MR. ALEIX OBIOL (Orcid ID : 0000-0002-5475-9827)
DR. CATERINA R GINER (Orcid ID : 0000-0002-7267-0260)

Article type : Resource Article

A metagenomic assessment of microbial eukaryotic diversity in the global ocean

Aleix Obiol[1]*, Caterina R. Giner[1,2], Pablo Sánchez[1], Carlos M. Duarte[3], Silvia G. Acinas[1], Ramon Massana[1]*

[1] *Department of Marine Biology and Oceanography, Institut de Ciències del Mar (ICM-CSIC), Barcelona, Catalonia, Spain*

[2] *Current address: Institute for the Oceans and Fisheries, University of British Columbia, Vancouver, Canada*

[3] *King Abdullah University of Science and Technology (KAUST), Red Sea Research Center (RSRC), Thuwal, Saudi Arabia*

***Corresponding authors:**
Aleix Obiol (obiol@icm.csic.es) and Ramon Massana (ramonm@icm.csic.es)
Institut de Ciències del Mar (ICM-CSIC)
Passeig Marítim de la Barceloneta 37-49
08003 Barcelona, Catalonia, Spain

Running title: Diversity of marine protists by metagenomics

Abstract

Surveying microbial diversity and function is accomplished by combining complementary molecular tools. Among them, metagenomics is a PCR free approach that contains all genetic information from microbial assemblages and is today performed at a relatively large scale and reasonable cost, mostly based on very short reads. Here we investigated the potential of metagenomics to provide taxonomic reports of marine microbial eukaryotes. We prepared a curated database with reference sequences of the V4 region of 18S rDNA clustered at 97% similarity and used this database to extract and classify metagenomic reads. More than half of them were unambiguously affiliated to a unique reference whilst the rest could be assigned to a given taxonomic group. The overall diversity reported by metagenomics was similar to that obtained by amplicon sequencing of the V4 and V9 regions of the 18S rRNA gene, although either one or both of these amplicon surveys performed poorly for groups like Excavata, Amoebozoa, Fungi and Haptophyta. We then studied the diversity of picoeukaryotes and nanoeukaryotes using 91 metagenomes from surface down to bathypelagic layers in different oceans, unveiling a clear taxonomic separation between size fractions and depth layers. Finally, we retrieved long rDNA sequences from assembled metagenomes that improved phylogenetic reconstructions of particular groups. Overall, this study shows metagenomics as an excellent resource for taxonomic exploration of marine microbial eukaryotes.

***Keywords*:** metagenomics, marine protists, diversity, global ocean, amplicon sequencing

Introduction

Marine microbial eukaryotes are key components of planktonic ecosystems in all ocean biomes (Caron, Countway, Jones, Kim, & Schnetzer, 2012). They are, along with cyanobacteria, responsible for nearly half of the global primary production (Falkowski, 2012), and play important roles in food-web dynamics as grazers and parasites (Edgcomb, 2016; Jürgens & Massana, 2008), carbon export to the deep ocean (Guidi et al., 2016), and nutrient remineralization (Worden et al., 2015). A number of studies in the early 2000s, based on 18S ribosomal DNA (rDNA) amplicon data, hinted at their huge diversity and relevant novelty in different oceanic regions (Diez, Pedrós-Alió, & Massana, 2001; Edgcomb, Kysela, Teske, de Vera Gomez, & Sogin, 2002; López-García, Rodríguez-Valera, Pedrós-Alió, & Moreira, 2001; Moon-Van Der Staay, De Wachter, & Vaulot, 2001), recently confirmed by large-scale surveys (de Vargas et al., 2015; Pernice et al., 2016). However, rDNA amplicon data are dependent on PCR, which is known to introduce biases in microbial diversity estimates (Acinas, Sarma-Rupavtarm, Klepac-Ceraj, & Polz, 2005; Balzano, Abs, & Leterme, 2015; Sinclair, Osman, Bertilsson, & Eiler, 2015), potentially affecting both the number and relative abundance of the species and taxonomic groups present. In addition, the short reads of high-throughput sequencing (HTS) surveys require the choice of a given 18S rDNA region to amplify, with the hypervariable regions V9 (Amaral-Zettler, McCliment, Ducklow, & Huse, 2009) and V4 (Stoeck et al., 2010) being most used, which in some cases yield different results (Giner et al., 2016; Stoeck et al., 2010).

An alternative to amplicon-based HTS approaches (metabarcoding) to studying microbial diversity involves exploiting the taxonomic information contained in metagenomes. These use massive shotgun sequencing of genomic DNA extracted from microbial assemblages with the goal of assessing their functional metabolic potential. Given the usefulness and general application of the 18S rDNA, it follows that the identification of 18S rDNA sequences within the metagenomes provides a path to resolve microeukaryotic diversity free of the potential biases of PCR-dependent methods. Indeed, this technique was already used with shotgun Sanger sequencing data derived from the Global Ocean Survey, GOS (Not, del Campo, Balagué, de Vargas, & Massana, 2009; Piganeau, Desdevises, Derelle, & Moreau, 2008). In these studies, however, the modest sequencing depth attainable at the time allowed the retrieval of a low signal, with only 116 18S rDNA fragments found in the complete GOS dataset. The development of high-throughput sequencing platforms and the reduction of sequencing costs (Goodwin, McPherson, & McCombie, 2016) have allowed a drastic increase in sequencing depth, thus granting the retrieval of a significant number of short 18S rDNA metagenomic reads from a given sample

(metagenomic Illumina Tags or miTags; Logares *et al.* 2014). The term miTags coined by Logares et al. (2014) has been shortened hereafter as mTags to make it independent of the sequencing technology used. Several tools have been developed for the extraction of these reads (Bengtsson et al., 2011; Gruber-Vodicka, Seah, & Pruesse, 2019; Hartmann, Howes, Abarenkov, Mohn, & Nilsson, 2010; Huang, Gilna, & Li, 2009), generally based on 16S/18S rDNA Hidden Markov Model (HMM) profiles.

Although some studies have used HMM profiles to assess eukaryotic diversity in different environments (Bahram et al., 2018; Guajardo-Leiva, Pedrós-Alió, Salgado, Pinto, & Díez, 2018; Pernice et al., 2016; Saghaï et al., 2015), retrieving a precise taxonomical classification of the short metagenomic reads (100-250 bp) remains challenging (Breitwieser, Lu, & Salzberg, 2017), especially when targeting the 18S rDNA gene that contains a mosaic of highly conserved and highly variable regions (Neefs, Van de Peer, De Rijk, Chapelle, & De Wachter, 1993). Some bioinformatic tools have tried to address this concern by keeping the highest unambiguous level in hierarchical taxonomic classifications (Bengtsson-Palme et al., 2015; Guo, Cole, Zhang, Brown, & Tiedje, 2016), but these are still highly dependent on good reference databases for a correct taxonomic assignment (Pedrós-Alió, Acinas, Logares, & Massana, 2018).

Here we attempted to expand the taxonomy assessment potential of metagenomic reads by extracting and classifying them using the *eukaryotesV4* database, a custom database of eukaryotic V4 18S rDNA sequences built for this study. We first assessed the resolution level that this method can provide using 91 marine metagenomes collected during the Malaspina 2010 Circumglobal Expedition (Duarte, 2015). We then compared the obtained results with the more common amplicon sequencing of the V4 region (using the data from Giner et al., 2020; **Figure S1**), and the V9 region (newly obtained here). The mentioned paper from Giner et al. (2020) focused on vertical changes of picoeukaryotic (0.2-3 µm) diversity in the global ocean assessed by V4-metabarcoding and here we complement this study using V9-metabarcoding, metagenomic data, and add the still unexplored nanoeukaryotic fraction (3-20 µm) of the Malaspina dataset. Finally, we increased the taxonomic information of microbial eukaryotes by retrieving long sequences of the ribosomal DNA operon from assembled metagenomes. Overall, our study reveals that the analysis of metagenomes using a well curated rDNA database yields very good reports of the taxonomic groups present in marine assemblages, together with broadly comparable results with metabarcoding.

Materials and Methods

Building a custom 18S rDNA-V4 region database

A custom database of the V4 region of 18S rDNA (**Table S1**), *eukaryotesV4*, was created to retrieve and taxonomically classify metagenomic reads (mTags from now on). The database was first built using 97% clustered V4 sequences from previous environmental high-throughput sequencing (HTS) studies in European coastal systems: Blanes Bay Microbial Observatory (Giner et al., 2019) and BioMarKs project (Massana et al., 2015), and the water column of the global ocean sampled during the Malaspina cruise (Giner et al., 2020). Then, the database was complemented with trimmed 97% clustered V4 sequences from SILVA SSU 128 (Quast et al., 2013) that were not found in environmental HTS datasets. This trimming was performed using cutadapt v1.16 (Martin, 2011) with the universal eukaryotic forward primer from Stoeck et al. (2010) and reverse primer from Balzano et al. (2015), with an error rate of 0.2. All sequences were manually curated to discard possible chimeras and were classified at three taxonomic levels based on previous data, exhaustive inspection in multiple phylogenetic trees and iterative testing with environmental datasets to detect and correct problematic cases (i.e. distant references sequences retrieving the same mTag). These three levels were: (i) OTU_{97} level (Operational Taxonomic Units of sequences clustered at 97% similarity), (ii) taxonomic group (in general a formal Class), and (iii) supergroup. The largest effort was the classification at the taxonomic group level, which comprised 136 groups (**Table S1**). The final *eukaryotesV4* database contains 25,849 sequences, 43% of which derive from environmental datasets.

