School of Doctoral Studies in Biological Sciences
UNIVERSITY OF SOUTH BOHEMIA, FACULTY OF SCIENCE

Diversity and biogeography of diplonemid and kinetoplastid protists in global marine plankton

Ph.D. Thesis

M.Sc. Olga Flegontova

Supervisor: Aleš Horák, Ph.D.

Biology Centre CAS, Institute of Parasitology

České Budějovice, 2017

This thesis should be cited as:

Flegontova, O, 2017: Diversity and biogeography of diplonemid and kinetoplastid protists in global marine plankton. Ph.D. Thesis. University of South Bohemia, Faculty of Science, School of Doctoral Studies in Biological Sciences, České Budějovice, Czech Republic, 121 pp.

Annotation

The PhD thesis is composed of three published papers, one manuscript submitted for publication and one manuscript in preparation. The main research goal was investigating the diversity of marine planktonic protists using the metabrcoding approach. The worldwide dataset of the *Tara* Oceans project including small subunit ribosomal DNA metabarcodes (the V9 region) was used. The first paper investigated general patterns of protist abundance and diversity in a global set of samples from the photic zone of the ocean. Diplonemids, a sub-group of Euglenozoa, emerged as one of the most diverse lineages. The second paper was a mini-review highlighting this unexpected result on diplonemids. The third paper provided a detailed characteristic of the diversity, abundance and community structure of diplonemids and revealed them as the most species-rich eukaryotic clade in the plankton. The submitted manuscript was focused on the same topics related to planktonic kinetoplastids, the sister-group of diplonemids. The manuscript in preparation will describe the creation of a curated reference database of excavate 18S rDNA sequences, including those of diplonemids and kinetoplastids, that will be indispensable for analyses of environmental high-throughput metabarcoding data. Our studies provide essentially the first global survey for diplonemid and kinetoplastids protists, which were overlooked previously in marine biodiversity and ecology studies.

Declaration [in Czech]

Prohlašuji, že svoji disertační práci jsem vypracovala samostatně pouze s použitím pramenů a literatury uvedených v seznamu citované literatury.

Prohlašuji, že v souladu s § 47b zákona č. 111/1998 Sb. v platném znění souhlasím se zveřejněním své disertační práce, a to v nezkrácené podobě elektronickou cestou ve veřejně přístupné části databáze STAG provozované Jihočeskou univerzitou v Českých Budějovicích na jejích internetových stránkách, a to se zachováním mého autorského práva k odevzdanému textu této kvalifikační práce. Souhlasím dále s tím, aby toutéž elektronickou cestou byly v souladu s uvedeným ustanovením zákona č. 111/1998 Sb. zveřejněny posudky školitele a oponentů práce i záznam o průběhu a výsledku obhajoby kvalifikační práce. Rovněž souhlasím s porovnáním textu mé kvalifikační práce s databází kvalifikačních prací Theses.cz provozovanou Národním registrem vysokoškolských kvalifikačních prací a systémem na odhalování plagiátů.

České Budějovice, 14.08.2017 .....

Olga Flegontova

This thesis originated from a partnership of Faculty of Science, University of South Bohemia, and Institute of Parasitology, Biology Centre of the CAS, supporting doctoral studies in the Molecular and Cell Biology and Genetics study programme.

Přírodovědecká fakulta
Faculty of Science

BIOLOGICKÉ CENTRUM AV ČR, v. v. i.

Financial support

This work was supported by:
ERC CZ, grant LL1205
Internal grant of the Biology Centre of the CAS, grant 320/9564
The Grant Agency of the University of South Bohemia, grant 04−088/2014/P

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Aleš Horák for his patient guidance of my research, for his contagious optimism, kindness, easy communication, and for providing opportunities to present my research at a number of conferences. Without his help, my research and writing of this thesis would not be possible.

Second, I am grateful to my husband, the best friend, colleague and co-author of some of my papers Pavel Flegontov, who motivated me to start my PhD program, supported me extensively throughout my studies, helped me a lot with organizing scientific expeditions and proofread the thesis.

Third, I would like to acknowledge Julius Lukeš who gave me the opportunity to became a protistologist and join his team eight years ago, and whose great interest in diplonemids has driven my project forward.

I would also like to thank all the people from Jula's lab, both former and present lab-members, for their help in my work, for creating friendly atmosphere in the institute, and for all the fun and partying nights we have had together in the last eight years.

In addition, none of the work presented in this thesis would originate without numerous collaborators who helped us to generate data and provided advice, namely, Shruti Malviya, Chris Bowler, Colomban de Vargas, and Kasia Piwosz.

Last but not least, I thank my parents and close friends, who have supported me throughout my life and made my life in Ceske Budejovice happy, easy and unforgettable.

List of papers and author's contribution

The thesis is based on the following papers (listed chronologically):

I. de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury JM, Bittner L, Chaffron S, Dunthorn M, Engelen S, **Flegontova O**, Guidi L, Horák A, Jaillon O, Lima-Mendez G, Lukeš J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S; *Tara* Oceans Coordinators, Acinas SG, Bork P, Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E (2015) Eukaryotic plankton diversity in the sunlit ocean. ***Science***. 348(6237):1261605 (IF = 37.205).

Olga Flegontova participated in data analysis and interpretation.

II. Lukeš J, **Flegontova O**, Horák A. (2015) Diplonemids. ***Current Biology***. 25(16):R702-704 (IF = 8.851).

Olga Flegontova prepared figures and edited the manuscript.

III. **Flegontova O***, Flegontov P*, Malviya S*, Audic S, Wincker P, de Vargas C, Bowler C, Lukeš J, Horák A (2016) Extreme diversity of diplonemid eukaryotes in the ocean. ***Current Biology***. 26(22):3060-3065 (IF = 8.851).

Olga Flegontova participated in study design, analyzed and interpreted the data and contributed to writing of the manuscript.

IV. **Flegontova O**, Flegontov P, Malviya S, Poulain J, de Vargas C, Bowler C, Lukeš J, Horák A (submitted manuscript) Neobodonids are dominant kinetoplastids in the global ocean.

Olga Flegontova participated in study design, analyzed and interpreted the data and contributed to writing of the manuscript.

V. **Flegontova O***, Karnkowska A*, Kolisko M*, Lax G*, Maritz JM*, Panek T*, Carlton JM, Cepicka I, Horak A, Keeling PJ, Lukes J, Simpson AGB, Tai V (manuscript in preparation). Excavata in EukRef, a novel curated database of small subunit rRNA gene sequences.

Olga Flegontova participated in data analysis and interpretation and contributed to writing of the manuscript.

* These authors contributed equally

Co-author agreement

Aleš Horák, the supervisor of this Ph.D. thesis and co-author of all presented papers, fully acknowledges the contribution of Olga Flegontova.

…….
Aleš Horák, Ph.D.

Contents

1 Introduction

1.1 Environmental metabarcoding: a new tool for biodiversity studies

High-throughput sequencing of environmental DNA isolated from whole communities has revolutionized many scientific fields, from microbial ecology to archaeology (Lozupone et al. 2012, Worden et al. 2015, del Campo et al. 2016, Pedersen et al. 2016, Seersholm et al. 2016). Studies of microbes relying on environmental DNA fall into two major types: sequencing whole community genomes or transcriptomes (termed metagenomes and metatranscriptomes) (Bragg & Tyson 2014) or sequencing short highly variable amplicons termed tags or barcodes (Taberlet et al. 2012). The former approach reveals metabolic potential of the community and is especially useful for studies focused on prokaryotes: organisms with relatively uniform morphology and behavior, but with extremely diverse metabolic capacities (Keeling & del Campo 2017). The latter approach reveals organism diversity and community composition at an unprecedented resolution, if a large number of tags is screened against a reference database allowing taxonomic assignments. Sequencing of full-length small-subunit ribosomal RNA genes has further revealed previously unknown major eukaryotic lineages such as marine alveolate groups I and II (Lopez-Garcia et al. 2001, Moon-van der Staay et al. 2001), marine stramenopiles (Lin et al. 2012, Massana et al. 2014), and picozoans (Not et al. 2007, Seenivasan et al. 2013), and has greatly expanded our understanding of diversity within known groups such as prasinophytes (Viprey et al. 2008). Although amplicon-based metabarcoding has become a tool of choice in environmental microbiology, similar tags can be extracted from high-coverage metagenomes or metatranscriptomes, to avoid biases associated with primer specificity and amplification efficiency (Logares et al. 2014a). Below we describe popular metabarcoding approaches applied to microbial eukaryotes (protists), with a special emphasis on biases of the methods.

A genetic marker suitable for barcoding diverse eukaryotic communities must contain both highly conserved and variable regions, otherwise either design of universal primers or accurate diversity assessment would be compromised. Virtually the only gene or transcript used for this purpose is the small-subunit ribosomal RNA (SSU or 18S rRNA), which contains nine hyper-variable regions embedded within conserved sequence (Leray and Knowlton 2016). Pioneering studies on bacteria and archaea used a very short V6 region (Sogin et al. 2006; Huber et al. 2007), but longer and more informative V9 (Amaral-Zettler et

al. 2009) and V4 regions (Stoeck et al. 2010; Pawlowski et al. 2012) were introduced later for eukaryotes. Currently, the variable regions most widely used for eukaryotes are V1-2, V4, and V9 (Leray and Knowlton 2016).

A typical metabarcoding experiment includes the following steps: 1) an optional size fractionation, usually applied to freshwater or marine planktonic samples, helps to reduce community complexity and to gain more detailed information on the distribution of organisms in samples among respective size classes; 2) DNA isolation, or RNA isolation followed by cDNA synthesis; 3) amplification of the SSU rDNA/rRNA region chosen as a tag/barcode; 4) ligation of indexed sequencing adapters and library preparation; 5) sequencing on either the Roche 454 or Illumina MiSeq or HiSeq platforms; 6) quality filtering of the sequencing reads; 7) merging of identical reads into barcodes, also named "ribotypes", with associated abundance values (read counts); 8) removal of the rarest barcodes that likely represent sequencing errors or chimaeras; 9) clustering of similar barcodes into operational taxonomic units (OTUs) using a certain relative or absolute distance threshold; 10) taxonomic assignment using a similarity search vs. a reference database (taxonomic assignment may be performed before or after clustering); 11) analysis of relative abundance, OTU richness, biogeography, or OTU co-occurrence patterns for whole communities or for taxonomic groups of interest. Choices made at some of these steps may profoundly affect study outcome and interpretation.

The first crucial decision to make is the choice of the starting material, DNA or RNA. DNA is stable and easy to isolate from different substrates, therefore it is widely used in biodiversity studies, including the recent global examples (de Vargas et al. 2015; Pernice et al. 2016). However, DNA as a source material has several important drawbacks. First, it is preserved in most environments for a long time (Nielsen et al. 2007), and the signal might reflect the abundance of dead cells. This problem is especially acute in the marine benthic sediments (Danovaro et al. 2005; Dell'Anno and Danovaro 2005; Stoeck et al. 2007), and thus RNA was preferred in a major study focused on European coastal benthic sites (Forster et al. 2016a). In contrast, protist community composition analyzed in planktonic DNA and RNA samples was similar (Logares et al. 2014b; Massana et al. 2015). Dissolved DNA found in the plankton was shown to be derived mostly from the pico- and nano-plankton size fractions, and was attributed at least partially to cell breakage during filtration (Massana et al. 2015).

Second, even if "dead DNA" is not a concern, eukaryotic cells differ by several orders of magnitude in the number of rRNA gene copies they carry (Zhu et al. 2005; Medinger et al.

2010), and this number is correlated not only with cell size, but also with genome size (Prokopowich et al. 2003). The number of ribosomes per cell is much better correlated with cell volume, and thus is more suitable for measuring relative biomass. For example, dinoflagellates and related syndinians, also named marine alveolates (MALV), are mostly small cells up to 20 μm in size, but with an exceptionally high rDNA copy number (Zhu et al. 2005; Medinger et al. 2010; Siano et al. 2010; Massana et al. 2015). These protists are abundant in the marine plankton, but their relative abundance was shown to be inflated even more in DNA-based studies because of this copy-number bias (Massana et al. 2015; Giner et al. 2016; Piredda et al. 2017). Thus, RNA-based metabarcoding (Massana et al. 2015; Giner et al. 2016) or fluorescent in situ hybridization or FISH (Siano et al. 2010; Giner et al. 2016) reflect true cell abundance better than DNA-based metabarcoding.

Another group of biases arises at the amplification step (von Wintzingerode et al. 1997). First, universal primers targeting the V9 or V4 hyper-variable SSU rRNA regions are not truly universal, i.e. much more efficient for some eukaryotic clades as compared to others (Amaral-Zettler et al. 2009; Hong et al. 2009; Edgcomb et al. 2011). For instance, the V4 primers are known to work poorly for excavates and foraminiferans (Pawlowski et al. 2011; Pernice et al. 2016), important protist groups in the marine plankton (de Vargas et al. 2015). None of the widely used barcodes is perfect, as demonstrated by Giner et al. (2016): a cDNA-based analysis of V4 metabarcodes produced relative abundance values closer to those estimated by FISH in five planktonic samples, while V9 performed better in four samples. The results were dependent on community composition. The V4 region, in contrast to V9, is known to be highly variable in length across major eukaryotic clades (Pawlowski et al. 2011). Thus, not only differential primer specificity, but also amplicon length variability might make PCR less efficient for certain clades. This problem was highlighted in a study targeting the V4 region in bathypelagic samples (Pernice et al. 2016): the abundance of excavates (mostly belonging to the diplonemid clade) assessed by metagonome-derived tags ("mitags") was 11%, but only 1% as estimated using V4 tags. The authors attributed this discrepancy to the primer specificity bias and the amplicon length bias. However, the problem was likely exacerbated by an amplicon length cutoff of 600 nt introduced in the study (Pernice et al. 2016). According to our unpublished data, the V4 region usually exceeds this limit in diplonemids, and most diplonemid tags were likely discarded at the read processing step. Such a long region is very difficult to sequence reliably with the existing 454 or Illumina MiSeq technologies: the former can produce reads up to ~800 nt, but is tricky to use and is no longer

supported by the manufacturer, while the longest paired Illumina reads reach 300 nt and cover just ~500 nt amplicons. In summary, although the V4 region allows higher taxonomic resolution and contains more phylogenetic information, the shorter V9 region (Amaral-Zettler et al. 2009) might recover less biased community composition (Pawlowski et al. 2011). Indeed, our V9-based studies have revealed a high relative abundance and staggering diversity of marine pelagic diplonemids (de Vargas et al. 2015; Flegontova et al. 2016). Combination of both V4 and V9 barcodes (Stoeck et al. 2010; Logares et al. 2014b; Giner et al. 2016; Piredda et al. 2017), or the usage of metagenome or metatranscriptome-derived 18S rRNA tags (Logares et al. 2014a; Pernice et al. 2016) represent more laborious and expensive, but more robust alternative approaches.

The Roche 454 high-throughput sequencing method, often referred to as "pyrosequencing", was used for all pioneering metabarcoding studies (for example, Sogin et al. 2006; Huber et al. 2007; Amaral-Zettler et al. 2009; Stoeck et al. 2010; Edgcomb et al. 2011; Pawlowski et al. 2011; Logares et al. 2012a) since it was the first high-throughput method to appear (Margulies et al., 2005) and provided sufficiently long reads (Amaral-Zettler et al. 2009; Logares et al. 2012b), unlike early Illumina versions. Nowadays, the Illumina MiSeq technology approaches 454 in read length, if 300 nt paired-end reads are merged into longer reads of ~500 nt. And the high rate of indel errors in homopolymer tracts makes the 454 technology less suitable for diversity studies (Huse et al. 2007). Thus, at present the Illumina technology is the mainstay of metabarcoding studies (Logares et al. 2014b; de Vargas et al. 2015; Mahe et al. 2017; Piredda et al. 2017), but studies based on the 454 technology are still common (Egge et al. 2015; Massana et al. 2015; Forster et al. 2016a; Pernice et al. 2016). Illumina HiSeq with its 100+100 nt or 150+150 nt reads is a much cheaper alternative for the shorter V9 region (130 nt on average). Illumina HiSeq is also the only viable technology for metatranscriptomic/metagenomic studies that require very large read outputs (Logares et al. 2014a). Both sequencing technologies (Illumina MiSeq and 454) applied to the V4 region produced very much similar relative abundances of 60 taxonomic groups in planktonic picoeukaryotic communities (Giner et al. 2016), demonstrating that technology-specific biases are negligible.

Pipelines of most metabarcoding studies typically involve some sort of metabarcode clustering which considerably reduces the amount of data passed to the taxonomic assignment and further steps, and in some cases, produces diversity patterns closely approximating those of species or genera. Three principal approaches to metabarcode clustering exist (Forster et al.

2016b): 1) the simplest and historically popular approach is centroid-based clustering implemented in USEARCH (Edgar 2010) and similar programs, which first selects an OTU centroid amplicon (usually the most abundant amplicon is chosen first), and then assigns to the OTU all amplicons falling within a certain global similarity threshold, for instance, 97%; 2) sequence similarity networks (Forster et al. 2015) also rely on a global similarity threshold, but OTUs are constructed in an iterative way, joining all amplicons within the threshold distance, then repeating this step until the graph stops growing; 3) another approach implemented in the program Swarm (Mahe et al. 2014, 2015) constructs similar networks, but instead of a relative sequence identity threshold uses an absolute number of mutations (substitutions or indels) as a distance measure.

The centroid clustering approaches, although popular (Massana et al. 2015; Dupont et al. 2016; Forster et al. 2016a; Pernice et al. 2016; Debroas et al. 2017; Piredda et al. 2017), have two major disadvantages. First, global sequence similarity thresholds are arbitrary, one threshold is not suitable for all clades with widely different evolution rates, and there is no agreement on which threshold is the best (Nebel et al. 2011; Brown et al. 2015; Forster et al. 2016b). Various studies used either 95% (Amaral-Zettler et al. 2009; Edgcomb et al. 2011; Logares et al. 2014b; Giner et al. 2016; Debroas et al. 2017; Piredda et al. 2017), 97% (Stoeck et al. 2010; Edgcomb et al. 2011; Logares et al. 2015; Massana et al. 2015; Dupont et al. 2016; Forster et al. 2016a; Giner et al. 2016; Pernice et al. 2016; Piredda et al. 2017), or 99% (Stoeck et al. 2010; Edgcomb et al. 2011; Logares et al. 2014b; Egge et al. 2015; Logares et al. 2015; Massana et al. 2015) similarity thresholds. As seen from this list, some studies tested two and more thresholds. Second, the sequence input order affects OTU composition (Koeppel & Wu 2013; Mahe et al. 2014). We expect that network-based methods approximate true limits of sequence diversity and correspond much better to taxonomic boundaries (Mahe et al. 2015; Schmidt et al. 2015; Forster et al. 2016b). OTU boundaries are not expected to match species boundaries, and also depend on the marker region. V9 or V4 OTUs clustered using Swarm with default settings or using the 97% similarity threshold usually correspond to genera and higher taxonomic ranks, but not to protistan or metazoan species (Massana et al. 2015; Leray and Knowlton 2016). V4 OTUs clustered using the 99% similarity threshold correspond to protistan species much better (Massana et al. 2015).

However, these network-based methods have their own problems. First, sequence similarity networks share the same problem of ad hoc global similarity thresholds. Second, in a high-coverage amplicon dataset sequencing errors or extremely rare OTUs, which are

numerous in any community (Logares et al. 2014b, 2015), would form sprawling graphs, and special approaches are required to define meaningful sub-graphs. Graphs are broken along abundance valleys, where the read count associated with barcodes steadily falls and then rises along a path in the graph (Mahe et al. 2014, 2015). An opposite problem of over-splitting appears, if a path in the graph is missing due to a low-abundance barcode being filtered out from the dataset or totally missing (Mahe et al. 2015). These problems illustrate the ad hoc nature of all algorithms, including the network-based ones. Third, the distance of one substitution or a single nucleotide indel, the preferred distance value in the linkage clustering algorithm employed by Swarm (Mahe et al. 2015), apparently leads to over-splitting of OTUs on datasets of long V4 amplicons (Forster et al. 2016b). Not all functions of the Swarm software are available for larger distance thresholds, and the calculation speed is much slower (Mahe et al. 2015). Despite these problems, the Swarm approach has gained popularity and was used in the *Tara* Oceans project (de Vargas et al. 2015; Lima-Mendez et al. 2015; Flegontova et al. 2016; Malviya et al. 2016; Biard et al. 2017) and in other studies (Filker et al. 2015; Oikonomou et al. 2015; Mahe et al. 2017).

The last crucial data processing step is taxonomic assignment. In principle, two approaches are possible here: 1) performing OTU clustering on a dataset including both reference and newly obtained sequences (Forster et al. 2016b), but in this case too many sequences would remain unassigned; 2) a similarity search for barcodes or OTUs against a reference database, using BLAST, ggsearch, or related approaches. At this step sequences with a similarity lower than 80% (de Vargas et al. 2015) or ~85% (Massana et al. 2015; Pernice et al. 2016) are labeled as unassigned, and others are annotated using the best hit or a group of best hits. The lowest non-contradictory taxonomic rank is usually chosen among those of the hits and assigned to the query. Curated databases of SSU rRNA sequences play a key role in taxonomic assignment since taxonomic annotations in the NCBI database are notoriously uninformative and error-ridden (Guillou et al. 2013). The earliest SSU rRNA databases, including the most widely known SILVA database (Pruesse et al. 2007), focused on prokaryotes only. The SILVA database includes eukaryotic and organellar sequences in later releases, but inherits their inadequate annotation from NCBI (Guillou et al. 2013). To facilitate metabarcoding studies, a dedicated database was created for protists and other eukaryotes – the PR2 database (Guillou et al. 2013). This database became widely used in the microbial ecology community, including large-scale projects such as Malaspina, BioMarKs, and *Tara* Oceans (Massana et al. 2015; Forster et al. 2016a; Pernice et al. 2016; de Vargas et

al. 2015). The PR2 database was constructed from GenBank, EMBL and WGS-EMBL sequences annotated as rRNAs and passing the 799 nt length threshold. Special procedures were employed to remove chimeric sequences, introns, and misannotated prokaryotic and organellar sequences. In total, about 137,000 nuclear-encoded rRNA sequences received annotations composed of eight ranks. As expected, the database is heavily biased towards Opisthokonta (54% of total number of sequences), Alveolata, and Archaeplastida (15 and 12%, respectively). Only 30% sequences are nearly complete. In total, 64% of sequences include the V4 region and only 12% include the V9 region, and these numbers vary a lot clade by clade (Guillou et al. 2013).

1.2 Composition of Euglenozoa and their morphological and molecular traits

In 1981 Cavalier-Smith, recognizing phylogenetic relationships between two traditional protozoan classes kinetoplastids and euglenids, created the kingdom Euglenozoa, one of nine eukaryotic kingdoms in his classification of eukaryotes (Cavalier-Smith 1981). Fundamental cellular traits shared by kinetoplastids and euglenids and recognized by Cavalier-Smith (1981, 1993) included: 1) mitochondrial cristae shaped like a flattened disc with a narrow neck, clearly distinguishing Euglenozoa from the other eight kingdoms; 2) two anterior cilia with dense intraciliary paraxonemal rods and generally simple non-tubular mastigonemes (sometimes one cilium occurs in kinetoplastids or, very rarely, three or four cilia occur in euglenids); 3) three asymmetric microtubular ciliary roots; 4) lack of a cell wall, but presence of a pellicle, a single-layer corset of peripheral evenly spaced microtubules lying under the cell membrane; 5) stacked Golgi cisternae; 6) peroxisomes (in euglenids) or glycosomes (in kinetoplastids); 7) closed mitosis with endonuclear spindle. The grouping of euglenids and kinetoplastids within the phylum Euglenozoa was confirmed by subsequent studies (Corliss 1984; Kivic & Walne 1984; Patterson 1988; Triemer & Ott 1990; Triemer & Farmer 1991), and later the phylum Euglenozoa was expanded by adding diplonemids (Cavalier-Smith 1993) and newly discovered symbiontids (Simpson 1997; Yubuki et al. 2009). Thus, according to the present-day classification of eukaryotes, the phylum Euglenozoa consists of four well-recognized groups: Euglenida, Symbiontida, Kinetoplastea, and Deplonemea (Adl et al. 2012).

Simpson (1997), refining the concept of Euglenozoa, proposed three morphological synapomorphies of the phylum: 1) unique structure of intraciliary paraxonemal rods; 2) presence of three ciliary roots; 3) presence of extrusomes. Apart from euglenids and kinetoplastids, rod-like paraxonemal structures occur in dinoflagellates and some stramenopiles; however, in biflagellated kinetoplastids and euglenids a distinct and common pattern can be perceived: the anterior cilium has paraxonemal rods in the form of a tubular lattice, while the recurrent cilium has paraxonemal rods in the form of a three-dimensional lattice. Previously, most diplonemids were regarded as lacking paraxonemal rods, however those species investigated had very short cilia, while *Hemistasia* (Yabuki & Tame 2015), *Rhynchopus*, and newly described genera *Lacrima*, *Sulcionema*, and *Flectonema* all have paraxonemal rods (Tashyreva et al. submitted).

Three microtubular ciliary roots of euglenids and kinetoplastids nucleate from around the basal bodies. The Dorsal Root (in euglenids), or Fibre dorsal (in kinetoplastids), originates from the anterior cilium and supports the ciliary pocket. The Intermediate Root (in euglenids), or Fibre ventral (in kinetoplastids), originates between the basal bodies and also supports the ciliary pocket. The Ventral Root (in euglenids), or MTR (microtubule-reinforced region in kinetoplastids), originates from the recurrent cilium and loops over to become continuous with the ingestion apparatus (Brugerolle et al. 1979; Solomon et al. 1987). Diplonemids have a similar root pattern, though lack the connection between the ventral root and the ingestion apparatus (Montegut-Felkner & Triemer 1994; Montegut-Felkner & Triemer 1996). The third synapomorphy, tubular thick-walled extrusomes were observed in some but not in all euglenids, kinetoplastids, and diplonemids (Simpson 1997). Principal morphological characteristics which define the phylum Euglenzoa and help to distinguish its four major groups are listed in Table 1.

Relationships between euglenids, kinetoplastids, and diplonemids were not resolved using the 18S rRNA and *cox1* genes. Some authors proposed sisterhood relationships for diplonemids and kinetoplastids (Maslov et al. 1999; Moreira et al. 2004), others for euglenids and diplonemids (Moreira et al. 2001; Von der Heyden et al. 2004) or for kinetoplastids and euglenids (Busse & Preisfeld 2002). Later, a phylogenetic analysis based on the cytosolic isoforms of heat shock proteins (hsp) 90 and 70 supported the close relationship between diplonemids and kinetoplastids to the exclusion of euglenids (Simpson & Roger 2004).

Table 1. General morphological characters of the four subgroups of Euglenozoa.

	Euglenida	**Symbiontida**	**Kinetoplastea**	**Diplonemea**
lifestyle	mostly free-living, aerobic	free-living, anaerobic	free-living or parasitic, aerobic	free-living or parasitic, aerobic
anterior feeding apparatus	MTR/pocket type	MTR/pocket type	MTR/pocket type	MTR/pocket type
plicate mouth	present		lacking	present
anterior cilia	1-2, rarely 3-4	2	1-2	2
properties of the ciliary apparatus:				
ciliary pocket associated with the feeding apparatus	present	present	present	present
two functional kinetosomes	present	present	present	present
three asymmetrically arranged microtubular roots	present	present	present	present
intraciliary heteromorphic paraxonemal rods	present	present	present	present or lacking
non-tubular mastigonemes	present		present	lacking
mitochondrial cristae	discoid	reduced or absent	discoid	plate-like
kinetoplast	lacking		present	lacking
pellicular strips	present	lacking	lacking	lacking
plastids	present or lacking	lacking	lacking	lacking
epicellular bacteria	lacking	present	lacking	lacking
glycosomes	lacking	lacking	present	present
metaboly	present or lacking		lacking	pronounced
tubular extrusomes	present or lacking	present or lacking	present or lacking	present or lacking
apical papilla	lacking	lacking	lacking	present

Now let us turn to molecular traits common to euglenozoans. Kinetoplastids, especially parasitic trypanosomatids, received by far the most attention, and our knowledge of the eugloenozoan molecular biology is heavily biased by studies of this group. If we consider unpublished genomic data on diplonemids and euglenids, only a handful of molecular synapomorphies are characteristic for euglenozoans as a whole (Table 2). First, all euglenozoans have a trans-splicing machinery that attaches a short universal spliced leader sequence to at least a fraction of nuclear transcripts (Frantz et al. 2000; Santana et al. 2001; Sturm et al. 2001; Dykova et al. 2003; Gawryluk et al. 2016). The trans-splicing machinery adds a short (around 30-40 nucleotides), capped spliced leader sequence to the 5' end of the mRNA. The spliced leader is conserved within a given genome, but varies in size and sequence among species. Trans-splicing is catalyzed by the classic spliceosome, just the 5'

end of the intron is located on the spliced leader molecule, and the 3' end within the transcript. Thus, mRNAs for protein-coding genes begin with a short stretch of the spliced leader sequence, followed by the 5' untranslated region and the coding region.

Table 2. Molecular biological characters of the three major subgroups of Euglenozoa.

	Euglenida	**Kinetoplastea**	**Diplonemea**
nuclear spliced-leader RNA	present	present	present
nuclear polycistronic transcription	lacking	present	lacking
nuclear introns	intron-rich	intron-poor	intron-rich
nuclear canonical introns	abundant	rare	abundant or lacking
nuclear non-canonical introns	rare	lacking	abundant or lacking
mitochondrial DNA	linear DNA fragments with one or two genes, high recombination rate	conventional circular genomes (maxicircles) and minicircles encoding antisense guide RNAs	one gene fragment per circle
U-insertion RNA-editing in mitochondria	lacking	present	present
glycolysis	in cytosol	in glycosomes	in glycosomes

As known in other organisms (Lukeš et al. 2009), the emergence of trans-splicing facilitates the origin of polycistronic transcription since this process allows easy cleavage of polycistronic molecules. Diplonemids and euglenids have spliced leader RNAs, but apparently not all transcripts are polycistronic and require trans-splicing (G. Burger, M. Field, unpublished data). But in kinetoplastids all protein-coding transcripts are capped with the spliced leader sequence and all are organized into huge polycistronic units (Clayton 2016). Genes in these units are functionally unrelated, but show a high degree of conservation in gene order across trypanosomatids (El-Sayed et al. 2005). This peculiar genome organization and expression in kinetoplastids has profound implications for the regulation of gene expression. In trypanosomatids, there is only a handful of promoters and transcription factors leading to the general lack of control over transcription initiation (Clayton 2016). In *Trypanosoma brucei*, virtually all nuclear DNA seems to be permanently transcribed. Consequently, control levels in trypanosomatids, and likely in other kinetoplastids (Jackson et

al. 2016), are limited to RNA processing, export, degradation, as well as translation initiation and protein stability (Clayton 2016).

Another unique feature of gene expression in kinetoplastids is the almost complete absence of cis-spliced introns (Simpson et al. 2002; Clayton 2016). Diplonemids and euglenids are different in this respect (Canaday et al. 2001). The genome of *Euglena* is large, about 1 Gbp, and contains both canonical introns and a small proportion of poorly characterized non-canonical introns that lack GT-AG splice boundaries (M. Field, unpublished data). The genome of *Diplonema papillatum* is about 200 Mbp in size, and contains a lot of canonical introns (G. Burger, unpublished data). Other marine diplonemids, as revealed by sequencing ten single-cell genomes (Gawryluk et al. 2016), also have bloated, gene-sparse nuclear genomes, but, in contrast to *Diplonema* and *Euglena*, these genomes have a high density of non-canonical introns. Non-canonical introns lack GT-AG splice boundaries; instead, introns frequently have short (3-6 bp) direct repeats, which are partly exonic and partly intronic, and extensive base pairing interactions might form between the ends of introns. The U1 snRNA (required for binding the 5' splice site in cis-spliced canonical introns) may be used strictly in spliced leader addition in these organisms. The non-canonical introns may represent an active class of mobile elements that are spliced at the RNA level (Gawryluk et al. 2016).

Most euglenozoans have a single branched mitochondrion (Roy et al. 2007), usually with a high DNA content. Structure and expression of mitochondrial genomes tends to be extremely variable across eukaryotes (Burger et al. 2003), and euglenozoans are no exception. In *Euglena*, the mitochondrial genome has reduced gene content, and is represented by a chaotic array of linear DNA fragments. Some fragments contain one or two full-length genes, other fragments contain just short gene pieces, likely non-functional. This genome organization is apparently maintained by an active recombination system (Dobakova et al. 2015). In diplonemids and kinetoplastids, the gene content is conserved, but genome structure and expression are different. The only unifying motif is a complex RNA processing machinery that performs U-insertion/deletion RNA editing in kinetoplastids, or U-insertion RNA editing coupled with trans-splicing in diplonemids. The U-insertion/deletion RNA editing is common to all kinetoplastids, in some cases involves hundreds of Us inserted or deleted at dozens of sites across a transcript, and this process is guided by short antisense RNA molecules serving as templates (so-called guide RNAs) (Kable et al. 1997). In diplonemids, trans-splicing is a more prominent feature of the mitochondrial genetic system.

Every mature transcript is accurately assembled from several independently transcribed modules, with stretches of Us inserted between some of them (Marande et al. 2005; Marande & Burger 2007; Kiethega et al. 2013; Vlcek et al. 2011; Valach et al. 2014; Burger et al. 2016; Moreira et al. 2016; Valach et al. 2016; Yabuki et al. 2016). Deamination RNA editing has also been found in diplonemid mitochondria (Moreira et al. 2016). It was hypothesized that trans-splicing and concomitant U-insertion is guided by antisense RNA molecules similar to kinetoplastid guide RNAs (Flegontov et al. 2011), however no such molecules were found, and the mechanism remains obscure (Valach et al. 2016).

Another molecular synapomorphy of kinetoplastids and diplonemids is the compartmentalization of glycolysis in unique peroxisome-derived organelles called glycosomes (Opperdoes and Borst 1977; Opperdoes et al. 1988; Makiuchi et al. 2011), while glycolysis in euglenids and all other eukaryotes takes place in the cytosol. Cavalier-Smith has even used this synapomorphy as a defining character for a new subphylum Glycomonada, sister to subphylum Euglenoida (Cavalier-Smith 2016).

1.3 Phylogeny of kinetoplastids

The order Kinetoplastida, later renamed into Kinetoplastea (Cavalier-Smith 1993; Adl et al. 2012), was first defined by Honigberg in 1963, and it consisted of uniflagellate obligatory parasites, trypanosomatids, and biflagellate bodonids, both free-living and parasitic. The name of this order was derived from a characteristic structure, termed the kinetoplast. The kinetoplast is a part of the mitochondrion that is firmly associated with the basal bodies of the cilia and contains a large mass of mitochondrial DNA termed kinetoplast DNA, or kDNA (Vickerman 1976).

A phylogenetic analysis based on sequences of 18S rRNA and cytoplasmic hsp90 revealed that: 1) monophyletic trypanosomatids branch within paraphyletic bodonids falling into several clades (Callahan et al. 2002; Simpson et al. 2002); 2) trypanosomatids are most closely related to the bodonid clade comprising *Bodo saltans*, *B*. *edax*, and *B*. cf. *uncinatus* (Callahan et al. 2002; Simpson et al. 2002); 3) *Ichthyobodo* forms the most basal branch among all kinetoplastids (Callahan et al. 2002). This result suggested that trypanosomatids acquired the parasitic lifestyle independently of any parasitic bodonids.

Incorporation into the 18S rRNA phylogeny of sequences from *Ichthyobodo*, *Perkinsela*, and environmental kinetoplastid sequences from deep-sea hydrothermal vents

broke the long basal branch between kinetoplastids and their closest outgroups (diplonemids and euglenids), forming the basis for a revised classification of Kinetoplastea (Moreira et al. 2004). Kinetoplastids were divided into three groups with maximal support: 1) the most basal environmental clade including only sequences from deep-sea hydrothermal vents (Lopez-Garcia et al. 2003); 2) the clade consisting of the *Ichthyobodo* and *Perkinsela* genera, which was named Prokinetoplastina; 3) the clade composed of the other apical kinetoplastid genera, named Metakinetoplastina. Metakinetoplastina in turn was subdivided into four orders (Table 3): 1) Eubodonida (*Bodo saltans*, *Bodo edax*, *Bodo* cf. *uncinatus*); 2) Parabodonida (*Bodo caudatus*, *Parabodo nitrophilus*, *Cryptobia*, *Procryptobia*, *Trypanoplasma*); 3) Neobodonida (*Bodo designis*, *Bodo saliens*, *Cruzella*, *Dimastigella*, *Rhynchobodo*, *Rhynchomonas*); 4) Trypanosomatida (*Trypanosoma*, *Leishmania*, *Crithidia*, *Leptomonas*). The phylogenetic relationships of these orders within Metakinetoplastina were not resolved (Moreira et al. 2004). Given that the genus *Bodo* appeared to be spread among three orders, the generic name *Bodo* was retained for the type species *B. saltans* (Ehrenberg 1830), and *B. caudatus* was renamed into *Parabodo caudatus*, and *B. designis* and *B. saliens* were renamed into *Neobodo designis* and *Neobodo saliens*, respectively. An analogous situation was found for the genus *Cryptobia*, since *C. bullocki*, *C. catostomi*, and *C. salmositica* (marine fish parasites) form a monophyletic group with *Trypanoplasma* (a fish blood parasite), but not with the type species *C. helicis*, an endoparasite of snails. Therefore, the marine cryptobias were reclassified within the genus *Trypanoplasma* (Moreira et al. 2004).

Table 3. Morphological and molecular characters of the five major subgroups of Kinetoplastea.

	Prokinetoplastina	**Parabodonida**	**Neobodonida**	**Eubodonida**	**Trypanosomatida**
lifestyle	parasitic or endosymbiotic	free-living or parasitic	free-living or parasitic	free-living	parasitic
feeding strategy		phagotrophic or osmotrophic	phagotrophic	phagotrophic	osmotrophic
cilia	2	2	2	2	1
anterior cilium		lacks hairs	lacks hairs	non-tubular hairs	absent
posterior cilium		lacks hairs, attached to the body or free	lacks hairs, attached to the body or free	lacks hairs, free from the body	lacks hairs, attached to the body
cytostome		anterolateral, with prominent preoral	apical, associated with conspicuous	anterolateral, surrounded by	if present, close to flagellar pocket,

		ridge	development of preciliary rostrum	lappets, preoral ridge absent	lacks associated oral structures
mitochonrial DNA	polykinetoplastic	pankinetoplastic or eukinetoplastic	polykinetoplastic or eukinetoplastic	eukinetoplastic	eukinetoplastic
maxicircles and minicircles	not concatenated	not concatenated	not concatenated	not concatenated	concatenated into a single network
minicircles		supercoiled or fused into huge megacircle	relaxed	relaxed	relaxed

The phylogeny by Moreira et al. (2004) was confirmed by von der Heyden et al. (2004), who sequenced 18S rRNAs of 34 free-living bodonids. Considerable genetic diversity was found within species defined on morphological grounds, especially within *Neobodo designis*, *Parabodo caudatus*, *Rhynchomonas nasuta* and above all *Bodo saltans*. Neobodonids appeared as the most diverse clade, found in soil, freshwater, and marine water. *Neobodo designis* (*Bodo designis* Skuja 1948) and *Rhynchomonas nasuta* Stokes 1888 were long recognized as extremely common and widespread species, found in tropical and temperate regions, in marine, estuarine, freshwater, and soil habitats (Larsen & Patterson 1990; Patterson & Lee).

Below we describe major clades of kinetoplastids, their ecology and some molecular traits (Table 3). Trypanosomatida is by far the most well-studied clade since it includes important human pathogens *Trypanosoma* and *Leishmania*, and the second largest clade after neobodonids according to the number of 18S rRNA sequences in GenBank. Trypanosomatids contain the highest number of formally described species (Maslov et al. 2013) and include *Trypanosoma brucei*, a model species for molecular biology studies. Trypanosomatids are obligatory endoparasites of terrestrial insects (usually hemipterans and dipterans), and many of them switched to dixenous (two-host) life cycles in insects and vertebrates (Lukeš et al. 2014, Maslov et al. 2013). Some members of the genus *Trypanosoma* circulate between freshwater fish and leeches (Lukeš et al. 2014, Maslov et al. 2013).

All trypanosomatids have a unique type of kDNA called eukinetoplastic: two distinct classes of relaxed (not supercoiled as in typical mitochondria and bacteria) mitochondrial chromosomes, maxicircles and minicircles, are catenated and packed into a single dense

network (Lukeš et al. 2002). Maxicircles exist in several dozens of copies per mitochondrion, have uniform sequence, and contain typical mitochondrial genes. Minicircles exist in thousands of copies per mitochondrion (Shapiro & Englund 1995) and encode one to four guide RNAs per circle, depending on the taxon (Simpson Larry 1997). In contrast to kDNA of trypanosomatids with its uniform morphology, minicircles and maxicircles of bodonids are present in a plethora of forms, however none of which is a network (Lukeš et al. 2002). Like trypanosomatids, many bodonids have eukinetoplastic mitochondrial DNA: kDNA is focused in a single dense disk, but maxi- and minicircles are not concatenated into a network (Blom et al. 1998). In others bodonids, kDNA is distributed throughout the mitochondrion, either as one diffuse entity (pankinetoplastic) or as distinct nodules (polykinetoplastic) (Table 3). The network of trypanosomatids has likely developed to ensure faithful replication (Klingbeil 2001) and protect from loss of guide RNA classes essential for RNA editing. In bodonids, which lack this sophisticated replication mechanism (Klingbeil 2001), huge kDNA associates have developed to avoid accidental minicircle losses (Lukeš et al. 2002).

Eubodonids include only one described morphospecies, *Bodo saltans*, a genetically diverse entity suggested for splitting into multiple species (Callahan et al. 2002; Moreira et al. 2004; von der Heyden 2004). This clade is most closely related to trypanosomatids (Jackson et al. 2016). *B. saltans* has eukinetoplastic mitochondrial DNA with relaxed minicircles, each encoding two guide RNAs (Blom et al. 1998). Eubodonids are free-living bacteriovorous protists found in soil, in freshwater and marine habitats (von der Heyden 2004), and *Bodo saltans* appeared among 20 most commonly seen zooflagellates (Patterson & Lee 2000).

Neobodonida is the largest group of kinetoplastids represented in GenBank. An overwhelming majority of neobodonids are free-living marine flagellates, benthic or pelagic, using bacteria as food (Lukeš et al. 2014). Only *Azumiobodo* is parsitic (Hirose et al. 2012), and *Dimastigella trypaniformis* is a commensal of the intestine of a termite (Stolba et al. 2001). Mitochondrial DNAs of *Bodo designis*, *B. saliens*, and *Rhynchomonas* (Swale 1973) are eukinetoplastic, while those of *Cruzella* (Zíková et al. 2003), *Dimastigella* (Breunig et al. 1993; Stolba et al. 2001) and *Rhynchobodo* (Brugerolle 1985) are polykinetoplastic. Since we revised the neobodonid taxonomy in our studies, it is described in detail in the section "Summary of results and discussion".

The smallest bodonid clade is Parabodonida, comprising *Parabodo* spp., *Procryptobia sorokini*, *Cryptobia helicis*, a parasite of snails, and *Trypanoplasma* spp. living in fish blood (Moreira et al. 2004, von der Heyden et al. 2004). *Parabodo caudatus*, *Cryptobia*, and

Trypanoplasma have pankinetoplastidic mitochondrial DNA (Lukeš et al. 2002). While minicircles of *P. caudatus* and *Cryptobia* are monomeric and supercoiled (Hajduk et al. 1986; Lukeš et al. 1998), minicircles of *Trypanoplasma* are fused into a 200kb-long megacircle (Maslov & Simpson 1994).

The Prokinetoplastina clade includes two genera, *Ichthyobodo* and *Perkinsela*. *Ichthyobodo* is an ectoparasite that is typically considered infecting the skin and gills of freshwater fish, but has also been observed as a pathogen of marine fish, amphibians and invertebrates (Callahan et al. 2002; Todal et al. 2004; Isaksen et al. 2012). *Perkinsela* is an obligatory endosymbiont of three genera of amoebae (*Paramoeba*, *Neoparamoeba*, and *Janickina*), which were isolated from water and sand as well as from vertebrate and invertebrate hosts, but most of them are ectoparasites of marine fishes (Dykova et al. 2008). The type species *Perkinsiella amoebae* (subsequently *Perkinsiella* was renamed to *Perkinsela* since the genus name had already been taken by an insect, Dykova et al. 2008) was described as a relative of kinetoplastid flagellates, which lacks cilia, kinetosomes, endoplasmic reticulum and Golgi apparatus, but possesses usually two nuclei and a single giant mitochondrion containing a huge amount of DNA, even more abundant than its nuclear DNA (Hollande 1980). Another unusual feature of *Perkinsela* is the extremely reduced number of subpellicular microtubules in an incomplete microtubular corset (Dykova et al. 2003; Dykova et al. 2008) and its chaotic mitochondrial genome with a reduced gene content, composed of linear fragments similar to *Euglena* (David et al. 2015).

Few environmental sequences branch as a sister-clade to kinetoplastids (Moreira et al. 2004), but it remains unknown whether these organisms share synapomorphies of kinetoplastids. These sequences were detected for the first time at a marine hydrothermal vent chimney (Lopez-Garcia et al. 2003). Later representatives of this clade were found in three other studies of extreme marine biomes: deep-sea sediments and plankton down to 5,189 m (Lecroq et al. 2009; Scheckenbach et al. 2010; Salani et al. 2012).

1.4 Phylogeny of diplonemids

Before 2009 Diplonemea, in contrast to their sister-group Kinetoplastea, was a small, poorly studied clade with only two described genera (*Diplonema* and *Rhynchopus*), which are predatory, parasitic, or commensalic marine protists (Adl et al. 2012). During the last eight years the picture of diplonemid diversity has changed substantially. First, the genus

Hemistasia was re-described as a diplonemid (Yabuki & Tame 2015). Second, two novel diplonemid groups were discovered through environmental sequencing and named deep-sea pelagic diplonemids I and II, or DSPD I and DSPD II (Lara et al. 2009).

Historically, the family Diplonemidae (Cavalier-Smith 2016) included three genera: *Diplonema*, *Rhynchopus*, and *Isonema*. The type species of Diplonemea, *Diplonema breviciliata*, was described in 1913 by Griessmann, and placed into the euglenid family Astasiidae as a colorless marine flagellate. *D. breviciliata* had an elongate pear-shaped body, 14-23 μm in length, with two short non-motile cilia that arise from a pronounced anterior ciliary pocket. It was usually found gliding on surfaces (Griessmann 1913). The second species, *Rhynchopus amitus*, was described from Baltic Sea plankton as an unusual highly metabolic colorless euglenid of the family Rhynchopodaceae (Skuja 1948). *R. amitus* had also two short cilia, and an anterior papilla separating the ingestion apparatus from the ciliary pocket. The third species, *Isonema nigricans*, was described from a polluted marine habitat, and its electron-microscopical observations again suggested a taxonomic position near euglenids (Schuster et al. 1968). The fourth species, *I. papillatum* described by Porter was found associated with eelgrass: the protists enter plant cells and scavenge degenerating cytoplasm (Porter 1973). Later, two new species were described: *Diplonema ambulator* causing the *Cryptocoryne* plant disease and *D. metabolicum* feeding on leaves of the *Halophila* seagrass (Larsen & Patterson 1990). Triemer and Ott realized that *Diplonema* and *Isonema* should be merged into one genus (Triemer & Ott 1990).

Later, a detailed morphological description was provided for four new *Rhynchopus* species: *R. coscinodiscivorus* (Schnepf 1994), *R. euleeides* (Roy et al. 2007), *R. serpens*, and *R. humris* (Tashyreva et al. submitted). In contrast to *Diplonema*, *Rhynchopus* species produce two distinct cell types: big trophic cells with very short cilia concealed in the flagellar pocket, which appear in nutrient-rich media and move by gliding, and smaller cells equipped with two long cilia that appear during starvation (Roy et al. 2007; Tashyreva et al. submitted). While *Diplonema* has extensive flat cristae, *Rhynchopus* has “abnormal” mitochondria almost devoid of cristae. *R. coscinodiscivorus* was found feeding on the cytoplasm of the planktonic diatom *Coscinodiscus* (Schnepf 1994). A *Rhynchopus* species very closely related to *R. humris* (Tashyreva et al. submitted) was reported to infect the *Nephrops norvegicus* lobster (von der Heyden et al. 2004), and several other isolates were found to parasitize crabs, lobsters and clams (Kent et al. 1987; von der Heyden et al. 2004). It seems that parasitism, at

least transient, is a common life strategy for this genus. In this context, the trophic cell type is interpreted as the parasitic stage, and the flagellated cell type as the infective stage.

Recently, the family Diplonemidae was extended by two genera isolated from surface marine plankton: *Lacrima* with closely related diverse environmental sequences and *Sulcionema*, the most basal species-poor branch of the family (Tashyreva et al. submitted). *Diplonema* sp. ATCC50224 was re-described as *Flectonema* as it formed a separate branch in the 18S rRNA tree (Tashyreva et al. submitted).

The family Hemistasiidae (Cavalier-Smith 2016) contains two species, *Hemistasia phaeocysticola* and *H. amylophagus.* The taxonomic position of *Hemistasia* until recently had remained unclear. *Hemistasia* was initially described as a dinoflagellate under the name *Oxyrrhis phaeocysticola* (Scherffel 1900). Griessmann (1913), giving a description as a highly metabolic cell with a spiral groove and two flagella, considered *Hemistasia* as a euglenid relative. Later *Hemistasia* was tentatively affiliated with diplonemids due to its prominent apical papillum and giant flat mitochondrial cristae (Patterson 1994), while others, detecting a polykinetoplast and the appearance resembling *Rhynchobodo*, placed *Hemistasia* into kinetoplastids (Elbrachter et al. 1996). Yabuki and Tame (2015) have re-described this species and assigned it to diplonemids according to their 18S rRNA tree. *Hemistasia* differs from Diplonemidae in having: 1) prominent tubular extrusomes; 2) mitochondria with giant flat cristae; 3) kinetoplast-like material visible by electron microscopy; 4) a large food vacuole; 5) smooth cortical alveoli (Yabuki & Tame 2015; Cavalier-Smith 2016). *Hemistasia* preys on diatoms, dinoflagellates, haptophytes, copepods, etc. (Elbrachter et al. 1996).

