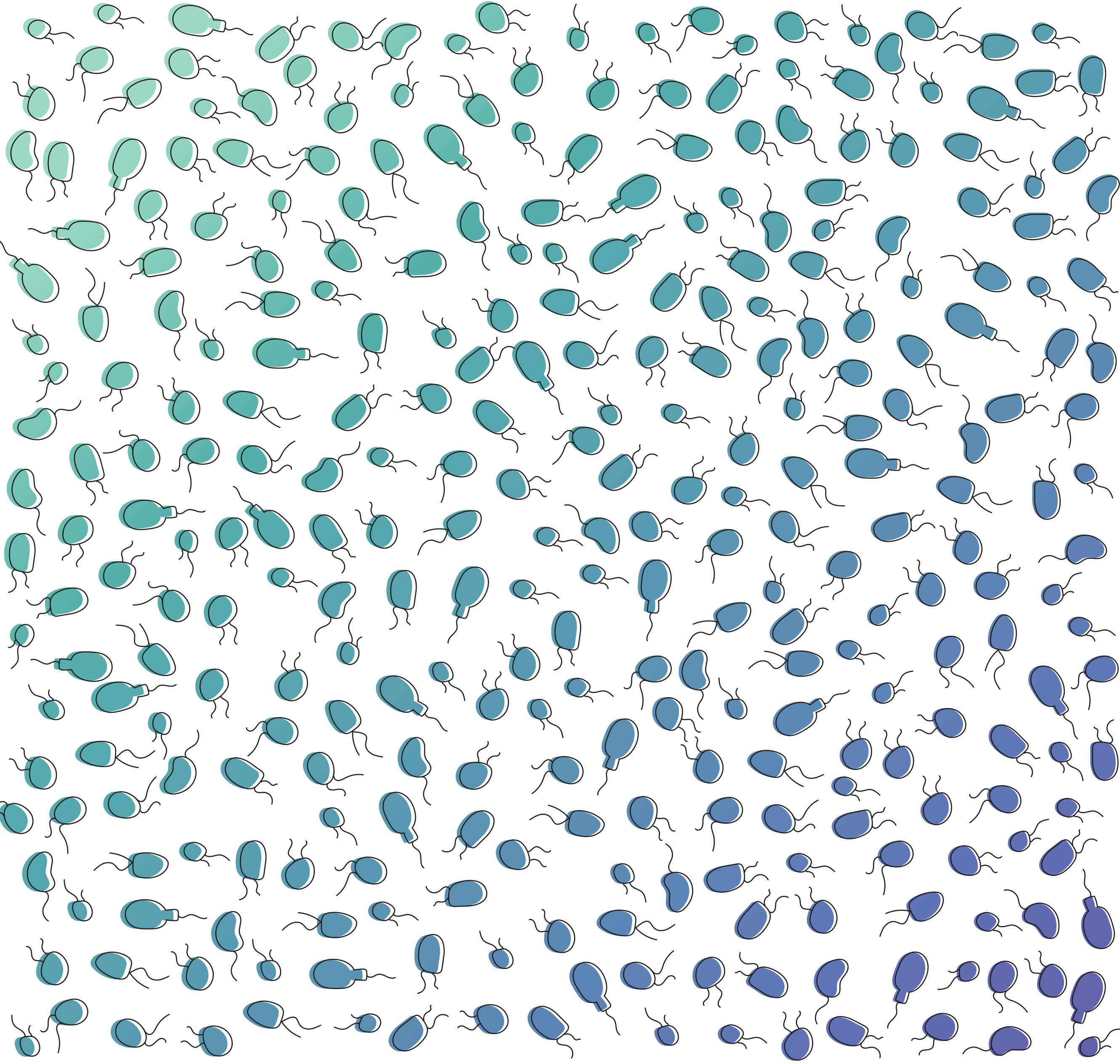
UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Return of the marine heterotrophic flagellates: diversity, distribution and gene expression patterns

Aleix Obiol Plana



PhD thesis

Return of the marine heterotrophic flagellates:

diversity, distribution and gene expression patterns

Aleix Obiol Plana

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Programa de doctorat en Ciències del Mar

Barcelona, setembre 2022

Return of the marine heterotrophic flagellates: diversity, distribution and gene expression patterns

Aleix Obiol Plana
Departament de Biologia Marina i Oceanografia
Institut de Ciències del Mar

Director:

Dr. Ramon Massana i Molera
Departament de Biologia Marina i Oceanografia
Institut de Ciències del Mar

Tutor acadèmic:

Dr. César Mösso Aranda
Departament d’Enginyeria Civil i Ambiental
Universitat Politècnica de Catalunya

Institut
de Ciències
del Mar

EXCELENCIA
SEVERO
OCHOA

Barcelona, 19th September 2022

During the realization of this thesis, Aleix Obiol Plana held a PhD fellowship FPI (BES-2017-080765), funded by the Spanish Ministry of Science and Innovation and was a student of UPC's PhD program in Marine Sciences. This thesis was also supported by ALLFLAGS (CTM2016-75083-R), SINGEK (H2020-MSCA-ITN-2015-675752), Malaspina-2010 (CSD2008–00077) and DIVAS (PID2019-108457RB-I00) projects and the "Severo Ochoa Centre of Excellence" accreditation (CEX2019-000928-S).

A la meva mare,
al meu pare,
a l’Alba
i a la Mariona

“Retxes de sol atravessen blaus marins,
ses algues tornen verdes i brillen ses estrelles,
que ja s’ha fet de nit i es plàncton s’il·lumina
i canten ses balenes a 30.000 quilòmetres d’aquí.”

- Antònia Font, Batiscafo Katiuscas

Acknowledgements

Ara fa set anys que vaig decidir marxar del grup de recerca on havia estat fent el màster, on hauria pogut continuar fent un doctorat. En aquell moment, iniciava una etapa plena de dubtes, valorant si havia pres la decisió correcta i, encara més important, pensant què volia fer a partir de llavors. En aquell moment, com en tants d'altres, vaig tenir el suport incondicional de la meva família (la meva mare, el meu pare i la meva germana), i el meu primer agraïment és per vosaltres.

Dit això, moltes gràcies Ramon per haver-me donat l'oportunitat de començar a treballar a l'ICM i, sobretot, haver confiat en mi. M'has guiat en aquest món de la ciència, has valorat sempre la meva feina i el meu criteri, m'has tingut en compte per altres projectes i m'has donat llibertat per explorar i aprendre. Has fet que tots aquests anys de tesi hagin funcionat molt bé, sense gaires entrebancs, i crec que hem fet un bon equip. T'ho agraeixo sincerament, ha estat una sort tenir-te com a director.

He procurat escriure tota la tesi fent servir el plural perquè aquesta és fruit d'un treball col·lectiu. De la gent amb qui he treballat, tu Irene n'has estat una peça clau. M'has ensenyat i m'has ajudat moltíssim, treballar amb tu és fantàstic i fas que tot sigui fàcil, així que gràcies. També gràcies Vanessa, per les hores de laboratori però també per totes les hores travessant l'Atlàntic, i a la resta de l'equip POSEIDON: Caro, em va agradar tenir l'oportunitat de conèixer-te i ajudar-nos mútuament; i Isabel, gràcies per fer que tot anés bé, pel teu suport i consells. Seguim, David gràcies per respondre tots els meus dubtes i aportar una visió diferent de l'habitual, he après amb tu i ha estat una sort tenir-te al grup. Gràcies Javi per entrar sempre al despatx del Ramon com només tu ho saps fer i fer-me partícip de moments hilarants; gràcies també per aportar sempre els teus coneixements de mestre protistòleg i per aconsellar-me amb l'estada. Gràcies Eli per tota l'ajuda al laboratori quan jo no en sabia res, gràcies Jeff per confiar en mi fent recomptes de microscopi, gràcies Dolors per contribuir sempre amb la teva experiència, gràcies Pep pels comentaris que mai et deixen indiferent (i per la teva veu de baix a la coral!), gràcies Clara per facilitar-ho tot i descobrir-me el món de Blanes. Gràcies Pablo per la paciència que tens, per respondre els dubtes d'un complet ignorant primer i per aguantar totes les peticions referents a marbits després. Gràcies Marta S per sempre donar-li una volta més i donar bones idees. Gràcies Fran, Lídia, Ina, Aurélie, Ramiro, Oriol, Susanne i la resta de gent amb qui he col·laborat aquests anys.

Passem ara a la banda més personal de l'ICM, aquella gent que fa la rutina molt més divertida i alegre. El primer agraïment, com no pot ser d'una altra manera, per a l'Adrià. No m'imagino tot això sense tenir-te al costat. M'has ensenyat noves maneres de fer, m'has transmès el teu esperit crític i m'has fet reflexionar. Amb diferència, ets la persona amb qui més hores he compartit tots aquests anys i ha estat una gran sort, ens ho hem passat molt bé, ets un gran amic. Seguim amb tu, Patri, la que ho aguantes tot, la que es preocupa per tothom. Que bé compartir despatx amb tu, que divertits els moments dins i fora de l'ICM, has estat un gran suport i ets una molt bona amiga. Ja parlarem, ja. Joan, quin gran descobriment, ets autèntic, amb tu sempre s'hi està bé, m'encanta la passió com vius les coses. Al final vaig venir a Cadaqués, poc us ho pensàveu! Marina, sempre interessant-te per l'altra gent, super constant i fent les coses bé, has estat la meva gran companya en aquesta recta final, ànims que ara a tu et queda molt poc també i ja ho tens, t'ho mereixes tot! Andreu, el guardià dels *hot topics* i l'ànima dels divendres, va ser molt bonic quan ens vas explicar perquè mai no volies que marxéssim de la taula. Carlota, la meva connexió amb Txèquia per parlar de tot aquell món. Ari, pots preguntar-me tots els dubtes que tinguis de *ggplot* que sempre te'ls contestaré. Sara, super escaladora, gràcies per intentar-me arrencar repetides vegades de l'ordinador a l'hora de dinar (sense èxit majoritàriament, però la intenció és el que compta), ànims amb el català que ho estàs fent molt bé. Miguel, endavant que tu també ja ho tens, encara recordo el viatge en camió a Cartagena amb l'Ana, que divertit. Queralt, encara espero una ampolla de vidre decorada! O potser millor no, que ja sabem com acabarà... Maria, la reina dels caquis i la descobridora musical, la teva presència encara hi és, al despatx. Marta R, ets una inspiració en el disseny, m'encanta el que fas (només cal veure com he maquetat aquesta tesi). He de parar. Gràcies Deju, Manu, Guillem (*ufff*), Isa, Anna, Dani, Ari R, Claudia i a la resta de gent amb qui ens hem creuat pel camí.

De la meva estada a České Budějovice, thank you Monika for everything. I went for a short stay and came back with a friend. You made me feel at home from the moment you rescued me from the wrong institute on day one. It was super fun living with you and going with the flow (*my maaan!*). Thanks also Martin (boss), Eric (subboss) and Jeff (jefois), I am super grateful for your welcome, I felt involved in the lab from the very first week.

De l'equip de l'Autònoma, vull recordar la Isabel Esteve, que és qui em va introduir en el món de la ciència, i la passió per l'ecologia microbiana que la caracteritzava. Guardo com un tresor les classes magistrals al despatx i l'entusiasme amb què explicava les coses. Tampoc no m'oblido de tu, Laia, vas ser un suport molt necessari en un moment on jo anava molt perdut, gràcies.

Moltes gràcies també a dues persones sense les quals jo ara mateix no estaria escrivint aquests agraïments. En primer lloc, gràcies Paqui per les teves bones idees i consells, tens permís per fer-me totes les preguntes tecnològiques que vulguis. I després, moltes gràcies a l'Antonio Gómez, per ensenyar-me les fites del camí quan jo no hi veia ni un corriol. També gràcies Vie i Betta pels consells acadèmics, a tu també Vie pels moments de muntanya, i Andreu, per a tu continuaré sent l'autèntic ordinadors. I Núria, quan arribi el moment pensa't bé això de ser científica! També gràcies a la gent de l'esplai, de Sambossa i del barri en general, a tot aquest teixit associatiu tan potent que tenim.

Gràcies al grup d'amics que hem creat aquests últims anys, ha estat una gran sort acabar-nos trobant. David, don Heinz estaria orgullós de tu; Nina, ets una artista, moltes gràcies per la masterclass i els consells maquetadors, m'han salvat totalment; Marc, prou excuses amb la cama i anem a escalar ja; Laia, la teva determinació no té rival; Gerard, em decotaràs la tesi? M'heu ajudat a trobar una passió que ara mateix és un punt central a la meva vida i fer-ho al vostre costat és un regal.

I per acabar, gràcies a la meva companya d'aventures. Si ara intentés fer uns agraïments convencionals em quedaria molt curt, així que no ho faré. Només diré que tu sempre hi ets, Marion, i això és meravellós.

Tots aquests anys han estat un aprenentatge constant, soc una persona molt diferent de la que va entrar plena de nervis fa sis anys a l'ICM. Continuo tenint molts dels dubtes que tenia llavors, però també hi ha moltes altres coses que ara les veig clares. Estic molt agraït del camí que m'ha portat fins aquí. Endavant amb tot el que hagi de venir.

Poblenou, 18 de setembre de 2022

Summary

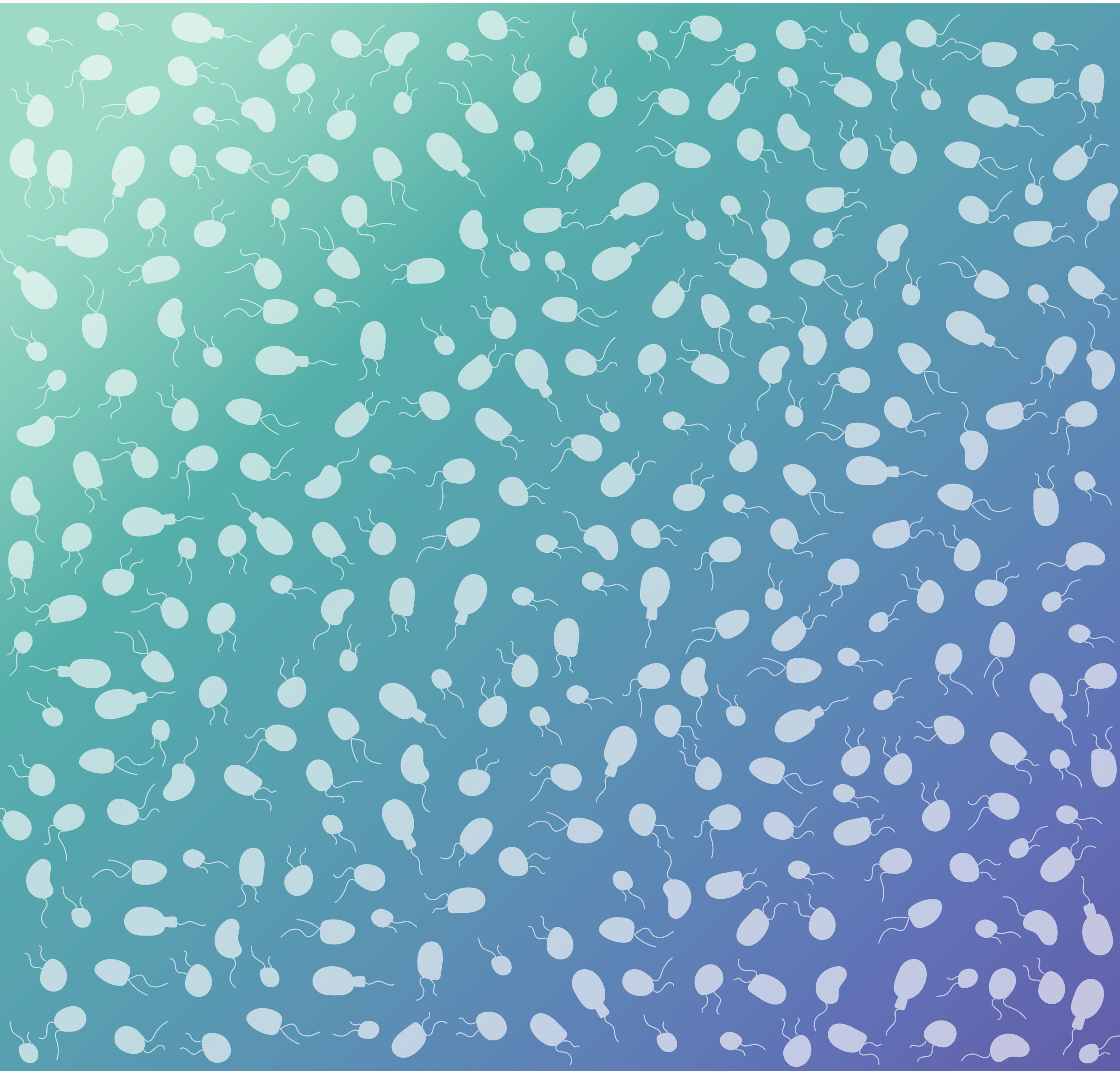
Marine heterotrophic flagellates (HF) are very small (2-5 µm) unpigmented protists that are dominant bacterial grazers in the ocean, where they link the transfer of carbon from bacterial cells to higher trophic levels. Through their bacterivorous activity, they also act as nutrient recyclers that allow for regenerated primary production, and they are partially responsible of keeping bacterial abundances in the ocean fairly constant. HFs are widespread throughout the eukaryotic tree of life, ubiquitous in the plankton and display a high functional diversity. During the last decades of the twentieth century, a growing interest in this functional group occurred, and studies were performed to characterize their ecological role. However, this initial attention diminished due to the difficulty to study natural HF species, as they possess few morphological traits for identification and generally remain uncultured. Consequently, HFs have been often neglected in marine surveys, to the extent of becoming one of the most understudied components of the marine microbiome. With the advent of high-throughput sequencing and the reduction of sequencing costs, studying these protists at a high-resolution level became feasible. This thesis represents a return to the study of HFs using these newly developed tools. We first investigated the distribution patterns of eukaryotic diversity along the water column of the ocean by metagenomics and compared the results with metabarcoding approaches. This analysis revealed a clear separation of taxonomic groups between pico- (0.2-3 µm) and nanoplanktonic (3-20 µm) fractions, as well as between photic (0-200 m) and aphotic (>200 m) regions. While some groups were not well represented by metabarcoding approaches due to technical biases, HFs were generally not affected by them. We then studied the diversity and distribution of HFs in the ocean using global metabarcoding data sets. With this, we identified a few dozens of HF species, most of them uncultured, as the dominant in surface and deep ocean regions. Many of these dominant species were present at relatively constant abundances, while others were influenced by temperature or displayed patchy distributions. Finally, we jumped from global patterns to study the gene expression of HFs in natural assemblages growing by bacterivory in unamended incubations. The obtained results using metatranscriptomics sequencing showed similar functional dynamics between experiments done at different times of a seasonal cycle, with marked differences between incubation times. Genes related to cysteine peptidases as well as some glycoside hydrolases emerged as key components involved in the process of bacterivory. Overall, this thesis returns HFs back to the spotlight and creates a solid foundation on which to perform renewed research on the ecology and functional role of this group.

Resum

Els flagel·lats heterotròfics marins (HF) són protists no pigmentats de mida molt petita (2-5 μm). Són els principals bacterívors a l'oceà i tenen un paper clau a la cadena tròfica com a enllaç en la transferència de carboni de les cèl·lules bacterianes cap a nivells tròfics superiors. A través de la seva activitat, també actuen com a remineralitzadors de nutrients, tot permetent una contínua producció primària, i són responsables parcialment de mantenir estable la concentració de bacteris a l'oceà. Els HFs es troben al llarg de tot l'arbre eucariota de la vida, tenen una gran presència en ambients planctònics i mostren una gran diversitat funcional. Durant les últimes dècades del segle XX, l'atenció cap aquest grup funcional va créixer de forma notable, i es van dur a terme estudis per caracteritzar millor el seu paper ecològic. Tanmateix, aquest interès va anar disminuint a causa de la dificultat per estudiar-ne les espècies naturals, atès que aquestes tenen pocs trets morfològics per a la seva identificació i generalment no han pogut ser cultivades. Com a conseqüència, els HFs han estat sovint oblidats en les campanyes oceanogràfiques, fins al punt de convertir-se en un dels components menys estudiats del microbioma marí. Amb el naixement de les tècniques de seqüenciació massiva i la reducció dels seus costos, l'estudi detallat d'aquests protists ha passat a ser possible. Aquesta tesi representa un retorn a l'estudi dels HF utilitzant aquestes noves eines. En primer lloc, hem investigat els patrons de distribució de la diversitat eucariota al llarg de la columna d'aigua de l'oceà per mitjà de dades de metagenòmica i els hem comparat amb els resultats obtinguts a partir de seqüències del gen ribosomal 18S (metabarcoding). Aquesta anàlisi mostra una clara separació taxonòmica entre les fraccions pico (0.2-3 μm) i nanoplanctòniques (3-20 μm), així com entre les regions fòtiques (0-200 m) i afòtiques (>200 m). Tot i que alguns dels grups no estan ben representats en l'anàlisi feta a través del gen 18S a causa de biaxos tècnics, la majoria dels HF no es veuen afectats per aquests. A continuació, hem estudiat la diversitat i distribució dels HFs a l'oceà utilitzant conjunts de dades globals de metabarcoding. Amb això, hem identificat unes poques espècies d'HFs, majoritàriament no cultivades, com les dominants en aigües superficials i profundes de l'oceà. Moltes d'aquestes espècies dominants es troben presents en abundàncies relativament constants, mentre que d'altres estan influenciades per la temperatura o mostren distribucions irregulars. Finalment, hem passat de l'anàlisi de patrons globals a l'estudi de l'expressió gènica de comunitats naturals d'HFs mitjançant una sèrie d'incubacions on la bacterivoria ha estat estimulada. Els resultats obtinguts fent servir dades de metatranscriptòmica mostren una dinàmica funcional similar entre experiments duts a terme en diferents estacions de l'any, amb diferències marcades entre els temps d'incubació. Els gens relacionats amb les peptidases de cisteïna, així com algunes hidrolases glicosídiques es presenten com a components clau implicats en el procés de bacterivoria. Amb tot, amb aquesta tesi hem tornat a posar els HFs al centre d'atenció i hem creat una base sòlida sobre la qual realitzar una investigació renovada de l'ecologia i el paper funcional d'aquest grup.

Contents

INTRODUCTION



Introduction

The marine microbiome and the carbon cycle

The marine environment is the largest ecosystem on Earth, covering more than 70% of its surface with an average water column of 3682 m and containing 97% of the world's water (Charette & Smith, 2010). Although not visible to the naked eye, life in the ocean is eminently microbial. Microbes, namely microbial eukaryotes (protists) and prokaryotes (bacteria and archaea), represent approximately two-thirds of the total biomass of marine organisms (Bar-On & Milo, 2019) and are the engines driving the biogeochemical cycles of the planet (Falkowski et al., 2008). In fact, around half of the total global carbon fixation on Earth occurs in the ocean, being performed by cyanobacteria and microalgae (Falkowski, 2012), which form the base of ocean food chains (Worden et al., 2015). Furthermore, biological nitrogen fixation in the ocean is solely performed by a variety of prokaryotes (Zehr & Kudela, 2011).

The current view of the importance of microorganisms in the ocean is less than 50 years old (Williams & Ducklow, 2019). Marine microbiology started as a research field back in the late 19th century, when scientists from other biology-related fields began to participate in oceanographic expeditions (Zobell, 1946). During 1930s, the first studies on the role of bacteria in the ocean started, with different approaches between Western (mainly United States) and Eastern (mainly Russia) laboratories. While the former relied on plate counts to quantify the abundance of bacteria in the ocean, the latter implemented direct counts using microscopy. Thus, Eastern scientists obtained more accurate bacterial abundances, up to 3 orders of magnitude higher than Western scientists. Despite having a more accurate view of the importance of bacteria in the sea (Sorokin, 1978), their work had little impact on researchers in the West. This, coupled with the latter being reluctant to trust microscopic evidence, hindered earlier conceptual advances in microbial interactions in the sea. It was not until the mid 1970s that Pomeroy (1974) created the conceptual model on which today's marine microbiology is based, and his ideas finally matured in the early 1980s, when Azam et al. (1983) coined the term 'microbial loop'.

The microbial loop concept starts with the active uptake of dissolved organic matter (DOM) by heterotrophic bacteria in the plankton (**Fig. 1**). With its incorporation into bacterial biomass, DOM becomes available to higher trophic levels after being consumed by heterotrophic flagellates, which in turn are ingested by larger protists and microzooplankton (**Fig. 1**). Today we know that a significant fraction of this bacterial biomass is consumed by viral lysis (Fuhrman & Noble, 1995) or by pigmented protists (Zubkov & Tarran, 2008). Among other sources, DOM is released to the environment through microbial lysis by viruses, sloppy feeding and phytoplankton exudation (Lønborg et al., 2020). Phytoplankton cells, at the base of the marine food web, grow in the sunlit part of the ocean, also known as epipelagic zone (0-200m). Some of the organic matter produced there can escape from predators and the microbial loop and sink into the dark ocean. In the mesopelagic zone (200-1000m), most of this organic carbon is consumed by bacteria and can return to the atmosphere as carbon dioxide in a timespan of

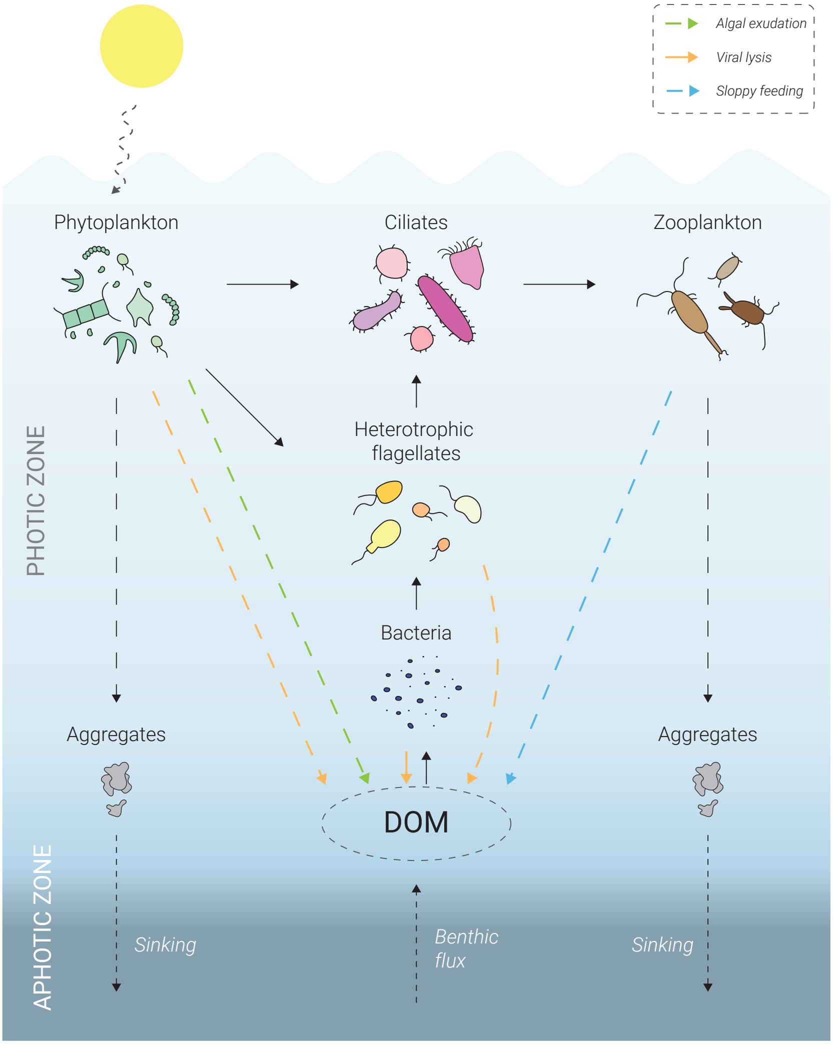


Figure 1. The marine microbial food web. A schematic depiction of the trophic interactions occurring in marine pelagic ecosystems. Phytoplankton form the base of the food web as primary producers. Fixed nutrients are transformed to dissolved organic matter (DOM) through various mechanisms and enter the food chain via the microbial loop, with its incorporation into bacterial biomass. Heterotrophic flagellates feed on bacteria and channel carbon up to higher trophic levels (i.e., ciliates and zooplankton). Part of the fixed carbon can be exported to the deep ocean through particle sinking from phytoplankton cells or fecal pellets from zooplankton in a process known as the biological carbon pump.

months to years (Arístegui et al., 2009). Nevertheless, the carbon that circumvents microbial processing in the mesopelagic zone is exported to the bathypelagic region (1000-4000m) or even the seafloor sediment, where it can be sequestered for centuries (Arístegui et al., 2009) in a process known as the biological carbon pump (**Fig. 1**).

The omics era

The huge difference in numbers between bacterial counts coming from culturing and those from direct microcopy was named the "Great Plate Count Anomaly" (Staley & Konopka, 1985). In fact, it was proposed that only around 0.1-1% of bacterial and protist organisms can be easily cultured (Amann et al., 1995; Caron et al., 1989), and these often do not represent the abundant taxa in the ocean (del Campo et al., 2013a; Pedrós-Alió, 2006). Molecular techniques such as amplification and sequencing of the small subunit of the ribosomal RNA gene (18S rDNA in eukaryotes and 16S rDNA in prokaryotes) led to the discovery of extremely relevant taxa in the oceans (Acinas et al., 2022), such as *Pelagibacter ubique* (Giovannoni et al., 1990), which is the most abundant bacterium in the ocean, marine archaea (DeLong, 1992; Fuhrman et al., 1992) or the abundant and widespread marine stramenopiles (MAST) (Massana et al., 2004). The development of high-throughput sequencing (HTS) platforms revolutionized the field of microbial ecology, as they dramatically increased the sequencing depth while keeping low costs (Goodwin et al., 2016). With this, amplification and sequencing of the 18S/16S rRNA gene (known as metabarcoding or amplicon sequencing; **Fig. 2**) became the standard to characterize the diversity of microbial communities in marine environments and to allow for community comparisons (Burki et al., 2021). The advent of HTS also consolidated omics approaches, first implemented more than 20 years ago (Béjà, 2000), as a way to study the whole gene repertoire of the community rather than focusing on specific marker genes. Omics techniques based on nucleic acids are metagenomics, metatranscriptomics and single cell genomics (**Fig. 2**). Metagenomics aims to sequence the DNA content of a microbial community (all genomes), while metatranscriptomics does the same but targeting messenger RNA (all genes being expressed). The main advantage of these techniques is that they do not rely on PCR amplification, so preventing putative amplification biases (Acinas et al., 2005) and taxonomic biases derived from the used primers (McNichol et al., 2021). Finally, single cell genomics is based on isolating single cells, and then amplifying and sequencing their genomes (Richards et al., 2019) (**Fig. 2**). Other omics techniques are metaproteomics, which is based on the quantification and identification of the proteins from a microbial community (Kleiner, 2019), and metabolomics, where metabolites – the substrates, intermediates and products of cell metabolism – are targeted (Fiehn, 2002).

The first global effort to characterize microbial communities in the surface ocean using an omics approach was the Global Ocean Sampling Expedition (GOS) (Venter, 2004), but this did not use HTS tools. Later, more expeditions were developed to circumnavigate the globe while sampling water from surface to the deep ocean, such as Tara Oceans (Karsenti et al., 2011) or Malaspina 2010 (Duarte, 2015) initiatives, just to name two of them (see Acinas et al. (2022) for a review). Apart from global environmental surveys, other collaborative efforts were carried out to sequence microbial diversity, such as the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP; Keeling et al. (2014)), aiming to obtain transcriptomes of cultured microbial eukaryotes. The beauty of science is that all these initia-

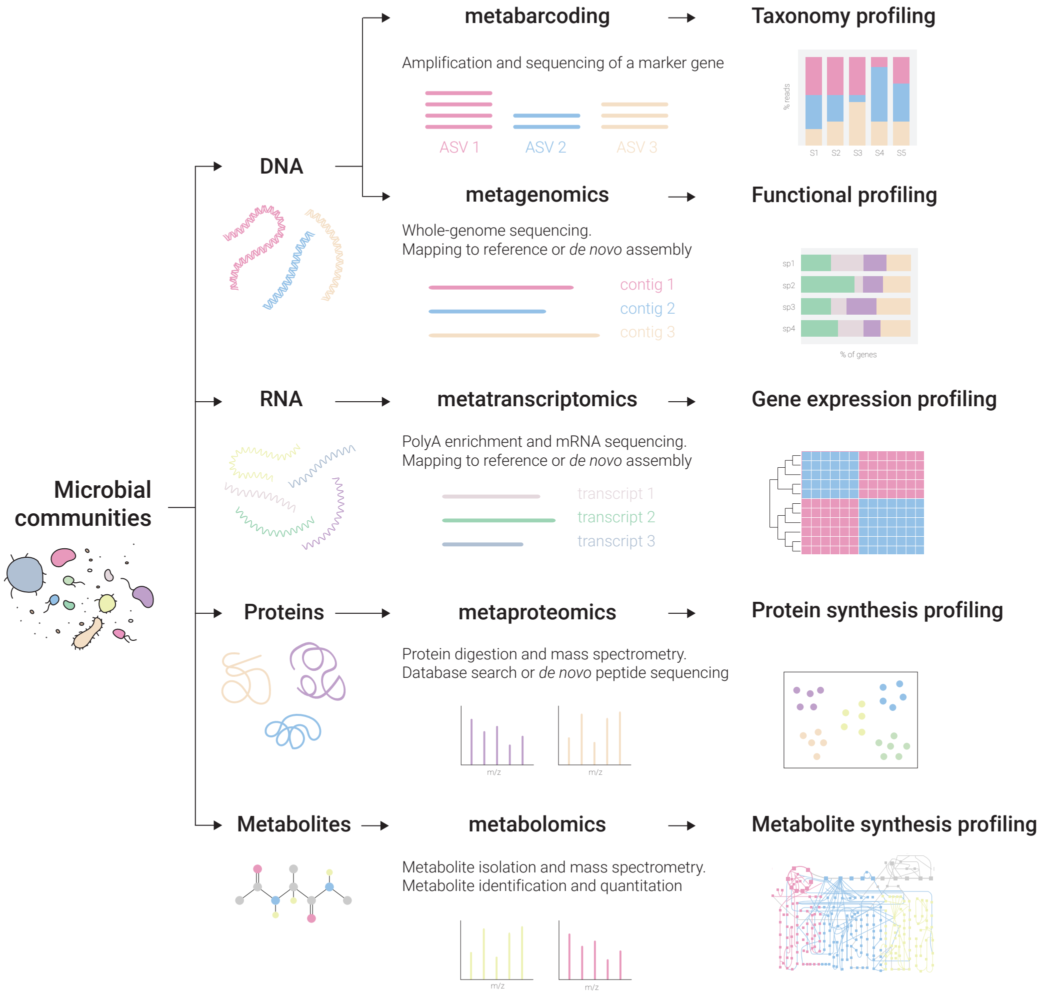

Microbial communities
DNA
metabarcoding
Taxonomy profiling
Amplification and sequencing of a marker gene
ASV 1
ASV 2
ASV 3
% reads
S1
S2
S3
S4
S5
metagenomics
Functional profiling
Whole-genome sequencing.
Mapping to reference or *de novo* assembly
contig 1
contig 2
contig 3
sp1
sp2
sp3
sp4
% of genes
RNA
metatranscriptomics
Gene expression profiling
PolyA enrichment and mRNA sequencing.
Mapping to reference or *de novo* assembly
transcript 1
transcript 2
transcript 3
Proteins
metaproteomics
Protein synthesis profiling
Protein digestion and mass spectrometry.
Database search or *de novo* peptide sequencing
m/z
m/z
Metabolites
metabolomics
Metabolite synthesis profiling
Metabolite isolation and mass spectrometry.
Metabolite identification and quantitation
m/z
m/z

Figure 2. Overview of different omics techniques. Various omics approaches can be performed to study microbial communities. These depend on the starting biological material (from DNA to metabolites) and use different processing and analytical techniques. Although not an omics approach, metabarcoding is also included here due to its wide use in microbial ecology studies. Figure adapted from Zuñiga et al. (2017).

tives made (or should make) all their sequencing data publicly available, thus representing a precious data resource that can be used by other researchers to further study microbial communities in the ocean.

