Report

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

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,

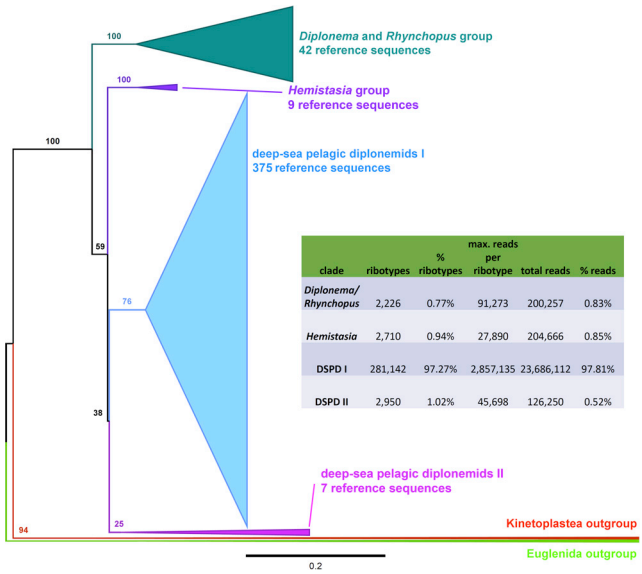


Figure 1. A Maximum-Likelihood Tree Based on 433 Diplonemid 18S rRNA Sequences Longer than 500 bp and Kinetoplastid and Euglenid Outgroups

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

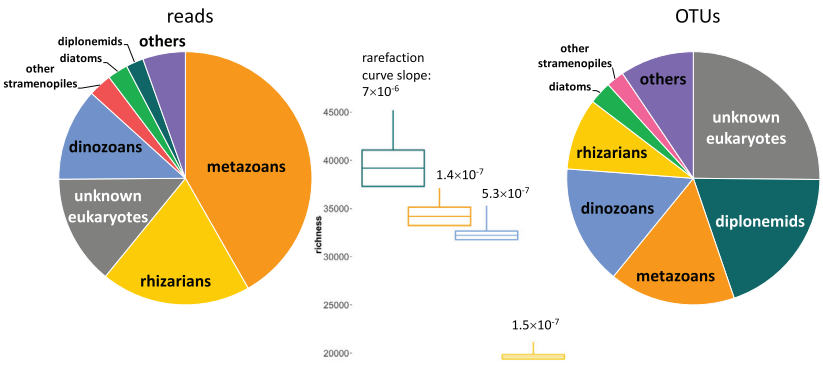


Figure 2. Fractions of Richness and Abundance Corresponding to Six Clades of Planktonic Eukaryotes, which Are Most Diverse in the Extended *Tara* Oceans Dataset