Two environmental clades, DSPD I and DSPD II (deep sea pelagic diplonemids), were established by Lara and others in 2009, when roughly a hundred unique diplonemid 18S rRNA sequences were amplified from diverse planktonic samples (from 5 to 3000 m depth). The first 18S rRNA sequence of the DSPD I clade was retrieved from a planktonic 0.2-5 μm size fraction taken at the depth of 3000 m in the Drake passage (Lopez-Garcia et al. 2001). Subsequent studies described in the next section expanded the set of DSPD environmental sequences, and the formal name Eupelagonemidae was proposed for the DSPD I clade (Okamoto et al., submitted). Hemistasiidae forms a clade with Eupelagonemidae, and Diplonemidae forms a deeper branch in an SSU rRNA tree (Yabuki and Tame 2015). However, sparse sequence sampling and low bootstrap support values make that result unreliable.

Any cultured strains are lacking for Eupelagonemidae, but ten cells ranging from 10 to 30 μm in size were isolated from depths of 50-160 m, photographed under a light microscope and subjected to single-cell genome amplification (Gawryluk et al. 2016). Remarkably, those 10 cells represented 25% of all heterotrophic flagellates identified by sequencing, consistent with the idea that diplonemids are abundant in marine plankton. All the cells studied have very large genomes (>100 Mbp), and thus parasitic lifestyle is unlikely for them. Microscopic and genomic evidence suggest that at least some eupelagonemids prey upon eukaryotes: prasinophytes and haptophytes (Gawryluk et al. 2016).

1.5 Environmental studies of kinetoplastids and diplonemids

In early environmental studies, relying on low-throughput cloning and Sanger sequencing of 18S rDNA, the following novel clades of kinetoplastids and diplonemids were discovered: eupelagonemids (Lopez-Garcia et al. 2001, 2007), basal kinetoplastids KIN1 and prokinetoplastids PRO1 (Lopez-Garcia et al. 2003), neobodonids NEO1, NEO2, and NEO3 (von der Heyden & Cavalier-Smith 2005). The clade nomenclature used here follows reference trees constructed by us and shown in Figures 1 and 2 in section 3.5, *Manuscript II*.

Von der Heyden and Cavalier-Smith (2005) amplified 39 18S rRNA sequences from diverse environmental samples (marine, freshwater and soil) using kinetoplastid-specific primers. An overwhelming majority of sequences belonged to Neobodonida, some to Eubodonida, only four to Prokinetoplastina, and none to Parabodonida and Trypanosomatida. It was shown that strains of the *Neobodo designis* morphospecies fall into exclusively marine and freshwater lineages (von der Heyden & Cavalier-Smith 2005).

A series of clone-based studies was focused on hydrothermal vents and other deep-sea environments. Lopez-Garcia et al. (2003) obtained 37 18S rRNA sequences from a hydrothermal vent chimney at the Mid-Atlantic Ridge: from hydrothermal sediment; from microcolonizers exposed to a hydrothermal fluid source, and from hydrothermal fluid/seawater mixtures. Seven of 37 sequences belonged to kinetoplastids: two sequences from sediments to the clade KIN1; one sequence from sediments to PRO1; one sequence from microcolonizers to Parabodonida (*Procryptobia*); and three sequences from microcolonizers to Neobodonida (*Bodo saliens*, *Cruzella*, and NEO2) (Lopez-Garcia et al. 2003). Kinetoplastid diversity in this small dataset was outnumbered only by alveolates and metazoans, and no sequences of diplonemids and euglenids were detected. Absence of

kinetoplastid sequences from seawater suggest that these kinetoplastids from the hydrothermal environment are rather benthic than planktonic.

A later more extensive study of the Lost City hydrothermal vent field in the Atlantic has also found kinetoplastids on carbonates and eupelagonemids in the plankton only (Lopez-Garcia et al. 2007). After the alveolates and fungi, Euglenozoa was the most represented phylum detected in that study. The eupelagonemid clade was recognized, but not named. A clone-based study of a much larger scale reported 923 protistan 18S rRNA sequences from the photic and bathypelagic zones (depth of 2500 m) in the western North Atlantic (Countway et al. 2007). Euglenozoans emerged as one of the most abundant protist groups in deep plankton, but not in the photic zone. Kinetoplastids and diplonemids were not considered separately in that study, but our re-analysis of the sequences revealed that all belong to Eupelagonemidae.

As discussed above, the first targeted study of eupelagonemids was performed by Lara et al. in 2009. Although no assessment of their relative abundance was performed in that study, 95 sequences ~1200 nt in length were obtained for eupelagonemids, extending their known diversity tenfold. The authors looked for the presence of diplonemids in planktonic samples from the Marmara Sea, the Ionian Sea, the South and North Atlantic, and in a sample from the East Pacific Rise. Some freshwater samples were also tested. Diplonemid sequences were amplified from all deep-sea samples (depth 500-3000 m), but no amplicons were retrieved from 6 of 9 surface samples (depth 5-100 m) and from freshwater samples. Diplonemids had pan-oceanic distribution with occurrence of identical phylotypes in geographically distant environments and in very different water masses. Diplonemids showed a marked stratified distribution through the water column, being very scarce or absent in surface waters. There was a high diversity of diplonemid phylotypes within a single sample.

Extending the qualitative results above, substantial relative abundance of eupelagonemids (up to 25% of all protists) was revealed in several mesopelagic and bathypelagic samples from the South Pacific (in a 3-10 μm size fraction) (Sauvadet et al. 2010). In total, 377 clones were sequenced for this size fraction. Kinetoplastids (neobodonids) were under-represented in the libraries due to a primer specificity bias (Sauvadet et al. 2010). Another large-scale clone-based study of abyssal sediment-overlaying water in the South Atlantic (Scheckenbach et al. 2010) supported the notion that kinetoplastids and especially diplonemids are abundant in this environment. 763 clones were obtained from the depths of 5000 - 5600 m. Few eupelagonemid sequences were by far the most abundant diplonemids,

and diverse less abundant sequences of Diplonemidae were also obtained. Among kinetoplastids, *Rhynchomonas* was by far the most abundant. A study of the hadopelagic zone at the depth of 6000 m found that eupelagonemids are relatively abundant (~5% of 339 eukaryotic clones) even in that extreme environment (Eloe et al. 2011).

A follow-up study on abyssal plains conducted by Salani et al. (2012) targeted specifically kinetoplastid 18S rDNA sequences and retrieved 1364 clones clustered into 317 OTUs at the 99% percent identity threshold and 177 OTUs at the 97% threshold. Rank-abundance curves showed a result typical for microbial communities, where just few OTUs account for >50% of reads, and singleton OTUs are numerous. The most represented taxa were *Rhynchomonas*, *Ichthyobodo*, and *Neobodo*. No members of Eubodonida, Parabodonida, and Tryposomatida were retrieved. Reanalysis of this dataset using our EukRef tree (see Figure 1 in section 3.5, *Manuscript II*) has yielded different results due to a much narrower definition of *Neobodo* used by us. Here percentages of clones deposited in GenBank by Salani et al. (2012) are listed: *Rhynchomonas*, 47%; unclassified Neobodonida, 17%; *Rhynchobodo*, 8%; *Dimastigella*, 7%; *Klosteria*, 4%; unclassified Prokinetoplastina, 14%; the most basal environmental clade KIN1, 1%; NEO1, NEO3, and *Cruzella*, 1% each. Parabodonids, eubodonids, trypanosomatids, *Neobodo*, NEO2, and PRO1 were either not detected or had a negligible percentage.

FISH with a specific probe demonstrated that kinetoplastids are abundant at depths down to 5000 m in the North Atlantic: while absolute abundance of kinetoplastids and eukaryotes decreased with depth, relative abundance of kinetoplastids increased up to 27% in the deepest pelagic layer (Morgan-Smith et al. 2011). A larger set of specific FISH probes used in a follow-up study (Morgan-Smith et al. 2013) was used to measure relative abundance of kinetoplastids, diplonemids, fungi, MALV II, labyrinthulomycetes, and marine stramenopiles (MAST) 4. Kinetoplastids were detected throughout the water column, accounting for, on average, 7-12% of FISH-labeled eukaryotes in each water mass. Surprisingly, diplonemids were less abundant, comprising 1-3% of eukaryotes. It was suggested that the abundance of MALV is greatly overestimated and the abundance of kinetoplastids greatly underestimated in metabarcoding studies (Morgan-Smith et al. 2013).

Kinetoplastids and diplonemids had been neglected in many metabarcoding studies (for example, Massana et al. 2015; Pernice et al. 2016; Mahe et al. 2017), either due to primer specificity and amplicon-length biases characteristic for the V4 barcode (see the first section), or due to low taxonomic resolution of early studies of the V9 region, lumping all excavates

together. The conclusion that kinetoplastids are overlooked by universal primers targeting eukaryotic SSU rRNAs was made by Mukherjee et al. (2015). In that clone-based study, kinetoplastids, counted with a FISH probe, contributed up to 12 and 36% of total eukaryotes in the epilimnion and hypolimnion of a freshwater lake, respectively. These freshwater kinetoplastids belonged to the *Rhynchomonas* and NEO1 sub-clades of neobodonids (see Figure 1 in section 3.5, *Manuscript II*). High abundance of diplonemids in the bathypelagic zone (11% of eukaryotes) was demonstrated by Pernice et al. (2016) using the metagenomic "mitag" approach (Logares et al. 2014a). Except for this study, pioneering studies exploring diplonemid and kinetoplastid abundance and diversity with the metabarcoding approach were performed with our participation and are described in the next section: a global survey of the photic zone relying on the V9 barcode (de Vargas et al. 2015) and our follow-up studies focused on diplonemids (Flegontova et al. 2016) and kinetoplastids (Flegontova et al. submitted).

In summary, previous environmental studies, mostly based on clone libraries and not on high-throughput metabarcoding approaches, suggest that eupelagonemids might represent around 10% of eukaryotes in deep-sea plankton, an extremely cell-poor but huge environment by volume. The relative abundance of kinetoplastids in the pelagic environment is difficult to estimate since FISH-based studies and clone-based or metabarcoding studies provide widely different estimates of abundance, which is attributed to primer specificity biases. In any case, *Rhynchomonas* and other neobodonids are by far the most abundant kinetoplastids in the ocean, with unclassified Prokinetoplastina being less abundant, but detectable. The other clades have negligible abundance in the ocean.

Research objectives

- Using the methods of high-throughput metabarcoding, analyze distribution of diplonemid and kinetoplastid protists across marine planktonic size fractions and depth zones.

- Perform a global survey of biogeography for diplonemids and kinetoplastids.

- Analyze co-occurrence of diplonemid and kinetoplastid operational taxonomic units with other organisms, in order to gain insights into their lifestyle and trophic strategies.

- Build a curated reference database of diplonemid and kinetoplastid small subunit rRNA gene sequences, to be used in metabarcoding studies.

3 Summary of results and discussion

Here we summarize results of our key studies, but first we present a revised taxonomy of kinetoplastids and diplonemids that serves as a framework for the metabarcoding studies (Flegontova et al. manuscript in preparation). Currently, a new reference database of SSU rRNA sequences is being developed, aiming to take care of the biases in the PR^2 database (Guillou et al. 2013) and to surpass it in scope (http://eukref.org/; Berney et al. 2017). This database is named EukRef and forms a part of UniEuk, an even larger database aiming to build a phylogenetic framework integrating both metabarcoding datasets (EukBank) and longer SSU sequences. The first metabarcode to be integrated into UniEuk is the V4 SSU rRNA region (Berney et al. 2017). In the EukRef project we were responsible for improving taxonomic annotations of kinetoplastid and diplonemid sequences. SSU rRNA sequences longer than 500 nt were extracted from the GenBank database by an iterative BLAST search on a clade-by-clade basis. A maximum likelihood phylogenetic tree constructed using these sequences reflects the established taxonomy and has allowed us to correct annotation of many deposited sequences and to define several new environmental clades with high support, to be published in an upcoming paper.

Our final tree consisted of 2,339 sequences forming the Kinetoplastea clade with bootstrap support of 94% and 433 sequences falling into the Diplonemea clade with bootstrap support of 100% (see Figures 1 and 2 in section 3.5, *Manuscript II*). While 92% sequences from the former clade were correctly annotated as kinetoplastids, 65% sequences from the latter clade were poorly annotated as uncultured eukaryotes, uncultured marine eukaryotes, or uncultured euglenozoans. Our tree supported the division of kinetoplastids into three major groups (Figure 1 in section 3.5): Metakinetopastina, Prokinetoplastina, and the basal environmental clade (Moreira et al. 2004; von der Heyden et al. 2004). The basal environmental clade, containing sequences from extreme marine biomes, such as a hydrothermal vent chimney, deep-sea sediments and deep-sea plankton (Lopez-Garcia et al. 2003; Lecroq et al. 2009; Scheckenbach et al. 2010; Salani et al. 2012), we named KIN1. The Prokinetoplastina clade includes three well-supported sub-clades (Figure 1 in section 3.5): *Ichthyobodo*, *Perkinsela*, and an environmental-only clade. The environmental-only Prokinetoplastina clade, named by us PRO1, includes few sequences from diverse marine and

freshwater biomes (Lopez-Garcia et al. 2003; von der Heyden et al. 2005; Sauvadet et al. 2010; Scheckenbach et al. 2010; Salani et al. 2012).

All four traditionally recognized Metakinetoplastina sub-clades: Neobodonida, Parabodonida, Eubodonida, and Trypanosomatida (Moreira et al. 2004) have been recovered in our tree, albeit with low bootstrap supports (Figure 1 in section 3.5). Neobodonids represent by far the most diverse and abundant kinetoplastid group in the ocean (see section 1.5) and in our EukRef database. We mostly supported the phylogenetic results by von der Heyden and Cavalier-Smith (2005) who separated neobodonids into nine sub-clades. Six of them had a genus annotation (*Cruzella*, *Dimastigella*, *Klosteria*, *Neobodo*, *Rhynchobodo*, *Rhynchomonas*), and three sub-clades lacking formal description were named by us NEO1, NEO2, and NEO3. Another neobodonid sub-clade is *Azumiobodo*, a parasite of ascidians and the only parasitic neobodonid described (Hirose et al. 2012). In summary, we suggested re-annotating 92% neobodonid sequences with incomplete or incorrect taxonomy in the GenBank database.

Parabodonida and Eubodonida, as compared to Neobodonida, are small groups in both the GenBank database and in marine environmental surveys. While four parabodonid genus-level clades could be distinguished (free-living *Parabodo* and *Procryptobia*, and parasitic *Cryptobia* and *Trypanoplasma*), various eubodonids were often described as one species, *Bodo saltans*. However, *B. saltans* is a cluster of genetically diverse organisms with uniform morphology, and it was suggested for splitting into multiple species (Moreira et al. 2004; Heyden & Cavalier-Smith 2005). Trypanosomatida was the second largest kinetoplastid subgroup in our dataset. A majority of them are coming from cultures and are annotated to the genus and species level, thus no correction of their annotation was needed.

Subgroups of diplonemids defined previously were recovered with high bootstrap support in our tree: Diplonemidae, Hemistasiidae, Eupelagonemidae (Okamoto et al. submitted) formerly known as deep-sea pelagic diplonemids or DSPD I (Lara et al. 2009), and DSPD II (Figure 2 in section 3.5). This tree was the first one including all four clades and a broad spectrum of SSU rRNA diversity, but phylogenetic relationships of the four diplonemid groups remained unresolved. Diplonemidae emerges as the earliest branching clade in another 18S rRNA tree (Tashyreva et al. submitted).

Below we briefly present our global metabarcoding studies targeting marine planktonic diplonemids and kinetoplastids. Initially we took part in a large-scale study investigating eukaryotic diversity in hundreds of size-fractionated planktonic samples collected during the

Tara Oceans expedition in 2009-2012 (de Vargas et al. 2015). The V9 region of SSU rDNA was used as the principal barcode. Only samples taken in the surface and deep chlorophyll maximum zones were analyzed. At most sampling locations, the following size fractions were obtained: picoplankton (0.8-5 μm), nanoplankton (5-20 μm), microplankton (20-180 μm), and mesoplankton (180-2000 μm). Total eukaryotic diversity encompassed 150,000 operational taxonomic units generated with a linkage clustering approach (Mahe et al. 2014, 2015), and was saturated at the global level. About 30% of OTUs were not assigned to any existing phylum, and novel diversity was found within most established protist clades. Heterotrophic protist groups were found to be more abundant and diverse than photosynthetic ones. Diplonemids unexpectedly emerged as a diverse and abundant eukaryotic group on a par with such well-known major clades as collodarians, MALV, and ciliates. Diplonemids comprised 12,000 OTUs. Remarkably, diplonemid diversity was not saturated, unlike that of other large clades. In contrast to diplonemids, just about 150 kinetoplastid OTUs were found in the global dataset.

In a later mini-review (Lukeš et al. 2015) we highlighted the findings reported in de Vargas et al. (2015) and speculated on the possible lifestyle of diplonemids. We stressed that diplonemids in the photic zone represent the third most OTU-rich group after dinozoans and metazoans, and the sixths most abundant group after metazoans, rhizarians, dinozoans, diatoms, and other stramenopiles. Although abundance estimates derived from metabarcoding experiments are affected by various biases (primer specificity biases, highly variable copy number of rRNA genes per cell), this result is striking in any case.

Next, we performed a detailed investigation of the *Tara* Oceans V9 metabarcoding dataset focused on diplonemids (Flegontova et al. 2016). Based on previous studies (see section 1.5), we suspected that diplonemids, similar to other heterotrophic groups, prefer the deeper non-photic zone of the ocean that was not investigated by de Vargas et al. (2015). Although absolute cell counts drop with depth (Morgan-Smith et al. 2013), we expected to find a high relative abundance of diplonemids in the deep. We have extended the original metabarcoding dataset with 516 samples, including 61 coming from the mesopelagic zone (200-1000 m). As we expected, the inclusion of the mesopelagic zone, where diplonemids comprise 14% of eukaryotic reads on average, has affected the diversity estimates drastically. In contrast to 12,000 diplonemid OTUs found in the photic zone (de Vargas et al. 2015), we found about 45,000 OTUs in the extended dataset, which made diplonemids the most OTU-rich eukaryotic group in the plankton, surpassing even metazoans and dinozoans.

About 36% of diplonemid OTUs were confined to the mesopelagic zone and a great majority of OTUs was extremely rare: 100 OTUs represented 93% of diplonemid reads. This type of rank abundance curve is typical for some, but not all, protist groups (dinoflagellates, diatoms, pelagophytes, see Keeling and del Campo 2017). In the extended dataset, the diversity of diplonemids reached saturation on the global scale. No biogeographic patterns were revealed for the whole clade, and no striking correlations with the abundance of other organisms were found in a global interactome dataset (Lima-Mendez et al. 2015). Thus, the feeding strategy of diplonemids remains elusive. Several among 100 most abundant OTUs occurred predominantly in the mesoplankton, which suggested they were parasites of larger organisms. However, a great majority of diplonemids occurred mostly in the pico- and nanoplankton fractions.

Finally, we used similar approaches to investigate diversity and biogeography of kinetoplastids (Flegontova et al. submitted manuscript), using the same extended *Tara* Oceans metabarcoding dataset as above (Flegontova et al. 2016). In general, kinetoplastids followed the same patterns as diplonemids: they were much more abundant and diverse in the mesopelagic zone, demonstrated no biogeography on the level of the whole clade, and just 14 OTUs accounted for 94% of reads. However, kinetoplastids were 5-10 times less abundant throughout the water column (0.2% per sample on average) as compared to diplonemids, and were ~100 times less diverse, with just 512 OTUs found. According to the size fractionation data, a majority of kinetoplastids are smaller than 5 μm, cf. 20 μm for diplonemids. An overwhelming majority of marine kinetoplastids belongs to free-living neobodonids (70% of OTUs and 98% of reads). In contrast to previous studies based on clone libraries (see section 1.5), the predominant genus was *Neobodo*, and not *Rhynchomonas*. Scrutinizing size fraction distribution for the 14 most abundant OTUs, we found three putatively parasitic OTUs; and for two of them putative hosts were revealed through a simple OTU co-occurrence analysis. The putative hosts are a planktonic appendicularian and a copepod. Before our study, just one parasitic genus *Azumiobodo* was described among neobodonids. *Azumiobodo* spp. infect benthic tunicates (ascidians), while the novel parasitic neobodonid parasitizes planktonic tunicates, appendicularians.

To conclude, our metabarcoding studies have changed the picture of diplonemid diversity drastically (de Vargas et al. 2015; Flegontova et al. 2016) and provided a deeper insight into kinetoplastid diversity in marine environments (Flegontova et al. submitted manuscript). It became clear that the OTU diversity of the DSPD I (Eupelagonemidae) clade is a hundred times higher than the diversity of Diplonemidae, Hemistasiidae, and DSPD II

(Flegontova et al. 2016). Most surprisingly, it exceeds not only the diversity of kinetoplastids and all other excavates, but also the diversity of any other eukaryotic supergroup in the marine plankton worldwide: Alveolata, Rhizaria, Stramenopiles, Opisthokonta, Amoebozoa, and Archaeplastida (de Vargas et al. 2015; Lukeš et al. 2015; Flegontova et al. 2016). As compared to Diplonemidae, DSPD I diplonemids have very short branches in 18S rRNA trees (see Figure 2 in section 3.5, *Manuscript II*; Gawryluk et al. 2016; Tashyreva et al. submitted), and therefore they have likely experienced explosive speciation relatively recently.

4 References

Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V, Heiss A, Hoppenrath M, Lara E, Le Gall L, Lynn DH, McManus H, Mitchell EA, Mozley-Stanridge SE, Parfrey LW, Pawlowski J, Rueckert S, Shadwick L, Schoch CL, Smirnov A, Spiegel FW (2012) The revised classification of eukaryotes. *J Eukaryot Microbiol*, **59**, 429-493.

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM (2009) A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One*, **4**, e6372.

Berchtold M, Philippe H, Breunig A, Brugerolle G, Konig H (1994) The phylogenetic position of *Dimastigella trypaniformis* within the parasitic kinetoplastids. *Parasitol Res*, **80**, 672-679.

Berney C, Ciuprina A, Bender S, Brodie J, Edgcomb V, Kim E, Rajan J, Parfrey LW, Adl S, Audic S, Bass D, Caron DA, Cochrane G, Czech L, Dunthorn M, Geisen S, Glockner FO, Mahe F, Quast C, Kaye JZ, Simpson AGB, Stamatakis A, del Campo J, Yilmaz P, de Vargas C (2017) UniEuk: Time to speak a common language in protistology! *J Eukaryot Microbiol*, **64**(3), 407-411.

Biard T, Bigeard E, Audic S, Poulain J, Gutierrez-Rodriguez A, Pesant S, Stemmann L, Not F (2017) Biogeography and diversity of Collodaria (Radiolaria) in the global ocean. *ISME J*, **11**, 1331-1344.

Blom D, de Haan A, van den Berg M, Sloof P, Jirků M, Lukeš J, Benne R (1998) RNA editing in the free-living bodonid *Bodo saltans*. *Nucleic Acids Res*, **26**, 1205-1213.

Bragg L, Tyson GW (2014) Metagenomics using next-generation sequencing. *Methods Mol Biol*, **1096**, 183-201.

Breunig A, Konig H, Brugerolle G, Vickerman K, Hertel H (1993) Isolation and ultrastructural features of a new strain of *Dimastigella trypaniformis* Sandon 1928 (Bodonina, Kinetoplastida) and comparison with a previously isolated strain. *Eur J Protistol*, **29**, 416-424.

Brown EA, Chain FJ, Crease TJ, MacIsaac HJ, Cristescu ME (2015) Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? *Ecol Evol*, **5**, 2234-2251.

Brugerolle G, Lom J, Nohynkova E, Joyon L (1979) Comparison et evolution des structures cellulaires chez plusieurs especes de bodonides et cryptobiides appurtenant aux genres *Bodo*, *Cryptobia* et Trypanoplasma (Kinetoplastida, Mastigophora). *Protistologica*, **15**, 197-221.

Brugerolle G (1985) Des trichocystes chez les bodonides, un caractere phylogenetique supplementaire entre Kinetoplastida et Euglenida. *Protistologica*, **21**, 339-348.

Burger G, Gray MW, Lang BF (2003) Mitochondrial genomes: anything goes. *Trends Genet*, **19**, 709-716.

Burger G, Moreira S, Valach M (2016) Genes in hiding. *Trends Genet*, **32**, 553-565.

Busse I, Preisfeld A (2002) Phylogenetic position of *Rhynchopus* sp. and *Diplonema ambulator* as indicated by analyses of euglenozoan small subunit ribosomal DNA. *Gene*, **284**, 83-91.

Callahan HA, Litaker RW, Noga EJ (2002) Molecular taxonomy of the suborder Bodonina (Order Kinetoplastida), including the important fish parasite, *Ichthyobodo necator*. *J Eukaryot Microbiol*, **49**, 119-128.

Canaday J, Tessier LH, Imbault P, Paulus F (2001) Analysis of *Euglena gracilis* alpha-, beta- and gamma-tubulin genes: introns and pre-mRNA maturation. *Mol Genet Genomics*, **265**, 153-160.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972-1973.

Cavalier-Smith T (1981) Eukaryote kingdoms: seven or nine? *Biosystems*, **14**, 461-481.

Cavalier-Smith T (1993) Kingdom Protozoa and its 18 phyla. *Microbiol Rev*, **57**, 953-994.

Cavalier-Smith T (2016) Higher classification and phylogeny of Euglenozoa. *Eur J Protistol*, **56**, 250-276.

Clayton CE (2016) Gene expression in kinetoplastids. *Current Opinion in Microbiology*, **32**, 46-51.

Corliss JO (1984) The kingdom Protista and its 45 phyla. *BioSystems*, **17**, 87-126.

Countway PD, Gast RJ, Dennett MR, Savai P, Rose JM, Caron DA (2007) Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). *Environ Microbiol*, **9**, 1219-1232.

Danovaro R, Corinaldesi C, Dell'Anno A, Fabiano M, Corselli C (2005) Viruses, prokaryotes and DNA in the sediments of a deep-hypersaline anoxic basin (DHAB) of the Mediterranean Sea. *Environ Microbiol*, **7**, 586-592.

David V, Flegontov P, Gerasimov E, Tanifuji G, Hashimi H, Logacheva MD, Maruyama S, Onodera NT, Gray MW, Archibald JM, Lukeš J (2015) Gene loss and error-prone RNA editing in the mitochondrion of *Perkinsela*, an endosymbiotic kinetoplastid. *mBio*, **6**, e01498-15.

d'Avila-Levy CM, Volotao AC, Araujo FM, de Jesus JB, Motta MC, Vermelho AB, Santos AL, Branquinha MH (2009) *Bodo* sp., a free-living flagellate, expresses divergent proteolytic activities from the closely related parasitic trypanosomatids. *J Eukaryot Microbiol*, **56**, 454-458.

Debroas D, Domaizon I, Humbert JF, Jardillier L, Lepere C, Oudart A, Taib N (2017) Overview of freshwater microbial eukaryotes diversity: a first analysis of publicly available metabarcoding data. *FEMS Microbiol Ecol*, **93**, fix023.

del Campo J, Ruiz-Trillo I (2013) Environmental survey meta-analysis reveals hidden diversity among unicellular opisthokonts. *Mol Biol Evol*, **30**, 802-805.

del Campo J, Guillou L, Hehenberger E, Logares R, Lopez-Garcia P, Massana R (2016) Ecological and evolutionary significance of novel protist lineages. *Eur J Protistol*, **55**, 4-11.

Dell'Anno A, Danovaro R (2005) Extracellular DNA plays a key role in deep-sea ecosystem functioning. *Science*, **309**, 2179.

de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury JM, Bittner L, Chaffron S, Dunthorn M, Engelen S, Flegontova O, Guidi L, Horák A, Jaillon O, Lima-Mendez G, Lukeš J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S; *Tara* Oceans Coordinators, Acinas SG, Bork P, Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science*, **348**, 1261605.

Dobakova E, Flegontov P, Skalicky T, Lukeš J (2015) Unexpectedly streamlined mitochondrial genome of the Euglenozoan *Euglena gracilis*. *Genome Biol Evol*, **7**, 3358-3367.

Dolezel D, Jirků M, Maslov DA, Lukeš J (2000) Phylogeny of the bodonid flagellates (Kinetoplastida) based on small-subunit rRNA gene sequences. *Int J Syst Evol Microbiol*, **5**, 1943-1951.

Dupont AOC, Griffiths RI, Bell T, Bass D (2016) Differences in soil micro-eukaryotic communities over soil pH gradients are strongly driven by parasites and saprotrophs. *Environ Microbiol*, **18**, 2010-2024.

Dykova I, Fiala I, Lom J, Lukeš J (2003) *Perkinsiella* amoebaelike endosymbionts of *Neoparamoeba* spp., relatives of the kinetoplastid *Ichthyobodo*. *Eur J Protistol*, **39**, 37-52.

Dykova I, Fiala I, Peckova H (2008) *Neoparamoeba* spp. and their eukaryotic endosymbionts similar to *Perkinsela amoebae* (Hollande, 1980): Coevolution demonstrated by SSU rRNA gene phylogenies. *Eur J Protistol*, **44**, 269-277.

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460-2461.

Edgcomb V, Orsi W, Bunge J, Jeon S, Christen R, Leslin C, Holder M, Taylor GT, Suarez P, Varela R, Epstein S (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J*, **5**, 1344-1356.

Egge ES, Johannessen TV, Andersen T, Eikrem W, Bittner L, Larsen A, Sandaa RA, Edvardsen B (2015) Seasonal diversity and dynamics of haptophytes in the Skagerrak, Norway, explored by high-throughput sequencing. *Mol Ecol*, **24**, 3026-3042.

Elbrachter M, Schnepf E, Balzer I (1996) *Hemistasia phaecysticola* (Scherffel) comb. nov., redescription of a free-living, marine, phagotrophic kinetoplastid flagellate. *Arch Protistenkd*, **147**, 125-136.

Eloe EA, Shulse CN, Fadrosh DW, Williamson SJ, Allen EE, Bartlett DH (2011) Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. *Environ Microbiol Rep*, **3**, 449-458.

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, Caler E, Renauld H, Worthey EA, Hertz-Fowler C, et al. (2005) Comparative genomics of trypanosomatid parasitic protozoa. *Science*, **309**, 404-409.

Eyden BP (1977) Morphology and ultrastructure of *Bodo designis* Skuja 1948. *Protistologica*, **13**, 169-179.

Filker S, Gimmler A, Dunthorn M, Mahe F, Stoeck T (2015) Deep sequencing uncovers protistan plankton diversity in the Portuguese Ria Formosa solar saltern ponds. *Extremophiles*, **19**, 283-295.

Flegontov P, Gray MW, Burger G, Lukeš J (2011) Gene fragmentation: a key to mitochondrial genome evolution in Euglenozoa? *Curr Genet*, **57**, 225-232.

Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, Bowler C, Lukeš J, Horák A (2016) Extreme diversity of diplonemid eukaryotes in the ocean. *Curr Biol*, **26**, 3060-3065.

Forster D, Bittner L, Karkar S, Dunthorn M, Romac S, Audic S, Lopez P, Stoeck T, Bapteste E (2015) Testing ecological theories with sequence similarity networks: marine ciliates exhibit similar geographic dispersal patterns as multicellular organisms. *BMC Biol*, **13**, 16.

Forster D, Dunthorn M, Mahe F, Dolan JR, Audic S, Bass D, Bittner L, Boutte C, Christen R, Claverie JM, Decelle J, Edvardsen B, Egge E, Eikrem W, Gobet A, Kooistra WH, Logares R, Massana R, Montresor M, Not F, Ogata H, Pawlowski J, Pernice MC, Romac S, Shalchian-Tabrizi K, Simon N, Richards TA, Santini S, Sarno D, Siano R, Vaulot D, Wincker P, Zingone A, de Vargas C, Stoeck T (2016a) Benthic protists: the under-charted majority. *FEMS Microbiol Ecol*, **92**, fiw120.

Forster D, Dunthorn M, Stoeck T, Mahe F (2016b) Comparison of three clustering approaches for detecting novel environmental microbial diversity. *PeerJ*, **4**, e1692.

Frantz C, Ebel C, Paulus F, Imbault P (2000) Characterization of trans-splicing in Euglenoids. *Curr Genet*, **37**, 349-355.

Gawryluk RMR, del Campo J, Okamoto N, Strassert JFH, Lukeš J, Richards TA, Worden AZ, Santoro AE, Keeling PJ (2016) Morphological identification and single-cell genomics of marine diplonemids. *Curr Biol*, **26**, 3053-3059.

Giner CR, Forn I, Romac S, Logares R, de Vargas C, Massana R (2016) Environmental sequencing provides reasonable estimates of the relative abundance of specific picoeukaryotes. *Appl Environ Microbiol*, **82**, 4757-4766.

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, de Vargas C, Decelle J, del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WH, Lara E, le Bescot N, Logares R, Mahe F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet AL, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R (2013) The protist ribosomal reference database (PR2): a catalog of unicellular eukaryote small subunit rRNA sequences with curated taxonomy. *Nucleic Acids Res*, **41**, D597-604.

Griessmann K (1913) Ober marine Flagellaten. *Archivffir Protistenkunde*, **32**, 1-78.

Hajduk SL, Siqueira AM, Vickerman K (1986) Kinetoplast DNA of *Bodo caudatus*: a noncatenated structure. *Mol Cell Biol*, **6**, 4372-4378.

Hirose E, Nozawa A, Kumagai A, Kitamura S (2012) *Azumiobodo hoyamushi* gen. nov. et sp. nov. (Euglenozoa, Kinetoplastea, Neobodonida): a pathogenic kinetoplastid causing the soft tunic syndrome in ascidian aquaculture. *Dis Aquat Organ*, **97**, 227-235.

Hollande A (1980) Identification du parasome (Nebenkern) de *Janickina pigmentifera* à un symbionte (*Perkinsiella amoebae* nov gen – nov sp.) apparenté aux flagellés Kinetoplastidiés. *Protistologica*, **16**, 613-625.

Hong S, Bunge J, Leslin C, Jeon S, Epstein SS (2009) Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J*, **3**, 1365-1373.

Honigberg BM (1963) A contribution to systematics of the non-pigmented flagellates. In *Progress in Protozoology: proceedings of the first International Congress on protozoology held at Prague* (ed. J Ludvik, J Lom & J Vavra). Academic Press.

Huber JA, Welch DM, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML (2007) Microbial population structures in the deep marine biosphere. *Science*, **318**, 97-100.

Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol*, **8**, R143.

Isaksen TE, Karlsbakk E, Repstad O, Nylund A (2012) Molecular tools for the detection and identification of *Ichthyobodo* spp. (Kinetoplastida), important fish parasites. *Parasitol Int*, **61**, 675-683.

Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, Hartley C, Sanders M, Wastling JM, Dacks JB, Acosta-Serrano A, Field MC, Ginger ML, Berriman M (2016) Kinetoplastid phylogenomics reveals the evolutionary innovations associated with the origins of parasitism. *Curr Biol*, **26**, 1-12.

Kable ML, Heidmann S, Stuart KD (1997) RNA editing: getting U into RNA. *Trends Biochem Sci*, **22**, 162-166.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*, **30,** 772-780.

Keeling PJ, del Campo J (2017) Marine protists are not just big bacteria. *Curr Biol*, **27**, R541-R549.

Kent ML, Elston RA (1987) An *Isonema*-like flagellate (Protozoa Mastigophora) infection in larval geoduck clams, *Panope abrupta*. *J Invert Pathol*, **50**, 221-229.

Kiethega GN, Yan Y, Turcotte M, Burger G (2013) RNA-level unscrambling of fragmented genes in *Diplonema* mitochondria. *RNA Biol*, **10**, 301-313.

Kivic PA, Walne PL (1984) An evaluation of a possible phylogenetic relationship between the Euglenophyta and Kinetoplastida. *Origins Life*, **13**, 269-288.

Klingbeil MM, Drew ME, Liu Y, Morris JC, Motyka SA, Saxowsky TT, Wang Z, Englund PT (2001) Unlocking the secrets of trypanosome kinetoplast DNA network replication. *Protist*, **152**, 255-262.

Koeppel AF, Wu M (2013) Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Res*, **41**, 5175-5188.

Lara E, Moreira D, Vereshchaka A, Lopez-Garcia P (2009) Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. *Environ Microbiol*, **11**, 47-55.

Larsen J, Patterson JL (1990) Some flagellates (Protista) from tropical marine sediments. *J Nat Hist*, **24**, 801-937.

Lecroq B, Gooday AJ, Cedhagen T, Sabbatini A, Pawlowski J (2009) Molecular analyses reveal high levels of eukaryotic richness associated with enigmatic deep-sea protists (Komokiacea). *Mar Biodiv*, **39**, 45-55.

Leray M, Knowlton N (2016) Censusing marine eukaryotic diversity in the twenty-first century. *Philos Trans R Soc Lond B Biol Sci*, **371**, 1702.

Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, Chaffron S, Ignacio-Espinosa JC, Roux S, Vincent F, Bittner L, Darzi Y, Wang J, Audic S, Berline L, Bontempi G, Cabello AM, Coppola L, Cornejo-Castillo FM, D'Ovidio F, DeMeester L, Ferrera I, Garet-Delmas MJ, Guidi L, Lara E, Pesant S, Royo-Llonch M, Salazar G, Sanchez P, Sebastian M, Souffreau C, Dimier C, Picheral M, Searson S, Kandels-Lewis S, *Tara* Oceans Coordinators, Gorsky G, Not F, Ogata H, Speich S, Stemmann L, Weissenbach J, Wincker P, Acinas SG, Sunagawa S, Bork P, Sullivan MB, Karsenti E, Bowler C, de Vargas C, Raes J (2015) Determinants of community structure in the global plankton interactome. *Science*, **348**, 1262073.

Lin YC, Campbell T, Chung CC, Gong GC, Chiang KP, Worden AZ (2012) Distribution patterns and phylogeny of marine stramenopiles in the north Pacific Ocean. *Appl Environ Microbiol*, **78**, 3387-3399.

Logares R, Audic S, Santini S, Pernice MC, de Vargas C, Massana R (2012a) Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. *ISME J*, **6**, 1823-1833.

Logares R, Haverkamp TH, Kumar S, Lanzen A, Nederbragt AJ, Quince C, Kauserud H (2012b) Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of current platforms and bioinformatics approaches. *J Microbiol Methods*, **91**, 106-113.

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, Hingamp P, Ogata H, de Vargas C, Lima-Mendez G, Raes J, Poulain J, Jaillon O, Wincker P, Kandels-Lewis S, Karsenti E, Bork P, Acinas SG (2014a) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol*, **16**, 2659-2671.

Logares R, Audic S, Bass D, Bittner L, Boutte C, Christen R, Claverie JM, Decelle J, Dolan JR, Dunthorn M, Edvardsen B, Gobet A, Kooistra WH, Mahe F, Not F, Ogata H, Pawlowski J, Pernice MC, Romac S, Shalchian-Tabrizi K, Simon N, Stoeck T, Santini S, Siano R, Wincker P, Zingone A, Richards TA, de Vargas C, Massana R (2014b) Patterns of rare and abundant marine microbial eukaryotes. *Curr Biol*, **24**, 813-821.

Logares R, Mangot JF, Massana R (2015) Rarity in aquatic microbes: placing protists on the map. *Res Microbiol*, **166**, 831-841.

Lopez-Garcia P, Rodriguez-Valera F, Pedros-Alio C, Moreira D (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature*, **409**, 603-607.

Lopez-Garcia P, Philippe H, Gail F, Moreira D (2003) Autochthonous eukaryotic diversity in hydrothermal sediment and experimental microcolonizers at the Mid-Atlantic Ridge. *PNAS*, **100**, 697-702.

Lopez-Garcia P, Vereshchaka A, Moreira D (2007) Eukaryotic diversity associated with carbonates and fluid-seawater interface in Lost City hydrothermal field. *Environ Microbiol*, **9**, 546-554.

Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R (2012) Diversity, stability and resilience of the human gut microbiota. *Nature*, **489**, 220-230.

Lukeš J, Jirků M, Doležel D, Králová I, Hollar L, Maslov DA (1997) Analysis of ribosomal RNA genes suggests that trypanosomes are monophyletic. *J Mol Evol*, **44**, 521-527.

Lukeš J, Jirků M, Avliyakulov N, Benada O (1998) Pankinetoplast DNA structure in a primitive bodonid flagellate, *Cryptobia helicis*. *EMBO J*, **17**, 838-846.

Lukeš J, Guilbride DL, Votýpka J, Zíková A, Benne R, Englund PT (2002) Kinetoplast DNA network: evolution of an improbable structure. *Eukaryot Cell*, **1**, 495-502.

Lukeš J, Leander BS, Keeling PJ (2009) Cascades of convergent evolution: The corresponding evolutionary histories of euglenozoans and dinoflagellates. *PNAS*, **106**, 9963-9970.

Lukeš J, Skalický T, Týč J, Votýpka J, Yurchenko V (2014) Evolution of parasitism in kinetoplastid flagellates. *Mol Biochem Parasitol*, **195**, 115-122.

Lukeš J, Flegontova O, Horák A (2015) Diplonemids. *Curr Biol*, **25**, R702-R704.

Mahe F, Rognes T, Quince C, de Vargas C, Dunthorn M (2014) Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ*, **2**, e593.

Mahe F, Rognes T, Quince C, de Vargas C, Dunthorn M (2015) Swarm v2: highly scalable and high-resolution amplicon clustering. *PeerJ*, **3**, e1420.

Mahe F, de Vargas C, Bass D, Czech L, Stamatakis A, Lara E, Singer D, Mayor J, Bunge J, Sernaker S, Siemensmeyer T, Trautmann I, Romac S, Berney C, Kozlov A, Mitchell EAD, Seppey CVW, Egge E, Lentendu G, Wirth R, Trueba G, Dunthorn M (2017) Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nature Ecology & Evolution*, **1**, 0091.

Makiuchi T, Annoura T, Hashimoto M, Hashimoto T, Aoki T, Nara T (2011) Compartmentalization of a glycolytic enzymein *Diplonema*, a non-kinetoplastid euglenozoan. *Protist*, **162**, 482-489.

Malviya S, Scalco E, Audic S, Vincent F, Veluchamy A, Poulain J, Wincker P, Iudicone D, de Vargas C, Bittner L, Zingone A, Bowler C (2016) Insights into global diatom distribution and diversity in the world's ocean. *Proc Natl Acad Sci USA*, **113**, E1516-25.

Marande W, Lukeš J, Burger G (2005) Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. *Eukaryot Cell*, **4**, 1137-1146.

Marande W, Burger G (2007) Mitochondrial DNA as a genomic jigsaw puzzle. *Science*, **318**, 415.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380.

Maslov DA, Simpson L (1994) RNA editing and genomic organization in the cryptobiid kinetoplastid protozoan, *Trypanoplasma borreli*. *Mol Cell Biol*, **14**, 8174-8182.
Maslov DA, Yasuhira S, Simpson L (1999) Phylogenetic affinities of *Diplonema* within the Euglenozoa as inferred from the SSU rRNA gene and partial COI protein sequences. *Protist*, **150**, 33-42.
Maslov DA, Votypka J, Yurchenko V, Lukeš J (2013) Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. *Trends Parasitol*, **29**, 43-52.
Massana R, del Campo J, Sieracki ME, Audic S, Logares R (2014) Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J*, **8**, 854-866.
Massana R, Gobet A, Audic S, Bass D, Bittner L, Boutte C, Chambouvet A, Christen R, Claverie JM, Decelle J, Dolan JR, Dunthorn M, Edvardsen B, Forn I, Forster D, Guillou L, Jaillon O, Kooistra WH, Logares R, Mahe F, Not F, Ogata H, Pawlowski J, Pernice MC, Probert I, Romac S, Richards T, Santini S, Shalchian-Tabrizi K, Siano R, Simon N, Stoeck T, Vaulot D, Zingone A, de Vargas C (2015) Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ Microbiol*, **17**, 4035-4049.
Medinger R, Nolte V, Pandey RV, Jost S, Ottenwalder B, Schlotterer C, Boenigk J (2010) Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Mol Ecol*, **1**, 32-40.
Montegut-Felkner AE, Triemer RE (1994) Phylogeny of *Diplonema ambulatur* (Larsen & Patterson). 1. Homologies of the flagellar apparatus. *Europ J Protistol*, **30**, 227-237.
Montegut-Felkner AE, Triemer RE (1996) Phylogeny of *Diplonema ambulatur* (Larsen & Patterson). 2. Homologies of the feeding apparatus. *Europ J Protistol*, **32**, 64-76.
Moon-van der Staay SY, de Wachter R, Vaulot D (2001) Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*, **409**, 607-610.
Moreira D, Lopez-Garcia P, Rodriguez-Valera F (2001) New insights into the phylogenetic position of diplonemids: G+C content bias, differences of evolutionary rate and a new environmental sequence. *Int J Syst Evol Microbiol*, **51**, 2211-2219.
Moreira D, Lopez-Garcia P, Vickerman K (2004) An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. *Int J Syst Evol Microbiol*, **54**, 1861-1875.
Moreira S, Valach M, Aoulad-Aissa M, Otto C, Burger G (2016) Novel modes of RNA editing in mitochondria. *Nucleic Acids Res*, **44**, 4907-4919.
Morgan-Smith D, Herndl GJ, van Aken HM, Bochdansky AB (2011) Abundance of eukaryotic microbes in the deep subtropical North Atlantic. *Aquat Microb Ecol*, **65**, 103-115.
Morgan-Smith D, Clouse MA, Herndl GJ, Bochdansky AB (2013) Diversity and distribution of microbial eukaryotes in the deep tropical and subtropical North Atlantic Ocean. *Deep-Sea Research I*, **78**, 58-69.
Mukherjee I, Hodoki Y, Nakano S (2015) Kinetoplastid flagellates overlooked by universal primers dominate in the oxygenated hypolimnion of Lake Biwa, Japan. *FEMS Microbiol Ecol*, **91**, fiv083.
Nebel M, Pfabel C, Stock A, Dunthorn M, Stoeck T (2011) Delimiting operational taxonomic units for assessing ciliate environmental diversity using small-subunit rRNA gene sequences. *Environ Microbiol Rep*, **3**, 154-158.
Nielsen KM, Johnsen PJ, Bensasson D, Daffonchio D (2007) Release and persistence of extracellular DNA in the environment. *Environ Biosafety Res*, **6**, 37-53.

Not F, Valentin K, Romari K, Lovejoy C, Massana R, Tobe K, Vaulot D, Medlin LK (2007) Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes. *Science*, **315**, 253-255.

Oikonomou A, Filker S, Breiner HW, Stoeck T (2015) Protistan diversity in a permanently stratified meromictic lake (Lake Alatsee, SW Germany). *Environ Microbiol*, **17**, 2144-2157.

Opperdoes FR, Borst P (1977) Localization of nine glycolytic enzymes in a microbody-like organelle in *Trypanosoma brucei*: the glycosome. *FEBS Lett*, **80**, 360-364.

Opperdoes FR, Nohynkova E, van Schaftingen E, Lambeir AM, Veenhuis M, van Roy J (1988) Demonstration of glycosomes (microbodies) in the bodonid flagellate *Trypanoplasma borelli* (Protozoa, Kinetoplastida). *Mol Biochem Parasitol*, **30**, 155-163.

Patterson DJ (1988) The evolution of protozoa. *Mem Inst Oswaldo Cruz*, **83**, 580-600.

Patterson DJ (1994) Protozoa: evolution and systematics. In *Progress in Protozoology* (ed. K Hausmann & N Holsmann), pp. 1-14. Stuttgart, Germany.

Patterson DJ, Lee WJ (2000) Geographic distribution and diversity of free-living heterotrophic flagellates. In *The Flagellates – Unity, diversity and evolution* (ed. BSC Leadbeater & JC Green), pp. 269-287. London, UK: Taylor & Francis.

Pawlowski J, Christen R, Lecroq B, Bachar D, Shahbazkia HR, Amaral-Zettler L, Guillou L (2011) Eukaryotic richness in the abyss: insights from pyrotag sequencing. *PLoS One*, **6**, e18169.

Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, Bowser SS, Cepicka I, Decelle J, Dunthorn M, Fiore-Donno AM, Gile GH, Holzmann M, Jahn R, Jirků M, Keeling PJ, Kostka M, Kudryavtsev A, Lara E, Lukeš J, Mann DG, Mitchell EA, Nitsche F, Romeralo M, Saunders GW, Simpson AGB, Smirnov AV, Spouge JL, Stern RF, Stoeck T, Zimmermann J, Schindel D, de Vargas C (2012) CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biol*, **10**, e1001419.

Pedersen MW, Ruter A, Schweger C, Friebe H, Staff RA, Kjeldsen KK, Mendoza ML, Beaudoin AB, Zutter C, Larsen NK, Potter BA, Nielsen R, Rainville RA, Orlando L, Meltzer DJ, Kjær KH, Willerslev E (2016) Postglacial viability and colonization in North America's ice-free corridor. *Nature*, **537**, 45-49.

Pernice MC, Giner CR, Logares R, Perera-Bel J, Acinas SG, Duarte CM, Gasol JM, Massana R (2016) Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. *ISME J*, **10**, 945-958.

Piredda R, Tomasino MP, D'Erchia AM, Manzari C, Pesole G, Montresor M, Kooistra WH, Sarno D, Zingone A (2017) Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. *FEMS Microbiol Ecol*, **93**, fiw200.

Porter D (1973) *Isonema papillatum* sp. n., a new colorless marine flagellate: a light- and electronmicroscopic study. *Protozool*, **20**, 351-356.

Price MN, Dehal PS, Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.

Prokopowich CD, Gregory TR, Crease TJ (2003) The correlation between rDNA copy number and genome size in eukaryotes. *Genome*, **46**, 48-50.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, Glockner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*, **35**, 7188-7196.

Roy J, Faktorova D, Benana O, Lukeš J, Burger G (2007) Description of *Rhynchopus euleeides* n. sp. (Diplonemea), a free-living marine euglenozoan. *J Eukaryot Microbiol*, **54**, 137-145.

Salani FS, Arndt H, Hausmann K, Nitsche F, Scheckenbach F (2012) Analysis of the community structure of abyssal kinetoplastids revealed similar communities at larger spatial scales. *ISME J*, **6**, 713-723.

Santana DM, Lukeš J, Sturm NR, Campbell DA (2001) Two sequence classes of kinetoplastid 5S ribosomal RNA gene revealed among bodonid spliced leader RNA gene arrays. *FEMS Microbiol Lett*, **204**, 233-237.

