Eukaryotic Microbiology | ISOP | Published by the International Society of Protistologists

Transcriptomics of Diphyllatea (CRuMs) from South Pacific crater lakes confirm new cryptic clades

Luis Javier Galindo[1,2,3] | **Varsha Mathur**[1] | **Hadleigh Frost**[4] | **Guifré Torruella**[5] | **Thomas A. Richards**[1] | **Nicholas A. T. Irwin**[1,4,6]

[1]Department of Biology, University of Oxford, Oxford, UK

[2]Instituto Universitario de Investigación del Agua, Universidad de Granada, Granada, Spain

[3]Departamento de Ecología, Facultad de Ciencias, Universidad de Granada, Granada, Spain

[4]Merton College, University of Oxford, Oxford, UK

[5]Institut de Biologia Evolutiva (CSIC-UPF), Barcelona, Catalonia, Spain

[6]Gregor Mendel Institute (GMI), Austrian Academy of Sciences, Vienna BioCenter (VBC), Vienna, Austria

Correspondence
Luis Javier Galindo and Nicholas A. T. Irwin, Department of Biology, University of Oxford, Oxford, UK.
Email: luisjagg92@gmail.com and nicholas.irwin@gmi.oeaw.ac.at

Abstract

The Diphyllatea (CRuMs) are heterotrophic protists currently divided into three distinct clades (Diphy I–III). Diphy I are biflagellates in the genus *Diphylleia*, whereas Diphy II and III represent cryptic clades comprising *Collodictyon*-type quadriflagellates that were recently distinguished based on rRNA gene phylogenies. Here, we isolated Diphyllatea from freshwater crater lakes on two South Pacific islands and generated high-quality transcriptomes from species representing each clade, including the first transcriptomic data from Diphy III. Phylogenomic analyses support the separation of Diphy II and III, while transcriptome completeness highlights the utility of these data for future studies. Lastly, we discuss the biogeography and ecology of Diphyllatea on these remote islands.

KEYWORDS
biogeography, *Collodictyon*, crater lakes, Diphyllatea, phylogenomics, single-cell, transcriptomics

INTRODUCTION

CRUMS is a eukaryotic supergroup composed of morphologically diverse heterotrophic organisms (Brown et al., 2018; Zhao et al., 2012). The CRuMs branch sister to the Amorphea (Brown et al., 2018; Burki et al., 2020), making them an important clade for reconstructing early eukaryotic evolution (Blaz et al., 2023). The diversity of cellular forms within CRuMs includes filopodial amoebas within the Rigifilida (Mikrjukov & Mylnikov, 2001; Yabuki et al., 2013), gliding nanoflagellates within the Mantamonadida (Blaz et al., 2023; Cavalier-Smith et al., 2014; Glücksman et al., 2011), and large multiflagellated predatory swimming protists within the class Diphyllatea (Brugerolle, 2006; Brugerolle et al., 2002; Orr et al., 2018). Based on morphological observations and SSU (small subunit) rRNA gene phylogenies, the class Diphyllatea (order Diphylleida) has been proposed to group the families Diphylleidae and Sulcomonadidae. The Diphylleidae includes the biflagellate, *Diphylleia rotans*, and the quadriflagellate, *Collodictyon triciliatum* (Cavalier-smith, 2013; Ruggiero et al., 2015; Zhao et al., 2012), whereas

Luis Javier Galindo and Varsha Mathur contributed equally to this work.

the Sulcomonadidae is represented by the biflagellate species, *Sulcomonas lacustris* (Brugerolle, 2006).

A recent advance in the taxonomy of Diphyllatea came following the isolation and morphological characterization of 11 new isolates (Orr et al., 2018). Orr et al. (2018) sequenced the SSU rRNA genes of these organisms and generated the most complete phylogeny, in terms of diversity, of the group to date. Based on this phylogeny, the authors proposed that Diphyllatea is composed of three major clades: Diphy I, II, and III. Diphy I corresponds to the biflagellated *Diphylleia*, while clades II and III are sister groups comprising cryptic quadriflagellated species with a *Collodictyon triciliatum*-type morphology. Importantly, this work identified Diphy III as a distinct clade at the same taxonomic level as Diphy I and II. Given SSU rRNA gene sequence divergence between Diphy II and III (79–89% sequence identity), the authors argued that these two cryptic clades, which were morphologically indistinguishable using light microscopy, could represent different genera (Orr et al., 2018). However, the SSU rRNA gene can misrepresent microbial species diversity (Piganeau et al., 2011), and there is little molecular data available from the Diphyllatea for confirmatory phylogenomic analyses or wider comparative genomic studies. Currently, there is only one high-quality transcriptome available for *D. rotans* (Brown et al., 2018) (Diphy I clade) and a few 454 pyrosequencing contigs from *C. triciliatum* (Diphy II clade) (Zhao et al., 2012). Accordingly, we lack the data required to confirm the phylogenetic relationships within the Diphyllatea and understand CRuMs evolution more broadly.