Microbial eukaryotes

The eukaryotic cell

Contrary to prokaryotic cells, eukaryotic cells have a membrane-bound nucleus, a highly compartmentalized cytoplasm containing membrane-bound organelles (such as the mitochondrion or chloroplast) and a cytoskeleton that shapes the cell and can form complex structures (Massana & Logares, 2013). While prokaryotes display a high diversity in their metabolism, microbial unicellular eukaryotes (or protists) have a wide range of morphologies, cell sizes and behaviour (Keeling & del Campo, 2017). In fact, protists' size range spans more than 5 orders of magnitude (Caron et al., 2012), going from less than 1 μm to more than 1 cm. According to their size, planktonic protists are commonly divided into pico- (0.2-3 μm), nano- (3-20 μm) and microplankton (20-200 μm). Protists can also be divided by their nutritional modes into heterotrophic (predatory, osmotrophic or parasitic), phototrophic and mixotrophic (organisms capable of both heterotrophy and phototrophy) (Bachy et al., 2022).

The process by which the eukaryotic cell emerged – eukaryogenesis – still represents a key question in evolutionary biology. To date, the most accepted scenario is that eukaryotes derive from a host cell related to Asgard archaea (Eme et al., 2017) that incorporated a bacterium closely related to alphaproteobacterial lineages, which after several changes became the mitochondrion (Gabaldón, 2021). Later, the incorporation of a cyanobacterium by a heterotrophic eukaryotic host gave rise to the first photosynthetic plastid in a process called primary endosymbiosis (López-García et al., 2017). Virtually all the eukaryotic plastids derive from this endosymbiotic event, with the only known exception being *Paulinella chromatophora* (Marin et al., 2005). In the subsequent eukaryotic evolution there have been multiple secondary endosymbiotic events (Worden et al., 2015) in which a heterotrophic eukaryotic host engulfed a photosynthetic eukaryote. This originated secondary plastids, which can be found in many lineages of the eukaryotic tree of life.

A brief history of protistology

Protists were first described nearly 350 years ago by Antonie van Leeuwenhoek, arguably the first microbiologist in history, using a hand-made microscope (Leeuwenhoek, 1677). The complexity of protist shapes inspired Ernst Haeckel's world famous illustrations (Haeckel, 1887), which, together with his remarkable studies, had a big impact on both the general public and the scientific community. However, it was Victor Hensen, a contemporary of Haeckel, the first to treat protists as key players in the base of marine food webs (Smetacek, 1999), and the one who coined the term "plankton". His works were never translated to English (contrary to Haeckel's), and his findings were not embraced by other scientists. After Leeuwenhoek's first observations, researchers started classifying protists into different taxa based on morphology and trophic strategies, placing photosynthetic ones close to plants ("algae") and heterotrophic ones near animals ("protozoa"). Nevertheless, the myriad of complex morphologies displayed by protists together with inconsistent descriptors to resolve the relationship between the different taxonomic groupings led to the placement of protists in a kingdom of 'lower' organisms below animals, plants and fungi (Bachy et al., 2022). This separation was challenged by the use of electron microscopy to study the ultrastructure of eukaryotes (Corliss, 1984), a work started by Irene Manton in phototrophic protists and later continued by Dorothy Pitelka for heterotrophic protists. With this new technique, researchers could identify cell components that were essentially the same between protists and multicellular organisms, such as mitochondrion or chloroplasts (Taylor, 2003). The first molecular studies using 18S rRNA as a marker gene started to solve critical evolutionary relationships (Taylor, 2003), and started to put together, in eukaryotic "supergroups, groups that apparently had no relation. It was the advent of phylogenomics that finally brought a clearer picture of the tree of eukaryotes, and allowed assembling several groups previously described by ultrastructure into a small number of 'supergroups' (Baldauf, 2003). Some of these "supergroups" were solely formed by unicellular microorganisms and others contained both multicellular organisms and microbial lineages. Thus, the separation between "higher" and "lower" eukaryotes was abandoned for good.

The eukaryotic tree of life

Today we know that most of the phylogenetic diversity within the domain Eukarya is constituted by microbial forms (Caron et al., 2012) and that multicellular organisms (such as animals and plants) are present in a few branches of the eukaryotic tree of life (Burki et al., 2020). The phylogenetic diversity of eukaryotes can be divided into major lineages, often referred to as "supergroups" (Burki et al., 2020). Currently, these are divided into TSAR, Haptista, Cryptista, Archaeplastida, Amorphea, CRuMs, Discoba, Metamonada, Hemimastigophora and orphan taxa (**Fig. 3**).

TSAR constitutes a mega-assemblage that groups Telonemia, Stramenopiles, Alveolata and Rhizaria (Strassert et al., 2019). Telonemia is a supergroup of flagellated organisms from which only 7 described species exist (Shalchian-Tabrizi et al., 2006; Tikhonenkov et al., 2022). Telonemia represents the sister group to the SAR clade, originally described by phylogenomic approaches 15 years ago (Burki et al., 2007) and comprising the highest diversity within protists (del Campo et al., 2014).

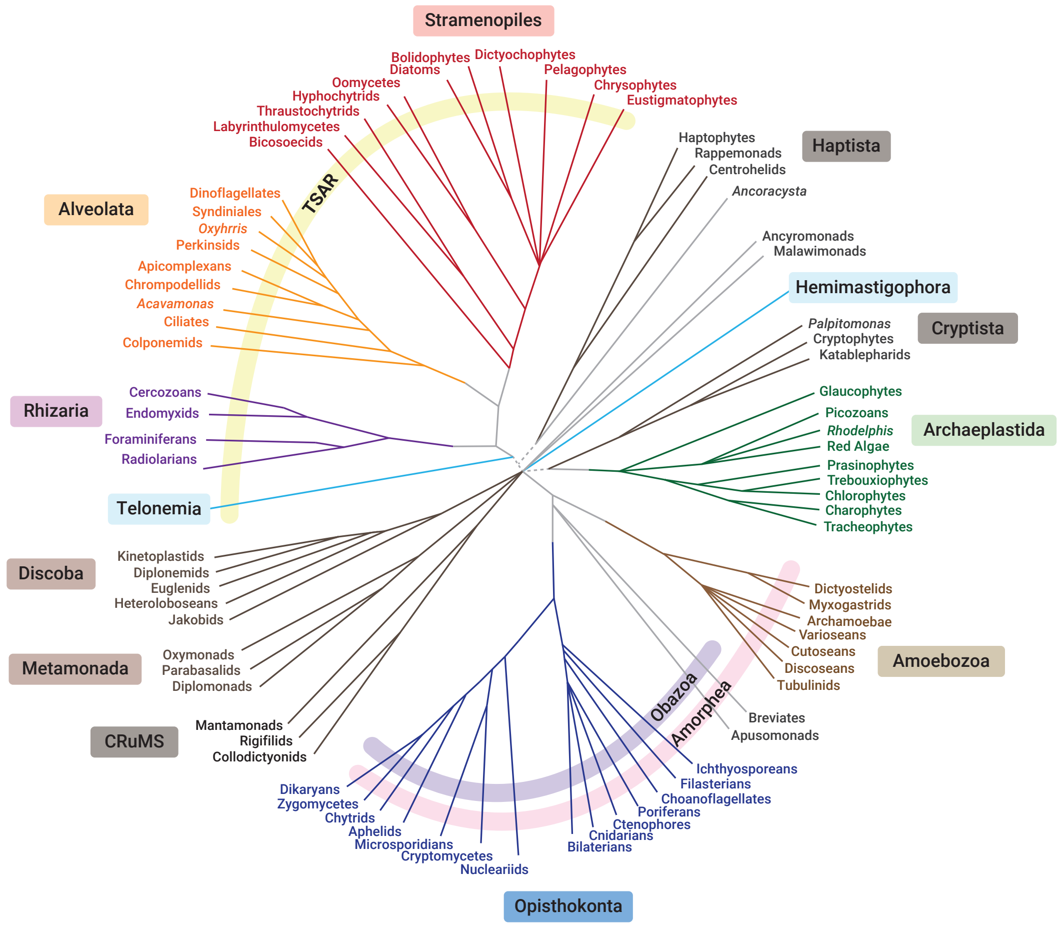


Figure 3. The eukaryotic tree of life. A schematic tree of eukaryotes based on a consensus of phylogenomic studies together with morphological and cell biological information. Figure adapted from Keeling & Burki (2019) and updated with information from Burki et al. (2020) and Schön et al. (2021).

Stramenopiles are formed by a large radiation of chloroplast-containing groups (placed in Ochrophyta) together with basal groups of heterotrophic protists (Massana et al., 2014). Most Stramenopiles, also known as heterokonts, have 2 flagella of unequal length, with the longer one having hair-like structures called mastigonemes (Adl et al., 2019). The Ochrophyta contains relevant groups such as the well-known diatoms, responsible for 40-45% of total oceanic primary production (Mann, 1999); the chrysophytes, pelagophytes, bolidophytes and dictyochophytes, abundant phototrophic picoeukaryotes in the surface ocean (Massana, 2011), some of which also contain mixotrophic species (Stoecker et al., 2017); and the giant kelp, which can be one of the largest organisms on Earth (Caron et al., 2012). Ochrophyta plastids, originated by secondary or even tertiary endosymbiosis with red algae (Azuma et al., 2022), are bound by four membranes, with the outermost one being fused with the endoplasmic reticulum and nuclear membranes (Andersen, 2004). In the heterotrophic part of Stramenopiles radiation, there are pseudofungi groups such as labyrinthulomycetes, saprotrophic organisms highly present in bathypelagic marine snow (Bochdansky et al., 2017); and oomycetes, like the notorious *Phytophthora*, the causative agent of the potato blight that led Ireland into the Great Famine during the mid-nineteenth century (Bachy et al., 2022). Other relevant heterotrophic flagellates are bicosoecids, generally associated to particles (Schoenle et al., 2020) and common in the benthos (Caron et al., 2012); *Blastocystis*, a potentially mutualistic parasite in the human gut microbiome (del Campo et al., 2020); and many MAST groups (Massana et al., 2014), which were first described in environmental surveys and still remain largely uncultured, with only a few exceptions (Cavalier-Smith & Scoble, 2013; Shiratori et al., 2017). Although generally considered as heterotrophic, most of MAST functional roles are yet to be characterized. In the case of MAST-3 (Nanomonadea), parasitic and bacterivorous modes have been described for *Solenicola setigera* (Gómez et al., 2011) and *Incisomonas marina* (Cavalier-Smith & Scoble, 2013), respectively. For MAST-4, MAST-1 and MAST-7, there is microscopic evidence of their phagotrophic activity (Massana et al., 2009; Rodríguez-Martínez et al., 2022).

Alveolata contains three main groups, namely ciliates, dinoflagellates and apicomplexans. Their main characteristic is the presence of alveoli, membrane-bound cavities, in the inner part of the plasma membrane (Caron et al., 2012). Ciliates are a major group of predators and grazers and represent a significant trophic link between heterotrophic flagellates and zooplankton (Sherr & Sherr, 2002). Dinoflagellates display a high diversity of species and trophic lifestyles, from autotrophs and mixotrophs to grazers and parasites, and they can host intracellular symbionts or be endosymbionts themselves (Taylor et al., 2008). They are one of the most abundant groups of large microalgae in the plankton, with a pivotal role in primary production and famous for the production of algal blooms (Simpson & Eglit, 2016). Predatory dinoflagellates are important grazers in the ocean, preying from bacteria and picoeukaryotes to diatoms, ciliates and even some metazoans (Jeong et al., 2010). Dinoflagellates also comprise the marine alveolates (MALV, also known as Syndiniales), first discovered by environmental surveys (López-García et al., 2001) and very abundant in the ocean where they act mostly as parasites (Guillou et al., 2008). Apicomplexans are mainly parasites of a wide range of animals (del Campo et al., 2019), such as fish or even humans, as the infamous *Plasmodium*, the agent of malaria. The origin of plastids in alveolates mostly derives from secondary or higher-order endosymbiotic events, with the exception of a few lineages that incorporate them through kleptoplasty (Stoecker et al., 2009).

Rhizaria includes a wide range of mostly heterotrophic amoeboid protists that have pseudopodia, which they use mainly for feeding (Adl et al., 2019), and flagellated forms (Burki

& Keeling, 2014). They can be divided into foraminifera, radiolarians and cercozoans. The former 2 groups produce the skeletal structures that can pictured in the above-mentioned Haeckel's illustrations (Caron et al., 2012) and are also important in paleobiology (Burki & Keeling, 2014). Cercozoans include several groups of cercomonad flagellates, some apparent amoebas, and the photosynthetic algae chlorarachniophytes, the plastids of which are from secondary endosymbiotic origin with green algae. The only heterotrophic member of the chlorarachniophytes is *Minorisa minuta,* a tiny heterotrophic flagellate abundant in coastal waters (del Campo et al., 2013b). A characteristic trait of phototrophic chlorarachniophytes (and cryptophytes, see below) is the presence of a nucleomorph, which is a reduced version of the endosymbiont nucleus (Curtis et al., 2012).

Haptista contains 2 main lineages: haptophytes and centrohelids. Haptophytes are phototrophic microorganisms that can form high-density blooms. One of its most famous members are the coccolithophores (e.g., *Emiliania huxleyi*), cells covered by calcium carbonate plates that have a unique role in biogeochemical cycles (Bachy et al., 2022) and their microfossils represent important chalk deposits (as in the Cliffs of Dover). Non-calcifying haptophytes are also important constituents of phytoplankton, and some its species can be key bacterial grazers (Unrein et al., 2014). Centrohelids have a 'heliozoan' amoeboid shape with axopodia radiating from a spherical cell body (Burki et al., 2020), with either silica scales or organic spicules (Cavalier-Smith & Chao, 2012). There are only 2 described centrohelid species with naked cells, belonging to the genus *Oxnerella (*Cavalier-Smith & Chao, 2012).

Cryptista are formed by cryptophytes, katablepharids and the single genus *Palpitomonas*. Cryptophytes have a secondary plastid and a nucleomorph and they are of particular interest for the study of their evolution (Sibbald & Archibald, 2020). Katablepharids are enigmatic heterotrophic protists that are being targeted for the study of plastid acquisition (Okamoto et al., 2009).

Archaeplastida are mostly phototrophic cells with plastids coming from a primary endosymbiotic event. They are divided into green algae (from which land plants evolved), red algae and glaucophytes. No heterotrophic component was described within Archaeplastida until the recent description of *Rhodelphis*, a sister group to red algae formed by phagotrophic species containing non-photosynthetic plastids (Gawryluk et al., 2019). Additionally, a recent phylogenomic study placed Picozoa (Not et al., 2007), a group of heterotrophic flagellates without plastid formerly treated as orphan taxa, as a second closely-related lineage to red algae, thus challenging the current view of plastid evolution (Schön et al., 2021).

Amorphea groups opisthokonts and amoebozoans together, as well as two small lineages of heterotrophic flagellates, apusomonads and breviates. The latter two lineages cluster together with opisthokonts to form the Obazoa (Brown et al., 2013).

Opisthokonta are divided into two main lineages: Nucletmycea (or Holomycota), which includes fungi (Grossart et al., 2019) and their unicellular relatives; and the Holozoa, containing animals (Metazoa) and their unicellular relatives (Torruella et al., 2012). One of these unicellular holozoans are choanoflagellates, bacterivorous protists that display a characteristic microvillar feeding collar (Simpson & Eglit, 2016) and represent the closest relatives to animals. Thus, they are a key lineage in the study of the transition to multicellularity (Sebé-Pedrós et al., 2017). Other closely-related taxa being targeted in this quest are filastereans and ichtyosporeans (Sebé-Pedrós et al., 2017), which include several fish parasites.

Amoebozoa, as the name suggests, contains amoeboid protists. These present lobose pseudopodia, or lamellipodia (used for feeding and locomotion) that are much broader than the pseudopodia present in Rhizaria (Simpson & Eglit, 2016). They also include flagellated forms, as well as human parasites such as *Entamoeba histolytica* (causative agent of amoebic disentery). Amoebozoa are common in marine sediments, from which several species have been isolated (Kudryavtsev et al., 2018; Kudryavtsev & Pawlowski, 2013, 2015), and known planktonic forms are mostly associated to particles (Kudryavtsev et al., 2022; Rogerson et al., 2003).

CRuMs is a grouping of former orphan taxa that were recently found to branch together (Brown et al., 2018). It includes Collodictyonids, Rigifilida and the genus *Mantamonas*, all free-living protists with different morphologies.

Discoba and Metamonada were formerly treated as one supergroup named Excavata. However, no strong evidence exists that supports this grouping (Burki et al., 2020). Discoba includes abundant and diverse heterotrophic flagellates such as diplonemids (Flegontova et al., 2016) and kinetoplastids (Flegontova et al., 2018), as well as the photosynthetic euglenids. Metamonada is formed by anaerobic protists including several pathogens like *Giardia* and *Trichomonas* and symbionts in animal guts (Bachy et al., 2022).

Hemimastigophora represents one of the deepest-branching lineages in eukaryotes, containing free-living protists from soils, and was recently proposed as a new supergroup (Lax et al., 2018).

Orphan taxa groups organisms and lineages with an unclear taxonomical position in the eukaryotic tree. These include *Ancoracysta*, malawimonads and ancyromonads, all of which are free-living heterotrophic flagellates (Burki et al., 2020).

The study of protists: limitations and advances

Accessing reference genomes for protists is key to investigate both the ecology and the evolution of these taxa. However, databases of eukaryotic reference genomes currently have three critical biases. First, the big majority of the sequenced genomes belong to multicellular organisms (animals, fungi and plants). Second, genomes from phototrophic (easier to culture and maintain), parasitic or economically important microbial species greatly outnumber those from heterotrophic protists (del Campo et al., 2014) (**Fig. 4**). And third, a high fraction of protist diversity is not available in culture (del Campo et al., 2013a). Apart from these biases, another problem arises for already available genomes, and that is that a large number of the predicted genes are completely unknown. In fact, it is estimated that around 40-60% of genes from microbial systems completely lack a functional annotation, and efforts are being made to bridge this knowledge gap (Vanni et al., 2022).

Advances in different omics fields, as well as new innovative culturing efforts, currently represent powerful tools to tackle the above-mentioned problems. In the field of genomics, the Earth BioGenome Project (Blaxter et al., 2022) now aims to sequence, catalog, and characterize the genomes of all of Earth's eukaryotic biodiversity. Meanwhile, metagenome-assembled genomes, formerly used only in prokaryotes due to their simpler genomes (Massana & López-Escardó, 2022), are now being obtained en masse for eukaryotes (Alexander et al., 2022; Delmont et al., 2022; Duncan et al., 2022) thanks to the reanalysis of published global data sets. Single-cell genomics is also allowing for obtaining genomes from lineages that refuse culturing (Gawryluk et al., 2016; Labarre et al., 2020; Latorre et al., 2021; Seeleuthner

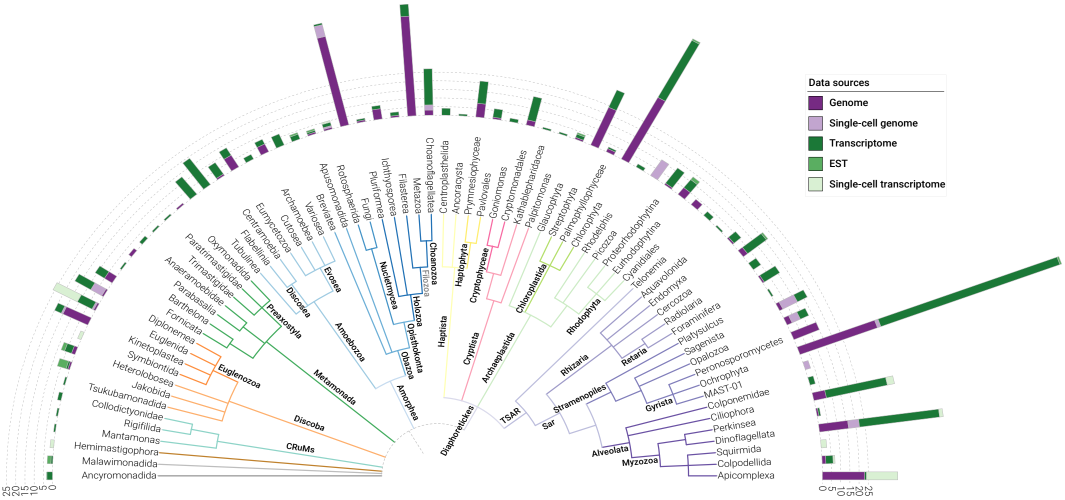


Figure 4. The EukProt database. Schematic eukaryotic tree of life displaying the distribution of the 993 genomic data sets and their sources available in EukProt v3. Figure extracted from Richter et al. (2022).

et al., 2018), and in the case of Picozoa helping to find their place in the eukaryotic tree of life (Schön et al., 2021). Metatranscriptomics and transcriptomics represent powerful tools for phylogenomic analyses (Keeling et al., 2014; Tice et al., 2021) and to assess the ecological functional roles of eukaryotic taxa (Carradec et al., 2018; Lambert et al., 2022). Additionally, metabarcoding surveys keep being useful to uncover the huge diversity of uncultured protists and describe their biogeographical patterns (de Vargas et al., 2015; Giner et al., 2020; Pernice et al., 2016), and the obtained public data sets represent a vast wealth of information for further exploration (Vaulot et al., 2022).

Marine heterotrophic flagellates

Heterotrophic flagellates (HF; **Fig. 5**) are small unpigmented protists that are dominant bacterial grazers in the ocean, therefore having pivotal roles in global biogeochemical cycles. First, given their small size (2-5 µm), they represent a link between bacteria and higher trophic levels, thus maintaining the carbon transfer in marine food webs (as highlighted by Azam et al. (1983)); second, they recycle inorganic nutrients such as phosphorus and nitrogen that can then be used again for regenerated primary production (Goldman & Caron, 1985; Sherr & Sherr, 2002); and third, they keep bacterial abundances in the ocean fairly constant (Pernthaler, 2005), a role that share with viruses (Fuhrman & Noble, 1995) and mixotrophic protists (Zubkov & Tarran, 2008). HFs are distributed in the plankton at concentrations 10^2-10^4 cells ml^{-1} and represent around 20% of all microbial eukaryotes in the epipelagic zone of the ocean (Jürgens & Massana, 2008). Collectively, HFs form an extremely taxonomically diverse assemblage, with species scattered through all the major eukaryotic supergroups (see previous section). Contrary to larger protists (>20 µm) like dinoflagellates and ciliates, most HFs cannot be easily discriminated through light microscopy, as they generally display a simple morphology without conspicuous morphological traits. That is, a round or oval cell with a single nucleus and 1-2 flagella (**Fig. 5**). Consequently, and despite their huge diversity, HFs have been historically treated as a single functional group that behaves similarly to certain environmental factors (Sanders et al., 1992).

The study of HFs through history

The first studies evaluating the role of HFs in seawater date back from the beginning of the twentieth century. These attempted to quantify HFs in the sea (Lohmann, 1911) as well as isolate and describe some of its forms (Griessmann, 1914). HFs were nevertheless mostly ignored by ecologists until the realization that bacterial abundances in the ocean remained fairly constant despite high prokaryotic production rates (Fenchel, 1986). Following incoming evidence of HF being able to predate on bacteria (Haas & Webb, 1979; Sorokin, 1979), it was thus proposed that HFs could be responsible for maintaining these high bacterial numbers

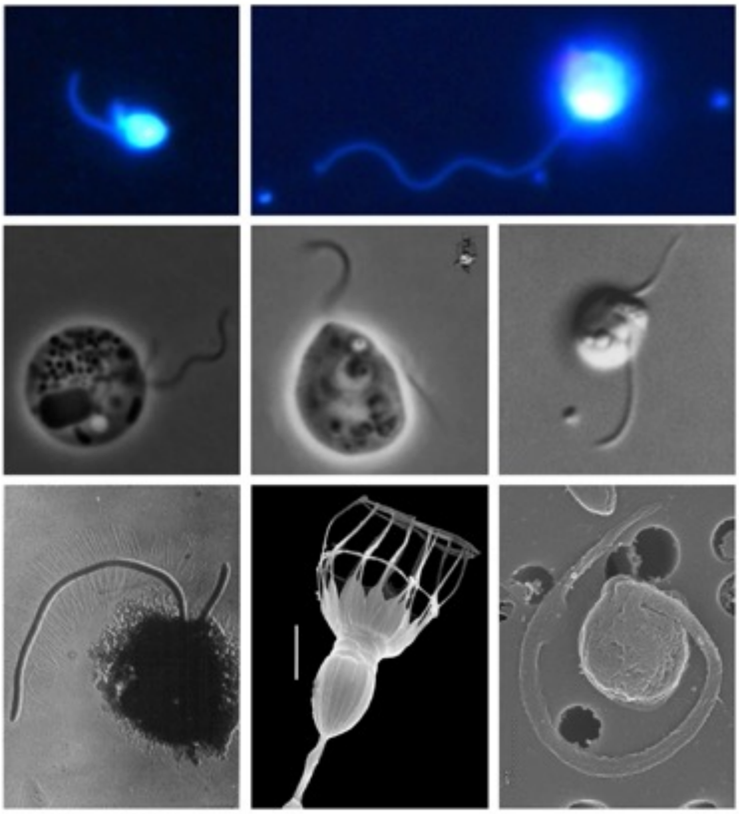


Figure 5. Heterotrophic flagellates under the microscope. HFs (2-5 µm) as seen by epifluorescence (upper images), phase contrast (middle), and electronic microscopy (lower). Micrographs extracted from Massana (2020).

in check, and that they could represent the "missing link" between bacteria and larger suspension feeders (such as ciliates and zooplankton) in pelagic food webs (Fenchel, 1982a-d). This last idea crystallized in the work by Azam et al. (1983) where the microbial loop concept was described, and represented a turning point for the consideration of the ecological significance of HFs in the ecosystem. As a result, studying HFs gained some popularity, subsequent works studied the importance of HF predation in food webs (Boenigk et al., 2001; Boenigk & Arndt, 2002; Jürgens & DeMott, 1995; Jürgens & Matz, 2002; Pernthaler, 2005; Sherr & Sherr, 2002), and some available HF strains were studied in detail regarding their autecology (Caron et al., 1986, 1990; Eccleston-Parry & Leadbeater, 1994; Geider & Leadbeater, 1988; Goldman & Caron, 1985). Collectively, these studies showed a high functional diversity (see next section) within HFs (Arndt et al., 2000) and established their key role in the ecosystem (Jürgens & Massana, 2008). In the early 2000s, the use of molecular techniques revolutionized again the view of HFs. Environmental surveys using metabarcoding identified a massive new diversity of HFs (Díez et al., 2001; López-García et al., 2001; Moon-Van Der Staay et al., 2001), and new taxonomic groups were described, such as MAST (Massana et al., 2004) or Picozoa (Not et al., 2007). Besides this enormous diversity, the use of FISH gave a new dimension to microscopy-based studies, and it allowed retrieving ecological information on *in situ* abundance, feeding modes, grazing preferences, and growth rates of specific lineages (Piwosz et al., 2021). However, after this initial peak of popularity, interest in HFs began to decrease due to the difficulty to study natural HF species. These possess few conspicuous morphological traits to identify them through microscopy and most of them remain uncultured (Weber et al., 2017). As mentioned in the previous section, this last problem is shared with the majority of protists, but the fact that HFs need a specific prey to grow makes culturing them even more complicated. HFs were therefore often neglected in marine surveys and became one of the most understudied components of the marine microbiome (del Campo et al., 2014). Finally, the advent of HTS and the reduction of sequencing costs (Goodwin et al., 2016) opened up new possibilities in the study of HFs and assessing these protists at a high-resolution level became feasible.

Functional diversity within HFs

HF assemblages include cells with different evolutionary histories, feeding mechanisms, optimal prey spectra, and behavior (Jürgens & Massana, 2008). Regarding their modes of movement, HF forms can be gliding or free-swimming, as well as be temporarily or permanently attached to a substrate (Fenchel, 1987). In terms of feeding modes, these comprise filter-feeding (e.g. choanoflagellates), direct interception feeding (e.g. chrysophytes) or raptorial feeding (e.g. kinetoplastids) (Jeuck & Arndt, 2013). One of the key factors in HF prey selectivity is bacterial size (Chrzanowski & Šimek, 1990), and niche differentiation of HF species occurs even in narrow prey size ranges, with species showing different size-dependent feeding efficiency curves (Jürgens & Matz, 2002). This existing size-selective grazing has implications in bacterial populations, such as shifting their size structure towards smaller or larger cell sizes (Pernthaler, 2005) as well as altering the behavioral traits of bacterial communities, which can develop grazing-resistant morphologies like cell aggregates or filamentous forms (Hahn, 2002). It has also been demonstrated that evolutionarily closely-related and morphologically similar HF species display different prey preferences that can differentially modify the abundance and composition of bacterial communities (Glücksman et al., 2010). At the same time, shifts in bacterial community structure can impact the structure of HF assemblages (Šimek et al., 2013), as different bacterial prey can have different nutritional values to predators (Šimek et al., 2018), and even the same bacterial strain may not have the same nutritional quality for different HF species (Chrzanowski & Foster, 2014). Besides different feeding behaviors, HFs also display different responses to abiotic factors, such as salinity or temperature, with some species displaying wide ranges of tolerance and others being restricted to narrower gradients (Arndt et al., 2000). As it happens with bacterial prey, HFs can also exhibit morphological and behavioral strategies to avoid predation by larger heterotrophic protists. Some of these strategies can be the presence of loricae in many bicosoecids and choanoflagellates or scales in chrysophytes, as well as the attachment to larger particles (Arndt et al., 2000).

The process of phagocytosis in HFs

Bacterivory is performed through phagocytosis, an ancient evolutionarily-conserved process (Boulais et al., 2010) with deep implications in the origin and evolution of the eukaryotic cell (Mills, 2020). This process starts with the internalization of the prey through the invagination of the plasma membrane, which creates a vesicle called phagosome (**Fig. 6**). For digestion, this phagosome strongly acidifies its lumen and fuses with a lysosome, which carries digestive enzymes, thus forming the phagolysosome. After absorption of the released nutrients into the cytoplasm, the phagosome is fused again with the plasma membrane and the remaining material is egested to the extracellular space (Flannagan et al., 2012). Phagocytosis is a well characterized process in metazoans due to its primary role in the immune response, but less is known for HF and other protists that rely on this mechanism for their nutrition (Rosales & Uribe-Querol, 2017). Even though the first observations of this process from the ciliate Paramecium date back from the late 1940s (Mast, 1947), very few studies were carried out to study phagocytosis as a feeding mechanism in microbial eukaryotes, with the exception of the

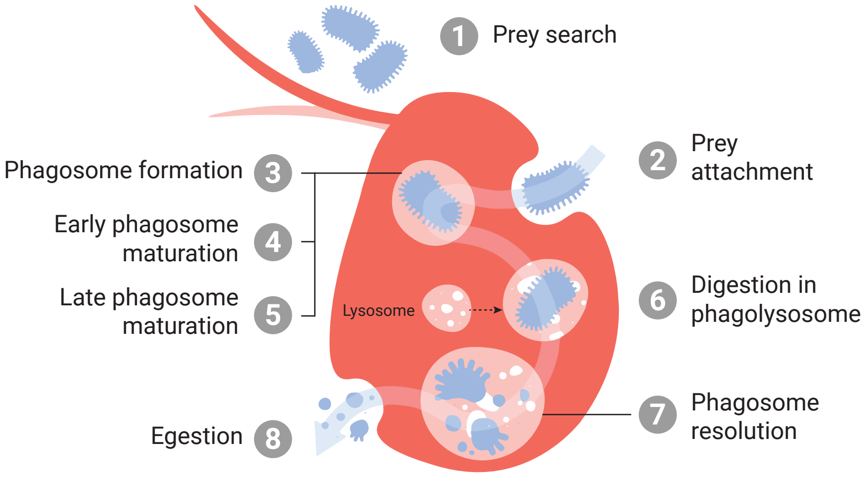
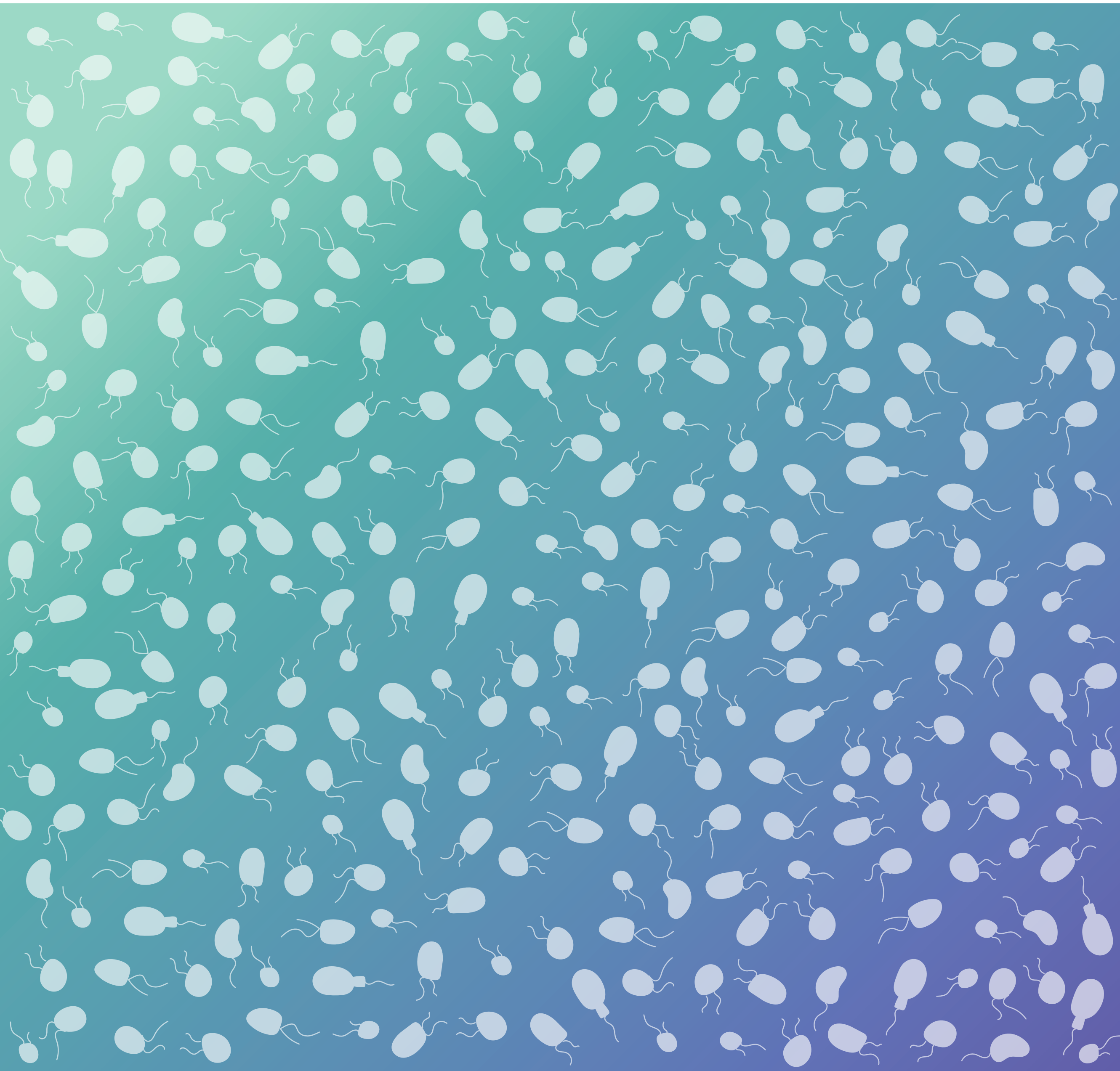


Figure 6. The process of phagocytosis. Schematic overview of phagocytosis in HFs. Adapted from original figure by Marta Royo-Llonch (SHOOK Studio).

amoeba *Dictyostelium discoideum* (Bozzaro et al., 2008). Given its primary role in bacterivory and, extensively, to global biogeochemical cycles, it is crucial to better understand this process in non-model grazers. Phagocytosis is a promising case where a purely eukaryotic function can be analyzed by gene expression, and recent transcriptomics with non-model HFs (Massana et al., 2021; Prokopchuk et al., 2022) and metatranscriptomics experiments with natural samples (Labarre et al., 2020) have been carried out to assess the expression of phagocytosis-related genes. These approaches coupled with extensive databases such as EukProt (Richter et al., 2022), MATOU (Carradec et al., 2018), MGT (Vorobev et al., 2020) or sMAGs (Delmont et al., 2022) are now a promising way to explore the function of HFs in the ocean and disentangle their different and diverse roles in the ocean.