Sauvadet AL, Gobet A, Guillou L (2010) Comparative analysis between protist communities from the deep-sea pelagic ecosystem and specific deep hydrothermal habitats. *Environ Microbiol*, **12**, 2946-2964.

Scheckenbach F, Hausmann K, Wylezich C, Weitere M, Arndt H (2010) Large-scale patterns in biodiversity of microbial eukaryotes from the abyssal sea floor. *Proc Natl Acad Sci USA*, **107**, 115-120.

Scherffel A (1900) *Phaeocystis globosa* nov. spec. nebst einigen Betrachtungen über die Phylogenie niederer, insbesondere brauner Organismen. *Wiss Meeresunters*, *Abt Helgoland*, **4**, 1-29.

Schmidt TS, Matias Rodrigues JF, von Mering C (2015) Limits to robustness and reproducibility in the demarcation of operational taxonomic units. *Environ Microbiol*, **17**, 1689-1706.

Schnepf E (1994) Light and electron microscopical observations in *Rhynchopus coscinodiscivorus* spec. nov., a color-less, phagotrophic euglenozoon with concealed flagella. *Arch Protistenkd*, **144**, 63-74.

Schuster FL, Goldstein S, Hershenov B (1968) Ultrastructure of a flagellate *Isonema nigrigricans* nov. gen. nov. sp. from a polluted marine habitat. *Protistologica*, **4**, 141-149.

Seenivasan R, Sausen N, Medlin LK, Melkonian M (2013) Picomonas judraskeda gen. et sp. nov.: the first identified member of the Picozoa phylum nov., a widespread group of picoeukaryotes, formerly known as 'picobiliphytes'. *PLoS One*, **8**, e59565.

Seersholm FV, Pedersen MW, Soe MJ, Shokry H, Mak SS, Ruter A, Raghavan M, Fitzhugh W, Kjær KH, Willerslev E, Meldgaard M, Kapel CM, Hansen AJ (2016) DNA evidence of bowhead whale exploitation by Greenlandic Paleo-Inuit 4,000 years ago. *Nat Commun*, **7**, 13389.

Shapiro TA, Englund PT (1995) The structure and replication of kinetoplast DNA. *Annu Rev Microbiol*, **49**, 117-143.

Siano R, Alves-de-Souza C, Foulon E, Bendif EM, Simon N, Guillou L, Not F (2010) Distribution and host diversity of Amoebophryidae parasites across oligotrophic waters of the Mediterranean Sea. *Biogeosciences*, **8**, 267-278.

Simpson AGB (1997a) The identity and composition of the Euglenozoa. *Arch. Protistenkd.*, **148**, 318-328.

Simpson L (1997b) The genomic organization of guide RNA genes in kinetoplastid protozoa: several conundrums and their solutions. *Mol Biochem Parasitol*, **86**, 133-141.

Simpson AGB, Lukeš J, Roger AJ (2002) The evolutionary history of kinetoplastids and their kinetoplasts. *Mol Biol Evol*, **19**, 2071-2083.

Simpson AGB, Roger AJ (2004) Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. *Mol Phylogenet Evol*, **30**, 201-212.

Skuja H (1948) Taxonomie des Phytoplanktons einiger Seen in Uppland, Schweden. *Symb Bot Ups*, **10**, 1-399.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proc Natl Acad Sci USA*, **103**, 12115-12120.

Solomon JA, Walne PL, Kivic PA (1987) *Entosiphon sulcatum* (Euglenophyceae): flagellar roots of the basal body complex and reservoir region. *J Phycol*, **23**, 85-98.
Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312-1313.
Stoeck T, Zuendorf A, Breiner HW, Behnke A (2007) A molecular approach to identify active microbes in environmental eukaryote clone libraries. *Microb Ecol*, **53**, 328-339.
Stoeck T, Bass D, Nebel M, Christen R, Jones MD, Breiner HW, Richards TA (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol*, **1**, 21-31.
Stolba P, Jirků M, Lukeš J (2001) Polykinetoplast DNA structure in *Dimastigella trypaniformis* and *Dimastigella mimosa* (Kinetoplastida). *Mol Biochem Parasitol*, **113**, 323-326.
Sturm NR, Maslov DA, Grisard EC, Campbell DA (2001) *Diplonema* spp. possess spliced leader RNA genes similar to the Kinetoplastida. *J Eukaryot Microbiol*, **48**, 325-331.
Swale EMF (1973) A study of the colourless flagellate *Rhynchomonas nasuta* (Stokes) Klebs. *Biol J Linn Soc*, **5**, 255-264.
Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Mol Ecol*, **21**, 2045-2050.
Tashyreva D, Prokopchuk G, Yabuki A, Kaur B, Faktorova D, Votypka J, Kusaka C, Fujikura K, Shiratori T, Ishida KI, Horák A, Lukeš J (submitted) Phylogeny and morphology of new diplonemids from Japan.
Todal JA, Karlsbakk E, Isaksen TE, Plarre H, Urawa S, Mouton A, Hoel E, Koren CWR, Nylund A (2004) *Ichthyobodo necator* (Kinetoplastida) – a complex of sibling species. *Dis Aquat Org*, **58**, 9-16.
Triemer RE, Ott DW (1990) Ultrastructure of *Diplonema ambulator* Larsen & Patterson (Euglenozoa) and its relationship to *Isonema*. *Eur J Protistol*, **25**, 316-320.
Triemer RE, Farmer MA (1991) An ultrastructural comparison of the mitotic apparatus, feeding apparatus, flagellar apparatus and cytoskeleton in euglenoids and kinetoplastids. *Protoplasma*, **164**, 91-104.
Valach M, Moreira S, Kiethega GN, Burger G (2014) Trans-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria. *Nucleic Acids Res*, **42**, 2660-2672.
Valach M, Moreira S, Faktorova D, Lukeš J, Burger G (2016) Post-transcriptional mending of gene sequences: Looking under the hood of mitochondrial gene expression in diplonemids. *RNA Biol*, **13**, 1204-1211.
Vickerman K (1976) The diversity of the kinetoplastid flagellates. In *Biology of the Kinetoplastida* (ed. WHR Lumsden & DA Evans), pp. 1-34. London, UK: Academic Press.
Viprey M, Guillou L, Ferreol M, Vaulot D (2008) Wide genetic diversity of picoplanktonic green algae (Chloroplastida) in the Mediterranean Sea uncovered by a phylum-biased PCR approach. *Environ Microbiol*, **10**, 1804-1822.
Vlček C, Marande W, Teijeiro S, Lukeš J, Burger G (2011) Systematically fragmented genes in a multipartite mitochondrial genome. *Nucleic Acids Res*, **39**, 979-988.
von der Heyden S, Chao EE, Vickerman K, Cavalier-Smith T (2004) Ribosomal RNA phylogeny of Bodonid and Diplonemid flagellates and the evolution of Euglenozoa. *J Eukaryot Microbiol*, **51**, 402-416.
von der Heyden S, Cavalier-Smith T (2005) Culturing and environmental DNA sequencing uncover hidden kinetoplastid biodiversity and a major marine clade within ancestrally freshwater *Neobodo designis*. *Int J Syst Evol Microbiol*, **55**, 2605-2621.

von Wintzingerode F, Gobel UB, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev*, **21**, 213-229.

Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ (2015) Environmental science. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. *Science*, **347**, 1257594.

Yabuki A, Tame A (2015) Phylogeny and reclassification of *Hemistasia phaeocysticola* (Scherffel) Elbrächter & Schnepf, 1996. *J Eukaryot Microbiol*, **62**, 426-429.

Yabuki A, Tanifuji G, Kusaka C, Takishita K, Fujikura K (2016) Hyper-eccentric structural genes in the mitochondrial genome of the algal parasite *Hemistasia phaeocysticola*. *Genome Biol Evol*, **8**, 2870-2878.

Yubuki N, Edgcomb VP, Bernhard JM, Leander BS (2009) Ultrastructure and molecular phylogeny of *Calkinsia aureus*: cellular identity of a novel clade of deep-sea euglenozoans with epibiotic bacteria. *BMC Microbiol*, **9**, 16.

Zhu F, Massana R, Not F, Marie D, Vaulot D (2005) Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol*, **52**, 79-92.

Zíková A, Vancová M, Jirků M, Lukeš J (2003) *Cruzella marina* (Bodonina, Kinetoplastida): non-catenated structure of poly-kinetoplast DNA. *Exp Parasitol*, **104**, 159-161.

5 Original publications

5.1 Paper I

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury JM, Bittner L, Chaffron S, Dunthorn M, Engelen S, **Flegontova O**, Guidi L, Horák A, Jaillon O, Lima-Mendez G, Lukeš J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S; *Tara* Oceans Coordinators, Acinas SG, Bork P, Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E (2015) Eukaryotic plankton diversity in the sunlit ocean. ***Science***. 348(6237):1261605 (IF = 37.205).

Abstract

Marine plankton support global biological and geochemical processes. Surveys of their biodiversity have hitherto been geographically restricted and have not accounted for the full range of plankton size.We assessed eukaryotic diversity from 334 size-fractionated photic-zone plankton communities collected across tropical and temperate oceans during the circumglobal *Tara* Oceans expedition.We analyzed 18S ribosomal DNA sequences across the intermediate plankton-size spectrum from the smallest unicellular eukaryotes (protists, >0.8 micrometers) to small animals of a few millimeters. Eukaryotic ribosomal diversity saturated at ~150,000 operational taxonomic units, about one-third of which could not be assigned to known eukaryotic groups. Diversity emerged at all taxonomic levels, both within the groups comprising the ~11,200 cataloged morphospecies of eukaryotic plankton and among twice as many other deep-branching lineages of unappreciated importance in plankton ecology studies. Most eukaryotic plankton biodiversity belonged to heterotrophic protistan groups, particularly those known to be parasites or symbiotic hosts.

OCEAN PLANKTON

Eukaryotic plankton diversity in the sunlit ocean

Colomban de Vargas,[1,2]*† Stéphane Audic,[1,2]† Nicolas Henry,[1,2]† Johan Decelle,[1,2]† Frédéric Mahé,[3,1,2]† Ramiro Logares,[4] Enrique Lara,[5] Cédric Berney,[1,2] Noan Le Bescot,[1,2] Ian Probert,[6,7] Margaux Carmichael,[1,2,8] Julie Poulain,[9] Sarah Romac,[1,2] Sébastien Colin,[1,2,8] Jean-Marc Aury,[9] Lucie Bittner,[10,11,8,1,2] Samuel Chaffron,[12,13,14] Micah Dunthorn,[3] Stefan Engelen,[9] Olga Flegontova,[15,16] Lionel Guidi,[17,18] Aleš Horák,[15,16] Olivier Jaillon,[9,19,20] Gipsi Lima-Mendez,[12,13,14] Julius Lukeš,[15,16,21] Shruti Malviya,[8] Raphael Morard,[22,1,2] Matthieu Mulot,[5] Eleonora Scalco,[23] Raffaele Siano,[24] Flora Vincent,[13,8] Adriana Zingone,[23] Céline Dimier,[1,2,8] Marc Picheral,[17,18] Sarah Searson,[17,18] Stefanie Kandels-Lewis,[25,26] *Tara* Oceans Coordinators‡ Silvia G. Acinas,[4] Peer Bork,[25,27] Chris Bowler,[8] Gabriel Gorsky,[17,18] Nigel Grimsley,[28,29] Pascal Hingamp,[30] Daniele Iudicone,[23] Fabrice Not,[1,2] Hiroyuki Ogata,[31] Stephane Pesant,[32,22] Jeroen Raes,[12,13,14] Michael E. Sieracki,[33,34] Sabrina Speich,[35,36] Lars Stemmann,[17,18] Shinichi Sunagawa,[25] Jean Weissenbach,[9,19,20] Patrick Wincker,[9,19,20]* Eric Karsenti[26,8]*

Marine plankton support global biological and geochemical processes. Surveys of their biodiversity have hitherto been geographically restricted and have not accounted for the full range of plankton size. We assessed eukaryotic diversity from 334 size-fractionated photic-zone plankton communities collected across tropical and temperate oceans during the circumglobal *Tara* Oceans expedition. We analyzed 18*S* ribosomal DNA sequences across the intermediate plankton-size spectrum from the smallest unicellular eukaryotes (protists, >0.8 micrometers) to small animals of a few millimeters. Eukaryotic ribosomal diversity saturated at ~150,000 operational taxonomic units, about one-third of which could not be assigned to known eukaryotic groups. Diversity emerged at all taxonomic levels, both within the groups comprising the ~11,200 cataloged morphospecies of eukaryotic plankton and among twice as many other deep-branching lineages of unappreciated importance in plankton ecology studies. Most eukaryotic plankton biodiversity belonged to heterotrophic protistan groups, particularly those known to be parasites or symbiotic hosts.

The sunlit surface layer of the world's oceans functions as a giant biogeochemical membrane between the atmosphere and the ocean interior (*1*). This biome includes plankton communities that fix CO_2 and other elements into biological matter, which then enters the food web. This biological matter can be remineralized or exported to the deeper ocean, where it may be sequestered over ecological to geological time scales. Studies of this biome have typically focused on either conspicuous phyto- or zooplankton at the larger end of the organismal size spectrum or microbes (prokaryotes and viruses) at the smaller end. In this work, we studied the taxonomic and ecological diversity of the intermediate size spectrum (from 0.8 μm to a few millimeters), which includes all unicellular eukaryotes (protists) and ranges from the smallest protistan cells to small animals (*2*). The ecological biodiversity of marine planktonic protists has been analyzed using Sanger (*3*–*5*) and high-throughput (*6*, *7*) sequencing of mainly ribosomal DNA (rDNA) gene markers, on relatively small taxonomic and/or geographical scales, unveiling key new groups of phagotrophs (*8*), parasites (*9*), and phototrophs (*10*). We sequenced 18*S* rDNA metabarcodes up to local and global saturations from size-fractionated plankton communities sampled systematically across the world tropical and temperate sunlit oceans.

A global metabarcoding approach

To explore patterns of photic-zone eukaryotic plankton biodiversity, we generated ~766 million raw rDNA sequence reads from 334 plankton samples collected during the circumglobal *Tara* Oceans expedition (*11*). At each of 47 stations, plankton communities were sampled at two water-column depths corresponding to the main hydrographic structures of the photic zone: subsurface mixed-layer waters and the deep chlorophyll maximum (DCM) at the top of the thermocline. A low-shear, nonintrusive peristaltic pump and plankton nets of various mesh sizes were used on board *Tara* to sample and concentrate appropriate volumes of seawater to theoretically recover complete local eukaryotic biodiversity from four major organismal size fractions: piconanoplankton (0.8 to 5 μm), nanoplankton (5 to 20 μm), microplankton (20 to 180 μm), and mesoplankton (180 to 2000 μm) [see (*12*) for detailed *Tara* Oceans field sampling strategy and protocols].

We extracted total DNA from all samples, polymerase chain reaction (PCR)–amplified the hypervariable V9 region of the nuclear gene that encodes 18*S* rRNA (*13*), and generated an average of 1.73 ± 0.65 million sequence reads (paired-end Illumina) per sample (*11*). Strict bioinformatic quality control led to a final data set of 580 million reads, of which ~2.3 million were distinct,

[1]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [2]Sorbonne Universités, Université Pierre et Marie Curie (UPMC) Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [3]Department of Ecology, University of Kaiserslautern, Erwin-Schroedinger Street, 67663 Kaiserslautern, Germany. [4]Department of Marine Biology and Oceanography, Institute of Marine Science (ICM)–Consejo Superior de Investigaciones Científicas (CSIC), Passeig Marítim de la Barceloneta 37-49, Barcelona E08003, Spain. [5]Laboratory of Soil Biology, University of Neuchâtel, Rue Emile-Argand 11, 2000 Neuchâtel, Switzerland. [6]CNRS, FR2424, Roscoff Culture Collection, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Sorbonne Universités, UPMC Paris 06, FR 2424, Roscoff Culture Collection, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [8]Ecole Normale Supérieure, Institut de Biologie de l'ENS (IBENS), and Inserm U1024, and CNRS UMR 8197, Paris, F-75005 France. [9]Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91000 Evry, France. [10]CNRS FR3631, Institut de Biologie Paris-Seine, F-75005, Paris, France. [11]Sorbonne Universités, UPMC Paris 06, Institut de Biologie Paris-Seine, F-75005, Paris, France. [12]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [13]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [14]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [15]Institute of Parasitology, Biology Centre, Czech Academy of Sciences, Branišovská 31, 37005 České Budějovice, Czech Republic. [16]Faculty of Science, University of South Bohemia, Branišovská 31, 37005 České Budějovice, Czech Republic. [17]CNRS, UMR 7093, Laboratoire d'Océanographie de Villefranche-sur-Mer (LOV), Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [18]Sorbonne Universités, UPMC Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [19]CNRS, UMR 8030, CP5706, Evry, France. [20]Université d'Evry, UMR 8030, CP5706, Evry, France. [21]Canadian Institute for Advanced Research, 180 Dundas Street West, Suite 1400, Toronto, Ontario M5G 1Z8, Canada. [22]MARUM, Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany. [23]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [24]Ifremer, Centre de Brest, DYNECO/Pelagos CS 10070, 29280 Plouzané, France. [25]Structural and Computational Biology, European Molecular Biology Laboratory (EMBL), Meyerhofstraße 1, 69117 Heidelberg, Germany. [26]Directors' Research, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [27]Max-Delbrück-Centre for Molecular Medicine, 13092 Berlin, Germany. [28]CNRS UMR 7232, Biologie Intégrative des Organismes Marins (BIOM), Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [29]Sorbonne Universités Paris 06, Observatoire Océanologique de Banyuls (OOB) UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [30]Aix Marseille Université, CNRS IGS UMR 7256, 13288 Marseille, France. [31]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan. [32]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [33]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA. [34]National Science Foundation, Arlington, VA 22230, USA. [35]Department of Geosciences, Laboratoire de Météorologie Dynamique (LMD), Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris Cedex 05, France. [36]Laboratoire de Physique des Océans, Université de Bretagne Occidentale (UBO)–Institut Universitaire Européen de la Mer (IUEM), Place Copernic, 29820 Plouzané, France.

***Corresponding author. E-mail: vargas@sb-roscoff.fr (C.d.V.); pwincker@genoscope.cns.fr (P.W.); karsenti@embl.de (E.K.)**

†These authors contributed equally to this work. ‡*Tara* Oceans Coordinators and affiliations appear at the end of this paper.

hereafter denoted "metabarcodes." We then clustered metabarcodes into biologically meaningful operational taxonomic units (OTUs) (*14*) and assigned a eukaryotic taxonomic path to all metabarcodes and OTUs by global similarity analysis with 77,449 reference, Sanger-sequenced V9 rDNA barcodes covering the known diversity of eukaryotes and assembled into an in-house database called *V9_PR2* (*15*). Beyond taxonomic assignation, we inferred basic trophic and symbiotic ecological modes (photo- versus heterotrophy; parasitism, commensalism, mutualism for both hosts and symbionts) to *Tara* Oceans reads and OTUs on the basis of their genetic affiliation to large monophyletic and monofunctional groups of reference barcodes. We finally inferred large-scale ecological patterns of eukaryotic biodiversity across geography, taxonomy, and organismal size fractions based on rDNA abundance data and community similarity analyses and compared them to current knowledge extracted from the literature.

The extent of eukaryotic plankton diversity in the photic zone of the world ocean

Sequencing of ~1.7 million V9 rDNA reads from each of the 334 size-fractionated plankton samples was sufficient to approach saturation of eukaryotic richness at both local and global scales (Fig. 1, A and B). Local richness represented, on average, 9.7 ± 4% of global richness, the latter approaching saturation at ~2 million eukaryotic metabarcodes or ~110,000 OTUs (*16*). The global pool of OTUs displayed a good fit to the truncated Preston log-normal distribution (*17*), which, by extrapolation, suggests a total photic-zone eukaryotic plankton richness of ~150,000 OTUs, of which ~40,000 were not found in our survey (Fig. 1C). Thus, we estimate that our survey unveiled ~75% of eukaryotic ribosomal diversity in the globally distributed water masses analyzed. The extrapolated ~150,000 total OTUs is much higher than the ~11,200 formally described species of marine eukaryotic plankton (see below) and probably represents a highly conservative, lower-boundary estimate of the true number of eukaryotic species in this biome, given the relatively limited taxonomic resolution power of the 18*S* rDNA gene. Our data indicate that eukaryotic taxonomic diversity is higher in smaller organismal size fractions, with a peak in the piconanoplankton (Fig. 1A), highlighting the richness of tiny organisms that are poorly characterized in terms of morphotaxonomy and physiology (*18*). A first-order, supergroup-level classification of all *Tara* Oceans OTUs demonstrated the prevalence (at the biome scale and across the >four orders of size magnitude sampled) of protist rDNA biodiversity with respect to that of classical multicellular eukaryotes, i.e., animals, plants, and fungi (Fig. 2A). Protists accounted for >85% of total eukaryotic ribosomal diversity, a ratio that may well hold true for other marine, freshwater, and terrestrial oxygenic ecosystems (*19*). The latest estimates of total marine eukaryotic biodiversity based on statistical extrapolations from classical taxonomic knowledge predict the existence of 0.5 to 2.2 million species [including all benthic and planktonic systems from reefs to deep-sea vents (*20*, *21*)] but do not take into account the protistan knowledge gap highlighted here. Simple application of our animal–to–other eukaryotes ratio of ~13% to the robust prediction of the total number of metazoan species from (*20*) would imply that 16.5 million and 60 million eukaryotic species potentially inhabit the oceans and Earth, respectively.

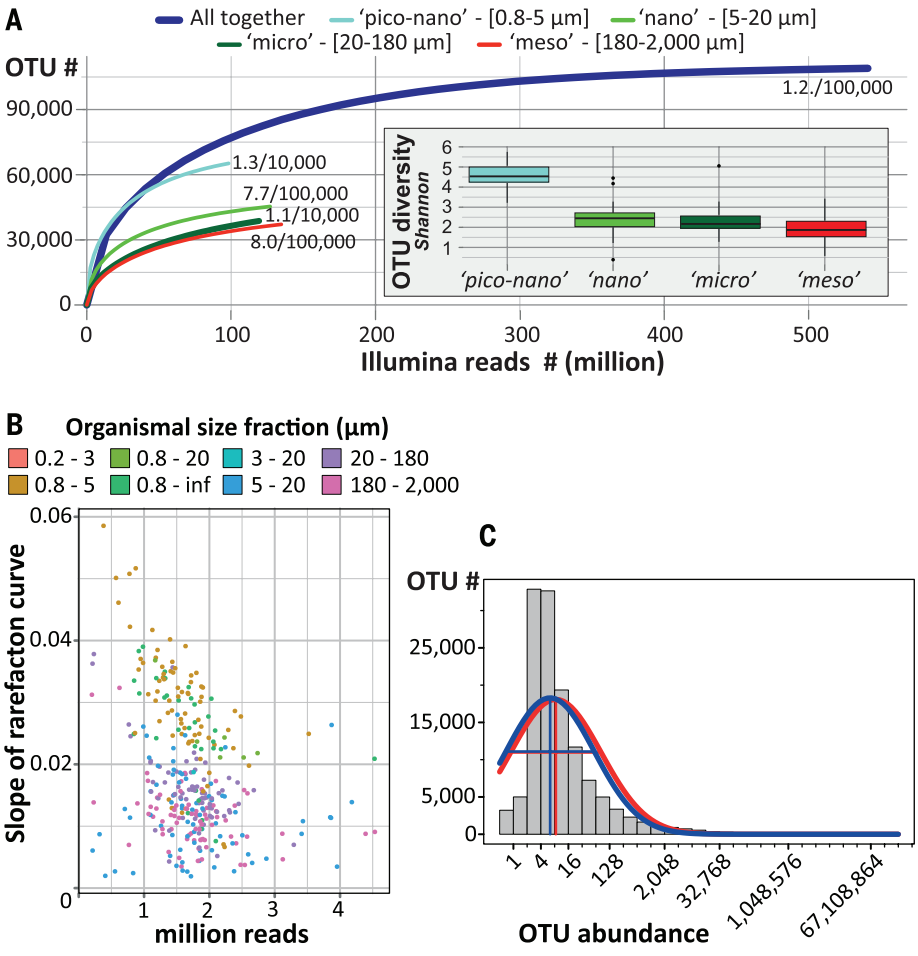


Fig. 1. Photic-zone eukaryotic plankton ribosomal diversity. (**A**) V9 rDNA OTUs rarefaction curves and overall diversity (Shannon index, inset) for each plankton organismal size fraction. Proximity to saturation is indicated by weak slopes at the end of each rarefaction curve (e.g., 1.2/100,000 means 1.2 novel metabarcodes obtained every 100,000 rDNA reads sequenced). (**B**) Saturation slope versus number of V9 rDNA reads for all of the 334 samples (dots) analyzed herein. A slope of 0.02 indicates that two novel barcodes can be recovered if 100 new reads are sequenced. Samples are colored according to size fraction. (**C**) Global OTU abundance distribution and fit to the Preston log-normal model. Most OTUs in our data set were represented by 3 to 16 reads, whereas fewer OTUs presented less or more abundances. Quasi-Poisson fit to octaves (red curve) and maximized likelihood to $\log_2$ abundances (blue curve) approximations were used to fit the OTU abundance distribution to the Preston log-normal model. Overall, the global (A) and local (B) saturation values indicate that our extensive sampling effort (in terms of spatiotemporal coverage and sequencing depth) uncovered the majority of eukaryotic ribosomal diversity within the photic layer of the world's tropical to temperate oceans. Calculation of the Preston veil, which infers the number of OTUs that we missed (or were veiled) during our sampling (~40,000), confirmed that we captured most of the protistan richness, thus allowing extraction of holistic and general patterns of eukaryotic plankton biodiversity from our data set.

Phylogenetic breakdown of photic-zone eukaryotic biodiversity

About one-third of eukaryotic ribosomal diversity in our data set did not match any reference barcode in the extensive *V9_PR2* database ("unassigned" category in Fig. 2A). This unassignable diversity represented only a small proportion (2.6%) of total reads and increased in both richness and abundance in smaller organismal size fractions, suggesting that it corresponds mostly to rare and minute taxa that have escaped previous characterization. Some may also correspond to divergent rDNA pseudogenes, known to exist in eukaryotes (*22*, *23*) or sequencing artefacts (*24*), although both of these would be expected to be present in equal proportion in all

size fractions [details in (*16*)]. The remaining ~87,000 assignable OTUs were classified into 97 deep-branching lineages covering the full spectrum of cataloged eukaryotic diversity amongst the seven recognized supergroups and multiple lineages of uncertain placement (*15*) whose origins go back to the primary radiation of eukaryotic life in the Neoproterozoic. Although highly represented in the *V9_PR2* reference database, several well-known lineages adapted to terrestrial, marine benthic, or anaerobic habitats (e.g., Embryophyta; apicomplexan and trypanosome parasites of land plants and animals; amoeboflagellate Breviatea; and several lineages of Amoebozoa, Excavata, and Cercozoa) were not detected in our metabarcoding data set, suggesting the absence of contamination during the PCR and sequencing steps on land and reducing the number of deep branches of eukaryotic plankton to 85 (Fig. 3).

We then extracted the metabarcodes assigned to morphologically well-known planktonic eukaryotic taxa from our data set and compared them with the conventional, 150 year-old morphological view of marine eukaryotic plankton that includes ~11,200 cataloged species divided into three broad categories: ~4350 species of phytoplankton (microalgae), ~1350 species of protozooplankton (relatively large, often biomineralized, heterotrophic protists), and ~5500 species of metazooplankton (holoplanktonic animals) (*25*–*27*). A congruent picture of the distribution of morphogenetic diversity among and within these organismal categories emerged from our data set (Fig. 2B), but typically, three to eight times more rDNA OTUs were found than described morphospecies in the best-known lineages within these categories. This is within the range of the number of cryptic species typically detected in globally-distributed pelagic taxa using molecular data (*28*, *29*). The general congruency between genetic and morphological data in the cataloged compartment of eukaryotic plankton suggests that the protocols used, from plankton sampling to DNA sequencing, recovered the known eukaryotic biodiversity without major qualitative or quantitative biases. However, OTUs related to morphologically described taxa represented only a minor part of the total eukaryotic plankton ribosomal and phylogenetic diversity. Overall, <1% of OTUs were strictly identical to reference sequences, and OTUs were, on average, only ~86% similar to any V9 reference sequence (Fig. 3F) (*16*). This shows that most photic-zone eukaryotic plankton V9 rDNA diversity had not been previously sequenced from cultured strains, single-cell isolates, or even environmental clone library surveys. The *Tara* Oceans metabarcode data set added considerable phylogenetic information to previous protistan rDNA knowledge, with an estimated mean tree-length increase of 453%, reaching >100% in 43 lineages (*16*). Even in the best-referenced groups such as the diatoms (1232 reference sequences) (Fig. 3B), we identified many new rDNA sequences, both within known groups and forming new clades (*16*).

Eleven "hyperdiverse" lineages each contained >1000 OTUs, together representing ~88 and

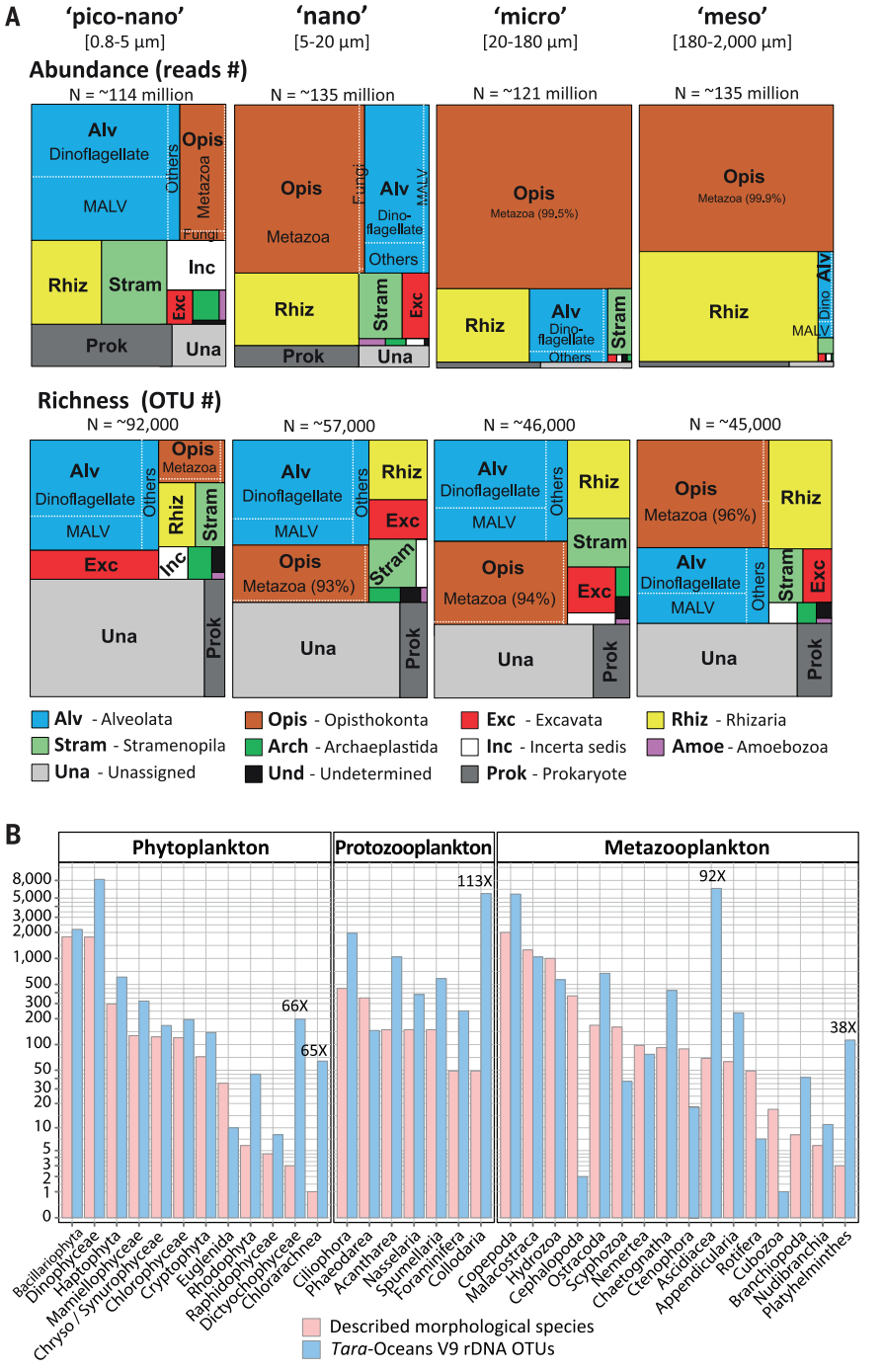


Fig. 2. Unknown and known components of eukaryotic plankton biodiversity. (**A**) Phylogenetic breakdown of the entire metabarcoding data set at the eukaryotic supergroup level. All *Tara* Oceans V9 rDNA reads and OTUs were classified among the seven recognized eukaryotic supergroups plus the known but unclassified deep-branching lineages (incertae sedis). The tree maps display the relative abundance (upper part) and richness (lower part) of the different eukaryotic supergroups in each organismal size fraction. Note that ~5% of barcodes were assigned to prokaryotes, essentially in the piconano fraction, witnessing the universality of the eukaryotic primers used. Barcodes are "unassigned" when sequence similarity to a reference sequence is <80% and "undetermined" when eukaryotic supergroups could not be discriminated (at similarity >80%). (**B**) Ribosomal DNA diversity associated with the morphologically known and cataloged part of eukaryotic plankton. The total number of morphologically described species in the literature [red bars, based on (*25*–*27*)] and the corresponding total number of *Tara* Oceans V9 rDNA OTUs (blue bars) are indicated for each of the 35 classical lineages of eukaryotic phyto-, protozoo-, and metazooplankton. The five classical groups that were found to be substantially more diverse than previously thought (from 38- to 113-fold more OTUs than morphospecies) are highlighted. Note that in the classical morphological view, phyto- and metazooplankton comprise ~88% of total eukaryotic plankton diversity.

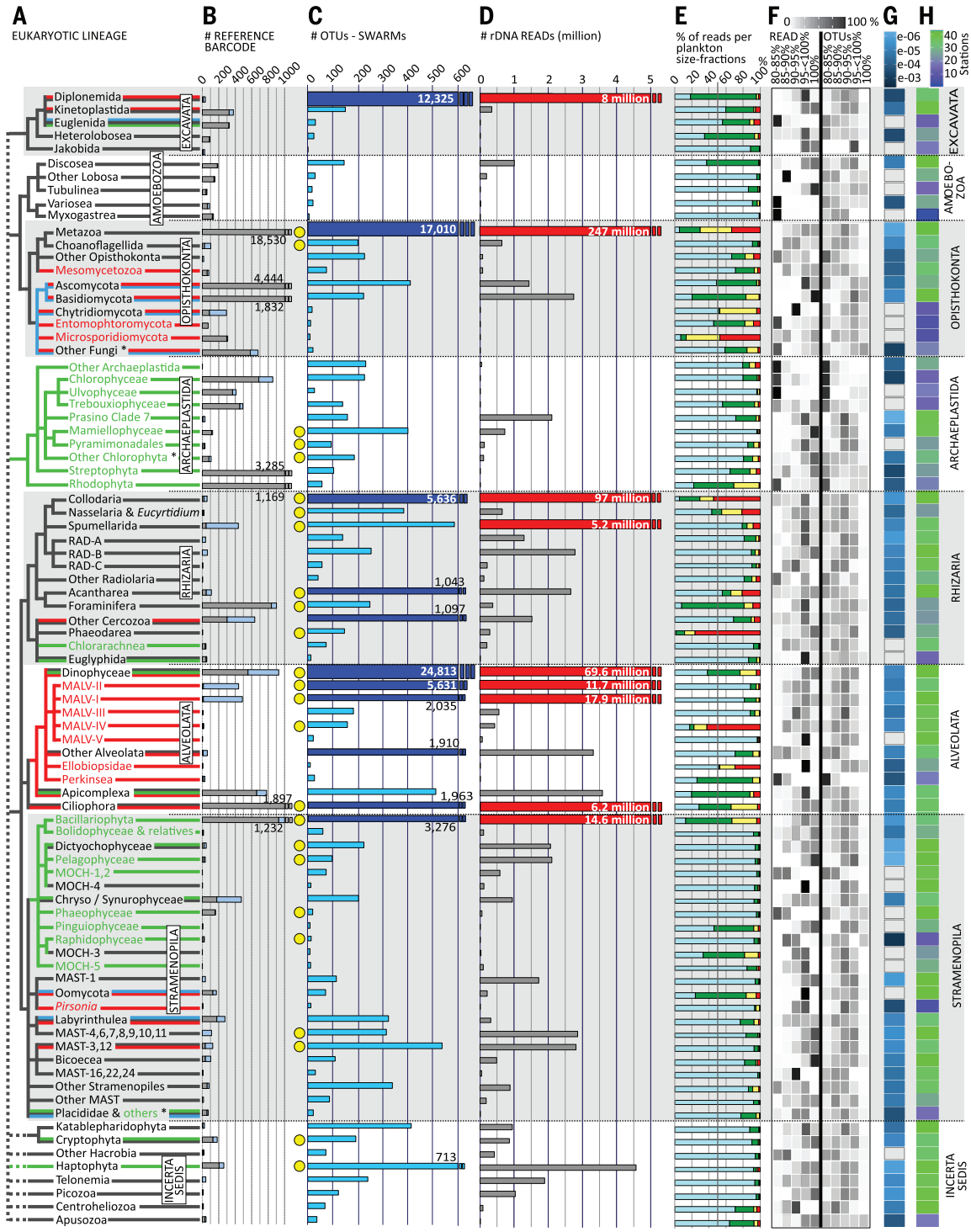


Fig. 3. Phylogenetic distribution of the assignable component of eukaryotic plankton ribosomal diversity. (**A**) Schematic phylogeny of the 85 deep-branching eukaryotic lineages represented in our global oceans metabarcoding data set, with broad ecological traits based on current knowledge: red, parasitic; green, photoautotrophic; blue, osmo- or saprotrophic; black, mostly phagotrophic lineages. Lineages known only from environmental sequence data were colored in black by default. For simplicity, three branches (denoted by asterisks) artificially group a few distinct lineages [details in (*15*)]. (**B**) Number of reference V9 rDNA barcodes used to annotate the metabarcoding data set (gray, with known taxonomy at the genus and/or species level; light blue, from previous 18S rDNA environmental clone libraries). (**C**) *Tara* Oceans V9 rDNA OTU richness. Dark blue thicker bars indicate the 11 hyperdiverse lineages containing >1000 OTUs. Yellow circles highlight the 25 lineages that have been recognized as important in previous marine plankton biodiversity and ecology studies using morphological and/or molecular data [see also (*15*)]. (**D**) Eukaryotic plankton abundance expressed as numbers of rDNA reads (the red bars indicate the nine most abundant lineages with >5 million reads). (**E**) Proportion of rDNA reads per organismal size fraction. Light blue, piconano-; green, nano-; yellow, micro-; red, mesoplankton. (**F**) Percentage of reads and OTUs with 80 to 85%, 85 to 90%, 90 to 95%, 95 to <100%, and 100% sequence similarity to a reference sequence. (**G**) Slope of OTU rarefaction curves. (**H**) Mean geographic occupancy (average number of stations in which OTUs were observed, weighted by OTU abundance).

~90% of all OTUs and reads, respectively (Fig. 3C). Among these, the only permanently phototrophic taxa were diatoms (Fig. 4A) and about one-third of dinoflagellates (Fig. 4, B to F), together comprising ~15 and ~13% of hyperdiverse OTUs and reads, respectively (*30*). Most hyperdiverse photic-zone plankton belonged to three supergroups—the Alveolata, Rhizaria, and Excavata—about which we have limited biological or ecological information. The Alveolata, which consist mostly of parasitic [marine alveolates (MALVs)] (Fig. 4F) and phagotrophic (ciliates and most dinoflagellates) taxa, were by far the most diverse supergroup, comprising ~42% of all assignable OTUs. The Rhizaria are a group of amoeboid heterotrophic protists with active pseudopods displaying a broad spectrum of ecological behavior, from phagotrophy to parasitism and mutualism (symbioses) (*31*). Rhizarian diversity peaked in the Retaria (Fig. 4, C and D) a subgroup including giant protists that build complex skeletons of silicate (Polycystinea), strontium sulfate (Acantharia) (Fig. 4C), or calcium carbonate (Foraminifera) and thus comprise key microfossils for paleoceanography. Unsuspected rDNA diversity was recorded within the Collodaria (5636 OTUs), polycystines that are mostly colonial, poorly silicified, or naked and live in obligatory symbiosis with photosynthetic dinoflagellates (Fig. 4D) (*32*, *33*). Arguably, the most surprising component of novel biodiversity was the >12,300 OTUs related to reference sequences of diplonemids, an excavate lineage that has only two described genera of flagellate grazers, one of which parasitizes diatoms and crustaceans (*34*, *35*). Their ribosomal diversity was not only much higher than that observed in classical plankton groups such as foraminifers, ciliates, or diatoms (50-fold, 6-fold, and 3.8-fold higher, respectively) but was also far from richness saturation (Fig. 3E). Eukaryotic rDNA diversity peaked especially in the few lineages that extend across larger size fractions (i.e., metazoans, rhizarians, dinoflagellates, ciliates, diatoms) (Fig. 3E). Larger cells or colonies not only provide protection against predation via size-mediated avoidance and/or construction of composite skeletons but also provide support for complex and coevolving relationships with often specialized parasites or mutualistic symbionts.

Beyond this hyperdiverse, largely heterotrophic eukaryotic majority, our data set also highlighted the phylogenetic diversity of poorly known phagotrophic (e.g., 413 OTUs of Katablepharidophyta, 240 OTUs of Telonemia), osmotrophic (e.g., 410 OTUs of Ascomycota, 322 OTUs of Labyrinthulea), and parasitic (e.g., 384 OTUs of gregarine apicomplexans, 160 OTUs of Ascetosporea, 68

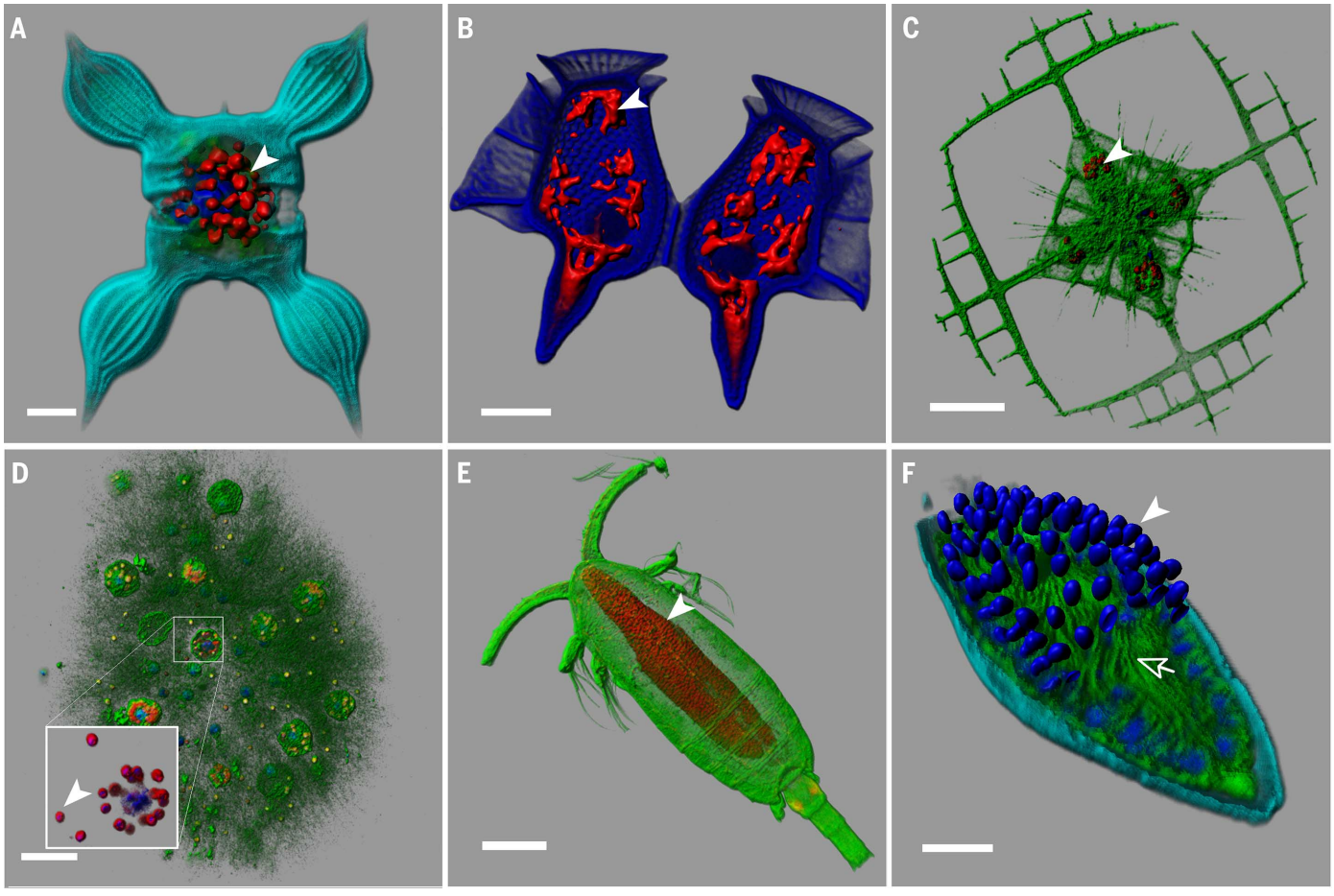


Fig. 4. Illustration of key eukaryotic plankton lineages. (**A**) Stramenopila; a phototrophic diatom *Chaetoceros bulbosus*, with its chloroplasts in red (arrowhead). Scale bar, 10 μm. (**B**) Alveolata; a heterotrophic dinoflagellate *Dinophysis caudata* harboring kleptoplasts [in red (arrowhead)]. Scale bar, 20 μm (*75*). (**C**) Rhizaria; an acantharian *Lithoptera* sp. with endosymbiotic haptophyte cells from the genus *Phaeocystis* [in red (arrowhead)]. Scale bar, 50 μm (*41*). (**D**) Rhizaria; inside a colonial network of Collodaria, a cell surrounded by several captive dinoflagellate symbionts of the genus *Brandtodinium* (arrowhead). Scale bar, 50 μm (*33*). (**E**) Opisthokonta; a copepod whose gut is colonized by the parasitic dinoflagellate *Blastodinium* [red area shows nuclei (arrowhead)]. Scale bar, 100 μm (*51*). (**F**) Alveolata; a cross-sectioned, dinoflagellate cell infected by the parasitoid alveolate *Amoebophrya* (MALV-II). Each blue spot (arrowhead) is the nucleus of future free-living dinospores; their flagella are visible in green inside the mastigocoel cavity (arrow). Scale bar, 5 μm. The cellular membranes were stained with DiOC6 (green); DNA and nuclei were stained with Hoechst (blue) [the dinoflagellate theca in (B) was also stained by this dye]. Chlorophyll autofluorescence is shown in red [except for in (E)]. An unspecific fluorescent painting of the cell surface (light blue) was used to reveal cell shape for (A) and (F). All specimens come from *Tara* Oceans samples preserved for confocal laser scanning fluorescent microscopy. Images were three-dimensionally reconstructed with Imaris (Bitplane).

OTUs of Ichthyosporea) protist groups. Amongst the 85 major lineages presented in the phylogenetic framework of Fig. 3, less than one-third (~25) have been recognized as important in previous marine plankton biodiversity and ecology studies using morphological and/or molecular data (Fig. 3C) (*15*). The remaining ~60 branches had either never been observed in marine plankton or were detected through morphological description of one or a few species and/or the presence of environmental sequences in geographically restricted clone library surveys (*15*). This understudied diversity represents ~25% of all taxonomically assignable OTUs (>21,500) and covers broad taxonomic and geographic scales, thus representing a wealth of new actors to integrate into future plankton systems biology studies.

Insights into photic-zone eukaryotic plankton ecology

Functional annotation of taxonomically assigned V9 rDNA metabarcodes was used as a first attempt to explore ecological patterns of eukaryotic diversity across broad spatial scales and organismal size fractions, focusing on fundamental trophic modes (photo- versus heterotrophy) and symbiotic interactions (parasitism to mutualism). Heterotroph (protists and metazoans) V9 rDNA metabarcodes were substantially more diverse (63%) and abundant (62%) than phototroph metabarcodes that represented <20% of OTUs and reads across all size fractions and geographic sites, with an increasing heterotroph-to-phototroph ratio in the micro- and mesoplankton (Fig. 5A, confirmed in 17 non–size-fractionated samples (*30*). These results challenge the classical morphological view of plankton diversity, biased by a terrestrial ecology approach, whereby phyto- and metazooplankton (the plant-animal paradigm) are thought to comprise ~88% of eukaryotic plankton diversity (Fig. 2B) and heterotrophic protists are typically reduced in food-web modeling to a single entity, often idealized as ciliate grazers.

An unsuspected richness and abundance of metabarcodes assigned to monophyletic groups of heterotrophic protists that cannot survive without endosymbiotic microalgae was found in larger size fractions ("photosymbiotic hosts" in Fig. 5A). Their abundance and even diversity were sometimes greater than those of all metazoan metabarcodes, including those from copepods. Most of these cosmopolitan photosymbiotic hosts were found within the hyperdiverse radiolarians Acantharia (1043 OTUs) and Collodaria (5636 OTUs) (Figs. 3, 4B, and 5D), which have often been overlooked in traditional morphological surveys of plankton-net–collected material because of their delicate gelatinous and/or easily dissolved structures but are known to be very abundant from microscope-based and in situ imaging studies (*36*–*38*). All 95 known colonial collodarian species described since the 19th century (*39*) harbor intracellular symbiotic microalgae, and these key players for plankton ecology are protistan analogs of photosymbiotic corals in tropical coastal reef ecosystems with no equivalent in terrestrial ecology. In addition to their contribution to total primary production (*36*, *38*), these diverse, biologically complex, often biomineralized, and relatively long-lived giant mixotrophic protists stabilize carbon in larger size fractions and probably increase its flux to the ocean interior (*38*). Conversely, the microalgae that are known obligate intracellular partners in open-ocean photosymbioses (*33*, *40*–*42*) (Fig. 5B) were neither very diverse nor highly abundant and occurred evenly across organismal size fractions (Fig. 5C). However, their relative contribution was greatest in the mesoplankton category (10%) (Fig. 5C), where the known photosymbionts of pelagic rhizarians were found (together with their hosts) (Fig. 5B). The stable and systematic abundance of photosymbiotic microalgae across size fractions [a pattern not shown by nonphotosymbiotic microalgae (*30*)] suggests that pelagic photosymbionts maintain free-living and potentially actively growing populations in the piconano- and nanoplankton, representing an accessible pool for recruitment by their heterotrophic hosts. This appears to contrast with photosymbioses in coral reefs and terrestrial systems, where symbiotic microalgal populations mainly occur within their multicellular hosts (*43*).