Sampling, DNA extraction and sequencing

As part of the Malaspina 2010 Circumnavigation Expedition (Duarte, 2015), we visited 10 stations distributed across the world's major oceans: 3 in the Atlantic Ocean, 3 in the Indian Ocean and 4 in the Pacific Ocean (**Figure S1; Table S2**). At each station, seawater samples from 7 depths (surface, deep chlorophyll maximum, and 2-3 depths at the mesopelagic and bathypelagic regions) were collected by means of Niskin bottles attached to a rosette coupled with a CTD profiler, which measured conductivity, temperature, fluorescence, salinity and dissolved oxygen along the water column. About 12 L of seawater were prefiltered through a 200 µm nylon mesh and then sequentially filtered with a peristaltic pump through a 20 µm nylon mesh followed by a 3 µm and a 0.2 µm pore-size 142 mm Millipore polycarbonate filters. Filters were immediately flash-frozen into liquid nitrogen and stored at -80ºC until processed in the lab. Samples for amplicon sequencing were similarly collected except that filtering was carried out using 47 mm diameter filters.

DNA extracts for metagenomics were obtained with the phenol-chloroform protocol (Massana, Murray, Preston, & DeLong, 1997). For the nanoeukaryotic fraction (3-20 µm) we obtained 25 samples from 4 stations, and for the picoeukaryotic fraction (0.2-3 µm) we obtained 66 samples from 10 stations (**Figure S1; Table S2**). Whole metagenome sequencing was performed using a PCR free protocol at CNAG (Barcelona, Spain; http://cnag.cat/). Short-insert paired-end libraries were prepared with the Illumina TruSeq Sample Preparation kit (Illumina Inc.) and sequenced using the HiSeq 2000 Illumina platform (2x101 bp) for all picoeukaryotic samples and 6 nanoeukaryotic samples. For the remaining 19 nanoeukaryotic samples, short-insert paired-end libraries were prepared with KAPA HyperPrep kit (Roche Kapa Biosystems) and sequenced with either NovaSeq 6000 or HiSeq 4000 Illumina platforms (2x151 bp). Sequencing yielded 30.4 ± 20.6 Gbp per sample (average ± standard deviation). The full functional exploitation of the metagenomes is currently in preparation.

DNA extracts for amplicon sequencing and data for metabarcoding of the V4 region derive from a recent study targeting the picoeukaryotic fraction (Giner et al., 2020). We used the same DNA extracts to amplify the V9 region of the 18S rDNA (~130 bp) using primers 1389F and 1510R (Amaral-Zettler et al., 2009). A touchdown PCR amplification protocol with 35 cycles was performed for both regions. Sequencing of amplicons was done with the MiSeq Illumina platform (2x250 bp) at RTLGenomics (Lubbock, TX, USA; http://rtlgenomics.com/). Amplicon data of the V4 and V9 regions was obtained along the vertical profile of 4 stations (**Figure S1; Table S2**).

Metagenomics data processing for mTags extraction and classification

Metagenomic raw reads were trimmed for TruSeq adapters and filtered for phred scores of ≥20 and length ≥45 base pairs either with trimmomatic v0.35 (Bolger, Lohse, & Usadel, 2014) for HiSeq runs, or with cutadapt v1.16 (Martin, 2011) for NovaSeq runs. The pipeline used to extract and assign taxonomy to V4 metagenomic reads (mTags) is shown in **Figure 1A**. Reads longer than 70 bp were mapped against a 92% clustered version of *eukaryotesV4* (10,188 reference sequences) using BLAST v2.7.1 (Altschul, Gish, Miller, Myers, & Lipman, 1990). Sequences with a hit to a reference sequence with >90% similarity and >70% query alignment were retrieved from the metagenomes with seqtk's v.1.3 *subseq* option (https://github.com/lh3/seqtk). As most metagenomes yielded 101 bp reads, mTags of 151 bp from some metagenomes were trimmed at the 3' end to 101 bp with seqkit's v0.10.1 *subseq* option (Shen, Le, Li, & Hu, 2016) to make results comparable. All mTags were then mapped against the *eukaryotesV4* database with the *usearch_local* command of USEARCH v9.2 (Edgar, 2010) with a 97% similarity threshold and options *-strand both, -mincols 70*, and *top_hits_only*, which yielded all top scoring hits for each

read. Based on this score list, mTags were classified as: (i) those with a single hit (OTU_{97} level), (ii) those with >1 hit to sequences of the same taxonomic group (Group level), (iii) those with >1 hit to sequences of different groups but same supergroup (Supergroup level) and (iv) those with hits to sequences from different supergroups (Ambiguous level). When both reads of an Illumina pair matched the database, the best assignment was considered and counted as one. After the classification, mTags from Charophyta, Metazoa, nucleomorphs and Ulvophyceae were removed (5.5% of total mTags). The final table contained 302,269 mTags from 66 picoeukaryotic samples and 25 nanoeukaryotic samples that were assigned to 4,723 OTU_{97} and 84 higher-rank levels.

mTags were also extracted from the metagenomes using Hidden Markov Models (HMM) as used in previous studies (Bengtsson-Palme et al., 2015; Guo et al., 2016; Logares et al., 2014), using the *hmmsearch* in HMMER v3.2 (hmmer.org) as implemented in Logares et al. (2014), with *e-value* 10 and a custom HMM profile prepared from an aligned version of *eukaryotesV4* database. The taxonomy assignment of the extracted reads was done as before.

Amplicon data processing

In both V4 and V9 amplicon datasets, raw reads were trimmed for amplification primers with cutadapt v1.16 (Martin, 2011) and processed using the software package DADA2 v1.4 (Callahan et al., 2016) with the parameters *truncLen 220,210* and *maxEE 6,8* for V4 samples and *truncLen 110,90* and *maxEE 4,6* for V9 samples. For each dataset, an amplicon sequence variant (ASV) table was obtained. Samples contained 51,421 reads on average (standard error 4,860) and ASVs present in only 1 sample with less than 10 reads were removed. Taxonomic assignment of V4 ASVs was performed using *blastn* command in BLAST v2.7.1 (Altschul et al., 1990) against *eukaryotesV4*. Taxonomy of V9 ASVs was assigned using BLAST against PR^2 (Guillou et al., 2013) and NCBI nt databases and was formatted to match *eukaryotesV4* taxonomic levels. ASVs classified as Archaea, Bacteria, Charophyta, Ulvophyceae, Metazoa, and nucleomorphs were removed (1.3% of reads in V4 and 47% in V9 amplicons, mostly coming from amplified prokaryotic taxa in the latter). The V4 final table contained 23 samples, 6,037 ASVs and 1,760,294 reads, while the V9 final table contained 23 samples, 3,310 ASVs and 605,053 reads.

Metagenomes assembly, 18S rDNA contigs retrieval and phylogenetic analyses

Metagenomic samples were assembled using MEGAHIT v1.1.3 (Li, Liu, Luo, Sadakane, & Lam, 2015) with *meta-large* preset and a minimum contig length of 500 bp. Assembled contigs containing the eukaryotic rDNA operon were retrieved using *blastn* command in BLAST against *eukaryotesV4* database. In order to identify the location of 18S and 28S genes in the contigs,

these were mapped against PR^2 (Guillou et al., 2013) and SILVA LSU 132 (Quast et al., 2013), respectively, using BLAST. A full list with all extracted contigs with more than 1000 bp of 18S rDNA gene is found in **Table S3**.

Phylogenetic trees were built using contigs containing >1500 bp of 18S rDNA and complete 18S versions of sequences from *eukaryotesV4* database extracted from SILVA SSU 128 (Quast et al., 2013). These were aligned using MAFFT v7.402 (Katoh & Standley, 2013) in *auto* mode. Alignments were trimmed with trimAl v1.4.rev22 (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009), regions not shared among sequences were removed with AliView v1.25 (Larsson, 2014) and a maximum likelihood (ML) tree was constructed with RAxML v8.2.12 (Stamatakis, 2014), using GTRCATI model, by selecting the best topology out of 1000 alternative trees. Bootstrap analysis was done with 1000 pseudo-replicates.

Statistical analyses

Most of the analyses of this study were conducted in R statistical environment v3.6.0 (R Core Team, 2019). Using *vegan* package v.2.5-5 (Oksanen et al., 2019), Bray-Curtis dissimilarities were computed on relative abundance OTU/ASV tables with function *vegdist()* and non-metric multidimensional scaling (NMDS) was performed with function *metaMDS()* on the dissimilarity matrixes obtained. These were also used to run PERMANOVA tests with *vegan*'s function *adonis2()*. General analyses were performed with the package *tidyverse* v1.2.1 (Wickham et al., 2019). All scripts used are located at GitHub (https://github.com/aleixop/Malaspina_Euk_mTags).

Results

mTags extraction and taxonomic classification

We prepared an exhaustive collection of V4-18S rDNA sequences from the complete eukaryotic domain. Sequences within the *eukaryotesV4* reference dataset have been clustered at 97% similarity (OTU_{97}), curated to remove any chimeric signal, and classified into main taxonomic groups (**Table S1**). Using this database, we evaluated the taxonomic assessment power of V4-containing fragments (mTags) retrieved from 91 marine metagenomes from different oceans and depths (**Figure S1**) following the described pipeline (**Figure 1A**). A total of 302,269 mTags averaging 99.4 bp in length were retrieved, of which 58.5% were assigned to a specific sequence in the database (OTU_{97} level) and 40.4% were classified at group level, while a very low proportion were classified at supergroup level (1.0%) or remained ambiguous (0.2%) (**Figure**

1B). Thus, the assignment to a given group was achieved in nearly 99% of the mTags. Classification precision was not homogeneous among the different supergroups (**Figure 1C**): Stramenopiles presented only 27.1% of mTags defined to the OTU_{97} level and 7.3% not defined to any given group, mainly due to a conserved V4 region within Ochrophyta, while on the other side, 88.6% of mTags from Amoebozoa were well defined.