AIMS AND OBJECTIVES



Aims and objectives

Heterotrophic flagellates (HF) represent a key component in the marine ecosystem. During the last decades of the twentieth century, studying them gained some popularity, and some aspects regarding their ecology were characterized. However, this initial attention was later followed by a general lack of interest, mainly due to the difficulty to assess them. With the advent of high-throughput techniques and the growing availability of global sequencing data sets, it is now feasible to study HFs at a high-resolution level. This thesis represents a return to their study, emphasizing their pivotal role in the ocean that seminal studies revealed 40 years ago. **The overall aim of this thesis is to better understand the ecology and functional role of HFs in the ocean** using global and local data sets and different omics approaches.

The thesis is divided into 3 chapters. In Chapter 1 (*A metagenomic assessment of microbial eukaryotic diversity in the global ocean*), we studied the abundance and distribution of microbial eukaryotes in the global ocean from surface to the deep ocean by means of metagenomics and compared it to metabarcoding This analysis revealed that HFs do not generally face technical biases when assessed through this latter approach. Thus, in Chapter 2 (*Oceanic heterotrophic flagellates are dominated by a few widespread taxa*), we investigated the distribution and diversity of HF communities along horizontal and vertical scales by means of metabarcoding. We identified the most abundant HF species in the ocean, most of which remain uncultured, and determined their biogeographical distribution. In Chapter 3 (*Gene expression dynamics of natural assemblages of heterotrophic flagellates during bacterivory*), we characterized the gene expression patterns of natural assemblages of uncultured HFs during phagotrophic feeding in unamended incubations through metatranscriptomics.

The general and specific objectives of this thesis are the following:

Objective 1. To determine the distribution of microbial eukaryotes in the ocean using metagenomics.

- To create a custom database of the 18S-V4 region of the rDNA covering the whole eukaryotic diversity.
- To develop a pipeline to extract and classify 18S-V4 fragments (mTags) from metagenomic data sets using the previous database.
- To compare the community profiles of microbial eukaryotes along depth and between pico- and nanosized fractions.
- To compare the community profiles of microbial eukaryotes obtained with metag-

enomics with the ones obtained through metabarcoding to detect putative technical biases in some taxonomic groups.

- To assess whether amplicon sequencing is a suitable tool to study the abundance and distribution of HFs in the ocean.

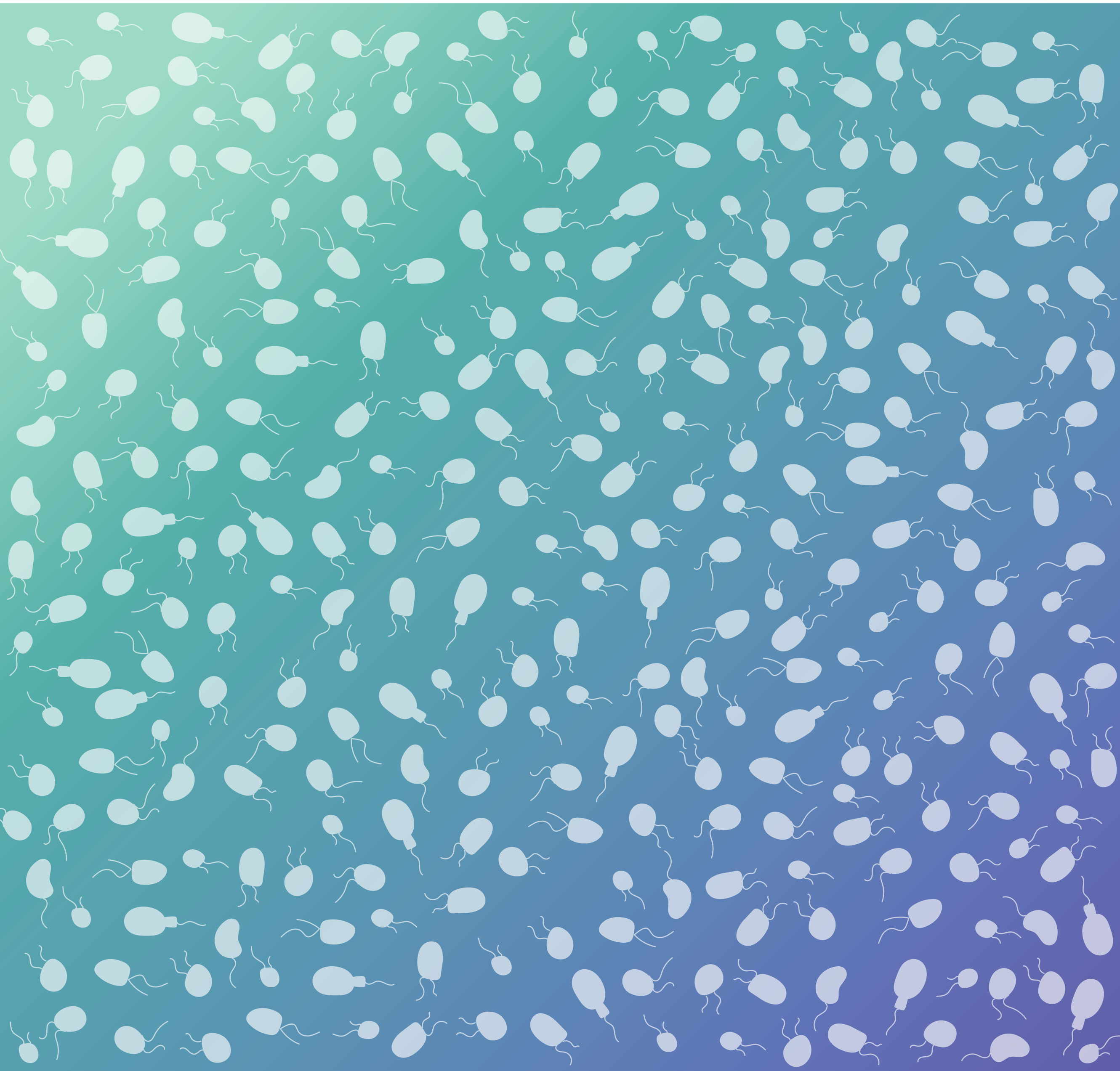
Objective 2. To identify the most abundant HF species in the ocean.

- To assess the diversity of HF assemblages along horizontal and vertical scales.
- To compare the HF community profile obtained using V4 amplicons from the Malaspina survey with other sequencing approaches of the same data set (V9 amplicons and mTags) and metabarcoding from other surveys (TARA oceans).
- To identify the environmental factors driving the distribution of HF species.
- To characterize the distribution patterns of abundant HF species.
- To study the co-occurrence patterns between prokaryotes and HF species.

Objective 3. To better understand the process of bacterivory in HFs.

- To promote the growth of natural HF assemblages in unamended incubations of surface seawater.
- To follow the dynamics of HFs, phototrophic flagellates and bacterial cells through epifluorescence microscopy and metatranscriptomic 18S-V4 mTags.
- To study gene expression of natural HF assemblages at different states of their growth.
- To compare gene expression and taxonomic patterns between incubations started with different natural assemblages.
- To select genes highly expressed in the HF growth state of the incubations, where bacterivory is highest.
- To investigate the expression levels of the selected genes among species with different trophic modes.

CHAPTER 1



CHAPTER 1

A metagenomic assessment of microbial eukaryotic diversity in the global ocean

Aleix Obiol, Caterina R. Giner, Pablo Sánchez, Carlos M. Duarte, Silvia G. Acinas, Ramon Massana

Molecular Ecology Resources (2020). DOI: 10.1111/1755-0998.13147

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.

1.1. Introduction

Marine microbial eukaryotes are key components of planktonic ecosystems in all ocean biomes (Caron et al., 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 (Díez et al., 2001; Edgcomb et al., 2002; López-García et al., 2001; Moon-Van Der Staay et al., 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 et al., 2005; Balzano et al., 2015; Sinclair et al., 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 et al., 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 et al., 2009; Piganeau et al., 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 data set. The development of high-throughput sequencing platforms and the reduction of sequencing costs (Goodwin et al., 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 et al., 2019; Hartmann et al., 2010; Huang et al., 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 et al., 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 et al., 2017), especially when targeting the 18S rDNA gene that contains a mosaic of highly conserved and highly variable regions (Neefs et al., 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 et al., 2016), but these are still highly dependent on good reference databases for a correct taxonomic assignment (Pedrós-Alió et al., 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 data set. 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.

1.2. Materials and Methods

1.2.1. 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 data sets. 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 data sets to detect and correct problematic cases (i.e. distant references sequences retrieving the same mTag). These three levels were: (1) OTU$_{97}$ level (Operational Taxonomic Units of sequences clustered at 97% similarity), (2) taxonomic group (in general a formal Class), and (3) 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 data sets.

1.2.2. 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: three in the Atlantic Ocean, three in the Indian Ocean and four 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 laboratory. 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 et al, 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 (2 x 101 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 (2 x 151 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 (2 x 250 bp) at RTLGenomics (http://rtlgenomics.com/). Amplicon data of the V4 and V9 regions was obtained along the vertical profile of 4 stations (**Figure S1**; **Table S2**).

1.2.3. 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 et al., 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 et al., 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 et al., 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: (a) those with a single hit (OTU$_{97}$ level), (b) those with >1 hit to sequences of the same taxonomic group (Group level), (c) those with >1 hit to sequences of different groups but same supergroup (Supergroup level) and (d) 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.

1.2.4. Amplicon data processing

In both V4 and V9 amplicon data sets, 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 data set, an amplicon sequence variant (ASV) table was obtained. Samples contained 51,421 reads on average (standard error 4,860) and ASVs present in only one 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.

1.2.5. Metagenomes assembly, 18S rDNA contigs retrieval and phylogenetic analyses

Metagenomic samples were assembled using MEGAHIT v1.1.3 (Li et al., 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 et al., 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.

1.2.6. 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).

1.3. Results

1.3.1. 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 data set were 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.

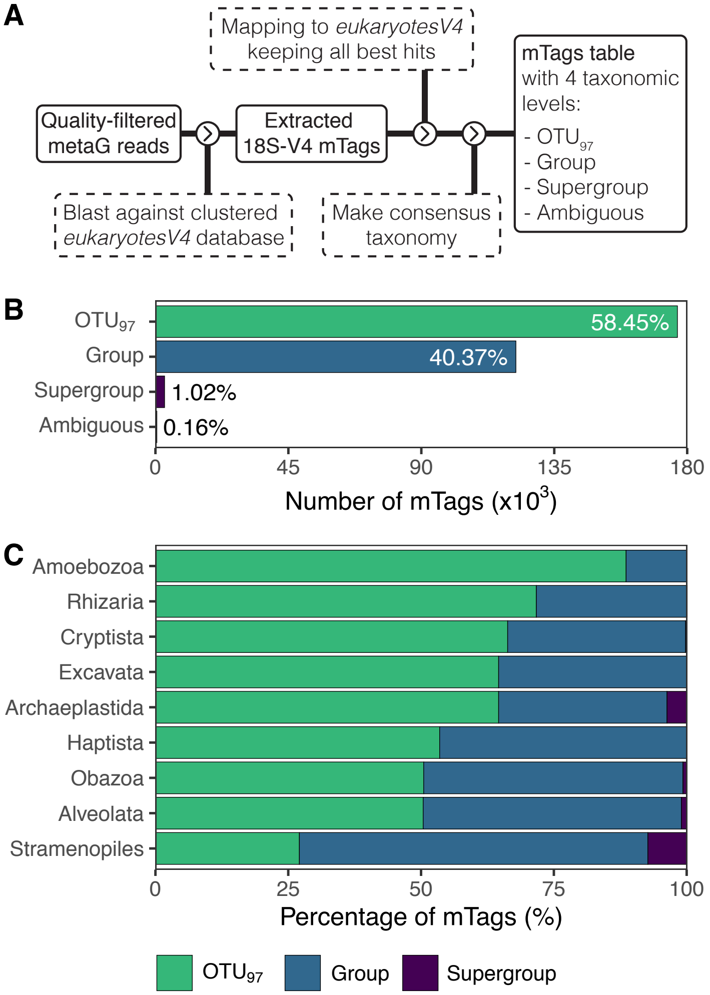


Figure 1. Pipeline for V4-18S rDNA mTags extraction from metagenomes (metaG) and classification and technical results. (**A**) Flow diagram of the pipeline used in this study. (**B**) Number of V4-18S mTags retrieved at the four defined taxonomic levels. (**C**) Relative abundance of the three corresponding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

1.3.2. 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 data set and, except for Prymnesiophyceae, had significantly lower relative abundances in the V9 data set. MALV-II and MALV-I, groups with very high relative abundances in the mTags survey, were significantly overrepresented with V4 amplicons (p < .05), but not with V9 amplicons. Groups equally represented in the three surveys (i.e. did not have significant differences; p > .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.

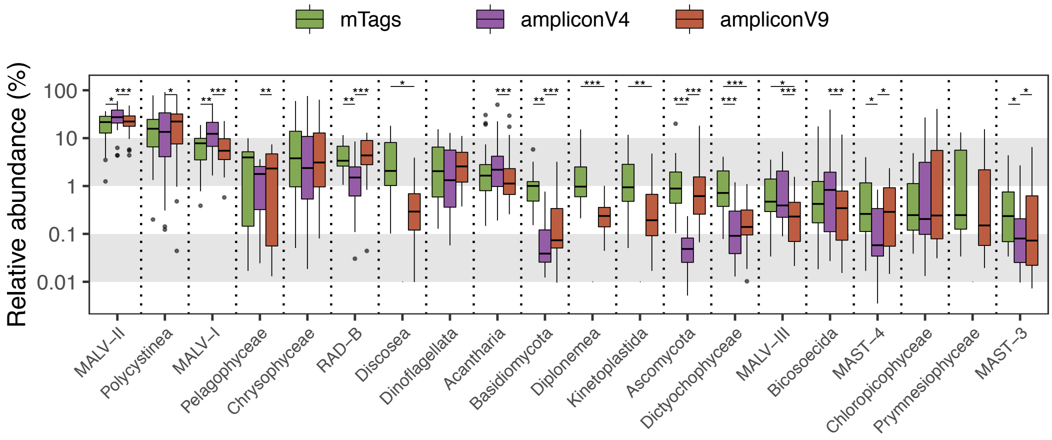


Figure 2. Distribution of the relative abundance of main taxonomic groups as seen by each of the three sequencing approaches (mTags, V4 amplicons and V9 amplicons) in a subset of 23 picoeukaryotic (0.2-3 µm) samples from four vertical profiles. Groups are ordered by decreasing median of the relative abundance by mTags. A $\log_{10}$ scale on the y-axis is used. Significant differences between approaches with Wilcoxon paired tests are shown (*: p < .05, **: p < .01, ***: p < .001).

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 < .001), while depth layer explained 36% of it (p < .001).

1.3.3. Nano and picoeukaryotic diversity assessed by mTags

We used the V4-18S mTags retrieved from the full data set 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- versus nanoeukaryotic fractions and photic versus 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 ($p < .001$), respectively, followed by differences in oceanic regions (7%, $p < .001$), 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.

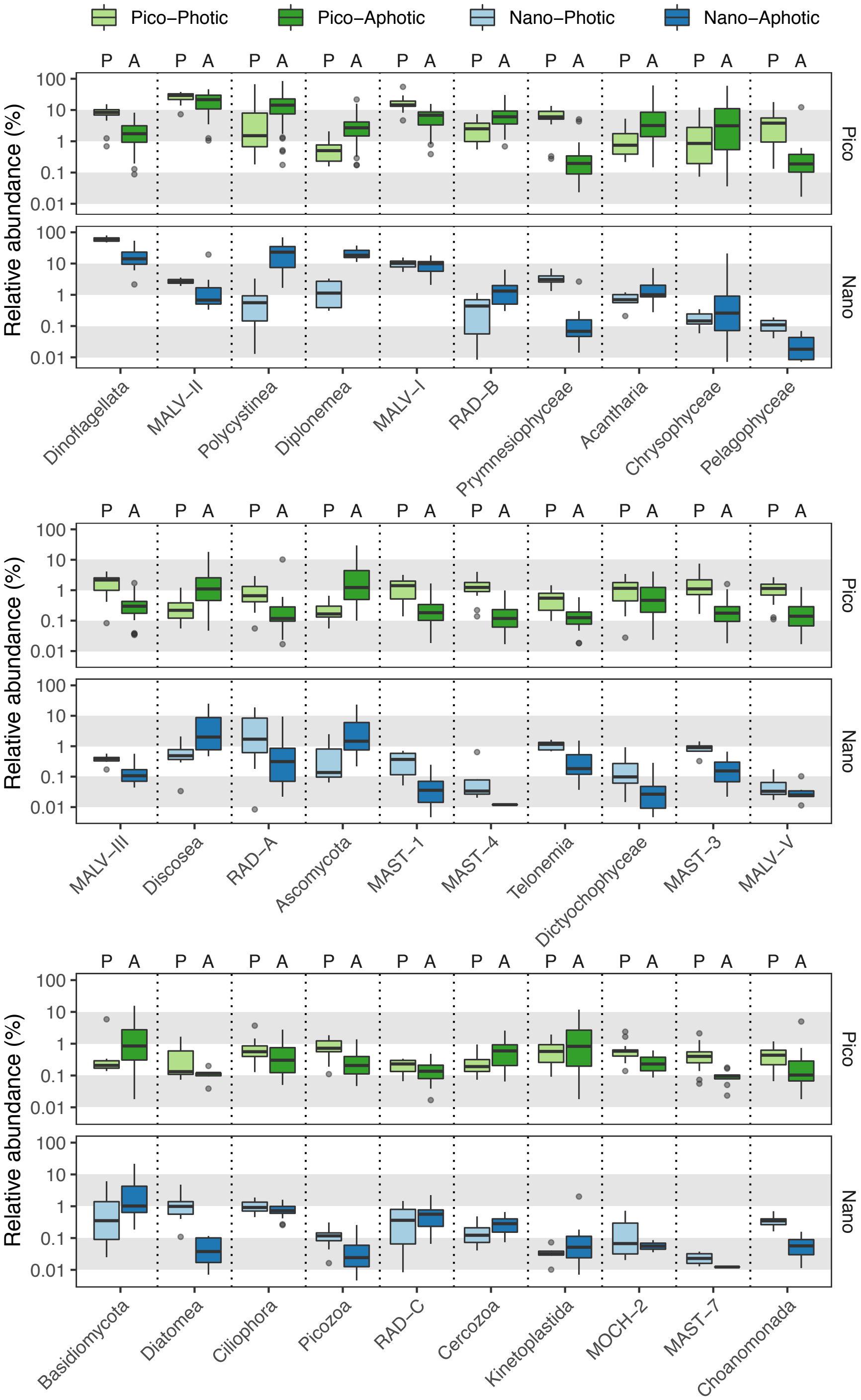


Figure 3. Distribution of the relative abundance of main eukaryotic groups from the pico- (0.2-3 µm) and nanoeukaryotic (3-20 µm) fractions in photic ("P") and aphotic ("A") layers of the ocean as seen by mTags. Groups are ordered by decreasing median relative abundances and $\log_{10}$ scale on the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

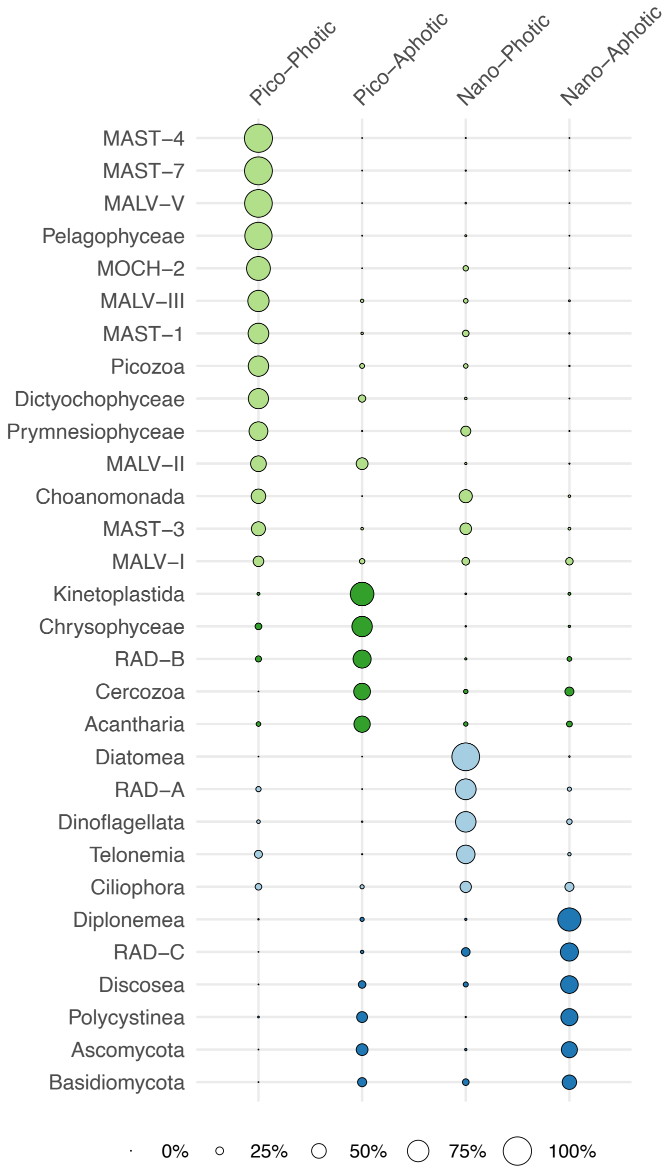


Figure 4. Summary of the relevance of main taxonomic groups in pico (0.2-3 μm)/nanoeukaryotic (3-20μm) fractions and photic/aphotic layers as seen by 18S mTags. The median of the relative abundance was calculated for each taxonomic group with samples from the four categories (pico-photic, pico-aphotic,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

Table 1. Overview of the taxonomic affiliation of the 724 contigs having at least 1,000 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.

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

(18S)(28S) **rDNA operon** ● **complete** ◐ **partial**

1.3.4. 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 >1,000 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 1,500 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 two 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.

1.4. 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 et al., 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

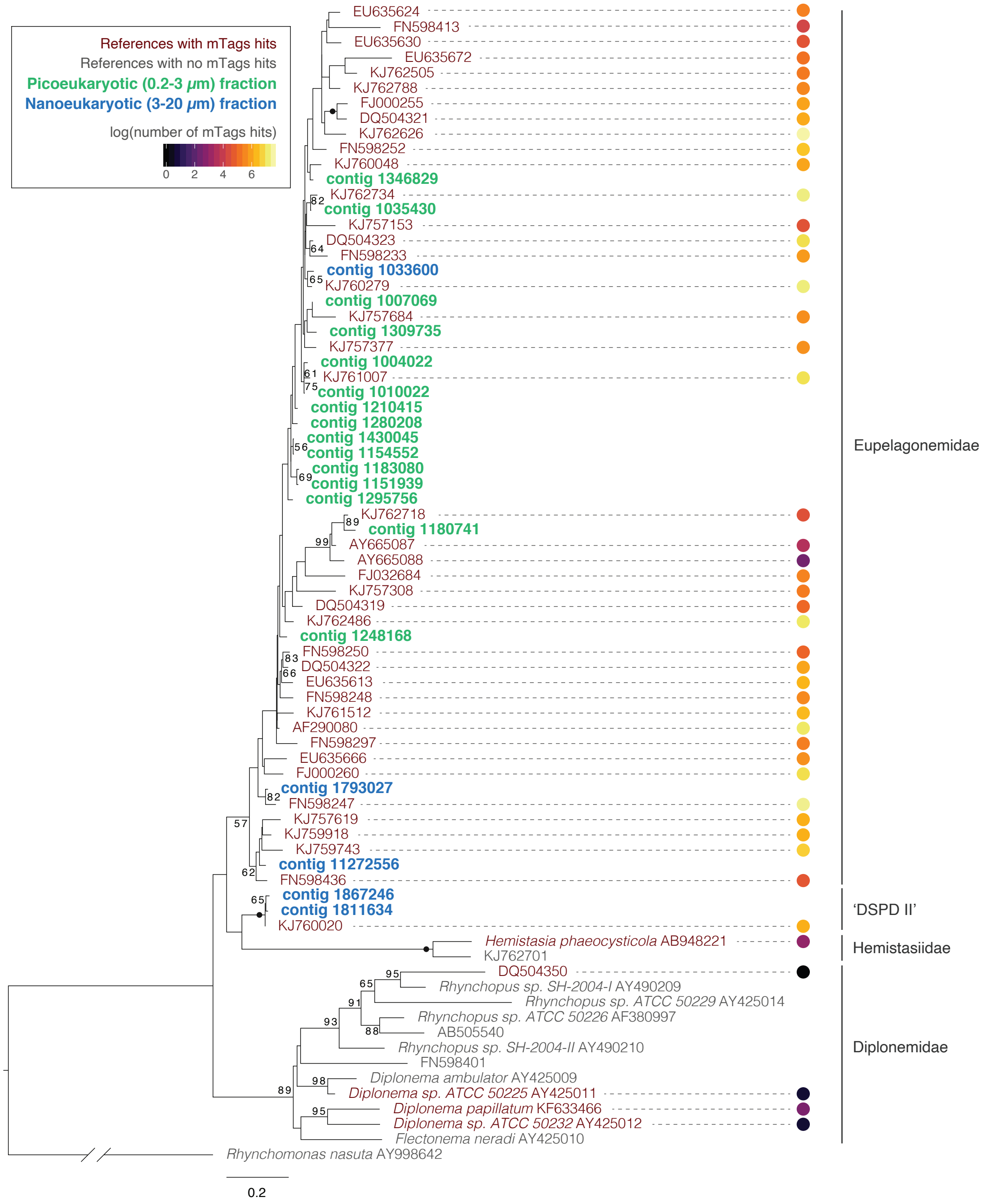


Figure 5. Maximum likelihood phylogenetic tree of Diplonemea using the 18S rDNA from metagenome contigs and reference sequences from the eukaryotesV4 database, derived from an alignment of 940 positions where only shared regions among sequences were kept. RAxML bootstraps are shown when the value is >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.

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 et al., 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 et al., 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 et al., 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 et al., 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 et al., 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 et al., 2016; Piredda et al., 2017; Tragin et al., 2018) and others reporting different performances depending on the taxa (Dunthorn et al., 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 probably limit their amplification. In fact, it was already known that the V4 region of Amoebozoa is not easily amplified (Lahr et al., 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 data set (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 et al., 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 data set (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 et al., 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.

1.5. 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).

1.6. Data availability 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.

1.7. 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 data sets. *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

Díez, 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 data sets. *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. *The 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

1.8. Supplementary figures

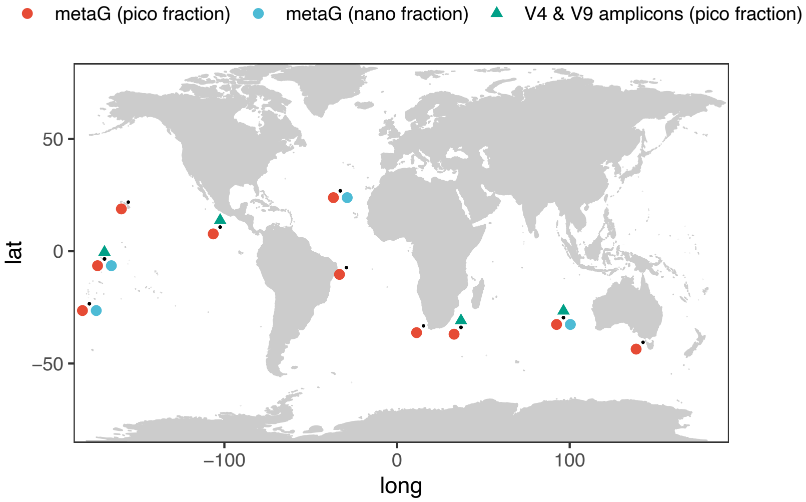


Figure S1. Map showing the geographic position (black dots) of the 10 sampling stations and the sequencing analysis done in each station.

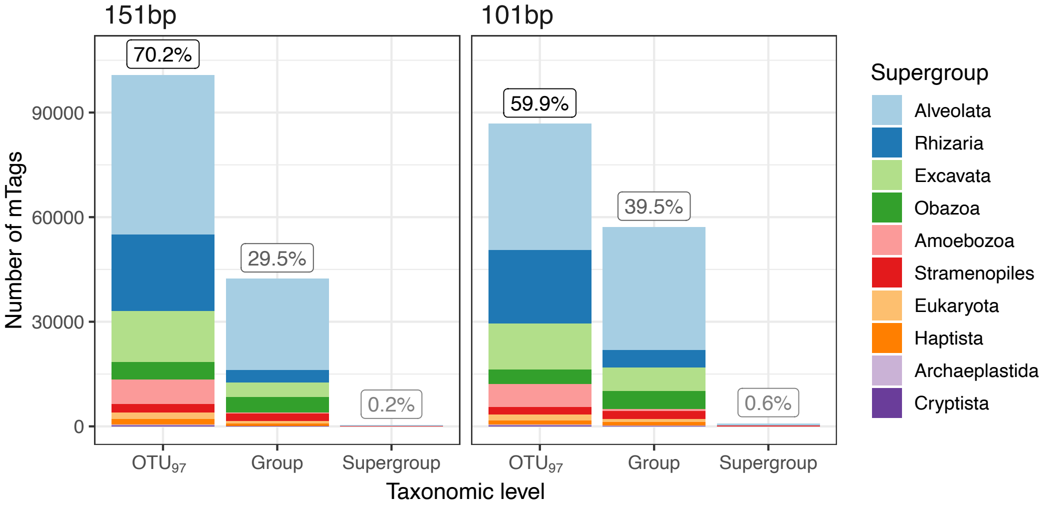


Figure S2. Differences on taxonomic resolution between using the original 151 bp sized mTags from some metagenomes, and the same data set trimmed at 101 bp.

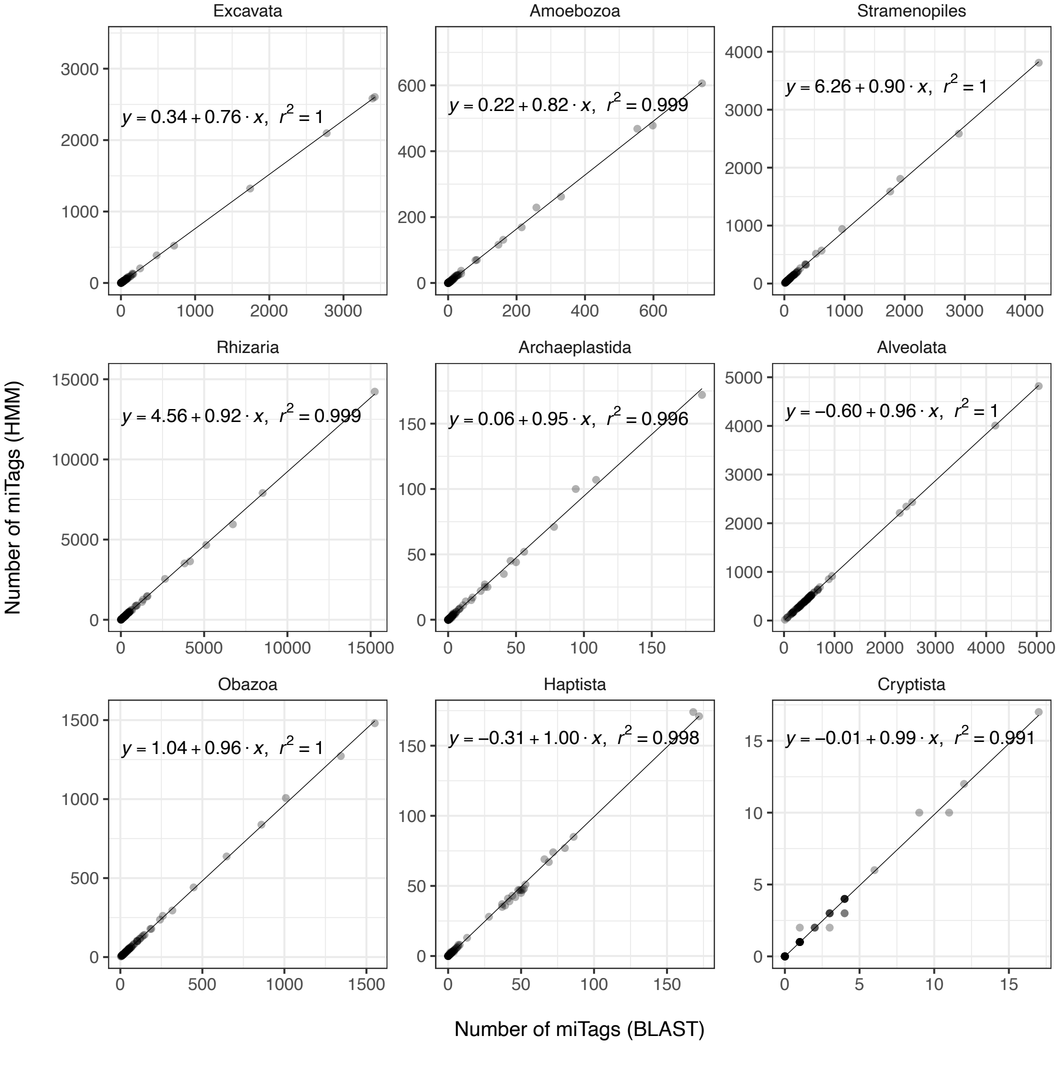


Figure S3. Comparison of 18S-V4 mTags extraction methods. Correlations between the number of mTags extracted by direct BLAST mapping and HMM profiling within the different supergroups, which are ordered by decreasing differences. Linear regression equations and r-squared coefficients are shown.

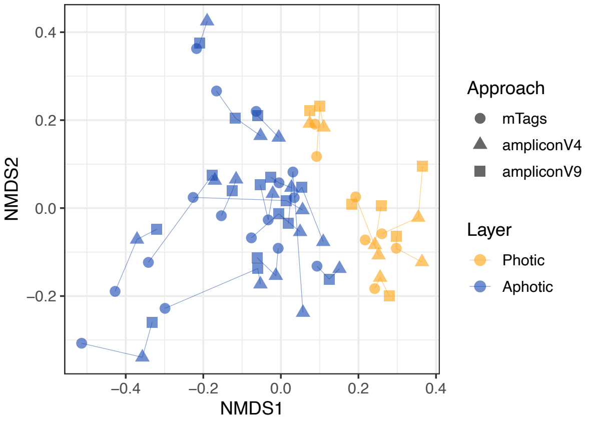
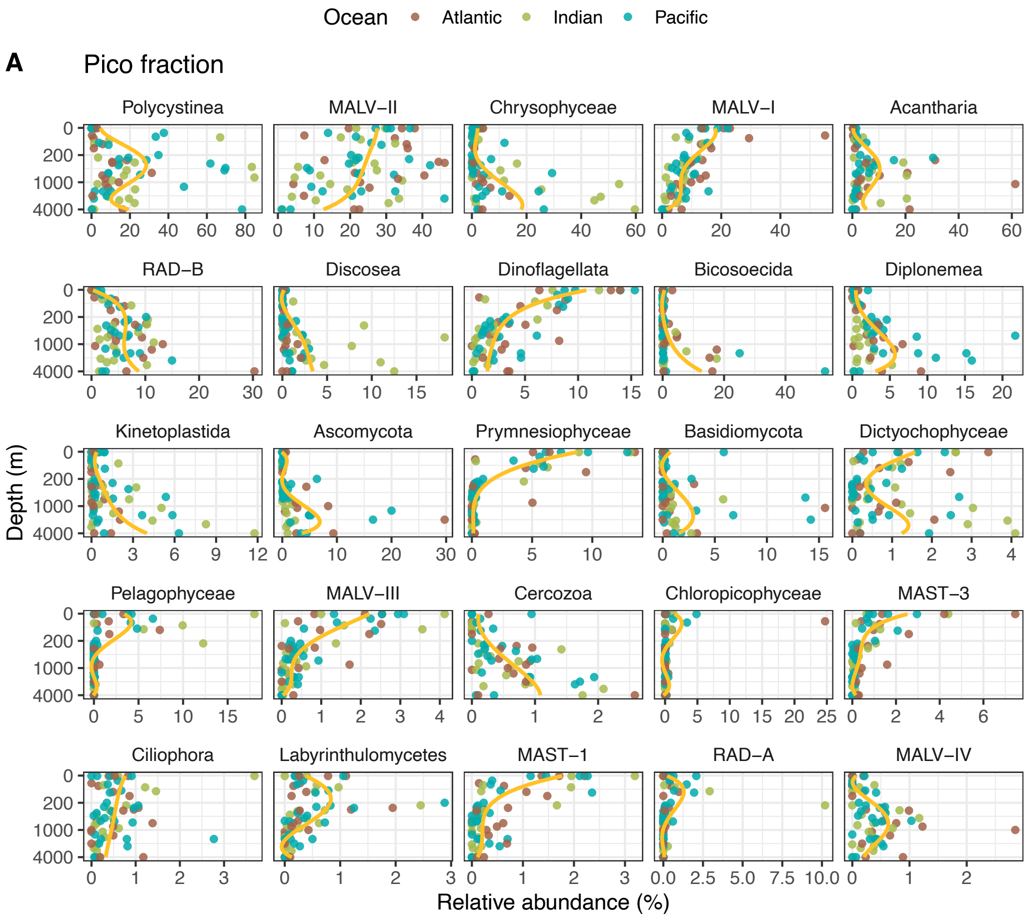


Figure S4. Comparison of 23 samples surveyed with 18S-V4 mTags, V4 and V9 amplicons in a NMDS plot based on Bray-Curtis dissimilarities of community structures defined by the relative abundances of taxonomic groups. Lines join the same environmental sample.