On the other end of the spectrum of biological interactions, rDNA metabarcodes affiliated to groups of known parasites were ~90 times more diverse than photosymbionts in the piconanoplankton, where they represented ~59% of total heterotrophic protistan ribosomal richness and ~53% of abundance (Figs. 4 and 5C), although this latter value may be inflated by a hypothetically higher rDNA copy number in some marine alveolate lineages (*18*). Parasites in this size fraction were mostly (89% of diversity and 88% of abundance across all stations) within the MALV-I and -II Syndiniales (*30*), which are known exclusively as parasitoid species that kill their hosts and release hundreds of small (2 to 10 μm), nonphagotrophic dinospores (*9*, *44*) that survive for only a few days in the water column (*45*). Abundant parasite-assigned metabarcodes in small size fractions (Fig. 5, B and C) suggest the existence of a large and diverse pool of free-living parasites in photic-zone piconanoplankton, mirroring phage ecology (*46*) and reflecting the extreme diversity and abundance of their known main hosts: radiolarians, ciliates, and dinoflagellates (Fig. 3) (*9*, *47*–*49*). Contrasting with the pattern observed for metabarcodes affiliated to purely phagotrophic taxa, the relative abundance and richness of putative parasite metabarcodes decreased in the nano- and microplanktonic size fractions but increased again in the mesoplankton (Fig. 5C), where parasites are most likely in their infectious stage within larger-sized host organisms. This putative in hospite parasites richness, equivalent to only 23% of that in the piconanoplankton, consisted mostly of a variety of alveolate taxa known to infect crustaceans: MALV-IV such as *Haematodinium* and *Syndinium*; dinoflagellates such as *Blastodinium* (Fig. 4E); and apicomplexan gregarines, mainly *Cephaloidophoroidea* (Fig. 5B) (*9*, *50*, *51*). This pattern contrasts with terrestrial systems where most parasites live within their hosts and are typically transmitted either vertically or through vectors because they generally do not survive outside their hosts (*52*). In the pelagic realm, free-living parasitic spores, like phages, are protected from dessication and dispersed by water diffusion and are apparently massively produced, which likely increases horizontal transmission rate.

Community structuring of photic-zone eukaryotic plankton

Clustering of communities by their compositional similarity revealed the primary influence of organism size ($P = 10^{-3}$, $r^2 = 0.73$) on community structuring, with piconanoplankton displaying stronger cohesiveness than larger organismal size fractions (Fig. 6A). Filtered size-fraction–specific communities separated by thousands of kilometers were more similar in composition than they were to communities from other size fractions at the same location. This was emphasized by the fact that ~36% of all OTUs were restricted to a single size category (*53*). Further analyses within each organismal size fraction indicated that geography plays a role in community structuring, with samples being partially structured according to basin of origin, a pattern that was stronger in larger organismal size fractions ($P = 0.001$ in all cases, $r^2 = 0.255$ for piconanoplankton, 0.371 for nanoplankton, 0.473 for microplankton, and 0.570 for mesoplankton) (Fig. 6B). Mantel correlograms comparing Bray-Curtis community similarity to geographic distances between all samples indicated significant positive correlations in all organismal size fractions over the first ~6000 km, the correlation breaking down at larger geographic distances (*54*). This positive correlation between community dissimilarity and geographic distance, expected under neutral biodiversity dynamics (*55*), challenges the classical niche model for photic-zone eukaryotic plankton biogeography (*56*). The significantly stronger community differentiation by ocean basin in larger organismal size fractions (Fig. 6B) suggests increasing dispersal limitation from piconano- to nano-, micro-, and mesoplankton. Thus, larger-sized eukaryotic plankton communities, containing the highest abundance and diversity of metazoans (Figs. 2A and 5B), were spatially more heterogeneous in terms of both taxonomic (Fig. 6) and functional (Fig. 5A) composition and abundance. The complex life cycle and behaviors of metazooplankton, including temporal reproductive and growth cycles and vertical migrations, together with putative rapid adaptive evolution processes to mesoscale oceanographic features (*57*), may explain the stronger geographic differentiation of mesoplanktonic communities. By contrast, eukaryotic communities in the piconanoplankton were richer (Fig. 1A) and more homogeneous in taxonomic composition (Fig. 6), representing a stable compartment across the world's oceans (*58*).

Even though protistan communities were diverse, the proportions of abundant (>1%) and

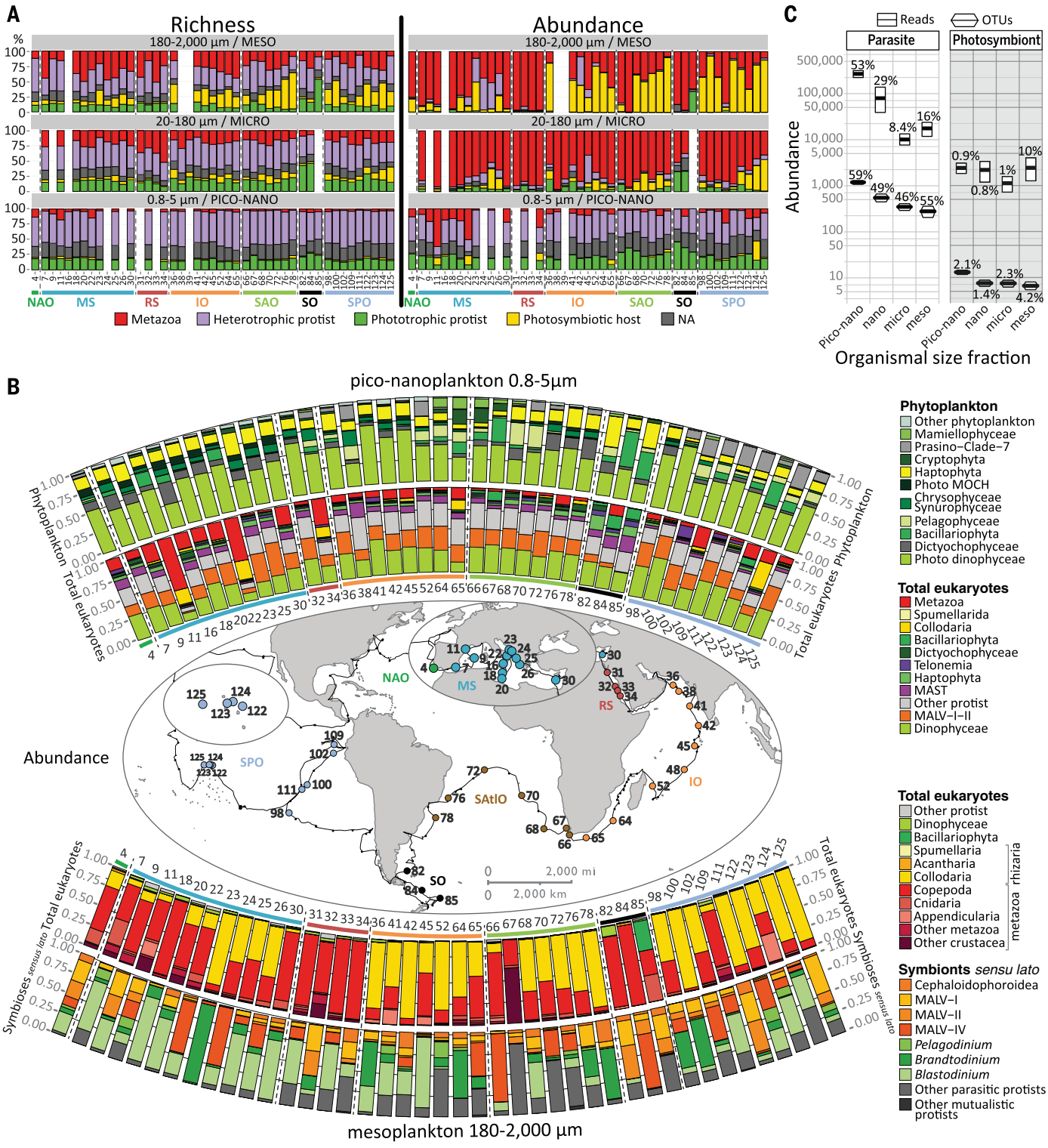


Fig. 5. Metabarcoding inference of trophic and symbiotic ecological diversity of photic-zone eukaryotic plankton. (**A**) Richness (OTU number) and abundance (read number) of rDNA metabarcodes assigned to various trophic taxo-groups across plankton organismal size fractions and stations. Note that the nano size fraction did not contain enough data to be used in this biogeographical analysis [for all size-fraction data, see (*30*)]. NA, not applicable. (**B**) Relative abundance of major eukaryotic taxa across *Tara* Oceans stations for (i) phytoplankton and all eukaryotes in piconanoplankton (above the map) and (ii) all eukaryotes and protistan symbionts (sensu lato) in mesoplankton (below the map). Note the pattern of inverted relative abundance between collodarian colonies (Fig. 4) and copepods in, respectively, the oligotrophic and eutrophic and mesotrophic systems. The dinoflagellates *Brandtodinium* and *Pelagodinium* are endophotosymbionts in Collodaria (*33*) and Foraminifera (*40*, *42*), respectively. (**C**) Richness and abundance of parasitic and photosymbiotic (microalgae) protists across organismal size fractions. The relative contributions (percent) of parasites to total heterotrophic protists and of photosymbionts to total phytoplankton are indicated above each symbol.

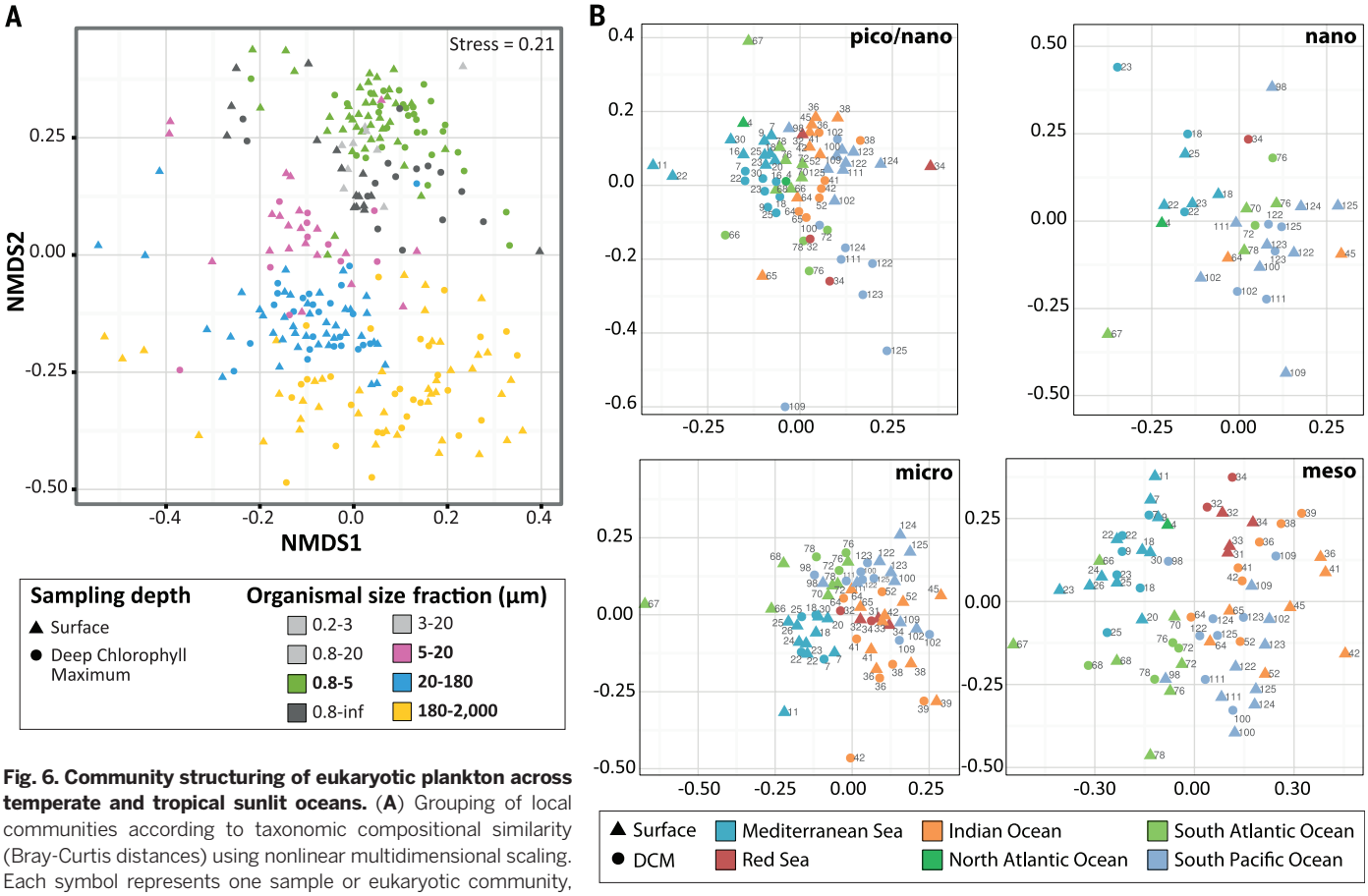


Fig. 6. Community structuring of eukaryotic plankton across temperate and tropical sunlit oceans. (**A**) Grouping of local communities according to taxonomic compositional similarity (Bray-Curtis distances) using nonlinear multidimensional scaling. Each symbol represents one sample or eukaryotic community, corresponding to a particular depth (shape) and organismal size fraction (color). (**B**) Same as in (A), but the different plankton organismal size fractions were analyzed independently, and communities are distinguished by depth (shape) and ocean basins' origin (color). An increasing geographic community differentiation along increasing organismal size fractions is visible and confirmed by the Mantel test [$P = 10^{-3}$, R_m = 0.36, 0.49, 0.50, and 0.51 for the highest piconano- to mesoplankton correlations in Mantel correlograms; see also (*54*)]. In addition, samples from the piconanoplankton only were discriminated by depth (surface versus DCM; P = 0.001, r^2 = 0.2). The higher diversity and abundance of eukaryotic phototrophs in this fraction (Fig. 5A) may explain overall community structuring by light and, thus, depth.

rare (<0.01%) OTUs were more or less constant across communities, as has been observed in coastal waters (*6*). Only 2 to 17 OTUs (i.e., 0.2 to 8% of total OTUs per and across sample) dominated each community (*54*), suggesting that a small proportion of eukaryotic taxa are key for local plankton ecosystem function. On a worldwide scale, an occurrence-versus-abundance analysis of all ~110,000 *Tara* Oceans OTUs revealed the hyperdominance of cosmopolitan taxa (Fig. 7A). The 381 (0.35% of the total) cosmopolitan OTUs represented ~68% of the total number of reads in the data set. Of these, 269 (71%) OTUs had >100,000 reads and accounted for nearly half (48%) of all rDNA reads (Fig. 7A), a pattern reminiscent of hyperdominance in the largest forest ecosystem on Earth, where only 227 tree species out of an estimated total of 16,000 account for half of all trees in Amazonia (*59*). The cosmopolitan OTUs belonged mainly (314 of 381) to the 11 hyperdiverse eukaryotic planktonic lineages (Fig. 3C) and were essentially phagotrophic (40%) or parasitic (21%), with relatively few (15%) phytoplanktonic taxa (*54*). Of the cosmopolitan OTUs, which represent organisms that are likely among the most abundant eukaryotes on Earth, 25% had poor identity (<95%) to reference taxa, and 11 of these OTUs could not even be affiliated to any available reference sequence (Fig. 7B) (*54*).

Conclusions and perspectives

We used rDNA sequence data to explore the taxonomic and ecological structure of total eukaryotic plankton from the photic oceanic biome, and we integrated these data with existing morphological knowledge. We found that eukaryotic plankton are more diverse than previously thought, especially heterotrophic protists, which may display a wide range of trophic modes (*60*) and include an unsuspected diversity of parasites and photosymbiotic taxa. Dominance of unicellular heterotrophs in plankton ecosystems likely emerged at the dawn of the radiation of eukaryotic cells, together with arguably their most important innovation: phagocytosis. The onset of eukaryophagy in the Neoproterozoic (*61*) probably led to adaptive radiation in heterotrophic eukaryotes through specialization of trophic modes and symbioses, opening novel serial biotic ecological niches. The extensive codiversification of relatively large heterotrophic eukaryotes and their associated parasites supports the idea that biotic interactions, rather than competition for resources and space (*62*), are the primary forces driving organismal diversification in marine plankton systems. Based on rDNA, heterotrophic protists may be even more diverse than prokaryotes in the planktonic ecosystem (*63*). Given that organisms in highly diverse and abundant groups, such as the alveolates and rhizarians, can have genomes more complex than those of humans (*64*), eukaryotic plankton may contain a vast reservoir of unknown marine planktonic genes (*65*). Insights are developing into how heterotrophic protists contribute to a multilayered and integrated ecosystem. The protistan parasites and mutualistic symbionts increase connectivity and complexity of pelagic food webs (*66*, *67*) while contributing to the carbon quota of their larger, longer-lived, and often biomineralized symbiotic hosts, which themselves contribute to carbon export when they die. Decoding the ecological and evolutionary rules governing plankton diversity remains essential for understanding how the

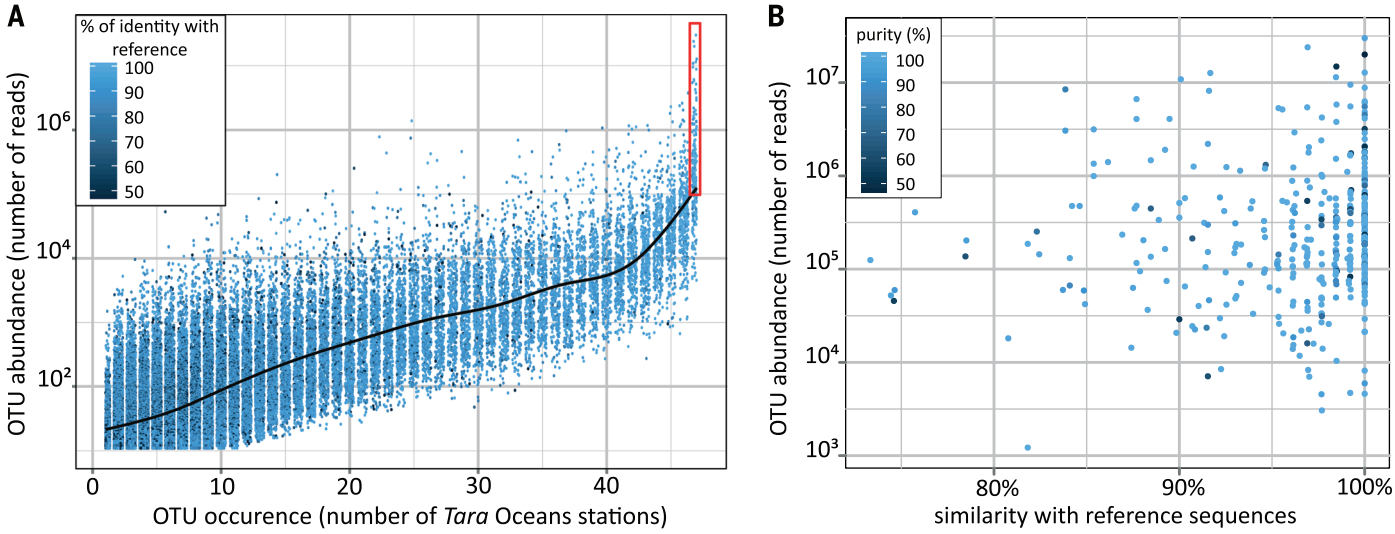


Fig. 7. Cosmopolitanism and abundance of eukaryotic marine plankton. (**A**) Occurrence-versus-abundance plot including the ~110,000 *Tara* Oceans V9 rDNA OTUs. OTUs are colored according to their identity with a reference sequence, and a fitted curve indicates the median OTU size value for each OTU geographic occurrence value. The red rectangle encloses the cosmopolitan and hyperdominant (>10^5 reads) OTUs. (**B**) Similarity to reference barcode and taxonomic purity [a measure of taxonomic assignment consistency defined as the percentage of reads within an OTU assigned to the same taxon; see (*13*)] of the 381 cosmopolitan OTUs, along their abundance (*y* axis).

critical ocean biomes contribute to the functioning of the Earth system.

Materials and methods

V9-18S rDNA for eukaryotic metabarcoding

We used universal eukaryotic primers (*68*) to PCR-amplify (25 cycles in triplicate) the V9-18*S* rDNA genes from all *Tara* Oceans samples. This barcode presents a combination of advantages for addressing general questions of eukaryotic biodiversity over extensive taxonomic and ecological scales: (i) It is universally conserved in length (130 ± 4 base pairs) and simple in secondary structure, thus allowing relatively unbiased PCR amplification across eukaryotic lineages followed by Illumina sequencing. (ii) It includes both stable and highly variable nucleotide positions over evolutionary time frames, allowing discrimination of taxa over a substantial phylogenetic depth. (iii) It is extensively represented in public reference databases across the eukaryotic tree of life, allowing taxonomic assignment among all known eukaryotic lineages (*13*).

Biodiversity analyses

Our bioinformatic pipeline included quality checking (Phred score filtering, elimination of reads without perfect forward and reverse primers, and chimera removal) and conservative filtering (removal of metabarcodes present in less than three reads and two distinct samples). The ~2.3 million metabarcodes (distinct reads) were clustered using an agglomerative, unsupervised single-linkage clustering algorithm, allowing OTUs to reach their natural limits while avoiding arbitrary global clustering thresholds (*13*, *14*). This clustering limited overestimation of biodiversity due to errors in PCR amplification or DNA sequencing, as well as intragenomic polymorphism of rDNA gene copies (*13*). *Tara* Oceans metabarcodes and OTUs were taxonomically assigned by comparison to the 77,449 reference barcodes included in our *V9_PR2* database (*15*). This database derives from the Protist Ribosomal Reference (PR2) database (*69*) but focuses on the V9 region of the gene and includes the following reorganizations: (i) extension of the number of ranks for groups with finer taxonomy (e.g., animals), (ii) expert curation of the taxonomy and renaming in novel environmental groups and dinoflagellates, (iii) resolution of all taxonomic conflicts and inclusion of environmental sequences only if they provide additional phylogenetic information, and (iv) annotation of basic trophic and/or symbiotic modes for all reference barcodes assigned to the genus level [see (*53*) and (*15*) for details]. The *V9_PR2* reference barcodes represent 24,435 species and 13,432 genera from all known major lineages of the tree of eukaryotic life (*15*). Metabarcodes with ≥80% identity to a reference V9 rDNA barcode were considered assignable. Below this threshold it is not possible to discriminate between eukaryotic supergroups, given the short length of V9 rDNA sequences and the relatively fast rate accumulation of substitution mutations in the DNA. In addition to assignment at the finest-possible taxonomic resolution, all assignable metabarcodes were classified into a reference taxonomic framework consisting of 97 major monophyletic groups comprising all known high-rank eukaryotic diversity. This framework, primarily based on a synthesis of protistan biodiversity (*19*), also included all key but still unnamed planktonic clades revealed by previous environmental rDNA clone library surveys (*70*) [e.g., marine alveolates (MALV), marine stramenopiles (MAST), marine ochrophytes (MOCH), and radiolarians (RAD)] (*15*). Details of molecular and bioinformatics methods are available on a companion Web site at http://taraoceans.sb-roscoff.fr/EukDiv/ (*53*). We compiled our data into two databases including the taxonomy, abundance, and size fraction and biogeography information associated with each metabarcode and OTU (*71*).

Ecological inferences

From our *Tara* Oceans metabarcoding data set, we inferred patterns of eukaryotic plankton functional ecology. Based on a literature survey, all reference barcodes assigned to at least the genus level that recruited *Tara* Oceans metabarcodes were associated to basic trophic and symbiotic modes of the organism they come from (*15*) and used for a taxo-functional annotation of our entire metabarcoding data set with the same set of rules used for taxonomic assignation (*53*). False positives were minimized by (i) assigning ecological modes to all individual reference barcodes in *V9_PR2*; (ii) inferring ecological modes to metabarcodes related to monomodal reference barcode(s) (otherwise transferring them to a "NA, nonapplicable" category); and (iii) exploring broad and complex trophic and symbiotic modes that involve fundamental reorganization of the cell structure and metabolism, emerged relatively rarely in the evolutionary history of eukaryotes, and most often concern all known species within monophyletic and ancient groups [see (*15*) for details]. In case of photo- versus heterotrophy, >75% of the major, deep-branching eukaryotic lineages considered (Fig. 3) are monomodal and recruit ~87 and ~69% of all *Tara* Oceans V9 rDNA reads and OTUs, respectively. For parasitism, ~91% of *Tara* Oceans metabarcodes are falling within monophyletic and major groups containing exclusively parasitic species (essentially within the major MALVs groups). Although biases could arise in functional annotation of metabarcodes

relatively distant from reference barcodes in the few complex polymodal groups (e.g., the dinoflagellates that can be phototrophic, heterotrophic, parasitic, or photosymbiotic), a conservative analysis of the trophic and symbiotic ecological patterns presented in Fig. 3, using a ≥99% assignation threshold, shows that these are stable across organismal size fractions and space, independently of the similarity cutoff (80 or 99%), demonstrating their robustness across evolutionary times (*30*).

Note that rDNA gene copy number varies from one to thousands in single eukaryotic genomes (*72*, *73*), precluding direct translation of rDNA read number into abundance of individual organisms. However, the number of rDNA copies per genome correlates positively to the size (*73*) and particularly to the biovolume (*72*) of the eukaryotic cell it represents. We compiled published data from the last ~20 years, confirming the positive correlation between eukaryotic cell size and rDNA copy number across a wide taxonomic and organismal size range [see (*74*); note, however, the ~one order of magnitude of cell size variation for a given rDNA copy number]. To verify whether our molecular ecology protocol preserved this empirical correlation, light microscopy counts of phytoplankton belonging to different eukaryotic supergroups (coccolithophores, diatoms, and dinoflagellates) were performed from nine *Tara* Oceans stations from the Indian, Atlantic, and Southern oceans; transformed into biomass and biovolume data; and then compared with the relative number of V9 rDNA reads found for the identified taxa in the same samples (*74*). Results confirmed the correlation between biovolume and V9 rDNA abundance data (r^2 = 0.97, $P = 1 \times 10^{-16}$), although we cannot rule out the possibility that some eukaryotic taxa may not follow the general trend.

REFERENCES AND NOTES

1. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998). doi: 10.1126/science.281.5374.237; pmid: 9657713
2. D. A. Caron, P. D. Countway, A. C. Jones, D. Y. Kim, A. Schnetzer, Marine protistan diversity. *Annu. Rev. Mar. Sci.* **4**, 467–493 (2012). doi: 10.1146/annurev-marine-120709-142802; pmid: 22457984
3. P. López-García, F. Rodríguez-Valera, C. Pedrós-Alió, D. Moreira, Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**, 603–607 (2001). doi: 10.1038/35054537; pmid: 11214316
4. S. Y. Moon-van der Staay, R. De Wachter, D. Vaulot, Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610 (2001). doi: 10.1038/35054541; pmid: 11214317
5. B. Díez, C. Pedrós-Alió, R. Massana, Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl. Environ. Microbiol.* **67**, 2932–2941 (2001). doi: 10.1128/AEM.67.7.2932-2941.2001; pmid: 11425705
6. R. Logares *et al*., Patterns of rare and abundant marine microbial eukaryotes. *Curr. Biol.* **24**, 813–821 (2014). doi: 10.1016/j.cub.2014.02.050; pmid: 24704080
7. V. Edgcomb *et al*., Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J.* **5**, 1344–1356 (2011). doi: 10.1038/ismej.2011.6; pmid: 21390079
8. R. Massana *et al*., Phylogenetic and ecological analysis of novel marine stramenopiles. *Appl. Environ. Microbiol.* **70**, 3528–3534 (2004). doi: 10.1128/AEM.70.6.3528-3534.2004; pmid: 15184153
9. L. Guillou *et al*., Widespread occurrence and genetic diversity of marine parasitoids belonging to *Syndiniales* (*Alveolata*). *Environ. Microbiol.* **10**, 3349–3365 (2008). doi: 10.1111/j.1462-2920.2008.01731.x; pmid: 18771501
10. H. Liu *et al*., Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 12803–12808 (2009). doi: 10.1073/pnas.0905841106; pmid: 19622724
11. Companion Web site: Figure W1 and Database W1 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
12. S. Pesant *et al*., Open science resources for the discovery and analysis of *Tara* Oceans data. http://biorxiv.org/content/early/2015/05/08/019117 (2015).
13. Companion Web site: Text W1 and Figure W2 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
14. F. Mahé, T. Rognes, C. Quince, C. de Vargas, M. Dunthorn, Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ* **2**, e593 (2014). doi: 10.7717/peerj.593; pmid: 25276506
15. Companion Web site: Database W2, Database W3, and Database W6 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
16. Companion Web site: Text W3, Text W4, Text W5, Figure W4, Figure W5, Figure W6, and Figure W7 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
17. F. W. Preston, The commonness, and rarity, of species. *Ecology* **29**, 254–283 (1948). doi: 10.2307/1930989
18. R. Massana, Eukaryotic picoplankton in surface oceans. *Annu. Rev. Microbiol.* **65**, 91–110 (2011). doi: 10.1146/annurev-micro-090110-102903; pmid: 21639789
19. J. Pawlowski *et al*., CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLOS Biol.* **10**, e1001419 (2012). doi: 10.1371/journal.pbio.1001419; pmid: 23139639
20. C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, B. Worm, How many species are there on Earth and in the ocean? *PLOS Biol.* **9**, e1001127 (2011). doi: 10.1371/journal.pbio.1001127; pmid: 21886479
21. W. Appeltans *et al*., The magnitude of global marine species diversity. *Curr. Biol.* **22**, 2189–2202 (2012). doi: 10.1016/j.cub.2012.09.036; pmid: 23159596
22. L. M. Márquez, D. J. Miller, J. B. MacKenzie, M. J. H. Van Oppen, Pseudogenes contribute to the extreme diversity of nuclear ribosomal DNA in the hard coral *Acropora*. *Mol. Biol. Evol.* **20**, 1077–1086 (2003). doi: 10.1093/molbev/msg122; pmid: 12777522
23. S. R. Santos, R. A. Kinzie III, K. Sakai, M. A. Coffroth, Molecular characterization of nuclear small subunit (18S)-rDNA pseudogenes in a symbiotic dinoflagellate (*Symbiodinium*, Dinophyta). *J. Eukaryot. Microbiol.* **50**, 417–421 (2003). doi: 10.1111/j.1550-7408.2003.tb00264.x; pmid: 14733432
24. J. Decelle, S. Romac, E. Sasaki, F. Not, F. Mahé, Intracellular diversity of the V4 and V9 regions of the 18S rRNA in marine protists (radiolarians) assessed by high-throughput sequencing. *PLOS ONE* **9**, e104297 (2014). doi: 10.1371/journal.pone.0104297; pmid: 25090095
25. A. Sournia, M.-J. Chrétiennot-Dinet, M. Ricard, Marine phytoplankton: How many species in the world ocean? *J. Plankton Res.* **13**, 1093–1099 (1991). doi: 10.1093/plankt/13.5.1093
26. P. H. Wiebe *et al*., Deep-sea sampling on CMarZ cruises in the Atlantic Ocean – An introduction. *Deep-Sea Res. Part II* **57**, 2157–2166 (2010). doi: 10.1016/j.dsr2.2010.09.018
27. D. Boltovskoy, Diversity and endemism in cold waters of the South Atlantic: Contrasting patterns in the plankton and the benthos. *Sci. Mar.* **69**, 17–26 (2005).
28. C. de Vargas, R. Norris, L. Zaninetti, S. W. Gibb, J. Pawlowski, Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 2864–2868 (1999). doi: 10.1073/pnas.96.6.2864; pmid: 10077602
29. K. M. K. Halbert, E. Goetze, D. B. Carlon, High cryptic diversity across the global range of the migratory planktonic copepods *Pleuromamma piseki* and *P. gracilis*. *PLOS ONE* **8**, e77011 (2013). doi: 10.1371/journal.pone.0077011; pmid: 24167556
30. Companion Web site: Figure W8, Figure W9, Figure W10, and Figure W14 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
31. F. Burki, P. J. Keeling, Rhizaria. *Curr. Biol.* **24**, R103–R107 (2014). doi: 10.1016/j.cub.2013.12.025; pmid: 24502779
32. N. R. Swanberg, thesis, Massachussetts Institute of Technology (1974).
33. I. Probert *et al*., *Brandtodinium* gen. nov. and *B. nutricula* comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with polycystine radiolarians. *J. Phycol.* **50**, 388–399 (2014). doi: 10.1111/jpy.12174
34. S. von der Heyden, E. E. Chao, K. Vickerman, T. Cavalier-Smith, Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of euglenozoa. *J. Eukaryot. Microbiol.* **51**, 402–416 (2004). doi: 10.1111/j.1550-7408.2004.tb00387.x; pmid: 15352322
35. E. Schnepf, Light and electron microscopical observations in *Rhynchopus coscinodiscivorus* spec. nov., a Colorless, phagotrophic euglenozoon with concealed flagella. *Arch. Protistenkd.* **144**, 63–74 (1994). doi: 10.1016/S0003-9365(11)80225-3
36. M. R. Dennett, Video plankton recorder reveals high abundances of colonial Radiolaria in surface waters of the central North Pacific. *J. Plankton Res.* **24**, 797–805 (2002). doi: 10.1093/plankt/24.8.797
37. L. Stemmann *et al*., Global zoogeography of fragile macrozooplankton in the upper 100-1000 m inferred from the underwater video profiler. *ICES J. Mar. Sci.* **65**, 433–442 (2008). doi: 10.1093/icesjms/fsn010
38. A. F. Michaels, D. A. Caron, N. R. Swanberg, F. A. Howse, C. M. Michaels, Planktonic sarcodines (Acantharia, Radiolaria, Foraminifera) in surface waters near Bermuda: Abundance, biomass and vertical flux. *J. Plankton Res.* **17**, 131–163 (1995). doi: 10.1093/plankt/17.1.131
39. E. Haeckel, "Report on the Radiolaria collected by H.M.S. Challenger during the years 1873-1876" in *Report on the Scientific Results of the Voyage of* H.M.S. Challenger *During the Years 1873-76. Zoology.* (Neill, Edinburgh, 1887).
40. R. Siano, M. Montresor, I. Probert, F. Not, C. de Vargas, *Pelagodinium* gen. nov. and *P. béii* comb. nov., a dinoflagellate symbiont of planktonic foraminifera. *Protist* **161**, 385–399 (2010). doi: 10.1016/j.protis.2010.01.002; pmid: 20149979
41. J. Decelle *et al*., An original mode of symbiosis in open ocean plankton. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 18000–18005 (2012). doi: 10.1073/pnas.1212303109; pmid: 23071304
42. Y. Shaked, C. de Vargas, Pelagic photosymbiosis: rDNA assessment of diversity and evolution of dinoflagellate symbionts and planktonic foraminiferal hosts. *Mar. Ecol. Prog. Ser.* **325**, 59–71 (2006). doi: 10.3354/meps325059
43. J. Decelle, New perspectives on the functioning and evolution of photosymbiosis in plankton: Mutualism or parasitism? *Commun. Integr. Biol.* **6**, e24560 (2013). doi: 10.4161/cib.24560; pmid: 23986805
44. R. Siano *et al*., Distribution and host diversity of Amoebophryidae parasites across oligotrophic waters of the Mediterranean Sea. *Biogeosciences* **8**, 267–278 (2011). doi: 10.5194/bg-8-267-2011
45. D. Coats, M. Park, Parasitism of photosynthetic dinoflagellates by three strains of *Amoebophyra* (Dinophyta): Parasite survival, infectivity, generation time, and host specificity. *J. Phycol.* **528**, 520–528 (2002). doi: 10.1046/j.1529-8817.2002.01200.x
46. K. E. Wommack, R. R. Colwell, Virioplankton: Viruses in aquatic ecosystems. *Microbiol. Mol. Biol. Rev.* **64**, 69–114 (2000). doi: 10.1128/MMBR.64.1.69-114.2000; pmid: 10704475
47. A. Skovgaard, Dirty tricks in the plankton: Diversity and role of marine parasitic protists. *Acta Protozool.* **53**, 51–62 (2014).
48. J. Bråte *et al*., Radiolaria associated with large diversity of marine alveolates. *Protist* **163**, 767–777 (2012). doi: 10.1016/j.protis.2012.04.004; pmid: 22658831
49. T. R. Bachvaroff, S. Kim, L. Guillou, C. F. Delwiche, D. W. Coats, Molecular diversity of the syndinean genus *Eudubosquella* based on single-cell PCR analysis. *Appl. Environ. Microbiol.* **78**, 334–345 (2012). doi: 10.1128/AEM.06678-11; pmid: 22081578
50. S. Rueckert, T. G. Simdyanov, V. V. Aleoshin, B. S. Leander, Identification of a divergent environmental DNA sequence clade using the phylogeny of gregarine parasites (Apicomplexa) from crustacean hosts. *PLOS ONE* **6**, e18163 (2011). doi: 10.1371/journal.pone.0018163; pmid: 21483868
51. A. Skovgaard, S. A. Karpov, L. Guillou, The parasitic dinoflagellates *Blastodinium* spp. inhabiting the gut of marine, planktonic copepods: Morphology, ecology, and unrecognized species diversity. *Front. Microbiol.* **3**, 305 (2012). doi: 10.3389/fmicb.2012.00305; pmid: 22973263
52. H. McCallum *et al*., Does terrestrial epidemiology apply to marine systems? *Trends Ecol. Evol.* **19**, 585–591 (2004). doi: 10.1016/j.tree.2004.08.009

53. Companion Web site: detailed Material and Methods, Database W9, and Figure W11 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
54. Companion Web site: Figure W12, Figure W13, Database W7, and Database W8 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
55. M. Holyoak, M. A. Leibold, R. D. Holt, Eds., *Metacommunities: Spatial Dynamics and Ecological Communities* (University of Chicago Press, Chicago, 2005).
56. L. G. M. Baas Becking, *Geobiologie of Inleiding tot de Milieukunde* (W. P. Van Stockum and Zoon, The Hague, Netherlands, 1934).
57. K. T. C. A. Peijnenburg, E. Goetze, High evolutionary potential of marine zooplankton. *Ecol. Evol.* **3**, 2765–2781 (2013). doi: 10.1002/ece3.644; pmid: 24567838
58. V. Smetacek, Microbial food webs. The ocean's veil. *Nature* **419**, 565 (2002). doi: 10.1038/419565a; pmid: 12374956
59. H. ter Steege *et al*., Hyperdominance in the Amazonian tree flora. *Science* **342**, 1243092 (2013). doi: 10.1126/science.1243092; pmid: 24136971
60. D. Vaulot, K. Romari, F. Not, Are autotrophs less diverse than heterotrophs in marine picoplankton? **10**, 266–267 (2002).
61. A. H. Knoll, Paleobiological perspectives on early eukaryotic evolution. *Cold Spring Harb. Perspect. Biol.* **6**, 1–14 (2014). doi: 10.1101/cshperspect.a016121; pmid: 24384569
62. V. Smetacek, A watery arms race. *Nature* **411**, 745 (2001). doi: 10.1038/35081210; pmid: 11459035
63. S. Sunagawa *et al*., Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
64. M. J. Oliver, D. Petrov, D. Ackerly, P. Falkowski, O. M. Schofield, The mode and tempo of genome size evolution in eukaryotes. *Genome Res.* **17**, 594–601 (2007). doi: 10.1101/gr.6096207; pmid: 17420184
65. H. Abida *et al*., Bioprospecting marine plankton. *Mar. Drugs* **11**, 4594–4611 (2013). doi: 10.3390/md11114594; pmid: 24240981
66. G. Lima-Mendez *et al*., Determinants of community structure in the global plankton interactome. *Science* **348**, 1262073 (2015).
67. K. D. Lafferty, A. P. Dobson, A. M. Kuris, Parasites dominate food web links. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11211–11216 (2006). doi: 10.1073/pnas.0604755103; pmid: 16844774
68. L. A. Amaral-Zettler, E. A. McCliment, H. W. Ducklow, S. M. Huse, A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PLOS ONE* **4**, e6372 (2009). doi: 10.1371/journal.pone.0006372; pmid: 19633714
69. L. Guillou *et al*., The Protist Ribosomal Reference database (PR^2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* **41**, D597–D604 (2013). doi: 10.1093/nar/gks1160; pmid: 23193267
70. R. Massana, J. del Campo, M. E. Sieracki, S. Audic, R. Logares, Exploring the uncultured microeukaryote majority in the oceans: Reevaluation of ribogroups within stramenopiles. *ISME J.* **8**, 854–866 (2014). doi: 10.1038/ismej.2013.204; pmid: 24196325
71. Companion Web site: Database W4 and Database W5 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
72. A. Godhe *et al*., Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl. Environ. Microbiol.* **74**, 7174–7182 (2008). doi: 10.1128/AEM.01298-08; pmid: 18849462
73. F. Zhu, R. Massana, F. Not, D. Marie, D. Vaulot, Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol. Ecol.* **52**, 79–92 (2005). doi: 10.1016/j.femsec.2004.10.006; pmid: 16329895
74. Companion Web site: Text W2 and Figure W3 (available at http://taraoceans.sb-roscoff.fr/EukDiv/).
75. M. Kim, S. Nam, W. Shin, D. W. Coats, M. Park, *Dinophysis caudata* (dinophyceae) sequesters and retains plastids from the mixotrophic ciliate prey *Mesodinium rubrum*. *J. Phycol.* **48**, 569–579 (2012). doi: 10.1111/j.1529-8817.2012.01150.x

ACKNOWLEDGMENTS

We thank the following people and sponsors for their commitment: CNRS (in particular, the GDR3280); EMBL; Genoscope/CEA; UPMC; VIB; Stazione Zoologica Anton Dohrn; UNIMIB; Rega Institute; KU Leuven; Fund for Scientific Research – The French Ministry of Research, the French Government "Investissements d'Avenir" programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), and MEMO LIFE (ANR-10-LABX-54); PSL* Research University (ANR-11-IDEX-0001-02); ANR (projects POSEIDON/ANR-09-BLAN-0348, PROMETHEUS/ANR-09-PCS-GENM-217, PHYTBACK/ANR-2010-1709-01, and TARA-GIRUS/ANR-09-PCS-GENM-218); EU FP7 (MicroB3/No.287589, IHMS/HEALTH-F4-2010-261376); European Research Council Advanced Grant Awards to C. Bowler (Diatomite:294823); Gordon and Betty Moore Foundation grant 3790 to M.B.S.; Spanish Ministry of Science and Innovation grant CGL2011-26848/BOS MicroOcean PANGENOMICS and TANIT (CONES 2010-0036) grant from the Agency for Administration of University and Research Grants (AGAUR) to S.G.A.; and Japan Society for the Promotion of Science KAKENHI grant 26430184 to H.O. We also thank the following for their support and commitment: A. Bourgois, E. Bourgois, R. Troublé, Région Bretagne, G. Ricono, the Veolia Environment Foundation, Lorient Agglomération, World Courier, Illumina, the Electricité de France Foundation, Fondation pour la Recherche sur la Biodiversité, the Prince Albert II de Monaco Foundation, and the *Tara* schooner and its captains and crew. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org). We also acknowledge assistance from European Bioinformatics Institute (EBI) (in particular, G. Cochrane and P. ten Hoopen) as well as the EMBL Advanced Light Microscopy Facility (in particular, R. Pepperkok). We thank F. Gaill, B. Kloareg, F. Lallier, D. Boltovskoy, A. Knoll, D. Richter, and E. Médard for help and advice on the manuscript. We declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. Data described herein are available at http://taraoceans.sb-roscoff.fr/EukDiv/, at EBI under the project IDs PRJEB402 and PRJEB6610, and at PANGAEA (see table S1). The data release policy regarding future public release of *Tara* Oceans data is described in (*12*). All authors approved the final manuscript. This article is contribution number 24 of *Tara* Oceans. The supplementary materials contain additional data.

Tara Oceans Coordinators

Silvia G. Acinas,[1] Peer Bork,[2] Emmanuel Boss,[3] Chris Bowler,[4] Colomban de Vargas,[5,6] Michael Follows,[7] Gabriel Gorsky,[8,9] Nigel Grimsley,[10,11] Pascal Hingamp,[12] Daniele Iudicone,[13] Olivier Jaillon,[14,15,16] Stefanie Kandels-Lewis,[2,17] Lee Karp-Boss,[3] Eric Karsenti,[4,17] Uros Krzic,[18] Fabrice Not,[5,6] Hiroyuki Ogata,[19] Stephane Pesant,[20,21] Jeroen Raes,[22,23,24] Emmanuel G. Reynaud,[25] Christian Sardet,[26,27] Mike Sieracki,[28] Sabrina Speich,[29,30] Lars Stemmann,[8] Matthew B. Sullivan,[31]* Shinichi Sunagawa,[2] Didier Velayoudon,[32] Jean Weissenbach,[14,15,16] Patrick Wincker[14,15,16]

[1]Department of Marine Biology and Oceanography, ICM-CSIC, Passeig Marítim de la Barceloneta, 37-49, Barcelona E08003, Spain. [2]Structural and Computational Biology, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [3]School of Marine Sciences, University of Maine, Orono, ME 04469, USA. [4]Ecole Normale Supérieure, IBENS, and Inserm U1024, and CNRS UMR 8197, Paris, F-75005 France. [5]CNRS, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [6]Sorbonne Universités, UPMC Paris 06, UMR 7144, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France. [7]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [8]CNRS, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [9]Sorbonne Universités, UPMC Paris 06, UMR 7093, LOV, Observatoire Océanologique, F-06230, Villefranche-sur-Mer, France. [10]CNRS UMR 7232, BIOM, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [11]Sorbonne Universités Paris 06, OOB UPMC, Avenue du Fontaulé, 66650 Banyuls-sur-Mer, France. [12]Aix Marseille Université, CNRS IGS, UMR 7256, 13288 Marseille, France. [13]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy. [14]CEA, Institut de Génomique, GENOSCOPE, 2 rue Gaston Crémieux, 91057 Evry, France. [15]CNRS, UMR 8030, CP5706 Evry, France. [16]Université d'Evry, UMR 8030, CP5706 Evry, France. [17]Directors' Research, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [18]Cell Biology and Biophysics, EMBL, Meyerhofstraße 1, 69117 Heidelberg, Germany. [19]Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. [20]PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany. [21]MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany. [22]Department of Microbiology and Immunology, Rega Institute, KU Leuven, Herestraat 49, 3000 Leuven, Belgium. [23]Center for the Biology of Disease, VIB, Herestraat 49, 3000 Leuven, Belgium. [24]Department of Applied Biological Sciences, Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium. [25]Earth Institute, University College Dublin, Dublin, Ireland. [26]CNRS, UMR 7009 Biodev, Observatoire Océanologique, F-06230 Villefranche-sur-Mer, France. [27]Sorbonne Universités, UPMC Univ Paris 06, UMR 7009 Biodev, F-06230 Observatoire Océanologique, Villefranche-sur-Mer, France. [28]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA. [29]Department of Geosciences, LMD, Ecole Normale Supérieure, 24 rue Lhomond, 75231 Paris, Cedex 05, France. [30]Laboratoire de Physique des Océan UBO-IUEM Place Copernic 29820 Plouzané, France. [31]Department of Ecology and Evolutionary Biology, University of Arizona, 1007 East Lowell Street, Tucson, AZ 85721, USA. [32]DVIP Consulting, Sèvres, France.
*Present address: Department of Microbiology, Ohio State University, Columbus, OH 43210, USA.

SUPPLEMENTARY MATERIALS

www.sciencemag.org/content/348/6237/1261605/suppl/DC1
Table S1
Appendix S1

23 September 2014; accepted 27 February 2015
10.1126/science.1261605

www.sciencemag.org/content/348/6237/1261605/suppl/DC1

Supplementary Materials for

Eukaryotic plankton diversity in the sunlit ocean

Colomban de Vargas,* Stéphane Audic, Nicolas Henry, Johan Decelle, Frédéric Mahé, Ramiro Logares, Enrique Lara, Cédric Berney, Noan Le Bescot, Ian Probert, Margaux Carmichael, Julie Poulain, Sarah Romac, Sébastien Colin, Jean-Marc Aury, Lucie Bittner, Samuel Chaffron, Micah Dunthorn, Stefan Engelen, Olga Flegontova, Lionel Guidi, Aleš Horák, Olivier Jaillon, Gipsi Lima-Mendez, Julius Lukeš, Shruti Malviya, Raphael Morard, Matthieu Mulot, Eleonora Scalco, Raffaele Siano, Flora Vincent, Adriana Zingone, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, *Tara* Oceans Coordinators, Silvia G. Acinas, Peer Bork, Chris Bowler, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Fabrice Not, Hiroyuki Ogata, Stephane Pesant, Jeroen Raes, Michael E. Sieracki, Sabrina Speich, Lars Stemmann, Shinichi Sunagawa, Jean Weissenbach, Patrick Wincker,* Eric Karsenti*

*Corresponding author. E-mail: vargas@sb-roscoff.fr (C.d.V.); pwincker@genoscope.cns.fr (P.W.); karsenti@embl.de (E.K.)

Published 22 May 2015, Science **348**, 1261605 (2015)
DOI: 10.1126/science.1261605

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/348/6237/1261605/suppl/DC1)

Caption for Table S1.