To address this, we generated four new transcriptomes representing each of the three clades of Diphyllatea, with the aim of reconstructing the evolutionary relationships within the group and improving the availability of CRuMs transcriptomic data. To accomplish this, we isolated single cells from water samples derived from four freshwater volcanic crater lakes on two geographically distant South Pacific islands, Uvea (Wallis & Futuna) and Upolu (Samoa). These transcriptomes include one representative of Diphy I, two from Diphy II, and the first transcriptome from Diphy III. These high-quality transcriptomes will provide a useful resource for future genomic studies and permitted phylogenomic analyses, confirming that Diphy III is a distinct clade sister to Diphy II. Lastly, we hypothesize about how these freshwater protists colonized the remote and isolated crater lakes in the South Pacific and discuss their potential ecological role as microbial predators and grazers.

MATERIALS AND METHODS

Sample collection, cell isolation, transcriptome sequencing, and assembly

Water samples were collected from freshwater volcanic lakes on the islands of Uvea (Wallis and Futuna) and Upolu (Samoa) between the 24th of April and 15th of May of 2023 as part of an effort to characterize the protist diversity of volcanic crater lakes in the South Pacific. The water samples used in this study were collected from lakes Kikila (surface water, first 10 cm; 13°17′48.3″ S 176°11′24.7″ W) and Lanutavake (3 m depth; 13°19′16.2″ S 176°12′50.6″ W) on the island of Uvea (Wallis and Futuna) and from lakes Lanoto'o (at 1.5 m depth; 13°54′39.6″ S 171°49′40.1″ W) and Pue (at 3 m depth; 13°56′13.0″ S 171°45′14.7″ W) on the island of Upolu (Samoa) (Figure 1A). Water was collected using a 5 L Niskin water sampler deployed from an inflatable kayak. For each sample, 40 mL of lake water was stored in a 25 cm^2 vented cell culture flask at room temperature and taken to the laboratory for further analysis.