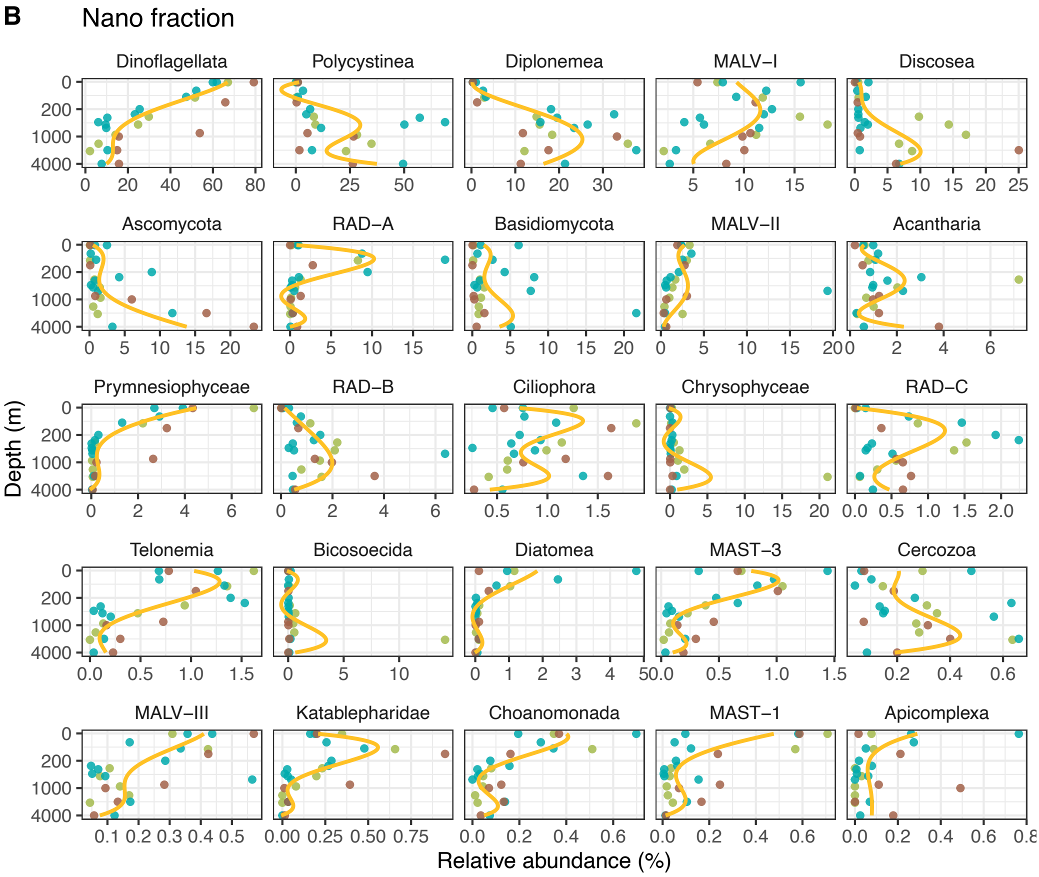


Figure S5. Vertical distribution of main taxonomic groups along the water column as represented by mTags for pico- (**A**) and nanoeukaryotic (**B**) fractions (0.2-3 µm and 3-20 µm, respectively). Each dot represents relative abundance of that group in a specific sample and is colored by ocean. Depth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

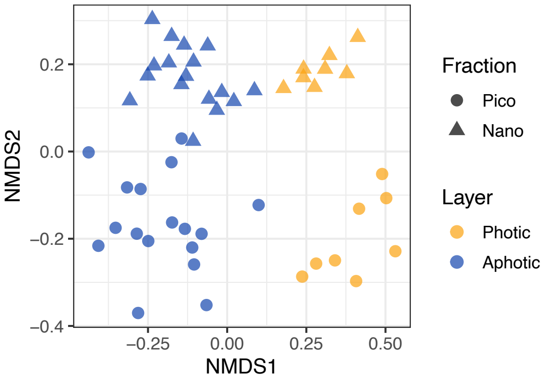


Figure S6. Comparison of the community structure of pico- (0.2-3 µm) and nanoeukaryotic (3-20 µm) fractions along the global ocean as seen in non-metric multidimensional scaling based on Bray-Curtis dissimilarities. Relative abundances of OTU-defined mTags are used.

1.9. Supplementary tables

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.

Group	Supergroup	OTU$_{97}$ sequences	Sequence lengths (bp)		
			Average	Min	Max
Apicomplexa	Alveolata	615	373	305	487
Ciliophora	Alveolata	1975	364	277	519
Colpodellida	Alveolata	28	371	351	383
Colponema	Alveolata	7	379	377	381
Dinoflagellata	Alveolata	1384	377	300	393
Ellobiopsidae	Alveolata	9	365	359	372
InSedAlveolata	Alveolata	138	376	331	400
MALV-I	Alveolata	605	376	300	402
MALV-II	Alveolata	2306	378	307	411
MALV-III	Alveolata	82	379	349	390
MALV-IV	Alveolata	47	381	377	389
MALV-V	Alveolata	14	369	310	379
Perkinsidae	Alveolata	94	379	356	437
X-Cell	Alveolata	56	396	364	422
Archamoebae	Amoebozoa	9	371	332	391
Cavosteliida	Amoebozoa	8	385	357	423
Dictyostelia	Amoebozoa	8	405	381	422
Discosea	Amoebozoa	152	485	384	741
Fractovitelliida	Amoebozoa	3	383	382	386
Gracilipodida	Amoebozoa	17	392	379	465
InSedAmoebozoa	Amoebozoa	37	370	342	457
Myxogastria	Amoebozoa	84	438	417	514
Protosporangiida	Amoebozoa	6	432	382	547
Protosteliida	Amoebozoa	20	378	368	386
Schizoplasmodiida	Amoebozoa	13	391	372	401
Tubulinea	Amoebozoa	87	390	301	469
Charophyta	Archaeplastida	730	380	351	407
Chlorodendrophyceae	Archaeplastida	8	381	377	391
Chlorophyceae	Archaeplastida	156	379	350	432
Chloropicophyceae	Archaeplastida	18	376	357	381
Glaucocystophyceae	Archaeplastida	2	384	383	384
Glaucophyta	Archaeplastida	3	386	381	396
InSedArchaeplastida	Archaeplastida	122	380	347	427
Mamiellophyceae	Archaeplastida	100	377	360	398
Nephroselmidophyceae	Archaeplastida	17	380	377	383
Other-Chlorophyta	Archaeplastida	1	380	380	380
Palmophyllophyceae	Archaeplastida	19	378	371	381
Pedinophyceae	Archaeplastida	7	375	369	380
Prasino-Clade-IX	Archaeplastida	8	372	365	376
Pycnococcaceae	Archaeplastida	3	380	378	381
Pyramimonadales	Archaeplastida	44	379	377	384
Rhodophyta	Archaeplastida	388	382	349	470
Trebouxiophyceae	Archaeplastida	81	380	353	406
Ulvophyceae	Archaeplastida	79	378	364	394
Cryptomonadales	Cryptista	109	380	355	432
Goniomonas	Cryptista	4	422	367	484
Katablepharidae	Cryptista	21	381	372	467
Palpitomonas	Cryptista	1	375	375	375
Ancyromonadida	Eukaryota	39	382	369	401
InSedEukaryota	Eukaryota	324	360	303	454
Malawimonadidae	Eukaryota	1	386	386	386
Mantamonas	Eukaryota	2	385	384	386
Nucleomorph	Eukaryota	18	397	358	455
Picozoa	Eukaryota	31	378	355	384
Rigifilida	Eukaryota	6	388	386	390
Telonemia	Eukaryota	32	376	361	386
Diplonemea	Excavata	56	535	492	549
Euglenida	Excavata	194	627	510	765
Fornicata	Excavata	14	367	354	387
Heterolobosea	Excavata	24	475	332	506
Jakobida	Excavata	18	411	359	483
Kinetoplastida	Excavata	220	547	451	744
Parabasalia	Excavata	159	270	256	316
Preaxostyla	Excavata	11	405	279	611
Centrohelida	Haptista	159	399	379	471
Pavlovophyceae	Haptista	25	377	364	388
Prymnesiophyceae	Haptista	63	380	372	387
Apusomonadida	Obazoa	132	382	345	411
Ascomycota	Obazoa	601	375	265	452
BasalFungi	Obazoa	704	384	275	546
Basidiomycota	Obazoa	416	380	350	408
Breviatea	Obazoa	24	379	367	391
Choanomonada	Obazoa	241	383	352	502
Discicristoidea	Obazoa	39	407	373	567
Filasterea	Obazoa	5	376	355	382
Ichthyosporea	Obazoa	47	385	302	528
InSedObazoa	Obazoa	67	386	343	451
MarineOpisthokonts	Obazoa	27	376	370	384
Metazoa	Obazoa	7220	427	287	789
Acantharia	Rhizaria	434	368	337	384
Cercozoa	Rhizaria	1440	392	346	575
Chlorarachniophyta	Rhizaria	56	385	366	422
Foraminifera	Rhizaria	15	647	332	764
InSedRhizaria	Rhizaria	35	388	368	398
Polycystinea	Rhizaria	318	382	354	394
RAD-A	Rhizaria	64	386	384	389
RAD-B	Rhizaria	66	385	367	395
RAD-C	Rhizaria	28	378	376	381
Aurearenophyceae	Stramenopiles	1	396	396	396
Bicosoecida	Stramenopiles	165	371	317	429
Blastocystis	Stramenopiles	45	393	367	503
Bolidomonas	Stramenopiles	34	382	378	385
Cantina	Stramenopiles	1	378	378	378
Chrysomerophyceae	Stramenopiles	4	388	386	389
Chrysophyceae	Stramenopiles	299	383	333	451
Developayella	Stramenopiles	6	385	382	387
Diatomea	Stramenopiles	822	379	337	475
Dictyochophyceae	Stramenopiles	94	392	382	429
Eustigmatales	Stramenopiles	20	388	382	401
Hyphochytriales	Stramenopiles	5	387	377	391
InSedStramenopiles	Stramenopiles	134	389	359	450
Labyrinthulomycetes	Stramenopiles	414	381	312	417
MAST-1	Stramenopiles	28	381	348	386
MAST-10	Stramenopiles	2	380	379	382
MAST-11	Stramenopiles	3	380	377	383
MAST-12	Stramenopiles	65	387	382	396
MAST-16	Stramenopiles	1	389	389	389
MAST-2	Stramenopiles	4	382	381	383
MAST-20	Stramenopiles	2	380	378	382
MAST-22	Stramenopiles	8	384	378	386
MAST-23	Stramenopiles	2	388	388	389
MAST-24	Stramenopiles	2	386	386	387
MAST-25	Stramenopiles	5	375	373	377
MAST-3	Stramenopiles	170	384	337	430
MAST-4	Stramenopiles	6	379	372	383
MAST-6	Stramenopiles	17	385	381	387
MAST-7	Stramenopiles	19	378	373	381
MAST-8	Stramenopiles	26	386	383	391
MAST-9	Stramenopiles	36	379	375	396
MOCH-1	Stramenopiles	21	384	381	386
MOCH-2	Stramenopiles	33	383	373	388
MOCH-3	Stramenopiles	2	388	388	389
MOCH-4	Stramenopiles	3	388	387	388
MOCH-5	Stramenopiles	7	383	380	390
Olisthodiscus	Stramenopiles	1	387	387	387
Opalinata	Stramenopiles	4	394	347	436
Pelagophyceae	Stramenopiles	34	383	339	389
Peronosporomycetes	Stramenopiles	104	391	380	405
Phaeophyceae	Stramenopiles	25	395	390	405
Phaeothamniophyceae	Stramenopiles	3	388	387	388
Picophagea	Stramenopiles	3	386	384	388
Pinguiochrysidales	Stramenopiles	8	389	382	394
Pirsonia	Stramenopiles	21	388	384	392
Placidida	Stramenopiles	6	386	367	398
Raphidophyceae	Stramenopiles	9	392	388	405
Xanthophyceae	Stramenopiles	25	395	377	413

Table S2. Analyzed samples and their associated metadata.

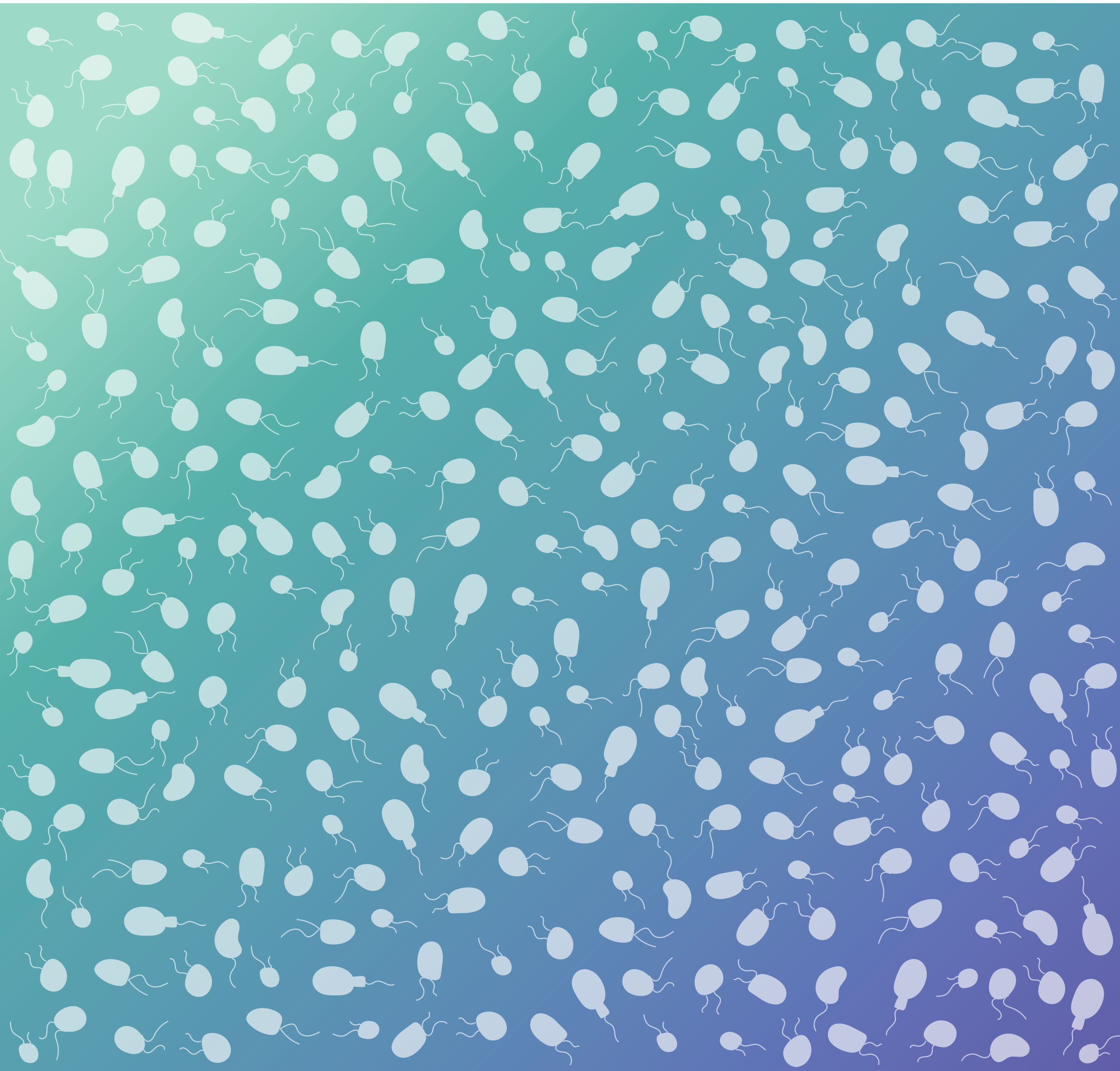
Station	Depth	Layer	Layer2	temp	cond	salinity	O2	Ocean	Sub ocean	Long prov	lat	long	mTags pico	mTags nano	amplicon V4	amplicon V9
19	3	Photic	Surface	27.11	5.74	36.48	4.33	AO	SAO	SATL	-7.283	-29.322	x			
19	76	Photic	DSL	26.60	5.70	36.59	4.40	AO	SAO	SATL	-7.283	-29.322	x			
19	120	Photic	DCM	21.73	5.16	36.50	3.88	AO	SAO	SATL	-7.283	-29.322	x			
19	400	Aphotic	Mesopelagic	8.64	3.68	34.78	1.61	AO	SAO	SATL	-7.283	-29.322	x			
19	800	Aphotic	Mesopelagic	4.68	3.31	34.46	3.44	AO	SAO	SATL	-7.283	-29.322	x			
19	1600	Aphotic	Bathypelagic	4.15	3.34	34.96	5.05	AO	SAO	SATL	-7.283	-29.322	x			
19	4000	Aphotic	Bathypelagic	1.96	3.22	34.85	5.40	AO	SAO	SATL	-7.283	-29.322	x			
44	3	Photic	Surface	20.66	4.93	35.56	4.85	AO	SAO	BENG	-33.23	15.34	x			
44	56	Photic	DCM	18.30	4.69	35.53	4.90	AO	SAO	BENG	-33.23	15.34	x			
44	350	Aphotic	Mesopelagic	9.41	3.74	34.72	4.50	AO	SAO	BENG	-33.23	15.34	x			
44	725	Aphotic	Mesopelagic	4.53	3.28	34.37	4.37	AO	SAO	BENG	-33.23	15.34	x			
44	1200	Aphotic	Bathypelagic	3.02	3.19	34.59	3.76	AO	SAO	BENG	-33.23	15.34	x			
49	86	Photic	DCM	19.23	4.79	35.57	4.64	IO	IO	ISSG	-33.91	37.042	x		x	x
49	450	Aphotic	Mesopelagic	13.54	4.20	35.29	4.56	IO	IO	ISSG	-33.91	37.042	x		x	x
49	800	Aphotic	Mesopelagic	10.08	3.83	34.82	4.65	IO	IO	ISSG	-33.91	37.042	x		x	x
49	1200	Aphotic	Bathypelagic	5.31	3.38	34.43	4.18	IO	IO	ISSG	-33.91	37.042	x		x	x
49	3000	Aphotic	Bathypelagic	2.27	3.21	34.81	NA	IO	IO	ISSG	-33.91	37.042	x		x	x
49	4000	Aphotic	Bathypelagic	1.11	3.14	34.74	4.67	IO	IO	ISSG	-33.91	37.042	x		x	x
63	3	Photic	Surface	21.74	5.07	35.79	4.64	IO	IO	ISSG	-29.57	96.407	x	x	x	x
63	115	Photic	DCM	16.00	4.47	35.68	5.07	IO	IO	ISSG	-29.57	96.407	x	x	x	x
63	420	Aphotic	Mesopelagic	10.80	3.89	34.89	5.21	IO	IO	ISSG	-29.57	96.407	x	x	x	x
63	650	Aphotic	Mesopelagic	9.21	3.73	34.67	5.01	IO	IO	ISSG	-29.57	96.407	x	x	x	x
63	950	Aphotic	Mesopelagic	4.99	3.33	34.37	4.22	IO	IO	ISSG	-29.57	96.407	x	x	x	x
63	1790	Aphotic	Bathypelagic	2.69	3.19	34.68	3.31	IO	IO	ISSG	-29.57	96.407	x	x	x	x
63	2600	Aphotic	Bathypelagic	1.79	3.15	34.73	3.83	IO	IO	ISSG	-29.57	96.407	x	x	x	x
76	3	Photic	Surface	15.75	4.36	34.99	5.23	SAB	SAB	SSTC	-40.55	142.5	x			
76	70	Photic	DCM	13.43	4.15	35.11	5.19	SAB	SAB	SSTC	-40.55	142.5	x			
76	275	Aphotic	Mesopelagic	10.85	3.90	34.95	5.32	SAB	SAB	SSTC	-40.55	142.5	x			
76	550	Aphotic	Mesopelagic	8.65	3.67	34.59	4.85	SAB	SAB	SSTC	-40.55	142.5	x			
76	860	Aphotic	Mesopelagic	5.22	3.35	34.40	4.16	SAB	SAB	SSTC	-40.55	142.5	x			
76	2800	Aphotic	Bathypelagic	1.76	3.15	34.73	3.90	SAB	SAB	SSTC	-40.55	142.5	x			
76	3300	Aphotic	Bathypelagic	1.47	3.15	34.73	4.09	SAB	SAB	SSTC	-40.55	142.5	x			
83	3	Photic	Surface	26.67	5.52	35.28	4.22	PO	SPO	SPSG	-23.38	-178.21	x	x		
83	110	Photic	DCM	23.62	5.26	35.69	4.23	PO	SPO	SPSG	-23.38	-178.21	x	x		
83	200	Aphotic	Mesopelagic	20.86	4.97	35.63	3.84	PO	SPO	SPSG	-23.38	-178.21	x	x		
83	350	Aphotic	Mesopelagic	16.73	4.53	35.41	4.03	PO	SPO	SPSG	-23.38	-178.21	x	x		
83	750	Aphotic	Mesopelagic	6.45	3.45	34.37	4.51	PO	SPO	SPSG	-23.38	-178.21	x	x		
83	2000	Aphotic	Bathypelagic	2.45	3.17	34.61	3.15	PO	SPO	SPSG	-23.38	-178.21	x			
83	2500	Aphotic	Bathypelagic	2.39	3.19	34.62	3.12	PO	SPO	SPSG	-23.38	-178.21	x	x		
92	3	Photic	Surface	28.13	5.69	35.39	4.22	PO	NPO	PEQD	-3.41	-169.46	x	x	x	x
92	65	Photic	DCM	28.03	5.69	35.42	4.22	PO	NPO	PEQD	-3.41	-169.46	x	x	x	x
92	450	Aphotic	Mesopelagic	8.47	3.65	34.65	1.67	PO	NPO	PEQD	-3.41	-169.46	x	x	x	x
92	580	Aphotic	Mesopelagic	7.42	3.56	34.59	1.48	PO	NPO	PEQD	-3.41	-169.46	x	x		
92	650	Aphotic	Mesopelagic	6.71	3.49	34.56	1.72	PO	NPO	PEQD	-3.41	-169.46	x	x		
92	1500	Aphotic	Bathypelagic	3.00	3.20	34.60	2.25	PO	NPO	PEQD	-3.41	-169.46	x		x	x
92	4000	Aphotic	Bathypelagic	1.40	3.16	34.70	3.57	PO	NPO	PEQD	-3.41	-169.46	x	x	x	x
101	3	Photic	Surface	24.65	5.24	34.76	4.50	PO	NPO	NPTG	21.89	-155.66	x			
101	125	Photic	DCM	23.46	5.18	35.23	4.57	PO	NPO	NPTG	21.89	-155.66	x			
101	320	Aphotic	Mesopelagic	12.83	4.01	34.22	3.22	PO	NPO	NPTG	21.89	-155.66	x			
101	500	Aphotic	Mesopelagic	7.10	3.48	34.09	2.05	PO	NPO	NPTG	21.89	-155.66	x			
101	730	Aphotic	Mesopelagic	5.07	3.33	34.32	0.67	PO	NPO	NPTG	21.89	-155.66	x			
101	2000	Aphotic	Bathypelagic	2.20	3.15	34.62	1.90	PO	NPO	NPTG	21.89	-155.66	x			
101	4000	Aphotic	Bathypelagic	1.47	3.17	34.69	3.16	PO	NPO	NPTG	21.89	-155.66	x			
120	3	Photic	Surface	29.28	5.53	33.46	4.18	PO	NPO	PNEC	10.759	-102.44	x		x	x
120	37	Photic	DCM	26.14	5.29	34.02	3.40	PO	NPO	PNEC	10.759	-102.44	x		x	x
120	280	Aphotic	Mesopelagic	11.12	3.90	34.75	0.12	PO	NPO	PNEC	10.759	-102.44	x		x	x
120	780	Aphotic	Mesopelagic	5.74	3.41	34.56	0.10	PO	NPO	PNEC	10.759	-102.44	x		x	x
120	2000	Aphotic	Bathypelagic	2.26	3.16	34.65	1.86	PO	NPO	PNEC	10.759	-102.44	x		x	x
120	2800	Aphotic	Bathypelagic	1.85	3.16	34.67	2.32	PO	NPO	PNEC	10.759	-102.44	x			
141	3	Photic	Surface	23.87	5.52	37.53	4.49	AO	NAO	NASE	26.911	-32.837	x	x		
141	150	Photic	DCM	20.05	5.05	36.99	4.62	AO	NAO	NASE	26.911	-32.837	x	x		
141	430	Aphotic	Mesopelagic	14.52	4.37	35.99	3.98	AO	NAO	NASE	26.911	-32.837	x			
141	900	Aphotic	Mesopelagic	8.40	3.72	35.27	3.35	AO	NAO	NASE	26.911	-32.837	x	x		
141	1000	Aphotic	Mesopelagic	7.63	3.65	35.25	3.59	AO	NAO	NASE	26.911	-32.837	x	x		
141	2500	Aphotic	Bathypelagic	3.18	3.29	34.99	5.25	AO	NAO	NASE	26.911	-32.837	x	x		
141	4000	Aphotic	Bathypelagic	2.44	3.27	34.91	5.22	AO	NAO	NASE	26.911	-32.837	x	x		

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. Available at: https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2F1755-0998.13147&file=men13147-sup-0002-TableS1-S4.xlsx

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	Pico-Photic		Pico-Aphotic		Nano-Photic		Nano-Aphotic	
	median	mean ± stderr	median	mean ± stderr	median	mean ± stderr	median	mean ± stderr
Dinoflagellata	8.4	8.6 ± 0.9	1.7	2.1 ± 0.3	60.9	60.6 ± 3.6	14.2	16.5 ± 3
MALV-II	29.2	27.1 ± 1.8	21.5	21.2 ± 1.8	2.6	2.7 ± 0.2	0.7	2.2 ± 1.1
Polycystinea	1.4	8.9 ± 3.9	14.3	21.4 ± 3.4	0.6	0.8 ± 0.4	23.2	24.7 ± 5.1
Diplonemea	0.4	0.5 ± 0.1	2.7	4 ± 0.7	1.1	1.5 ± 0.4	18.4	21.5 ± 2.1
MALV-I	14.6	17.3 ± 2.4	6.8	6.6 ± 0.6	10.2	10.1 ± 1.1	9.9	8.9 ± 1.1
RAD-B	1.9	2.3 ± 0.5	6.1	6.7 ± 0.7	0.5	0.5 ± 0.1	1.3	1.6 ± 0.4
Prymnesiophyceae	6	6.8 ± 0.8	0.1	0.4 ± 0.1	3.1	3.4 ± 0.6	0.1	0.3 ± 0.2
Acantharia	0.7	1.3 ± 0.3	3.1	8 ± 1.7	0.7	0.7 ± 0.1	1	1.7 ± 0.4
Chrysophyceae	0.9	1.9 ± 0.6	2.9	9.6 ± 2.2	0.1	0.1 ± 0	0.2	1.6 ± 1.2
Pelagophyceae	2.5	3.7 ± 1	0	0.3 ± 0.3	0.1	0.1 ± 0	0	0 ± 0
MALV-III	2.2	1.9 ± 0.2	0.3	0.3 ± 0.1	0.4	0.4 ± 0	0.1	0.1 ± 0
Discosea	0	0.1 ± 0.1	0.8	2.2 ± 0.5	0.5	0.8 ± 0.3	2	6 ± 1.7
RAD-A	0.4	0.7 ± 0.2	0	0.3 ± 0.2	1.9	5.1 ± 2.3	0.3	1 ± 0.5
Ascomycota	0	0.1 ± 0	1	3.2 ± 0.8	0.1	0.6 ± 0.3	1.5	4.9 ± 1.6
MAST-1	1.4	1.3 ± 0.2	0.1	0.2 ± 0	0.4	0.4 ± 0.1	0	0.1 ± 0
MAST-4	1.2	1.4 ± 0.2	0	0.1 ± 0	0	0.1 ± 0.1	0	0 ± 0
Telonemia	0.5	0.5 ± 0.1	0	0.1 ± 0	1.2	1.1 ± 0.1	0.2	0.4 ± 0.1
Dictyochophyceae	1.2	1.3 ± 0.2	0.4	0.8 ± 0.2	0.1	0.2 ± 0.1	0	0 ± 0
MAST-3	1.1	1.8 ± 0.4	0.1	0.2 ± 0	0.9	0.9 ± 0.1	0.2	0.2 ± 0
MALV-V	1.1	1 ± 0.2	0	0.1 ± 0	0	0 ± 0	0	0 ± 0
Basidiomycota	0	0.4 ± 0.3	0.6	1.9 ± 0.5	0.4	1.4 ± 0.7	1	3.3 ± 1.3
Diatomea	0	0.2 ± 0.1	0	0 ± 0	1	1.4 ± 0.5	0	0.1 ± 0
Ciliophora	0.5	0.7 ± 0.2	0.3	0.5 ± 0.1	0.9	1.1 ± 0.2	0.7	0.8 ± 0.1
Picozoa	0.6	0.8 ± 0.1	0.1	0.2 ± 0	0.1	0.1 ± 0	0	0 ± 0
RAD-C	0	0.1 ± 0	0.1	0.1 ± 0	0.3	0.5 ± 0.2	0.6	0.7 ± 0.2
Cercozoa	0	0.1 ± 0	0.6	0.7 ± 0.1	0.1	0.2 ± 0.1	0.3	0.3 ± 0
Kinetoplastida	0	0.3 ± 0.1	0.5	1.7 ± 0.4	0	0 ± 0	0	0.2 ± 0.1
MOCH-2	0.4	0.5 ± 0.1	0	0 ± 0	0.1	0.2 ± 0.1	0	0 ± 0
MAST-7	0.4	0.5 ± 0.1	0	0 ± 0	0	0 ± 0	0	0 ± 0
Choanomonada	0.4	0.4 ± 0.1	0	0.2 ± 0.1	0.3	0.4 ± 0.1	0	0.1 ± 0

CHAPTER 2



CHAPTER 2

Oceanic heterotrophic flagellates are dominated by a few widespread taxa

Aleix Obiol, Imer Muhovic, Ramon Massana

Limnology and Oceanography (2021). DOI: 10.1002/lno.11956

Abstract

Marine heterotrophic flagellates (HF) form a diverse and ecologically relevant functional group of bacterial grazers and nutrient remineralizers in oceanic waters. Despite playing a crucial role in marine biogeochemical cycles, there is still a lack of information on which specific taxa dominate HF assemblages and what are their patterns of distribution in a global context. In the present work we addressed this issue by analyzing amplicon sequencing data sets retrieved from samples taken in tropical and subtropical oceanic regions at depths from surface to 4000 m. Only a few dozens of widespread taxa, mostly affiliating to MAST clades, Picozoa, Bicosoecida and Chrysophyceae, seemed to dominate surface HF assemblages. The majority of these dominant HFs were present at relatively constant abundances, while others were influenced by temperature or displayed a patchy distribution. In the deep ocean, only a handful of taxa belonging to Bicosoecida and Chrysophyceae, together with Diplonemea and Kinetoplastida, explained most of the HF signal. Co-occurrence networks between HF and prokaryotic taxa at the surface ocean revealed two main clusters influenced by temperature that did not seem to show specific patterns of interaction. However, some correlations emerged outside these thermal groups that could represent new prey-predator interactions. Overall, we identified the putatively most ecologically relevant HF taxa in the ocean, which become promising targets for further experimental and genomic studies.

2.1. Introduction

Marine heterotrophic flagellates (HFs) are minute unpigmented eukaryotes (2-20 µm in size) that are found in the plankton at concentrations of 10^2-10^4 cells ml^{-1}. They represent around 20% of total eukaryotic organisms in the photic zone of the oceans (Jürgens & Massana, 2008). Collectively, they form a highly diverse assemblage, with species affiliated with all major eukaryotic supergroups (Adl et al., 2019; Jeuck & Arndt, 2013; Schön et al., 2021). Consideration of their importance in ocean ecosystems changed dramatically about 40 yr ago, when it was shown that these microorganisms were active bacterial grazers and were part of marine food webs (Azam et al., 1983). Further studies confirmed that HFs, and particularly those in the 2-5 µm size range, were major agents of prokaryotic mortality in planktonic systems (Fenchel, 1986), along with viruses (Fuhrman & Noble, 1995), and mixotrophic protists (Zubkov & Tarran, 2008). In these planktonic systems, HFs are crucial in channeling carbon to higher trophic levels and in regenerating inorganic nutrients such as nitrogen and phosphorus that would otherwise be kept within bacterial biomass (Jürgens & Massana, 2008; Pernthaler, 2005; Sherr & Sherr, 2002).

Despite their significance in the environment, studying HFs presents some challenges that have impeded obtaining a detailed image of the diversity and function of natural HF assemblages. First, contrary to larger eukaryotic microorganisms, many HF cells lack conspicuous morphological traits that could be used for taxonomy, thus hindering their identification through light microscopy. Second, the available cultured strains do not generally represent the dominant species in the environment, many of which still remain uncultured (del Campo et al., 2013). So, ecophysiological studies on cultured strains (functional and numerical responses, recycling capacity, environmental responses) may poorly account for *in situ* ecological performances. In fact, HF assemblages have been usually treated as a black box in terms of their ecological activity in marine ecosystems, ignoring they are formed by species with distinct ecophysiologies. Molecular diversity studies have been fundamental to open this black box.

The first molecular diversity surveys of marine protists, including HFs, cloned and sequenced the 18S ribosomal DNA gene as a phylogenetic marker (Díez et al., 2001; Edgcomb et al., 2002; López-García et al., 2001; Moon-Van Der Staay et al., 2001). These revealed a large taxonomic diversity in natural assemblages and allowed new uncultured lineages to be described (Guillou et al., 2008; Massana et al., 2004; Not et al., 2007). Recent global oceanic expeditions, using high-throughput sequencing, also surveyed eukaryotic diversity (de Vargas et al., 2015; Giner et al., 2020; Obiol et al., 2020), expanding the initial picture and creating precious resources for further detailed studies. Recent publications have used the released data to assess the diversity and biogeography of specific groups, such as diatoms (Malviya et al., 2016), dinoflagellates (Le Bescot et al., 2016), green algae (Lopes dos Santos et al., 2017; Metz et al., 2019), kinetoplastids (Flegontova et al., 2018) and ciliates (Canals et al., 2020). An equivalent study targeting the taxonomically heterogeneous HF assemblages in the water column of the oceans is still missing.