List of all environmental DNA samples analyzed in this study and their metadata. *Sample sequence identifier*: an internal identifier; *Sample label*: a label allowing fast identification of the *Tara* Oceans station, sampling depth, and organismal size-fraction for each genetic sample; *INSDC run accession number:* the identifier under which raw unmatched paired end *Illumina* rDNA sequences have been deposited in public nucleotide databases. *Corresponding nucleotides data published at ENA*: the url to access the corresponding files at the European Bioinformatics Institute (EBI); *PANGAEA sample identifier*: the accession number under which contextual data for this sample is accessible at *PANGAEA* (http://www.pangaea.de); *Corresponding contextual data published at PANGAEA*: the url to access to contextual data for this sample at PANGAEA; *Station identifier [TARA_station#]*: identifier of the *Tara* Oceans station; *Date/Time [yyyy-mm-ddThh:mm]*: date and time of sampling; *Latitude [degrees North]*: latitude of sampling; *Longitude [degrees East]*: longitude of sampling; *Sampling depth [m]*: depth of sampling; *Environmental Feature*: Environmental Ontology description of the sample (http://environmentontology.org/); *Size fraction lower threshold [micrometer]*: lower size limit of the filtering process; *Size fraction upper threshold [micrometer]*: upper size limit of the filtering process; *Marine pelagic biomes (Longhurst 2007) Ocean and sea regions (IHO General Sea Areas 1953) [MRGID registered at www.marineregions.com]*: Ocean and sea region name and identifier; *Marine pelagic biomes (Longhurst 2007) [MRGID registered at www.marineregions.com]*: pelagic biome name and identifier. Further information on the bioinformatic cleaning and filtration process is available in Database W1 (http://taraoceans.sb-roscoff.fr/EukDiv) or in PANGAEA at http://doi.pangaea.de/10.1594/PANGAEA.843017.

Appendix S1. List of curated data sets used in this study and available at PANGAEA

1. Registry of selected samples from the *Tara* Oceans Expedition (2009-2013)

Tara Oceans Consortium, Coordinators; *Tara* Oceans Expedition, Participants (2014) Registry of selected samples from the *Tara* Oceans Expedition (2009-2013). doi:10.1594/PANGAEA.840721

http://doi.pangaea.de/10.1594/PANGAEA.840721

2. Contextual environmental data of selected samples from the *Tara* Oceans Expedition (2009-2013)

Chaffron, S. et al. (2014) Contextual environmental data of selected samples from the *Tara* Oceans Expedition (2009-2013). doi:10.1594/PANGAEA.840718

http://doi.pangaea.de/10.1594/PANGAEA.840718

3. Contextual biodiversity data of selected samples from the *Tara* Oceans Expedition (2009-2013)

Chaffron, S. et al. (2014) Contextual biodiversity data of selected samples from the *Tara* Oceans Expedition (2009-2013). doi:10.1594/PANGAEA.840698

http://doi.pangaea.de/10.1594/PANGAEA.840698

4. Database W1 at http://taraoceans.sb-roscoff.fr/EukDiv/

De Vargas, C. et al. (2015) List of size fractionated eukaryotic plankton community samples and associated metadata (Database W1). doi:10.1594/PANGAEA.843017

http://doi.pangaea.de/10.1594/PANGAEA.843017

5. Database W4 at http://taraoceans.sb-roscoff.fr/EukDiv/

De Vargas, C. et al. (2015) Total V9 rDNA information organized at the metabarcode level (Database W4). doi:10.1594/PANGAEA.843018

http://doi.pangaea.de/10.1594/PANGAEA.843018

6. Database W5 at http://taraoceans.sb-roscoff.fr/EukDiv/

De Vargas, C. et al. (2015) Total V9 rDNA information organized at the OTU level (Database W5). doi:10.1594/PANGAEA.843022

http://doi.pangaea.de/10.1594/PANGAEA.843022

5.2 Paper II

Lukeš J, **Flegontova O**, Horák A. (2015) Diplonemids. ***Current Biology***. 25(16):R702-704 (IF = 8.851).

Abstract

Lukeš et al. introduce an enigmatic group of unicellular eukaryotes called the diplonemids, which according to recent surveys may be widespread in marine ecosystems.

students. A degree was not a prerequisite for doing well in life. Now young people are told they need secondary education if they are to get a good job. Many professors still see their primary educational role as educating students with similar values to themselves in preparation for academic careers. There is a disconnect between what society sees as the role of the faculty and how the faculty see their role. Pressures for education to serve utilitarian ends are decried as a degradation of academic values. Probably more students than ever before, measured as a proportion of the general population, are studying literature and the arts but, rather than celebrating, faculty in these fields are conscious of losing ground relative to other disciplines within the academy. The nature of research has also changed. A hundred years ago, most scientific research was relatively cheap and supported by private or university funds. Faculty did much of the work themselves. Now, expensive research is supported by government funds with benchwork performed by the indentured labor of graduate students and postdocs. The head of laboratory functions as a kind of Chief Executive Officer directing this labor.

With more expected of universities, there are pressures for universities to be more accountable, accompanied by a managerial revolution that seeks objective metrics of productivity in aid of the efficient allocation of resources. The problem with metrics is that they assess what is easy to measure and are rapidly corrupted as individuals modify their behaviors to conform, or to appear to conform, to whatever metric provides material rewards. Activity is easier to measure than thought and counting is quicker than reading. All these requirements eat into the time of the faculty while expanding the size of the managerial class. Universities are seeing the same trends as the broader society, increasing inequality, less time, and a greater proportion of goods expropriated by managers. Advancement of knowledge and education of the young are public goods and extending the reach of the invisible hand may not be the best way to supply these goods.

Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA.
E-mail: dhaig@oeb.harvard.edu

Quick guide

Diplonemids

Julius Lukeš[1,2], Olga Flegontova[1], and Aleš Horák[1]

What are diplonemids and where do they belong? Diplonemids have been classically described as heterotrophic biflagellated unicellular eukaryotes (protists) from the kingdom Euglenozoa (part of the supergroup Excavata), which also contains important pathogens of humans, livestock and plants called kinetoplastids (with *Trypanosoma*, *Leishmania* and *Phytomonas* as the most notorious representatives) and mostly photosynthetic euglenids (represented, for example, by ubiquitous *Euglena*). Compared to these widespread, diverse and important kin, diplonemids were until very recently only rarely found in marine or freshwater environments and only half a dozen species of two genera had been described. Diplonemids are generally considered to be predatory eukaryovores, although parasitic and possibly also symbiotic life strategies are described for some species. The flagship species, *Diplonema papillatum*, is a sack-shaped cell equipped with two short, thin flagella and, together with a few other diplonemid members, is available from American Type Culture Collection.

Honestly, if we were to pick candidates for exciting protists just a few months ago, diplonemids would be at the bottom of our list. Indeed, even specialized protistological textbooks usually devote just a paragraph or two to these obscure flagellates, which have consistently been studied by a single lab, the group of Gertraud Burger in Montreal. But diplonemids recently emerged as one of the most diverse and abundant eukaryotes. And the amazing thing is that we barely know what they look like or what they do. How could such an apparently important group remain totally overlooked for such a long time? The answer lies in the environment they occupy, which is primarily the depths of the ocean.

Are there any molecular features unique to diplonemids? Like their sister group the kinetoplastids, diplonemids harbor a huge mitochondrial genome, composed of thousands of circular DNA molecules, which are either relaxed and interlocked into a single network, or free and supercoiled. We know a lot about mitochondrial RNA editing and processing in the pathogenic *Trypanosoma brucei*, and it seemed likely that similar mechanisms would be in place in related diplonemids. However, diplonemids developed another unique way of dealing with their mitochondrial transcripts. While in *T. brucei* mitochondrial mRNAs are heavily edited by multiple post-transcriptional insertions and/or deletions of uridines, pretty much the same handful of transcripts is processed in a dramatically different manner in *D. papillatum* and *Rhynchopus* spp. No intact full-size gene has ever been found in their mitochondrial genomes, with each circular DNA molecule encoding just a single gene fragment. In a puzzling mechanism, the individual fragments are transcribed and spliced together by an extensive, yet totally uncharacterized *trans*-splicing machinery. By gradual addition of fragments, a mature and translatable molecule is generated. The machinery must be extremely precise, able to pick among dozens of different gene fragments, splicing the neighbors together in an exact manner. This is already a very twisted and unprecedented way of generating transcripts of just about a dozen mitochondrial-encoded genes, yet it is further complicated by limited RNA editing. It can be safely said that so far this is the most baroque example of maturation of any organellar transcript.

What is the real diversity of diplonemids? The environmental sequencing revolution at the turn of this century revealed the existence of two previously unknown yet abundant eukaryotic clades. The first comprises important parasites of plankton related to classic dinoflagellates called Marine Alveolate Group I and II (with five lineages being recognized today). The second group is known as Picozoa (originally picobiliphytes), miniscule heterotrophic flagellates of unclear life strategy. Somewhat in the shadow of these important discoveries, the

analysis of 18S rRNA sequences from the Drake Passage planktonic samples revealed the existence of an environmental clade related to classic diplonemids. This new lineage gradually started expanding with sequences obtained from the mesopelagic to abyssopelagic layers of the Atlantic Ocean and Mediterranean Sea.

The aim of the international Tara Oceans expedition (2009–2012) and Tara Oceans Polar Circle expedition (2013) was a holistic assessment of eukaryotic diversity from planktonic samples collected across the tropical, temperate and polar worlds. Although novel diversity emerged at all taxonomic levels, diplonemids stood out as they represented one of the most diverse and abundant eukaryotic groups (Figure 1A,B). The analysis is based on ~800 million short V9 barcode sequences, a fragment of the 18S rRNA gene that is phylogenetically informative due to its variability. In this survey, diplonemid barcodes constituted the 6th most abundant eukaryotic group in marine plankton (Figure 1A). They were present in the photic layer of all 45 worldwide-distributed sampling stations, but their abundance clearly grew with oceanic depth. According to a detailed analysis of this V9 dataset, the mesopelagic layer, typically ranging from 200 to 1,000 meters, contributed more than 80% of the global diplonemid abundance, and diplonemids comprised up to 58% of all eukaryotic barcodes of the mesopelagic zone at some stations. Such abundance was certainly unexpected, but what is even more surprising is that, from the perspective of sequence diversity, diplonemids rank 3rd only after the well-studied dinoflagellates and metazoans (Figure 1B).

Based on a rather conservative definition of an operational taxonomic unit, the dataset contains ~12,300 diplonemid species. This is a true bonanza given that we have no idea what any of these organisms look like. Importantly, the phylogenies also show that the classic diplonemids from textbooks, for which at least some molecular and morphological data are available, constitute a branch that is a rather distant sister group to this

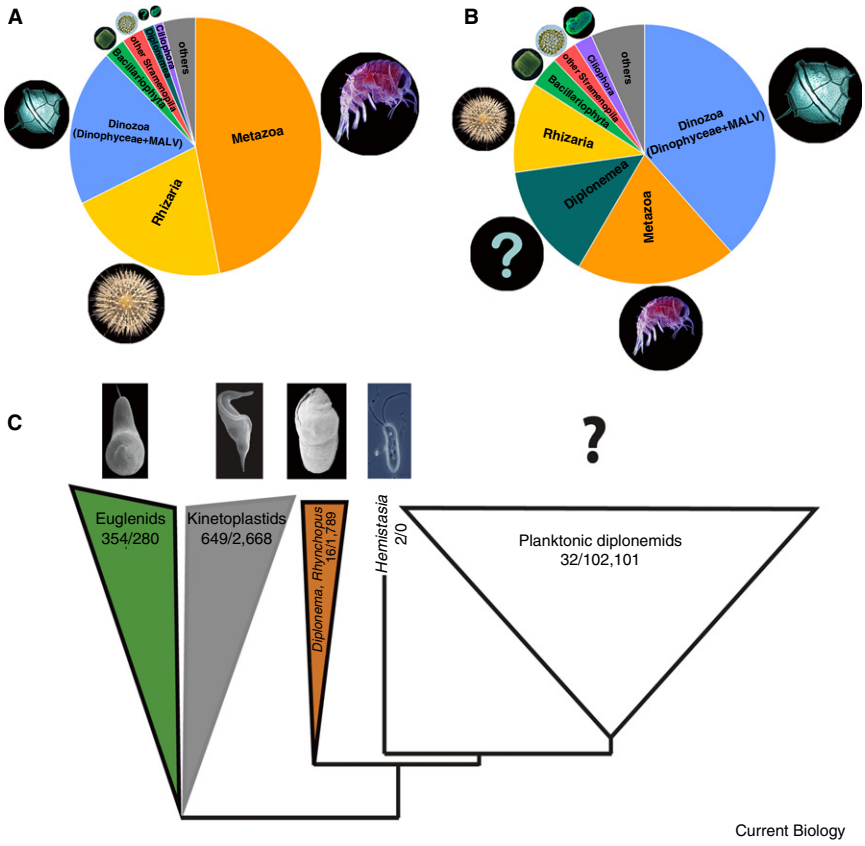


Figure 1. Diplonemid abundance and diversity.
(A) Pie chart showing the 7 most abundant eukaryotic planktonic lineages according to the counts of the V9 sequence, a fragment of the 18S rRNA gene. (B) Pie chart showing the 7 richest eukaryotic planktonic lineages according to operational taxonomic unit (OTU) counts. OTUs are defined with the linkage clustering 'Swarm' approach. (C) Schematized maximum likelihood phylogeny of diplonemid evolution based on V9 sequences. Digits below taxa names show numbers of reference V9 sequences available in public databases/numbers of unique V9 reads as revealed by a global metabarcoding survey of the photic ocean zone.

extremely diverse marine diplonemid-like planktonic clade (Figure 1C).

What is the lifestyle of planktonic diplonemids? The stunning extent of diversity and abundance is based mostly on the V9 barcode of the 18S rRNA gene, which for interspecific comparison appears to be as suitable for diplonemids as for other protists. Hence, we are facing an unusual challenge. Based on sequences, there is a well-defined group of protists in the world's oceans that we know very little about, in particular whether they are free-living, commensals or parasites. Out of the possible life strategies, we could obviously exclude only phototrophy (both from phylogenetic and ecological reasons). A few studies hint to their parasitic lifestyle but if most diplonemids are indeed parasites they would have to infect the majority of marine eukaryotes, likely other protists for the most part. Alternatively, multiple diplonemid species could infect a single host species, but this would contradict the evolutionary trends seen in other parasitic groups, where exploring new hosts is a driving force of speciation.

There are some clues speaking against the parasitic lifestyle of diplonemids. Firstly, preliminary data indicate that the abundance of diplonemids increases with depth, and is still significant in very deep layers of the ocean, down to 5,000 meters, which supports an even less diverse palette of putative hosts. Secondly, an *in silico* analysis of the same global dataset of V9 barcodes from sunlit oceans offers an insight

into possible interactions of planktonic species based on mutual exclusion/co-occurrence of their barcodes. It reveals a plethora of interactions for major marine protist parasites such as syndinians and apicomplexans. In both cases numerous connections tie these parasitic protists with the expected host spectrum. However, even though diplonemids ranked as the 6th most abundant eukaryotic group, they show very little putative interactions with both eukaryotic and prokaryotic components of the plankton community. Thus, the issue of diplonemid lifestyle can only be resolved by obtaining new data, isolating marine diplonemids and analyzing them in the lab.

Is any representative of planktonic diplonemids available in culture? A search for the identity of the planktonic diplonemids and their role in the ocean ecosystem recently yielded an unexpected result with the establishment in culture and concurrent redescription of *Hemistasia phaeocysticola*. Due to the lack of molecular data for the last two decades this euglenozoan ended up with the orphaned *incertae sedis* status. In the currently available extensive 18S rRNA dataset, *Hemistasia* emerged within the robust monophyletic clade of planktonic diplonemids, which constitutes a sister group to the classic diplonemids of the genera *Diplonema* and *Rhynchopus*. Interestingly, *Hemistasia* is a widely distributed, although virtually ignored, predator or parasite of diatoms, dinoflagellates and haptophytes, as well as metazoans, in particular the copepods. Although there is a considerable genetic distance between *Hemistasia* and most planktonic diplonemids (Figure 1C), it is the only available representative of one of the most abundant and diverse marine eukaryotes.

Have diplonemids been sequenced? More than a dozen genomes of pathogenic kinetoplastids (*Trypanosoma*, *Leishmania*, and *Phytomonas* spp.) have been sequenced, but no genomes are published for diplonemids and euglenozoans. We only have estimates that these flagellates carry genomes several times larger than those of the above-mentioned parasites, which range from 20 to 35 Mbp. It will be interesting to find out whether this huge difference is reflected in higher gene number, as it is somewhat counterintuitive that a free-living protist would need fewer genes than its parasitic relative. From the fragmentary information currently available it seems that the common feature of all euglenozoans, namely the addition of a small RNA molecule called spliced leader RNA onto every nuclear transcript, is conserved also in diplonemids.

What should we do next? The study of diplonemids faces two major challenges. While at least one member of the genera *Diplonema* and *Rhynchopus* is available in culture, an easy-to-grow strain representing the hyper-diverse marine clade is much needed. Next, in order to obtain deeper insight into the vagaries of diplonemids, a genetically tractable strain amenable to methods of reverse and forward genetics will have to be generated. The realization of these goals, together with the recent revelations from the TARA expedition, will help rescue the diplonemids from obscurity and bring them into the spotlight.

Where can I learn more?

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., *et al.* (2015). Eukaryotic plankton diversity in the sunlit global ocean. Science *348*, 1261605.

Lara, E., Moreira, D., Vereshchaka, A., and López-García, P. (2009). Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. Environ. Microbiol. *11*, 47–55.

Lima-Mendéz, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J.C., *et al.* (2015). Determinants of community structure in the global plankton interactome. Science *348*, 1262073

Marande, W., and Burger, G. (2007). Mitochondrial DNA as a genomic jigsaw puzzle. Science *318*, 415.

Sturm, N.R., Maslov, D.A., Grisard, E.C., and Campbell, D.A. (2001). *Diplonema* spp. possess spliced leader RNA genes similar to the Kinetoplastida. J. Eukaryot. Microbiol. *48*, 325–331.

Yabuki, A., and Tame, A. (2015). Phylogeny and reclassification of *Hemistasia phaeocysticola* (Scherffel) Elbrächter & Schnepf, 1996. J. Eukaryot. Microbiol. *62*, 426–429.

[1]Biology Centre, Institute of Parasitology, Czech Academy of Sciences and Faculty of Sciences, University of South Bohemia, 37005 České Budějovice (Budweis), Czech Republic. [2]Canadian Institute for Advanced Research, Toronto, ON M5G 1Z8, Canada.

Correspondence

North American velvet ants form one of the world's largest known Müllerian mimicry complexes

Joseph S. Wilson[1,*], Joshua P. Jahner[2], Matthew L. Forister[2], Erica S. Sheehan[1], Kevin A. Williams[3], and James P. Pitts[4]

Color mimicry is often celebrated as one of the most straightforward examples of evolution by natural selection, as striking morphological similarity between species evolves in response to a shared predation pressure [1]. Recently, a large North American mimetic complex was described that included 65 species of *Dasymutilla* velvet ants (Hymenoptera: Mutillidae) [2]. Beyond those 65 species, little is known about how many species participate in this unique Müllerian complex, though several other arthropods are thought to be involved as Müllerian mimics (spider wasps [3]) and Batesian mimics (beetles, antlions, and spiders; see references in [2]). Müllerian mimicry is similarity in appearance or phenotype among harmful species, while Batesian mimicry is similarity in which not all species are harmful. Here, we investigate the extent of the velvet ant mimicry complex beyond *Dasymutilla* by examining distributional and color pattern similarities in all of the 21 North American diurnal velvet ant genera, including 302 of the 361 named species (nearly 84%), as well as 16 polymorphic color forms and an additional 33 undescribed species. Of the 351 species and color forms that were analyzed (including undescribed species), 336 exhibit some morphological similarities and we hypothesize that they form eight distinct mimicry rings (Figure 1A; Supplemental information). Two of these eight mimicry rings, red-headed *Timulla* and black-headed *Timulla*, were not documented in earlier assessments of mimicry in velvet ants [2–4], and are newly described here. These findings identify one of the largest known Müllerian mimicry

5.3 Paper III

Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, Bowler C, Lukeš J, Horák A (2016) Extreme diversity of diplonemid eukaryotes in the ocean. ***Current Biology***. 26(22):3060-3065 (IF = 8.851).

Abstract

The world's oceans represent by far the largest biome, with great importance for the global ecosystem. The vast majority of ocean biomass and biodiversity is composed of microscopic plankton. Recent results from the *Tara* Oceans metabarcoding study revealed that a significant part of the plankton in the upper sunlit layer of the ocean is represented by an understudied group of heterotrophic excavate flagellates called diplonemids. We have analyzed the diversity and distribution patterns of diplonemid populations on the extended set of *Tara* Oceans V9 18S rDNA metabarcodes amplified from 850 size-fractionated plankton communities sampled across 123 globally distributed locations, for the first time also including samples from the mesopelagic zone, which spans the depth from about 200 to 1,000 meters. Diplonemids separate into four major clades, with the vast majority falling into the deep-sea pelagic diplonemid clade. Remarkably, diversity of this clade inferred from metabarcoding data surpasses even that of dinoflagellates, metazoans, and rhizarians, qualifying diplonemids as possibly the most diverse group of marine planktonic eukaryotes. Diplonemids display strong vertical separation between the photic and mesopelagic layers, with the majority of their relative abundance and diversity occurring in deeper waters. Globally, diplonemids display no apparent biogeographic structuring, with a few hyperabundant cosmopolitan operational taxonomic units (OTUs) dominating their communities. Our results suggest that the planktonic diplonemids are among the key heterotrophic players in the largest ecosystem of our biosphere, yet their roles in this ecosystem remain unknown.

Current Biology

Extreme Diversity of Diplonemid Eukaryotes in the Ocean

Highlights

- Pelagic diplonemids are the most diverse planktonic eukaryotes in the ocean
- They are depth stratified and are more abundant and diverse in the deep ocean
- Diplonemids are cosmopolitan, with no clear biogeographic structuring
- They may be key players in the ocean ecosystem, yet their role remains unknown

Authors

Olga Flegontova, Pavel Flegontov, Shruti Malviya, ..., Chris Bowler, Julius Lukeš, Aleš Horák

Correspondence

ogar@paru.cas.cz

In Brief

Flegontova et al. present detailed analysis of global diversity and distribution of diplonemids, the most diverse planktonic eukaryotes. They find diplonemids to be virtually ubiquitous but much more abundant and diverse in the deep ocean. The results suggest that they are among the key players of the ocean ecosystem, yet their role remains unknown.

Flegontova et al., 2016, Current Biology *26*, 3060–3065
November 21, 2016
http://dx.doi.org/10.1016/j.cub.2016.09.031

CellPress

Current Biology
Report

Extreme Diversity of Diplonemid Eukaryotes in the Ocean

Olga Flegontova,[1,2,11] Pavel Flegontov,[1,4,11] Shruti Malviya,[3,11,12] Stephane Audic,[5,6] Patrick Wincker,[7,8,9] Colomban de Vargas,[5,6] Chris Bowler,[3] Julius Lukeš,[1,2,10] and Aleš Horák[1,2,13,*]

[1]Institute of Parasitology, Biology Centre, Czech Academy of Sciences, 37005 České Budějovice, Czech Republic
[2]Faculty of Science, University of South Bohemia, 37005 České Budějovice, Czech Republic
[3]Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), 75005 Paris, France
[4]Faculty of Science, University of Ostrava, 71000 Ostrava, Czech Republic
[5]Station Biologique de Roscoff, 29680 Roscoff, France
[6]Sorbonne Universités, 75005 Paris, France
[7]Genoscope, CEA, 91000 Évry, France
[8]CNRS UMR 8030, 91000 Évry, France
[9]Université d'Evry Val d'Essonne, 91000 Évry, France
[10]Canadian Institute for Advanced Research, Toronto, ON M5G 1Z8, Canada
[11]Co-first author
[12]Present address: Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore 560065, India
[13]Lead Contact
*Correspondence: ogar@paru.cas.cz
http://dx.doi.org/10.1016/j.cub.2016.09.031

SUMMARY

The world's oceans represent by far the largest biome, with great importance for the global ecosystem [1–4]. The vast majority of ocean biomass and biodiversity is composed of microscopic plankton. Recent results from the *Tara* Oceans metabarcoding study revealed that a significant part of the plankton in the upper sunlit layer of the ocean is represented by an understudied group of heterotrophic excavate flagellates called diplonemids [5, 6]. We have analyzed the diversity and distribution patterns of diplonemid populations on the extended set of *Tara* Oceans V9 18S rDNA metabarcodes amplified from 850 size-fractionated plankton communities sampled across 123 globally distributed locations, for the first time also including samples from the mesopelagic zone, which spans the depth from about 200 to 1,000 meters. Diplonemids separate into four major clades, with the vast majority falling into the deep-sea pelagic diplonemid clade. Remarkably, diversity of this clade inferred from metabarcoding data surpasses even that of dinoflagellates, metazoans, and rhizarians, qualifying diplonemids as possibly the most diverse group of marine planktonic eukaryotes. Diplonemids display strong vertical separation between the photic and mesopelagic layers, with the majority of their relative abundance and diversity occurring in deeper waters. Globally, diplonemids display no apparent biogeographic structuring, with a few hyperabundant cosmopolitan operational taxonomic units (OTUs) dominating their communities. Our results suggest that the planktonic diplonemids are among the key heterotrophic players in the largest ecosystem of our biosphere, yet their roles in this ecosystem remain unknown.

RESULTS AND DISCUSSION

Diplonemids, a protist clade with just three genera and about a dozen of species described, were far from the focus of scientific community [5, 7–9]. In the last two decades, however, reports about an environmental clade called deep-sea pelagic diplonemids (DSPDs) and related to the known diplonemids were growing, especially from the deeper waters [10, 11]. Still, diplonemids have evaded a broader recognition until a recent global metabarcoding survey into the diversity of plankton revealed diplonemids as one of the most diverse and abundant eukaryotes of the sunlit ocean [5, 6]. In the present study, we have extended the original *Tara* Oceans metabarcoding dataset based on the V9 region of 18S rDNA [6] with 516 samples including 61 coming from the mesopelagic zone, thus increasing the dataset size 2.5 times. We aim to provide a detailed analysis of patterns of diplonemid diversity and distribution that might help uncover their ecological role. A map of sampling stations is provided in Figure S1; a list of samples and basic sequencing read statistics is in Table S1. The original set of reads was filtered by considering ribotypes present in at least two stations and represented by more than two reads, to avoid potential biases associated with sequencing errors [6], producing a dataset of 24.2 million reads, 289,028 ribotypes, and 45,197 operational taxonomic units (OTUs) assigned to diplonemids (in total, the samples contained 1.15×10^9 reads assigned to eukaryotes; Table S1B).

Phylogeny of Diplonemids

In order to investigate the phylogenetic hierarchy of diplonemids and the distribution of their barcodes among the known diversity,

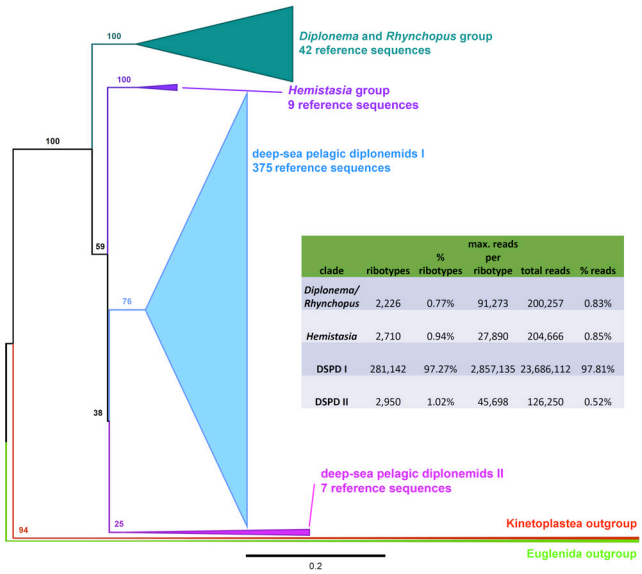


Figure 1. A Maximum-Likelihood Tree Based on 433 Diplonemid 18S rRNA Sequences Longer than 500 bp and Kinetoplastid and Euglenid Outgroups

A maximum-likelihood tree based on 433 diplonemid 18S rRNA sequences (>500 bp) extracted from GenBank with the EukRef approach [12] (http://eukref.org/curation-pipeline-overview/) and kinetoplastid and euglenid outgroups. For reducing the tree size, only seed sequences representing clusters with the 97% identity threshold were included. The tree was constructed with RAxML, the GTR+CAT+I model, and 1,000 rapid bootstrap replicates. Major diplonemid clades and their bootstrap support values are shown: the *Diplonema/Rhynchopus* clade of "classic" diplonemids also observed in deep-sea environments [13, 14], deep-sea pelagic diplonemids II, DSPD II clade [11], the *Hemistasia* clade, and the largest deep-sea pelagic diplonemids I, DSPD I clade [11]. An overwhelming majority of diplonemid metabarcodes from this study falls into the DSPD I clade (see inset). While major diplonemid clades have a moderate or high bootstrap support (from 76 to 100), their branching order is largely unresolved (support from 38 to 59), and the internal topology of the DSPD I clade is especially poorly resolved (data not shown). See also Figure S2.

we created a maximum-likelihood reference tree with exhaustive sampling of all available diplonemid 18S rRNA sequences longer than 500 bp (see Supplemental Experimental Procedures). Diplonemids were recovered as a robust monophyletic clade subdivided into four major lineages (Figure 1). Subsequent mapping of short diplonemid V9 barcodes on the reference tree revealed no novel phylogenetic structuring; i.e., all the barcodes fell into one of the four existing clades.

The vast majority of barcodes and reads alike (~97%) was assigned to the DSPD I clade [11], whereas "classic" diplonemids (*Diplonema* and *Rhynchopus* and associated environmental sequences), the *Hemistasia* clade [9] and the DSPD II clade [11], each accounted for approximately 1% of the observed diversity and abundance. Relationships among these four clades remain poorly resolved (Figure 1), and the branching order within the DSPD I clade also remains largely unresolved. Such a weak structuring of DSPD I clade, as revealed by phylogenetic analysis, combined with the high species richness and relatively low sequence divergence suggests a relatively recent massive radiation.

Diplonemids Are the Most Diverse Planktonic Eukaryotes Abundant in the Deep Ocean

Since their discovery in 2001, DSPDs were found mostly in the deeper oceanic layers [10, 11, 13–16], with reports from the photic zone being rare [11]. Given this focus on deep-sea habitats, the extent of global diplonemid abundance and diversity from the photic zone [6] was highly surprising. Therefore, we included mesopelagic samples and compared the diversity and abundance of diplonemids across all three sampled zones.

With 45,197 diplonemid OTUs found in our extended *Tara* Oceans dataset, diplonemids are the most diverse planktonic eukaryotes. They comprise 19.6% of all eukaryotic OTUs, followed by metazoans (16.1%), dinozoans (15.3%), and rhizarians (9.2%) (Figure 2). These four groups account for ~60% of total eukaryotic diversity of the plankton, and the ranking of clades by diversity was robust in the resampled datasets (Figure 2). On a subset of 334 *Tara* Oceans samples from the photic zone de Vargas et al. [6] have demonstrated that, unlike other hyperdiverse planktonic clades, the diversity of diplonemids is far from saturation. However, the diversity of diplonemids, as well as metazoans, dinozoans, and rhizarians is now saturated in the substantially extended set of 850 *Tara* Oceans samples used in our study, with the slopes of OTU rarefaction curves in the 10^{-7} to 10^{-6} range (Figures 2 and 3).

The relative abundance of diplonemids (diplonemid read count divided by total eukaryotic read count) clearly increased with depth, reaching an average of 14% in the mesopelagic zone versus ~1% in the upper zones, and this difference in abundance was significant according to ANOVA combined with Tukey's honest significance test (Figure S3A). Among 32 stations containing samples from the mesopelagic zone and surface and/or deep chlorophyll maximum (DCM), only one station displayed a higher abundance of diplonemids in the photic layer. Diplonemids were most abundant in the smallest size fraction (0.8–5 μm) (Figure S3B). Reassuringly, we found comparable relative abundance values in nine DNA-RNA sample pairs matched by station and size fraction (average 1.3% versus 1%, ANOVA p value adjusted for multiple testing = 0.70). This result suggests that the reported abundance of diplonemids is not significantly affected by amplification of DNA derived from dead cells. Moreover, DNA and RNA samples did not significantly differ in richness (p value 0.72). Absolute and relative richness of diplonemids (diplonemid OTU count divided by total eukaryotic OTU count) followed the same trends as relative abundance with respect to depth and size fractions (Figures S3C–S3F), and removal of samples treated with whole-genome amplification prior to generation of V9 amplicons (Table S1) did not change the picture (Table S2).

Next, we analyzed the depth and size fraction distribution of the 100 most abundant OTUs (Figure S2), all of which occurred across three depth zones and, remarkably, represented 92.6% of all diplonemid reads. Ninety seven of 100 most abundant OTUs belonged to the DSPD I clade, and just one OTU belonged to each of the other clades (Figure S2). The "classic diplonemid" and *Hemistasia* OTUs occurred mostly in the surface zone, while a single abundant OTU of the DSPD II clade occurred predominantly in the mesopelagic zone. Only six of these most abundant OTUs were found predominantly among

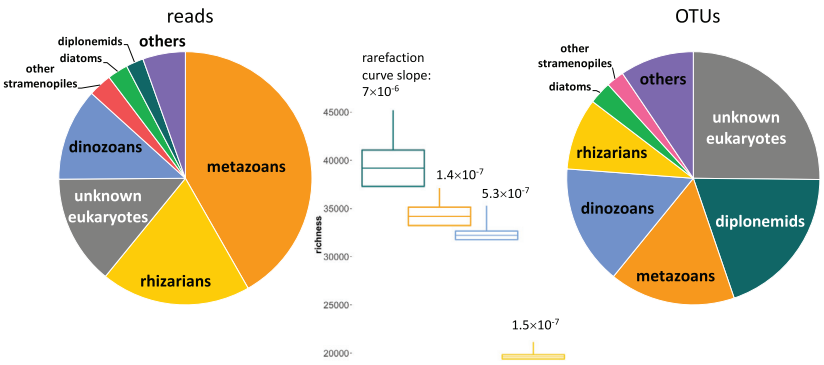


Figure 2. Fractions of Richness and Abundance Corresponding to Six Clades of Planktonic Eukaryotes, which Are Most Diverse in the Extended *Tara* Oceans Dataset

These clades are (1) DSPD I diplonemids; (2) metazoans; (3) dinozoans, which include dinoflagellates and related, mostly parasitic, environmental clades of marine alveolates (MALVs); (4) rhizarians; (5) diatoms; and (6) other stramenopiles. The boxplots in the middle show OTU counts for the four top clades: diplonemids, metazoans, dinozoans, and rhizarians color coded in the same way as in the pie charts. The upper whisker extends up to the OTU count observed in the extended *Tara* Oceans dataset of 850 samples; the crossbar shows a mean OTU count in 1,000 datasets subjected to bootstrapping of samples (see Supplemental Experimental Procedures), and the hinges show SD of the mean. Corresponding slopes of OTU rarefaction curves are shown beside each boxplot.

the meso-plankton fraction (180–2,000 μm), suggesting possible association with very large protists or small metazoans (Figure S2). The remaining OTUs were present primarily in the mesopelagic zone (78 OTUs according to ANOVA), and in the smallest size fractions (up to 20 μm).

The only metabarcoding dataset with a comparably global sampling of deeper oceanic layers has been published only recently [17]. Their analysis of V4 18S rDNA barcodes reveals just 1.5% of excavate (mostly diplonemid) sequences in the bathypelagic layer (depths from 1,000 to 4,000 m), a considerably lower amount compared to the results presented here (14% on average in the mesopelagic zone). Notably, 8% of V4-based OTUs in that study belonged to Excavata (mainly to diplonemids). Unfortunately, the dataset of Pernice et al. [17] differs from ours in many important aspects (bathypelagic versus meso- to epipelagic zones; V4 versus V9 regions of 18S rDNA, different bioinformatics protocols), which does not allow a detailed comparison. According to our pilot analysis, the V4 region of diplonemid 18S rDNAs is about 600 bp or longer (data not shown), while 454 barcodes in the range from 150 to 600 bp were used by Pernice et al. [17]. Diplonemid V4 barcodes might thus have been filtered out at an initial stage of their analysis. Pernice et al. [17] suggested that a negative bias against long amplicons and poor performance of universal V4 primers explains the poor representation of diplonemids among their "pyrotags."

Moreover, Pernice et al. [17] report a significant amount of excavate (and "particularly diplonemid") barcodes among the metagenomic data (10.7% of reads matching 18S rDNA sequences), which suggests diplonemids are a very significant component of plankton even in the deepest oceanic layers. In another study, based on fluorescence in situ hybridization, diplonemids accounted for up to 15% of eukaryotic cells in the bathypelagic zone [18]. We also looked for diplonemid V4 and V9 sequences in the *Tara* Oceans metagenomic dataset, yet since it does not reach the depth and global coverage of the metabarcoding data, such a comparison is not possible at the moment. And the lack of relevant reference genomes makes precise taxonomic binning of the bulk of metagenomic reads unfeasible.

Diplonemid Communities Are Stratified According to Depth

Above, we show that diplonemids are a significant part of surface plankton communities, but their distribution is centered toward the deeper layers. We therefore examined whether diplonemids from the photic zone are different from the mesopelagic ones.

Indeed, the diversity of diplonemids is highly stratified according to depth, with just 1,883 OTUs (4.2%) shared across all three depth zones. A significant fraction of OTUs (16,088; 35.6%) was present exclusively in the mesopelagic zone (Figure 4A), despite only 7.7% samples and 7% eukaryotic reads coming from this zone (Table S1B). This difference in diplonemid community composition is supported also by non-metric multidimensional scaling analysis (Figure 4B) based on pairwise Bray-Curtis distances among samples. Even though the separation is not as clearly pronounced here, mesopelagic samples stand apart from the mixed cluster of surface and DCM samples. It is worth mentioning that a vast majority of strictly depth-specific OTUs was rare, while the most abundant OTUs were cosmopolitan and present at all depths, albeit with largely varying abundance across the depth gradient (76 of 100 most abundant OTUs were distributed predominantly in the mesopelagic) (Figure S2).

Diplonemids Are Cosmopolitan with No Clear Biogeographic Pattern

It is generally accepted that protistan communities are stratified along gradients of biotic and/or abiotic factors, such as light, oxygen concentration, temperature, pressure, salinity, and nutrients [19–21], which predetermine their distribution. The classic dispersal model of microscopic eukaryotes as postulated by Finlay [22] assumes that the immense abundance of individuals is sufficient to overcome geographic barriers of dispersal. This results in ubiquity of most species, also expressed as "everything is everywhere." According to this model, the presence of particular species in a given environment is a function of micro-niche spectrum rather than geographic distance. The model also predicts low global species number and high local richness. However, de Vargas et al. [6] found the numbers of planktonic taxa severely underestimated and showed significant overall correlation between community composition and geographic

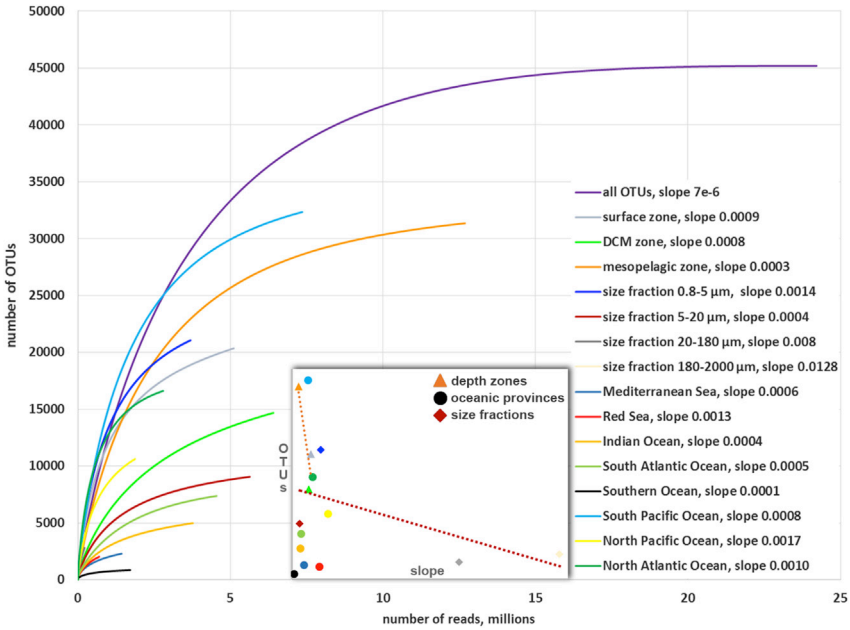


Figure 3. Rarefaction Curves for OTUs

OTU count versus read number. Slopes are indicated in the legend on the right. Curves were constructed for the full dataset, for the depth zone and size fraction subsets, and for the oceanic provinces. The lowest slopes of OTU rarefaction curves were observed in the mesopelagic zone (slope 0.0003), and in the nano-plankton fraction (0.0004). Much higher slope values in two larger size fractions reflect low abundance of diplonemids in the corresponding samples. The piconano-plankton fraction (0.8–5 µm) and the North Pacific and North Atlantic Oceans demonstrate a very high and unsaturated diversity of diplonemids, with slopes in the 10^{-3} range. On the other hand, the diversity in the Southern Ocean is closer to saturation (slope 0.0001) despite a much more limited sampling (Table S1). For depth zones and size fractions diversity saturation tends to increase with richness, but there is no clear trend for oceanic provinces: see the inset showing a plot of total OTU counts versus rarefication curve slopes and trend lines. See also Figures S2 and S3 and Table S1.

distance when considering all eukaryotes together. Dispersal abilities of plankton, especially on the larger side of the size spectrum, seem to be limited by increasing distance.

However, only several studies with global sampling provide compelling evidence of "biogeography" in particular protist groups (see [23] for review). Naturally, the extensive *Tara* Oceans metabarcoding dataset seems to be an ideal tool for testing biogeographic nature of planktonic distribution. First such a case has just recently been reported from diatoms by Malviya et al. [24]. They report a complex biogeographical pattern for diatoms with only a few cosmopolitan ribotypes displaying high abundance and an even distribution across stations. Unlike diatoms, distribution of diplonemids reveals no such a pattern. In general, we found a majority of richness comprising rare OTUs present at less than ten sampling sites (Figure 4C). The more abundant an OTU, the more ubiquitous was its distribution, and most of them occurred in stations with a high evenness statistic (Figure 4C).

In the surface zone, the largest number of OTUs, approximately 6,300, was shared between the South Pacific and North Atlantic Oceans, and the number of OTUs unique to any single oceanic province was low, with the North Atlantic having the highest count (~1,600; Figure S4A). In the DCM zone, however, the South Pacific had by far the highest number of unique OTUs (~2,100) (Figure S4B). Unlike the South Pacific, the other oceanic provinces had marginal counts of unique OTUs in the DCM zone. Similarly, the South Pacific and the South and North Pacific Oceans combined had the highest counts of unique OTUs in the mesopelagic zone (~7,500 and ~4,800, respectively), while the other four provinces had very low counts of unique OTUs (Figure S4C). South Pacific and North Atlantic Oceans thus harbor most of the diplonemid diversity. However, relative abundance and diversity statistics (richness, relative richness, Shannon index, evenness) generally demonstrated no statistically significant differences across oceanic provinces (Figure S1; Table S3).

Diplonemids Are Heterotrophs of Unknown Ecological Role

The diversity, abundance, and ubiquity of DSPD I diplonemids presented above clearly speak for their importance in the global ocean ecosystem. So far, no DSPD I diplonemid has been formally described, and we know nothing about their biology. However, we can try to employ existing data to gain at least indirect evidence about their ecological role.

From the wide range of possible life strategies, phototrophy and mixotrophy can be excluded due to the abundance of diplonemids in the deep ocean. Looking at their closest kins, represented by "classic" diplonemids, *Hemistasia* and the kinetoplastid flagellates, does not provide useful hints because those groups evolved a wide array of life strategies ranging from feeding on bacteria, predation, to parasitism and/or obligate symbiosis [7, 9, 25, 26]. The extreme species richness of the DSPD I diplonemids could stem from their diverse trophic interactions, ranging from bacterivory or parasitism, as seen in related kinetoplastids [7, 24], eukaryovory characteristic for the sister-branching *Hemistasia* [9], or even the so-far-overlooked grazing on viruses [27–30].

Recent results suggest that many species-rich planktonic groups (e.g., within the dinozoans and metazoans) are predominantly parasites [6]. Among the 100 most abundant diplonemid OTUs, six OTUs were mostly present in the largest size fraction and may represent symbionts or parasites of very large protists or small metazoans (Figure S2). An in silico analysis based on mutual exclusion/co-occurrence patterns of barcodes [31] could, in principle, provide an overview of putative parasitic and symbiotic species interactions involving diplonemids.

The species interactome framework is currently available for the partial *Tara* Oceans dataset [6], which includes diplonemid barcodes from the photic zone only. The interactome (http://www.raeslab.org/companion/ocean-interactome.html) contains only 36 diplonemid ribotypes (belonging to 36 OTUs) meeting the stringent inclusion criteria [31] (see Supplemental Experimental

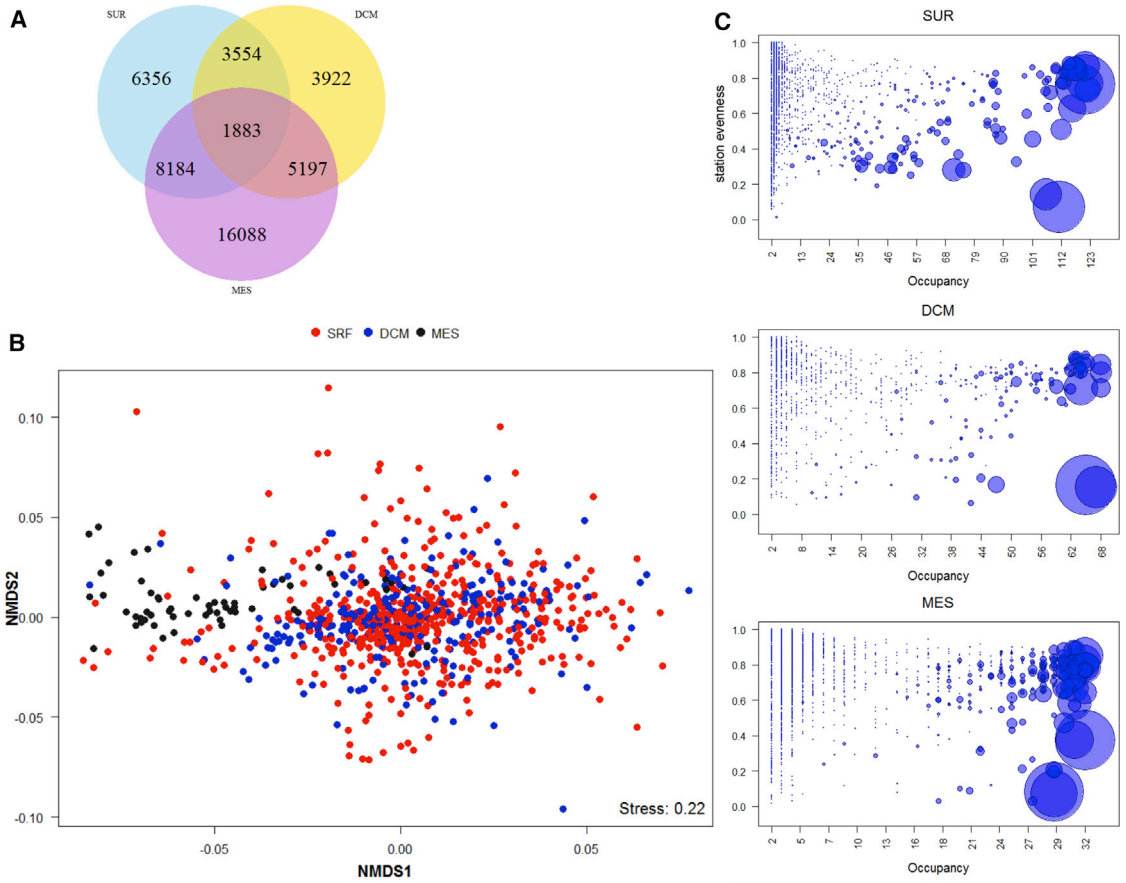


Figure 4. Depth Stratification of Diplonemid Diversity

(A) A Venn diagram of OTUs encountered in different depth zones: SUR, surface; DCM, deep chlorophyll maximum; MES, mesopelagic.

(B) Non-metric multi-dimensional scaling (NMDS) of pairwise Bray-Curtis distances among samples reveals mesopelagic communities as outliers. The depth zones are coded by color and abbreviated as follows: SRF, surface; DCM, deep chlorophyll maximum; MES, mesopelagic.

(C) Cosmopolitan and rare OTUs in three depth zones: SUR, surface; DCM, deep chlorophyll maximum; MES, mesopelagic. Occupancy values, i.e., the number of stations where an OTU was found, are plotted on the x axis, and average station evenness for these stations is plotted on the y axis. Bubble size represents a read count for a given OTU.

See also Figures S1–S4 and Tables S2 and S3.

Procedures for details), with 13 ribotypes belonging to the top 100 abundant OTUs in our dataset. The diplonemid interactome includes 1,008 positive correlations (co-presence of a diplonemid ribotype with another ribotype) and 95 negative correlations (mutual exclusion), and the following clades featured most frequently among positive correlations: parasitic dinoflagellates of the Syndiniales group (235 correlations), bacteria (193 correlations), and parasitic or bacterivorous marine stramenopiles (MAST; 89 correlations). Notably, two diplonemid ribotypes correlated mostly with bacteria (nine of 13 and 58 of 127 interactions). The following clades featured most frequently in negative correlations: crustaceans (23 correlations), dinoflagellates (16 correlations), and radiolarians (ten correlations). This represents a rather poor signal with no obvious pattern compared to a plethora of putative interactions detected for major marine protist parasites such as marine alveolates (MALVs, also known as syndinians) and apicomplexans with their expected host spectra [31]. The enigma of diplonemids' role in the ocean ecosystems could thus be unequivocally resolved only by new data, including single-cell genomics ([32] this issue of *Current Biology*) and transcriptomics, introduction of marine diplonemids into culture, and investigation of their life style in the laboratory.

ACCESSION NUMBERS

The project number for the sequences reported in this paper is EBI: PRJEB16766. The individual accession numbers are EBI: ERS1431784–ERS1432633.

SUPPLEMENTAL INFORMATION

Supplemental Information includes four figures, three tables, and Supplemental Experimental Procedures and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2016.09.031.

AUTHOR CONTRIBUTIONS

A.H., P.F., and J.L. designed the study. P.W., C.d.V., S.A., and C.B. provided the data. O.F., P.F., S.M., S.A., and A.H. performed data analysis. A.H., P.F., O.F., J.L., and C.B. wrote the manuscript.