Diphyllatea cells were identified based on morphology using an Olympus CKX53 inverted microscope, and single cells were isolated using glass capillary micropipettes. We generated two samples per lake, which each contained pools of 3–7 morphologically consistent cells for downstream cDNA synthesis using the Smart-Seq2 protocol (Picelli et al., 2014). The cells were washed at least three times in sterile molecular-grade water (Fisher Scientific), photographed when possible, and transferred to a 0.2 mL thin-walled PCR tube containing 2 µL of cell lysis buffer (0.2% Triton X-100 and RNase inhibitor (Invitrogen)). The samples were incubated with free dNTPs and the oligo-dT primer (5′–AAGCAGTGGTATCAACGCAGAGTACT$_{30}$VN-3′), which anneals to RNA molecules containing a poly(A) tail. Next, the reverse transcription (RT) reaction was performed using Superscript II reverse transcriptase (Invitrogen) and a template-switching oligo (TSO) (5′-AAGCAGTGGTATCAACGCAGAGTACATrGrG+G-3′). The resulting cDNA was amplified using the IS PCR primer (5′ AAGCAGTGGTATCAACGCAGAGT-3′) and KAPA HiFi HotStart ReadyMix (2X; KAPA Biosystems) for 18 PCR cycles. All steps were performed using reaction mixes and thermal cycler conditions described in the Smart-Seq2 protocol (Picelli et al., 2014). Finally, the cDNA was purified using Ampure XP beads (Beckman Coulter), and the cDNA concentration was quantified with a Qubit 2.0 fluorometer (Thermo Fisher Scientific Inc.). Prior to high-throughput sequencing, 2 µL of the final cDNA product was used as a template for PCR amplification of the 18S rRNA gene. GoTaq DNA Polymerase (Thermo Fisher Scientific) and the general eukaryotic primers 82F (5′-GAAACTGCGAATGGCTC-3′) and 1520R (5′-CYGCAGGTTCACCTAC-3′) were used with an annealing temperature of 55°C (30 s) and an extension time of 2 min (72°C). PCR products were sequenced using Sanger sequencing (performed by Eurofins), and sequences were identified by performing BLASTn searches against the nonredundant NCBI database (Pruitt et al., 2007). After identification, sequencing libraries were prepared using the Novogene NGS RNA library prep set protocol (PT042) and sequenced on a
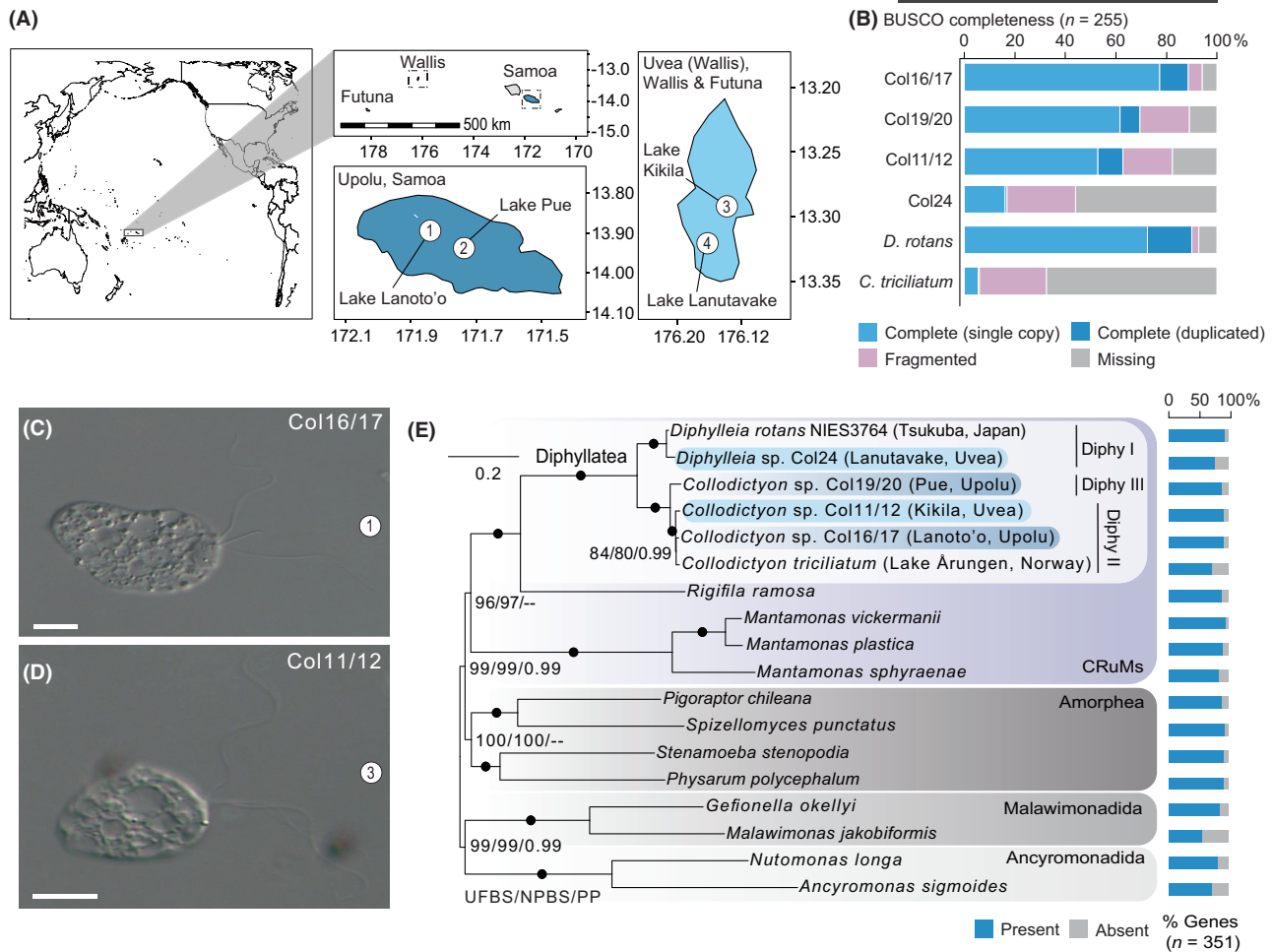
FIGURE 1 (A) Maps depicting the locations of the sampled lakes on the islands of Upolu (Samoa) and Uvea (Wallis & Futuna). Numbers correspond to each of the lakes: Lake Lanoto'o (1), Lake Pue (2), Lake Kikila (3) and Lake Lanutavake (4). (B) BUSCO completeness bar charts for all the transcriptomic data available for the Diphyllatea based on the eukaryotic_odb10 dataset. (C, D) DIC light micrographs of *Collodictyon* sp. cells (Diphy II) from Lakes Lanoto'o (C) and Kikila (D); scale bars 10 μm. (E) Maximum-likelihood phylogenomic tree of CRuMs and their relatives constructed using a dataset comprising 126,700 amino acid positions derived from 351 concatenated proteins, representing 18 CRuMs species and their relatives. The phylogeny was inferred using the PMSF approximation of the LG+C60+G model and statistical support was assessed using ultrafast bootstraps (UFBS) and nonparametric bootstraps (NPBS). Bayesian posterior probabilities (PP) are also shown and were inferred using the CAT-Poisson model. Nodes with full statistical support are indicated with black circles and the scale bar represents the average number of substitutions per site. Light and dark blue coloring corresponds to the island of Uvea (Wallis & Futuna) and Upolu (Samoa), respectively.