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

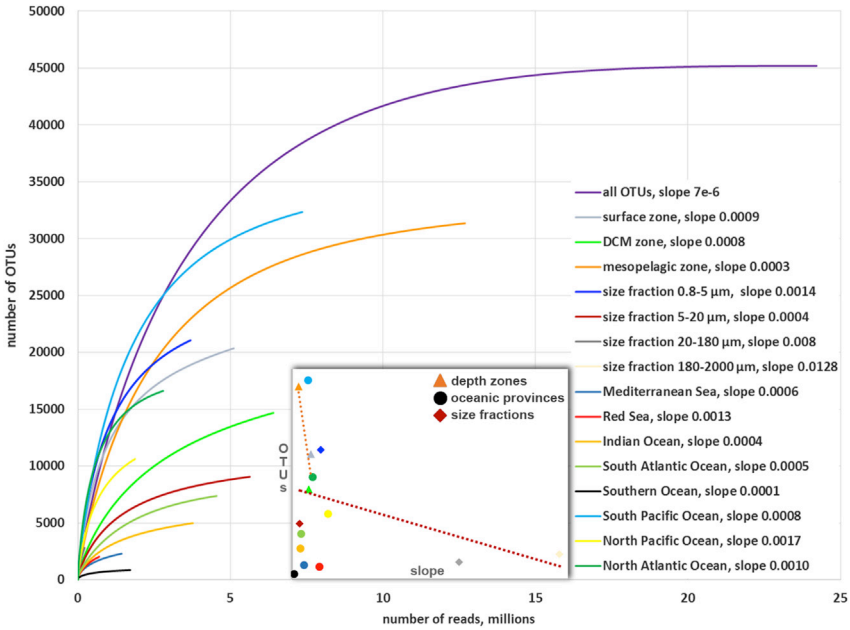


Figure 3. Rarefaction Curves for OTUs

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

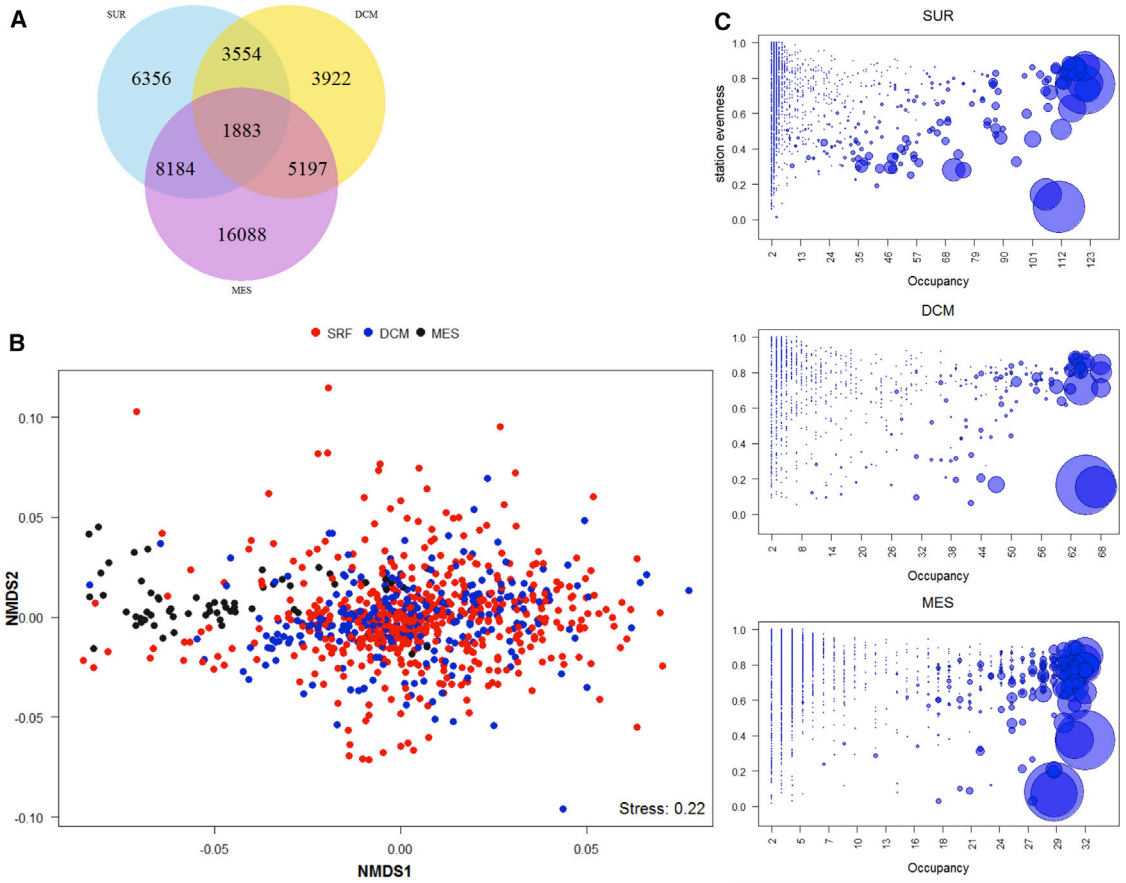


Figure 4. Depth Stratification of Diplonemid Diversity

(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.

