289

Heeg, J.S. & Wolf, M. (2015) ITS2 and 18S rDNA sequence-structure phylogeny of chlorella and allies (Chlorophyta, Trebouxiophyceae, Chlorellaceae). *Plant Gene*, 4, 20–28. Available from: https://doi.org/10.1016/j.plgene.2015.08.001

Hepperle, D. (2004) Align Ver.07/04©. multisequence alignment editor and preparation/manipulation of phylogenetic datasets. Win32-Version. Available from: http://www.sequentix.de. [Accessed 10th June 2022].

Jukes, T.H. & Cantor, C.R. (1969) Evolution of protein molecules. In: Munro, H.N. (Ed.) *Mammalian Protein Metabolism*, Vol. 3. New York, NY: Academic Press, pp. 21–132. Available from: https://doi.org/10.1016/B978-1-4832-3211-9.50009-7

Karnkowska, A., Bennett, M.S., Watza, D., Kim, J.I., Zakryś, B. & Triemer, R.E. (2015) Phylogenetic relationships and morphological character evolution of photosynthetic Euglenids (Excavata) inferred from taxon-rich analyses of five genes. *The Journal of Eukaryotic Microbiology*, 62, 362–373. Available from: https://doi.org/10.1111/jeu.12192

Karnkowska-Ishikawa, A., Milanowski, R., Triemer, R.E. & Zakryś, B. (2012) Taxonomic revisions of morphologically similar species from two *Euglenoid genera*: euglena (*E. granulata* and *E. velata*) and *Euglenaria* (*Eu. anabaena*, *Eu. caudata*, and *Eu. clavata*). *Journal of Phycology*, 48, 729–739. Available from: https://doi.org/10.1111/j.1529-8817.2012.01140.x

Keller, A., Förster, F., Müller, T., Dandekar, T., Schultz, J. & Wolf, M. (2010) Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. *Biology Direct*, 5, 4. Available from: https://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-5-4

Kim, J.I., Linton, E.W. & Shin, W. (2015) Taxon-rich multigene phylogeny of the photosynthetic Euglenoids (Euglenophyceae). *Frontiers in Ecology and Evolution*, 3, 98. Available from: https://doi.org/10.3389/fevo.2015.00098

Kim, J.I. & Shin, W. (2008) Phylogeny of the Euglenales inferred from plastid LSU rDNA sequences. *Journal of Phycology*, 44, 994–1000. Available from: https://doi.org/10.1111/j.1529-8817.2008.00536.x

Kim, J.I. & Shin, W. (2014) Molecular phylogeny and cryptic diversity of the genus Phacus (Phacaceae, Euglenophyceae) and the descriptions of seven new species. *Journal of Phycology*, 50, 948–959. Available from: https://doi.org/10.1111/jpy.12227

Kim, J.I., Shin, W. & Triemer, R.E. (2010) Multigene analyses of photosynthetic Euglenoids and new family, Phacaceae (Euglenales). *Journal of Phycology*, 46, 1278–1287. Available from: https://doi.org/10.1111/j.1529-8817.2010.00910.x

Kolisko, M., Flegontova, O., Karnkowska, A., Lax, G., Maritz, J.M., Pánek, T. et al. (2020) EukRef-excavates: seven curated SSU ribosomal RNA gene databases. *Database*, 2020, baaa080. Available from: https://doi.org/10.1093/database/baaa080

Kostygov, A., Karnkowska, A., Votýpka, J., Tashyreva, D., Maciszewski, K., Yurchenko, V. et al. (2021) Euglenozoa: taxonomy, diversity and ecology, symbioses and viruses. *Open Biology*, 11, 200407. Available from: https://doi.org/10.1098/rsob.200407

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H. et al. (2007) ClustalW and ClustalX version 2.0. *Bioinformatics*, 23, 2947–2948. Available from: https://doi.org/10.1093/bioinformatics/btm404

Lim, H.C., Teng, S.T., Lim, P.T., Wolf, M. & Leaw, C.P. (2016) 18S rDNA phylogeny of Pseudonitzschia (Bacillariophyceae) inferred from sequence-structure information. *Phycologia*, 55, 134–146. Available from: https://doi.org/10.2216/15-78.1

Linton, E.W., Karnkowska-Ishikawa, A., Kim, J.I., Shin, W., Bennett, M.S., Kwiatowski, J. et al. (2010) Reconstructing Euglenoid evolutionary relationships using three genes: nuclear SSU and LSU, and chloroplast SSU rDNA sequences and the description of *Euglenaria* gen. nov. (Euglenophyta). *Protist*, 161, 603–619. Available from: https://doi.org/10.1016/j.protis.2010.02.002

Linton, E.W., Nudelman, M.A., Conforti, V. & Triemer, R.E. (2000) A molecular analysis of the euglenophytes using SSU rDNA.