single lane of an Illumina Novoseq 6000 using 150-bp paired-end reads. All raw reads have been deposited in NCBI Bioproject PRJNA1100983.

The quality of the raw sequencing reads was assessed using FastQC v0.12.1 (Andrews, 2010). Low-quality reads, adapters, and primer sequences were trimmed using Trimmomatic v0.39, and the remaining reads were assembled with rnaSpades v3.13.0 (Bolger et al., 2014; Bushmanova et al., 2019). To remove bacterial contamination, both BLASTn searches against the NCBI nt database and Diamond v2.1.7 BLASTx searches against UniProt reference proteomes were conducted (Buchfink et al., 2021; Consortium, 2022). The bacterial reads were filtered out using BlobToolKit v4.2.1 and Bowtie v2.5.1 (Laetsch & Blaxter, 2017; Langmead et al., 2009). Given our SSU rRNA gene sequencing results, we combined

assembled transcripts derived from similar strains (SSU rRNA gene sequence similarity >99%) that were obtained from the same lake sample. Accordingly, we generated the following four merged Diphyllatea transcriptomes: Col11/12 (Lake Kikila), Col16/17 (Lake Lanoto'o), Col19/20 (Lake Pue), and Col24 (Lake Lanutavake). Protein coding regions were predicted using TransDecoder v5.7.1 (https://github.com/TransDecoder/TransDecoder) and clustered with CD-HIT v4.8.1 (Li & Godzik, 2006) at 99% identity. To maximize sensitivity for capturing ORFs, Diamond v2.1.7 BLASTx searches against the UniProt database were carried out to retain ORFs with homology to known proteins (Buchfink et al., 2021; Consortium, 2022). The completeness of the predicted proteomes was assessed using BUSCO v5.5.0 (Manni et al., 2021; Simão et al., 2015)

using the eukaryota_odb10 lineage dataset. Assembled transcripts and proteomes are available in a Figshare data repository (https://figshare.com/projects/Diphyllatea_from_the_South_Pacific/200110).

SSU rRNA gene phylogeny and phylogenomic analyses

To identify the species isolated for transcriptomics, the Diphyllatea SSU rRNA gene dataset developed by Orr et al. (2018) was expanded with SSU sequences extracted from the sequenced transcriptomes. These 38 sequences were aligned with MAFFT v.7 (Katoh & Standley, 2013) and trimmed with trimAL v.1.2 (Capella-Gutiérrez et al., 2009) using a gap-threshold of 30%, resulting in an alignment with 1078 sites. The resulting alignment was used to generate a maximum likelihood phylogeny with IQ-TREE v.1.6.11 (Nguyen et al., 2015) using the GTR+F+I+R2 model, selected using ModelFinder (Kalyaanamoorthy et al., 2017) based on Bayesian information criteria (BIC). Statistical support was inferred from 1000 ultra-fast bootstraps.