ACKNOWLEDGMENTS

This work was funded by Czech Grant Agency project P506-12-P9 (to A.H.) and 14-23986S (to J.L.); Gordon and Betty Moore Foundation grant GBMF4983 to J.L.; European Union Framework Programme 7 (MicroB3/No. 287589) and European Research Council Advanced Grant Award (Diatomite: 294823) to C.B.; EU Operational Programme Research and Development for Innovation project CZ.1.05/2.1.00/19.0388, Moravian-Silesian region projects MSK2013-DT1, MSK2013-DT2, MSK2014-DT1, and the Institution Development Program of the University of Ostrava to P.F.; University of South Bohemia grant 04-088/2014/P to O.F.; French Government "Investissements d'Avenir" programs Oceanomics (ANR-11-BTBR-0008), France Génomique (ANR-10-INBS-09-08), Memo Life (ANR-10-LABX-54), Paris Sciences et Lettres (PSL*) Research University (ANR-11-IDEX-0001-02), and Agence Nationale de la Recherche project Prometheus (ANR-09-PCS-GENM-217). This article is contribution #47 of *Tara* Oceans.

Received: June 27, 2016
Revised: September 2, 2016
Accepted: September 15, 2016
Published: November 21, 2016

REFERENCES

1. Daniel, B., and Hain, M.P. (2012). The biological productivity of the ocean. Nat. Educ. *3*, 20.
2. Takao, S., Hirawake, T., Wright, S.W., and Suzuki, K. (2012). Variations of net primary productivity and phytoplankton community composition in the Indian sector of the Southern Ocean as estimated from ocean color remote sensing data. Biogeosciences *9*, 3875–3890.
3. Falkowski, P.G., Fenchel, T., and Delong, E.F. (2008). The microbial engines that drive Earth's biogeochemical cycles. Science *320*, 1034–1039.
4. Field, C.B., Behrenfeld, M.J., Randerson, J.T., and Falkowski, P. (1998). Primary production of the biosphere: Integrating terrestrial and oceanic components. Science *281*, 237–240.
5. Lukeš, J., Flegontova, O., and Horák, A. (2015). Diplonemids. Curr. Biol. *25*, R702–R704.
6. de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., et al.; Tara Oceans Coordinators (2015). Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science *348*, http://dx.doi.org/10.1126/science.1261605.
7. Simpson, A.G.B. (1997). The identity and composition of the Euglenozoa. Arch. Protistenkd. *148*, 318–328.
8. Adl, S.M., Simpson, A.G.B., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., Brown, M.W., Burki, F., Dunthorn, M., Hampl, V., et al. (2012). The revised classification of eukaryotes. J. Eukaryot. Microbiol. *59*, 429–493.
9. Yabuki, A., and Tame, A. (2015). Phylogeny and reclassification of *Hemistasia phaeocysticola* (Scherffel) Elbrächter & Schnepf, 1996. J. Eukaryot. Microbiol. *62*, 426–429.
10. López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C., and Moreira, D. (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature *409*, 603–607.
11. Lara, E., Moreira, D., Vereshchaka, A., and López-García, P. (2009). Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. Environ. Microbiol. *11*, 47–55.
12. del Campo, J., and Ruiz-Trillo, I. (2013). Environmental survey meta-analysis reveals hidden diversity among unicellular opisthokonts. Mol. Biol. Evol. *30*, 802–805.
13. Scheckenbach, F., Hausmann, K., Wylezich, C., Weitere, M., and Arndt, H. (2010). Large-scale patterns in biodiversity of microbial eukaryotes from the abyssal sea floor. Proc. Natl. Acad. Sci. USA *107*, 115–120.
14. Eloe, E.A., Shulse, C.N., Fadrosh, D.W., Williamson, S.J., Allen, E.E., and Bartlett, D.H. (2011). Compositional differences in particle-associated and free-living microbial assemblages from an extreme deep-ocean environment. Environ. Microbiol. Rep. *3*, 449–458.
15. Sauvadet, A.L., Gobet, A., and Guillou, L. (2010). Comparative analysis between protist communities from the deep-sea pelagic ecosystem and specific deep hydrothermal habitats. Environ. Microbiol. *12*, 2946–2964.
16. Countway, P.D., Gast, R.J., Dennett, M.R., Savai, P., Rose, J.M., and Caron, D.A. (2007). Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). Environ. Microbiol. *9*, 1219–1232.
17. Pernice, M.C., Giner, C.R., Logares, R., Perera-Bel, J., Acinas, S.G., Duarte, C.M., Gasol, J.M., and Massana, R. (2016). Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. ISME J. *10*, 945–958.
18. Morgan-Smith, D., Clouse, M.A., Herndl, G.J., and Bochdansky, A.B. (2013). Diversity and distribution of microbial eukaryotes in the deep tropical and subtropical North Atlantic Ocean. Deep Sea Res. Part I Oceanogr. Res. Pap. *78*, 58–69.
19. Diehl, S. (2002). Phytoplankton, light, and nutrients in a gradient of mixing depths. Theor. Ecol. *83*, 386–398.
20. Arístegui, J., and Gasol, J. (2009). Microbial oceanography of the dark ocean's pelagic realm. Limnol. Oceanogr. *54*, 1501–1529.
21. Orcutt, B.N., Sylvan, J.B., Knab, N.J., and Edwards, K.J. (2011). Microbial ecology of the dark ocean above, at, and below the seafloor. Microbiol. Mol. Biol. Rev. *75*, 361–422.
22. Finlay, B.J. (2002). Global dispersal of free-living microbial eukaryote species. Science *296*, 1061–1063.
23. Foissner, W. (2006). Biogeography and dispersal of micro-organisms: A review emphasizing protists. Acta Protozool. *45*, 111–136.
24. Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker, P., Iudicone, D., de Vargas, C., Bittner, L., et al. (2016). Insights into global diatom distribution and diversity in the world's ocean. Proc. Natl. Acad. Sci. USA *113*, E1516–E1525.
25. Triemer, R.E., and Ott, D.W. (1990). Ultrastructure of *Diplonema ambulator* Larsen & Patterson (Euglenozoa) and its relationship to *Isonema*. Eur. J. Protistol. *25*, 316–320.
26. Lukeš, J., Skalický, T., Týč, J., Votýpka, J., and Yurchenko, V. (2014). Evolution of parasitism in kinetoplastid flagellates. Mol. Biochem. Parasitol. *195*, 115–122.
27. Bettarel, Y., Sime-Ngando, T., Bouvy, M., Arfi, R., and Amblard, C. (2005). Low consumption of virus-sized particles by heterotrophic nanoflagellates in two lakes of the French Massif Central. Aquat. Microb. Ecol. *39*, 205–209.
28. Danovaro, R., Dell'Anno, A., Corinaldesi, C., Magagnini, M., Noble, R., Tamburini, C., and Weinbauer, M. (2008). Major viral impact on the functioning of benthic deep-sea ecosystems. Nature *454*, 1084–1087.
29. Dell'Anno, A., Corinaldesi, C., and Danovaro, R. (2015). Virus decomposition provides an important contribution to benthic deep-sea ecosystem functioning. Proc. Natl. Acad. Sci. USA *112*, E2014–E2019.
30. Gonzalez, J.M., and Suttle, C.A. (1993). Grazing by marine nanoflagellates on viruses and virus-sized particles: Ingestion and digestion. Mar. Ecol. Prog. Ser. *94*, 1–10.
31. Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J.C., Roux, S., Vincent, F., et al. (2015). Ocean plankton. Determinants of community structure in the global plankton interactome. Science *348*, http://dx.doi.org/10.1126/science.1262073.
32. Gawryluk, R.M.R., del Campo, J., Okamoto, N., Strassert, J.F.H., Lukeš, J., Richards, T.A., Worden, A.Z., Santoro, A.E., and Keeling, P.J. (2016). Morphological identification and single-cell genomics of marine diplonemids. Curr. Biol. *26*, this issue, 3053–3059.

Current Biology, Volume *26*

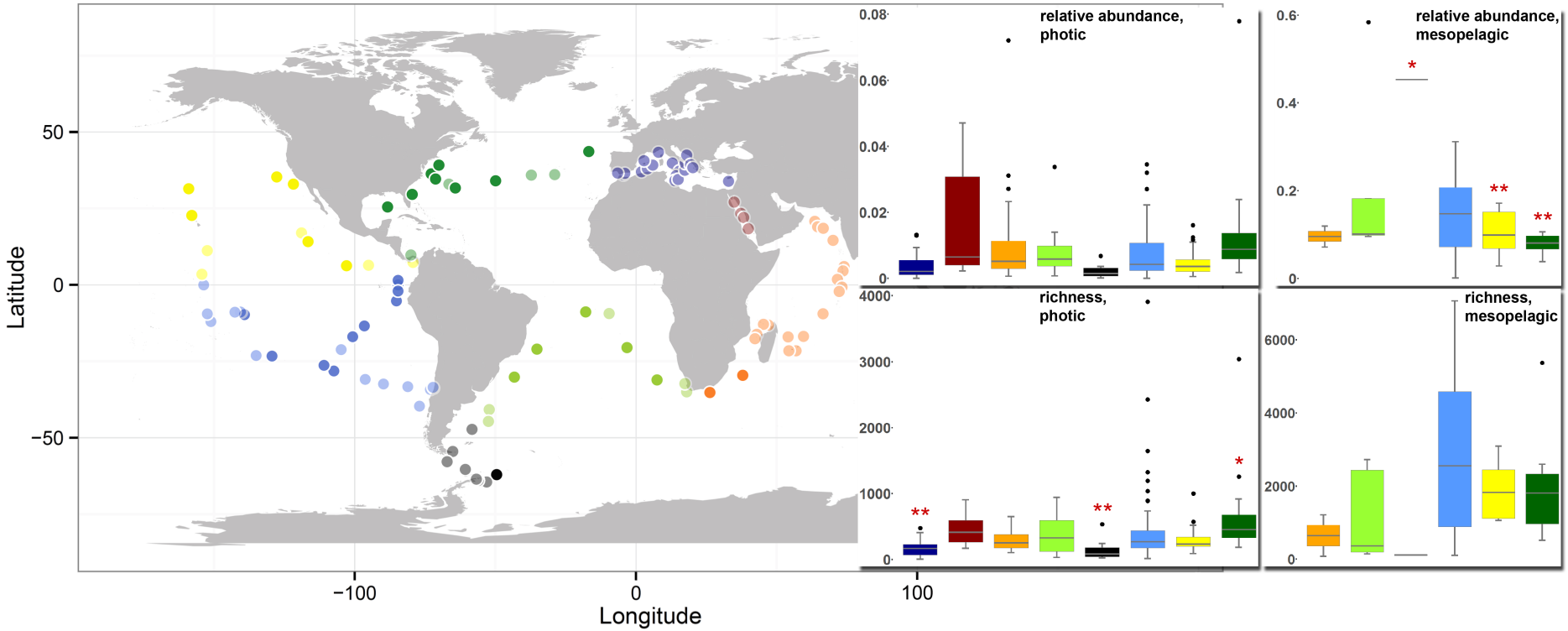
Supplemental Information

Extreme Diversity of Diplonemid Eukaryotes in the Ocean

Olga Flegontova, Pavel Flegontov, Shruti Malviya, Stephane Audic, Patrick Wincker, Colomban de Vargas, Chris Bowler, Julius Lukeš, and Aleš Horák

Supplemental Figures

Figure S1.



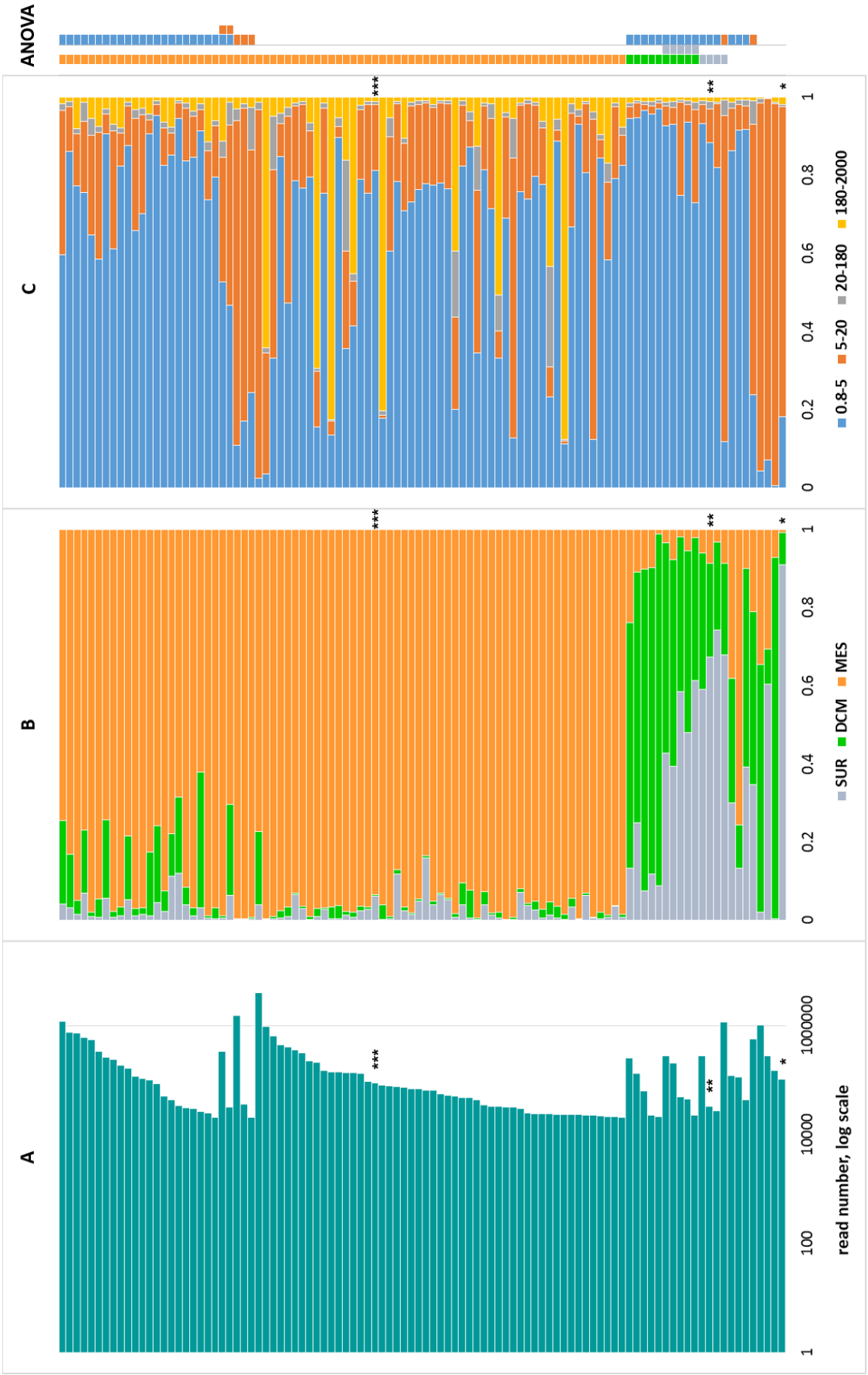

A
B
C
ANOVA
read number, log scale
1
100
10000
1000000
SUR
DCM
MES
0.8-5
5-20
20-180
180-2000
0
0.2
0.4
0.6
0.8
1
***
**
*

Figure S2.

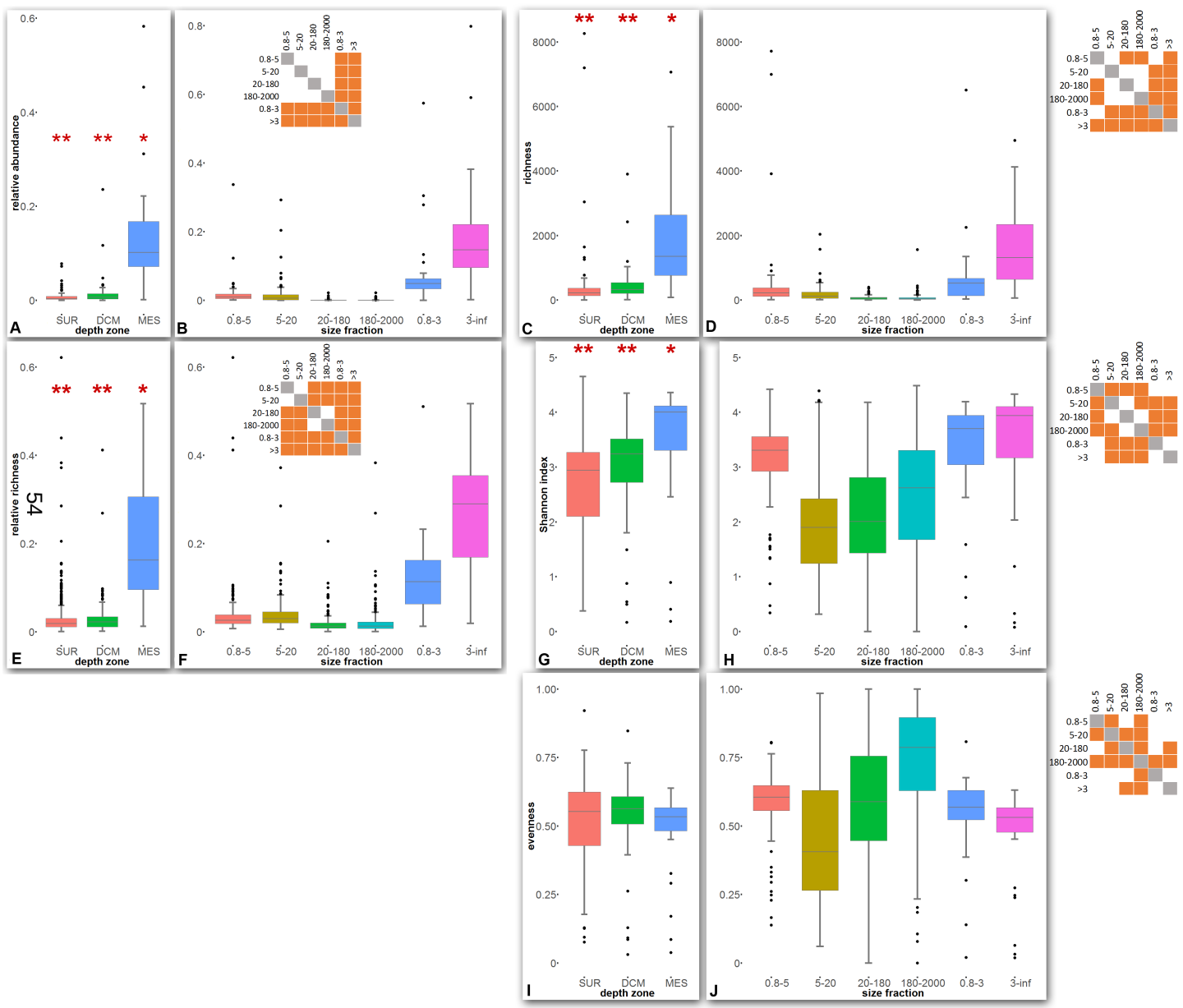
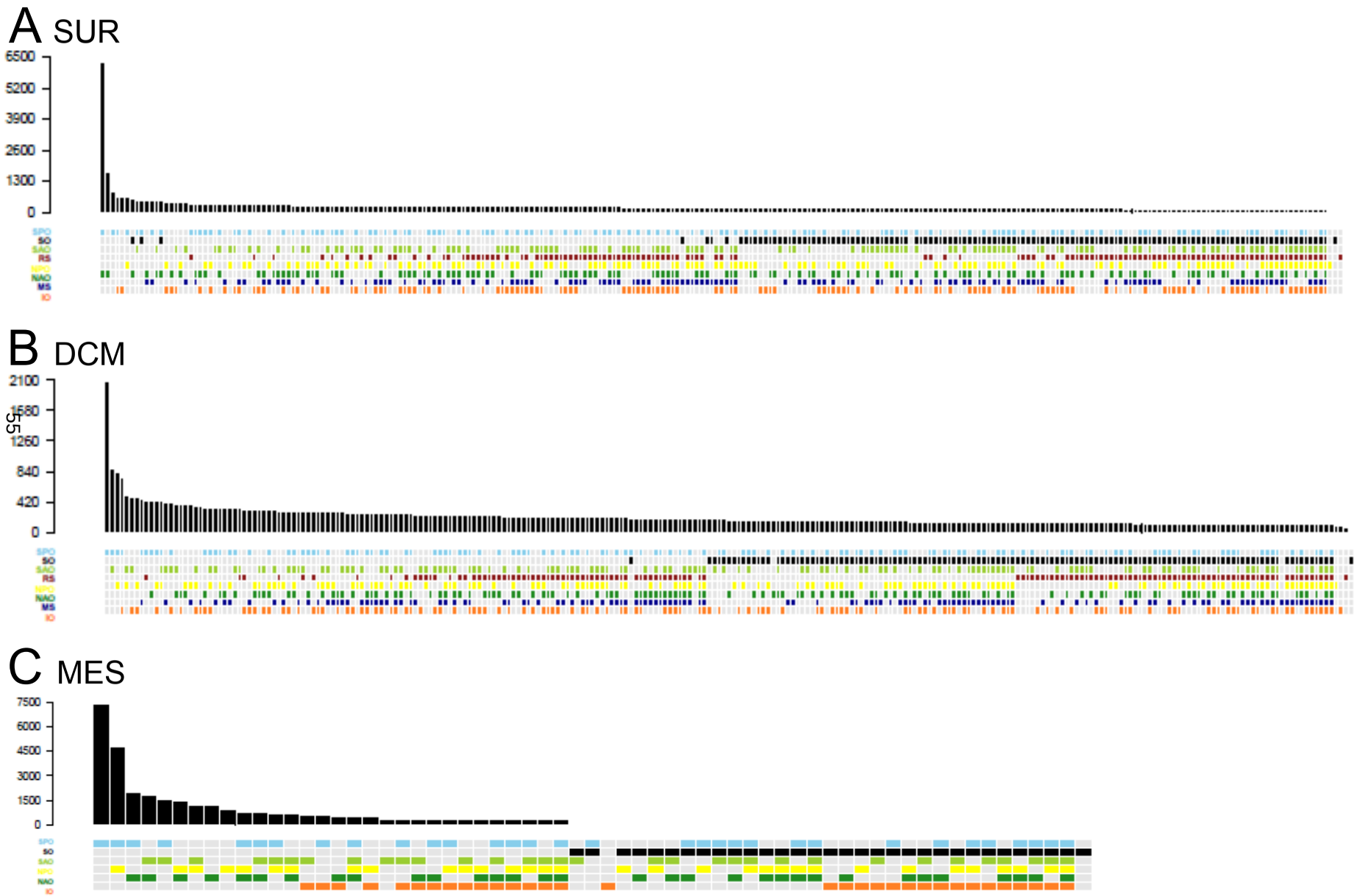

relative abundance
relative richness
richness
Shannon index
evenness
depth zone
size fraction
SUR
DCM
MES
0.8-5
5-20
20-180
180-2000
0.8-3
>3
A
B
C
D
E
F
G
H
I
J
54

Figure S4.


A SUR
6500
5200
3900
2600
1300
0
SPO
SO
SAO
RS
NPO
NAO
MS
IO
B DCM
2100
1680
1260
840
420
0
SPO
SO
SAO
RS
NPO
NAO
MS
IO
C MES
7500
6000
4500
3000
1500
0
SPO
SO
SAO
NPO
NAO
IO

Supplemental Figure legends

Figure S1. Related to Figure 4. Locations of 123 sampling stations on the world map. Oceanic provinces are color-coded in the following way: dark-blue, the Mediterranean Sea; red, the Red Sea; orange, the Indian Ocean; light-green, the South Atlantic Ocean; black, the Southern Ocean; light-blue, the South Pacific Ocean; yellow, the North Pacific Ocean; and dark-green, the North Atlantic Ocean. Stations lacking mesopelagic samples are shown in semi-transparent colors. Relative abundance of diplonemid ribotypes (percentage of diplonemid reads among eukaryotic reads) and richness (OTU count) across the oceanic provinces are shown using box plots. For making the plots, all samples at a given station and depth zone were merged. Due to large differences in relative abundance between the photic and mesopelagic zones, separate plots were produced for the latter. The box plot shows the median (crossbar), the first and third quartiles (hinges) and values within 1.5 inter-quartile range from the hinge (whiskers). Outliers are shown with black circles. Pairs formed by oceanic provinces marked with single and double asterisks are significantly different according to ANOVA combined with Tukey's honest significance test (p-value adjusted for multiple testing < 0.05). For example, the North Atlantic Ocean is different from the Mediterranean Sea or the Southern Ocean according to richness in the photic zone.

Figure S2. Related to Figures 1, 3 and 4A,B .Total read counts (**A**) and normalized relative abundance values found in depth zones (**B**) and size fractions (**C**) for 100 most abundant OTUs. The analysis of size fraction distribution was confined to the photic zone since samples of four fractions were generally available, as compared to just two fractions for the mesopelagic zone (Table S1A). For any OTU, relative abundance in each depth zone or size fraction was calculated, then normalized by the sum of relative abundances across all zones/fractions. This naïve approach was used for visualization only, while the analysis of variance (ANOVA) combined with Tukey's honest significance test was used to test whether a particular OTU was significantly more abundant in a given zone/fraction as compared to other zones/fractions (p-value adjusted for multiple testing < 0.05). The right-hand panel shows color codes for zones/fractions in which a given OTU occurred predominantly according to ANOVA. OTUs are sorted by their abundance patterns determined with ANOVA, and then by read counts in the descending order. All OTUs belonged to the DSPD I clade, with the exception of OTUs marked with: one asterisk, the *Diplonema/Rhynchopus* clade; two asterisks, the *Hemistasia* clade; three asterisks, the DSPD II clade.

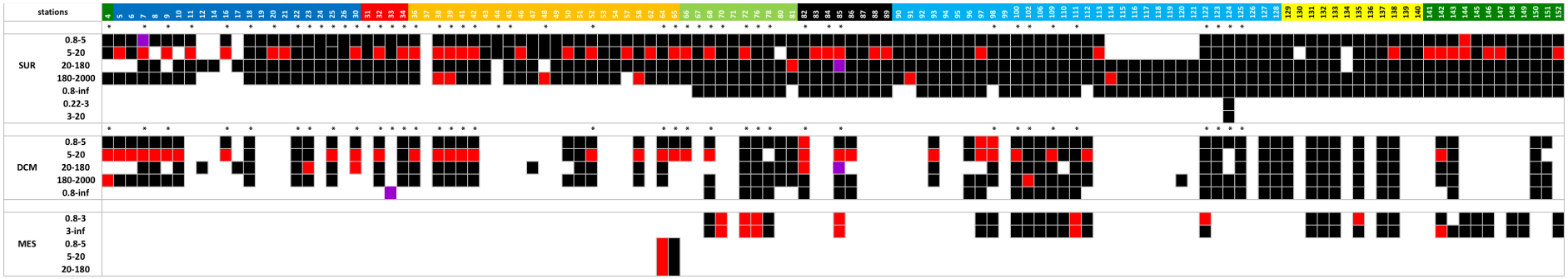
Figure S3. Related to Figures 3 and 4A,B. Box plots illustrating relative abundance of diplonemid ribotypes (**A, B,** calculated as diplonemid read count divided by eukaryotic read count), richness (**C, D,** OTU count), relative richness (**E, F,** diplonemid OTU count / eukaryotic OTU count), Shannon index (**G, H**) and evenness (**I, J**) of diplonemid communities across three depth zones (**A, C, E, G, I**) and six size fractions (**B, D, F, H, J**). Various size fractions were merged for a given depth zone; and surface and DCM zone samples were merged for a given size fraction (for calculating relative richness samples were not merged). Two fractions, 0.8-3 μm and >3 μm, correspond to the mesopelagic zone, while the other correspond to the surface and DCM zones. The box plot shows the median (crossbar), the first and third quartiles (hinges) and values within 1.5 inter-quartile range from the hinge (whiskers). Outliers are shown with black circles. Pairs formed by depth zones marked with single and double asterisks are significantly different according to ANOVA combined with Tukey's honest significance test (p-value adjusted for multiple testing < 0.05). For size fractions, significantly different pairs are marked in orange in the matrices beside each panel.

Figure S4. Related to Figure 4C. Counts of OTUs encountered in the surface (**A**), DCM (**B**), and mesopelagic (**C**) zones in all possible combinations of oceanic provinces: dark-blue, Mediterranean Sea (MS); red, Red Sea (RS); orange, Indian Ocean (IO); light-green, South Atlantic Ocean (SAO); black, Southern Ocean (SO); light-blue, South Pacific Ocean (SPO); yellow, North Pacific Ocean (NPO); and dark-green, North Atlantic Ocean (NAO).

Supplemental Tables

Table S1.

A



B

			samples								eukaryotic reads								diplonemid reads							
			ALL	SUR	DCM	MES	ALL, %	SUR, %	DCM, %	MES, %	ALL	SUR	DCM	MES	ALL, %	SUR, %	DCM, %	MES, %	ALL	SUR	DCM	MES	ALL, %	SUR, %	DCM, %	MES, %
full dataset	oceanic provinces	MS	111	67	44	0	13.1%	13.0%	16.1%	0.0%	1.53E+08	9.58E+07	5.73E+07	0.00E+00	13.3%	14.3%	14.4%	0.0%	1.44E+06	3.84E+05	1.05E+06	0.00E+00	5.9%	7.5%	16.4%	0.0%
		RS	24	14	10	0	2.8%	2.7%	3.7%	0.0%	4.05E+07	2.27E+07	1.78E+07	0.00E+00	3.5%	3.4%	4.5%	0.0%	7.01E+05	2.77E+05	4.23E+05	0.00E+00	2.9%	5.4%	6.6%	0.0%
		IO	124	80	42	2	14.6%	15.5%	15.4%	3.3%	1.85E+08	1.10E+08	7.18E+07	3.48E+06	16.1%	16.3%	18.1%	4.3%	3.77E+06	1.41E+06	2.06E+06	3.12E+05	15.6%	27.5%	32.1%	2.5%
		SAO	85	47	28	10	10.0%	9.1%	10.3%	16.4%	1.25E+08	6.62E+07	4.25E+07	1.62E+07	10.9%	9.8%	10.7%	20.1%	4.55E+06	3.88E+05	5.54E+05	3.61E+06	18.8%	7.6%	8.6%	28.4%
		SO	58	41	15	2	6.8%	7.9%	5.5%	3.3%	6.90E+07	4.43E+07	2.13E+07	3.43E+06	6.0%	6.6%	5.4%	4.2%	1.72E+06	7.49E+04	8.77E+04	1.56E+06	7.1%	1.5%	1.4%	12.3%
		SPO	245	147	78	20	28.8%	28.5%	28.6%	32.8%	3.61E+08	2.01E+08	1.30E+08	2.97E+07	31.4%	29.9%	32.7%	36.8%	7.36E+06	1.12E+06	1.68E+06	4.56E+06	30.4%	22.0%	26.2%	35.9%
		NPO	104	57	35	12	12.2%	11.0%	12.8%	19.7%	1.01E+08	5.66E+07	3.34E+07	1.13E+07	8.8%	8.4%	8.4%	13.9%	1.88E+06	2.11E+05	3.09E+05	1.36E+06	7.7%	4.1%	4.8%	10.7%
		NAO	99	63	21	15	11.6%	12.2%	7.7%	24.6%	1.16E+08	7.56E+07	2.32E+07	1.67E+07	10.0%	11.3%	5.9%	20.7%	2.80E+06	1.25E+06	2.52E+05	1.30E+06	11.6%	24.4%	3.9%	10.2%
		total	850	516	273	61					1.15E+09	6.72E+08	3.97E+08	8.09E+07					2.42E+07	5.11E+06	6.41E+06	1.27E+07				
	size fractions	0.8-5	169	105	64	0	23.0%	24.0%	26.8%	0.0%	2.14E+08	1.27E+08	8.65E+07	0	21.7%	22.7%	24.7%	0.0%	3.69E+06	1.91E+06	1.79E+06	0	16.7%	43.3%	33.2%	0.0%
		5-20	159	99	60	0	21.6%	22.7%	25.1%	0.0%	2.32E+08	1.37E+08	9.49E+07	0	23.5%	24.5%	27.1%	0.0%	5.64E+06	2.32E+06	3.32E+06	0	25.4%	52.6%	61.7%	0.0%
		20-180	170	116	54	0	23.1%	26.5%	22.6%	0.0%	2.26E+08	1.50E+08	7.67E+07	0	23.0%	26.8%	21.9%	0.0%	2.23E+05	8.96E+04	1.34E+05	0	1.0%	2.0%	2.5%	0.0%
		180-2000	178	117	61	0	24.2%	26.8%	25.5%	0.0%	2.37E+08	1.45E+08	9.17E+07	0	24.0%	26.0%	26.2%	0.0%	2.33E+05	8.80E+04	1.45E+05	0	1.0%	2.0%	2.7%	0.0%
		0.8-3	29	0	0	29	3.9%	0.0%	0.0%	49.2%	3.86E+07	0	0	3.86E+07	3.9%	0.0%	0.0%	49.9%	3.59E+06	0	0	3.59E+06	16.2%	0.0%	0.0%	29.0%
		3-inf	30	0	0	30	4.1%	0.0%	0.0%	50.8%	3.88E+07	0	0	3.88E+07	3.9%	0.0%	0.0%	50.1%	8.79E+06	0	0	8.79E+06	39.7%	0.0%	0.0%	71.0%
		total	735	437	239	59					9.86E+08	5.59E+08	3.50E+08	7.74E+07					2.22E+07	4.40E+06	5.39E+06	1.24E+07				
no WGA samples	oceanic provinces	MS	91	58	33	0	12.2%	12.4%	14.3%	0.0%	1.18E+08	7.89E+07	3.89E+07	0.00E+00	12.1%	13.3%	12.2%	0.0%	4.03E+05	2.21E+05	1.82E+05	0.00E+00	3.1%	7.3%	5.8%	0.0%
		RS	20	12	8	0	2.7%	2.6%	3.5%	0.0%	3.59E+07	2.08E+07	1.51E+07	0.00E+00	3.7%	3.5%	4.8%	0.0%	2.56E+05	6.78E+04	1.88E+05	0.00E+00	2.0%	2.2%	6.0%	0.0%
		IO	99	65	33	1	13.3%	13.9%	14.3%	2.1%	1.30E+08	7.86E+07	5.04E+07	1.33E+06	13.4%	13.3%	15.8%	2.3%	9.06E+05	5.04E+05	2.43E+05	1.59E+05	7.0%	16.6%	7.8%	2.4%
		SAO	73	43	26	4	9.8%	9.2%	11.3%	8.5%	1.05E+08	5.96E+07	3.90E+07	6.73E+06	10.9%	10.1%	12.3%	11.4%	1.48E+06	3.69E+05	4.47E+05	6.61E+05	11.5%	12.1%	14.3%	9.8%
		SO	44	35	9	0	5.9%	7.5%	3.9%	0.0%	4.81E+07	3.66E+07	1.15E+07	0.00E+00	5.0%	6.2%	3.6%	0.0%	8.40E+04	6.00E+04	2.40E+04	0.00E+00	0.7%	2.0%	0.8%	0.0%
		SPO	229	143	69	17	30.7%	30.6%	29.9%	36.2%	3.33E+08	1.96E+08	1.13E+08	2.47E+07	34.3%	33.0%	35.4%	41.8%	6.01E+06	1.12E+06	1.60E+06	3.29E+06	46.7%	36.8%	51.4%	49.0%
		NPO	102	56	35	11	13.7%	12.0%	15.2%	23.4%	9.98E+07	5.59E+07	3.34E+07	1.05E+07	10.3%	9.4%	10.5%	17.8%	1.80E+06	1.57E+05	3.09E+05	1.33E+06	14.0%	5.2%	9.9%	19.9%
		NAO	87	55	18	14	11.7%	11.8%	7.8%	29.8%	9.96E+07	6.65E+07	1.73E+07	1.58E+07	10.3%	11.2%	5.4%	26.8%	1.93E+06	5.39E+05	1.21E+05	1.27E+06	15.0%	17.8%	3.9%	18.9%
		total	745	467	231	47					9.70E+08	5.93E+08	3.18E+08	5.90E+07					1.29E+07	3.04E+06	3.12E+06	6.72E+06				
	size fractions	0.8-5	164	103	61	0	25.9%	26.5%	30.8%	0.0%	2.06E+08	1.25E+08	8.14E+07	0	25.5%	26.0%	30.0%	0.0%	3.17E+06	1.41E+06	1.76E+06	0	28.3%	60.5%	75.8%	0.0%
		5-20	88	60	28	0	13.9%	15.5%	14.1%	0.0%	1.10E+08	7.37E+07	3.63E+07	0	13.6%	15.4%	13.4%	0.0%	1.13E+06	7.45E+05	3.84E+05	0	10.1%	32.1%	16.5%	0.0%
		20-180	164	114	50	0	25.9%	29.4%	25.3%	0.0%	2.16E+08	1.47E+08	6.87E+07	0	26.7%	30.7%	25.3%	0.0%	1.56E+05	8.92E+04	6.67E+04	0	1.4%	3.8%	2.9%	0.0%
		180-2000	170	111	59	0	26.9%	28.6%	29.8%	0.0%	2.19E+08	1.34E+08	8.52E+07	0	27.1%	28.0%	31.4%	0.0%	1.95E+05	8.26E+04	1.12E+05	0	1.7%	3.6%	4.8%	0.0%
		0.8-3	22	0	0	22	3.5%	0.0%	0.0%	47.8%	2.88E+07	0	0	2.88E+07	3.6%	0.0%	0.0%	49.9%	1.94E+06	0	0	1.94E+06	17.3%	0.0%	0.0%	29.6%
		3-inf	24	0	0	24	3.8%	0.0%	0.0%	52.2%	2.89E+07	0	0	2.89E+07	3.6%	0.0%	0.0%	50.1%	4.61E+06	0	0	4.61E+06	41.2%	0.0%	0.0%	70.4%
		total	632	388	198	46					8.09E+08	4.80E+08	2.72E+08	5.76E+07					1.12E+07	2.32E+06	2.32E+06	6.56E+06				

Table S2.

	SUR incl. WGA	SUR without WGA	DCM incl. WGA	DCM without WGA	MES incl. WGA	MES without WGA
average	0.7%	0.6%	1.3%	0.9%	14.0%	10.9%
standard deviation	2.4%	1.2%	3.7%	1.4%	15.3%	8.7%
maximum value	33.7%	12.3%	40.7%	8%	79.8%	38.1%
number of samples	516	467	273	231	61	47

Table S3.

surface and DCM zones								
oceanic province	MS	RS	IO	SAO	SO	SPO	NPO	NAO
relative abundance	0.7%	1.8%	1.7%	0.8%	0.2%	0.8%	0.5%	1.3%
richness	167.9**	455.6	300.4	362.5	140.7**	621	300.4	1078*
relative richness	1.9%**	2.8%**	3.2%**	2.5%**	2.1%**	3.0%**	2.6%**	6.4%*
Shannon index	2.75	2.38	2.60	2.72	2.01*	3.05**	3.07**	3.34**
evenness	0.60*	0.41	0.48**	0.50	0.43**	0.54	0.56	0.52
number of samples	35	7	34	17	11	51	19	18
mesopelagic zone								
relative abundance	N/A	N/A	9.5%	21.2%	45.3%*	14.5%	10.4%**	7.9%**
richness	N/A	N/A	648.5	1171	114	2868	1888	2029
relative richness	N/A	N/A	14.4%	14.2%	5.6%	20.2%	23.2%	21.6%
Shannon index	N/A	N/A	3.46	2.57	0.41*	3.59**	3.92**	4.00**
evenness	N/A	N/A	0.61**	0.37	0.09*	0.49**	0.53**	0.54**
number of samples	N/A	N/A	2	5	1	10	6	8

Supplemental Tables legends

Table S1. A. Related to Figure 3. The table shows a summary of all samples used, with WGA samples shown in red. Size fractions and depth zones are indicated on the left, and stations on the top. Cases where both regular and WGA samples were available for a given depth zone and size fraction are highlighted in violet. Merged cells correspond to combined size fractions. Oceanic provinces are color-coded in the following way: dark-blue, Mediterranean Sea; red, Red Sea; orange, Indian Ocean; light-green, South Atlantic Ocean; black, Southern Ocean; light-blue, South Pacific Ocean; yellow, North Pacific Ocean; and dark-green, North Atlantic Ocean. Depth zones are abbreviated as follows: SUR, surface; DCM, deep chlorophyll maximum; MES, mesopelagic. Surface and DCM sampling stations reported previously in de Vargas et al. (2015) [S2] are marked with asterisks above the respective columns. **B.** Breakdown of samples, eukaryotic reads and diplonemid reads by oceanic provinces, depth zones, and size fractions. Cells contains respective sample and read counts or percentages. Data for the original dataset (on top) and for the dataset without whole-genome aplification (WGA) samples are shown (at the bottom of the table). Size fraction ranges are shown in µm. Depth zones: SUR, surface; DCM, deep chlorophyll maximum, MES, mesopelagic zone. Oceanic provinces: MS, Mediterranean Sea; RS, Red Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean.

Table S2. Related to Figure 4A,B. Relative abundance of diplonemid ribotypes in the dataset without WGA samples and in the original dataset. Size fractions were not merged for this analysis. Excluding all diplonemid reads coming from samples amplified using WGA (Table S1) results in 20.6 million reads, 244,081 diplonemid ribotypes and 40,507 OTUs vs. 24.2 million reads, 289,028 ribotypes and 45,197 OTUs in the full dataset and 12,325 OTUs in de Vargas et al. 2015 [S2].

Table S3. Related to Figure 4. Relative abundance and diversity statistics averaged across oceanic provinces in the photic and mesopelagic zones. The statistics were first calculated for separate depth zones (surface, DCM, and mesopelagic), with size fractions merged for a given zone, and then mean values were calculated. The oceanic provinces are abbreviated as follows: Mediterranean Sea (MS); Red Sea (RS); Indian Ocean (IO); South Atlantic Ocean (SAO); Southern Ocean (SO); South Pacific Ocean (SPO); North Pacific Ocean (NPO); North Atlantic Ocean (NAO). Pairs formed by oceanic provinces marked with single and double asterisks are significantly different according to ANOVA combined with Tukey's honest significance test (p-value adjusted for multiple testing < 0.05).

Supplemental Experimental Procedures

Dataset composition

We worked with the eukaryotic small subunit ribosomal DNA (18S rDNA) metabarcoding dataset obtained in frame of the *Tara* Oceans expedition [S1, S2]. The dataset included DNA sequencing reads of the V9 region of the 18S rRNA gene clustered into ribotypes (barcodes). Planktonic DNA samples were collected at 123 stations worldwide (Fig. S1) in eight oceanographic provinces, i.e., the Mediterranean Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO), South Pacific Ocean (SPO), North Pacific Ocean (NPO), and North Atlantic Ocean (NAO). Up to three depth zones were sampled per station: the surface (5-25 m), deep chlorophyll maximum (DCM, 17-185 m) and the mesopelagic zone (268-852 m). The surface and DCM zones included up to four size fractions (Table S1A): 0.8-5 µm (piconano-plankton), 5-20 µm (nano-plankton), 20-180 µm (micro-plankton), and 180-2,000 µm (meso-plankton), plus some additional size fractions in a few samples (>0.8 µm, 0.8-180 µm, 0.22-3 µm, 3-20 µm). Mesopelagic samples usually included up to two size fractions: 0.8-3 µm and >3 µm. DNA was extracted from all samples, and the hyper-variable V9 region of the nuclear 18S rDNA was PCR-amplified [S3]. Samples (105 in total) with low starting DNA concentration were treated with a whole-genome amplification procedure prior to amplification of the V9 region, as described in [S2].

The final dataset included 123 stations and 850 samples, containing approximately 1,150 million eukaryotic V9 reads (merged paired-end reads of the Illumina technology). For a fraction of samples (334 samples), data were taken from a previous publication focused on the photic zone [S2], and 516 samples from 76 locations are newly reported in this study (Table S1A). Identical reads were clustered into ribotypes (barcodes), which received taxonomic assignments through annotation against an expert-curated V9 reference database (for details, see [S2]) derived from the PR2 database [S4]. Subsequently ribotypes with abundance less than 3 reads were removed in order to avoid potential biases associated with sequencing errors, following the approach used by de Vargas et al. [S2]. The reference database contained 7 sequences belonging to the *Diplonema* genus, 6 sequences belonging to the *Rhynchopus* genus, and 38 environmental diplonemid sequences. As a result, 289,028 ribotypes having ≥85% identity to reference sequences were assigned to clade Diplonemea (phylum Euglenozoa, super-group Excavata), with read counts (abundance) per ribotype ranging from 3 to 2,857,135, and with a total read count of 24,217,285. OTUs were defined using the linkage clustering 'Swarm' approach [S5], resulting in 45,197 OTUs. All read clustering, OTU definition and taxonomic assignment protocols closely followed those used in de Vargas et al. (2015)[S2], to ensure compatibility with this large-scale study.

Phylogenetic analysis

For the phylogenetic analysis of nearly full-length 18S rDNA sequences we used the following approach. First, a core set of diplonemid and kinetoplastid 18S rDNA sequences taken from the PR2 database [S4] was used as the initial query for an iterative search in the GenBank database, implemented in the BlastCircle v.0.3 script [S6](http://eukref.org/curation-pipeline-overview/). Second, the output sequences were clustered with the 97% identity threshold using USEARCH, resulting in a set of 'seed' rDNAs, i.e. representatives of each sequence cluster. Third, MAFFT v.7.245 [S7] with the '--auto' option and trimAl v.1.2 [S8] with the '-gt 0.3' and '-st 0.001' options were used to make and prune sequence alignments, including a distant eukaryotic outgroup. Fourth, FastTree v.2.1.8 [S9] was used to make a preliminary maximum likelihood tree. Seeds and corresponding sequence clusters falling outside of Euglenozoa were removed subsequently, and the clustering, alignment, and tree building steps were repeated a number of times until no sequences falling between the outgroup and the Euglenozoa clade were left. The final alignment was used to build a maximum likelihood tree with RAxML v.8.2.3 [S10] with the following options: phylogenetic model GTR+CAT+I; 25 rate categories; model optimization precision, 0.001; a random starting tree; 1,000 random bootstrap replicates and 200 iterations of the maximum likelihood algorithm.

The resulting reference tree allowed us to define four major diplonemid clades (Fig. 1). Using these clade assignments, a reference database of diplonemid V9 SSU rDNA sequences was prepared, and clade assignments for V9 ribotypes from this study were obtained with the ggsearch36 software, according to de Vargas et al. (2015)[S2].

Global OTU distribution analysis

The final dataset, a matrix of V9 read counts for OTUs vs. samples, was used for calculating the following statistics in separate samples (or merged samples originating from the same depth or a size fraction): i/ relative abundance, i.e., the percentage of diplonemid V9 reads among eukaryotic V9 reads; ii/ relative richness, i.e., the percentage of diplonemid OTUs among eukaryotic OTUs; iii/ richness, i.e., the number of diplonemid OTUs; iv/ Shannon diversity index of the diplonemid community, v/ evenness of the diplonemid community. The statistics were plotted using R v.3.2.3, and the analysis of variance (ANOVA) combined with Tukey's honest significance test was used to compare the distributions across depth zones, size fractions, or oceanic provinces. Plotting on the world map was performed using an open source software QGIS v.2.8 (http://qgis.org/en/site/) with open-source maps. OTU rarefaction curves were computed with the rarefaction function in R v.3.2.3 (http://www.jennajacobs.org/R/rarefaction.txt), and curve slopes were estimated using the last ten of 100 points. Bootstrap resampling of the matrix of V9 read counts for OTUs vs. samples was performed using R package 'resample' (http://www.timhesterberg.net/r-packages; https://cran.r-project.org/web/packages/resample/index.html): sample columns were subjected to the bootstrap procedure with 1,000 replicates, and mean OTU counts in the collection of resampled datasets and standard deviations were estimated for the four hyper-diverse eukaryotic clades (diplonemids, metazoans, dinozoans, and rhizarians).

To visualize the level of similarity between different stations, we ordinated the stations using non-metric multidimensional scaling (NMDS) which is the most robust unconstrained ordination method in community ecology [S11]. We used the Bray–Curtis

distance to create a matrix of dissimilarity before running NMDS. The relation between occurrence, abundance and station evenness of each OTU was assessed, where occupancy was defined as the number of stations in which an OTU occurs and the station evenness was defined as the degree to which each OTU is distributed equally among the stations in which it occurs. This relationship was analyzed separately at all the three depth zones. Compositional similarity between stations and oceanic provinces were computed based on Hellinger-transformed abundance matrix and incidence matrix using Bray-Curtis and Jaccard indices respectively, as a measure of β-diversity. The agglomerative method used for hierarchical cluster analysis was the Ward clustering. All these analyses were conducted using open source R version 2.15.0 [S12].

Supplemental References

S1. Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., de Vargas, C., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J. M., et al. (2011). A holistic approach to marine Eco-systems biology. PLoS Biol. *9*, e1001177.

S2. de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. Science *348*, 1261605–1261605.

S3. Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W., and Huse, S. M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA Genes. PLoS One *4*, e6372.

S4. Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., De Vargas, C., Decelle, J., et al. (2013). The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Res. *41*, D597–D604

S5. Mahé, F., Rognes, T., Quince, C., de Vargas, C., and Dunthorn, M. (2014). Swarm : robust and fast clustering method for amplicon-based studies PeerJ, 1–12.

S6. Del Campo, J., and Ruiz-Trillo, I. (2013). Environmental survey meta-analysis reveals hidden diversity among unicellular opisthokonts. Mol. Biol. Evol. *30*, 802–805.

S7. Katoh, K., and Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usabiltiy. Mol. Biol. Evol. *30*, 772–780.

S8. Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973.

S9. Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One *5,* e9490.

S10. Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

S11. Minchin, P. R. (1987). An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio *69*, 89–107.

S12. R Developement Core Team (2015). R: A Language and Environment for Statistical Computing. R Found. Stat. Comput. *1*, 409.

5.4 Manuscript I

Flegontova O, Flegontov P, Malviya S, Poulain J, de Vargas C, Bowler C, Lukeš J, Horák A (submitted manuscript) Neobodonids are dominant kinetoplastids in the global ocean.