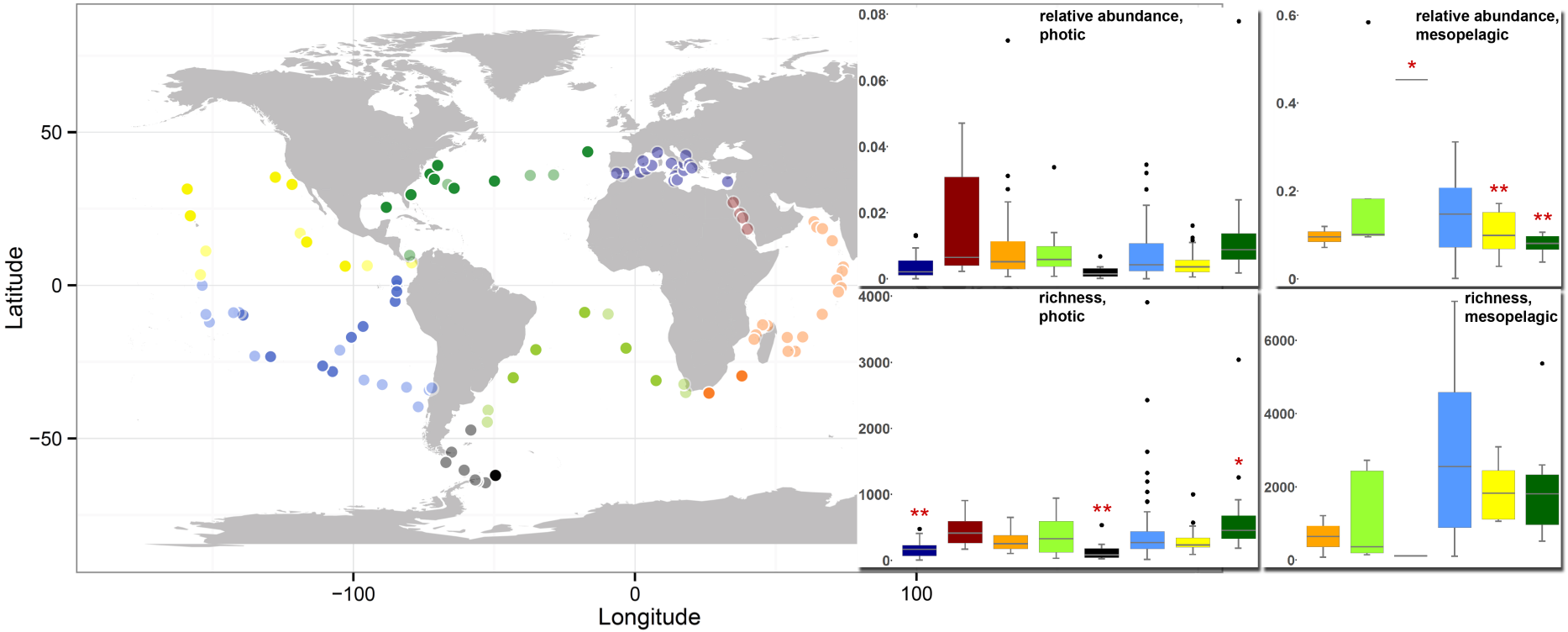
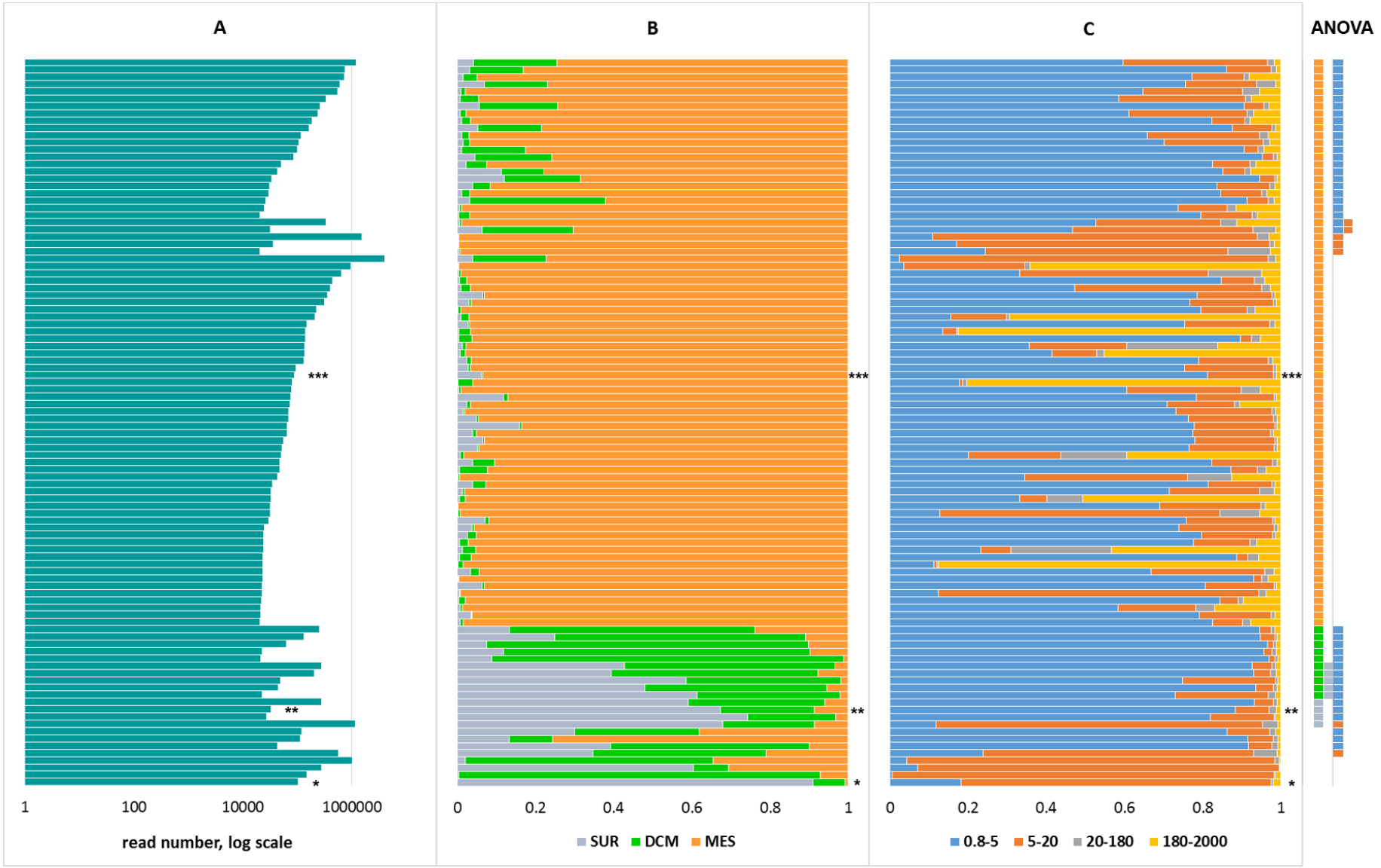