Using the transcriptomic data, we also performed a multigene phylogenomic analysis using a CRuMs-enriched dataset comprising 351 conserved protein markers from 18 species (Blaz et al., 2023; Lax et al., 2018). Each protein marker was aligned with MAFFT v.7 and trimmed with trimAL v.1.2 with the automated1 option. Single-protein trees were inferred from each trimmed alignment using FastTree v2.1.11 (Price et al., 2009). Alignments and trees were manually inspected to discard contaminants and paralogs and were edited using Geneious v11.0.18 (Kearse et al., 2012). Representative taxa were then selected (10 CRuMs +8 outgroups from Amorphea, Ancyromonadida, and Malawimonadida), and the resulting proteins were realigned and trimmed as stated above before finally being concatenated with alvert.py from the barrel-o-monkeys package (http://rogerlab.biochemistryandmolecularbiology.dal.ca/Software/Software.htm#Monkeybarrel). This resulted in a concatenated matrix with 126,700 amino acid positions.

The maximum likelihood (ML) phylogenomic trees were inferred using IQ-TREE v1.6 using the PMSF approximation (Wang et al., 2018) with a guide tree using the LG+C60+G model. Statistical support was generated from two reconstruction rounds with 1000 ultrafast bootstraps and 100 nonparametric bootstraps. All ML substitution models were selected using ModelFinder based on BIC. Bayesian inference was performed using PhyloBayes-MPI v1.5a (Lartillot et al., 2013). Four MCMC chains were run for 10,000 generations using the CAT-Poisson model, sampling one every tenth tree. Of the four chains, three properly converged (max difference=0.0277) as assessed using Tracer and PhyloBayes using a burn-in of 25% (Rambaut et al., 2018). Phylogenies were visualized using FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/). All alignments and trees are available in a Figshare data repository (https://figshare.com/projects/Diphyllatea_from_the_South_Pacific/200110).

RESULTS AND DISCUSSION

Four new transcriptomes of Diphyllatea

To investigate the phylogenetic relationships amongst the Diphyllatea, we sequenced seven transcriptomes of Diphyllatea, including two from Lake Kikila (Col11 and 12), two from Lake Lanoto'o (Col16 and 17), two from Lake Pue (Col19 and 20), and one from Lake Lanutavake (Col24) (Figure 1A). Based on relatedness (SSU>99% pairwise identity), SSU rRNA gene phylogeny (Figure S1), and sample locality, we merged the transcripts of Col11/12, Col16/17, and Col19/20, resulting in our final dataset of four transcriptomes, including Col24. After removing bacterial contamination, we retrieved high BUSCO presence values (89% to 94.2%) for three of our transcriptomic datasets, which is a significant improvement from the 22.6% BUSCO completeness score from the currently available 454 contigs of *C. triciliatum* (Blaz et al., 2023; Zhao et al., 2012) (Figure 1B; Table S1). In contrast, Col24 (our nonmerged sample) had a lower BUSCO presence of 44%. However, Col24 is closely related to *Diphylleia* (Diphy I; SSU pairwise identity of 99.7%), for which there is a good-quality transcriptome available (Brown et al., 2018) with a BUSCO presence of over 90% (Blaz et al., 2023). Our results highlight that single-cell transcriptome data can be improved by pooling isolated cells or data derived from individual cells, as has been observed in other studies (Keeling & Del Campo, 2017; Mangot et al., 2017). Given their quality, these new Diphyllatea transcriptomes will also provide a valuable resource for future comparative genomic studies.

The phylogeny of Diphyllatea

Phylogenetic analysis of the SSU rRNA gene from the new Diphyllatea transcriptomes was congruent with the analyses of Orr et al. (2018) and supported the presence of three distinct Diphyllatea clades (Figure S1). Col24 groups within Diphy I, along with the genus *Diphylleia*, whereas Col11, 12, 16, and 17 group within Diphy II alongside *Collodictyon*-type quadriflagellated species such as *C. triciliatum*. Notably, Col19 and 20 branch sister to Diphy II in the recently identified Diphy III clade. Although we could only obtain light microscopy images of cells from the Col 11/12 and 16/17 samples (Diphy II; Figure 1C,D), our microscopy observations without imaging confirmed that the cells isolated from Diphy I and III were biflagellated and quadriflagellated, respectively. These results therefore corroborate the identification of

Diphy III as an independent clade of cryptic quadriflagellated *Collodictyon*-type species.