Here, we used previously released sequencing data sets to address a series of crucial questions related to understanding marine HFs at a large scale: (1) Which are the main HF taxonomic groups in marine waters? (2) Are HF assemblages shaped by environmental factors? (3) Are there globally-dominant HF species and what are their spatial distributions? (4) Can we detect specific co-occurrence patterns among dominant HF species and prokaryotic taxa? To answer these questions, we analyzed V4 18S rDNA sequences assigned to HF taxonomic groups in 279 metabarcoding samples. These corresponded to the picoeukaryotic fraction (0.2-3 µm) and were collected during the Malaspina 2010 circumglobal expedition in surface waters (Logares et al., 2020) and down the water column (Giner et al., 2020). We also analyzed other oceanic data sets (V9 amplicons and metagenomes) to provide additional support to the emerging view. Our results revealed a large taxonomic diversity in HF marine assemblages, a contrasted community structure in different depth zones, and the existence of a few dominant and widespread taxa that may be ecologically relevant models whose ecological roles should be examined in further studies. We also highlighted co-occurrence patterns that could represent undescribed and promising prey-predator interactions. Overall, our findings put HFs back in the spotlight in which they were placed decades ago.

2.2. Materials and methods

2.2.1. Samples for sequencing and amplicon processing

Samples were collected in tropical and subtropical oceans during the Malaspina 2010 Circumnavigation Expedition (**Fig. S1**; **Table S1**). Seawater sampling, filtration to keep the pico-sized fraction (0.2-3 µm) and nucleic acids extractions are explained in the original publications for surface samples (Logares et al. 2020) and for vertical profiles (Giner et al., 2020). Here we considered 122 nucleic acid extracts from surface waters (3 m depth; DNA extracts only) and 179 from vertical profiles (88 DNA and 91 RNA extracts). The vertical profiles were collected at 13 stations sampled at surface, Deep Chlorophyll Maximum (DCM) and 2-3 depths at mesopelagic (200-1000 m) and bathypelagic (1000-4000m) zones. For the obtained extracts, the V4 region of the 18S rDNA (~380 bp) was amplified by Polymerase Chain Reaction (PCR) using eukaryotic universal primers (Stoeck et al., 2010) and sequenced with the MiSeq platform (2 x 250 bp).

We trimmed Illumina raw reads to remove amplification primers using cutadapt v1.16 (Martin, 2011) and processed them with DADA2 v1.12.1 (Callahan et al., 2016). We set DADA2's truncLen parameters to 240,210 for the surface data set and 210,190 for both DNA and RNA data sets from the vertical profiles, as these were processed in separate batches and presented different quality profiles. In the three runs we used values of 6,8 for maxEE and pool=TRUE. We merged the obtained tables based on the amplicon sequence variants (ASVs) delineated. We removed chimeric ASVs with the method pooled as implemented in DADA2's function removeBimeraDenovo, and removed ASVs shorter than 300 bp or not present in at least 2 samples. Then, we discarded 14 samples with less than 8000 reads. We added group level taxonomy (in general a formal Class) by comparing ASVs by BLAST (Altschul et al., 1990) to the eukaryotesV4 version 4 database (Obiol et al., 2020), based in the taxonomic outline of Adl et al. (2019). When an ASV was >95% identical in >300 bp alignment to an eukaryotesV4 reference sequence, we assigned the taxonomic group of the reference to the ASV. When identity was between 90-95%, we manually inspected ASVs in phylogenetic trees performed with RAxML-NG (Kozlov et al., 2019) using all ASVs from the same supergroup. For identities lower than 90%, we placed ASVs as *incertae sedis* (InSed). A total of 426 ASVs (0.7% of the reads) remained unclassified even at the supergroup level and some could potentially represent very distinctive novel HF taxa. Yet, we did not consider them, as we lacked any information on their cell identity. We removed ASVs assigned to Metazoa, Charophyta, Embryophyta and nucleomorphs (5.3% of the reads). The final protist table contained 16,629 ASVs and 287 samples (116 from the surface data set, 80 from DNA vertical profiles, and 91 from RNA vertical profiles), with an average of 91,561 (70,064 SD, standard deviation) reads per sample.

2.2.2. Creating a catalogue of oceanic heterotrophic flagellates

We created a subset table of heterotrophic flagellates by keeping ASVs from taxonomic groups possibly including HF morphotypes (listed in **Table S2**). We did not consider dinoflagellates, as their minimal size is about 5 µm, which is larger than the size fraction analyzed here. We placed the selected ASVs into previously published phylogenetic trees (**Fig. S2**) for the following groups: Bicosoecida (del Campo & Massana, 2011), Centrohelida (Shishkin et al., 2018), Cercozoa (Bass et al., 2018), Choanoflagellata (del Campo & Massana, 2011), Chrysophyceae (del Campo & Massana, 2011), Dictyochophyceae (Sekiguchi et al., 2002), Marine Stramenopiles (MAST; Massana et al., 2014), Picozoa (Moreira & López-García, 2014), and Telonemia (Shalchian-Tabrizi et al., 2007). For each group, we retrieved complete 18S rDNA sequences using the NCBI accession numbers shown in the trees, aligned them using mafft v7.402 (Katoh & Standley, 2013), added the ASV sequences to the alignment with mafft, and constructed maximum likelihood trees with RAxML-NG (Kozlov et al., 2019). Group-specific phylogenetic trees served to refine taxonomic placements and to identify ASVs close to plastidic species in groups encompassing colorless and pigmented taxa. Thus, we removed 137 Dictyochophyceae and 7 Cercozoa ASVs that were >97% identical to species known to have chloroplasts, while we did not detect any ASV closely related to a plastidic Chrysophyceae. We also visually inspected the alignments and removed 133 ASVs that were obvious chimeras or partial sequences.

We filtered the obtained preliminary HF ASV table to keep samples with more than 500 reads (we removed 8 samples), and ASVs that were present in at least 2 samples. The final HF table contained 1642 ASVs from 25 taxonomic groups, 279 samples (115 from surface, 73 from DNA vertical profiles, and 91 from RNA vertical profiles), and an average of 23,906 (34,319 SD) reads per sample. With this table we created a catalogue of oceanic HF ASVs (**Table S3**). We kept RNA samples ("V4 RNA") only for comparison purposes and used the DNA data set ("V4 DNA"), which contained 188 samples from both surface and vertical profiles, for the rest of the analyses (**Table S1**). For each DNA ASV, we performed a BLAST search against NCBI nt database excluding environmental sequences to find its closest cultured match (CCM) and added a name to its unique identifier. If identity to the CCM was >97%, we used species name and percentage of identity; otherwise, we used group name and 'sp' followed by a rank abundance index. For example, the first Chrysophyceae ASV in terms of overall abundance without a close CCM was named 'Chrysophyceae-sp1'.

2.2.3. A web application to map global microbial biogeographies

We developed an online visualization tool for marine microbial biogeography. MicroMap is a web application that provides an interface for submitting queries to an ASV database and creating global maps of ASV abundances. MicroMap incorporates the 18S rDNA protist database used here and the 16S rDNA surface data set published in Logares et al. (2020). In addition, MicroMap can incorporate a custom-made database constructed from the users' own ASV table and sample table. Queries can be done with 16S/18S rDNA sequences, with named taxonomic groups, or with known sequence IDs from the database. In the query with a DNA sequence, the application runs a BLAST search using the parameters configured in the home page. It takes the matching ASVs and collates them into a results file used to interactively display relative abundances in four water zones and abundance spectra plots. In queries based on taxonomic names, the application looks for ASVs classified with that name and treats all of them as a single result. In queries based on a sequence ID, the tool displays the results of that ASV. MicroMap is implemented in Laravel 5.8 using PHP 7.2, and the front end of the application runs in VueJS, with R scripts providing additional functionality for creating plots. The interactive map uses the Datamaps library implemented in the D3 framework. It is hosted at the ICM-CSIC on a Dell server, with 320 GB of RAM and two 2 Intel Xeon E5-2650v4 processors, and is accessible at https://micromap.icm.csic.es/.

2.2.4. Other data sets used for comparison and co-occurrence analyses

We analyzed two additional molecular data sets to support the HF diversity and distribution results obtained by 18S-V4 amplicons (**Table S1**): (1) 70 V9-18S rDNA amplicon sequencing samples (0.8-5 µm size fraction; "TARA V9") from surface and DCM waters from TARA Oceans (Callahan, 2017; de Vargas et al., 2015), and (2) 66 metagenomes ("mTags") and 34 V9-18S rDNA amplicon sequencing samples ("V9 DNA") from vertical profiles from the pico fraction (0.2-3 µm) from Malaspina (Obiol et al., 2020). For co-occurrence analyses with prokaryotes, we used the V4-V5 16S rDNA data set ("Prokaryotes") from Logares et al. (2020), which comprised 115 surface samples from the same DNA extracts as the 18S rDNA data set. All samples used are listed in **Table S1**.

2.2.5. Statistical analyses

We performed general processing of ASV tables using R v4.0.5 (R Core Team, 2020) and packages tidyverse v1.3.1 (Wickham et al., 2019) and phyloseq v1.34.0 (McMurdie & Holmes, 2013). We conducted principal coordinate analysis (PCoA), non-metric multidimensional scaling (NMDS) and permutational multivariate analysis of variance (PERMANOVA) with vegan v2.5.7 (Oksanen et al., 2019) using Bray-Curtis dissimilarity matrices computed with the package DivNet v0.3.6 (Willis & Martin, 2020). We performed differential abundance tests of HF and prokaryotic ASVs with scaled environmental variables with the package corncob v0.1.0 (Martin et al., 2020). To describe the distribution pattern of surface HF ASVs, we used the ratio SD/mean of the relative abundance over samples; an ASV was considered "patchy" when this was higher than 2.3. When the ratio was lower, an ASV was considered "warm" or "cold" if it had a significant estimate across temperature in the corncob test (estimate > 0.5, p < 0.01) and "equal" when the test was not significant. We carried out co-occurrence network analyses using sparCC (Friedman & Alm, 2012) as implemented in fastSpar v0.0.7 (Watts et al., 2019). We used EnDED (Deutschmann et al., 2019) as implemented in Deutschmann et al. (2021) on the obtained correlation matrix to remove associations driven by environmental variables (temperature, conductivity, dissolved oxygen and fluorescence). We carried out network processing with tidygraph v1.2.0 (Pedersen, 2020b) and ggraph v2.0.5 (Pedersen, 2020a). We performed DivNet, corncob and sparCC analyses on absolute counts tables, as these packages already take into consideration the compositional nature of microbiome data (Gloor et al., 2017).

2.3. Results

2.3.1. General overview of HF assemblages in the ocean

The processing of V4 18S rDNA reads derived from DNA extracts taken at the surface and from vertical profiles in Malaspina (**Fig. S1**) yielded 1030 ASVs representing HF taxa within 24 taxonomic groups (**Table S3**, **Fig. S2**). The contribution of HF reads to the total picoeukaryotic signal varied between water column zones (**Fig. 1A**). Most surface and DCM samples showed similar contributions (median values of 11% and 5%, respectively), while a much larger dispersion was found in deeper samples, with median values of 3% in the mesopelagic and as large as 28% in the bathypelagic layer. In all depth zones, Alveolates (mainly Marine Al-

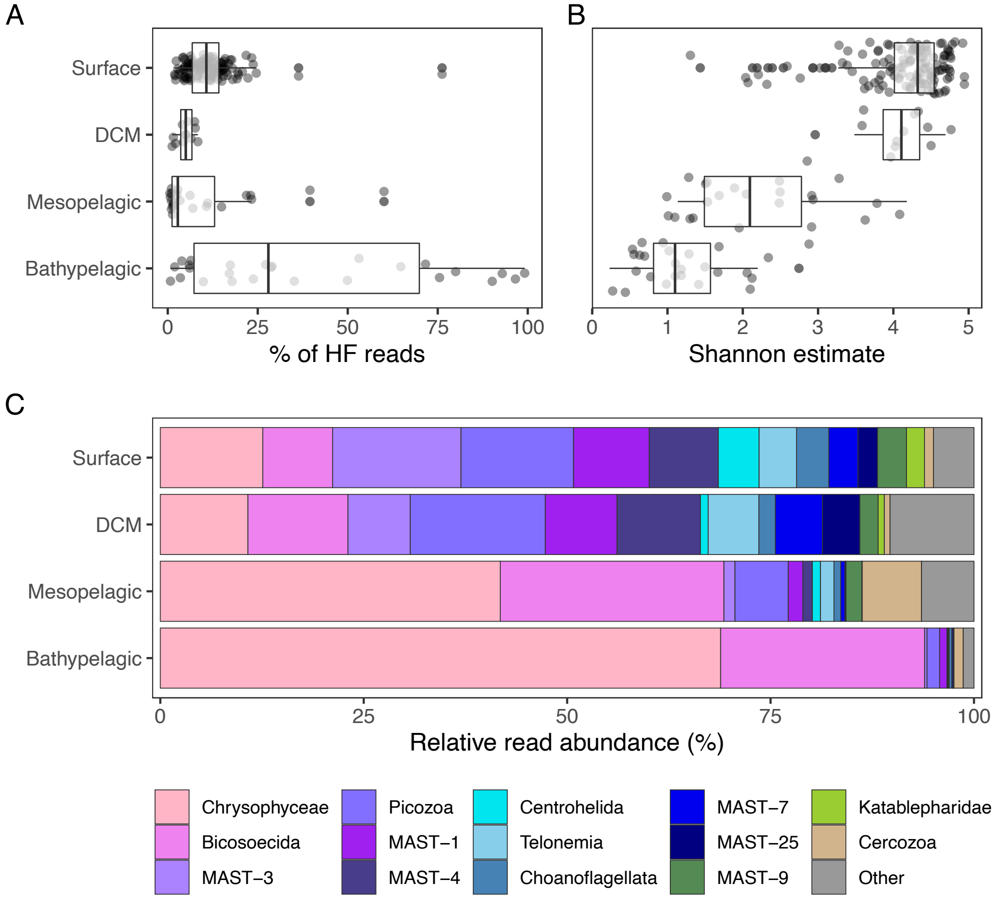


Figure 1. Contribution of HFs in four depth zones of the ocean water column derived from 188 Malaspina samples (V4 DNA data set). (**A**) Percentage of reads affiliating to HF taxa with respect to the total picoeukaryotic signal. (**B**) Shannon alpha-diversity indices of HF assemblages using the corresponding ASVs. (**C**) Relative read abundance of the main taxonomic groups of HFs.

veolates [MALV] clades) and other Rhizarians (mainly radiolarians) represented 58-89% of the protist signal (**Fig. S3**). Regarding Shannon estimates of alpha diversity based on HF ASVs, there was a clear decreasing trend with depth, with very different Shannon values at surface and the bathypelagic (median of 4.3 and 1.1, respectively) (**Fig. 1B**). The taxonomic affiliation of ASVs allowed delineation of the contribution of different taxonomic groups to HF assemblages. Surface and DCM samples were composed of several groups with similar overall relative abundances (MAST-3, MAST-1, MAST-4, Picozoa, Chrysophyceae, and Bicosoecida; 9-16% on average each; **Fig. 1C**). At the sample level, these were relatively stable across the ocean, with Chrysophyceae and Bicosoecida deviating from the general trend (**Fig. S4**). In mesopelagic and bathypelagic zones, samples were overwhelmingly dominated by Chrysophyceae and Bicosoecida (25-68% on average each; **Fig. 1C**) and showed distinct profiles in each station (**Fig. S5**).

We next compared the HF taxonomic composition derived from DNA samples (**Fig. 1C**) with that obtained from other oceanic data sets (**Table S1**). The RNA samples from the Malaspina survey yielded a highly similar HF composition in the four surveyed zones of the water column (**Fig. 2**), and ASVs that were detected in both DNA and RNA extracts represented 94-98% of the reads (**Fig. S6**). In general, ASVs unique to the RNA data set did not add new diversity, with notable exceptions within Choanoflagellata, InSedMAST (putative novel MAST clades), and MAST-3 (**Table S3**). We also processed two additional Malaspina data sets derived from DNA extracts: V9-18S amplicons and metagenomes (mTags). In surface samples, the relative read abundances of taxonomic groups were generally similar to the V4 DNA data set, except for a lower presence of Bicosoecida and the detection of Euglenozoa (Diplonemea and Kinetoplastida) at moderate abundances (**Fig. 2**). In aphotic samples, the presence of Euglenozoa increased substantially, to the point that Diplonemea was the most abundant group in mesopelagic metagenomes (34% on average). Finally, we analyzed the HF composition derived from V9-18S amplicons in photic waters of the TARA oceans survey (**Fig. 2**). Despite different sampling locations and size fractions analyzed (0.2-3 µm in Malaspina, and 0.8-5 µm in TARA), groups like MAST-1, -3, -4, and Picozoa showed similar relative abundances in both data sets. The most noticeable difference was the higher presence of Diplonemea in the TARA oceans survey (12% at surface and 32% in DCM). This was higher than that reported in the V9 and mTags Malaspina data sets and could be explained by the size fraction analyzed, as diplonemid cell sizes tend to exceed 3 µm.

2.3.2. Environmental drivers to HF community structure

We performed PERMANOVA analyses on the HF assemblages defined by ASVs to identify the environmental factors explaining the variance in their community structure (**Table S4**). In surface samples, temperature (16%) and the sampled ocean (13%) were the main factors

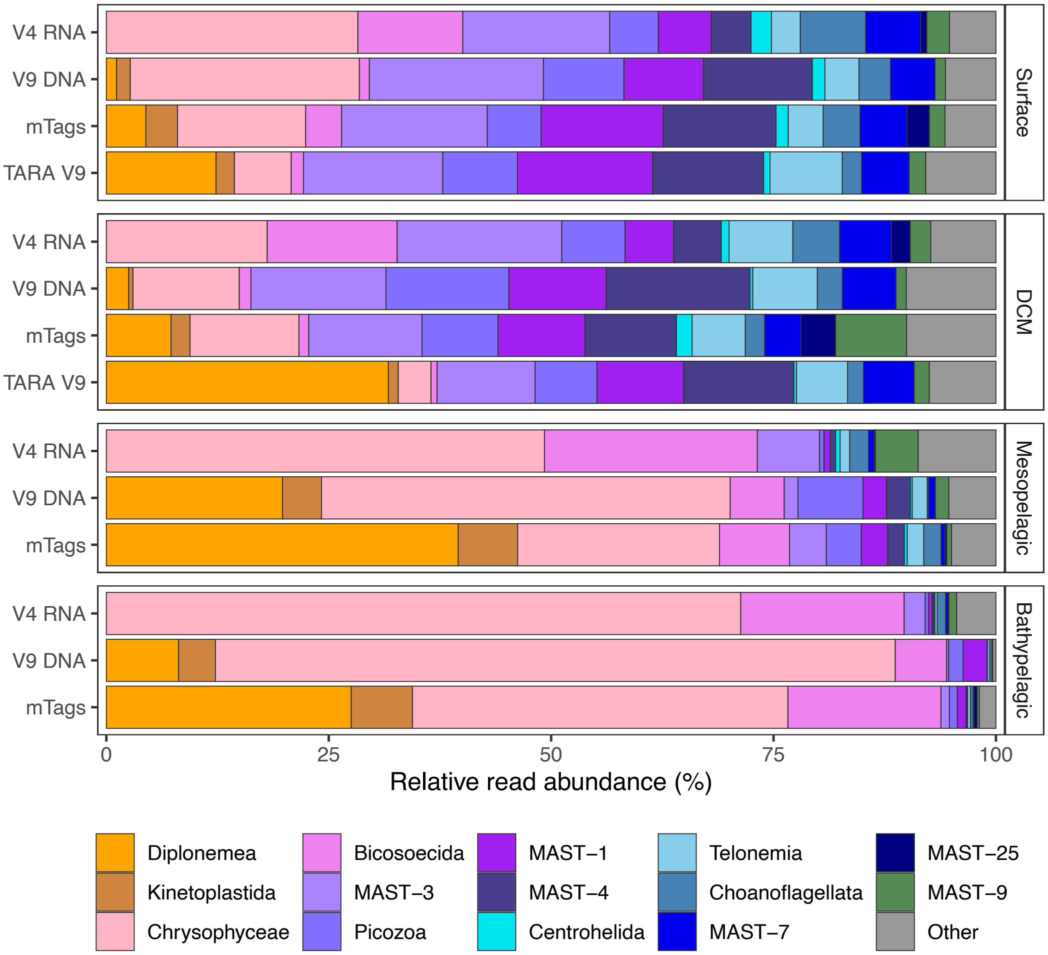


Figure 2. Relative read abundance of the main taxonomic groups of HFs in four depth zones of the ocean as seen by different oceanic data sets. V4 RNA: Malaspina 18S-V4 amplicons from RNA extracts; V9 DNA: Malaspina 18S-V9 amplicons from DNA extracts; mTags: Malaspina 18S-V4 metagenomic tags; TARA V9: TARA Oceans 18S-V9 amplicons from DNA extracts. See **Table S1** for details of the number of samples and the fraction analyzed in each data set. Dictyochophyceae and Cercozoa (that had pigmented taxa) and InSedMAST (only defined in the V4 tree) were not included in this comparison.

driving HF community variation, while in vertical profiles, the depth zone accounted for 22% of variance, and the sampled ocean for 19%. Within aphotic samples (mesopelagic and bathypelagic), the majority of the variance was explained by water masses (34%).

To further investigate how water temperature influenced HF assemblages in surface samples, we built a PCoA using Bray-Curtis dissimilarities among HF assemblages (**Fig. 3A**). Surface samples were distributed in the PCoA plot along the first axis (21% of variance explained) following a temperature gradient, thus highlighting the influence of this variable on the HF community structure. The effect of the sampled ocean was also evident when plotting the same dissimilarity matrix against the matrix of geographic distances among samples (**Fig. 3B**). Thus, geographically (and temporally) closer samples (<1000 km) were more similar, whereas dissimilarities increased up to distances of about 2000 km and remained fairly constant with samples that were farther apart.

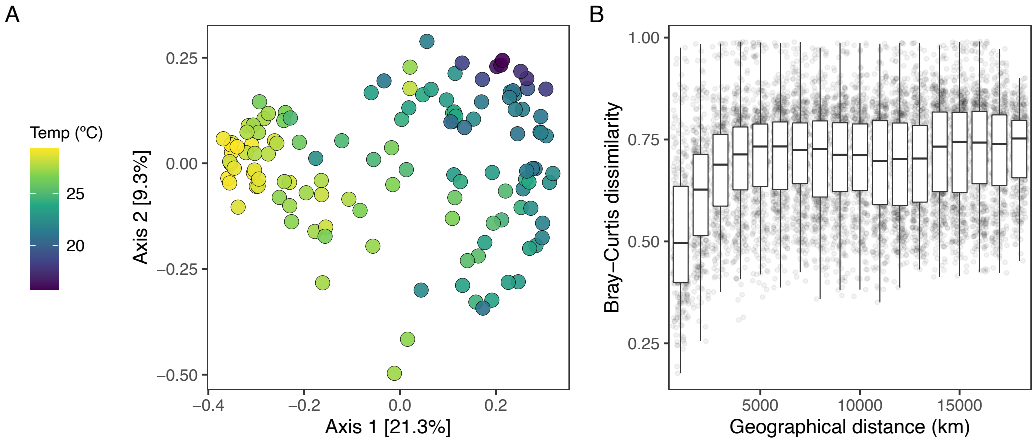


Figure 3. Variation of HF community structure in surface samples. (**A**) Principal Coordinate Analysis of surface samples colored by water temperature based on Bray-Curtis dissimilarities among samples. (**B**) Plot of Bray-Curtis dissimilarity among samples against geographical distances.

2.3.3. The dominant HFs in the ocean

We identified which HF ASVs dominated in the different depth zones by averaging their relative read abundances in all samples from that zone. The above-mentioned pattern of decreasing diversity with increasing depth was clearly reflected by the number of ASVs that explained the majority of reads in each depth zone (**Table S5**). Thus, while 52 ASVs accounted for 60% of the signal at the surface, only 6 did so in the mesopelagic, and 3 in the bathypelagic. In surface samples, the ASVs contributing to HF assemblages appeared with a reasonable evenness (**Table 1**), and the 10 most abundant surface ASVs accounted for 2.0-3.5% of the signal each: 1 ASV affiliated to Centrohelida (very distant to any cultured representative); 5 to MASTs (clades -1, -3 and -4); 1 to Chrysophyceae (*Spumella* sp. IOW86); 2 to Bicosoecida (*Cafeteria burkhardae* and *Caecitellus pseudoparvulus*); and 1 to Picozoa.

Some of these ASVs dominated at the DCM as well, where the most abundant ASVs were related to *Picomonas judraskeda* (7.4%) and *Caecitellus paraparvulus* (6.2%; **Table 1**). In deep waters, just 2 ASVs, namely *Spumella* IOW86 and *C. burkhardae*, dominated the HF assem-

Table 1. Dominant HFs in the four depth zones of the ocean. The 10 ASVs with highest relative read abundance in each depth zone are displayed, including their mean relative read abundance (%), SD and prevalence (percentage of samples where it has been detected). Numbers at the beginning of ASV names correspond to their unique identifiers in the data set.

ASV	Mean (%)	SD	Prevalence (%)
Surface			
54/ Centrohelida-sp1	3.5	3.8	97.6
75/ MAST-1D-sp1	3.3	2.2	98.4
1/ *Spumella* IOW86 (100%)	3.3	9.8	78.0
4/ *Cafeteria burkhardae* (100%)	3.2	7.6	80.3
31/ MAST-4A-sp1	3.1	3.0	93.7
51/ Picozoa-sp1	2.7	2.5	98.4
76/ MAST-1C-sp1	2.3	1.6	99.2
36/ *Caecitellus pseudoparvulus* (100%)	2.2	7.4	54.3
110/ MAST-4C-sp1	2.0	2.5	74.0
182/ MAST-3A-sp2	2.0	2.4	97.6
DCM			
138/ *Picomonas judraskeda* (97.4%)	7.4	5.3	100.0
9/ *Caecitellus paraparvulus* (97.7%)	6.2	14.3	83.3
1/ *Spumella* IOW86 (100%)	4.6	6.6	91.7
31/ MAST-4A-sp1	4.0	2.5	100.0
4/ *Cafeteria burkhardae* (100%)	3.1	8.0	75.0
75/ MAST-1D-sp1	2.8	1.5	100.0
312/ MAST-25-sp1	2.7	1.5	100.0
383/ Picozoa-sp2	2.4	1.6	100.0
12/ *Caecitellus paraparvulus* (100%)	2.2	4.9	50.0
763/ MOCH-4-sp1	1.9	1.8	91.7
Mesopelagic			
1/ *Spumella* IOW86 (100%)	29.3	31.7	100.0
4/ *Cafeteria burkhardae* (100%)	10.1	18.5	91.3
12/ *Caecitellus paraparvulus* (100%)	8.8	15.5	87.0
9/ *Caecitellus paraparvulus* (97.7%)	7.1	16.8	91.3
11/ *Spumella* IOW86 (99.5%)	4.3	12.7	60.9
138/ *Picomonas judraskeda* (97.4%)	3.9	6.1	95.7
16/ Chrysophyceae-sp1	3.8	7.8	82.6
130/ *Helkesimastix* sp. (99.7%)	2.0	9.4	17.4
123/ *Helkesimastix* sp. (99.7%)	1.5	6.8	17.4
1120/ Cercozoa-sp1	1.1	1.5	91.3
Bathypelagic			
1/ *Spumella* IOW86 (100%)	38.7	36.8	100.0
11/ Spumella IOW86 (99.5%)	17.6	32.7	76.9
4/ *Cafeteria burkhardae* (100%)	15.2	23.1	96.2
16/ Chrysophyceae-sp1	6.7	10.6	80.8
12/ *Caecitellus paraparvulus* (100%)	4.2	9.5	84.6
9/ *Caecitellus paraparvulus* (97.7%)	3.7	10.6	76.9
141/ *Paraphysomonas bandaiensis* (100%)	1.8	8.3	11.5
36/ *Caecitellus pseudoparvulus* (100%)	1.4	6.3	26.9
29/ *Planomonas micra* (100%)	1.0	2.1	65.4
686/ *Spumella* IOW86 (99.7%)	0.8	0.8	92.3

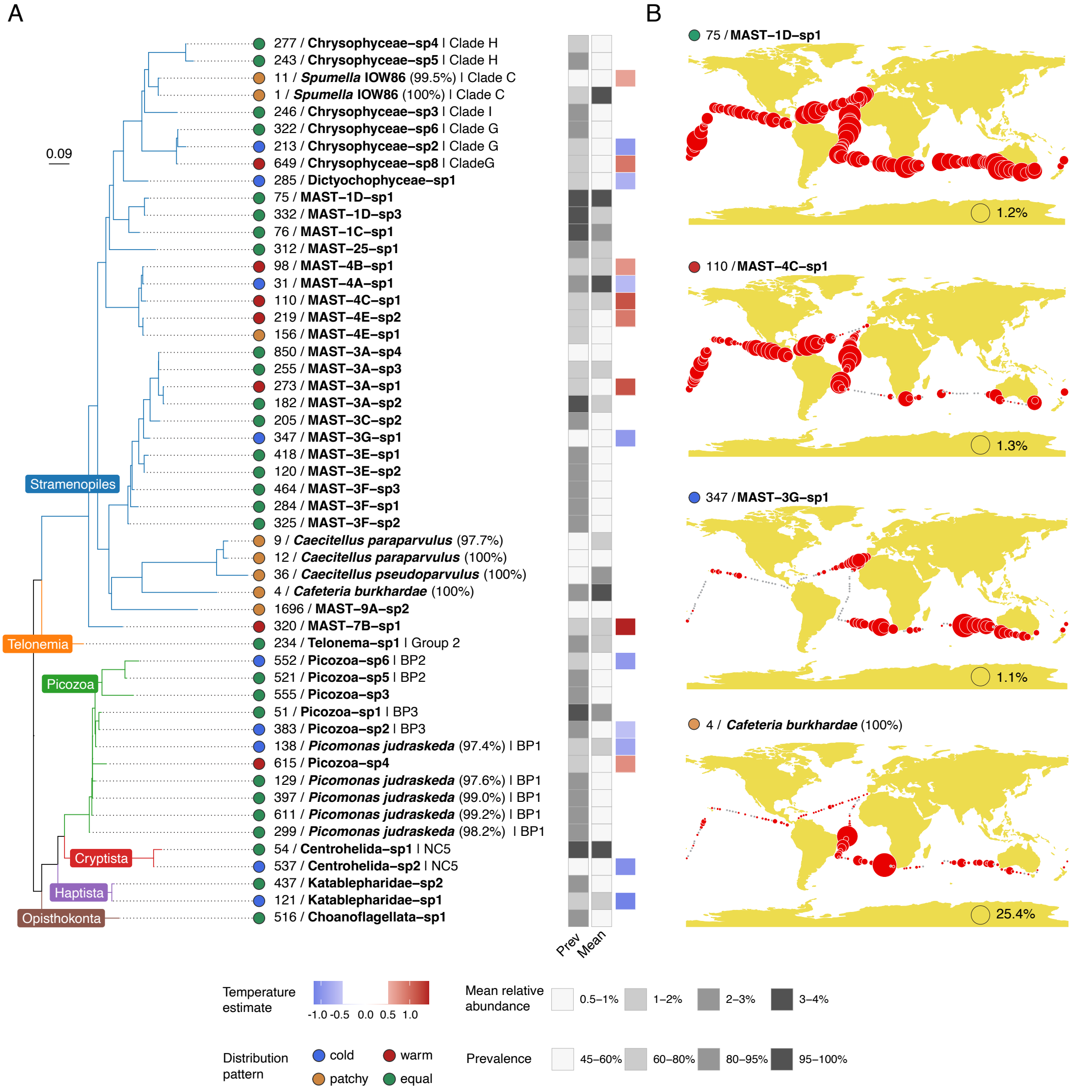


Figure 4. The 52 dominant HF ASVs in surface waters of the ocean. (**A**) Phylogenetic tree and broad distributional features of each dominant ASV. Prevalence and mean relative abundance are displayed, as well as their estimates of a differential abundance test with temperature. Taxa are then colored according to their distribution pattern. (**B**) Biogeography maps for four dominant HF ASVs exhibiting differentiated distribution patterns obtained by the MicroMap web application (https://micromap.icm.csic.es/). The area of red circles is proportional to the relative read abundance of each ASV in the whole picoeukaryotic pool (maximum value is displayed in each panel), while gray circles indicate absence. Numbers at the beginning of ASV names correspond to their unique identifiers in the data set.

blages and together accounted for 39.5% of the signal in mesopelagic and 53.9% in bathypelagic zones (**Table 1**). Looking at the list of dominant ASVs, it was evident that those related to cultured representatives increased with depth. A detailed list of the dominant ASVs in the four depth zones can be found in **Table S6**.

We then focused on the biogeography and distribution patterns of the 52 ASVs that explained 60% of the reads at the surface ocean (**Fig. 4**). These ASVs were found at comparable abundances in the RNA data set (**Fig. S7**) and also detected in the TARA survey (**Table S7**). Eleven ASVs were close (>97% identity) to cultured species within Chrysophyceae, Bicosoecida and Picozoa (albeit the reference culture is no longer available), 22 belonged to different environmental MAST clades, and the rest formed new taxa within groups known to contain cultured representatives (**Fig. 4A**). The majority of these dominant ASVs (31 out of 52) had a prevalence >75%. We classified each ASV into one of four distribution categories by analyzing (1) the ratio SD/Mean of their relative read abundances in all samples, (2) the differential abundance test across environmental factors, and (3) the heatmap clustering ASVs based on their normalized reads distribution (**Fig. S8**). ASVs were labeled as “equal” when they displayed relatively constant abundances across samples; “warm” or “cold” when showing preference for warmer or colder waters; and “patchy” when they appeared with peaks of high abundance in a few samples. More than half of the dominant surface ASVs (28) were labelled “equal”, 16 were associated with temperature (7 “warm” and 9 “cold”), and 8 were “patchy” (**Fig. 4A**). All Bicosoecida ASVs fell into this latter category, as well as *Spumella* sp. and MAST-4E-sp1.

Unsurprisingly, patchy ASVs tended to have the lowest prevalence (minimum of 47%). Analyzing the dominant ASVs in a phylogenetic tree (**Fig. 4A**), we detected closely related ones showing contrasted temperature distributions: MAST-4B-sp1 (“warm”) and MAST-4A-sp1 (“cold”) had 2 bp difference; and MAST-3A-sp1 (“warm”) and MAST-3A-sp2 (“equal”) had a single mismatch. We used the MicroMap web application developed here to construct global maps of the distribution of the dominant ASVs (**Fig. 4B** shows an example of each distribution pattern in surface samples). This provided a visual confirmation of the reported distribution characteristics: “cold” or “warm” ASVs occupied different regions of the cruise track (**Fig. S1**), “equal” ASVs showed a widespread distribution, and “patchy” ASVs displayed isolated peaks of high abundance.