Figure S2.


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

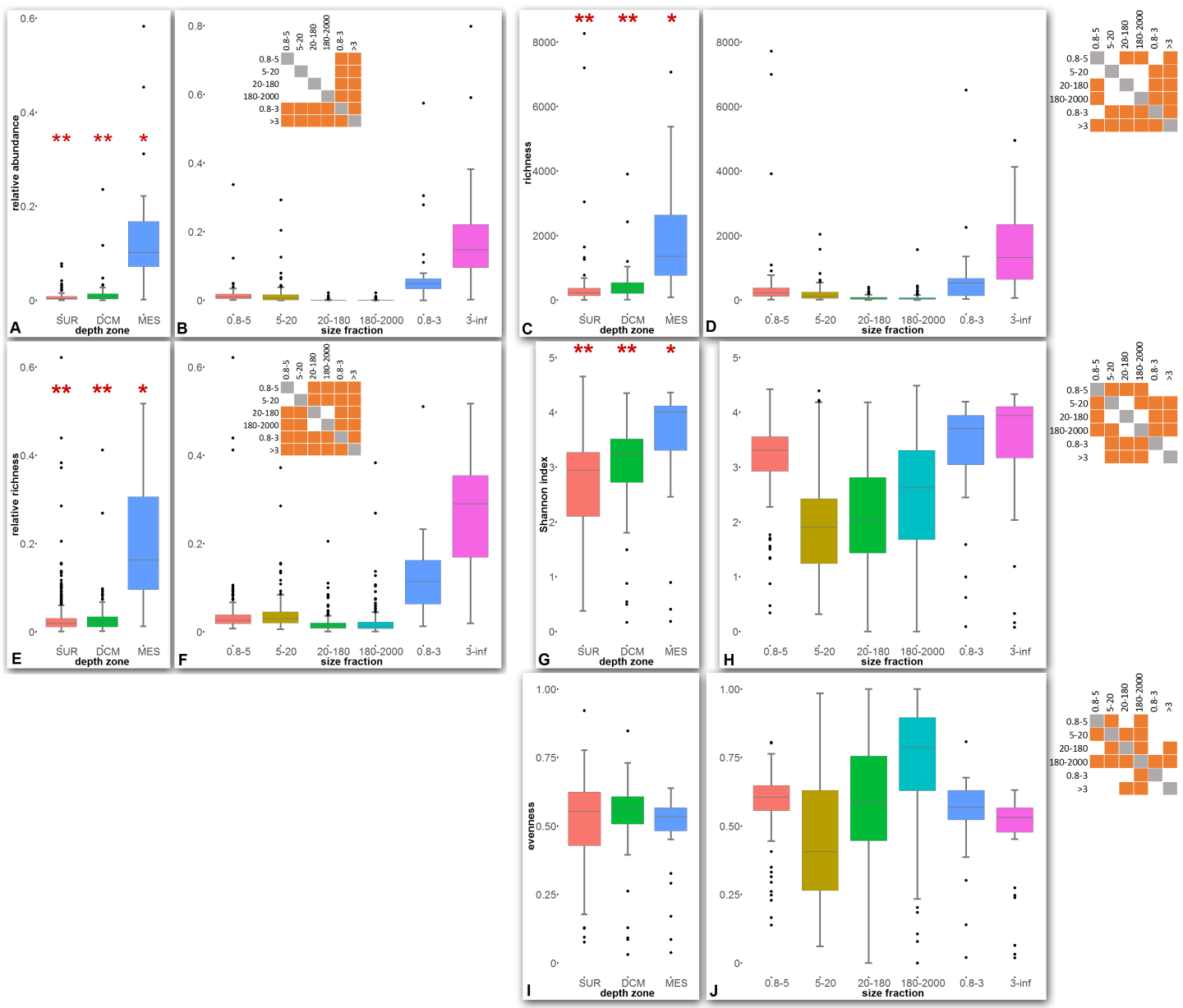
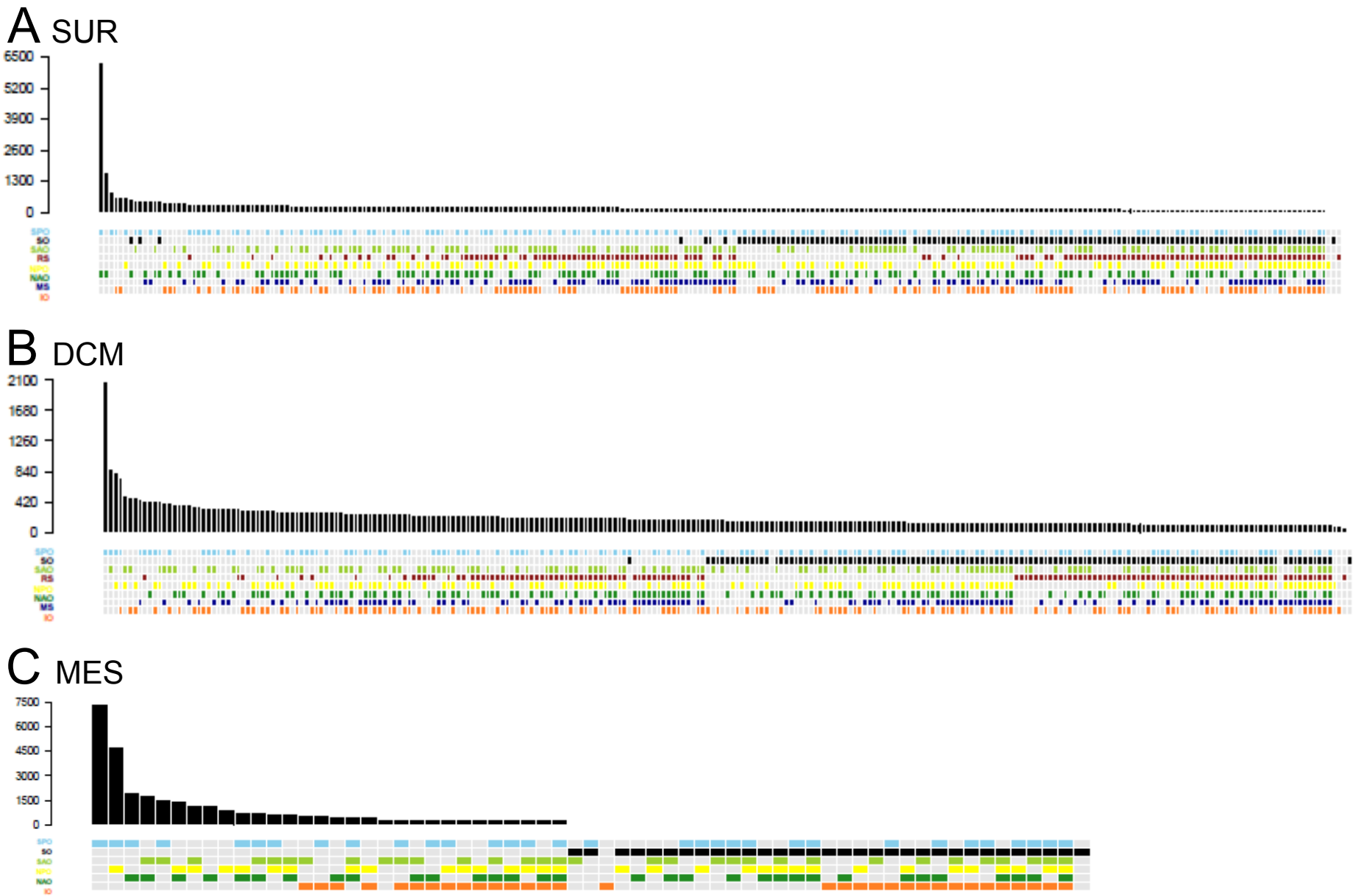