To support these results, we generated a concatenated multi-gene matrix for phylogenomic analysis. Both maximum likelihood and Bayesian inference analysis using the phylogenomic matrix strongly supported the placement of all four new transcriptomes within the three clades of Diphyllatea, confirming the genetic distinctiveness of Diphy III and its sister relationship with Diphy II (Figure 1E). Taken together, these results support the establishment of new taxonomic assignments for Diphy II-III, possibly as new genera as suggested by Orr et al. (2018). However, future work focusing on the isolation and morphological characterization of Diphy II-III species, particularly at an ultrastructural level, will be required before formal taxonomic classifications can be made. Moreover, it will be essential to obtain molecular data from *Sulcomonas lacustris* (Sulcomonadidae) to confirm its relationship within the Diphyllatea.

Diphyllatea in crater lakes of South Pacific islands

Diphyllatea are globally distributed in freshwater environments (Orr et al., 2018). However, it is interesting to find these exclusively freshwater protists in remote and isolated crater lakes in two different South Pacific islands, hundreds of kilometers away from each other and alternative terrestrial water systems. The long-range dispersal and biogeography of microbial eukaryotes has been extensively discussed previously and is thought to be influenced by factors such as cell size, environmental tolerance, and the ability to encyst (Bamforth, 1981; Bass et al., 2007; Finlay, 2002; Foissner, 2006, 2007). The introduction of Diphyllatea to these lakes must have involved long-distance travel and would have required overcoming significant environmental barriers, particularly given that the ancestor of Diphyllatea and Rigifilida was likely adapted to freshwater (Orr et al., 2018; Yabuki et al., 2013). However, the Diphyllatea are known to produce thick-walled resting cysts, which could have facilitated their dispersal (Foissner, 2007; Orr et al., 2018). The remoteness of these crater lakes complicates human-mediated dispersal, suggesting that their colonization could have been via natural mechanisms such as wind or animal vectors like marine birds. This dispersal ability, in combination with the known role of Diphyllatea as predators and grazers of (mainly) algae and cyanobacteria (e.g., Klaveness, 1995; Orr et al., 2018; Zhao et al., 2012), suggests that these protists could serve an important role in the establishment of trophic networks in emerging freshwater ecosystems. Lastly, our observations confirm that Diphyllatea represent a useful study group for tracking how freshwater protists can disperse across wide biogeographical territories and significant environmental boundaries.

ORCID

Luis Javier Galindo https://orcid.org/0000-0003-3642-3200
Varsha Mathur https://orcid.org/0000-0002-7695-034X
Guifré Torruella https://orcid.org/0000-0002-6534-4758
Thomas A. Richards https://orcid.org/0000-0002-9692-0973
Nicholas A. T. Irwin https://orcid.org/0000-0002-2904-8214

REFERENCES

Andrews, S. (2010) *FastQC: A quality control tool for high throughput sequence data.* http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bamforth, S.S. (1981) Protist biogeography. *The Journal of Protozoology*, 28, 2–9. Available from: https://doi.org/10.1111/j.1550-7408.1981.tb02794.x

Bass, D., Richards, T.A., Matthai, L., Marsh, V. & Cavalier-Smith, T. (2007) DNA evidence for global dispersal and probable endemicity of protozoa. *BMC Evolutionary Biology*, 7, 162. Available from: https://doi.org/10.1186/1471-2148-7-162

Blaz, J., Galindo, L.J., Heiss, A.A., Kaur, H., Torruella, G., Yang, A. et al. (2023) One high quality genome and two transcriptome datasets for new species of *Mantamonas*, a deep-branching eukaryote clade. *Scientific Data*, 10, 603. Available from: https://doi.org/10.1038/s41597-023-02488-2

Bolger, A.M., Lohse, M. & Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30, 2114–2120. Available from: https://doi.org/10.1093/bioinformatics/btu170

Brown, M.W., Heiss, A.A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A.K. et al. (2018) Phylogenomics places orphan protistan lineages in a novel eukaryotic super-group. *Genome Biology and Evolution*, 10, 427–433. Available from: https://doi.org/10.1093/gbe/evy014

Brugerolle, G. (2006) Description of a new freshwater heterotrophic flagellate *Sulcomonas lacustris* affiliated to the collodictyonids. *Acta Protozoologica*, 45, 175–182.