Journal of Phycology, 36, 740–746. Available from: https://doi.org/10.1046/j.1529–8817.2000.99226.x

Lukešová, S., Karlicki, M., Hadariová, L.T., Szabová, J., Karnkowska, A. & Hampl, V. (2020) Analyses of environmental sequences and two regions of chloroplast genomes revealed the presence of new clades of photosynthetic Euglenids in marine environments. *Environmental Microbiology Reports*, 12, 78–91. Available from: https://doi.org/10.1111/1758-2229.12817

Łukomska-Kowalczyk, M., Chaber, K., Fells, A., Milanowski, R. & Zakryś, B. (2021) Description of Flexiglena gen. nov. and new members of Discoplastis and Euglenaformis (Euglenida). *Journal of Phycology*, 57, 766–779. Available from: https://doi.org/10.1111/jpy.13107-20-159

Maddison, D.R., Swofford, D.L. & Maddison, W.P. (1997) Nexus: an extensible file format for systematic information. *Systematic Biology*, 46, 590–621. Available from: https://doi.org/10.1093/sysbio/46.4.590

Marin, B., Palm, A., Klingberg, M. & Melkonian, M. (2003) Phylogeny and taxonomic revision of plastid-containing euglenophytes based on SSU rDNA sequence comparisons and synapomorphic signatures in the SSU rRNA secondary structure. *Protist*, 154, 99–145. Available from: https://doi.org/10.1078/143446103764928521

Markert, S.M., Müller, T., Koetschan, C., Friedl, T. & Wolf, M. (2012) "Y" Scenedesmus (Chlorophyta, Chlorophyceae): the internal transcribed spacer 2 rRNA secondary structure revisited. *Plant Biology*, 14, 987–996. Available from: https://doi.org/10.1111/j.1438-8677.2012.00576.x

Matzov, D., Taoka, M., Nobe, Y., Yamauchi, Y., Halfon, Y., Asis, N. et al. (2020) Cryo-EM structure of the highly atypical cytoplasmic ribosome of *Euglena gracilis*. *Nucleic Acids Research*, 48, 11750–11761. Available from: https://doi.org/10.1093/nar/gkaa893

Milanowski, R., Kosmala, S., Zakryś, B. & Kwiatowski, J. (2006) Phylogeny of photosynthetic euglenophytes based on combined chloroplast and cytoplasmic SSU rDNA sequence analysis. *Journal of Phycology*, 42, 721–730. Available from: https://doi.org/10.1111/j.1529-8817.2006.00216.x

Müller, T., Rahmann, S., Dandekar, T. & Wolf, M. (2004) Accurate and robust phylogeny estimation based on profile distances: a study of the Chlorophyceae (Chlorophyta). *BMC Evolutionary Biology*, 4, 20. Available from: https://doi.org/10.1186/1471-2148-4-20

Müllner, A.N., Angeler, D.G., Samuel, R., Linton, E.W. & Triemer, R.E. (2001) Phylogenetic analysis of phagotrophic, phototrophic and osmotrophic Euglenoids by using the nuclear 18S rDNA sequence. *International Journal of Systematic and Evolutionary Microbiology*, 51, 783–791. Available from: https://doi.org/10.1099/00207713-51-3-783

Nudelman, M.A., Rossi, M.S., Conforti, V. & Triemer, R.E. (2003) Phylogeny of Euglenophyceae based on small subunit rDNA sequences: taxonomic implications. *Journal of Phycology*, 39, 226–235. Available from: https://doi.org/10.1046/j.1529-8817.2003.02075.x

Plieger, T. & Wolf, M. (2022) 18S and ITS2 rDNA sequence-structure phylogeny of Prototheca (Chlorophyta, Trebouxiophyceae). *Biologia*, 77, 569–582. Available from: https://doi.org/10.1007/s11756-021-00971-y

R Core Team. (2018) *R: a language and environment for statistical computing*. Vienna, Austria: R foundation for statistical computing. Available from: https://r-project.org/. [Accessed 15th June 2022].