2.3.4. Co-occurrence of HF and prokaryotes at the surface ocean

We performed a co-occurrence analysis between ASVs of HFs and prokaryotes retrieved from 113 surface samples of the Malaspina data set. After removing weak correlations (< 0.3) and putative indirect associations driven by environmental factors, the obtained network contained 479 HF and 462 prokaryotic nodes (ASVs) and a total of 26,835 edges (correla-

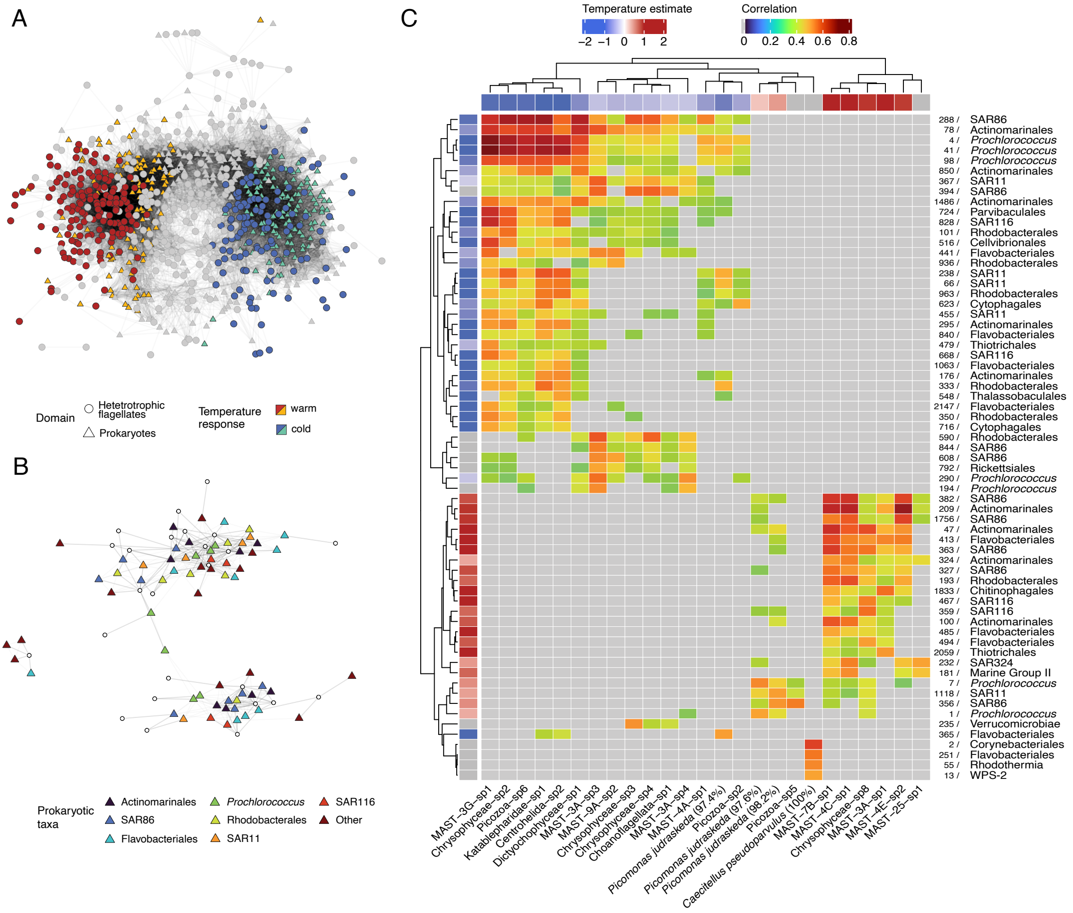


Figure 5. Co-occurrence patterns between HF ASVs and prokaryotic ASVs at the surface of the ocean. (**A**) Co-occurrrence network of HF and prokaryotic taxa. Each point represents an ASV, and these are colored by their temperature preference according to a differential abundance test (gray is a non-significant response to temperature). (**B**) Co-occurrence subnetwork of the dominant HF ASVs and prokaryotes, the latter colored by taxonomy. From the 52 dominant HF, only 25 display significant correlations and are thus shown in the subnetwork. (**C**) Heatmap displaying the correlations between dominant HFs (columns) and prokaryotes (rows); the estimates of a differential abundance test with temperature are displayed for each ASV.

tions). The topology of the network displayed 2 clearly differentiated clusters driven by temperature, as clearly seen when coloring the nodes by thermal preference (**Fig. 5A**), leaving a still substantial number of nonresponsive ASVs. This two-cluster scheme was maintained in the subnetwork formed by the dominant 52 HF ASVs and their correlations with prokaryotes (**Fig. 5B**). Most HF nodes were connected to several prokaryotic nodes, and these connections did not follow a specific taxonomic pattern. We then constructed a heatmap to visualize the specific co-occurrences, which clearly showed two clusters of HF nodes with opposite thermal preferences. These, in turn, were each associated with a cluster of prokaryotic nodes sharing the same thermal preference (**Fig. 5C**).

When looking at the composition of the correlating prokaryotes, the two thermal clusters included ASVs from similar groups: SAR11, SAR86, SAR116, Actinomarinales and *Prochlorococcus*. For the latter two groups, the ASVs shown in the networks accounted for most of their overall environmental signal. ASV_41 and ASV_4, both related to *Prochlorococcus*, showed the highest correlations with HF ASVs in the cluster with preference for higher temperatures, while in the cluster with preference for colder waters ASV_209, from *Candidatus* Actinomarina, and ASV_382, from the SAR86 clade, were the taxa with the strongest correlations. Apart from these two main thermal clusters, some other correlations appeared. Thus, *C. pseudoparvulus*, which had a patchy distribution in the ocean, was nonetheless strongly correlated to ASV_2 (*Rhodococcus*), ASV_55 (*Rubrivirga*) and ASV_13 (WPS-2). These associations appeared isolated from the rest of the nodes in the co-occurrence network (**Fig. 5B**).

2.4. Discussion

One of the main issues that prevented us from obtaining a detailed assessment on the ecology of heterotrophic flagellates, a key microbial component in marine ecosystems, is the lack of cultured strains that effectively represent the dominant HF species in the ocean. Besides the prevalent culturing bias, this was due to our general ignorance of which HF taxa dominate in the marine plankton, and whether or not this dominance is similar over oceanic biomes. In this study, we investigated the presence of marker genes of HF taxa over the tropical and subtropical ocean using mostly Malaspina data (Giner et al., 2020; Logares et al., 2020; Obiol et al., 2020) together with Tara Oceans data (de Vargas et al., 2015). While abundance of ribosomal marker genes may not always correlate directly with abundance of cells due to copy number variations (Zhu et al., 2005), these could be a good proxy for the *in situ* abundance of the HFs analyzed here, as small-sized cells are hypothesized to contain a low and constrained number of rDNA copies (Rodríguez-Martínez et al., 2009; Zhu et al., 2005). With this, we show the prevalence of particular taxonomic groups over horizontal and vertical scales, and identify a subset of dominant and widespread HF taxa in different depth zones of the water column of the ocean.

In surface waters, the dominant HFs resolved into two clearly differentiated groups with distinct distributions. The first group formed a constant "core" community of uncultured taxa belonging to MAST clades (mostly -1, -3, and -4), Picozoa, and Chrysophyceae (uncultured clades G, H, and I) that were found in most samples, some of them with a preference for colder or warmer waters. Previous microscopic counts confirmed the relatively homogeneous cell abundance in Malaspina surface samples of relevant clades like MAST-4, MAST-1C, and

MAST-7, which averaged between 10 and 50 cells ml^{-1} (Mangot et al., 2018). Moreover, these core surface taxa were scarce in the deep and dark ocean. A recent study using single amplified genomes detected rhodopsin genes in most MAST species investigated (Labarre et al., 2021), which could partially explain their success in photic waters and their virtual absence in the dark ocean. Although no cultured representative exists for the dominant MASTs, there are already partial genomes for 10 of them (Labarre et al., 2021), and this will be an invaluable resource for further research. The high number of picozoans among the dominant surface ASVs suggests that these likely play a significant role in the ecosystem, a role that is yet to be characterized (Burki et al., 2020). The only described picozoan species, *P. judraskeda*, originally isolated from coastal waters (Seenivasan et al., 2013) but today lost, was not detected in the open ocean. Finally, the most widespread and abundant ASV in surface waters was Centrohelida-sp1. This ASV, together with many others, formed an environmental clade phylogenetically distant to the described species of centrohelids (Shishkin et al., 2018). The prevalence of Centrohelida-sp1 in the picoeukaryotic fraction in surface waters hints at the existence of undescribed centrohelids that might be smaller than those reported (i.e., *Oxnerella micra*; Cavalier-Smith and Chao, 2012) and could be important bacterivores.

Parallel to the described "core" community, surface waters contained a second set of patchy, unevenly distributed HF taxa from the groups Chrysophyceae (genus *Spumella* in clade C), and Bicosoecida (species *C. burkhardae*, and *C. paraparvulus*). These ASVs were markedly dominant in deeper aphotic zones and coincided with a clear decrease in HF alpha diversity, a pattern already seen in studies targeting all microbial eukaryotes (Giner et al., 2020; Schnetzer et al., 2011). The low diversity and high dominance of these species, known to have a high tolerance to pressure (Živaljić et al., 2018) and to be well adapted to a particle-associated lifestyle (Jeuck & Arndt, 2013), may be explained by the fact that microbial life in the nutrient-poor deep ocean greatly relies on the particulate organic matter flux from upper layers (Nagata et al., 2010). So, the patchy distribution at the surface and deep ocean of *Spumella*, *Cafeteria* and *Caecitellus* spp. may be due to their living in or on particles as nutrient-rich hotspots. The fact that we detected them in the 0.2-3 µm size fraction could be due to the fragility of marine particles, which could break during filtration (Bochdansky et al., 2017). The particle-attached existence of these taxa could also explain their relative ease of culturing compared to free-living species, since standard culturing media (i.e. a rice grain with seawater) could somehow resemble nutrient-rich particles (del Campo et al., 2013). In fact, while in the deep ocean the majority of dominant ASVs here detected seem to be already in culture, no cultured representative exists for the dominant surface ASVs besides these few patchy species, highlighting the current bias that exists in protistan knowledge (Keeling & Burki, 2019).

Together with Chrysophyceae and Bicosoecida, Diplonemea (and to a lower extent Kinetoplastida), also appeared to be major contributors to deep HF assemblages. This was obvious when analyzing V9 amplicons and metagenomes, but unfortunately these euglenozoans did not appear in the V4-18S data set, as the combination of longer V4 regions (Pernice et al., 2016) and several mismatches with universal primers (Vaulot et al., 2021) prevent their correct amplification. Therefore, these euglenozoans do not appear in the catalogue of ASVs, a possibly minor issue for surface assemblages but an important limitation for the deep survey. The importance of deep diplonemids agrees with previous reports (Flegontova et al., 2016; Lara et al., 2009; Schoenle et al., 2021), albeit it has been also proposed that read abundances could overestimate cell abundances, given the tendency of Diplonemea to display high rDNA copy numbers (Mukherjee et al., 2020).

Environmental parameters measured during the Malaspina cruise could explain part of the variation displayed by HF assemblages along horizontal and vertical scales. Increasing depth in the water column resulted in a clear drop of HF diversity and the disappearance of many surface taxonomic groups, with deep ocean assemblages being dominated by only a few of them. In vertical profiles, the main explanatory variables were depth in the water column and the sampled ocean, together with water masses when aphotic samples were processed separately, as reported before for whole picoeukaryotic communities (Giner et al., 2020). In surface waters, we also detected the effect of the sampled ocean, but temperature was the primary environmental factor shaping sunlit HF communities. In global studies with strong latitudinal gradients, temperature has a marked effect in most planktonic groups (Ibarbalz et al., 2019), and HF assemblages are not an exception (Azovsky et al., 2016; Patterson & Lee, 2000). The temperature effect was also detected in studies targeting specific HF taxa (Boenigk et al., 2006; Flegontova et al., 2020; Latorre et al., 2021; Rodríguez-Martínez et al., 2013). Despite samples analyzed here having a relatively narrow temperature range (16-29ºC), we could detect a temperature effect on HF assemblages, highlighting the possible alterations that these could face by the expected increase of sea surface temperature due to global warming (Hutchins & Fu, 2017). The temperature gradient reported in the Malaspina survey was dominated by latitudinal changes and not by putative seasonality of the sampled sites, as could be inferred in **Fig. S1** (i.e., samples from the same latitude that were collected several months apart share highly similar temperature values). Nonetheless, in our work, environmental variables were incomplete predictors of community structure (at surface, they only predicted 36% of the variance). Biotic factors, such as the community composition of prokaryotic assemblages, could be playing an important role in shaping HF assemblages.

To reveal putative biotic interactions, we assessed the co-occurrence between HF and prokaryotic taxa in association networks. We did not aim to use co-occurrence as a direct proxy of interaction (Blanchet et al., 2020), but to identify potentially interesting biotic relationships that would then need to be experimentally validated (Carr et al., 2019). Even after removing associations that could arise from shared environmental preferences (Deutschmann et al., 2021; Röttjers & Faust, 2018), HF and prokaryotic ASVs were placed in two differentiated clusters formed by abundant taxa with different thermal preferences, a phenomenon already seen in previous studies (Fuhrman et al., 2008; Lima-Mendez et al., 2015; Pommier et al., 2007). At a broad scale, specific HF did not show preferential correlations with specific prokaryotic taxa, and this agrees with the perception that cell size is the main factor in prey vulnerability (Jürgens & Massana, 2008). However, some of the correlations displayed by taxa not belonging to the main "warm" and "cold" clusters could also suggest prey preference by some HFs. In fact, it has been shown that feeding natural HF communities with different bacterial strains can select distinct assemblages in freshwater manipulations (Šimek et al., 2018), and that closely related grazers can differently impact bacterial communities (Glücksman et al., 2010). Our results thus represent an opportunity to further study particular predator-prey interactions in new experimental scenarios.

Having identified the putatively most relevant HFs in the ocean, a renewed effort to culture them should be made, as these could be developed into new model organisms to be used to gain a better understanding of the role of HFs in the ecosystem. Single cell genomics is also a promising alternative to capture the genomes of these dominant species (Labarre et al., 2021), which can then be used in comparative genomics or to interpret complex metagenomic and metatranscriptomic data sets. Overall, our work paves the road for future studies on marine HFs, highlighting key species that may be playing crucial roles in the plankton.

2.5. Acknowledgements

This research was supported by the Spanish Ministry of Science and Innovation projects Malaspina-2010 (CSD2008–00077), ALLFLAGS (CTM2016-75083-R) and the "Severo Ochoa Centre of Excellence" accreditation (CEX2019-000928-S), and the European Union project SINGEK (H2020-MSCA-ITN-2015-675752). AO was supported by a Spanish FPI grant. We thank all scientists and crew that participated in the Malaspina 2010 expedition, Ramiro Logares for the initial stages of the MicroMap tool, and Ina M. Deutschmann for useful discussions on microbial networks. Bioinformatic analyses were performed at the Marbits platform (ICM-CSIC; https://marbits.icm.csic.es).

2.6. Data availability statement

Raw data used in this study can be found at the European Nucleotide Archive (http://www.ebi.ac.uk/ena) with accession numbers PRJEB23913 and PRJEB25224 (surface data sets for 18S and 16S, respectively; Logares et al., 2020), and PRJEB23771 (vertical profiles; Giner et al., 2020). TARA Oceans (Callahan, 2017; de Vargas et al., 2015) and Malaspina V9/mTags tables (Obiol et al., 2020) can be found at Zenodo with DOIs 10.5281/zenodo.581694 and 10.5281/zenodo.3629394, respectively. All tables and code used for data processing and analyses can be found at https://github.com/aleixop/Malaspina_HF.

2.7. References

Adl, S. M., Bass, D., Lane, C. E. et al. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology*, 66(1), 4–119. https://doi.org/10.1111/jeu.12691

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. https://doi.org/10.1016/S0022-2836(05)80360-2

Azam, F., Fenchel, T., Field, J. et al. (1983). The Ecological Role of Water-Column Microbes in the Sea. *Marine Ecology Progress Series*, 10, 257–263. https://doi.org/10.3354/meps010257

Azovsky, A. I., Tikhonenkov, D. V., & Mazei, Y. A. (2016). An Estimation of the Global Diversity and Distribution of the Smallest Eukaryotes: Biogeography of Marine Benthic Heterotrophic Flagellates. *Protist*, 167(5), 411–424. https://doi.org/10.1016/j.protis.2016.07.001

Bass, D., Tikhonenkov, D. V., Foster, R. et al. (2018). Rhizarian 'Novel Clade 10' Revealed as Abundant and Diverse Planktonic and Terrestrial Flagellates, including Aquavolon n. gen. *Journal of Eukaryotic Microbiology*, 65(6), 828–842. https://doi.org/10.1111/jeu.12524

Blanchet, F. G., Cazelles, K., & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, 23(7), 1050–1063. https://doi.org/10.1111/ele.13525

Bochdansky, A. B., Clouse, M. A., & Herndl, G. J. (2017). Eukaryotic microbes, principally fungi and labyrinthulomycetes, dominate biomass on bathypelagic marine snow. *The ISME Journal*, 11(2), 362–373. https://doi.org/10.1038/ismej.2016.113

Boenigk, J., Pfandl, K., Garstecki, T. et al. (2006). Evidence for Geographic Isolation and Signs of Endemism within a Protistan Morphospecies. *Applied and Environmental Microbiology*, 72(8), 5159–5164. https://doi.org/10.1128/AEM.00601-06

Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The New Tree of Eukaryotes. *Trends in Ecology & Evolution*, 35(1), 43–55. https://doi.org/10.1016/j.tree.2019.08.008

Callahan, B. J. (2017). ASV Tables inferred by DADA2 from the TARA Oceans v9 metabarcoding data set. *Zenodo*. https://doi.org/10.5281/zenodo.581694

Callahan, B. J., McMurdie, P. J., Rosen, M. J. et al. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13(7), 581–583. https://doi.org/10.1038/nmeth.3869

Canals, O., Obiol, A., Muhovic, I., Vaqué, D., & Massana, R. (2020). Ciliate diversity and distribution across horizontal and vertical scales in the open ocean. *Molecular Ecology*, 29(15), 2824–2839. https://doi.org/10.1111/mec.15528

Carr, A., Diener, C., Baliga, N. S., & Gibbons, S. M. (2019). Use and abuse of correlation analyses in microbial ecology. *The ISME Journal*, 13(11), 2647–2655. https://doi.org/10.1038/s41396-019-0459-z

Cavalier-Smith, T., & Chao, E. E. (2012). Oxnerella micra sp. n. (Oxnerellidae fam. n.), a Tiny Naked Centrohelid, and the Diversity and Evolution of Heliozoa. *Protist*, 163(4), 574–601. https://doi.org/10.1016/j.protis.2011.12.005

de Vargas, C., Audic, S., Henry, N. et al. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237), 1261605–1261605. https://doi.org/10.1126/science.1261605

del Campo, J., Balagué, V., Forn, I., Lekunberri, I., & Massana, R. (2013). Culturing Bias in Marine Heterotrophic Flagellates Analyzed Through Seawater Enrichment Incubations. *Microbial Ecology*, 66(3), 489–499. https://doi.org/10.1007/s00248-013-0251-y

del Campo, J., & Massana, R. (2011). Emerging diversity within chrysophytes, choanoflagellates and bicosoecids based on molecular surveys. *Protist*, 162(3), 435–448. https://doi.org/10.1016/j.protis.2010.10.003

Deutschmann, I. M., Krabberød, A. K., Benites, L. et al. (2021). Disentangling temporal associations in marine microbial networks. *Research Square*. https://doi.org/10.21203/rs.3.rs-404332/v1

Deutschmann, I. M., Lima-Mendez, G., Krabberød, A. K. et al. (2019). EnDED - Environmentally-Driven Edge Detection Program. *Zenodo*. https://doi.org/10.5281/zenodo.3271730

Díez, 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. https://doi.org/10.1128/AEM.67.7.2932-2941.2001

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. https://doi.org/10.1073/pnas.062186399

Fenchel, T. (1986). The Ecology of Heterotrophic Microflagellates. In K. C. Marshall (Ed.), *Advances in Microbial Ecology* (pp. 57–97). Springer US. https://doi.org/10.1007/978-1-4757-0611-6_2

Flegontova, O., Flegontov, P., Londoño, P. A. C. et al. (2020). Environmental determinants of the distribution of planktonic diplonemids and kinetoplastids in the oceans. *Environmental Microbiology*, 22(9), 4014–4031. https://doi.org/10.1111/1462-2920.15190

Flegontova, O., Flegontov, P., Malviya, S. et al. (2016). Extreme Diversity of Diplonemid Eukaryotes in the Ocean. *Current Biology*, 26(22), 3060–3065. https://doi.org/10.1016/j.cub.2016.09.031

Flegontova, O., Flegontov, P., Malviya, S. et al. (2018). Neobodonids are dominant kinetoplastids in the global ocean. *Environmental Microbiology*, 20(2), 878–889. https://doi.org/10.1111/1462-2920.14034

Friedman, J., & Alm, E. J. (2012). Inferring Correlation Networks from Genomic Survey Data. *PLoS Computational Biology*, 8(9), e1002687. https://doi.org/10.1371/journal.pcbi.1002687

Fuhrman, J. A., & Noble, R. T. (1995). Viruses and protists cause similar bacterial mortality in coastal seawater. *Limnology and Oceanography*, 40(7), 1236–1242. https://doi.org/10.4319/lo.1995.40.7.1236

Fuhrman, J. A., Steele, J. A., Hewson, I. et al. (2008). A latitudinal diversity gradient in planktonic marine bacteria. *Proceedings of the National Academy of Sciences*, 105(22), 7774–7778. https://doi.org/10.1073/pnas.0803070105

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

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Data sets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8(NOV), 1–6. https://doi.org/10.3389/fmicb.2017.02224

Glücksman, E., Bell, T., Griffiths, R. I., & Bass, D. (2010). Closely related protist strains have different grazing impacts on natural bacterial communities. *Environmental Microbiology*, 12(12), 3105–3113. https://doi.org/10.1111/j.1462-2920.2010.02283.x

Guillou, L., Viprey, M., Chambouvet, A. et al. (2008). Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environmental Microbiology*, 10(12), 3349–3365. https://doi.org/10.1111/j.1462-2920.2008.01731.x

Hutchins, D. A., & Fu, F. (2017). Microorganisms and ocean global change. *Nature Microbiology*, 2(6), 17058. https://doi.org/10.1038/nmicrobiol.2017.58

Ibarbalz, F. M., Henry, N., Brandão, M. C. et al. (2019). Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell*, 179(5), 1084-1097.e21. https://doi.org/10.1016/j.v.2019.10.008

Jeuck, A., & Arndt, H. (2013). A Short Guide to Common Heterotrophic Flagellates of Freshwater Habitats Based on the Morphology of Living Organisms. *Protist*, 164(6), 842–860. https://doi.org/10.1016/j.protis.2013.08.003

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). John Wiley & Sons, Inc. https://doi.org/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. https://doi.org/10.1093/molbev/mst010

Keeling, P. J., & Burki, F. (2019). Progress towards the Tree of Eukaryotes. *Current Biology*, 29(16), R808–R817. https://doi.org/10.1016/j.cub.2019.07.031

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, 35(21), 4453–4455. https://doi.org/10.1093/bioinformatics/btz305

Labarre, A., López-Escardó, D., Latorre, F. et al. (2021). Comparative genomics reveals new functional insights in uncultured MAST species. *The ISME Journal*, 15(6), 1767–1781. https://doi.org/10.1038/s41396-020-00885-8

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. https://doi.org/10.1111/j.1462-2920.2008.01737.x

Latorre, F., Deutschmann, I. M., Labarre, A. et al. (2021). Niche adaptation promoted the evolutionary diversification of tiny ocean predators. *Proceedings of the National Academy of Sciences*, 118(25), e2020955118. https://doi.org/10.1073/pnas.2020955118

Le Bescot, N., Mahé, F., Audic, S. et al. (2016). Global patterns of pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. *Environmental Microbiology*, 18(2), 609–626. https://doi.org/10.1111/1462-2920.13039

Lima-Mendez, G., Faust, K., Henry, N. et al. (2015). Determinants of community structure in the global plankton interactome. *Science*, 348(6237), 1262073–1262073. https://doi.org/10.1126/science.1262073

Logares, R., Deutschmann, I. M., Junger, P. C. et al. (2020). Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome*, 8(1), 1–17. https://doi.org/10.1186/s40168-020-00827-8

Lopes dos Santos, A., Gourvil, P., Tragin, M. et al. (2017). Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. *The ISME Journal*, 11(2), 512–528. https://doi.org/10.1038/ismej.2016.120

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. https://doi.org/10.1038/35054537

Malviya, S., Scalco, E., Audic, S. et al. (2016). Insights into global diatom distribution and diversity in the world’s ocean. *Proceedings of the National Academy of Sciences*, 113(11), E1516–E1525. https://doi.org/10.1073/pnas.1509523113

Mangot, J., Forn, I., Obiol, A., & Massana, R. (2018). Constant abundances of ubiquitous uncultured protists in the open sea assessed by automated microscopy. *Environmental Microbiology*, 20(10), 3876–3889. https://doi.org/10.1111/1462-2920.14408

Martin, B. D., Witten, D., & Willis, A. D. (2020). Modeling microbial abundances and dysbiosis with beta-binomial regression. *The Annals of Applied Statistics*, 14(1), 94–115. https://doi.org/10.1214/19-AOAS1283

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

Massana, R., Castresana, J., Balagué, V. et al. (2004). Phylogenetic and Ecological Analysis of Novel Marine Stramenopiles. *Applied and Environmental Microbiology*, 70(6), 3528–3534. https://doi.org/10.1128/AEM.70.6.3528

Massana, R., del Campo, J., Sieracki, M. E., Audic, S., & Logares, R. (2014). Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *The ISME Journal*, 8(4), 854–866. https://doi.org/10.1038/ismej.2013.204

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, 8(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Metz, S., Singer, D., Domaizon, I., Unrein, F., & Lara, E. (2019). Global distribution of Trebouxiophyceae diversity explored by high-throughput sequencing and phylogenetic approaches. *Environmental Microbiology*, 21(10), 3885–3895. https://doi.org/10.1111/1462-2920.14738

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. https://doi.org/10.1038/35054541

Moreira, D., & López-García, P. (2014). The rise and fall of picobiliphytes: How assumed autotrophs turned out to be heterotrophs. *BioEssays*, 36(5), 468–474. https://doi.org/10.1002/bies.201300176

Mukherjee, I., Salcher, M. M., Andrei, A. et al. (2020). A freshwater radiation of diplonemids. *Environmental Microbiology*, 22(11), 4658–4668. https://doi.org/10.1111/1462-2920.15209

Nagata, T., Tamburini, C., Arístegui, J. et al. (2010). Emerging concepts on microbial processes in the bathypelagic ocean – ecology, biogeochemistry, and genomics. *Deep Sea Research Part II: Topical Studies in Oceanography*, 57(16), 1519–1536. https://doi.org/10.1016/j.dsr2.2010.02.019

Not, F., Valentin, K., Romari, K. et al. (2007). Picobiliphytes: A Marine Picoplanktonic Algal Group with Unknown Affinities to Other Eukaryotes. *Science*, 315(5809), 253–255. https://doi.org/10.1126/science.1136264

Obiol, A., Giner, C. R., Sánchez, P. et al. (2020). A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Molecular Ecology Resources*, 20(3), 718–731. https://doi.org/10.1111/1755-0998.13147

Oksanen, J., Blanchet, F. G., Friendly, M. et al. (2019). vegan: Community Ecology Package. https://cran.r-project.org/package=vegan

Patterson, D. J., & Lee, W. J. (2000). Geographic distribution and diversity of free–living heterotrophic flagellates. In B. S. Leadbeater & J. C. Green (Eds.), *The flagellates: Unity, diversity and evolution* (pp. 267–287). Taylor & Francis.

Pedersen, T. L. (2020a). ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. https://cran.r-project.org/package=ggraph

Pedersen, T. L. (2020b). tidygraph: A Tidy API for Graph Manipulation. https://cran.r-project.org/package=tidygraph

Pernice, M. C., Giner, C. R., Logares, R. et al. (2016). Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. *The ISME Journal*, 10(4), 945–958. https://doi.org/10.1038/ismej.2015.170

Pernthaler, J. (2005). Predation on prokaryotes in the water column and its ecological implications. *Nature Reviews Microbiology*, 3(7), 537–546. https://doi.org/10.1038/nrmicro1180

Pommier, T., Canbäck, B., Riemann, L. et al. (2007). Global patterns of diversity and community structure in marine bacterioplankton. *Molecular Ecology*, 16(4), 867–880. https://doi.org/10.1111/j.1365-294X.2006.03189.x

R Core Team. (2020). R: A Language and Environment for Statistical Computing. https://www.r-project.org/

Rodríguez-Martínez, R., Labrenz, M., Del Campo, J. et al. (2009). Distribution of the uncultured protist MAST-4 in the Indian Ocean, Drake Passage and Mediterranean Sea assessed by real-time quantitative PCR. *Environmental Microbiology*, 11(2), 397–408. https://doi.org/10.1111/j.1462-2920.2008.01779.x

Rodríguez-Martínez, R., Rocap, G., Salazar, G., & Massana, R. (2013). Biogeography of the uncultured marine picoeukaryote MAST-4: temperature-driven distribution patterns. *The ISME Journal*, 7(8), 1531–1543. https://doi.org/10.1038/ismej.2013.53

Röttjers, L., & Faust, K. (2018). From hairballs to hypotheses–biological insights from microbial networks. *FEMS Microbiology Reviews*, 42(6), 761–780. https://doi.org/10.1093/femsre/fuy030

Schnetzer, A., Moorthi, S. D., Countway, P. D. et al. (2011). Depth matters: Microbial eukaryote diversity and community structure in the eastern North Pacific revealed through environmental gene libraries. *Deep Sea Research Part I: Oceanographic Research Papers*, 58(1), 16–26. https://doi.org/10.1016/j.dsr.2010.10.003

Schoenle, A., Hohlfeld, M., Hermanns, K. et al. (2021). High and specific diversity of protists in the deep-sea basins dominated by diplonemids, kinetoplastids, ciliates and foraminiferans. *Communications Biology*, 4(1), 1–10. https://doi.org/10.1038/s42003-021-02012-5

Schön, M. E., Zlatogursky, V. V., Singh, R. P., Poirier, C., Wilken, S., Mathur, V., Strassert, J. F. H., Pinhassi, J., Worden, A. Z., Keeling, P. J., Ettema, T. J. G., Wideman, J. G., & Burki, F. (2021). Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. *Nature Communications*, 12(1), 6651. https://doi.org/10.1038/s41467-021-26918-0

Seenivasan, R., Sausen, N., Medlin, L. K., & Melkonian, M. (2013). Picomonas judraskeda Gen. Et Sp. Nov.: The First Identified Member of the Picozoa Phylum Nov., a Widespread Group of Picoeukaryotes, Formerly Known as 'Picobiliphytes.' *PLoS ONE*, 8(3), e59565. https://doi.org/10.1371/journal.pone.0059565

Sekiguchi, H., Moriya, M., Nakayama, T., & Inouye, I. (2002). Vestigial Chloroplasts in Heterotrophic Stramenopiles Pteridomonas danica and Ciliophrys infusionum (Dictyochophyceae). *Protist*, 153(2), 157–167. https://doi.org/10.1078/1434-4610-00094

Shalchian-Tabrizi, K., Kauserud, H., Massana, R., Klaveness, D., & Jakobsen, K. S. (2007). Analysis of Environmental 18S Ribosomal RNA Sequences reveals Unknown Diversity of the Cosmopolitan Phylum Telonemia. *Protist*, 158(2), 173–180. https://doi.org/https://doi.org/10.1016/j.protis.2006.10.003

Sherr, E. B., & Sherr, B. F. (2002). Significance of predation by protists in aquatic microbial food webs. *Antonie van Leeuwenhoek*, 81(1), 293–308. https://doi.org/10.1023/A:1020591307260

Shishkin, Y., Drachko, D., Klimov, V. I., & Zlatogursky, V. V. (2018). Yogsothoth knorrus gen. n., sp. n. and Y. carteri sp. n. (Yogsothothidae fam. n., Haptista, Centroplasthelida), with Notes on Evolution and Systematics of Centrohelids. *Protist*, 169(5), 682–696. https://doi.org/10.1016/j.protis.2018.06.003

Šimek, K., Grujčić, V., Hahn, M. W. et al. (2018). Bacterial prey food characteristics modulate community growth response of freshwater bacterivorous flagellates. *Limnology and Oceanography*, 63(1), 484–502. https://doi.org/10.1002/lno.10759

Stoeck, T., Bass, D., Nebel, M. et al. (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. https://doi.org/10.1111/j.1365-294X.2009.04480.x

Vaulot, D., Geisen, S., Mahé, F., & Bass, D. (2021). pr2-primers: An 18S rRNA primer database for protists. *Molecular Ecology Resources*, 1755-0998.13465. https://doi.org/10.1111/1755-0998.13465

Watts, S. C., Ritchie, S. C., Inouye, M., & Holt, K. E. (2019). FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics*, 35(6), 1064–1066. https://doi.org/10.1093/bioinformatics/bty734

Wickham, H., Averick, M., Bryan, J. et al. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Willis, A. D., & Martin, B. D. (2020). Estimating diversity in networked ecological communities. *Biostatistics*. https://doi.org/10.1093/biostatistics/kxaa015

Zhu, F., Massana, R., Not, F., Marie, D., & Vaulot, D. (2005). Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiology Ecology*, 52(1), 79–92. https://doi.org/10.1016/j.femsec.2004.10.006

Živaljić, S., Schoenle, A., Nitsche, F. et al. (2018). Survival of marine heterotrophic flagellates isolated from the surface and the deep sea at high hydrostatic pressure: Literature review and own experiments. *Deep Sea Research Part II: Topical Studies in Oceanography*, 148(April 2017), 251–259. https://doi.org/10.1016/j.dsr2.2017.04.022

Zubkov, M. V., & Tarran, G. A. (2008). High bacterivory by the smallest phytoplankton in the North Atlantic Ocean. *Nature*, 455(7210), 224–226. https://doi.org/10.1038/nature07236