Brugerolle, G., Bricheux, G., Philippe, H. & Coffea, G. (2002) *Collodictyon triciliatum* and *Diphylleia rotans* (=*Aulacomonas submarina*) form a new family of flagellates (Collodictyonidae) with tubular mitochondrial cristae that is phylogenetically distant from other flagellate groups. *Protist*, 153, 59–70. Available from: https://doi.org/10.1078/1434-4610-00083

Buchfink, B., Reuter, K. & Drost, H.-G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18, 366–368. Available from: https://doi.org/10.1038/s41592-021-01101-x

Burki, F., Roger, A.J., Brown, M.W. & Simpson, A.G.B. (2020) The new tree of eukaryotes. *Trends in Ecology & Evolution*, 35, 43–55. Available from: https://doi.org/10.1016/j.tree.2019.08.008

Bushmanova, E., Antipov, D., Lapidus, A. & Prjibelski, A.D. (2019) rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *GigaScience*, 8, giz100. Available from: https://doi.org/10.1093/gigascience/giz100

Capella-Gutiérrez, S., Silla-Martínez, J.M. & Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25, 1972–1973. Available from: https://doi.org/10.1093/bioinformatics/btp348

Cavalier-Smith, T. (2013) Early evolution of eukaryote feeding modes, cell structural diversity, and classification of the protozoan phyla Loukozoa, Sulcozoa, and Choanozoa. *European Journal of Protistology*, 49, 115–178. Available from: https://doi.org/10.1016/j.ejop.2012.06.001

Cavalier-Smith, T., Chao, E.E., Snell, E.A., Berney, C., Fiore-Donno, A.M. & Lewis, R. (2014) Multigene eukaryote phylogeny reveals the likely protozoan ancestors of opisthokonts (animals, fungi, choanozoans) and Amoebozoa. *Molecular Phylogenetics and Evolution*, 81, 71–85. Available from: https://doi.org/10.1016/j.ympev.2014.08.012

Finlay, B.J. (2002) Global dispersal of free-living microbial eukaryote species. *Science*, 296, 1061–1063. Available from: https://doi.org/10.1126/science.1070710

Foissner, W. (2006) Biogeography and dispersal of micro-organisms: a review emphasizing protists. *Acta Protozoologica*, 45, 111–136.

Foissner, W. (2007) Dispersal and biogeography of protists: recent advances. *Japanese Journal of Protozoology*, 40, 1–16.

Glücksman, E., Snell, E.A., Berney, C., Chao, E.E., Bass, D. & Cavalier-Smith, T. (2011) The novel marine gliding zooflagellate genus *Mantamonas* (Mantamonadida ord. n.: Apusozoa). *Protist*, 162, 207–221. Available from: https://doi.org/10.1016/j.protis.2010.06.004

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A. & Jermiin, L.S. (2017) ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*, 14, 587–589. Available from: https://doi.org/10.1038/nmeth.4285

Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30, 772–780. Available from: https://doi.org/10.1093/molbev/mst010

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S. et al. (2012) Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28, 1647–1649. Available from: https://doi.org/10.1093/bioinformatics/bts199

Keeling, P.J. & Del Campo, J. (2017) Marine protists are not just big bacteria. *Current Biology*, 27, R541–R549. Available from: https://doi.org/10.1016/j.cub.2017.03.075

Klaveness, D. (1995) *Collodictyon triciliatum* H. J. Carter(1865)—a common but fixation-sensitive *Algivorous flagellate* from the limnopelagial. *Nordic Journal of Freshwater Research Drottningholm*, 70, 3–11.