Rahmann, S., Müller, T., Dandekar, T. & Wolf, M. (2006) Efficient and robust analysis of large phylogenetic datasets. In: Hsu, H. (Ed.) *Advanced data mining technologies in bioinformatics*. Hershey: IGI Global, pp. 104–117. Available from: https://doi.org/10.4018/978-1-59140-863-5.ch006

Saitou, N. & Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4, 406–425. Available from: https://doi.org/10.1093/oxfordjournals.molbev.a040454

Schliep, K.P. (2011) Phangorn: phylogenetic analysis in R. *Bioinformatics*, 27, 592–593. Available from: https://doi.org/10.1093/bioinformatics/btq706

Seibel, P.N., Müller, T., Dandekar, T., Schultz, J. & Wolf, M. (2006) 4SALE – a tool for synchronous RNA sequence and secondary structure alignment and editing. *BMC Bioinformatics*, 7, 498. Available from: https://doi.org/10.1186/1471-2105-7-498

Seibel, P.N., Müller, T., Dandekar, T. & Wolf, M. (2008) Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. *BMC Research Notes*, 1, 91. Available from: https://doi.org/10.1186/1756-0500-1-91

Selig, C., Wolf, M., Müller, T., Dandekar, T. & Schultz, J. (2008) The ITS2 database II: homology modelling RNA structure for molecular systematics. *Nucleic Acids Research*, 36, D377–D380. Available from: https://doi.org/10.1093/nar/gkm827

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30, 1312–1313. Available from: https://doi.org/10.1093/bioinformatics/btu033

Swofford, D.L. (2002) *PAUP*. Phylogenetic analysis using parsimony (*and other methods) version 4*. Sunderland, MA: Sinauer Associates.

Triemer, R.E., Linton, E., Shin, W., Nudelman, A., Monfils, A., Bennett, M. et al. (2006) Phylogeny of the Euglenales based upon combined SSU and LSU rDNA sequence comparisons and description of Discoplastis gen. nov. (Euglenophyta). *Journal of Phycology*, 42, 731–740. Available from: https://doi.org/10.1111/j.1529-8817.2006.00219.x

Wang, Y., Feng, J., Lv, J., Liu, Q., Nan, F., Liu, X. et al. (2021) Phylogenetic and morphological evolution of green euglenophytes based on 18S rRNA. *The Journal of Eukaryotic Microbiology*, 68, e12824. Available from: https://doi.org/10.1111/jeu.12824

Wolf, M., Achtziger, M., Schultz, J., Dandekar, T. & Müller, T. (2005) Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA*, 11, 1616–1623. Available from: https://doi.org/10.1261/rna.2144205

Wolf, M., Koetschan, C. & Müller, T. (2014) ITS2, 18S, 16S or any other RNA – simply aligning sequences and their individual secondary structures simultaneously by an automatic approach. *Gene*, 546, 145–149. Available from: https://doi.org/10.1016/j.gene.2014.05.065

Wolf, M., Ruderisch, B., Dandekar, T., Schultz, J. & Müller, T. (2008) ProfDistS: (profile-) distance based phylogeny on sequence-structure alignments. *Bioinformatics*, 24, 2401–2402. Available from: https://doi.org/10.1093/bioinformatics/btn453

Yamaguchi, A., Yubuki, N. & Leander, B.S. (2012) Morphostasis in a novel eukaryote illuminates the evolutionary transition from phagotrophy to phototrophy: description of Rapaza viridis n. gen. Et sp. (Euglenozoa, Euglenida). *BMC Evolutionary Biology*, 12, 29. Available from: https://doi.org/10.1186/1471-2148-12-29

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Rackevei, A.S., Karnkowska, A. & Wolf, M. (2023) 18S rDNA sequence–structure phylogeny of the *Euglenophyceae* (Euglenozoa, Euglenida). *Journal of Eukaryotic Microbiology*, 70, e12959. Available from: https://doi.org/10.1111/jeu.12959