2.8. Supplementary figures

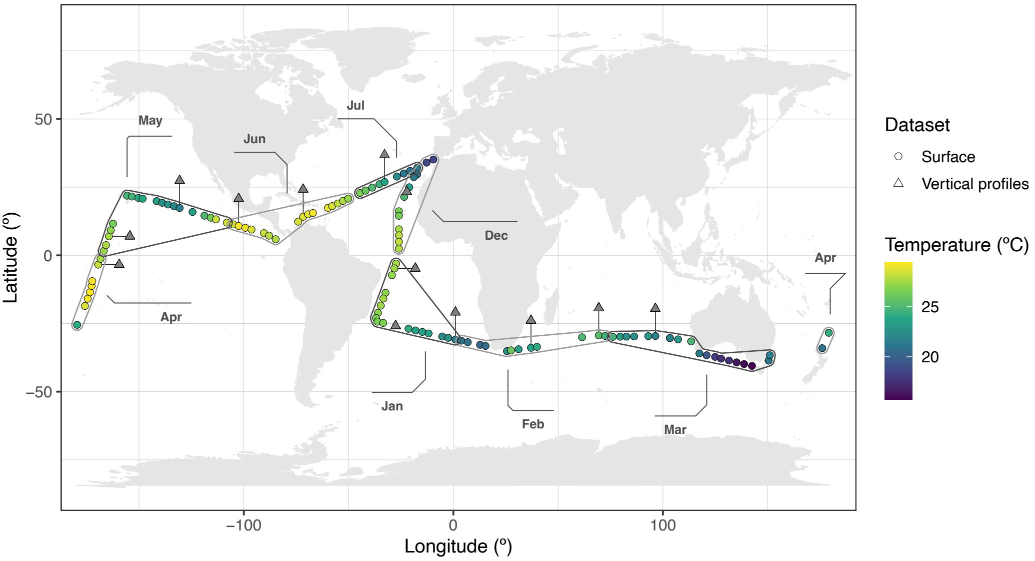
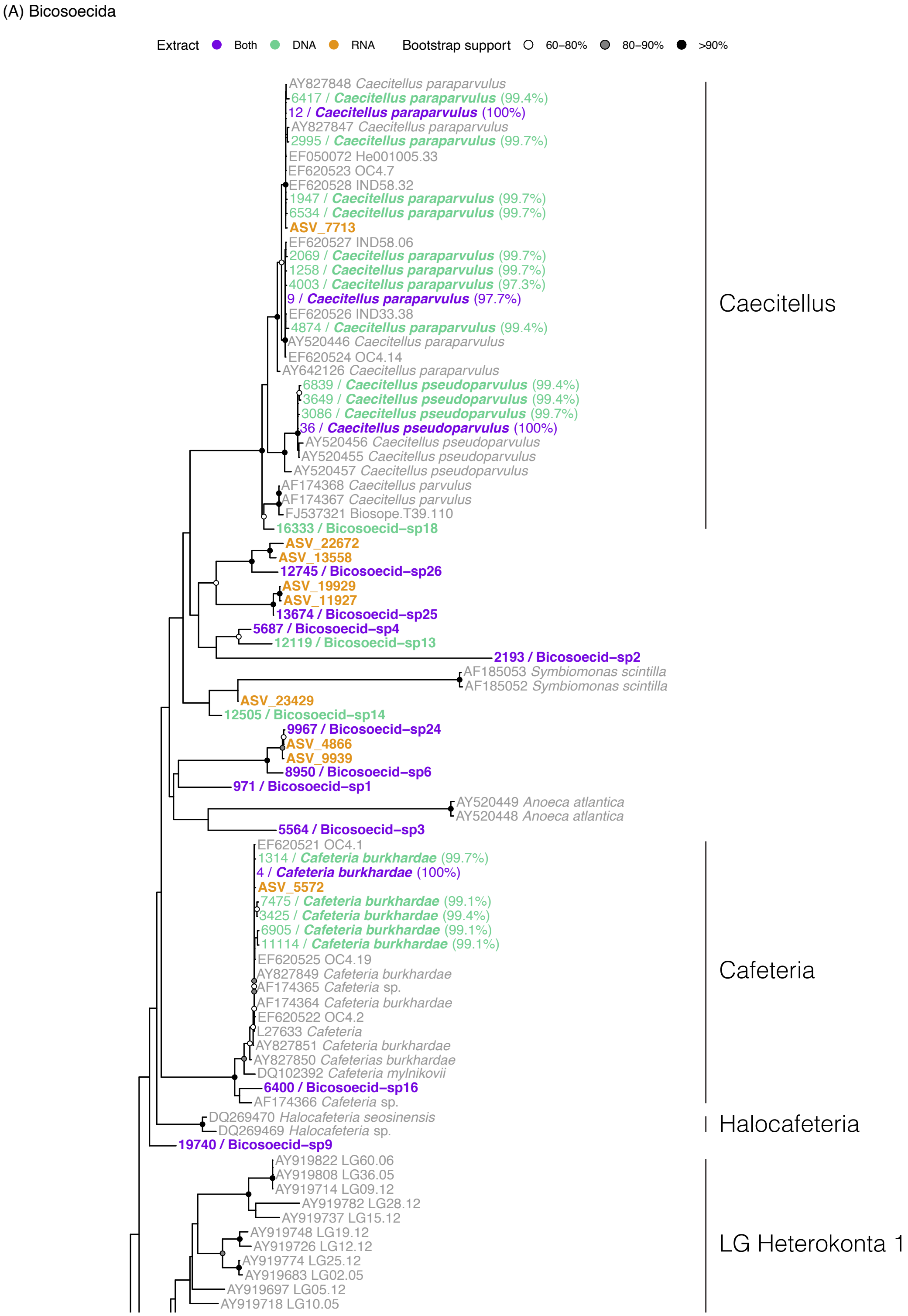
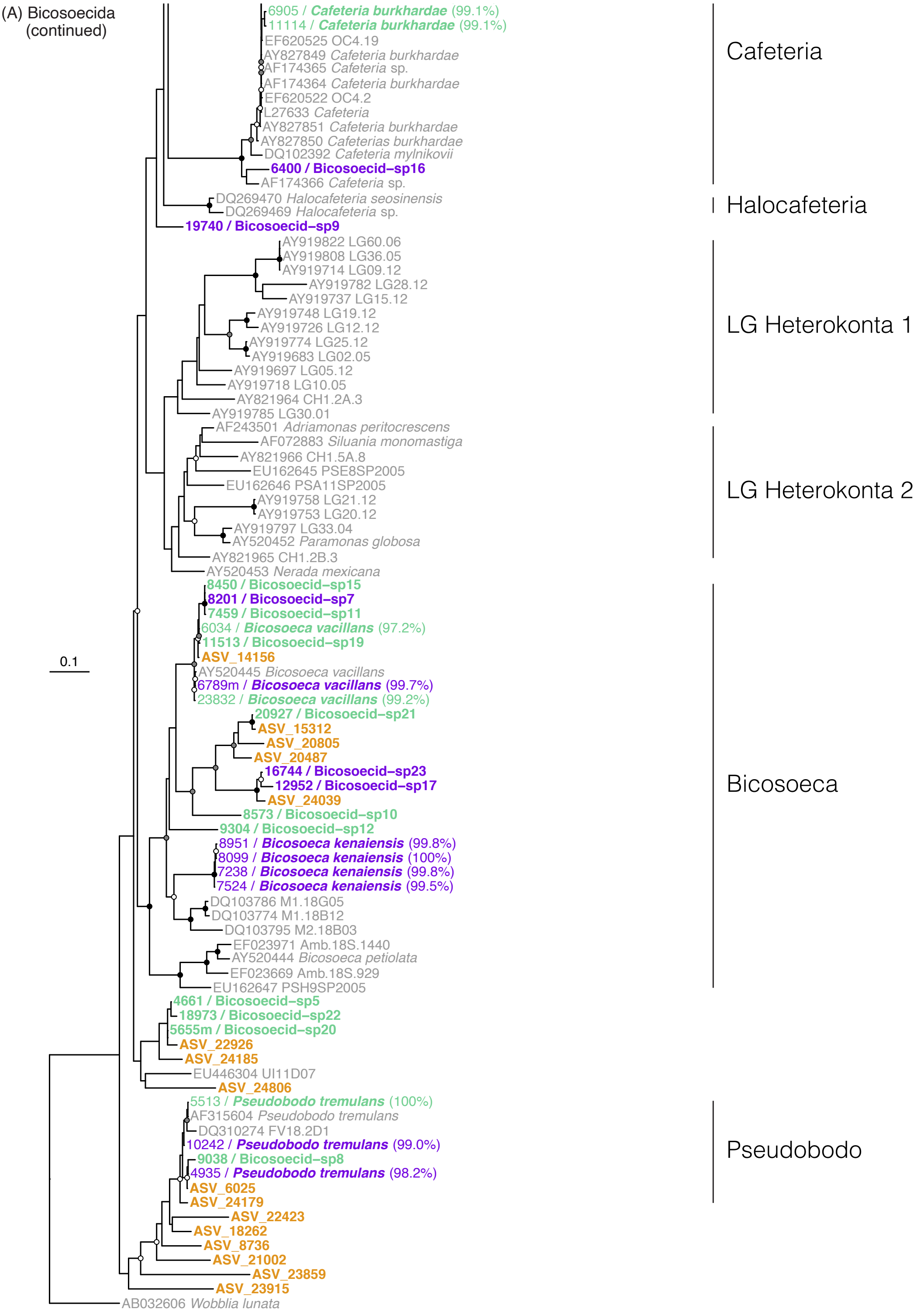
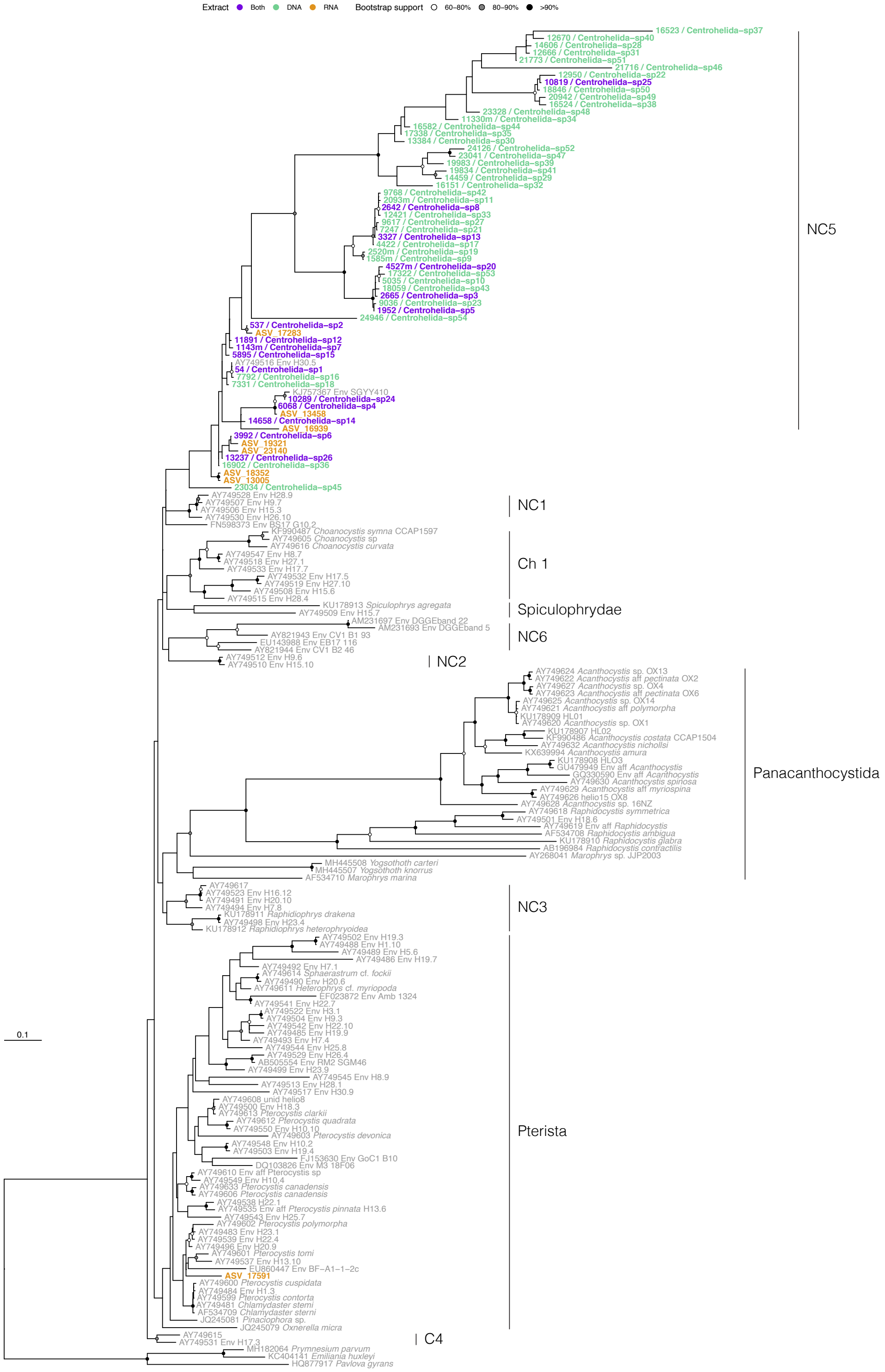


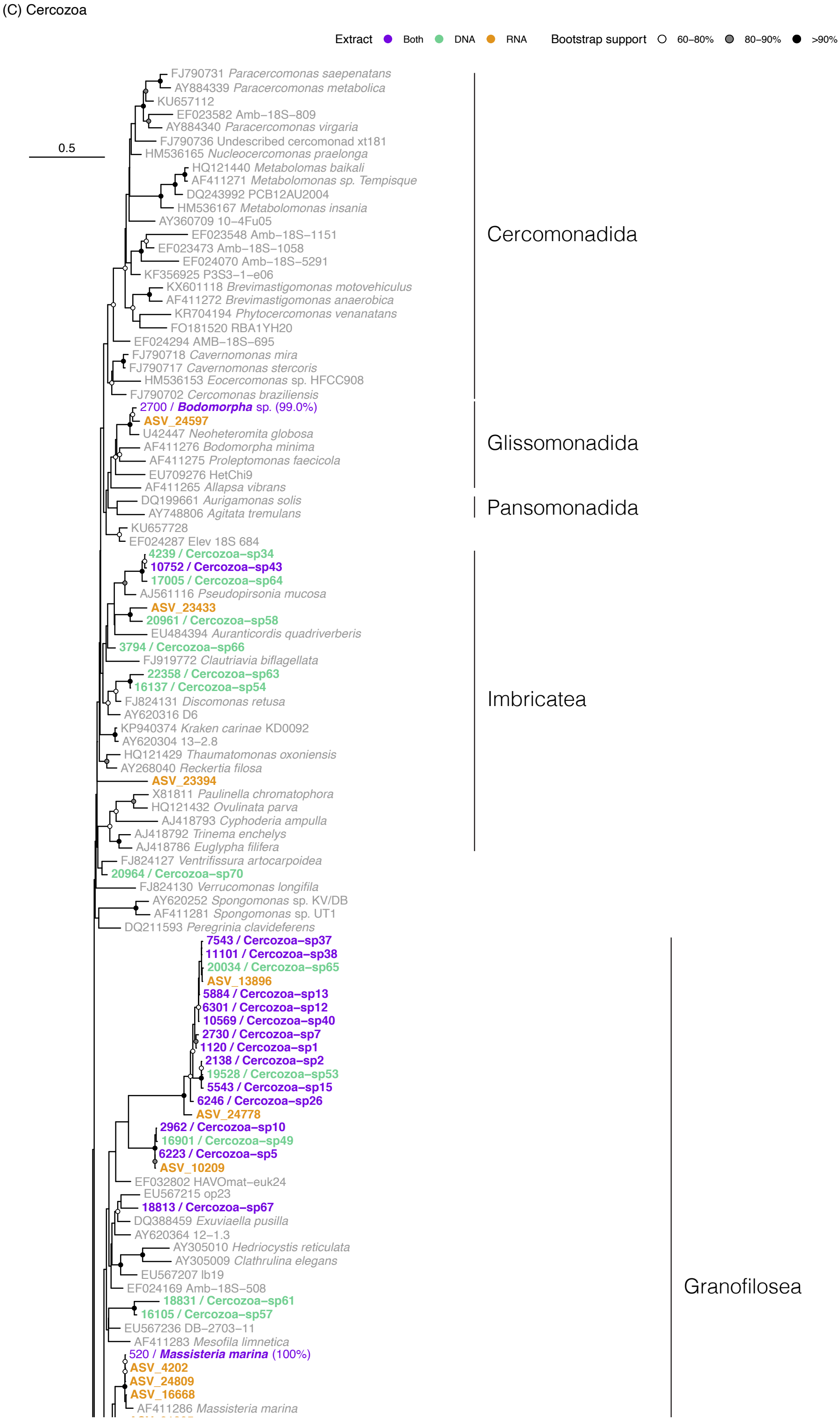
Figure S1. Map of sampling locations and sampling strategy down the water column (only surface or vertical profiles) for the Malaspina cruise. Samples are colored according to surface water temperature and grouped by the month of the year in which they were collected.


(A) Bicosoecida
Extract ● Both ● DNA ● RNA
Bootstrap support ○ 60–80% ● 80–90% ● >90%
AY827848 Caecitellus paraparvulus
6417 / Caecitellus paraparvulus (99.4%)
12 / Caecitellus paraparvulus (100%)
AY827847 Caecitellus paraparvulus
2995 / Caecitellus paraparvulus (99.7%)
EF050072 He001005.33
EF620523 OC4.7
EF620528 IND58.32
1947 / Caecitellus paraparvulus (99.7%)
6534 / Caecitellus paraparvulus (99.7%)
ASV_7713
EF620527 IND58.06
2069 / Caecitellus paraparvulus (99.7%)
1258 / Caecitellus paraparvulus (99.7%)
4003 / Caecitellus paraparvulus (97.3%)
9 / Caecitellus paraparvulus (97.7%)
EF620526 IND33.38
4874 / Caecitellus paraparvulus (99.4%)
AY520446 Caecitellus paraparvulus
EF620524 OC4.14
AY642126 Caecitellus paraparvulus
6839 / Caecitellus pseudoparvulus (99.4%)
3649 / Caecitellus pseudoparvulus (99.4%)
3086 / Caecitellus pseudoparvulus (99.7%)
36 / Caecitellus pseudoparvulus (100%)
AY520456 Caecitellus pseudoparvulus
AY520455 Caecitellus pseudoparvulus
AY520457 Caecitellus pseudoparvulus
AF174368 Caecitellus parvulus
AF174367 Caecitellus parvulus
FJ537321 Biosope.T39.110
16333 / Bicosoecid–sp18
Caecitellus
ASV_22672
ASV_13558
12745 / Bicosoecid–sp26
ASV_19929
ASV_11927
13674 / Bicosoecid–sp25
5687 / Bicosoecid–sp4
12119 / Bicosoecid–sp13
2193 / Bicosoecid–sp2
AF185053 Symbiomonas scintilla
AF185052 Symbiomonas scintilla
ASV_23429
12505 / Bicosoecid–sp14
9967 / Bicosoecid–sp24
ASV_4866
ASV_9939
8950 / Bicosoecid–sp6
971 / Bicosoecid–sp1
AY520449 Anoeca atlantica
AY520448 Anoeca atlantica
5564 / Bicosoecid–sp3
EF620521 OC4.1
1314 / Cafeteria burkhardae (99.7%)
4 / Cafeteria burkhardae (100%)
ASV_5572
7475 / Cafeteria burkhardae (99.1%)
3425 / Cafeteria burkhardae (99.4%)
6905 / Cafeteria burkhardae (99.1%)
11114 / Cafeteria burkhardae (99.1%)
EF620525 OC4.19
AY827849 Cafeteria burkhardae
AF174365 Cafeteria sp.
AF174364 Cafeteria burkhardae
EF620522 OC4.2
L27633 Cafeteria
AY827851 Cafeteria burkhardae
AY827850 Cafeterias burkhardae
DQ102392 Cafeteria mylnikovii
6400 / Bicosoecid–sp16
AF174366 Cafeteria sp.
Cafeteria
DQ269470 Halocafeteria seosinensis
DQ269469 Halocafeteria sp.
Halocafeteria
19740 / Bicosoecid–sp9
AY919822 LG60.06
AY919808 LG36.05
AY919714 LG09.12
AY919782 LG28.12
AY919737 LG15.12
AY919748 LG19.12
AY919726 LG12.12
AY919774 LG25.12
AY919683 LG02.05
AY919697 LG05.12
AY919718 LG10.05
LG Heterokonta 1

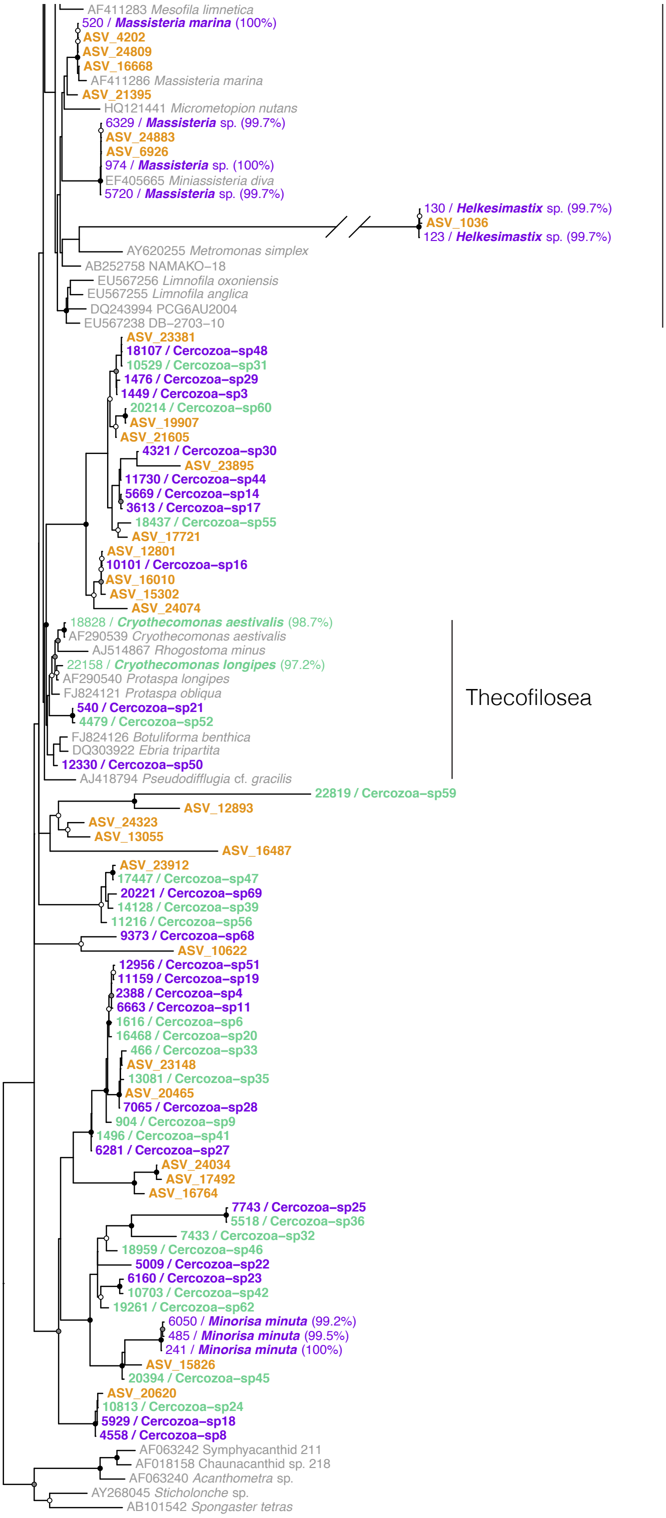

(A) Bicosoecida (continued)
6905 / Cafeteria burkhardae (99.1%)
11114 / Cafeteria burkhardae (99.1%)
EF620525 OC4.19
AY827849 Cafeteria burkhardae
AF174365 Cafeteria sp.
AF174364 Cafeteria burkhardae
EF620522 OC4.2
L27633 Cafeteria
AY827851 Cafeteria burkhardae
AY827850 Cafeterias burkhardae
DQ102392 Cafeteria mylnikovii
6400 / Bicosoecid–sp16
AF174366 Cafeteria sp.
DQ269470 Halocafeteria seosinensis
DQ269469 Halocafeteria sp.
19740 / Bicosoecid–sp9
AY919822 LG60.06
AY919808 LG36.05
AY919714 LG09.12
AY919782 LG28.12
AY919737 LG15.12
AY919748 LG19.12
AY919726 LG12.12
AY919774 LG25.12
AY919683 LG02.05
AY919697 LG05.12
AY919718 LG10.05
AY821964 CH1.2A.3
AY919785 LG30.01
AF243501 Adriamonas peritocrescens
AF072883 Siluania monomastiga
AY821966 CH1.5A.8
EU162645 PSE8SP2005
EU162646 PSA11SP2005
AY919758 LG21.12
AY919753 LG20.12
AY919797 LG33.04
AY520452 Paramonas globosa
AY821965 CH1.2B.3
AY520453 Nerada mexicana
8450 / Bicosoecid–sp15
8201 / Bicosoecid–sp7
7459 / Bicosoecid–sp11
6034 / Bicosoeca vacillans (97.2%)
11513 / Bicosoecid–sp19
ASV_14156
AY520445 Bicosoeca vacillans
6789m / Bicosoeca vacillans (99.7%)
23832 / Bicosoeca vacillans (99.2%)
20927 / Bicosoecid–sp21
ASV_15312
ASV_20805
ASV_20487
16744 / Bicosoecid–sp23
12952 / Bicosoecid–sp17
ASV_24039
8573 / Bicosoecid–sp10
9304 / Bicosoecid–sp12
8951 / Bicosoeca kenaiensis (99.8%)
8099 / Bicosoeca kenaiensis (100%)
7238 / Bicosoeca kenaiensis (99.8%)
7524 / Bicosoeca kenaiensis (99.5%)
DQ103786 M1.18G05
DQ103774 M1.18B12
DQ103795 M2.18B03
EF023971 Amb.18S.1440
AY520444 Bicosoeca petiolata
EF023669 Amb.18S.929
EU162647 PSH9SP2005
4661 / Bicosoecid–sp5
18973 / Bicosoecid–sp22
5655m / Bicosoecid–sp20
ASV_22926
ASV_24185
EU446304 UI11D07
ASV_24806
5513 / Pseudobodo tremulans (100%)
AF315604 Pseudobodo tremulans
DQ310274 FV18.2D1
10242 / Pseudobodo tremulans (99.0%)
9038 / Bicosoecid–sp8
4935 / Pseudobodo tremulans (98.2%)
ASV_6025
ASV_24179
ASV_22423
ASV_18262
ASV_8736
ASV_21002
ASV_23859
ASV_23915
AB032606 Wobblia lunata
0.1
Cafeteria
Halocafeteria
LG Heterokonta 1
LG Heterokonta 2
Bicosoeca
Pseudobodo

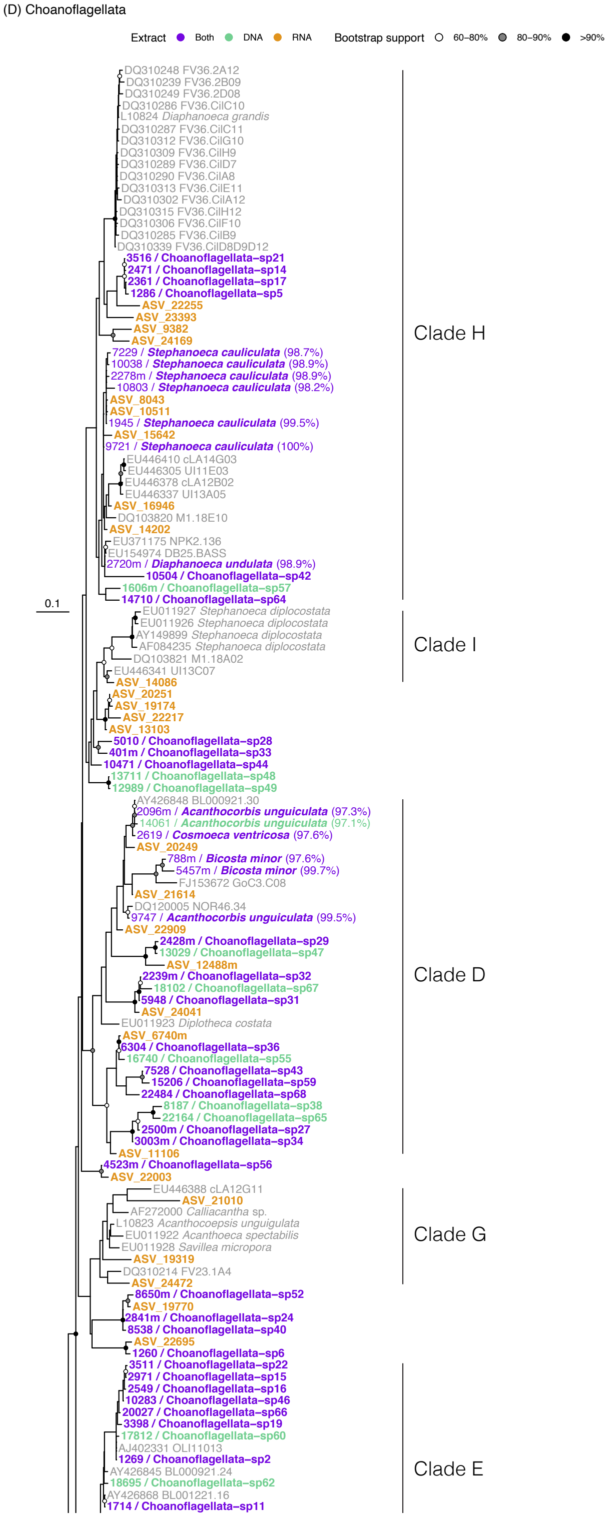
(B) Centrohelida


Extract ● Both ● DNA ● RNA
Bootstrap support ○ 60–80% ● 80–90% ● >90%
16523 / Centrohelida–sp37
12670 / Centrohelida–sp40
14606 / Centrohelida–sp28
12666 / Centrohelida–sp31
21773 / Centrohelida–sp51
21716 / Centrohelida–sp46
12950 / Centrohelida–sp22
10819 / Centrohelida–sp25
18846 / Centrohelida–sp50
20942 / Centrohelida–sp49
16524 / Centrohelida–sp38
23328 / Centrohelida–sp48
11330m / Centrohelida–sp34
16582 / Centrohelida–sp44
17338 / Centrohelida–sp35
13384 / Centrohelida–sp30
24126 / Centrohelida–sp52
23041 / Centrohelida–sp47
19983 / Centrohelida–sp39
19834 / Centrohelida–sp41
14459 / Centrohelida–sp29
16151 / Centrohelida–sp32
9768 / Centrohelida–sp42
2093m / Centrohelida–sp11
2642 / Centrohelida–sp8
12421 / Centrohelida–sp33
9617 / Centrohelida–sp27
7247 / Centrohelida–sp21
3327 / Centrohelida–sp13
4422 / Centrohelida–sp17
2520m / Centrohelida–sp19
1585m / Centrohelida–sp9
4527m / Centrohelida–sp20
17322 / Centrohelida–sp53
5035 / Centrohelida–sp10
18059 / Centrohelida–sp43
2665 / Centrohelida–sp3
9036 / Centrohelida–sp23
1952 / Centrohelida–sp5
24946 / Centrohelida–sp54
537 / Centrohelida–sp2
ASV_17283
11891 / Centrohelida–sp12
1143m / Centrohelida–sp7
5895 / Centrohelida–sp15
AY749516 Env H30.5
54 / Centrohelida–sp1
7792 / Centrohelida–sp16
7331 / Centrohelida–sp18
KJ757367 Env SGYY410
10289 / Centrohelida–sp24
6068 / Centrohelida–sp4
ASV_13458
14658 / Centrohelida–sp14
ASV_16939
3992 / Centrohelida–sp6
ASV_19321
ASV_23140
13237 / Centrohelida–sp26
16902 / Centrohelida–sp36
ASV_18352
ASV_13005
23034 / Centrohelida–sp45
NC5
AY749528 Env H28.9
AY749507 Env H9.7
AY749506 Env H15.3
AY749530 Env H26.10
FN598373 Env BS17 G10.2
NC1
KF990487 Choanocystis symna CCAP1597
AY749605 Choanocystis sp
AY749616 Choanocystis curvata
AY749547 Env H8.7
AY749518 Env H27.1
AY749533 Env H17.7
AY749532 Env H17.5
AY749519 Env H27.10
AY749508 Env H15.6
AY749515 Env H28.4
Ch 1
KU178913 Spiculophrys agregata
AY749509 Env H15.7
Spiculophrydae
AM231697 Env DGGEband 22
AM231693 Env DGGEband 5
AY821943 Env CV1 B1 93
EU143988 Env EB17 116
AY821944 Env CV1 B2 46
NC6
AY749512 Env H9.6
AY749510 Env H15.10
NC2
AY749624 Acanthocystis sp. OX13
AY749622 Acanthocystis aff pectinata OX2
AY749627 Acanthocystis sp. OX4
AY749623 Acanthocystis aff pectinata OX6
AY749625 Acanthocystis sp. OX14
AY749621 Acanthocystis aff polymorpha
KU178909 HL01
AY749620 Acanthocystis sp. OX1
KU178907 HL02
KF990486 Acanthocystis costata CCAP1504
AY749632 Acanthocystis nichollsi
KX639994 Acanthocystis amura
KU178908 HLO3
GU479949 Env aff Acanthocystis
GQ330590 Env aff Acanthocystis
AY749630 Acanthocystis spinosa
AY749629 Acanthocystis aff myriospina
AY749626 helio15 OX8
AY749628 Acanthocystis sp. 16NZ
AY749618 Raphidocystis symmetrica
AY749501 Env H18.6
AY749619 Env aff Raphidocystis
AF534708 Raphidocystis ambigua
KU178910 Raphidocystis glabra
AB196984 Raphidocystis contractilis
AY268041 Marophrys sp. JJP2003
Panacanthocystida
MH445508 Yogsothoth carteri
MH445507 Yogsothoth knorrus
AF534710 Marophrys marina
AY749617
AY749523 Env H16.12
AY749491 Env H20.10
AY749494 Env H7.8
KU178911 Raphidiophrys drakena
AY749498 Env H23.4
KU178912 Raphidiophrys heterophryoidea
NC3
AY749502 Env H19.3
AY749488 Env H1.10
AY749489 Env H5.6
AY749486 Env H19.7
AY749492 Env H7.1
AY749614 Sphaerastrum cf. fockii
AY749490 Env H20.6
AY749611 Heterophrys cf. myriopoda
EF023872 Env Amb 1324
AY749541 Env H22.7
AY749522 Env H3.1
AY749504 Env H9.3
AY749542 Env H22.10
AY749485 Env H19.9
AY749493 Env H7.4
AY749544 Env H25.8
AY749529 Env H26.4
AB505554 Env RM2 SGM46
AY749499 Env H23.9
AY749545 Env H8.9
AY749513 Env H28.1
AY749517 Env H30.9
AY749608 unid helio8
AY749500 Env H18.3
AY749613 Pterocystis clarkii
AY749612 Pterocystis quadrata
AY749550 Env H10.10
AY749603 Pterocystis devonica
AY749548 Env H10.2
AY749503 Env H19.4
FJ153630 Env GoC1 B10
DQ103826 Env M3 18F06
AY749610 Env aff Pterocystis sp
AY749549 Env H10.4
AY749633 Pterocystis canadensis
AY749606 Pterocystis canadensis
AY749538 H22.1
AY749535 Env aff Pterocystis pinnata H13.6
AY749543 Env H25.7
AY749602 Pterocystis polymorpha
AY749483 Env H23.1
AY749539 Env H22.4
AY749496 Env H20.9
AY749601 Pterocystis tomi
AY749537 Env H13.10
EU860447 Env BF–A1–1–2c
ASV_17591
AY749600 Pterocystis cuspidata
AY749484 Env H1.3
AY749599 Pterocystis contorta
AY749481 Chlamydaster sterni
AF534709 Chlamydaster sterni
JQ245081 Pinaciophora sp.
JQ245079 Oxnerella micra
Pterista
AY749615
AY749531 Env H17.3
C4
MH182064 Prymnesium parvum
KC404141 Emiliania huxleyi
HQ877917 Pavlova gyrans
0.1