Both read length and mTags extraction method influenced the final number of reads retrieved and taxonomically classified. We took advantage of the fact that a few samples were sequenced with a different Illumina technology that yielded longer reads (151 bp instead of 101 bp) to report that longer reads improve the resolution of the taxonomic classification, yielding a 10% increase in the number of OTU_{97}-defined reads (**Figure S2**). Additionally, extracting mTags following an HMM-based protocol instead of using BLAST resulted in an 8% decrease of the total number of mTags retrieved. HMM-based extraction runs substantially faster and is the commonly used protocol. Both extraction approaches were well correlated, with R^2 values being virtually 1 and slopes close to 1 in most taxonomic supergroups (**Figure S3**). Nevertheless, for some supergroups the slopes were somewhat lower (i.e. 0.76 in Excavata, 0.82 in Amoebozoa or 0.90 in Stramenopiles), indicating that up to 24% of the mTags of these groups were missed by the HMM extraction.

Comparison of mTags and amplicon sequencing

We compared the relative abundance of taxonomic groups derived from metagenomes with that obtained by V4 and V9 amplicon sequencing in a subset of 23 picoeukaryotic (0.2-3 µm) samples from 4 separate stations (**Figure S1; Table S2**). The most remarkable differences were found in Discosea, Diplonemea, Kinetoplastida, and Prymnesiophyceae (**Figure 2**). These groups were absent in the V4 dataset and, except for Prymnesiophyceae, had significantly lower relative abundances in the V9 dataset. MALV-II and MALV-I, groups with very high relative abundances in the mTags survey, were significantly overrepresented with V4 amplicons ($p < 0.05$), but not with V9 amplicons. Groups equally represented in the three surveys (i.e. did not have significant differences; $p > 0.05$) were Polycystinea, Pelagophyceae, Chrysophyceae, Dinoflagellata, Acantharia, Bicosoecida and Chloropicophyceae. Both amplicon approaches yielded lower relative abundances in the case of Dictyochophyceae, and V9 amplicons underrepresented MALV-III. Fungi (Ascomycota and Basidiomycota), MAST groups and RAD-B were significantly underrepresented by V4 amplicons.

Bray-Curtis dissimilarities between community structures based on the relative abundances of taxonomic groups in the three surveys were calculated and used for representing all samples in a

non-metric multidimensional scaling (NMDS) plot (**Figure S4**). The three community surveys from the same sample were always closely placed, and there was a clear separation between the photic and aphotic water layers. We performed a PERMANOVA test in order to interpret these patterns using the sequencing approach, depth layers and oceanic region as variables. Within these, the different sequencing approaches only explained 7% of the overall variance ($p < 0.001$), while depth layer explained 36% of it ($p < 0.001$).

Nano and picoeukaryotic diversity assessed by mTags

We used the V4-18S mTags retrieved from the full dataset of 91 metagenomes to assess pico- (0.2-3 µm) and nanoeukaryotic (3-20 µm) diversity in the global ocean and along the water column. Each taxonomic group displayed a distinct vertical distribution, which was consistent across oceanic regions (**Figure S5**). Within the picoeukaryotic fraction MALV-II and MALV-I dominated in both water layers, with a median relative abundance of 29% and 15% in the photic and 22% and 7% in the aphotic layer, respectively (**Figure 3; Table S4**). In the latter, Polycystinea was also present with high abundance (14%). Regarding the nano fraction, Dinoflagellata was highly abundant in photic layers (61%) and moderately abundant in aphotic ones (14%) and in the latter, Polycystinea (22%) and Diplonemea (19%) were also abundant (**Figure 3**). As expected, groups known to include picosized organisms such as Pelagophyceae, MALV-II and MAST-4 (in the photic layers) and Chrysophyceae (in the aphotic layers) were much more abundant in the smaller size fraction (**Figure 3**). On the other hand, groups including typically larger cell sizes like Diatomea or RAD-A (in the photic layers) were mostly found in the nanoeukaryotic fraction. Groups underepresented by amplicons were more present in aphotic layers; Kinetoplastida was primarily detected in the picoeukaryotic fraction and Diplonemea and Discosea in the nano fraction.

Taxonomic groups organized very well along pico- vs. nanoeukaryotic fractions and photic vs. aphotic layers, with most groups showing maximal relevance in one of the four resulting compartments (**Figure 4; Table S4)**, such as MAST-4 and MAST-7 in the pico-photic space and Diatomea in the nano-photic one. Conversely, a few groups covered the two size fractions of the same layer, such as Prymnesiophyceae, Choanomonada and MAST-3 in the photic layer or Polycystinea in the aphotic layer, and others were dispersed in the four categories, like MALV-I or Ciliophora (**Figure 4)**.

Clustering pico- and nanoeukaryotic samples by their Bray-Curtis dissimilarities using the V4-18S mTags identified at the OTU_{97} level (about 60% of total) revealed a clear separation between size

fractions and ocean layers in a NMDS plot (**Figure S6**). A PERMANOVA test using the variables size fraction, ocean layer, oceanic region and environmental parameters (temperature, salinity, dissolved oxygen and conductivity; **Table S2**) revealed ocean layer and size fraction as the main community structuring parameters, explaining 19% and 11% of the variance ($p < 0.001$), respectively, followed by differences in oceanic regions (7%, $p < 0.001$), indicating that communities within the same size fraction but from different geographic locations shared more similarities between them than with other size fraction communities in the same sampling station.

Phylogenetic analyses using long rDNA sequences from assembled contigs

Our metagenomics approach also allowed accessing long rDNA sequences for the most dominant groups. A total of 724 contigs containing >1000 bp of the 18S rDNA were obtained (**Table 1; Table S3**). Overall, 188 of these contigs encompassed a complete 18S, and 38 contigs seemed to have the complete rDNA operon (i.e. 18S and 28S genes). Looking at the identity of all retrieved 18S fragments against PR^2, nearly a third of them had a percentage identity lower than 95%, thus potentially expanding the taxonomic information of eukaryotic microbial diversity. Taxonomic groups most represented by contigs matched those most abundant in the mTags analysis (**Table 1**).

The diversity of one of these abundant groups, Diplonemea, which was largely overlooked by our V4 and V9 amplicon surveys, was further explored in a maximum likelihood phylogenetic tree with references from *eukaryotesV4* database (**Figure 5**). From a total of 20 contigs containing more than 1500 bp of 18S rDNA, 18 of them belonged to Eupelagonemidae and came from both pico- (0.2-3 µm) and nanoeukaryotic (3-20 µm) fractions. The other 2 contigs fell into DSPD II family and were retrieved from nano samples only. The mean percentage identity these contigs had against GenBank was 97.5%, and about a third of them appeared to be separated from reference sequences in the phylogenetic tree, confirming and expanding previous reports of a high phylogenetic diversity within the group. All reference sequences within Eupelagonemidae and DSPD II retrieved at least one mTag (**Figure 5**), highlighting the high diversity of Diplonemea in our oceanic samples.

Discussion

In this work we explored the taxonomic information contained in deeply-sequenced marine metagenomes to assess the global diversity of marine microbial eukaryotes, and compared it with

the results obtained by the commonly used 18S rDNA amplicon sequencing. One of the concerns of using Illumina-based metagenomic fragments (mTags) to assess the diversity of microbial communities is their short length (101 bp here), which potentially limits the taxonomic detail they provide. Previous research on prokaryotic 16S rDNA reported that fragments as small as 100 bp suffice for community analysis (Liu, Lozupone, Hamady, Bushman, & Knight, 2007; Logares et al., 2014), while our results reveal that these short fragments provide a highly accurate description of the taxonomic diversity of microbial eukaryotes at the group level, which is contingent on the availability of a good reference database (Pedrós-Alió et al., 2018). Using the hypervariable V4 region instead of the entire 18S that contains both conserved and variable regions (Neefs et al., 1993), nearly 60% of our retrieved mTags could be assigned to a given reference sequence in our 97%-clustered database, a number that increased when using 151 bp long fragments, showing the expected result of having less ambiguities with longer reads. Our highly-curated V4 region database, *eukaryotesV4*, turns out to be a simple, yet robust reference to correctly discriminate short metagenomic reads of microbial eukaryotes.

A clear advantage of metagenomic approaches over amplicon sequencing to address microbial diversity is that the former does not require a marker gene amplification and thus bypasses the biases that may accompany PCR steps (Acinas et al., 2005; Parada, Needham, & Fuhrman, 2016). There have been several studies comparing metagenomics and metabarcoding in different systems, mostly in prokaryotic communities, with some of them reporting a strong correlation (Fierer et al., 2012), comparable results at the phylum level (Poretsky, Rodriguez-R, Luo, Tsementzi, & Konstantinidis, 2014) or similar community structures in terms of presence or absence of specific taxa (Logares et al., 2014). However, other works have described metagenomics as a better-performing technique in assessing community structure (Shakya et al., 2013) and in revealing uncharacterized diversity (Eloe-Fadrosh, Ivanova, Woyke, & Kyrpides, 2016). Here, we detected similar overall compositions by both approaches, but particular taxonomic groups displayed different relative abundances among them and, in some extreme cases, some groups were only detected by metagenomics (discussed further below). Although it is important to remember that a community profiling without any biases is not attainable in sequencing experiments (McLaren, Willis, & Callahan, 2019), our results indicate that metagenomics yields a more complete image of overall diversity than metabarcoding when assessing eukaryotic communities at the group level, as nearly 99% of the reads are correctly defined to this level. Despite the very good classification at this level, 40% of the mTags matched with identical score to more than one OTU_{97} reference sequence in our database, highlighting the taxonomic limits of our approach. Even for the 60% mTags that were assigned to a unique OTU_{97}

sequence, these represented units clustered at 97%, a threshold at which we already lose taxonomic detail. We could use a database clustered at a higher identity (e.g. 99%), but then we would expect a much lower proportion of mTags unambiguously affiliated to an OTU_{99} reference. Therefore, amplicon sequencing clearly outperforms metagenomics in terms of fine-scale diversity recovery, as state-of-the-art tools are able to infer real biological variants differing by only one nucleotide (Nearing, Douglas, Comeau, & Langille, 2018) out from these data. Altogether, we advocate the use of amplicon sequencing when eukaryotic diversity has to be assessed in detail, and argue that their use is not needed when having metagenomic data and an overall image of the group diversity is sufficient.