Abstract

Kinetoplastid flagellates are composed of basal mostly free-living bodonids and derived obligatory parasitic trypanosomatids, which belong to the best-studied protists. Due to their omnipresence in aquatic environments and soil, the bodonids are of ecological significance. Here we present the first global survey of marine kinetoplastids and compare it with the strikingly different patterns of abundance and diversity in their sister clade, the diplonemids. Based on analysis of 18S rDNA V9 ribotypes obtained from 124 sampling sites collected during the *Tara* Oceans expedition, our results show generally low to moderate abundance and diversity of planktonic kinetoplastids. Although we have identified all major kinetoplastid lineages, 98% of kinetoplastid reads are represented by neobodonids, namely specimens of the *Neobodo* and *Rhynchomonas* genera, which make up 59% and 18% of all reads, respectively. Most kinetoplastids have small cell size (0.8 – 5 μm) and tend to be more abundant in the mesopelagic as compared to the euphotic zone. Some of the most abundant operational taxonomic units have distinct geographical distributions, and three novel putatively parasitic neobodonids were identified, along with their potential hosts.

environmental microbiology
environmental microbiology reports

Neobodonids are dominant kinetoplastids in the global ocean

Journal:	*Environmental Microbiology and Environmental Microbiology Reports*
Manuscript ID	EMI-2017-1071
Journal:	Environmental Microbiology
Manuscript Type:	EMI - Research article
Date Submitted by the Author:	11-Jul-2017
Complete List of Authors:	Flegontova, Olga; Biologicke centrum Akademie Ved Ceske Republiky, Institute of Parasitology; Jihoceska Univerzita v Ceskych Budejovicich Prirodovedecka Fakulta Flegontov, Pavel; Biologicke centrum Akademie Ved Ceske Republiky, Institute of Parasitology; Ostravska Univerzita v Ostrave Prirodovedecka fakulta, Life Science Research Centre Malviya, Shruti; National Centre for Biological Sciences, Tata Institute of Fundamental Research, Simons Centre for the Study of Living Machines; Ecole Normale Supérieure, Institut de Biologie de l'Ecole Normale Supérieure (IBENS) Poulain, Julie; Genoscope - Centre National de Séquençage, UMR CNRS 8030 de Vargas, Colomban; UPMC Univ Paris 06, UMR7144 - Equipe EPPO: Evolution du Plancton et PaléoOcéans; Sorbonne Universités Bowler, Chris ; Ecole Normale Superieure, Institut de Biologie de l'Ecole Normale Supérieure (IBENS) Lukes, Julius; Biologicke centrum Akademie Ved Ceske Republiky, Institute of Parasitology; Jihoceska Univerzita v Ceskych Budejovicich Prirodovedecka Fakulta; Canadian Institute for Advanced Research Horak, Ales; Biologicke centrum Akademie Ved Ceske Republiky, Institute of Parasitology; Jihoceska Univerzita v Ceskych Budejovicich Prirodovedecka Fakulta
Keywords:	protozoa, symbionts, parasites, kinetoplastids, diversity, plankton, ocean

Neobodonids are dominant kinetoplastids in the global ocean

Running title: Kinetoplastids in the oceans

Olga Flegontova[1,2], Pavel Flegontov[1,3], Shruti Malviya[4,5], Julie Poulain[6], Colomban de Vargas[7,8], Chris Bowler[5], Julius Lukeš[1,2,9,*] & Aleš Horák[1,2,*]

[1] *Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic*

[2] *Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic*

[3] *Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czech Republic*

[4] *Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India*

[5] *Ecole Normale Supérieure, PSL Research University, Institut de Biologie de l'Ecole Normale Supérieure (IBENS), CNRS UMR 8197, INSERM U1024, 46 rue d'Ulm, F-75005 Paris, France*

[6] *Genoscope, CEA, Évry, France*

[7] *Station Biologique de Roscoff, Roscoff, France.*

[8] *Sorbonne Universités, Paris, France.*

[9] *Canadian Institute for Advanced Research, Toronto, Canada*

[*] Corresponding authors

Corresponding address: Institute of Parasitology, Biology Centre, Czech Academy of Sciences, Branišovská 31, 37005 České Budějovice, Czech Republic

Tel. +420387775409, +420387775403

Fax. +420385310388

Email. ogar@paru.cas.cz

Conflict of interest: The authors declare no conflict of interest

Summary

Kinetoplastid flagellates are comprised of basal mostly free-living bodonids and derived obligatory parasitic trypanosomatids, which belong to the best-studied protists. Due to their omnipresence in aquatic environments and soil, the bodonids are of ecological significance. Here we present the first global survey of marine kinetoplastids and compare it with the strikingly different patterns of abundance and diversity in their sister clade, the diplonemids. Based on analysis of 18S rDNA V9 ribotypes obtained from 124 sampling sites collected during the *Tara* Oceans expedition, our results show generally low to moderate abundance and diversity of planktonic kinetoplastids. Although we have identified all major kinetoplastid lineages, 98% of kinetoplastid reads are represented by neobodonids, namely specimens of the *Neobodo* and *Rhynchomonas* genera, which make up 59% and 18% of all reads, respectively. Most kinetoplastids have small cell size (0.8 – 5 μm) and tend to be more abundant in the mesopelagic as compared to the euphotic zone. Some of the most abundant operational taxonomic units have distinct geographical distributions, and three novel putatively parasitic neobodonids were identified, along with their potential hosts.

Introduction

Kinetoplastid flagellates (Kinetoplastea) belong to the phylum Euglenozoa (Adl *et al.*, 2012). Basal kinetoplastid lineages are generally called bodonids, a polyphyletic assemblage of pear-shaped bi-flagellated protists. A small group of parasitic or endosymbiont species within bodonids, represented by the *Ichthyobodo* and *Perkinsela*, belongs to the clade Prokinetoplastina. However, the bulk of bodonids described so far belong to the clade Metakinetoplastina, which also includes the crown group of trypanosomatids (Moreira *et al.*, 2004). Within Metakinetoplastina, three lineages termed Eu-, Neo- and Parabodonida are recognized (Lukeš *et al.*, 2014). These bodonid lineages harbor mostly free-living bacteriovores from aquatic environments and soil, where they usually constitute a relatively minor group of uncertain ecological significance (Glaser *et al.*, 2014; Atkins *et al.*, 2000; López-García *et al.*, 2003).

The trypanosomatids, which encompass a majority of known species, have adopted commensal or parasitic life strategies (Lukeš *et al.*, 2014). Due to extreme diversification and host-parasite co-evolution, it seems plausible that every terrestrial vertebrate species harbors its own *Trypanosoma* species (Hamilton *et al.*, 2007). Although the medically and veterinarily important members of the genera *Trypanosoma* and *Leishmania* have received most attention, it is within insect hosts where most of the diversity of these terrestrial parasites seems to be hidden (Maslov *et al.*, 2013).

While local diversity in some oceanic ecosystems has been well documented, until recently no systematic study of eukaryotic planktonic biodiversity across the world's ocean, and across the full

range of organismal sizes, was available. One of the first studies to tackle this shortcoming was based on samples collected by the *Tara* Oceans expedition, by taking advantage of a huge dataset of 18S rDNA V9 metabarcode sequences to explore the taxonomic structure, ecological roles and mutual interactions of planktonic prokaryotes and eukaryotes (Brum *et al.*, 2015; de Vargas *et al.*, 2015; Lima-Mendez *et al.*, 2015; Villar *et al.*, 2015; Sunagawa *et al.*, 2015). A number of follow-up studies further focused on particular taxonomic groups of planktonic eukaryotes, analyzing in detail the V9 ribotypes and associated data (Le Bescot *et al.*, 2016; Malviya *et al.*, 2016; Mutsuo *et al.*, 2016; Flegontova *et al.*, 2016).

While a global analysis of kinetoplastid protists in marine habitats is lacking, there are a few reports from pelagic systems (von der Heyden and Cavalier-Smith, 2005; Salani *et al.*, 2012), deep-sea benthos (Atkins *et al.*, 2000; López-García *et al.*, 2003; Brown and Wolfe, 2006; Sauvadet *et al.*, 2010; Scheckenbach *et al.*, 2010) and hypersaline anoxic basins (Edgcomb *et al.*, 2011). With the exception of the latter niche, kinetoplastids generally seem to constitute a minor component of the plankton. However, this viewpoint has recently been challenged because the significant sequence divergence of the kinetoplastid 18S rRNA gene may make the universal primers typically used for metabarcoding unsuitable (Mukherjee *et al.*, 2015), a view further supported by the unexpectedly frequent appearance of these flagellates in FISH-based analyses of the mesopelagic and deeper layers (Morgan-Smith *et al.*, 2011). Indeed, free-living kinetoplastids have been identified as a dominant group of the hypolimnion of a freshwater lake ecosystem (Mukherjee *et al.*, 2015). These results suggest that bodonids may represent a major bacteriovorous component of the plankton that has so far escaped broader recognition (Mukherjee *et al.*, 2015)

On the other hand, we now know that heterotrophic protists constitute a much more diverse component of the plankton than formerly appreciated, significantly exceeding all photosynthetic eukaryotes in species number (de Vargas *et al.*, 2015; Worden *et al.*, 2015). The aim of the current study was therefore to investigate community structure, patterns of diversity and abundance, and possible ecological role of marine planktonic kinetoplastids, evaluated for the first time on a global scale in samples collected during the *Tara* Oceans expedition.

Materials and Methods

Dataset composition

We worked with the eukaryotic small subunit ribosomal RNA (18S rDNA) metabarcoding dataset obtained in the frame of the *Tara* Oceans expedition (Karsenti *et al.*, 2011; de Vargas *et al.*, 2015). The dataset included DNA sequencing reads of the V9 region of the 18S rRNA gene clustered into ribotypes. Planktonic DNA samples were collected at 124 stations worldwide (Suppl. Table 1) in eight

oceanographic provinces, namely the Mediterranean Sea (MS), Red Sea (RS), Indian Ocean (IO), South Atlantic Ocean (SAO), Southern Ocean (SO), South Pacific Ocean (SPO), North Pacific Ocean (NPO), and North Atlantic Ocean (NAO). Up to three depth zones were sampled per station: surface (SRF, 5-25 m), deep chlorophyll maximum (DCM, 17-185 m), and mesopelagic zone (MES, 347-852 m). At few stations, oxygen-depleted waters were sampled (OMZ, 268-595 m). The SRF and DCM zones included up to four size fractions (Suppl. Table. 1): 0.8-5 µm (piconano-plankton), 5-20 µm (nano-plankton), 20-180 µm (micro-plankton), and 180-2,000 µm (meso-plankton), plus some additional size fractions in a few samples (>0.8 µm, 0.8-20 µm). The OMZ included three size fractions: 0.8-5 µm, 5-20 µm, and 20-180 µm. Mesopelagic samples included two size fractions: 0.8-3 µm and >3 µm. DNA was extracted from all samples, and the hyper-variable V9 region of the nuclear 18S rDNA was PCR-amplified (Amaral-Zettler *et al.*, 2009). Identical reads were merged into ribotypes, which received taxonomic assignments through annotation against an expert-curated V9 reference database (for details, see de Vargas *et al.*, 2015) derived from the PR2 database (Guillou *et al.*, 2013). Subsequently ribotypes with abundance less than 3 reads were removed in order to avoid potential biases associated with sequencing errors, following the approach used by de Vergas et al. (2015). The ribotypes were clustered into OTUs using the linkage clustering 'Swarm' approach (Mahé *et al.*, 2014). From the resulting global dataset we extracted OTUs assigned to the Kinetoplastea phylum and refined clade assignments using an in-house 18S rRNA reference database for kinetoplastids and the ggsearch36 software.

In-house 18S rRNA reference database for kinetoplastids

For the phylogenetic analysis of nearly full-length 18S rRNA sequences we used the following approach. First, a core set of kinetoplastid 18S rRNA sequences taken from the PR2 database (Guillou *et al.*, 2013) was used as the initial query for an iterative search in the GenBank database, implemented in the BlastCircle v.0.3 script (eukref.org). Second, the output sequences were clustered with the 97% identity threshold using USEARCH (Edgar, 2010), resulting in a set of 'seed' rRNAs. Third, MAFFT v.7.245 (Katoh and Standley, 2013) with the '--auto' option and trimAl v.1.2 (Capella-Gutiérrez *et al.*, 2009) with the '-gt 0.3' and '-st 0.001' options were used to make and prune sequence alignments, including a distant eukaryotic outgroup. Fourth, FastTree v.2.1.8 (Price *et al.*, 2010) was used to make a preliminary maximum likelihood tree. Seeds and corresponding sequence clusters falling outside of Euglenozoa were removed subsequently, and the clustering, alignment, and tree building steps were repeated a number of times until no sequences falling between the outgroup and the Euglenozoa clade were left. The final alignment was used to build a maximum likelihood tree with RAxML v.8.2.3 (Stamatakis, 2014) with the following options: phylogenetic model GTR+CAT+I; 25 rate categories;

model optimization precision, 0.001; a random starting tree; 1,000 random bootstrap replicates and 200 iterations of the maximum likelihood algorithm.

Global OTU distribution analysis

The final dataset, a matrix of V9 read counts for OTUs vs. samples (Suppl. Table 1), was used for calculating the following statistics in separate samples or their combinations based on depth zones and size fraction: i/ relative abundance, i.e. the percentage of kinetoplastid V9 reads among eukaryotic V9 reads; ii/ richness, i.e. the number of kinetoplastid OTUs; iii/ Shannon diversity index, iv/ evenness. The one-way analysis of variance (ANOVA) combined with Tukey's honest significance test was used to compare the distributions across depth zones, size fractions, oceanic provinces, or three latitude zones: i/ tropical, 24°N-24°S; ii/ temperate 25-44°N and 25-44°S; iii/ Antarctic 44-65°S. Multi-way ANOVA assessed the influence of four variables listed above and their pairwise interactions on the abundance and diversity statistics. Plotting of various statistics on the world map was performed using an open source software QGIS v.2.8 (http://qgis.org/en/site/) with open-source maps. Compositional similarity between stations and oceanic provinces were computed based on Hellinger-transformed abundance matrix and incidence matrix using Bray-Curtis and Jaccard indices respectively, as a measure of β-diversity. The agglomerative method used for hierarchical cluster analysis was the Ward clustering.

Results

Abundance of kinetoplastids across depth zones and geographical regions

From the global meta-barcoding dataset from *Tara* Oceans we extracted 1 570 025 kinetoplastid reads belonging to 8 207 ribotypes clustered into 512 Operational Taxonomic Units (OTUs; see sample information and read counts for each OTU in Suppl. Table 1). Their diversity in the global dataset was comparable to that of rhodophytes or cryptophytes, was about 15% of the richness of another major terrestrial parasitic clade, apicomplexans but only about 1% compared to their sister lineage, diplonemids. The trypanosomatids, a lineage of terrestrial parasites, were essentially missing from our samples (Table 1). The same was true for other parasitic and/or endosymbiotic lineages (*Perkinsela* and *Ichthyobodo*). The vast majority of reads (99.8%), on the other hand, belonged to free living eu- (1.3%) and especially neobodonids (98.4%; of which 59.1% belong to the *Neobodo* and 18.2% to the *Rhynchomonas*). The distribution of richness among clades correlated well with the abundance (Pearson's $r = 0.946$), although the rare lineages listed above represented a much larger fraction of

kinetoplastid richness compared to their abundances (Table 1).

Rarefaction curves revealed that kinetoplastid diversity was saturated in the whole dataset, as it was for the four most diverse and abundant groups (Neobodonida, *Neobodo*, *Rhynchomonas*, and unknown Neobodonida), with rarefaction curve slopes ranging from 1×10^{-7} to 5×10^{-7} (Fig. 1). These values are comparable with the saturation of major planktonic eukaryotic groups, such as Metazoa, Dinozoa and Rhizaria. Kinetoplastid diversity was also approximately 10 times more saturated than that of diplonemids, their sister group, which were previously found to be highly diverse (Flegontova *et al.*, 2016).

Next, using one-way ANOVA, we explored the distribution of neobodonids as a whole, and of the 14 most abundant kinetoplastid OTUs (one eu- and 13 neobodonids) across three depth zones, six size fractions, three latitude zones, and eight oceanic provinces (Fig. 2). Globally, kinetoplastids represent only a small fraction of planktonic eukaryotes. Their relative abundance, i.e., the number of kinetoplastid reads divided by the number of total eukaryotic reads, averaged 0.2% per sample (ranging from 0% to 14.8%). Because neobodonids accounted for an overwhelming majority of marine kinetoplastids, the abundance of these groups mirrored the global patterns, in being significantly more abundant in the deeper mesopelagic (MES) zone below 200 metres as compared to the sunlit surface (SRF) and deep chlorophyll maximum (DCM) zones (see Methods for details). Their abundance was furthermore maximal in the smallest piconano-plankton size fractions (0.8-5 µm or 0.8-3 µm) in all zones. Thus, marine kinetoplastids are even smaller than the related diplonemids, which were almost equally abundant in the 0.8-5 µm and 5-20 µm fractions (Flegontova *et al.*, 2016). No statistically significant geographical patterns in abundance could be detected for the neobodonid group as a whole (Fig. 2).

When abundant OTUs were examined individually, we observed that the most abundant kinetoplastid OTU (belonging to *Neobodo*) as well as one eubodonid OTU were preferentially found in mesopelagic samples with low oxygen concentration (the oxygen minimum zone, OMZ). The single eubodonid taxon was found in the 5-20 µm fraction, but a majority of the abundant OTUs occurred mostly in piconano-plankton (below 5 µm), suggesting that the organisms are likely free-living. However, three OTUs among unclassified neobodonids showed significant enrichment in the largest meso-plankton fraction (180-2000 µm). From this size distribution we infer that these are probably novel and abundant parasitic taxa (see below for details).

We also analyzed geographic distributions of the 14 most abundant OTUs (Fig. 2). Their distribution across *Tara* Oceans stations is shown in Suppl. Fig. 1. The most abundant OTU, *Neobodo* OTU #324 accounting for 36% of all kinetoplastid reads, occurred predominantly in the Indian and South Pacific Oceans (Suppl. Fig. 1). A multi-way ANOVA analysis supports the conclusion that the

abundance of this OTU depends on two variables: size fraction and oceanic province (Fig. 3). *Neobodo* OTU #1514 occurred mostly in the tropical latitudes, but the effect is statistically significant in the mesopelagic zone only (Fig. 2). *Neobodo* OTU #2753 had a peculiar distribution, being found both in the Mediterranean Sea and in the Drake Passage (supported by one-way and multi-way ANOVA), and another *Neobodo* OTU #3211 occurred almost exclusively in the North Atlantic Ocean. Thus, four of the five most abundant *Neobodo* OTUs displayed distinct biogeographies. *Rhynchomonas* OTU #678 was prevalent at tropical latitudes (supported by one-way ANOVA), and *Rhynchomonas* OTU #3853 had a rather narrow geographic distribution, occurring almost exclusively at tropical latitudes of the Pacific Ocean and in the North Atlantic Ocean (supported by multi- and one-way ANOVA, see Figs. 2 and 3, Suppl. Fig. 1). Among six OTUs belonging to unknown neobodonids, only two demonstrated a geographic pattern: a putatively parasitic OTU #3742 was dominant in tropical regions, whereas OTU #3677 was more widely distributed in the North Atlantic, tropical Pacific and Indian Oceans (both cases are supported by one-way ANOVA). The only eubodonid OTU in our dataset, OTU #2803, was somewhat prevalent in the Mediterranean Sea although this effect was not statistically significant (Suppl. Fig. 1).

Diversity

We then examined the effect of depth, size fraction, latitudinal gradients and oceanic provinces on kinetoplastid diversity (Fig. 4). We analyzed all kinetoplastids, neobodonids only, or the most abundant neobodonid sub-clades (neobodonids account for about 70% of kinetoplastid OTUs, see Table 1). Diversity of all kinetoplastids and neobodonids exhibited very similar patterns (Fig. 4), and generally followed the same trends as their relative abundance, peaking in the MES zone and in the piconano-plankton size fraction. The same patterns were observed for all three major neobodonid sub-groups, apart from *Rhynchomonas*, whose richness was not significantly stratified by depth. However, when considering the number of OTUs unique to certain depth zones and size fractions, we observed that the surface zone had by far the largest number of unique OTUs: 37% of all kinetoplastid OTUs were unique to this zone (Suppl. Fig. 2A). In contrast, just 5% of OTUs were unique to the mesopelagic zone, even though average richness per station was much higher in this zone (Fig. 2). An explanation for this apparent paradox could be found by analysis of occupancy, because we found that the vast majority of surface-specific OTUs occurred in just a few stations (Fig. 5).

On the other hand, the distribution of unique OTUs across size fractions was better correlated with richness: 21% of kinetoplastid OTUs were unique to the piconano-plankton fraction (Suppl. Fig. 2B), and the same fraction demonstrated the highest richness (Fig. 2). Notably, about 3% of OTUs were unique to the micro-plankton fraction, and the same percentage was observed for the meso-

plankton. These OTUs may represent parasitic species of low abundance.

Evenness and richness followed different trends: evenness peaked in DCM samples for all kinetoplastids except *Neobodo*, for which it was maximal in the MES zone (Fig. 4). For all kinetoplastids except *Rhynchomonas*, evenness peaked in the nano- and micro-plankton size fractions, while it was the piconano- and nano-plankton in the case of *Rhynchomonas*. The richness of kinetoplastids, neobodonids, *Rhynchomonas*, and unknown neobodonids was significantly higher in tropical regions (Fig. 4), in particular in the South Pacific Ocean, although this effect may be due to a higher proportion of tropical samples from that region (72%). However, *Rhynchomonas* richness was the highest in the North Pacific and North Atlantic Oceans. Evenness demonstrated no statistically significant differences across latitudes or provinces (Fig. 2).

Ecological interactions

While the majority of kinetoplastid reads were found in the pico-nano size fraction, suggesting their small cell size, our results also showed a part of kinetoplastid diversity that was positively associated with larger fractions. These are putative candidates for parasitic/symbiotic lifestyles. Analysis of abundant OTUs revealed such associations for three neobodonid OTUs. The only parasitic neobodonid described so far is *Azumiobodo* (Hirose *et al.*, 2012; Kumagai *et al.*, 2013), which was poorly represented within our samples (about 7,000 reads in total). Significantly more abundant were the novel parasite candidates (with 15 000, 29 000 and 45 000 reads, respectively). By analyzing a global interaction network for planktonic OTUs within the photic zone (Lima-Mendez *et al.*, 2015), we found only the latter putatively parasitic OTU (OTU #2083) amongst the interactions meeting the inclusion criteria of this former work. The other two OTUs were lacking in the interactome since they were mainly found in the mesopelagic zone (Fig. 2). The OTU #2083 interacted with 26 taxa, mostly bacteria and alveolates, but possibly the most interesting interaction was a co-presence with a planktonic appendicularian species *Megalocercus huxleyi* occurring in the meso-plankton fraction, where OTU #2083 was also relatively abundant (Fig. 2). Notably, *Azumiobodo* parasitizes on ascidians, a closely related group of benthic animals (Hirose *et al.*, 2012; Kumagai *et al.*, 2013).

To find hosts of the other two putatively parasitic OTUs, we applied a simple approach: calculated Pearson's correlation coefficients for the 14 most abundant kinetoplastid OTUs (or kinetoplastid sub-clades) vs. 456 abundant metazoan OTUs. We used absolute abundance values (read counts) and considered metazoans represented by more than 10 000 reads. The best correlation among all kinetoplastid OTUs tested was between the putatively parasitic OTU #4802 and a copepod OTU belonging to the Calanoida group, $r = 1$, p-value = 0. Thus, we found possible hosts for two out of three putatively parasitic OTUs: an appendicularian and a copepod.

We then analyzed possible interactions of other kinetoplastids found in euphotic zone (SRF and DCM) samples. The interactome (http://www.raeslab.org/companion/ocean-interactome.html) contains only 12 kinetoplastid ribotypes (belonging to 12 OTUs) meeting the stringent inclusion criteria (Lima-Mendez et al., 2015). OTUs outside of neobodonids 'interact' with none or few (less than 10) other ribotypes, therefore no conclusive interpretation of the interactome is possible for these groups. For neobodonids, 208 positive (co-occurrence of a kinetoplastid ribotype with another ribotype) and 27 negative interactions (mutual exclusions) that cannot be explained by environmental factors affecting both interacting organisms (Suppl. Table 2) were found. Co-presence with bacteria and archaea, their main food source, was detected (42 interactions), as well as 34 co-occurrences with other bacteriovorous protists, such as ciliates, choanoflagellates, foraminiferans, radiolarians, and marine stramenopiles (MAST). The largest fraction of positive interactions (78 instances) involves various groups of Syndiniales (MALV) and other dinoflagellates. Most instances of mutual exclusion include crustaceans, cnidarians, molluscs, and ascidians (16 of 27 negative interactions), yet metazoans also participate in positive interactions (18 instances). We also expected to see a strong positive correlation between the intracellular symbiont *Perkinsela* and its host amoeba, however this correlation was not found in the interactome (Lima-Mendez *et al.*, 2015) because the abundance of *Perkinsela* was too low. However, using the more simple approach a very strong correlation was revealed between *Perkinsela* (all OTUs combined) and the most abundant *Paramoeba* OTU annotated as *Paramoeba branchifila*: $r = 0.98$, p-value = 0.

Discussion

There are generally two opposing views on the modes and limits of dispersal of protists. The first one champions the idea of a cosmopolitan, ubiquitous distribution, the main driving force being their short generation times, a high rate of dispersal, large population sizes and ability to form resistant cysts (Finlay and Fenchel, 2004; Boenigk *et al.*, 2012). This postulate of "everything is everywhere" is challenged by an alternative view that attaches biogeographies to at least some protists and finds their level of endemism comparably high with respect to other eukaryotes (Foissner, 2006). A detailed analysis of two closely related groups of excavates, kinetoplastids and diplonemids, in a truly global dataset composed of samples covering all oceanic provinces, allows interrogation of the above contrasting scenarios from a new perspective.

Both diplonemids and kinetoplastids are heterotrophic protists of similar size (mostly present in the piconano fraction) that are more abundant in the MES zone as compared to the euphotic zone, a pattern that is characteristic for a number of marine heterotrophs (Pernice *et al.*, 2015; Worden *et al.*, 2015). Although very little is known about the lifestyle of the former group, based on their distribution

296 across size fractions, only a small portion of the described OTUs are likely to be parasites (Gawryluk
297 *et al.*, 2016; Flegontova *et al.*, 2016). The same seems to be the case for marine kinetoplastids, because
298 the confirmed parasitic groups constitute a mere ~0.5% of reads: all Trypanosomatida, *Ichthyobodo*
299 (Prokinetoplastina), *Azumiobodo* (Neobodonida), *Trypanoplasma* and *Cryptobia* (Parabodonida).
300 Three potentially parasitic OTUs found in this study account for 5.7% of reads. With the available data
301 one would therefore assume that most oceanic diplonemids and kinetoplastids are free-living and may
302 thus have rather similar ecological roles. Yet unexpectedly, the patterns of their abundance and
303 diversity are dramatically different.

304 Diplonemids emerged recently as the most diverse and 6th most abundant eukaryotic taxon in
305 the global plankton (Gawryluk *et al.*, 2016; Flegontova *et al.*, 2016; David and Archibald, 2016). In
306 contrast, in previous reports kinetoplastids constituted just a minor component of pelagic communities,
307 being significantly more abundant in (abyssal) benthic communities (Salani *et al.*, 2012; Sauvadet *et*
308 *al.*, 2010; Atkins *et al.*, 2000; Brown and Wolfe, 2006; Scheckenbach *et al.*, 2010). The only marine
309 habitat with a significant content of kinetoplastids reported so far are hypersaline anoxic basins
310 (Edgcomb *et al.*, 2011). From the global set of 124 examined sampling sites, we have confirmed that,
311 with few exceptions, kinetoplastids indeed constitute a small component of the marine plankton, since
312 their average relative abundance among eukaryotes was only 0.2%. Our results do not confirm the
313 increased presence of kinetoplastids in oxygen-depleted habitats, although two of the 14 most
314 abundant kinetoplastid OTUs were significantly more abundant at OMZ sites compared to other zones
315 (Fig. 2). Because only nine OMZ samples (7% of sampling sites) were available in our global dataset,
316 we have refrained from their more detailed analysis and included them among the MES samples,
317 where they fit based on the sampling depth.

318 However, low global kinetoplastid counts may still be an underestimation of the reality. It was
319 recently demonstrated in freshwater habitats that in metabarcoding studies using universal primers
320 kinetoplastids went largely undetected, yet were significantly more abundant when a specific set of
321 oligonucleotides and/or *in situ* hybridization was employed (Mukherjee *et al.*, 2015). In fact, they
322 dominated the eukaryotic communities in the latter case. Indeed, these protists were reported as being
323 highly abundant in the Atlantic Ocean using kinetoplastid-specific FISH probes (Morgan-Smith *et al.*,
324 2011). Our results based on the V9 region of 18S rRNA are in agreement with other metabarcoding
325 studies (mentioned above), which using other sets of oligonucleotides showed the relative low
326 abundance of kinetoplastids in the global plankton. However until comparative studies are performed,
327 we cannot exclude that, due to their divergent 18S rRNA sequences, they remain heavily
328 underestimated in the clone libraries.

329 The bulk of kinetoplastid abundance and diversity in the plankton falls into the neobodonid

clade (~98% of all reads), while the endosymbiotic *Perkinsela* of the Prokinetoplastina clade is diverse (~8% of all kinetoplastid OTUs) but very rare in the plankton (only 0.2% of all kinetoplastid reads). Neobodonids are also morphologically the most diverse group among four bodonid clades (Lukeš *et al.*, 2014). It is worth noting that neobodonids are the main kinetoplastid clade in the soil (Ekelund *et al.*, 2001; Glaser *et al.*, 2014). Representatives of the genus *Neobodo* were dominant kinetoplastids in most stations sampled across all depth zones, including samples from the oxygen-depleted waters. The high abundance and global distribution of *Neobodo* species in our dataset was not unexpected, as they (and namely representatives of the *N. designis* complex) are known to be widely present in both marine and freshwater habitats (von der Heyden *et al.*, 2004; Lee and Patterson, 2002; Lee and Patterson, 1998; Scheckenbach *et al.*, 2006). Together with another neobodonid, *Rhynchomonas* spp., they were also virtually the only kinetoplastids for which any significant interactions could be retrieved from the global interactome of Lima-Mendez *et al*. (2015). Apart from the expected co-occurrence with their supposed bacterial prey and other bacterivorous protists, such as ciliates, choanoflagellates and MAST, we have found a surprisingly high number of interactions with various species of syndinians (MALV). These are typically parasites of various planktonic organisms. Although neobodonids were mostly found in samples from size fractions smaller than 20 μm and are thus possibly of very small cell size, we cannot exclude the possibility that neobodonids are target hosts of these parasitic syndinians.

Kinetoplastids are represented by hundreds of OTUs, yet just 14 abundant OTUs accounted for 93% of all reads. A total of 13 of the hyper-abundant OTUs are assigned to neobodonids and one to eubodonids. The pattern of relatively few dominant and globally distributed OTUs is very similar to that observed for diplonemids which, on the other hand, have diversified into tens of thousands of OTUs (Lukeš *et al.*, 2015; Flegontova *et al.*, 2016; David and Archibald, 2016). Distribution of several abundant neobodonid OTUs shows significant geographic signature, which may be caused by their physiology and/or association with specific prey. Unfortunately, we could not find any apparent interaction with other organisms explaining the aforementioned biogeographic signatures. This may be caused by the fact that while neobodonids were mostly found in the deeper layers, the *in-silico* interactome (Lima-Mendez *et al.*, 2015) is available only for samples from the euphotic zone, and thus does not include the majority of the kinetoplastid data.

Two other kinetoplastid groups are worth mentioning. One is the obligatory parasitic trypanosomatids, which constitute an absolute majority of the terrestrial diversity of kinetoplastids (Maslov *et al.*, 2013) and are likely one of the most diverse parasitic protists, yet are extremely rare in marine samples. This is not surprising because marine trypanosomatids are blood parasites mainly of fish (Woo, 2003; Lom and Dyková, 1992), a segment of marine life not specifically targeted by *Tara*

Oceans and thus missing from out dataset. The second case of Prokinetoplastina is more interesting. These early-branching kinetoplastids are represented by *Ichthyobodo* spp., ectoparasites of freshwater and marine fish, and *Perkinsela* spp., endosymbionts of amoebae (Dyková *et al.*, 2003; Simpson *et al.*, 2006; Tanifuji *et al.*, 2011; Feehan *et al.*, 2013), with the latter representing a vast majority of reads assigned to this clade, yet still only 0.2% of all kinetoplastids. These amoebae of the *Paramoeba* genus are ectoparasites of fish gills and thus are also expected to be missing in our samples due to the sampling strategy. *Perkinsela* (but not its host amoebae) fell below the abundance threshold set in the global interactome, but read counts for *Perkinsela* (all OTUs) and *Paramoeba* (all OTUs) are strongly correlated in our dataset (Pearson's $r = 0.93$), while any other kinetoplastid clade or any of the 14 highly abundant OTUs show no correlation with *Paramoeba* (|Pearson's r| up to 0.11). This strongly indicates that *Paramoeba* occurs frequently in cysts.

All in all, kinetoplastids follow a similar pattern as diplonemids: both groups show higher relative abundance in the MES zone and are dominated by just a handful of very abundant cosmopolitan OTUs. However, regarding their diversity there are significant differences. While diplonemids have undergone extreme (and likely recent) speciation into tens of thousands of OTUs of (mostly) low abundance, the same process appears not to have taken place in marine kinetoplastids. While the two major and opposing views on the distribution of protists, the "everything is everywhere" versus the "endemicity rules" theories fail to explain such a discrepancy, we should refrain from speculations until we learn more about the lifestyles of diplonemids. Notwithstanding, the sample richness and depth of sequencing presented here provide the first global and comprehensive insights into the qualitative and quantitative composition of kinetoplastids in the world's ocean.

Accession numbers

The project number for the sequences reported in this paper is EBI: XXXXXXXX.

Author contribution

AH, PF and JL designed the study; JP, CDV and CB provided the data; OF, PF, SM, and AH performed the data analysis; and AH, PF, OF, JL, and CB wrote the manuscript.

Acknowledgements

This work was supported by the ERC CZ LL1601 (to J.L.), and the Czech Grant Agency projects Nos. 15-17643S (to A.H.) and 14-23986S (to J.L.). C.B. acknowledges funding from the ERC Advanced Award ''Diatomite'', the Louis D Foundation, and the French Government ''Investissements

d'Avenir'' programmes MEMO LIFE (ANR-10-LABX-54), PSL* Research University (ANR-1253 11-IDEX-0001-02). CB also thanks the Radcliffe Institute of Advanced Study at Harvard University for a scholars fellowship during the 2016-2017 academic year. We thank the commitment of the following people and sponsors: CNRS (in particular Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEAthe French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008) and FRANCE GENOMIQUE (ANR-10-INBS-09-08), Agence Nationale de la Recherche, and European Union FP7 (MicroB3/No.287589), We also thank the support and commitment of agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the Eléctricité de France (EDF) Foundation, Fondation pour la recherche sur la biodiversité (FRB) , the Foundation Prince Albert II de Monaco, the Tara Foundation, its schooner and teams. We thank MERCATOR-CORIOLIS and ACRI-ST for providing daily satellite data during the expedition. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who graciously granted sampling permissions. *Tara* Oceans would not exist without continuous support from 23 institutes (http:// http://oceans.taraexpeditions.org/en/m/science/labs-involved/). The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the *Tara* Oceans expedition sampled in. This article is contribution number ZZZ of *Tara* Oceans.

Conflict of interest: The authors declare no conflict of interest

References

Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, *et al.* (2012). The revised classification of eukaryotes. *J Eukaryot Microbiol* **59**: 429–493.

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA Genes. *PLoS One* **4**. doi: 10.1371/journal.pone.0006372.

Atkins MS, Teske AP, Anderson OR. (2000). A survey of flagellate diversity at four deep-sea hydrothermal vents in the Eastern Pacific Ocean using structural and molecular approaches. *J Eukaryot Microbiol* **47**: 400–11.

Le Bescot N, Mahé F, Audic S, Dimier C, Garet MJ, Poulain J, *et al.* (2016). Global patterns of

pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. *Environ Microbiol* **18**: 609–26.

Boenigk J, Ereshefsky M, Hoef-Emden K, Mallet J, Bass D. (2012). Concepts in protistology: Species definitions and boundaries. *Eur J Protistol* **48**: 96–102.

Brown PB, Wolfe G V. (2006). Protist genetic diversity in the acidic hydrothermal environments of Lassen Volcanic National Park, USA. *J Eukaryot Microbiol* **53**: 420–31.

Brum JR, Ignacio-Espinoza JC, Roux S, Doulcier G, Acinas SG, Alberti A, *et al.* (2015). Ocean plankton. Patterns and ecological drivers of ocean viral communities. *Science* **348**: 1261498.

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

David V, Archibald JM. (2016). Evolution: Plumbing the depths of diplonemid diversity. *Curr Biol* **26**: R1290–R1292.

Dyková I, Fiala I, Lom J, Lukeš J. (2003). *Perkinsiella* amoebae-like endosymbionts of *Neoparamoeba* spp., relatives of the kinetoplastid *Ichthyobodo*. *Eur J Protistol* **39**: 37–52.

Edgar RC. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–1.

Edgcomb V, Orsi W, Bunge J, Jeon S, Christen R, Leslin C, *et al.* (2011). Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J* **5**: 1344–1356.

Ekelund F, Rønn R, Griffiths BS. (2001). Quantitative estimation of flagellate community structure and diversity in soil samples. *Protist* **152**: 301–314.

Feehan CJ, Johnson-Mackinnon J, Scheibling RE, Lauzon-Guay J-S, Simpson AGB. (2013). Validating the identity of *Paramoeba invadens*, the causative agent of recurrent mass mortality of sea urchins in Nova Scotia, Canada. *Dis Aquat Organ* **103**: 209–27.

Finlay BJ, Fenchel T. (2004). Cosmopolitan metapopulations of free-living microbial eukaryotes. *Protist* **155**: 237–44.

Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, *et al.* (2016). Extreme diversity of diplonemid eukaryotes in the Ocean. *Curr Biol* **26**: 3060–3065.

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

459 Gawryluk RMR, del Campo J, Okamoto N, Strassert JFH, Lukeš J, Richards TA, *et al.* (2016).
460 Morphological identification and single-cell genomics of marine diplonemids. *Curr Biol* **26**: 3053–
461 3059.

462 Glaser K, Kuppardt A, Krohn S, Heidtmann A, Harms H, Chatzinotas A. (2014). Primer pairs for the
463 specific environmental detection and T-RFLP analysis of the ubiquitous flagellate taxa Chrysophyceae
464 and Kinetoplastea. *J Microbiol Methods* **100**: 8–16.

465 Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, *et al.* (2013). The Protist Ribosomal
466 Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with
467 curated taxonomy. *Nucleic Acids Res* **41**: D597-604.

468 Hamilton PB, Gibson WC, Stevens JR. (2007). Patterns of co-evolution between trypanosomes and
469 their hosts deduced from ribosomal RNA and protein-coding gene phylogenies. *Mol Phylogenet Evol*
470 **44**: 15–25.

471 von der Heyden S, Cavalier-Smith T. (2005). Culturing and environmental DNA sequencing uncover
472 hidden kinetoplastid biodiversity and a major marine clade within ancestrally freshwater *Neobodo*
473 *designis*. *Int J Syst Evol Microbiol* **55**: 2605–21.

474 von der Heyden S, Chao EE, Vickerman K, Cavalier-Smith T. (2004). Ribosomal RNA phylogeny of
475 bodonid and diplonemid flagellates and the evolution of euglenozoa. *J Eukaryot Microbiol* **51**: 402–16.

476 Hirose E, Nozawa A, Kumagai A, Kitamura S. (2012). *Azumiobodo hoyamushi* gen. nov. et sp. nov.
477 (Euglenozoa, Kinetoplastea, Neobodonida): a pathogenic kinetoplastid causing the soft tunic syndrome
478 in ascidian aquaculture. *Dis Aquat Organ* **97**: 227–235.

479 Je Lee W, Patterson DJ. (1998). Diversity and geographic distribution of free-living heterotrophic
480 flagellates – analysis by PRIMER. *Protist* **149**: 229–244.

481 Karsenti E, Acinas SG, Bork P, Bowler C, de Vargas C, Raes J, *et al.* (2011). A holistic approach to
482 marine eco-systems biology. *PLoS Biol* **9** doi: 10.1371/journal.pbio.1001177.

483 Katoh K, Standley DM. (2013). MAFFT Multiple Sequence Alignment Software Version 7:
484 Improvements in performance and usabiltiy. *Mol Biol Evol* **30**: 772–780.

485 Kumagai A, Ito H, Sasaki R. (2013). Detection of the kinetoplastid *Azumiobodo hoyamushi*, the
486 causative agent of soft tunic syndrome, in wild ascidians *Halocynthia roretzi*. *Dis Aquat Organ* **106**:
487 267–71.

488 Lee WJ, Patterson DJ. (2002). Abundance and biomass of heterotrophic flagellates, and factors

controlling their abundance and distribution in sediments of Botany Bay. *Microb Ecol* **43**: 467–481.

Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, *et al.* (2015). Determinants of community structure in the global plankton interactome. *Science* 3**48**: 1262073_1-1262073_9.

Lom J, Dyková I. (1992). Protozoan parasites of fishes. Developments in aquaculture and fisheries science. Volume 26. Elsevier: Amsterdam.

López-García P, Philippe H, Gail F, Moreira D. (2003). Autochthonous eukaryotic diversity in hydrothermal sediment and experimental microcolonizers at the Mid-Atlantic Ridge. *Proc Natl Acad Sci U S A* **100**: 697–702.

Lukeš J, Flegontova O, Horák A. (2015). Diplonemids. *Curr Biol* **25**: R702–R704.

Lukeš J, Skalický T, Týč J, Votýpka J, Yurchenko V. (2014). Evolution of parasitism in kinetoplastid flagellates. *Mol Biochem Parasitol* **195**: 115–122.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2014). Swarm : robust and fast clustering method for amplicon-based studies PrePrints PrePrints. *PeerJ* 1–12.

Malviya S, Scalco E, Audic S, Vincent F, Veluchamy A, Poulain J, *et al.* (2016). Insights into global diatom distribution and diversity in the world's ocean. *Proc Natl Acad Sci U S A* **113**: E1516-25.

Maslov DA, Votýpka J, Yurchenko V, Lukeš J. (2013). Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. *Trends Parasitol* **29**: 43–52.

Moreira D, López-García P, Vickerman K. (2004). An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. *Int J Syst Evol Microbiol* **54**: 1861–1875.

Morgan-Smith D, Herndl G, van Aken H, Bochdansky A. (2011). Abundance of eukaryotic microbes in the deep subtropical North Atlantic. *Aquat Microb Ecol* **65**: 103–115.

Mukherjee I, Hodoki Y, Nakano S-I. (2015). Kinetoplastid flagellates overlooked by universal primers dominate in the oxygenated hypolimnion of Lake Biwa, Japan. *FEMS Microbiol Ecol* **91**. doi: 10.1093/femsec/fiv083.

Mutsuo I, Lopes dos Santos A, Gourvil P, Yoshikawa S, Kamiya M, Ohki K, *et al.* (2016). Diversity and oceanic distribution of Parmales and Bolidophyceae, a picoplankton group closely related to diatoms. *Isme J* **10**: 2419–34.

Pernice M, Giner C, Logares R. (2015). Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. *ISME J* **10**: 945–958.

519 Price MN, Dehal PS, Arkin AP. (2010). FastTree 2 - Approximately maximum-likelihood trees for
520 large alignments. *PLoS One* **5**. doi: 10.1371/journal.pone.0009490.

521 Salani FS, Arndt H, Hausmann K, Nitsche F, Scheckenbach F. (2012). Analysis of the community
522 structure of abyssal kinetoplastids revealed similar communities at larger spatial scales. *ISME J* **6**:
523 713–23.

524 Sauvadet AL, Gobet A, Guillou L. (2010). Comparative analysis between protist communities from
525 the deep-sea pelagic ecosystem and specific deep hydrothermal habitats. *Environ Microbiol* **12**: 2946–
526 2964.

527 Scheckenbach F, Hausmann K, Wylezich C, Weitere M, Arndt H. (2010). Large-scale patterns in
528 biodiversity of microbial eukaryotes from the abyssal sea floor. *Proc Natl Acad Sci U S A* **107**: 115–
529 120.

530 Scheckenbach F, Wylezich C, Mylnikov AP, Weitere M, Arndt H. (2006). Molecular comparisons of
531 freshwater and marine isolates of the same morphospecies of heterotrophic flagellates. *Appl Environ*
532 *Microbiol* **72**: 6638–43.

533 Simpson AGB, Stevens JR, Lukeš J. (2006). The evolution and diversity of kinetoplastid flagellates.
534 *Trends Parasitol* **22**: 168–174.

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

537 Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, *et al.* (2015). Ocean plankton.
538 Structure and function of the global ocean microbiome. *Science* **348**: 1261359.

539 Tanifuji G, Kim E, Onodera NT, Gibeault R, Dlutek M, Cawthorn RJ, *et al.* (2011). Genomic
540 characterization of *Neoparamoeba pemaquidensis* (Amoebozoa) and its kinetoplastid endosymbiont.
541 *Eukaryot Cell* **10**: 1143–6.

542 de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, *et al.* (2015). Eukaryotic plankton
543 diversity in the sunlit ocean. *Science (80-)* **348**: 1261605–1261605.

544 Villar E, Farrant GK, Follows M, Garczarek L, Speich S, Audic S, *et al.* (2015). Ocean plankton.
545 Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science* **348**:
546 1261447.

547 Woo PTK. (2003). *Cryptobia* (*Trypanoplasma*) salmositica and salmonid cryptobiosis. *J Fish Dis* **26**:
548 627–646.

549 Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ. (2015). Rethinking
550 the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science* **347**: 1257594–
551 1257594.

552

553

554 ## Table legends

555 **Table 1. Summary of kinetoplastid diversity and abundance by taxonomic groups**. The
556 taxonomic assignment of OTUs against an in-house reference database (see Methods) was performed
557 with ggsearch 36 according to de Vargas et al. (2015).

558

559 ## Figure legends

560 **Figure 1. Rarefaction curves for OTUs: OTU count vs. read number**. Slopes calculated for 10 last
561 data points are indicated in the legend on the right. Curves were constructed for the full Kinetoplastea
562 dataset, for the neobodonid clade and for its most abundant sub-groups: *Neobodo*, *Rhynchomonas*, and
563 unknown Neobodonida.

564

565 **Figure 2. Variation in average kinetoplastid abundance across depth zones, size fractions, and**
566 **geographical regions**. Only most abundant kinetoplastid clades and 14 most abundant OTUs were
567 considered. The bar plots show average relative abundance, with scale at the bottom of each column;
568 and pairs of the minus and asterisk symbols mark significant differences according to one-way
569 ANOVA. Because kinetoplastids were preferentially found in the smallest size fraction of 0.8-5 um
570 and in the mesopelagic zone, geographic variables were considered not only on the whole dataset, but
571 also separately on these subsets. Furthermore, because a different set of size fractions was taken in the
572 mesopelagic zone, the size variability was assessed in this zone separately. The following
573 abbreviations are used: SRF, surface zone; DCM, deep chlorophyll maximum zone; OMZ, oxygen
574 minimum zone; MES, mesopelagic zone; MS, Mediterranean Sea; RS, Red Sea; IO, Indian Ocean;
575 SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific
576 Ocean; NAO, North Atlantic Ocean.

577

578 **Figure 3. Factors driving abundance and diversity of kinetoplastids.** We performed a multi-way
579 ANOVA analysis to determine which variables drive relative abundance and diversity of
580 kinetoplastids and their most abundant sub-clades. The strongest influence we observed was size

fractions affecting abundance and diversity. Abundance was also significantly affected by depth, and in case of four OTUs also by oceanic provinces or latitude regions. On the other hand, diversity statistics were significantly affected by all four variables. *P*-values are coded as follows: full black, *p*-values < 0.001; chequered, *p*-values from 0.001 to 0.01; horizontal stripes, *p*-values from 0.01 to 0.05.

Figure 4. Variations in kinetoplastid diversity across depth zones, size fractions, and geographical regions. Only most abundant kinetoplastid clades were considered. The bar plots show various diversity indices (average richness, Shannon index, and evenness), and pairs of the minus and asterisk symbols mark significant differences according to one-way ANOVA. The analysis scheme and abbreviations are the same as in Fig. 2.

Figure 5. Analysis of cosmopolitan and rare OTUs. Occupancy values, i.e., the number of stations where an OTU was found, are plotted on the x axis, and average station evenness for these stations is plotted on the y axis. Bubble size represents a read count for a given OTU, and OTUs unique to one depth zone (**A**) or taxonomic group (**B**) are color-coded according to the legend.

Table 1. Summary of kinetoplastid diversity and abundance by taxonomic groups.

			richness, OTUs	richness, %	abundance, reads
Kinetoplastea			512	100	1570025
	Metakinetoplastina		440	85.9	1566578
		Neobodonida	360	70.3	1544278
		unknown Neobo	172	33.6	322763
		Neobodo	127	24.8	928002
		Rhynchomonas	36	7	285766
		Azumiobodo	13	2.5	7128
		Rhynchobodo	12	2.3	619
		Eubodonida	35	6.8	20864
		Parabodonida	17	3.3	715
		Parabodo	14	2.7	160
		Procryptobia	3	0.6	555
		Trypanosomatida	28	5.5	721
	Prokinetoplastina		72	14.1	3447
		Perkinsela	41	8	3094
		unknown Prokin	18	3.5	178
		Ichthyobodo	13	2.5	175

abundance, %
100
99.8
98.4
20.6
59.1
18.2
0.5
0
1.3
0
0
0
0
0.2
0.2
0
0

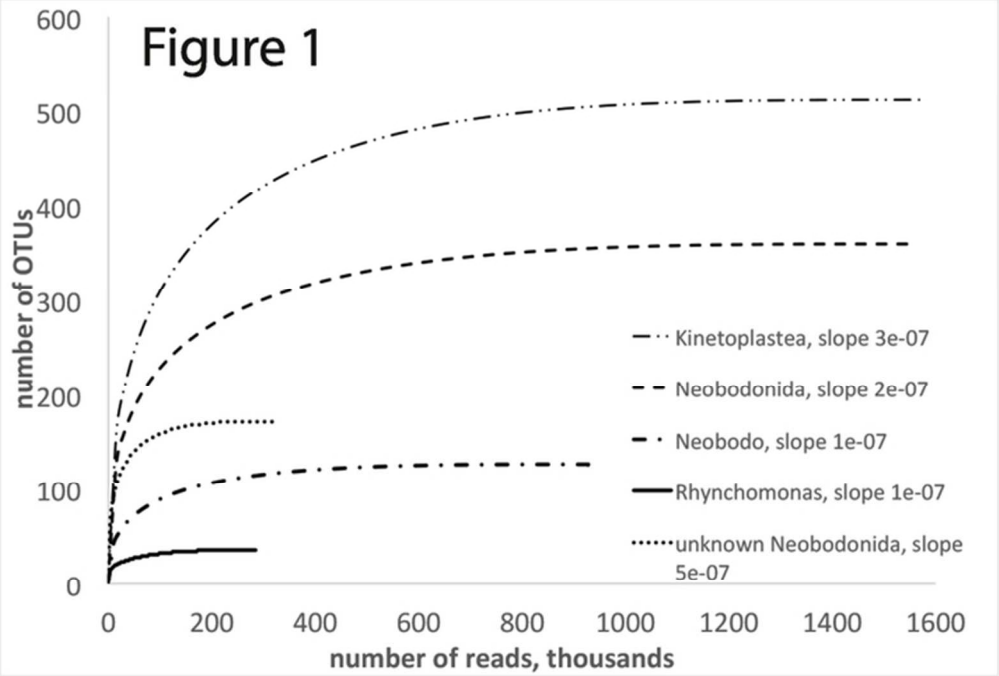


Figure 1. Rarefaction curves for OTUs: OTU count vs. read number. Slopes calculated for 10 last data points are indicated in the legend on the right. Curves were constructed for the full Kinetoplastea dataset, for the neobodonid clade and for its most abundant sub-groups: Neobodo, Rhynchomonas, and unknown Neobodonida.