Laetsch, D.R. & Blaxter, M.L. (2017) BlobTools: interrogation of genome assemblies. *F1000Research*, 6, 1287. Available from: https://doi.org/10.12688/f1000research.12232.1

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10, R25. Available from: https://doi.org/10.1186/gb-2009-10-3-r25

Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. (2013) PhyloBayes MPI: phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*, 62, 611–615. Available from: https://doi.org/10.1093/sysbio/syt022

Lax, G., Eglit, Y., Eme, L., Bertrand, E.M., Roger, A.J. & Simpson, A.G.B. (2018) *Hemimastigophora* is a novel supra-kingdom-level lineage of eukaryotes. *Nature*, 564, 410–414. Available from: https://doi.org/10.1038/s41586-018-0708-8

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. Available from: https://doi.org/10.1093/bioinformatics/btl158

Mangot, J., Logares, R., Sánchez, P., Latorre, F., Seeleuthner, Y., Mondy, S. et al. (2017) Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Scientific Reports*, 7, 1–12. Available from: https://doi.org/10.1038/srep41498

Manni, M., Berkeley, M.R., Seppey, M., Simão, F.A. & Zdobnov, E.M. (2021) BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, 38, 4647–4654. Available from: https://doi.org/10.1093/molbev/msab199

Mikrjukov, K.A. & Mylnikov, A.P. (2001) A study of the fine structure and the mitosis of a lamellicristate amoeba, *Micronuclearia podoventralis* gen. Et sp. nov. (Nucleariidae, Rotosphaerida). *European Journal of Protistology*, 37, 15–24.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. & Minh, B.Q. (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32, 268–274. Available from: https://doi.org/10.1093/molbev/msu300

Orr, R.J.S., Zhao, S., Klaveness, D., Yabuki, A., Ikeda, K., Watanabe, M.M. et al. (2018) Enigmatic Diphyllatea eukaryotes: culturing and targeted PacBio RS amplicon sequencing reveals a higher order taxonomic diversity and global distribution. *BMC Evolutionary Biology*, 18, 115. Available from: https://doi.org/10.1186/s12862-018-1224-z

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S. & Sandberg, R. (2014) Full-length RNA-seq from single cells using smart-seq2. *Nature Protocols*, 9, 171–181. Available from: https://doi.org/10.1038/nprot.2014.006

Piganeau, G., Eyre-Walker, A., Jancek, S., Grimsley, N. & Moreau, H. (2011) How and why DNA barcodes underestimate the diversity of microbial eukaryotes. *PLoS One*, 6, e16342. Available from: https://doi.org/10.1371/journal.pone.0016342

Price, M.N., Dehal, P.S. & Arkin, A.P. (2009) Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26, 1641–1650. Available from: https://doi.org/10.1093/molbev/msp077

Pruitt, K.D., Tatusova, T. & Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids*

Research, 35, D61–D65. Available from: https://doi.org/10.1093/nar/gkl842

Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. (2018) Posterior summarization in Bayesian phylogenetics using tracer 1.7. *Systematic Biology*, 67, 901–904. Available from: https://doi.org/10.1093/sysbio/syy032

Ruggiero, M.A., Gordon, D.P., Orrell, T.M., Bailly, N., Bourgoin, T., Brusca, R.C. et al. (2015) A higher level classification of all living organisms. *PLoS One*, 10, e0119248. Available from: https://doi.org/10.1371/journal.pone.0119248

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31, 3210–3212. Available from: https://doi.org/10.1093/bioinformatics/btv351

UniProt Consortium. (2022) UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51, D523–D531. Available from: https://doi.org/10.1093/nar/gkac1052

Wang, H.C., Minh, B.Q., Susko, E. & Roger, A.J. (2018) Modeling site heterogeneity with posterior mean site frequency profiles accelerates accurate phylogenomic estimation. *Systematic Biology*, 67, 216–235. Available from: https://doi.org/10.1093/sysbio/syx068

Yabuki, A., Ishida, K.-I. & Cavalier-Smith, T. (2013) *Rigifila ramosa* n. gen., n. sp., a filose apusozoan with a distinctive pellicle, is related to *Micronuclearia*. *Protist*, 164, 75–88. Available from: https://doi.org/10.1016/j.protis.2012.04.005

Zhao, S., Burki, F., Bråte, J., Keeling, P.J., Klaveness, D. & Shalchian-Tabrizi, K. (2012) *Collodictyon*—an ancient lineage in the tree of eukaryotes. *Molecular Biology and Evolution*, 29, 1557–1568. Available from: https://doi.org/10.1093/molbev/mss001

SUPPORTING INFORMATION

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