When looking at which of the amplified regions (i.e. V4 or V9) in the metabarcoding gave an image closer to the one yielded by metagenomics, there was not a clear winner, as both regions deviated from mTags in different ways, although it is worth noting that V9 was able to detect more groups than V4. There have been other studies carrying out a V4/V9 comparison, some of them yielding similar results in terms of community composition (Kim, Sprung, Duhamel, Filardi, & Kyoon Shin, 2016; Piredda et al., 2017; Tragin, Zingone, & Vaulot, 2018) and others reporting different performances depending on the taxa (Dunthorn, Klier, Bunge, & Stoeck, 2012; Forster et al., 2019; Giner et al., 2016; Pawlowski et al., 2011; Stoeck et al., 2010). In the present work some groups that were relatively abundant in the metagenomes did not appear or were underrepresented by amplicons. That was the case of Prymnesiophyceae in V4, known to have a critical mismatch in our reverse primer used here (Balzano et al., 2015; Piredda et al., 2017), and Diplonemea, Kinetoplastida and Discosea, groups that have typically longer V4 inserts (**Table S1**) that likely limit their amplification. In fact, it was already known that the V4 region of Amoebozoa is not easily amplified (Lahr, Grant, Nguyen, Lin, & Katz, 2011). The V9 region, initially chosen as the first high-throughput sequencing platforms could only work with very short lengths (Amaral-Zettler et al., 2009), is also known to cause some conflicts in taxonomic assignments (Pawlowski et al., 2011) and, more critically, makes it very difficult to place novel sequences within a phylogenetic context (Dunthorn et al., 2014). Therefore, both V4 and V9 amplicons have their limitations and provide complementary information that can be combined to improve community profiling analyses.

The taxonomic information retrieved from mTags using 91 samples from the Malaspina global scale survey revealed a clear separation between pico- (0.2-3 µm) and nanoeukaryotic (3-20 µm) communities. This size-driven differentiation of microbial eukaryotic assemblages was also observed in another large-scale study conducted in the photic region of the global ocean (de

Vargas et al., 2015) using amplicon sequencing, and here we report that this also happens in deeper aphotic waters. In relation to this separation, and of significance, is how the majority of taxonomic groups tend to occupy a different region in the size-depth layer space, thus stressing the importance of size fractionation, as cells from a given taxonomic group tend to belong to the same size class. This claims against treating the eukaryotic piconanoplankton as a uniform assemblage.

The image retrieved in our study on the picosized fraction was broadly similar to that reported in Giner *et al.* (2020) using V4 rDNA amplicons, with MALV groups dominating the photic zone along with Dinoflagellata, Pelagophyceae and Prymnesiophyceae, although this last group was only detected by metagenomics and represented a critical difference. In the aphotic layer, both approaches revealed a dominance of MALV, Polycystinea, Acantharia, and RAD-B, together with Diplonemea only detected by metagenomics. The fact that some taxonomic groups that are typically larger (e.g. Radiolaria) were found in the picoeukaryotic fraction could be explained by the presence of smaller life cycle stages or to filtration artifacts (Massana et al., 2015). Apart from Prymnesiophyceae and Diplonemea, we could retrieve a relatively important signal of Kinetoplastida in the picoeukaryotic fraction by means of metagenomics (median relative abundance 0.5%), a similar abundance already seen in Tara Oceans V9 amplicons but not detected in the metagenomes from that same dataset (Flegontova et al., 2018).

In the nano fraction, the photic layer was highly dominated by Dinoflagellata with more than 60% of median relative abundance, followed by MALV-I, Prymnesiophyceae and MALV-II. These high numbers of Alveolata groups in the sunlit part of the ocean were also reported in the 5-20 µm fraction of Tara Oceans amplicons (de Vargas et al., 2015). There, Rhizaria was also found to contribute largely to total reads, a trend that was not observed here, where the most abundant rhizarian group was RAD-B with a median relative abundance of 1%. In the aphotic part of the water column, the most dominant groups were Polycystinea, Diplonemea and Dinoflagellata. High abundances of these taxa have also been reported in regional (Countway et al., 2007; Zoccarato, Pallavicini, Cerino, Fonda Umani, & Celussi, 2016) and global (Pernice *et al.* 2016) deep water studies. In the latter, eukaryotic diversity was assessed on bathypelagic samples derived from the same Malaspina Expedition and their results, considering that the size fraction covered was the piconanoplankton (0.8-20 µm), are comparable to our results when combining the image given by aphotic samples of both pico- and nanoeukaryotic fractions. The detection of Diplonemea in all the above-mentioned works is explained by the fact that they used either PCR-based approaches not targeting the V4 region or metagenomes. The high presence of Discosea

in our dataset (2% median relative abundance), naked amoeboid protists still poorly assessed that have been found in both deep pelagic and benthic marine areas (Kudryavtsev & Pawlowski, 2013, 2015) hints at the relevance of these microorganisms in planktonic ecosystems, as well as Fungi, mainly Ascomycota and Basidiomycota, which seem to be important players in all aquatic ecosystems (Grossart et al., 2019).

Another type of valuable information on microbial eukaryotic diversity can be found in contigs in the assembled metagenomes containing long rDNA sequences. These contigs, sometimes encompassing full rDNA operons, allowed us to jump from the group and OTU_{97} levels reached by mTags to a high-resolution species level, and to expand the available taxonomic data for some eukaryotic groups. As a proof of concept, in this work we assessed the diversity within diplonemids, which have been recently reported as one of the most species-rich eukaryotic group in marine planktonic systems by means of amplicon sequencing (Flegontova et al., 2016) and were poorly represented in our amplicon comparison. As previous metabarcoding (Flegontova et al., 2016; Lara, Moreira, Vereshchaka, & López-García, 2009) and single cell (Gawryluk et al., 2016) studies found, the vast majority of the contigs we retrieved belonged to Eupelagonemidae (Okamoto et al., 2019; Tashyreva et al., 2018), formerly treated as DSPD II, a very diverse deep-branching monophyletic clade (Lara et al., 2009). A number of these recovered contigs were relatively distant to known reference sequences, thus confirming that part of the diversity of these microorganisms is yet to be explored.

Overall, our study reveals that the analysis of metagenomes using a well curated database provides very good taxonomic assignment of the groups dominating marine assemblages. Despite lower taxonomic resolution compared to amplicons, mTags outperformed these when defining community composition at the general group level, as they did not suffer from PCR biases. The obtained results on pico- and nanoeukaryotic diversity revealed a clear separation between size fractions and water depths in terms of community composition and allowed us to better define the ecological context of the main eukaryotic groups populating the global ocean.

Acknowledgments

This research was supported by the Spanish Ministry of Economy and Competitiveness projects Malaspina-2010 (CSD2008–00077) and ALLFLAGS (CTM2016-75083-R) and King Abdullah University of Science and Technology (KAUST) under contract OSR #3362. AO was supported by a Spanish FPI grant. We thank all scientists and crew that participated in the Malaspina 2010

expedition. We also thank Javier del Campo for useful discussions. Bioinformatic analyses were performed at the Marbits platform (ICM-CSIC; https://marbits.icm.csic.es).

References

Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V., & Polz, M. F. (2005). PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. *Applied and Environmental Microbiology*, *71*(12), 8966–8969. doi: 10.1128/AEM.71.12.8966-8969.2005

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Amaral-Zettler, L., McCliment, E. A., Ducklow, H. W., & 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*(7), 1–9. doi: 10.1371/journal.pone.0006372

Bahram, M., Hildebrand, F., Forslund, S. K., Anderson, J. L., Soudzilovskaia, N. A., Bodegom, P. M., … Bork, P. (2018). Structure and function of the global topsoil microbiome. *Nature*, *560*(7717), 233–237. doi: 10.1038/s41586-018-0386-6

Balzano, S., Abs, E., & Leterme, S. C. (2015). Protist diversity along a salinity gradient in a coastal lagoon. *Aquatic Microbial Ecology*, *74*(3), 263–277. doi: 10.3354/ame01740

Bengtsson-Palme, J., Hartmann, M., Eriksson, K. M., Pal, C., Thorell, K., Larsson, D. G. J., & Nilsson, R. H. (2015). metaxa2: improved identification and taxonomic classification of small and large subunit rRNA in metagenomic data. *Molecular Ecology Resources*, *15*(6), 1403–1414. doi: 10.1111/1755-0998.12399

Bengtsson, J., Eriksson, K. M., Hartmann, M., Wang, Z., Shenoy, B. D., Grelet, G.-A., … Nilsson, R. H. (2011). Metaxa: a software tool for automated detection and discrimination among ribosomal small subunit (12S/16S/18S) sequences of archaea, bacteria, eukaryotes, mitochondria, and chloroplasts in metagenomes and environmental sequencing datasets. *Antonie van Leeuwenhoek*, *100*(3), 471–475. doi: 10.1007/s10482-011-9598-6

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. doi: 10.1093/bioinformatics/btu170

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2017). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*, (June), 1–15. doi: 10.1093/bib/bbx120

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. doi: 10.1038/nmeth.3869

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. doi: 10.1093/bioinformatics/btp348

Caron, D. A., Countway, P. D., Jones, A. C., Kim, D. Y., & Schnetzer, A. (2012). Marine Protistan Diversity. *Annual Review of Marine Science*, *4*(1), 467–493. doi: 10.1146/annurev-marine-120709-142802

Countway, P. D., Gast, R. J., Dennett, M. R., Savai, P., Rose, J. M., & 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). *Environmental Microbiology*, *9*(5), 1219–1232. doi: 10.1111/j.1462-2920.2007.01243.x

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., … Velayoudon, D. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1261605–1261605. doi: 10.1126/science.1261605

Diez, B., Pedrós-Alió, C., & Massana, R. (2001). Study of Genetic Diversity of Eukaryotic Picoplankton in Different Oceanic Regions by Small-Subunit rRNA Gene Cloning and Sequencing. *Applied and Environmental Microbiology*, *67*(7), 2932–2941. doi: 10.1128/AEM.67.7.2932-2941.2001