60x41mm (300 x 300 DPI)

Figure 2

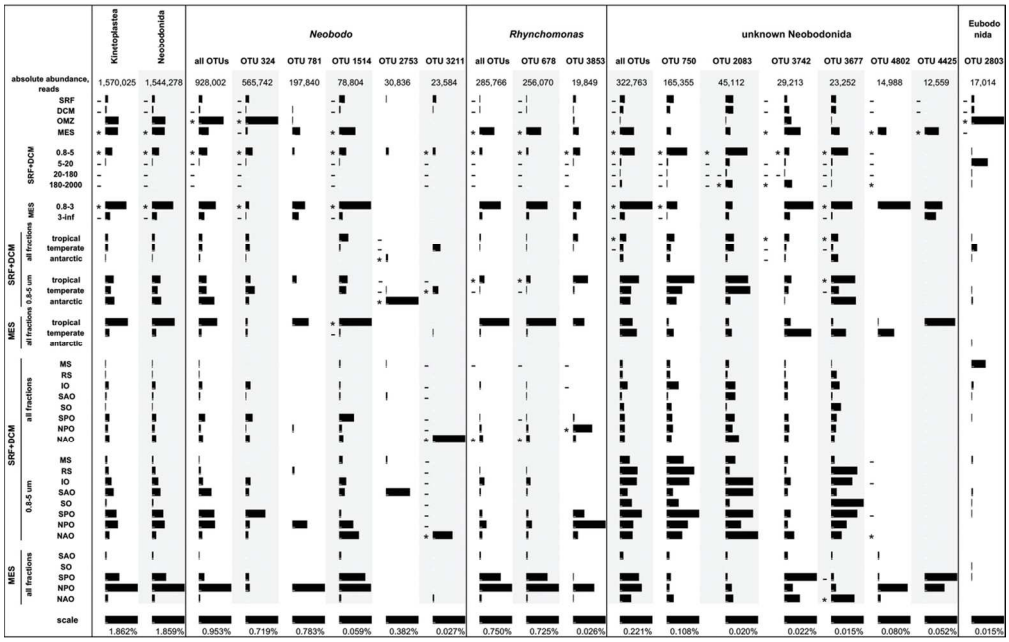


Figure 2. Variation in average kinetoplastid abundance across depth zones, size fractions, and geographical regions. Only most abundant kinetoplastid clades and 14 most abundant OTUs were considered. The bar plots show average relative abundance, with scale at the bottom of each column; and pairs of the minus and asterisk symbols mark significant differences according to one-way ANOVA. Because kinetoplastids were preferentially found in the smallest size fraction of 0.8-5 um and in the mesopelagic zone, geographic variables were considered not only on the whole dataset, but also separately on these subsets. Furthermore, because a different set of size fractions was taken in the mesopelagic zone, the size variability was assessed in this zone separately. The following abbreviations are used: SRF, surface zone; DCM, deep chlorophyll maximum zone; OMZ, oxygen minimum zone; MES, mesopelagic zone; MS, Mediterranean Sea; RS, Red Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean.

122x84mm (300 x 300 DPI)

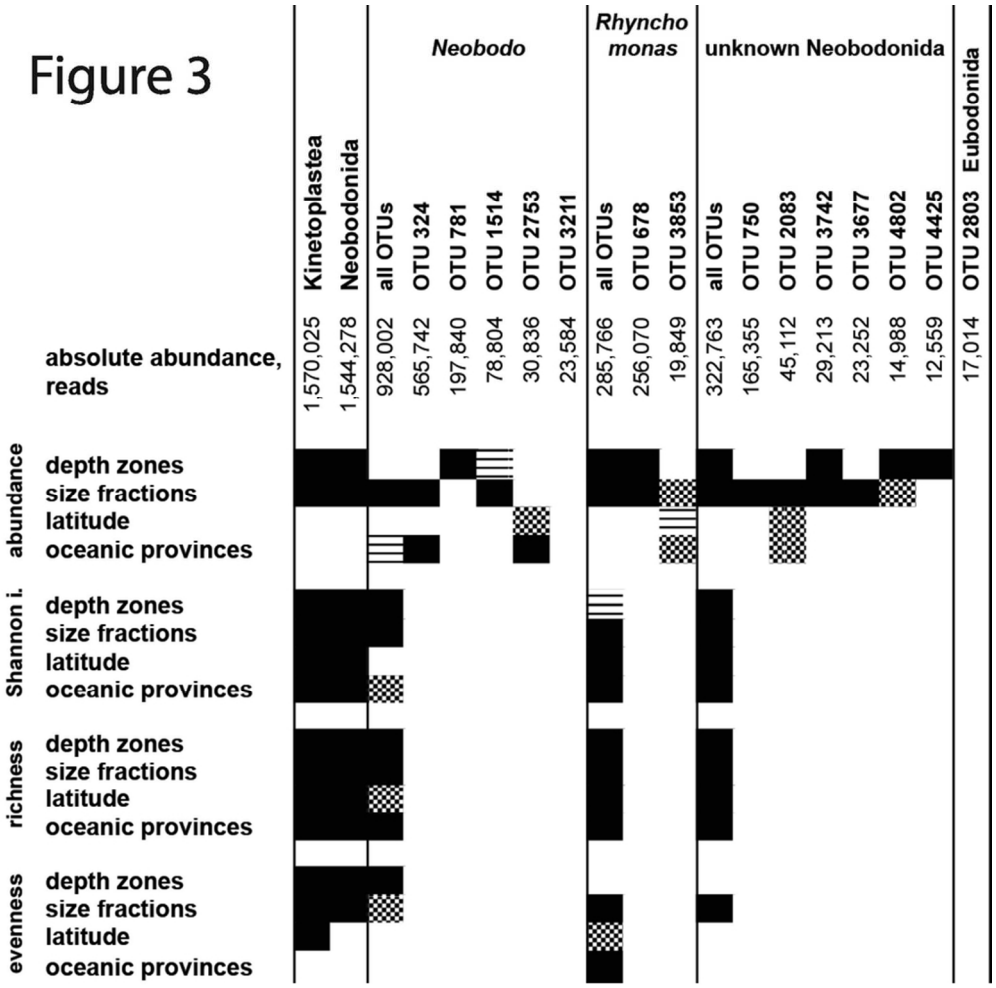


Figure 3. Factors driving abundance and diversity of kinetoplastids. We performed a multi-way ANOVA analysis to determine which variables drive relative abundance and diversity of kinetoplastids and their most abundant sub-clades. The strongest influence we observed was size fractions affecting abundance and diversity. Abundance was also significantly affected by depth, and in case of four OTUs also by oceanic provinces or latitude regions. On the other hand, diversity statistics were significantly affected by all four variables. P-values are coded as follows: full black, p-values < 0.001; chequered, p-values from 0.001 to 0.01; horizontal stripes, p-values from 0.01 to 0.05.

87x85mm (300 x 300 DPI)

Figure 4

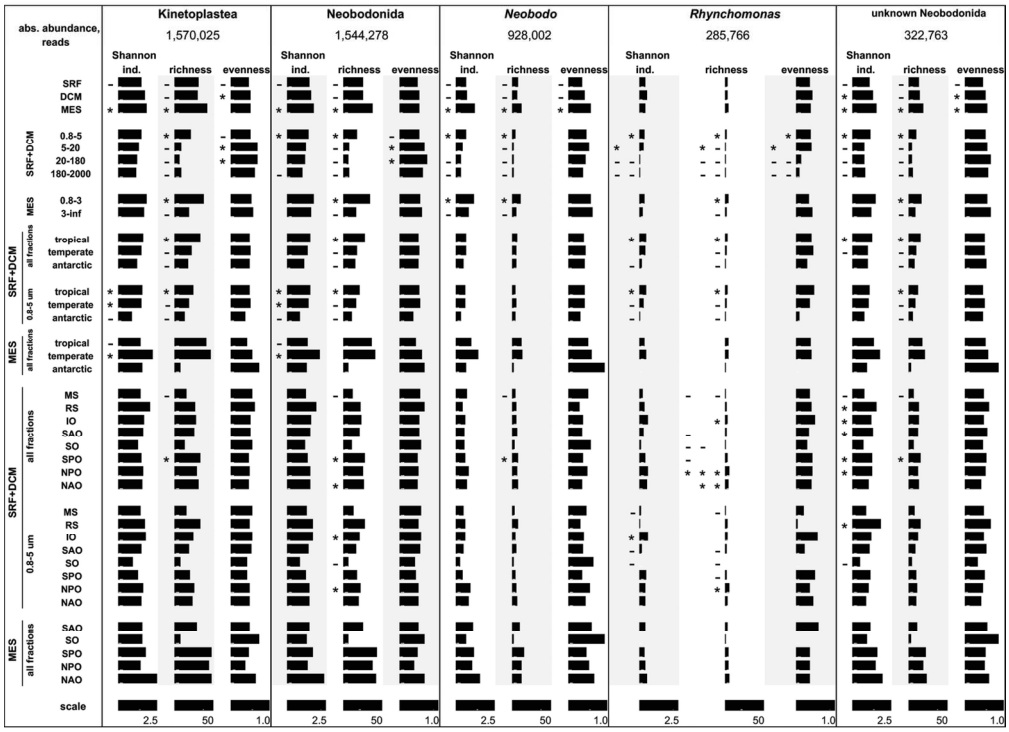


Figure 4. Variations in kinetoplastid diversity across depth zones, size fractions, and geographical regions. Only most abundant kinetoplastid clades were considered. The bar plots show various diversity indices (average richness, Shannon index, and evenness), and pairs of the minus and asterisk symbols mark significant differences according to one-way ANOVA. The analysis scheme and abbreviations are the same as in Fig. 2.

138x107mm (300 x 300 DPI)

Figure 5

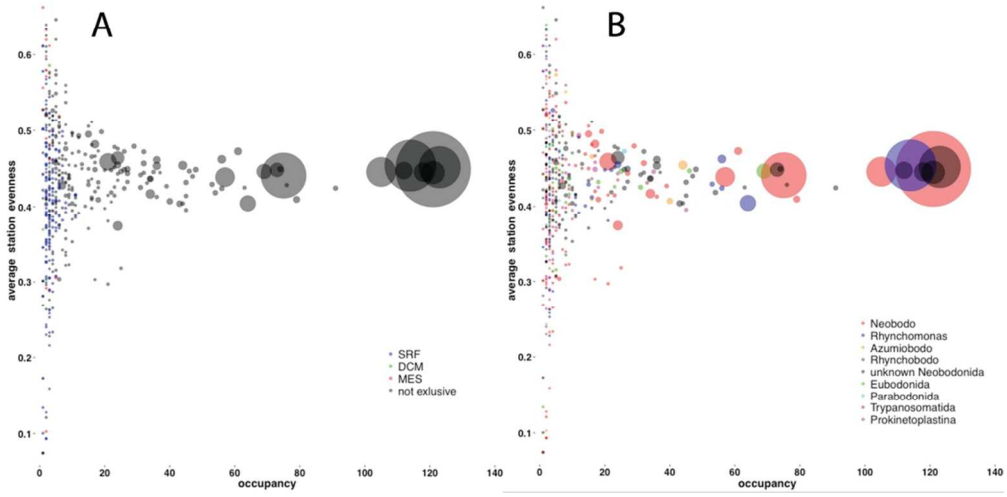


Figure 5. Analysis of cosmopolitan and rare OTUs. Occupancy values, i.e., the number of stations where an OTU was found, are plotted on the x axis, and average station evenness for these stations is plotted on the y axis. Bubble size represents a read count for a given OTU, and OTUs unique to one depth zone (A) or taxonomic group (B) are color-coded according to the legend.

99x55mm (300 x 300 DPI)

5.5 Manuscript II

Flegontova O, Karnkowska A, Kolisko M, Lax G, Maritz JM, Panek T, Carlton JM, Cepicka I, Horák A, Keeling PJ, Lukeš J, Simpson AGB, Tai V (manuscript in preparation) Excavata in EukRef, a novel curated database of small subunit rRNA gene sequences.

Below we present two chapters of the manuscript describing an updated reference taxonomy of Excavata, a part of the EukRef project (Berney et a. 2017). Additional chapters are to be contributed by the other co-authors.

Abstract

We provide a substantially revised annotation of kinetoplastid and diplonemid SSU rRNA sequences deposited in the GenBank database and a revised taxonomy for these protist groups, following previously published suggestions and a reference tree including all sequences longer than 500 nt. For Neobodonida, by far the largest clade of marine kinetoplastids, we kept original taxonomic annotations for 8% sequences, improved annotations for poorly annotated environmental sequences (90%), and corrected annotation for 2% sequences. For Trypanosomatida, a large clade of terrestrial kinetoplastid parasites, we kept the original taxonomic annotations for 92% sequences, provided annotations for 6% poorly annotated sequences, and corrected annotations for 2% of them. We suggest dividing the Diplonemea clade into four taxonomic groups: Diplonemidae, Hemistasiidae, DSPD I and DSPD II. Here we provide a more detailed annotation for 97% diplonemid sequences. Annotation of just 12 out of 433 diplonemid sequences remained unchanged, and one wrong annotation has been corrected.

Methods

For the construction of a curated reference database of excavate 18S rDNA sequences we used the following approach. First, for each major subclade of Excavata, e.g. Diplonemea, a core set of 18S rDNA sequences (reliably annotated and spanning the diversity of the group) was used as the initial query for an iterative search in the GenBank database, implemented in the BlastCircle v.0.3 script (del Campo & Ruiz-Trillo 2013) (http://eukref.org/curation-pipeline-overview/). Second, the output sequences were clustered with the 97% identity threshold using USEARCH, resulting in a set of 'seed' rDNAs, i.e. representatives of each

sequence cluster. Third, MAFFT v.7.245 (Katoh & Standley 2013) with the '--auto' option and trimAl v.1.2 (Capella-Gutierrez et al. 2009) with the '-gt 0.3' and '-st 0.001' options were used to make and prune sequence alignments, including a distant eukaryotic outgroup. Fourth, FastTree v.2.1.8 (Price et al. 2010) was used to make a preliminary maximum likelihood tree. Seeds and corresponding sequence clusters falling outside of the target clade (e.g. Diplonemea) were removed subsequently, and the clustering, alignment, and tree building steps were repeated a number of times until no sequences falling between the outgroup and the Excavata clade were left. The final alignment was used to build a maximum likelihood tree with RAxML v.8.2.3 (Stamatakis 2014) with the following options: phylogenetic model GTR+CAT+I; 25 rate categories; model optimization precision, 0.001; a random starting tree; 1,000 random bootstrap replicates and 200 iterations of the maximum likelihood algorithm.

Kinetoplastea

In our tree, the phylum Kinetoplastea has a bootstrap support of 94% and consists of 2,339 sequences (Figure 1). Most sequences (2,153) were correctly annotated as kinetoplastids, and 187 sequences were poorly annotated as 'uncultured eukaryote', or 'uncultured marine eukaryote', or 'uncultured euglenozoan'. The phylum Kinetoplastea is traditionally sub-divided into three highly supported clades (Moreira et al. 2004): i/ Metakinetoplastina (99% BS) comprising 2,089 sequences; ii/ Prokinetoplastina with 100% BS, composed of 235 sequences; iii/ a clade with 100% BS containing 15 environmental sequences, with no cultivable representative. The branching order of these three clades cannot be resolved in the 18S rDNA-based tree. Sequences of the third environmental-only clade were detected for the first time by Lopez-Garcia and co-authors (2003) in samples from a marine hydrothermal vent chimney; they appeared as a separate group at the base of Kinetoplastea. Representatives of this clade were later found in extreme marine biomes, such as deep-sea sediments and deep-sea plankton below 5,189 m. We suggest naming this clade as KIN1 (= Kinetoplastea clade 1).

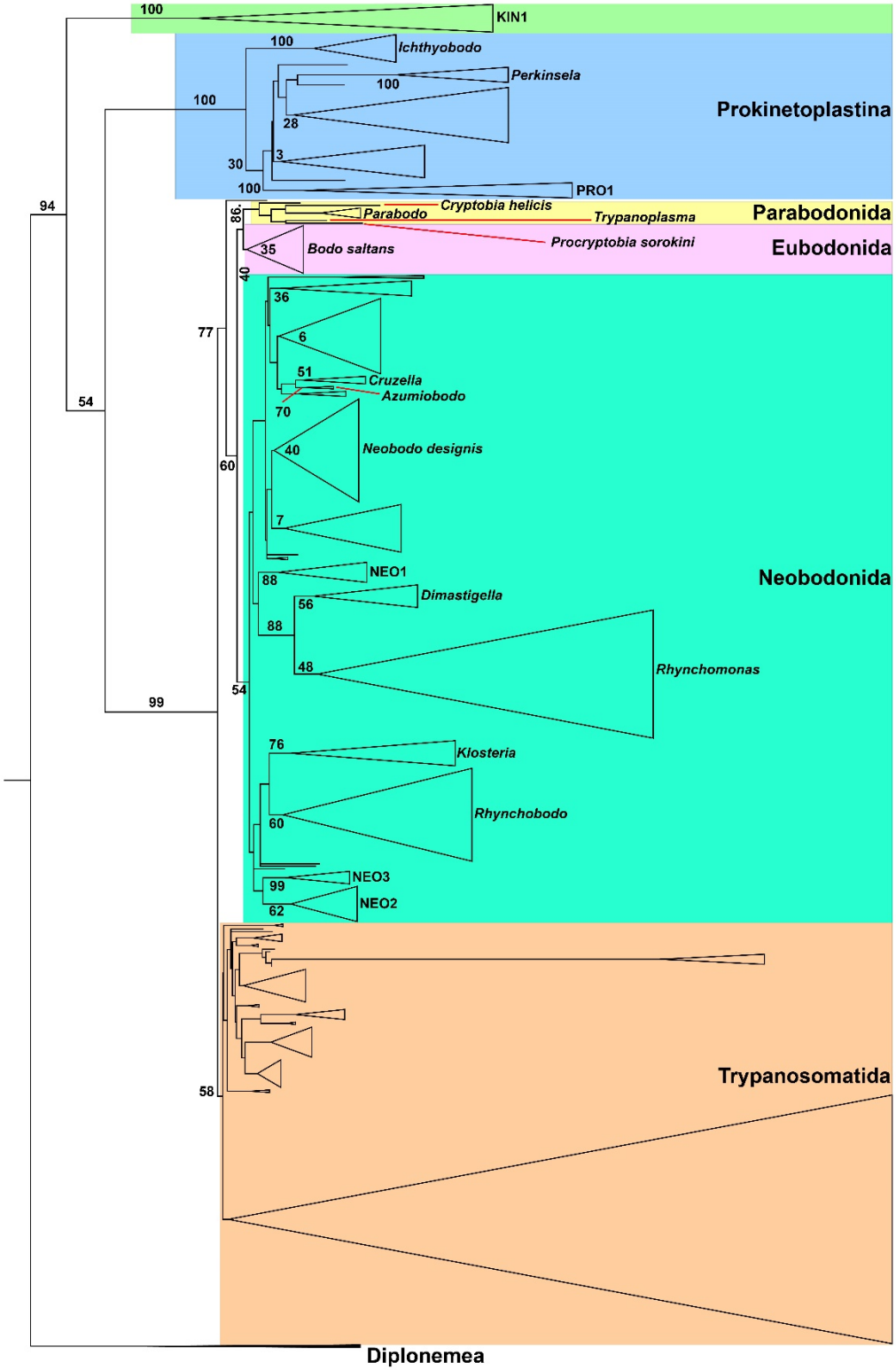


Figure 1. A maximum likelihood tree of 2,339 kinetoplastid sequences extracted form GenBank constructed using RAxML (model GTR+CAT+I+25 rate categories, 1000 rapid bootstrap replicates). For reducing the tree size, only seed sequences representing clusters with the 97% identity threshold were included. The diplonemid clade used as an outgroup is collapsed. Bootstrap support values were omitted at some nodes for the lack of space. The major kinetoplastid clades are highlighted in color.

The clade Prokinetoplastina can be further sub-divided into three distinct sub-clades with 100% BS: i/ 57 sequences, all of them annotated as *Ichthyobodo* sp.; ii/ 41 sequences annotated as *Perkinsiella*-like species or *Perkinsiella*-like organism (PLO); iii/ an environmental-only clade of nine sequences. All *Ichthyobodo* sequences were derived from gills or skin of captured fish in freshwater or marine biomes, while they were absent from environmental libraries, supporting the view that this genus is a strict ectoparasite of fish (Isaksen et al. 2012). *Perkinsiella* was originally described as an organelle of amoebae of the genera *Paramoeba*, *Neoparamoeba* and *Janickina*, which are themselves ectoparasites of marine fish, and only later was recognized as an intracellular aflagellar kinetoplastid (Hollande 1980; Tanifuji et al. 2011). The genus name *Perkinsiella* was already taken by an insect, therefore Dykova et al. (2008) renamed the genus to *Perkinsela*, and following this suggestion we renamed all sequences. *Perkinsela* spp. are obligatory endosymbionts, which is reflected by the fact that sequences were derived from amoeba-containing samples, while they were absent from environmental libraries. The environmental-only clade includes a handful of sequences from diverse biomes: deep-sea plankton, hydrothermal vents, a sulfidic anoxic fjord, and two freshwater lakes. One sequence was annotated as derived from a symbiont of an abyssal clam *Calyptogena magnifica*. Since this clade has remained unnamed, we propose to name it PRO1 (= Prokinetoplastina clade 1). The remaining 128 environmental sequences do not form well-supported clades within Prokinetoplastina, are not grouped with described genera, and were poorly annotated as uncultured kinetoplastids or eukaryotes. A great majority of these sequences were derived from deep-sea sediments or plankton, but five sequences came from freshwater and terrestrial biomes. Life style of these protists remains unknown.

The clade Metakinetoplastina has been sub-divided into four taxonomic groups: Eubodonida, Parabodonida, Neobodonida, and Trypanosomatida (Moreira et al. 2004). All four clades have been recovered in our tree, albeit with low boostrap supports and an unresolved branching order. The smallest and best supported clade, comprising 36 sequences, is Parabodonida (86% BS). We have further sub-divided parabodonids into five groups. The first one is genus *Parabodo* (100% BS; 13 sequences), which unites free-living organisms from terrestrial and freshwater biomes and potential parasites found in plant sap. Three sequences were mis-annotated as *Bodo* sp. or *Neobodo curvifilus*. *Parabodo* forms a well-supported clade with the second group: a single sequence belonging to *Cryptobia helicis*, an endoparasite of snails (Lukeš et al. 1998; von der Heyden et al. 2004). The third group of 15

sequences mostly from marine habitats has been assigned to *Procryptobia sorokini*. The fourth group is genus *Trypanoplasma*, parasitizing the blood of fish, and contains only 4 sequences. We have transferred 3 sequences from genus *Cryptobia* to *Trypanoplasma*, following Moreira et al. (2004). The fifth parabodonid group is represented by two basal-branching environmental sequences from oxygen-depleted marine sediment.

Eubodonids are free-living bacteriovorous protists found in soil, in freshwater and marine habitats, which were in our tree recovered as a monophyletic group but with a very low bootstrap support (35%). Most of 80 sequences were originally annotated as uncultured bodonids, uncultured eukaryotes, or *Bodo saltans* - a genetically diverse morphospecies likely representing multiple species (Moreira et al. 2004, von der Heyden 2004). Since in our tree eubodonids cannot be robustly split into sub-clades, and since sequences annotated as *B. saltans* are scattered in several branches, we decided to limit our annotation to the level of Eubodonida (one sequence was mis-annotated as *Neobodo designis*).

Neobodonida is the largest group of kinetoplastids represented in the database by 1,030 sequences, yet brought together with a low bootstrap support of 54%. About 90% of these sequences are poorly annotated as uncultured kinetoplastids or uncultured eukaryotes. Our tree does not resolve any high-order branching clades within neobodonids, but 10 clades with moderate or high support can be recognized: 9 of them described by von der Heyden and Cavalier-Smith (2005), and one (*Azumiobodo*) that was identified more recently (Hirose et al. 2012). The clades are listed below:

i/ *Azumiobodo* (70% BS; 3 sequences annotated as *Azumiobodo hoyamushi* (Hirose et al. 2012), for one sequence we changed the annotation from uncultured eukaryote to *Azumiobodo*);

ii/ *Cruzella* (51% BS; one sequence annotated as *Cruzella marina* (Dolezel et al. 2000), for six sequences we changed annotations from uncultured kinetoplastids to *Cruzella*);

iii/ *Dimastigella* (56% BS; 9 sequences annotated as *Dimastigella* (Berchtold et al. 1994), for 66 sequences we changed annotations from uncultured kinetoplastids to *Dimastigella*). Besides, one sequence was mis-annotated as *Rhynchobodo*, and one was annotated as *Phanerobia pelophila*. Following von der Heyden et al. (2004), we re-annotated the single *Phanerobia* sequence obtained in that study as *Dimastigella*;

iv/ *Klosteria* (76% BS; one sequence annotated as *Klosteria bodomorphis* (Nikolaev et al. 2003), for 30 sequences we changed annotations from uncultured kinetoplastids to *Klosteria*);

v/ *Neobodo* (40% BS; 39 sequences annotated as *Neobodo*, for 84 sequences we changed annotations from uncultured eukaryotes or uncultured kinetoplastids to *Neobodo*). Two sequences within the clade were mis-annotated as *Bodo*. Thirteen sequences, most of them originally annotated as *Neobodo designis*, fell into three other clades (NEO1, NEO2 and NEO3; see below), which corresponded to the clades defined by von der Heyden and Cavalier-Smith (2005). We decided to keep the name *Neobodo* only for the largest clade since *Neobodo designis* is a genetically diverse morphospecies, and same as *B. saltans* should be eventually split into multiple species (von der Heyden et al. 2004, von der Heyden and Cavalier-Smith 2005);

vi/ *Rhynchobodo* (60% BS; 3 sequences annotated as *Rhynchobodo* (Lukeš et al. 1997), for 74 sequences we changed annotations from uncultured kinetoplastids to *Rhynchobodo*). Two additional sequences originally annotated as *Rhynchobodo* fell outside of this clade. It was previously suspected that *Rhynchobodo* sp. ATCC50359 does not belong to genus *Rhynchobodo* (Simpson et al. 2002). An 18S rDNA-based tree has grouped this isolate with other *Rhynchobodo* species (von der Heyden et al. 2004), but according to our tree we re-annotated it as *Dimastigella. Rhynchobodo*-like strain NZ, isolated by von der Heyden et al. (2004) and assigned to the *Rhynchobodo* clade, fell out of this clade in our tree and was re-annotated as a neobodonid;

vii/ *Rhynchomonas* (48% BS; 28 sequences annotated as *Rhynchomonas nasuta*, for 439 sequences we changed annotations from uncultured kinetoplastids or uncultured eukaryotes to *Rhynchomonas*). One sequence falling into this clade was originally annotated as *Cryptaulax* sp. ATCC50746 (von der Heyden et al. 2004), however it was not included into phylogenetic trees in that study. We re-annotated it as *Rhynchomonas*;

viii/ NEO1 (88% BS; 6 sequences annotated as *Neobodo designis* were re-annotated as NEO1, for 24 sequences annotations were changed from uncultured eukaryotes or uncultured kinetoplastids to NEO1). This clade corresponds to a small clade of 5 *N. designis* sequences (von der Heyden and Cavalier-Smith 2005) that was separated from the major *N. designis* clade. This way a monophyletic clade was generated for *N. designis*;

ix/ NEO2 (62% BS; 4 sequences annotated as *Neobodo* sp. and one annotated as *Cryptaulaxoides*-like were re-annotated as NEO2, and for 25 sequences annotations were changed from uncultured kinetoplastids or uncultured eukaryotes to NEO2). This clade so far lacked any taxonomic status (von der Heyden et al. 2004, von der Heyden and Cavalier-Smith 2005) and is now labeled NEO2;

x/ NEO3 (99% BS; 3 sequences annotated as *Neobodo* sp. were re-annotated as NEO3, and for 18 sequences annotations were changed from uncultured kinetoplastids to NEO3). This clade so far lacked any taxonomic status (von der Heyden et al. 2004, von der Heyden and Cavalier-Smith 2005) and is now labeled NEO3.

There are multiple neobodonid sequences which we could not assign reliably to any clades described above, and therefore they remain annotated simply as Neobodonida. This diverse group comprises 159 sequences, most of them originally annotated as uncultured kinetoplastids, few as uncultured eukaryotes, and one as *Rhynchobodo* sp.

In conclusion, we provide a substantially revised taxonomy of Neobodonida, mostly following the suggestions by von der Heyden et al. (2004) and von der Heyden and Cavalier-Smith (2005). We kept original taxonomic annotations for 8% sequences, improved annotations for poorly annotated environmental sequences (90%), and corrected annotation for 2% sequences. The overwhelming majority of neobodonids are free-living marine flagellates, benthic or pelagic, using bacteria as a food source (Lukeš et al. 2014). Only one small clade, *Azumiobodo*, is a parasite of ascidians. Below we group the clades by their biomes: i/ exclusively marine (*Azumiobodo, Cruzella, Klosteria*); predominantly marine (*Dimastigella*, *Rhynchobodo*, *Rhynchomonas*, unclassified neobodonids); marine and freshwater or soil-dwelling (NEO2 and NEO3); predominantly freshwater or soil-dwelling (*Neobodo* sp. and NEO1).

Trypanosomatida is the second largest kinetoplastid subgroup in our dataset, consisting of a clade of 942 sequences and a single sequence (*Paratrypanosoma confusum*) branching as a sister clade to Eu-, Para-, and Neobodonida. That topology is apparently incorrect since *Paratrypanosoma* is a typical trypanosomatid judging by its molecular and morphological features, and branches as the most basal trypanosomatid according to an extensive phylogenetic analysis of a multi-protein dataset (Flegontov et al. 2013). Trypanosomatids are obligatory endoparasites of terrestrial insects (mostly hemipterans and dipterans), invertebrates (mostly leeches) and vertebrates (Lukeš et al. 2014, Maslov et al. 2013). All sequences in the database were annotated as trypanosomatids, a majority of them was

annotated to the genus and species level, and just 5 sequences came from environmental surveys. Our tree distinguishes most currently recognized genera or provisionally named groups, likely with generic status (Lukeš et al. 2014, Maslov et al. 2013), but higher order clades usually lack support. The well supported clades are listed below:

i/ *Angomonas* (100% BS; 16 sequences annotated as *Angomonas*, 5 sequences mis-annotated as *Herpetomonas* or *Strigomonas* were re-annotated as *Angomonas*). Forms a joint clade with *Strigomonas*. Monoxenous parasites of insects;

ii/ *Blastocrithidia* (98% BS; 5 sequences annotated as *Blastocrithidia*, one sequence mis-annotated as *Leptomonas* and 32 poorly annotated ones were re-annotated as *Blastocrithidia*). Members of this clade generally grow poorly in axenic cultures and were mostly recovered from environmental sequences of insect gut (Lukeš et al. 2014, Maslov et al. 2013). Monoxenous parasites of insects;

iii/ *Blechomonas* (99% BS, 17 sequences annotated as *Blechomonas*). Monoxenous parasites of fleas (Votypka et al. 2013);

iv/ *Herpetomonas* (56% BS; 35 sequences annotated as *Herpetomonas*, 4 sequences mis-annotated as *Leptomonas* and 4 poorly annotated ones were re-annotated as *Herpetomonas*). Monoxenous parasites of insects;

v/ Leishmaniinae (91% BS; 125 sequences annotated as *Crithidia, Endotrypanum, Leishmania, Leptomonas*, and *Lotmaria*; 4 sequences mis-annotated as *Angomonas, Blastocrithidia* or *Wallaceina* and 9 poorly annotated ones were re-annotated as Leishmaniinae). This recently proposed subfamily of trypanosomatids (Jirků et al. 2012) unites monoxenous parasites of insects and dixenous parasites of cold- and warm-blooded vertebrates (genus *Leishmania*), some of them pathogenic for humans. Due to extremely low divergence among their 18S rRNA genes, genera and species within Leishmaniinae cannot be distinguished based on this gene;

vi/ *Phytomonas* (17% BS; 7 sequences annotated as *Phytomonas*, two poorly annotated sequences were re-annotated as *Phytomonas*). This clade was usually recovered with good support (REF), but in our tree the clade is unsupported due to the long-branch attraction phenomenon, caused by a few *Trypanosoma* sequences incorrectly branching within *Phytomonas*. Dixenous parasites of plants and insects;

vii/ *Sergeia* (99% BS; one sequence annotated as *Sergeia podlipaevi*, one poorly annotated sequence was re-annotated as *Sergeia*). Monoxenous parasites of insects;

viii/ *Strigomonas* (100% BS; 6 sequences annotated as *Strigomonas*). Forms a well-supported clade with *Angomonas*. Monoxenous parasites of insects;

ix/ *Trypanosoma* (46% BS; 654 sequences annotated as various species of this genus). Due to a long-branch attraction artefact, a small clade of trypanosomes branches within *Phytomonas* in our 18S rRNA-based tree, which explains the low support of the main *Trypanosoma* clade. Dixenous parasites circulating between vertebrates and insects or leeches;

x/ *Wallacemonas* (100% BS; 4 sequences having an outdated annotation *Wallaceina*, one sequence mis-annotated as *Leptomonas*, and one poorly annotated sequence was re-annotated as *Wallacemonas* - a new genus name proposed by Kostygov et al. (2014)). Monoxenous parasites of insects;

xii/ TRY1 (100% BS; 3 poorly annotated sequences were re-annotated as TRY1). Monoxenous parasites of insects;

xiii/ TRY2 (95% BS; 3 poorly annotated sequences were re-annotated as TRY2). Monoxenous parasites of insects.

In summary, we kept the original taxonomic annotations for 92% sequences, provided annotations for 6% poorly annotated sequences, and corrected annotations for 2% of them.

Diplonemea

A total of 433 sequences robustly clustered (100% bootstrap support) within diplonemids according our 18S rRNA tree (Figure 2). A majority was mis-annotated as uncultured eukaryotes, uncultured marine eukaryotes, uncultured euglenozoans or uncultured euglenids, while 153 sequences were correctly deposited as diplonemids. The class Diplonemea can be further sub-divided into four well-supported clades: i/ a clade containing two described genera *Diplonema* and *Rhynchopus* (100% BS, 42 sequences); ii/ a *Hemistasia* clade (100% BS, 7 seq.) iii/ the largest clade containing only environmental sequences without any described representative (76% BS, 375seqs).; and finally iv/ a small environmental clade comprising seven sequences (100% BS). The mutual branching order of these four clades cannot be resolved on the basis of the 18S rDNA gene and the current sampling. We suggest assigning the following taxonomic names for these four clades: i/ Diplonemidae for the clade containing sequences of *Diplonema* and *Rhynchopus*; ii/

Hemistasiidae for the clade containing the sequence of *H. phaeocysticola* (Yabuki and Tame 2015); iii/ DSPD I for the large environmental clade, and iv/ DSPD II for the small environmental clade. The names DSPD I and DSPD II stand for deep-sea pelagic diplonemids and were assigned in the first description of these clades (Lara et al. 2009) and then used in recent studies of marine diplonemid diversity (Flegontova et al. 2016; Gawryluk et al. 2016).

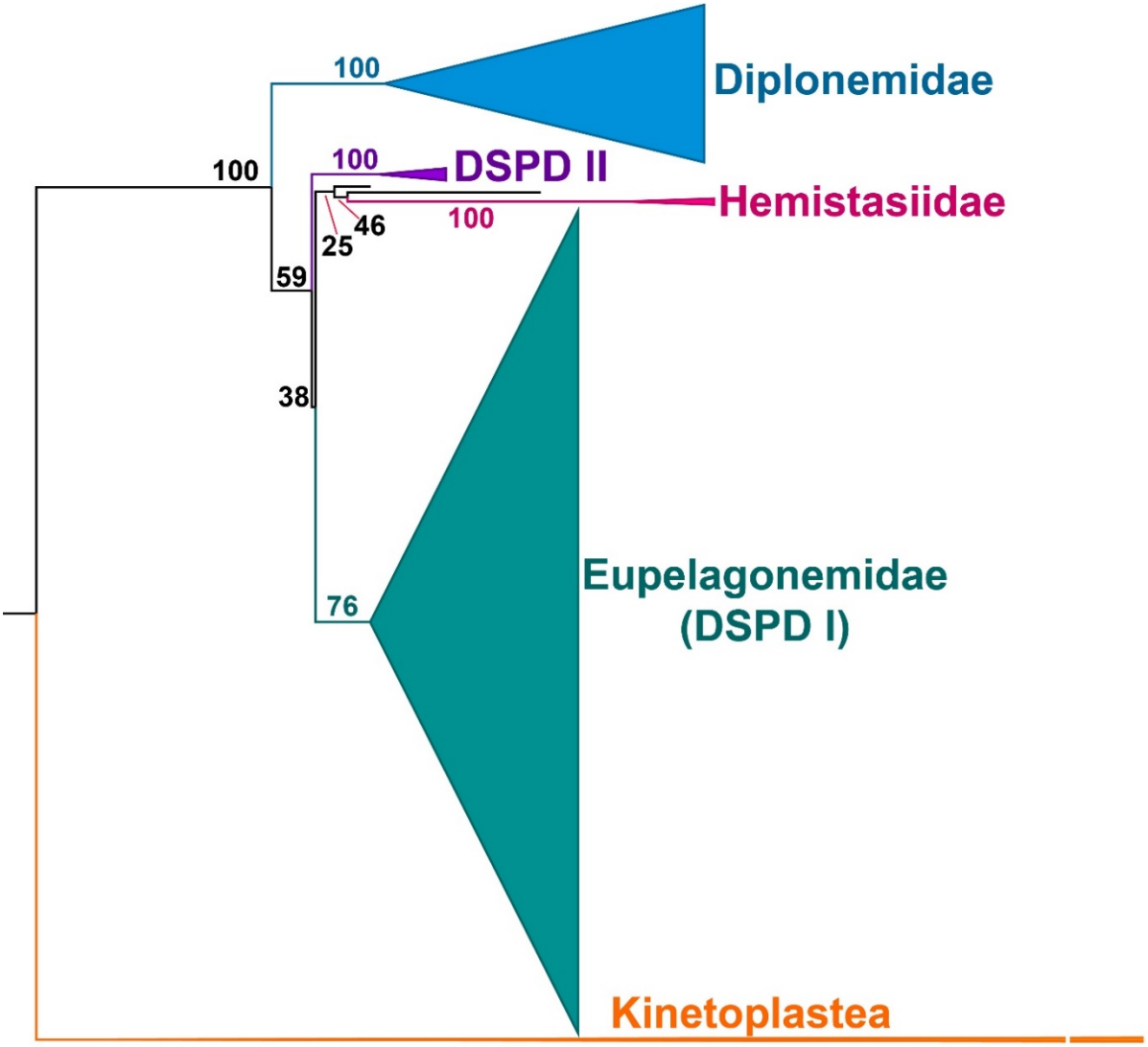


Figure 2. A maximum likelihood tree of 433 diplonemid sequences extracted form GenBank constructed using RAxML (model GTR+CAT+I+25 rate categories, 1000 rapid bootstrap replicates). For reducing the tree size, only seed sequences representing clusters with the 97% identity threshold were included. The kinetoplastid outgroup clade is collapsed. The major diplonemid clades are highlighted in color and labelled.

In the Diplonemidae group we observe one clade corresponding to genus *Rhynchopus* (96% BS, 7 sequences) and two separate clades of genus *Diplonema*: i/ 79% BS, 3 sequences (two annotated as *D. ambulator* and one annotated as *Diplonema* sp.); ii/ 35% BS, 4

sequences (two annotated as *D. papillatum*, one annotated as *Diplonema* sp., one mis-annotated as an uncultured euglenid). Six *Diplonema* sequences originated from cultures and one (mis-annotated as an uncultured euglenid) was derived from an oxygen-depleted marine habitat. All *Diplonema* cultures were isolated from marine water, some of them were free-living, while *D. papillatum* was isolated from the surface of common eelgrass *Zostera marina*. Five *Rhynchopus* sequences originated from cultures and two were environmental. All of them originated from marine biomes (plankton, hydrothermal plume, microbial mat), some of them were free-living, and two were blood parasites of a crustacean *Nephrops norvegicus*. The remaining 28 sequences of the Diplonemidae clade do not form reliable sub-clades, and all of them were derived from marine biomes (mostly planktonic, also benthic and from a hydrothermal plume).

All seven sequences of the Hemistasiidae group were found in marine plankton, and two of them originated from oxygen-depleted marine water. Only one sequence was annotated as *Hemistasia phaeocysticola*, and six sequences were poorly annotated as uncultured eukaryotes.

All 375 sequences from the DSPD I clade need a more detailed annotation inasmuch as 64% of them were annotated as eukaryotes without any lower-lever ranks, 34% were annotated up to the diplonemid level, and 2% up to the euglenozoan level. All sequences from the DSPD I clade were found in marine biomes, 99.5% belonged to plankton, and only few sequences were associated with hydrothermal plumes, oxygen-depleted marine water, marine sediments, and one was isolated from a bivalve *Bathymodiolus thermophilus*. The real diversity of this clade, ubiquitous in deep-sea environments worldwide, is most probably two orders of magnitude higher than reported in the current database (Flegontova et al. 2016). In fact, it emerged as the most diverse clade of planktonic eukaryotes, according to a meta-barcoding study based on the V9 region of 18S rDNA (Flegontova et al. 2016).

All seven sequences from the DSPD II clade were found in marine plankton as well. Six of them were annotated as uncultured marine diplonemids, and one as an uncultured eukaryote. Two environmental sequences formed deep branches of uncertain affiliation within the Diplonemea clade and could be annotated up to the Diplonemea level (originally they had been annotated as uncultured eukaryotes). Both of them originated from marine plankton, and one of them from an oxygen-depleted habitat.

We suggest dividing the Diplonemea clade into four taxonomic groups: Diplonemidae, Hemistasiidae, DSPD I and DSPD II. Here we provide a more detailed annotation for 97%

sequences, annotation of just 12 out of 433 sequences remained unchanged, and one wrong annotation has been corrected. 96.5% diplonemid sequences were environmental and only 3.5% were derived from cultured organisms. All 433 sequences originated from the marine or brakish biomes, with only 6 sequences derived from sediments, while an overwhelming majority came from the plankton. Information about any parasitic or symbiotic association with other organisms was not present in the database, with the exception of two cultures of *Rhynchopus* isolated from the blood of *N. norvegicus*, a culture of *D. papillatum* isolated from the surface of *Z. marina*, and one environmental sequence of the DSPD I clade that was isolated from *B. thermophilus*.

6 Curriculum vitae

Olga Flegontova

Date of birth:	13 November 1981
Nationality:	Russian
Home address:	Česká 16, České Budějovice 370 01, Czech Republic
Telephone number:	+ 420 604 697 819
E-mail:	olga@paru.cas.cz, oflegontova@mail.ru
Work address:	Biology Centre of the Czech Academy of Sciences, Institute of Parasitology & University of South Bohemia, Faculty of Science Branišovská 31, České Budějovice 370 05, Czech Republic

EDUCATION

2013 – present	**Ph.D. student** of Molecular and Cell Biology and Genetics Department of Molecular Biology and Genetics, Faculty of Science, **University of South Bohemia**, České Budějovice, Czech Republic Institute of Parasitology, **Czech Academy of Sciences**, Czech Republic Thesis: *Diversity and biogeography of diplonemid and kinetoplastid protists in global marine plankton* Supervisor: Aleš Horák
2000 – 2005	**B.S. and M.S., Physiology** Department of Neuroscience, Faculty of Biology, **Lomonosov Moscow State University**, Moscow, Russian Federation Thesis: *Effect of central interleukin 4 injection on neurodegeneration caused by β-amyloid peptide in rats* Supervisor: Natalia Gulyaeva

PROFESSIONAL EXPERIENCE

Employment:

2013 – present	**Graduate student**, Institute of Parasitology, **Biology Centre of the Czech Academy of Sciences**
2009 – 2013	**Researcher**, Institute of Parasitology, **Biology Centre of the Czech Academy of Sciences**
2005 – 2009	**Researcher**, Department of Molecular Basis of Human Genetics, **Institute of Molecular Genetics**, Moscow, Russian Federation

Research stays:

2014	**Visiting Student** (4-19/06) Oceanic Plankton Group, **Station Biologique de Roscoff**, Roscoff, France

---

PEER-REVIEWED PUBLICATIONS

Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, Bowler C, Lukeš J, Horák A (2016) Extreme diversity of diplonemid eukaryotes in the ocean. ***Curr Biol*** 26:3060-3065 (IF = 8.851).

Flegontov P, Butenko A, Firsov S, Kraeva N, Eliáš M, Field MC, Filatov D, **Flegontova O**, Gerasimov ES, Hlaváčová J, Ishemgulova A, Jackson AP, Kelly S, Kostygov AY, Logacheva MD, Maslov DA, Opperdoes FR, O'Reilly A, Sádlová J, Ševčíková T, Venkatesh D, Vlček Č, Volf P, Votýpka J, Záhonová K, Yurchenko V, Lukeš J (2016) Genome of *Leptomonas pyrrhocoris*: a high-quality reference for monoxenous trypanosomatids and new insights into evolution of *Leishmania*. ***Sci Rep*** 6:23704 (IF = 5.228).

Flegontov P, Changmai P, Zidkova A, Logacheva MD, Altınışık NE, **Flegontova O**, Gelfand MS, Gerasimov ES, Khrameeva EE, Konovalova OP, Neretina T, Nikolsky YV, Starostin G, Stepanova VV, Travinsky IV, Tříska M, Tříska P, Tatarinova TV (2016) Genomic study of the Ket: a Paleo-Eskimo-related ethnic group with significant ancient North Eurasian ancestry. ***Sci Rep*** 6:20768 (IF = 5.228).

Lukeš J, **Flegontova O**, Horák A (2015) Diplonemids. ***Curr Biol*** 25:R702-R704 (IF = 8.851).

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury J-M, Bittner L, Chaffron S, Dunthorn M, Engelen S, **Flegontova O**, Guidi L, Horák A, Jaillon O, Lima Mendez G, Lukeš J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S, *Tara* Oceans Coordinators, Acinas SG, Bork P,

Bowler C, Gaill F, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki M, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E (2015) Eukaryotic plankton diversity in the sunlit ocean. ***Science*** 348:1261605 (IF = 37.205).

Flegontova O, Lukeš J, Flegontov P (2012) Lack of evidence for integration of *Trypanosoma cruzi* minicircle DNA in South American human genomes. ***Int J Parasitol*** 42:437-441 (IF = 4.242).

Limborska SA, Khrunin AV, **Flegontova OV**, Tasitz VA, Verbenko DA (2011) Specificity of genetic diversity in D1S80 revealed by SNP-VNTR haplotyping. ***Ann Hum Biol*** 38:564-569 (IF = 1.148).

Flegontova OV, Khrunin AV, Lylova OI, Tarskaia LA, Spitsyn VA, Mikulich AI, Limborska SA (2009) Haplotype frequencies at the DRD2 locus in populations of the East European Plain. ***BMC Genet*** 10:62 (IF = 2.266).

Stepanichev MYu, **Flegontova OV**, Lazareva NA, Egorova LK, Gulyaeva NV (2006) [Influence of anti-inflammatory cytokine interleukin 4 on neurodegeneration in rats caused by beta-amyloid peptide.] ***Neyrochimiya*** 23:71-76 in Russian.

PRESENTATIONS AT CONFERENCES

2017	15th International Congress of Protistology, Prague, Czech Republic (30/07-04/08). Poster presentation.
	47th International Meeting of the Czech Society for Protozoology, Nové Hrady, Czech Republic (24-28/04). Oral presentation.
2016	ISEP/ISOP meeting Protist-2016, Moscow, Russia (6-10/06). Poster presentation.
	EMBO | EMBL Symposium: A New Age of Discovery for Marine Microeukaryotes, Heidelberg, Germany (26-29/01). Poster presentation.
2015	7th European Congress of Protistology, Seville, Spain (5-10/09). Oral presentation.
2014	ISEP/ISOP meeting Protist-2014, Banff, Canada (3-8/08). Oral and poster presentations.

HONORS, AWARDS, AND FUNDING

2014-2015	Grant Agency of the University of South Bohemia (reg. no. 04−088/2014/P, PI: Olga Flegontova). *Genomics of uncultured marine*

diplonemids from worldwide planktonic samples.

ATTENDED WORKSHOPS

2015	EukRef workshop, University of British Columbia, Vancouver, Canada (19-24/07).

RESEARCH CRUISES

2016	AREX2016, s/y Oceania, Institute of Oceanology, Polish Academy of Sciences (21/06-23/07).

SKILLS

Field work: collection of marine planktonic samples.

Laboratory work: more than ten years of experience with basic and advanced molecular biology methods and cultivation of protists.

Bioinformatics: good knowledge of Linux and data analysis approaches for phylogenomics, genomics, and environmental metabarcoding.

Programming: basic experience in Bash, Perl and R programming.

Languages: English (*TOEFL ITP* B1 level), Russian (native).

TEACHING AND STUDENTS

2016-2017	Teaching assistant at the Introduction to Bioinformatics and Introduction to Genomics courses, Faculty of Science, University of South Bohemia, České Budějovice