relative abundance
relative richness
richness
Shannon index
evenness
depth zone
size fraction
SUR
DCM
MES
0.8-5
5-20
20-180
180-2000
0.8-3
>3
A
B
C
D
E
F
G
H
I
J

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.

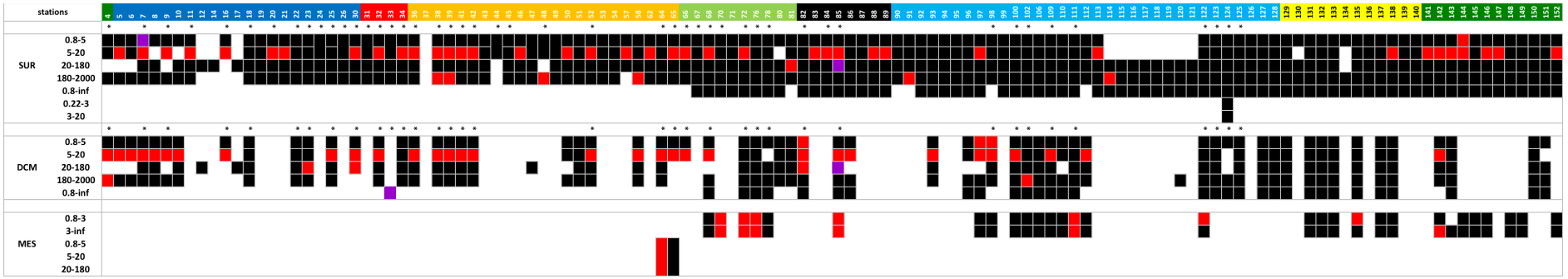
Figure S3. Related to Figures 3 and 4A,B. Box plots illustrating relative abundance of diplonemid ribotypes (**A, B,** calculated as diplonemid read count divided by eukaryotic read count), richness (**C, D,** OTU count), relative richness (**E, F,** diplonemid OTU count / eukaryotic OTU count), Shannon 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.