Duarte, C. M. (2015). Seafaring in the 21st century: The Malaspina 2010 circumnavigation expedition. *Limnology and Oceanography Bulletin*, *24*(1), 11–14. doi: 10.1002/lob.10008

Dunthorn, M., Klier, J., Bunge, J., & Stoeck, T. (2012). Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. *Journal of Eukaryotic Microbiology*, *59*(2), 185–187. doi: 10.1111/j.1550-7408.2011.00602.x

Dunthorn, M., Otto, J., Berger, S. A., Stamatakis, A., Mahé, F., Romac, S., … Zingone, A. (2014). Placing environmental next-generation sequencing amplicons from microbial eukaryotes into a phylogenetic context. *Molecular Biology and Evolution*, *31*(4), 993–1009. doi: 10.1093/molbev/msu055

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. doi: 10.1093/bioinformatics/btq461

Edgcomb, V. P. (2016). Marine protist associations and environmental impacts across trophic levels in the twilight zone and below. *Current Opinion in Microbiology*, *31*, 169–175. doi: 10.1016/j.mib.2016.04.001

Edgcomb, V. P., Kysela, D. T., Teske, A., de Vera Gomez, A., & Sogin, M. L. (2002). Benthic eukaryotic diversity in the Guaymas Basin hydrothermal vent environment. *Proceedings of the National Academy of Sciences*, *99*(11), 7658–7662. doi: 10.1073/pnas.062186399

Eloe-Fadrosh, E. A., Ivanova, N. N., Woyke, T., & Kyrpides, N. C. (2016). Metagenomics

uncovers gaps in amplicon-based detection of microbial diversity. *Nature Microbiology*, *1*, 15032. doi: 10.1038/nmicrobiol.2015.32

Falkowski, P. G. (2012). Ocean Science: The power of plankton. *Nature*, *483*(7387), S17–S20. doi: 10.1038/483S17a

Fierer, N., Leff, J. W., Adams, B. J., Nielsen, U. N., Bates, S. T., Lauber, C. L., … Caporaso, J. G. (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proceedings of the National Academy of Sciences*, *109*(52), 21390–21395. doi: 10.1073/pnas.1215210110

Flegontova, O., Flegontov, P., Malviya, S., Audic, S., Wincker, P., de Vargas, C., … Horák, A. (2016). Extreme Diversity of Diplonemid Eukaryotes in the Ocean. *Current Biology*, *26*(22), 3060–3065. doi: 10.1016/j.cub.2016.09.031

Flegontova, O., Flegontov, P., Malviya, S., Poulain, J., de Vargas, C., Bowler, C., … Horák, A. (2018). Neobodonids are dominant kinetoplastids in the global ocean. *Environmental Microbiology*, *20*(2), 878–889. doi: 10.1111/1462-2920.14034

Forster, D., Filker, S., Kochems, R., Breiner, H. W., Cordier, T., Pawlowski, J., & Stoeck, T. (2019). A Comparison of Different Ciliate Metabarcode Genes as Bioindicators for Environmental Impact Assessments of Salmon Aquaculture. *Journal of Eukaryotic Microbiology*, *66*(2), 294–308. doi: 10.1111/jeu.12670

Gawryluk, R. M. R., del Campo, J., Okamoto, N., Strassert, J. F. H., Lukeš, J., Richards, T. A., … Keeling, P. J. (2016). Morphological Identification and Single-Cell Genomics of Marine Diplonemids. *Current Biology*, *26*(22), 3053–3059. doi: 10.1016/j.cub.2016.09.013

Giner, C. R., Balagué, V., Krabberød, A. K., Ferrera, I., Reñé, A., Garcés, E., … Massana, R. (2019). Quantifying long-term recurrence in planktonic microbial eukaryotes. *Molecular Ecology*, *28*(5), 923–935. doi: 10.1111/mec.14929

Giner, C. R., Forn, I., Romac, S., Logares, R., de Vargas, C., & Massana, R. (2016). Environmental Sequencing Provides Reasonable Estimates of the Relative Abundance of Specific Picoeukaryotes. *Applied and Environmental Microbiology*, *82*(15), 4757–4766. doi: 10.1128/aem.00560-16

Giner, C. R., Pernice, M. C., Balagué, V., Duarte, C. M., Gasol, J. M., Logares, R., & Massana, R. (2020). Marked changes in diversity and relative activity of picoeukaryotes with depth in the world ocean. *The ISME Journal*, *14*(2), 437–449. doi: 10.1038/s41396-019-0506-9

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: Ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, *17*(6), 333–351. doi: 10.1038/nrg.2016.49

Grossart, H. P., Van den Wyngaert, S., Kagami, M., Wurzbacher, C., Cunliffe, M., & Rojas-

Jimenez, K. (2019). Fungi in aquatic ecosystems. *Nature Reviews Microbiology*. doi: 10.1038/s41579-019-0175-8

Gruber-Vodicka, H. R., Seah, B. K., & Pruesse, E. (2019). phyloFlash — Rapid SSU rRNA profiling and targeted assembly from metagenomes. *BioRxiv*, 521922. doi: https://doi.org/10.1101/521922

Guajardo-Leiva, S., Pedrós-Alió, C., Salgado, O., Pinto, F., & Díez, B. (2018). Active Crossfire Between Cyanobacteria and Cyanophages in Phototrophic Mat Communities Within Hot Springs. *Frontiers in Microbiology*, *9*. doi: 10.3389/fmicb.2018.02039

Guidi, L., Chaffron, S., Bittner, L., Eveillard, D., Larhlimi, A., Roux, S., … Gorsky, G. (2016). Plankton networks driving carbon export in the oligotrophic ocean. *Nature*, *532*(7600), 465–470. doi: 10.1038/nature16942

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., … Christen, R. (2013). The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), 597–604. doi: 10.1093/nar/gks1160

Guo, J., Cole, J. R., Zhang, Q., Brown, C. T., & Tiedje, J. M. (2016). Microbial Community Analysis with Ribosomal Gene Fragments from Shotgun Metagenomes. *Applied and Environmental Microbiology*, *82*(1), 157–166. doi: 10.1128/AEM.02772-15

Hartmann, M., Howes, C. G., Abarenkov, K., Mohn, W. W., & Nilsson, R. H. (2010). V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *Journal of Microbiological Methods*, *83*(2), 250–253. doi: 10.1016/j.mimet.2010.08.008

Huang, Y., Gilna, P., & Li, W. (2009). Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics*, *25*(10), 1338–1340. doi: 10.1093/bioinformatics/btp161

Jürgens, K., & Massana, R. (2008). Protistan Grazing on Marine Bacterioplankton. In D. L. Kirchman (Ed.), *Microbial Ecology of the Oceans* (2nd ed., pp. 383–441). doi: 10.1002/9780470281840.ch11

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. doi: 10.1093/molbev/mst010

Kim, E., Sprung, B., Duhamel, S., Filardi, C., & Kyoon Shin, M. (2016). Oligotrophic lagoons of the South Pacific Ocean are home to a surprising number of novel eukaryotic microorganisms. *Environmental Microbiology*, *18*(12), 4549–4563. doi: 10.1111/1462-2920.13523

Kudryavtsev, A., & Pawlowski, J. (2013). Squamamoeba japonica n. g. n. sp. (Amoebozoa): A

Deep-sea Amoeba from the Sea of Japan with a Novel Cell Coat Structure. *Protist*, *164*(1), 13–23. doi: 10.1016/j.protis.2012.07.003

Kudryavtsev, A., & Pawlowski, J. (2015). Cunea n. g. (Amoebozoa, Dactylopodida) with two cryptic species isolated from different areas of the ocean. *European Journal of Protistology*, *51*(3), 197–209. doi: 10.1016/j.ejop.2015.04.002

Lahr, D. J. G., Grant, J., Nguyen, T., Lin, J. H., & Katz, L. A. (2011). Comprehensive phylogenetic reconstruction of amoebozoa based on concatenated analyses of SSU-rDNA and actin genes. *PLoS ONE*, *6*(7), e22780. doi: 10.1371/journal.pone.0022780

Lara, E., Moreira, D., Vereshchaka, A., & López-García, P. (2009). Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. *Environmental Microbiology*, *11*(1), 47–55. doi: 10.1111/j.1462-2920.2008.01737.x

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276–3278. doi: 10.1093/bioinformatics/btu531

Li, D., Liu, C. M., Luo, R., Sadakane, K., & Lam, T. W. (2015). MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, *31*(10), 1674–1676. doi: 10.1093/bioinformatics/btv033

Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D., & Knight, R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Research*, *35*(18), e120–e120. doi: 10.1093/nar/gkm541

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmento, H., … Acinas, S. G. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*, *16*(9), 2659–2671. doi: 10.1111/1462-2920.12250

López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C., & Moreira, D. (2001). Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature*, *409*(6820), 603–607. doi: 10.1038/35054537

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10. doi: 10.14806/ej.17.1.200

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., … de Vargas, C. (2015). Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology*, *17*(10), 4035–4049. doi: 10.1111/1462-2920.12955

Massana, R., Murray, A., Preston, C., & DeLong, E. (1997). Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Applied and Environmental Microbiology*, *63*(1), 50–56.

McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *ELife*, *8*, 559831. doi: 10.7554/eLife.46923

Moon-Van Der Staay, S. Y., De Wachter, R., & Vaulot, D. (2001). Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*, *409*(6820), 607–610. doi: 10.1038/35054541

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. I. (2018). Denoising the Denoisers: an independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, *6*, e5364. doi: 10.7717/peerj.5364

Neefs, J.-M., Van de Peer, Y., De Rijk, P., Chapelle, S., & De Wachter, R. (1993). Compilation of small ribosomal subunit RNA structures. *Nucleic Acids Research*, *21*(13), 3025–3049. doi: 10.1093/nar/21.13.3025

Not, F., del Campo, J., Balagué, V., de Vargas, C., & Massana, R. (2009). New insights into the diversity of marine picoeukaryotes. *PLoS ONE*, *4*(9). doi: 10.1371/journal.pone.0007143

Okamoto, N., Gawryluk, R. M. R., del Campo, J., Strassert, J. F. H., Lukeš, J., Richards, T. A., … Keeling, P. J. (2019). A Revised Taxonomy of Diplonemids Including the Eupelagonemidae n. fam. and a Type Species, Eupelagonema oceanica n. gen. & sp. *Journal of Eukaryotic Microbiology*, *66*(3), 519–524. doi: 10.1111/jeu.12679

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … Wagner, H. (2019). *vegan: Community Ecology Package*. Retrieved from https://cran.r-project.org/package=vegan

Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology*, *18*(5), 1403–1414. doi: 10.1111/1462-2920.13023

Pawlowski, J., Christen, R., Lecroq, B., Bachar, D., Shahbazkia, H. R., Amaral-Zettler, L., & Guillou, L. (2011). Eukaryotic richness in the abyss: Insights from pyrotag sequencing. *PLoS ONE*, *6*(4). doi: 10.1371/journal.pone.0018169

Pedrós-Alió, C., Acinas, S. G., Logares, R., & Massana, R. (2018). Marine microbial diversity as seen by high-throughput sequencing. In J. M. Gasol & D. L. Kirchman (Eds.), *Microbial Ecology of the Oceans* (pp. 47–98). Wiley-Blackwell.

Pernice, M. C., Giner, C. R., Logares, R., Perera-Bel, J., Acinas, S. G., Duarte, C. M., … Massana, R. (2016). Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. *ISME Journal*, *10*(4), 945–958. doi: 10.1038/ismej.2015.170

Piganeau, G., Desdevises, Y., Derelle, E., & Moreau, H. (2008). Picoeukaryotic sequences in the Sargasso Sea metagenome. *Genome Biology*, *9*(1), R5. doi: 10.1186/gb-2008-9-1-r5

Piredda, R., Tomasino, M. P., D'Erchia, A. M., Manzari, C., Pesole, G., Montresor, M., … Zingone, A. (2017). Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. *FEMS Microbiology Ecology*, *93*(1), fiw200. doi: 10.1093/femsec/fiw200

Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D., & Konstantinidis, K. T. (2014). Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLoS ONE*, *9*(4). doi: 10.1371/journal.pone.0093827

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., … Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*(D1), 590–596. doi: 10.1093/nar/gks1219

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.r-project.org/

Saghaï, A., Zivanovic, Y., Zeyen, N., Moreira, D., Benzerara, K., Deschamps, P., … López-García, P. (2015). Metagenome-based diversity analyses suggest a significant contribution of non-cyanobacterial lineages to carbonate precipitation in modern microbialites. *Frontiers in Microbiology*, *6*(AUG), 797. doi: 10.3389/fmicb.2015.00797

Shakya, M., Quince, C., Campbell, J. H., Yang, Z. K., Schadt, C. W., & Podar, M. (2013). Comparative metagenomic and rRNA microbial diversity characterization using archaeal and bacterial synthetic communities. *Environmental Microbiology*, *15*(6), 1882–1899. doi: 10.1111/1462-2920.12086

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE*, *11*(10), e0163962. doi: 10.1371/journal.pone.0163962

Sinclair, L., Osman, O. A., Bertilsson, S., & Eiler, A. (2015). Microbial community composition and diversity via 16S rRNA gene amplicons: Evaluating the illumina platform. *PLoS ONE*, *10*(2), 1–18. doi: 10.1371/journal.pone.0116955

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi: 10.1093/bioinformatics/btu033

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H. W., & Richards, T. A. (2010). Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, *19*(SUPPL. 1), 21–31. doi: 10.1111/j.1365-294X.2009.04480.x

Tashyreva, D., Prokopchuk, G., Yabuki, A., Kaur, B., Faktorová, D., Votýpka, J., … Lukeš, J. (2018). Phylogeny and Morphology of New Diplonemids from Japan. *Protist*, *169*(2), 158–179. doi: 10.1016/j.protis.2018.02.001

Tragin, M., Zingone, A., & Vaulot, D. (2018). Comparison of coastal phytoplankton composition estimated from the V4 and V9 regions of the 18S rRNA gene with a focus on photosynthetic groups and especially Chlorophyta. *Environmental Microbiology*, *20*(2), 506–520. doi: 10.1111/1462-2920.13952

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi: 10.21105/joss.01686

Worden, A. Z., Follows, M. J., Giovannoni, S. J., Wilken, S., Zimmerman, A. E., & Keeling, P. J. (2015). Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science*, *347*(6223), 1257594–1257594. doi: 10.1126/science.1257594

Zoccarato, L., Pallavicini, A., Cerino, F., Fonda Umani, S., & Celussi, M. (2016). Water mass dynamics shape Ross Sea protist communities in mesopelagic and bathypelagic layers. *Progress in Oceanography*, *149*, 16–26. doi: 10.1016/j.pocean.2016.10.003

Data Accessibility Statement

- ***eukaryotesV4* database.** Available at https://github.com/aleixop/eukaryotesV4 with DOI 10.5281/zenodo.3522173.
- **Data processing, analysis scripts, mTags and ASV tables.** Available at https://github.com/aleixop/Malaspina_Euk_mTags with DOI 10.5281/zenodo.3629394.
- **Fasta files with retrieved mTags and contigs.** Available at https://github.com/aleixop/Malaspina_Euk_mTags with DOI 10.5281/zenodo.3629394.
- **V4 amplicon sequences.** Deposited at European Nucleotide Archive, accession number PRJEB23771, from a previous study (Giner et al., 2020)..
- **V9 amplicon sequences.** Deposited at European Nucleotide Archive, under accession number PRJEB36469.

Author contributions

CMD, SGA and RM designed research; CRG and AO processed the samples in the laboratory; CRG, PS and AO analyzed the data; AO and RM interpreted the results and wrote the paper. All authors contributed substantially to manuscript revisions.

Figure legends

Figure 1. Pipeline for V4-18S rDNA mTags extraction from metagenomes (metaG) and classification and technical results. (A) Flow diagram of the pipeline used in this study. (B) Number of V4-18S mTags retrieved at the four defined taxonomic levels. (C) Relative abundance of the three corresponding taxonomic levels within each supergroup.

Figure 2. Distribution of the relative abundance of main taxonomic groups as seen by each of the three sequencing approaches (mTags, V4 amplicons and V9 amplicons) in a subset of 23 picoeukaryotic (0.2-3 µm) samples from 4 vertical profiles. Groups are ordered by decreasing median of the relative abundance by mTags. A $\log_{10}$ scale on the y-axis is used. Significant differences between approaches with Wilcoxon paired tests are shown (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

Figure 3. Distribution of the relative abundance of main eukaryotic groups from the pico- (0.2-3 µm) and nanoeukaryotic (3-20 µm) fractions in photic ('P') and aphotic ('A') layers of the ocean as seen by mTags. Groups are ordered by decreasing median relative abundances and $\log_{10}$ scale on the y-axis is used.

Figure 4. Summary of the relevance of main taxonomic groups in pico(0.2-3 µm)/nanoeukaryotic(3-20µm) fractions and photic/aphotic layers as seen by 18S mTags. The median of the relative abundance was calculated for each taxonomic group with samples from the 4 categories (pico-photic, pico-aphotic, nano-photic, nano-aphotic) and dots represent these median values transformed to a 0-100 scale. These are colored based on the category where each taxonomic group is most relevant.

Figure 5. Maximum likelihood phylogenetic tree of Diplonemea using the 18S rDNA from metagenome contigs and reference sequences from the *eukaryotesV4* database, derived from an alignment of 940 positions where only shared regions among sequences were kept. RAxML bootstraps are shown when the value is >50 and black dots represent 100% support. Reference sequences with at least one hit in the mTags analysis are highlighted, as well as their total number of hits in logarithmic scale.

Table 1. Overview of the taxonomic affiliation of the 724 contigs having at least 1000 bp in the 18S rDNA, together with their coverage of the rest of the rDNA operon. In the circle pairs, left circle represents 18S rDNA gene and right circle the 28S rDNA gene, while black-colored circles represent full gene sequences and half colored ones represent partial gene sequences.

Supplementary Information

Figure S1. Map showing the geographic position (black dots) of the 10 sampling stations and the sequencing analysis done in each station.

Figure S2. Differences on taxonomic resolution between using the original 151 bp sized mTags from some metagenomes, and the same dataset trimmed at 101 bp.

Figure S3. Comparison of 18S-V4 mTags extraction methods. Correlations between the number of mTags extracted by direct BLAST mapping and HMM profiling within the different supergroups, which are ordered by decreasing differences. Linear regression equations and r-squared coefficients are shown.

Figure S4. Comparison of 23 samples surveyed with 18S-V4 mTags, V4 and V9 amplicons in a NMDS plot based on Bray-Curtis dissimilarities of community structures defined by the relative abundances of taxonomic groups. Lines join the same environmental sample.

Figure S5. Vertical distribution of main taxonomic groups along the water column as represented by mTags for pico- (A) and nanoeukaryotic (B) fractions (0.2-3 µm and 3-20 µm, respectively). Each dot represents relative abundance of that group in a specific sample and is colored by ocean. Depth axis has been modified to display the same distance in the three vertical ocean layers: epipelagic: 0-200m, mesopelagic: 200-1000m, bathypelagic: 1000-4000m.

Figure S6. Comparison of the community structure of pico- (0.2-3 µm) and nanoeukaryotic (3-20 µm) fractions along the global ocean as seen in non-metric multidimensional scaling based on Bray-Curtis dissimilarities. Relative abundances of OTU-defined mTags are used.

Table S1. Overview of the *eukaryotesV4* database, indicating the taxonomic affiliation of the reference sequences as well as their size in base pairs (bp; range and average). Supergroups and groups are ordered alphabetically.

Table S2. Analyzed samples and their associated metadata.

Table S3. List of contigs containing >1000 bp of 18S rDNA and their taxonomy. Total lengths of each contig, as well as coordinates for the 18S and 28S genes are displayed. For the 18S, closest match to all NCBI nt database ('ncbi all') and closest match to NCBI nt database excluding environmental sequences ('ncbi cultured') are given when available.

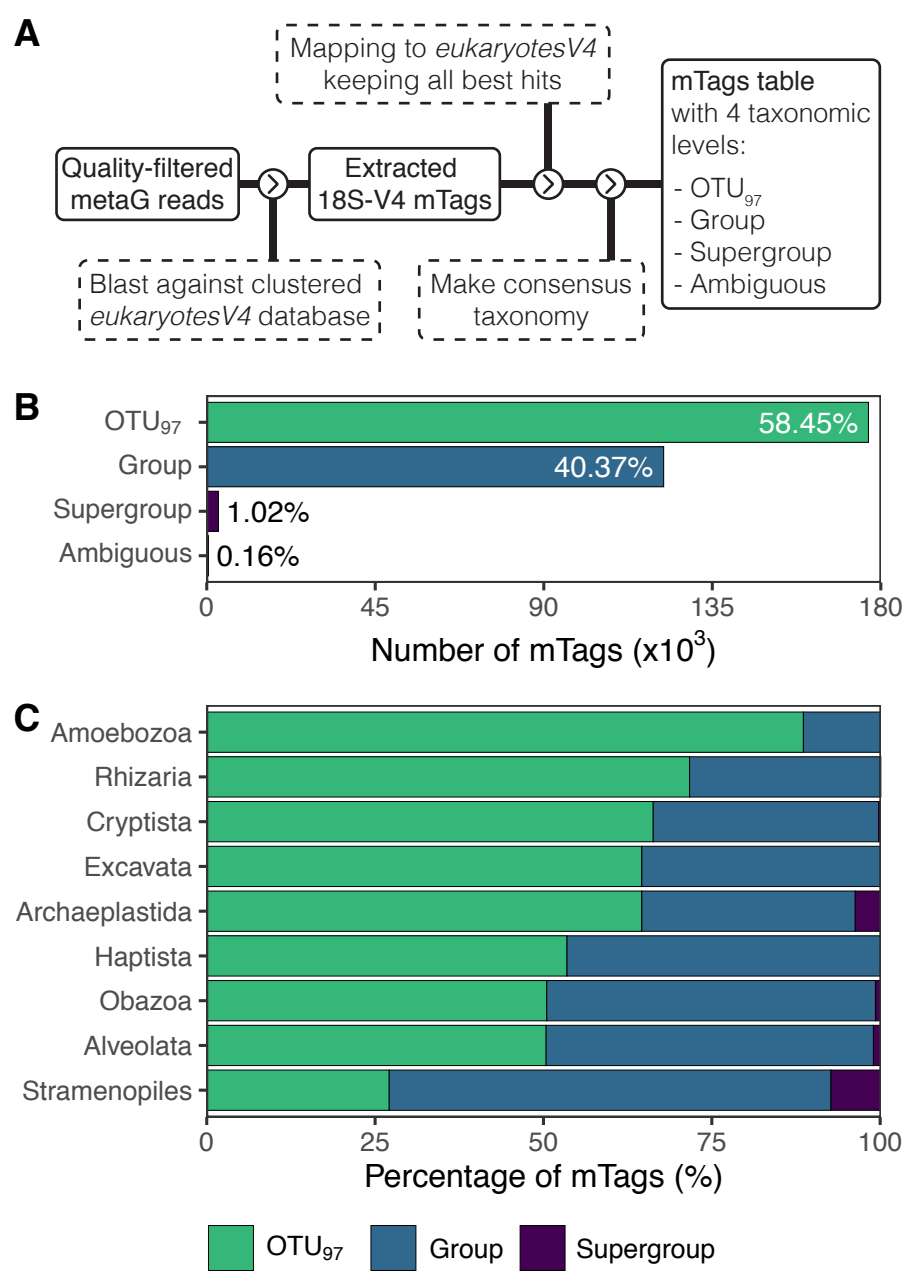
Table S4. Summary of median relative abundances and mean ± standard error for main taxonomic groups in pico(0.2-3 µm)/nanoeukaryotic(3-20 µm) fractions and photic/aphotic layers as seen by mTags.

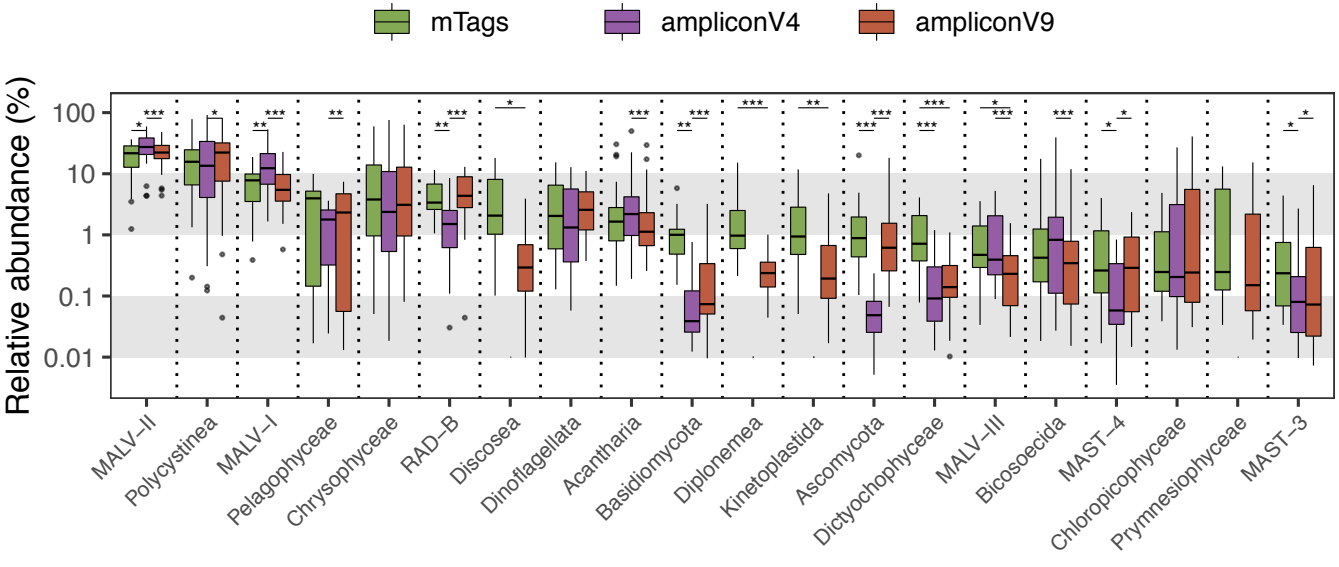
Group	Total contigs						
Polycystinea	158	79	11	10	37	2	19
MALV-I	87	62	6	6	13	–	–
Dinoflagellata	53	34	3	4	12	–	–
Diplonemea	51	43	4	4	–	–	–
Acantharia	47	35	6	2	1	–	3
Discosea	46	27	2	1	13	–	3
MALV-II	35	28	6	–	1	–	–
Chrysophyceae	29	17	1	–	2	1	8
Basidiomycota	28	13	8	3	3	1	–
RAD-B	28	20	4	3	1	–	–
Prymnesiophyceae	23	20	1	2	–	–	–
Kinetoplastida	20	8	2	2	6	–	2
Ascomycota	19	8	3	1	7	–	–
InSedAlveolata	11	9	1	–	1	–	–
Bicosoecida	10	3	1	–	4	–	2
Ciliophora	9	8	–	–	–	–	1
RAD-A	9	8	–	1	–	–	–
Pelagophyceae	8	7	–	–	–	1	–
Apicomplexa	7	3	1	–	3	–	–
RAD-C	7	6	1	–	–	–	–
InSedEukaryota	6	5	–	1	–	–	–
Cercozoa	5	4	1	–	–	–	–
Katablepharidae	4	4	–	–	–	–	–
Telonemia	4	4	–	–	–	–	–
Chloropicophyceae	3	1	–	2	–	–	–
Euglenida	3	–	–	1	2	–	–
Foraminifera	3	2	–	–	1	–	–
Choanomonada	2	2	–	–	–	–	–
Diatomea	2	2	–	–	–	–	–
Mamiellophyceae	2	2	–	–	–	–	–
Dictyochophyceae	1	1	–	–	–	–	–
Ellobiopsidae	1	1	–	–	–	–	–
MALV-III	1	1	–	–	–	–	–
MOCH-2	1	1	–	–	–	–	–
Prasino-Clade-IX	1	1	–	–	–	–	–
TOTAL	724	469	62	43	107	5	38

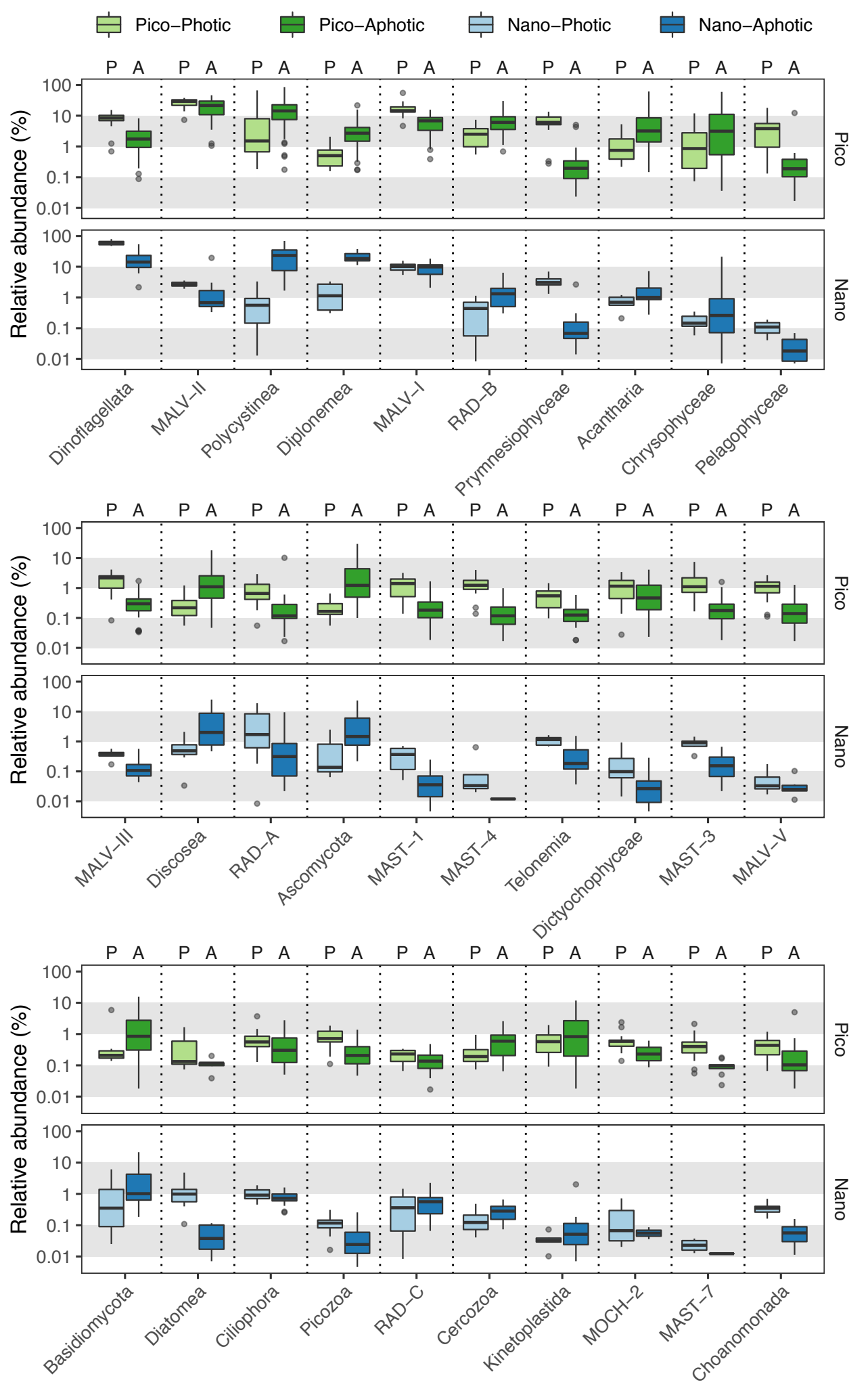
rDNA operon

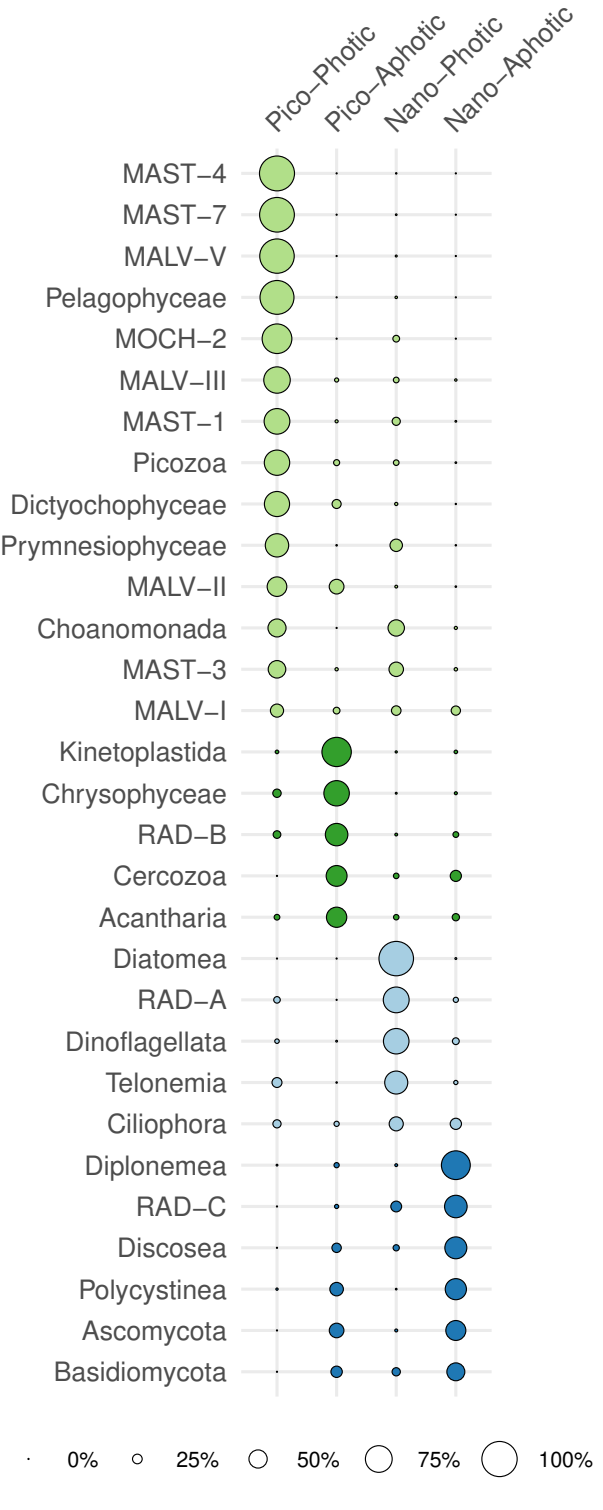

complete


partial


A
Mapping to eukaryotesV4 keeping all best hits
Quality-filtered metaG reads
Extracted 18S-V4 mTags
mTags table with 4 taxonomic levels:
- OTU97
- Group
- Supergroup
- Ambiguous
Blast against clustered eukaryotesV4 database
Make consensus taxonomy
B
OTU97
Group
Supergroup
Ambiguous
58.45%
40.37%
1.02%
0.16%
0
45
90
135
180
Number of mTags (x10^3)
C
Amoebozoa
Rhizaria
Cryptista
Excavata
Archaeplastida
Haptista
Obazoa
Alveolata
Stramenopiles
0
25
50
75
100
Percentage of mTags (%)
OTU97
Group
Supergroup


mTags
ampliconV4
ampliconV9
Relative abundance (%)
100
10
1
0.1
0.01
MALV−II
Polycystinea
MALV−I
Pelagophyceae
Chrysophyceae
RAD−B
Discosea
Dinoflagellata
Acantharia
Basidiomycota
Diplonemea
Kinetoplastida
Ascomycota
Dictyochophyceae
MALV−III
Bicosoecida
MAST−4
Chloropicophyceae
Prymnesiophyceae
MAST−3


Pico−Photic
Pico−Aphotic
Nano−Photic
Nano−Aphotic
P A
Relative abundance (%)
100
10
1
0.1
0.01
Pico
Nano
Dinoflagellata
MALV−II
Polycystinea
Diplonemea
MALV−I
RAD−B
Prymnesiophyceae
Acantharia
Chrysophyceae
Pelagophyceae
MALV−III
Discosea
RAD−A
Ascomycota
MAST−1
MAST−4
Telonemia
Dictyochophyceae
MAST−3
MALV−V
Basidiomycota
Diatomea
Ciliophora
Picozoa
RAD−C
Cercozoa
Kinetoplastida
MOCH−2
MAST−7
Choanomonada


Pico−Photic
Pico−Aphotic
Nano−Photic
Nano−Aphotic
MAST−4
MAST−7
MALV−V
Pelagophyceae
MOCH−2
MALV−III
MAST−1
Picozoa
Dictyochophyceae
Prymnesiophyceae
MALV−II
Choanomonada
MAST−3
MALV−I
Kinetoplastida
Chrysophyceae
RAD−B
Cercozoa
Acantharia
Diatomea
RAD−A
Dinoflagellata
Telonemia
Ciliophora
Diplonemea
RAD−C
Discosea
Polycystinea
Ascomycota
Basidiomycota
0%
25%
50%
75%
100%

References with mTags hits

References with no mTags hits

Picoeukaryotic (0.2-3 μm) fraction

Nanoeukaryotic (3-20 μm) fraction

log(number of mTags hits)

0 2 4 6

EU635624
FN598413
EU635630
EU635672
KJ762505
KJ762788
FJ000255
DQ504321
KJ762626
FN598252
KJ760048
contig 1346829
KJ762734
82
contig 1035430
KJ757153
DQ504323
64
FN598233
contig 1033600
65
KJ760279
contig 1007069
KJ757684
contig 1309735
KJ757377
contig 1004022
61
KJ761007
75
contig 1010022
contig 1210415
contig 1280208
contig 1430045
56
contig 1154552
contig 1183080
69
contig 1151939
contig 1295756
KJ762718
89
contig 1180741
99
AY665087
AY665088
FJ032684
KJ757308
DQ504319
KJ762486
contig 1248168
FN598250
83
DQ504322
66
EU635613
FN598248
KJ761512
AF290080
FN598297
EU635666
FJ000260
contig 1793027
82
FN598247
57
KJ757619
KJ759918
KJ759743
contig 11272556
62
FN598436
contig 1867246
65
contig 1811634
KJ760020
Hemistasia phaeocysticola AB948221
KJ762701
DQ504350
95
65
Rhynchopus sp. SH-2004-I AY490209
Rhynchopus sp. ATCC 50229 AY425014
91
93
Rhynchopus sp. ATCC 50226 AF380997
88
AB505540
Rhynchopus sp. SH-2004-II AY490210
FN598401
Diplonema ambulator AY425009
98
89
Diplonema sp. ATCC 50225 AY425011
Diplonema papillatum KF633466
95
Diplonema sp. ATCC 50232 AY425012
Flectonema neradi AY425010
Rhynchomonas nasuta AY998642

0.2

Eupelagonemidae

'DSPD II'

Hemistasiidae

Diplonemidae