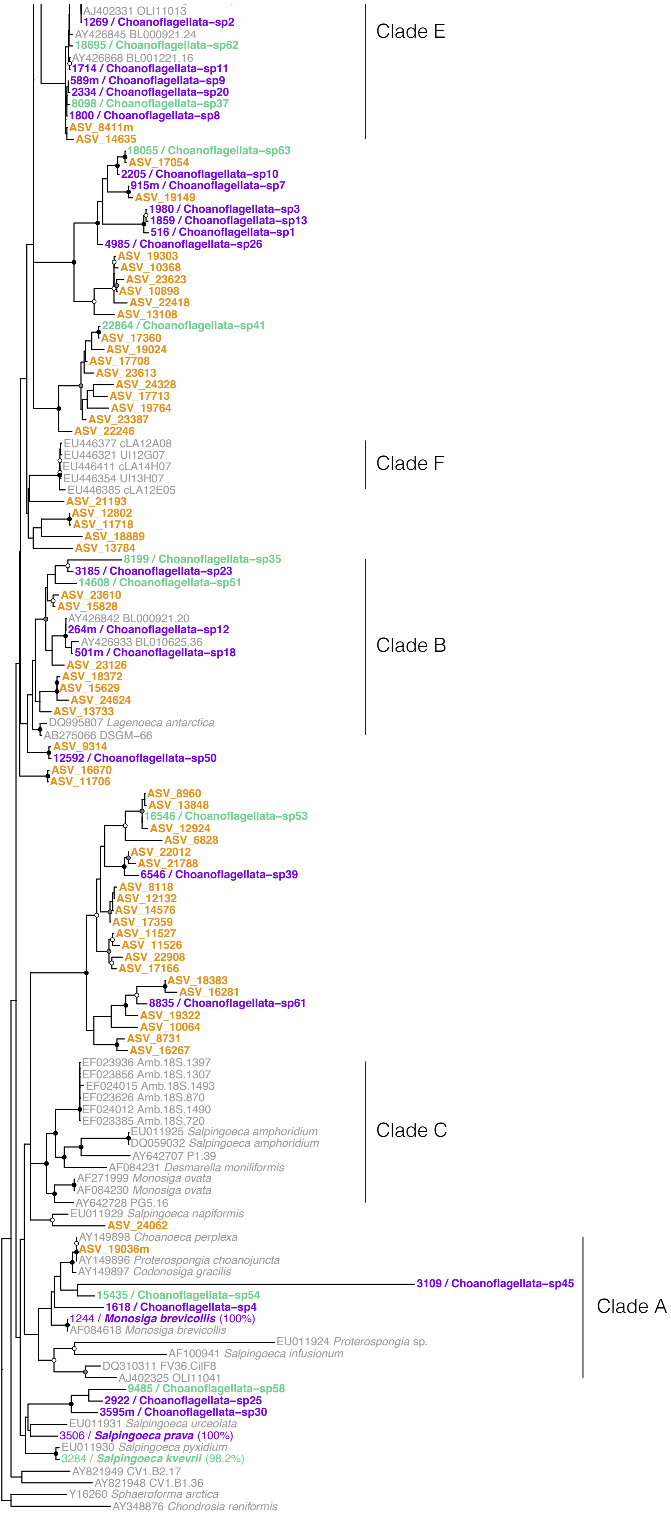

(C) Cercozoa
Extract ● Both ● DNA ● RNA
Bootstrap support ○ 60–80% ● 80–90% ● >90%
0.5
FJ790731 Paracercomonas saepenatans
AY884339 Paracercomonas metabolica
KU657112
EF023582 Amb–18S–809
AY884340 Paracercomonas virgaria
FJ790736 Undescribed cercomonad xt181
HM536165 Nucleocercomonas praelonga
HQ121440 Metabolomas baikali
AF411271 Metabolomonas sp. Tempisque
DQ243992 PCB12AU2004
HM536167 Metabolomonas insania
AY360709 10–4Fu05
EF023548 Amb–18S–1151
EF023473 Amb–18S–1058
EF024070 Amb–18S–5291
KF356925 P3S3–1–e06
KX601118 Brevimastigomonas motovehiculus
AF411272 Brevimastigomonas anaerobica
KR704194 Phytocercomonas venanatans
FO181520 RBA1YH20
EF024294 AMB–18S–695
FJ790718 Cavernomonas mira
FJ790717 Cavernomonas stercoris
HM536153 Eocercomonas sp. HFCC908
FJ790702 Cercomonas braziliensis
2700 / Bodomorpha sp. (99.0%)
ASV_24597
U42447 Neoheteromita globosa
AF411276 Bodomorpha minima
AF411275 Proleptomonas faecicola
EU709276 HetChi9
AF411265 Allapsa vibrans
DQ199661 Aurigamonas solis
AY748806 Agitata tremulans
KU657728
EF024287 Elev 18S 684
4239 / Cercozoa–sp34
10752 / Cercozoa–sp43
17005 / Cercozoa–sp64
AJ561116 Pseudopirsonia mucosa
ASV_23433
20961 / Cercozoa–sp58
EU484394 Auranticordis quadriverberis
3794 / Cercozoa–sp66
FJ919772 Clautriavia biflagellata
22358 / Cercozoa–sp63
16137 / Cercozoa–sp54
FJ824131 Discomonas retusa
AY620316 D6
KP940374 Kraken carinae KD0092
AY620304 13–2.8
HQ121429 Thaumatomonas oxoniensis
AY268040 Reckertia filosa
ASV_23394
X81811 Paulinella chromatophora
HQ121432 Ovulinata parva
AJ418793 Cyphoderia ampulla
AJ418792 Trinema enchelys
AJ418786 Euglypha filifera
FJ824127 Ventrifissura artocarpoidea
20964 / Cercozoa–sp70
FJ824130 Verrucomonas longifila
AY620252 Spongomonas sp. KV/DB
AF411281 Spongomonas sp. UT1
DQ211593 Peregrinia clavideferens
7543 / Cercozoa–sp37
11101 / Cercozoa–sp38
20034 / Cercozoa–sp65
ASV_13896
5884 / Cercozoa–sp13
6301 / Cercozoa–sp12
10569 / Cercozoa–sp40
2730 / Cercozoa–sp7
1120 / Cercozoa–sp1
2138 / Cercozoa–sp2
19528 / Cercozoa–sp53
5543 / Cercozoa–sp15
6246 / Cercozoa–sp26
ASV_24778
2962 / Cercozoa–sp10
16901 / Cercozoa–sp49
6223 / Cercozoa–sp5
ASV_10209
EF032802 HAVOmat–euk24
EU567215 op23
18813 / Cercozoa–sp67
DQ388459 Exuviaella pusilla
AY620364 12–1.3
AY305010 Hedriocystis reticulata
AY305009 Clathrulina elegans
EU567207 lb19
EF024169 Amb–18S–508
18831 / Cercozoa–sp61
16105 / Cercozoa–sp57
EU567236 DB–2703–11
AF411283 Mesofila limnetica
520 / Massisteria marina (100%)
ASV_4202
ASV_24809
ASV_16668
AF411286 Massisteria marina
Cercomonadida
Glissomonadida
Pansomonadida
Imbricatea
Granofilosea

(C) Cercozoa
(continued)

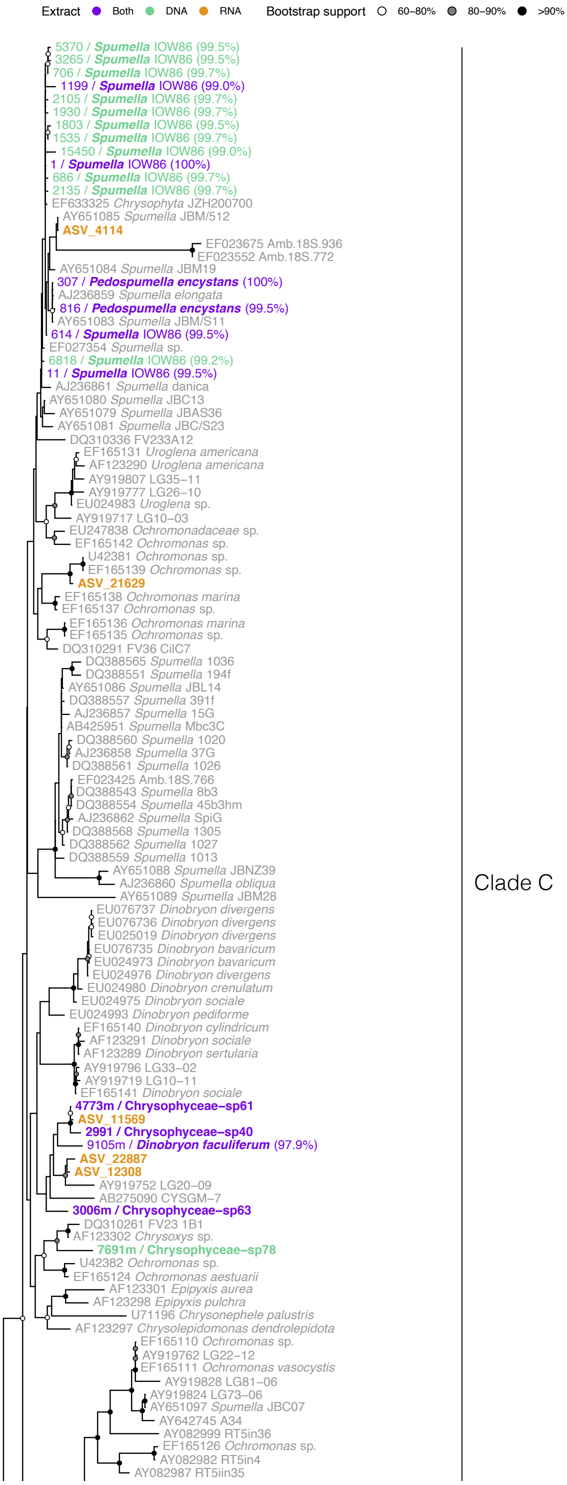



(D) Choanoflagellata
Extract ● Both ● DNA ● RNA
Bootstrap support ○ 60–80% ● 80–90% ● >90%
DQ310248 FV36.2A12
DQ310239 FV36.2B09
DQ310249 FV36.2D08
DQ310286 FV36.CilC10
L10824 Diaphanoeca grandis
DQ310287 FV36.CilC11
DQ310312 FV36.CilG10
DQ310309 FV36.CilH9
DQ310289 FV36.CilD7
DQ310290 FV36.CilA8
DQ310313 FV36.CilE11
DQ310302 FV36.CilA12
DQ310315 FV36.CilH12
DQ310306 FV36.CilF10
DQ310285 FV36.CilB9
DQ310339 FV36.CilD8D9D12
3516 / Choanoflagellata–sp21
2471 / Choanoflagellata–sp14
2361 / Choanoflagellata–sp17
1286 / Choanoflagellata–sp5
ASV_22255
ASV_23393
ASV_9382
ASV_24169
7229 / Stephanoeca cauliculata (98.7%)
10038 / Stephanoeca cauliculata (98.9%)
2278m / Stephanoeca cauliculata (98.9%)
10803 / Stephanoeca cauliculata (98.2%)
ASV_8043
ASV_10511
1945 / Stephanoeca cauliculata (99.5%)
ASV_15642
9721 / Stephanoeca cauliculata (100%)
EU446410 cLA14G03
EU446305 UI11E03
EU446378 cLA12B02
EU446337 UI13A05
ASV_16946
DQ103820 M1.18E10
ASV_14202
EU371175 NPK2.136
EU154974 DB25.BASS
2720m / Diaphanoeca undulata (98.9%)
10504 / Choanoflagellata–sp42
1606m / Choanoflagellata–sp57
14710 / Choanoflagellata–sp64
Clade H
0.1
EU011927 Stephanoeca diplocostata
EU011926 Stephanoeca diplocostata
AY149899 Stephanoeca diplocostata
AF084235 Stephanoeca diplocostata
DQ103821 M1.18A02
EU446341 UI13C07
ASV_14086
Clade I
ASV_20251
ASV_19174
ASV_22217
ASV_13103
5010 / Choanoflagellata–sp28
401m / Choanoflagellata–sp33
10471 / Choanoflagellata–sp44
13711 / Choanoflagellata–sp48
12989 / Choanoflagellata–sp49
AY426848 BL000921.30
2096m / Acanthocorbis unguiculata (97.3%)
14061 / Acanthocorbis unguiculata (97.1%)
2619 / Cosmoeca ventricosa (97.6%)
ASV_20249
788m / Bicosta minor (97.6%)
5457m / Bicosta minor (99.7%)
FJ153672 GoC3.C08
ASV_21614
DQ120005 NOR46.34
9747 / Acanthocorbis unguiculata (99.5%)
ASV_22909
2428m / Choanoflagellata–sp29
13029 / Choanoflagellata–sp47
ASV_12488m
2239m / Choanoflagellata–sp32
18102 / Choanoflagellata–sp67
5948 / Choanoflagellata–sp31
ASV_24041
EU011923 Diplotheca costata
ASV_6740m
6304 / Choanoflagellata–sp36
16740 / Choanoflagellata–sp55
7528 / Choanoflagellata–sp43
15206 / Choanoflagellata–sp59
22484 / Choanoflagellata–sp68
8187 / Choanoflagellata–sp38
22164 / Choanoflagellata–sp65
2500m / Choanoflagellata–sp27
3003m / Choanoflagellata–sp34
ASV_11106
Clade D
4523m / Choanoflagellata–sp56
ASV_22003
EU446388 cLA12G11
ASV_21010
AF272000 Calliacantha sp.
L10823 Acanthocoepsis unguigulata
EU011922 Acanthoeca spectabilis
EU011928 Savillea micropora
ASV_19319
DQ310214 FV23.1A4
ASV_24472
Clade G
8650m / Choanoflagellata–sp52
ASV_19770
2841m / Choanoflagellata–sp24
8538 / Choanoflagellata–sp40
ASV_22695
1260 / Choanoflagellata–sp6
3511 / Choanoflagellata–sp22
2971 / Choanoflagellata–sp15
2549 / Choanoflagellata–sp16
10283 / Choanoflagellata–sp46
20027 / Choanoflagellata–sp66
3398 / Choanoflagellata–sp19
17812 / Choanoflagellata–sp60
AJ402331 OLI11013
1269 / Choanoflagellata–sp2
AY426845 BL000921.24
18695 / Choanoflagellata–sp62
AY426868 BL001221.16
1714 / Choanoflagellata–sp11
Clade E

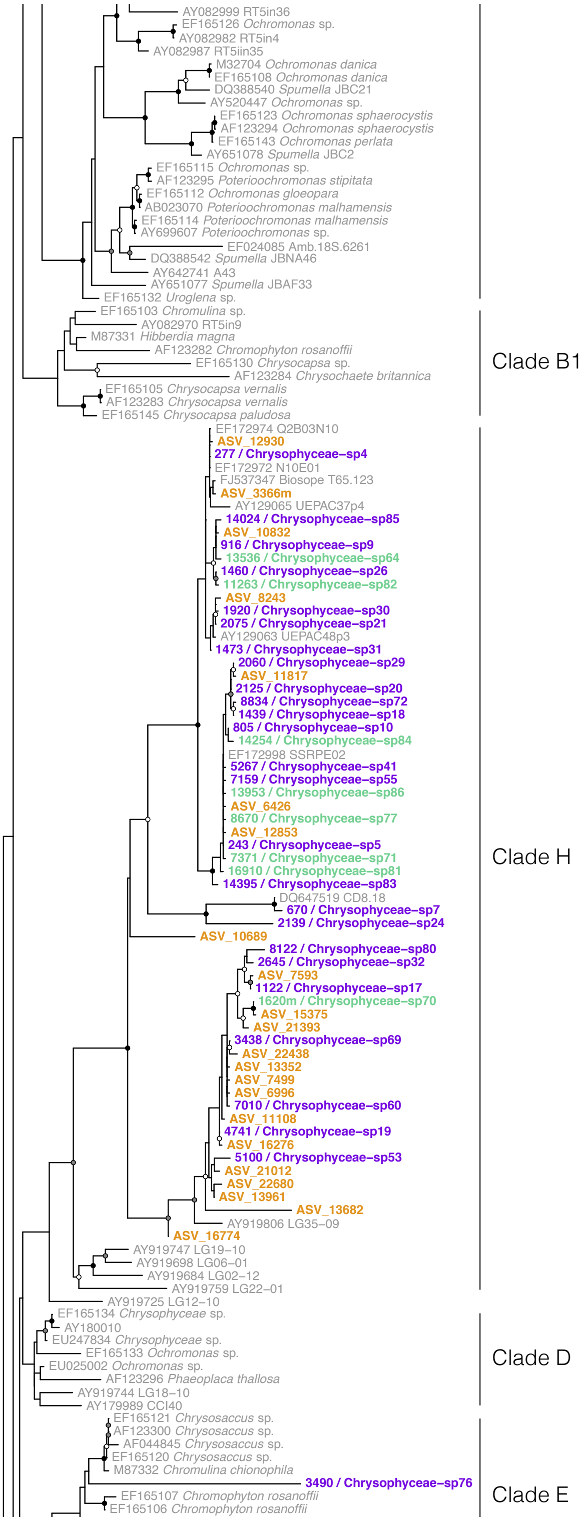
(D) Choanoflagellata (continued)


AJ402331 OLI11013
1269 / Choanoflagellata–sp2
AY426845 BL000921.24
18695 / Choanoflagellata–sp62
AY426868 BL001221.16
1714 / Choanoflagellata–sp11
589m / Choanoflagellata–sp9
2334 / Choanoflagellata–sp20
8098 / Choanoflagellata–sp37
1800 / Choanoflagellata–sp8
ASV_8411m
ASV_14635
Clade E
18055 / Choanoflagellata–sp63
ASV_17054
2205 / Choanoflagellata–sp10
915m / Choanoflagellata–sp7
ASV_19149
1980 / Choanoflagellata–sp3
1859 / Choanoflagellata–sp13
516 / Choanoflagellata–sp1
4985 / Choanoflagellata–sp26
ASV_19303
ASV_10368
ASV_23623
ASV_10898
ASV_22418
ASV_13108
22864 / Choanoflagellata–sp41
ASV_17360
ASV_19024
ASV_17708
ASV_23613
ASV_24328
ASV_17713
ASV_19764
ASV_23387
ASV_22246
EU446377 cLA12A08
EU446321 UI12G07
EU446411 cLA14H07
EU446354 UI13H07
EU446385 cLA12E05
Clade F
ASV_21193
ASV_12802
ASV_11718
ASV_18889
ASV_13784
8199 / Choanoflagellata–sp35
3185 / Choanoflagellata–sp23
14608 / Choanoflagellata–sp51
ASV_23610
ASV_15828
AY426842 BL000921.20
264m / Choanoflagellata–sp12
AY426933 BL010625.36
501m / Choanoflagellata–sp18
ASV_23126
ASV_18372
ASV_15629
ASV_24624
ASV_13733
DQ995807 Lagenoeca antarctica
AB275066 DSGM–66
Clade B
ASV_9314
12592 / Choanoflagellata–sp50
ASV_16670
ASV_11706
ASV_8960
ASV_13848
16546 / Choanoflagellata–sp53
ASV_12924
ASV_6828
ASV_22012
ASV_21788
6546 / Choanoflagellata–sp39
ASV_8118
ASV_12132
ASV_14576
ASV_17359
ASV_11527
ASV_11526
ASV_22908
ASV_17166
ASV_18383
ASV_16281
8835 / Choanoflagellata–sp61
ASV_19322
ASV_10064
ASV_8731
ASV_16267
EF023936 Amb.18S.1397
EF023856 Amb.18S.1307
EF024015 Amb.18S.1493
EF023626 Amb.18S.870
EF024012 Amb.18S.1490
EF023385 Amb.18S.720
EU011925 Salpingoeca amphoridium
DQ059032 Salpingoeca amphoridium
AY642707 P1.39
AF084231 Desmarella moniliformis
AF271999 Monosiga ovata
AF084230 Monosiga ovata
AY642728 PG5.16
EU011929 Salpingoeca napiformis
ASV_24062
Clade C
AY149898 Choanoeca perplexa
ASV_19036m
AY149896 Proterospongia choanojuncta
AY149897 Codonosiga gracilis
3109 / Choanoflagellata–sp45
15435 / Choanoflagellata–sp54
1618 / Choanoflagellata–sp4
1244 / Monosiga brevicollis (100%)
AF084618 Monosiga brevicollis
EU011924 Proterospongia sp.
AF100941 Salpingoeca infusionum
DQ310311 FV36.CilF8
AJ402325 OLI11041
Clade A
9485 / Choanoflagellata–sp58
2922 / Choanoflagellata–sp25
3595m / Choanoflagellata–sp30
EU011931 Salpingoeca urceolata
3506 / Salpingoeca prava (100%)
EU011930 Salpingoeca pyxidium
3284 / Salpingoeca kvevrii (98.2%)
AY821949 CV1.B2.17
AY821948 CV1.B1.36
Y16260 Sphaeroforma arctica
AY348876 Chondrosia reniformis

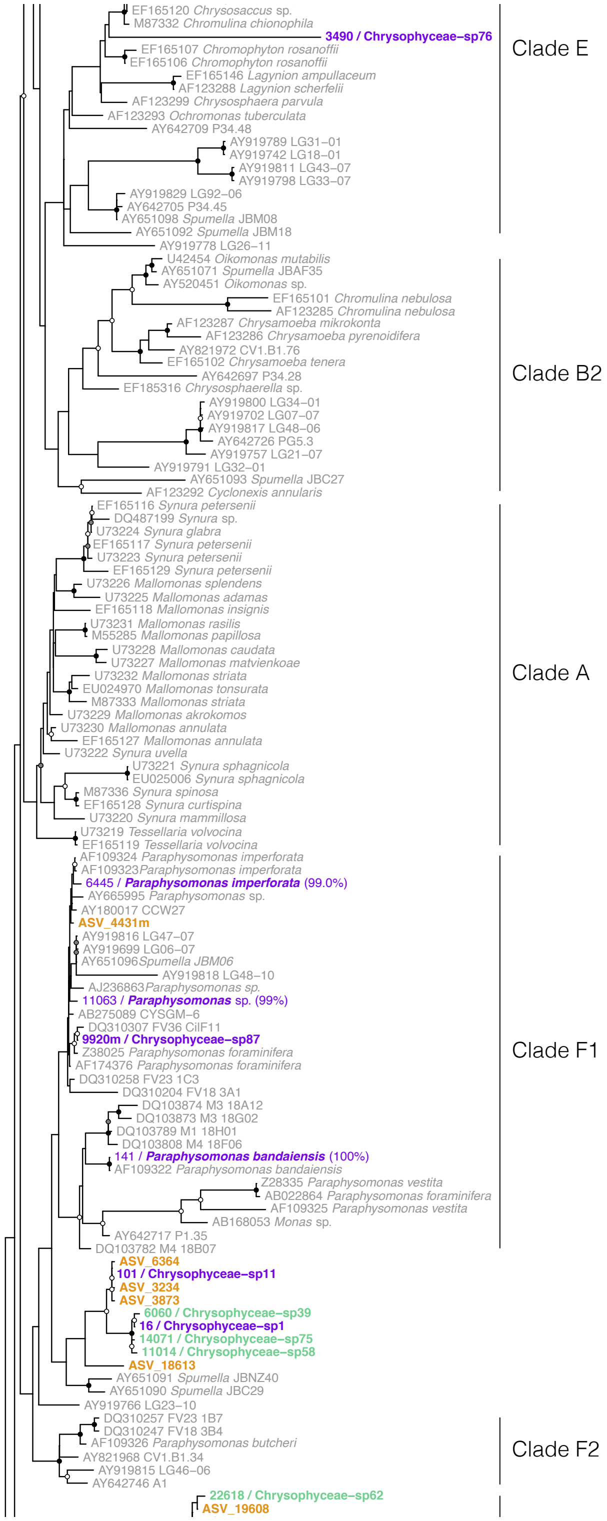
(E) Chrysophyceae


Extract ● Both ● DNA ● RNA
Bootstrap support ○ 60–80% ● 80–90% ● >90%
5370 / Spumella IOW86 (99.5%)
3265 / Spumella IOW86 (99.5%)
706 / Spumella IOW86 (99.7%)
1199 / Spumella IOW86 (99.0%)
2105 / Spumella IOW86 (99.7%)
1930 / Spumella IOW86 (99.7%)
1803 / Spumella IOW86 (99.5%)
1535 / Spumella IOW86 (99.7%)
15450 / Spumella IOW86 (99.0%)
1 / Spumella IOW86 (100%)
686 / Spumella IOW86 (99.7%)
2135 / Spumella IOW86 (99.7%)
EF633325 Chrysophyta JZH200700
AY651085 Spumella JBM/512
ASV_4114
EF023675 Amb.18S.936
EF023552 Amb.18S.772
AY651084 Spumella JBM19
307 / Pedospumella encystans (100%)
AJ236859 Spumella elongata
816 / Pedospumella encystans (99.5%)
AY651083 Spumella JBM/S11
614 / Spumella IOW86 (99.5%)
EF027354 Spumella sp.
6818 / Spumella IOW86 (99.2%)
11 / Spumella IOW86 (99.5%)
AJ236861 Spumella danica
AY651080 Spumella JBC13
AY651079 Spumella JBAS36
AY651081 Spumella JBC/S23
DQ310336 FV233A12
EF165131 Uroglena americana
AF123290 Uroglena americana
AY919807 LG35–11
AY919777 LG26–10
EU024983 Uroglena sp.
AY919717 LG10–03
EU247838 Ochromonadaceae sp.
EF165142 Ochromonas sp.
U42381 Ochromonas sp.
EF165139 Ochromonas sp.
ASV_21629
EF165138 Ochromonas marina
EF165137 Ochromonas sp.
EF165136 Ochromonas marina
EF165135 Ochromonas sp.
DQ310291 FV36 CilC7
DQ388565 Spumella 1036
DQ388551 Spumella 194f
AY651086 Spumella JBL14
DQ388557 Spumella 391f
AJ236857 Spumella 15G
AB425951 Spumella Mbc3C
DQ388560 Spumella 1020
AJ236858 Spumella 37G
DQ388561 Spumella 1026
EF023425 Amb.18S.766
DQ388543 Spumella 8b3
DQ388554 Spumella 45b3hm
AJ236862 Spumella SpiG
DQ388568 Spumella 1305
DQ388562 Spumella 1027
DQ388559 Spumella 1013
AY651088 Spumella JBNZ39
AJ236860 Spumella obliqua
AY651089 Spumella JBM28
EU076737 Dinobryon divergens
EU076736 Dinobryon divergens
EU025019 Dinobryon divergens
EU076735 Dinobryon bavaricum
EU024973 Dinobryon bavaricum
EU024976 Dinobryon divergens
EU024980 Dinobryon crenulatum
EU024975 Dinobryon sociale
EU024993 Dinobryon pediforme
EF165140 Dinobryon cylindricum
AF123291 Dinobryon sociale
AF123289 Dinobryon sertularia
AY919796 LG33–02
AY919719 LG10–11
EF165141 Dinobryon sociale
4773m / Chrysophyceae–sp61
ASV_11569
2991 / Chrysophyceae–sp40
9105m / Dinobryon faculiferum (97.9%)
ASV_22887
ASV_12308
AY919752 LG20–09
AB275090 CYSGM–7
3006m / Chrysophyceae–sp63
DQ310261 FV23 1B1
AF123302 Chrysoxys sp.
7691m / Chrysophyceae–sp78
U42382 Ochromonas sp.
EF165124 Ochromonas aestuarii
AF123301 Epipyxis aurea
AF123298 Epipyxis pulchra
U71196 Chrysonephele palustris
AF123297 Chrysolepidomonas dendrolepidota
EF165110 Ochromonas sp.
AY919762 LG22–12
EF165111 Ochromonas vasocystis
AY919828 LG81–06
AY919824 LG73–06
AY651097 Spumella JBC07
AY642745 A34
AY082999 RT5in36
EF165126 Ochromonas sp.
AY082982 RT5in4
AY082987 RT5iin35
Clade C

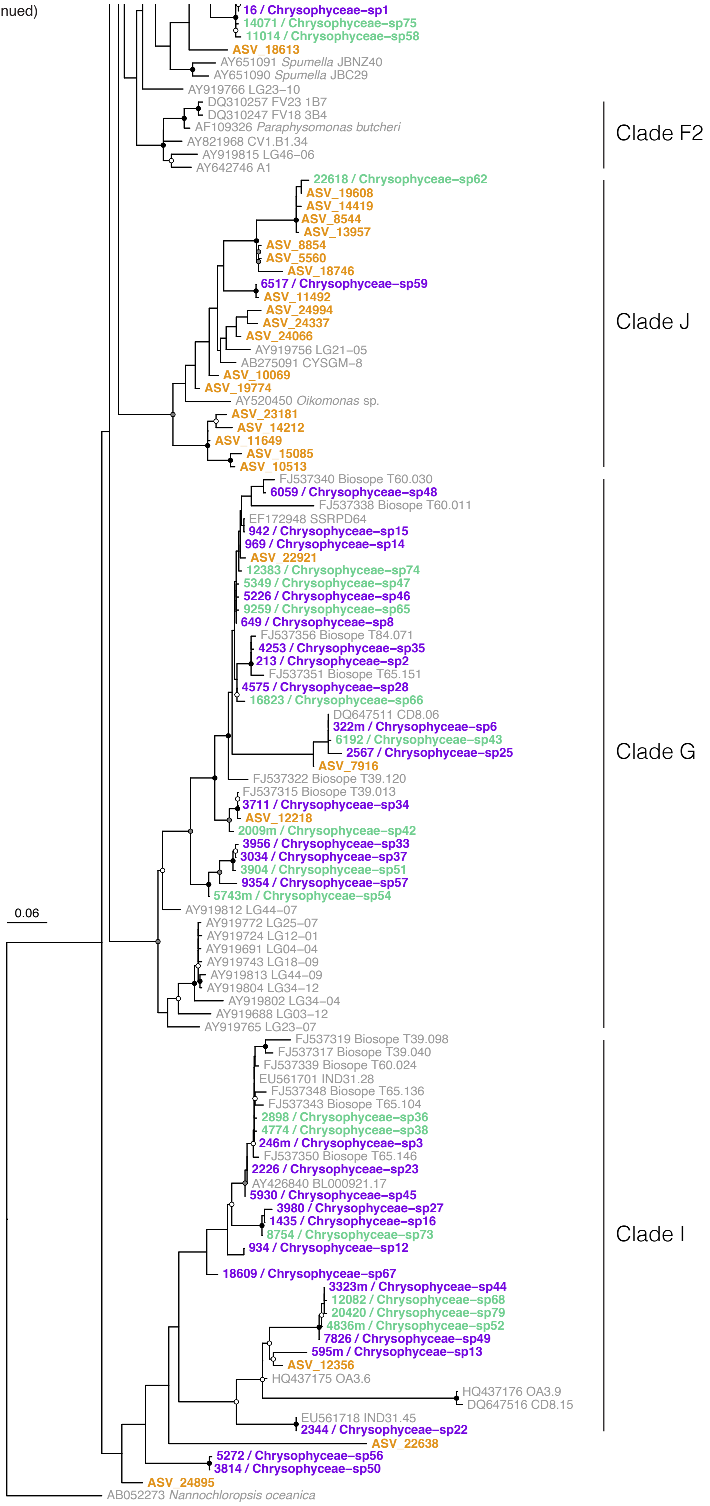
(E) Chrysophyceae (continued)


AY082999 RT5in36
EF165126 Ochromonas sp.
AY082982 RT5in4
AY082987 RT5iin35
M32704 Ochromonas danica
EF165108 Ochromonas danica
DQ388540 Spumella JBC21
AY520447 Ochromonas sp.
EF165123 Ochromonas sphaerocystis
AF123294 Ochromonas sphaerocystis
EF165143 Ochromonas perlata
AY651078 Spumella JBC2
EF165115 Ochromonas sp.
AF123295 Poterioochromonas stipitata
EF165112 Ochromonas gloeopara
AB023070 Poterioochromonas malhamensis
EF165114 Poterioochromonas malhamensis
AY699607 Poterioochromonas sp.
EF024085 Amb.18S.6261
DQ388542 Spumella JBNA46
AY642741 A43
AY651077 Spumella JBAF33
EF165132 Uroglena sp.
EF165103 Chromulina sp.
AY082970 RT5in9
M87331 Hibberdia magna
AF123282 Chromophyton rosanoffii
EF165130 Chrysocapsa sp.
AF123284 Chrysochaete britannica
EF165105 Chrysocapsa vernalis
AF123283 Chrysocapsa vernalis
EF165145 Chrysocapsa paludosa
Clade B1
EF172974 Q2B03N10
ASV_12930
277 / Chrysophyceae–sp4
EF172972 N10E01
FJ537347 Biosope T65.123
ASV_3366m
AY129065 UEPAC37p4
14024 / Chrysophyceae–sp85
ASV_10832
916 / Chrysophyceae–sp9
13536 / Chrysophyceae–sp64
1460 / Chrysophyceae–sp26
11263 / Chrysophyceae–sp82
ASV_8243
1920 / Chrysophyceae–sp30
2075 / Chrysophyceae–sp21
AY129063 UEPAC48p3
1473 / Chrysophyceae–sp31
2060 / Chrysophyceae–sp29
ASV_11817
2125 / Chrysophyceae–sp20
8834 / Chrysophyceae–sp72
1439 / Chrysophyceae–sp18
805 / Chrysophyceae–sp10
14254 / Chrysophyceae–sp84
EF172998 SSRPE02
5267 / Chrysophyceae–sp41
7159 / Chrysophyceae–sp55
13953 / Chrysophyceae–sp86
ASV_6426
8670 / Chrysophyceae–sp77
ASV_12853
243 / Chrysophyceae–sp5
7371 / Chrysophyceae–sp71
16910 / Chrysophyceae–sp81
14395 / Chrysophyceae–sp83
DQ647519 CD8.18
670 / Chrysophyceae–sp7
2139 / Chrysophyceae–sp24
ASV_10689
8122 / Chrysophyceae–sp80
2645 / Chrysophyceae–sp32
ASV_7593
1122 / Chrysophyceae–sp17
1620m / Chrysophyceae–sp70
ASV_15375
ASV_21393
3438 / Chrysophyceae–sp69
ASV_22438
ASV_13352
ASV_7499
ASV_6996
7010 / Chrysophyceae–sp60
ASV_11108
4741 / Chrysophyceae–sp19
ASV_16276
5100 / Chrysophyceae–sp53
ASV_21012
ASV_22680
ASV_13961
ASV_13682
AY919806 LG35–09
ASV_16774
AY919747 LG19–10
AY919698 LG06–01
AY919684 LG02–12
AY919759 LG22–01
AY919725 LG12–10
Clade H
EF165134 Chrysophyceae sp.
AY180010
EU247834 Chrysophyceae sp.
EF165133 Ochromonas sp.
EU025002 Ochromonas sp.
AF123296 Phaeoplaca thallosa
AY919744 LG18–10
AY179989 CCI40
Clade D
EF165121 Chrysosaccus sp.
AF123300 Chrysosaccus sp.
AF044845 Chrysosaccus sp.
EF165120 Chrysosaccus sp.
M87332 Chromulina chionophila
3490 / Chrysophyceae–sp76
EF165107 Chromophyton rosanoffii
EF165106 Chromophyton rosanoffii
Clade E

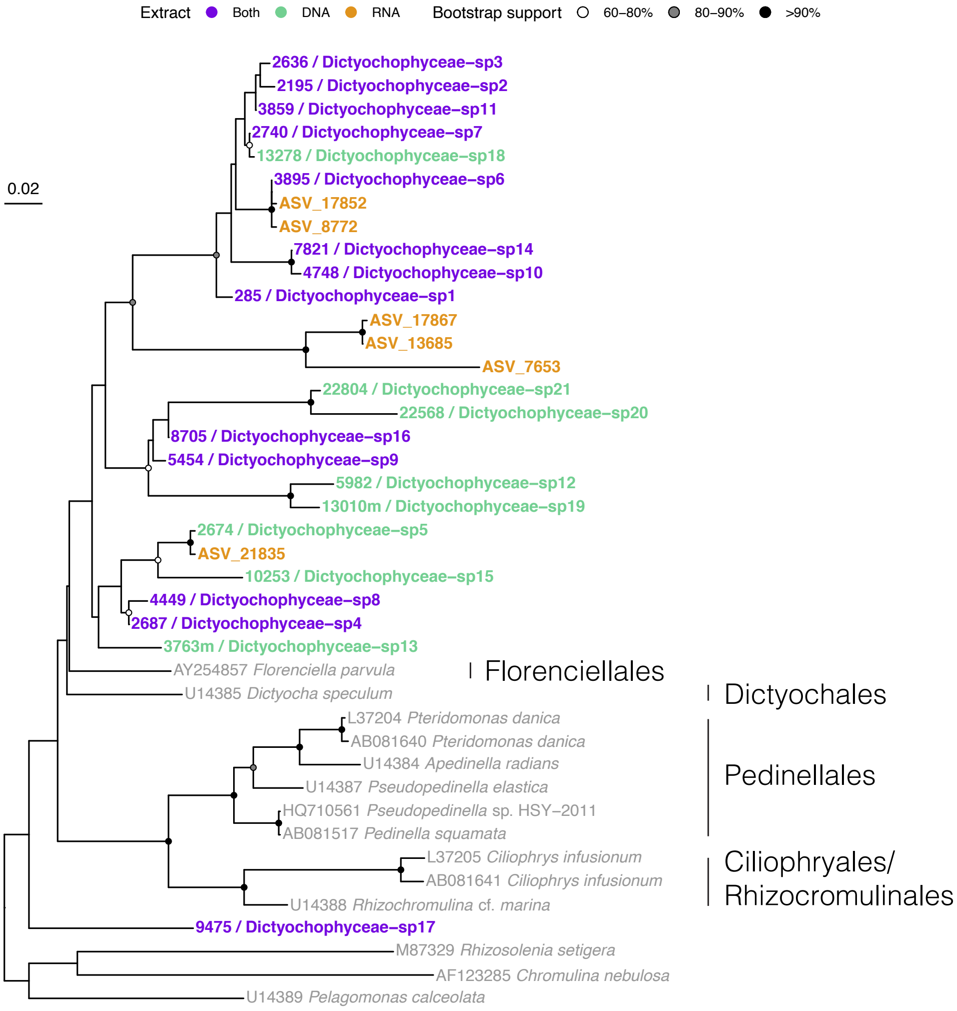
(E) Chrysophyceae (continued)


EF165120 Chrysosaccus sp.
M87332 Chromulina chionophila
3490 / Chrysophyceae–sp76
EF165107 Chromophyton rosanoffii
EF165106 Chromophyton rosanoffii
EF165146 Lagynion ampullaceum
AF123288 Lagynion scherfelii
AF123299 Chrysosphaera parvula
AF123293 Ochromonas tuberculata
AY642709 P34.48
AY919789 LG31–01
AY919742 LG18–01
AY919811 LG43–07
AY919798 LG33–07
AY919829 LG92–06
AY642705 P34.45
AY651098 Spumella JBM08
AY651092 Spumella JBM18
AY919778 LG26–11
Clade E
U42454 Oikomonas mutabilis
AY651071 Spumella JBAF35
AY520451 Oikomonas sp.
EF165101 Chromulina nebulosa
AF123285 Chromulina nebulosa
AF123287 Chrysamoeba mikrokonta
AF123286 Chrysamoeba pyrenoidifera
AY821972 CV1.B1.76
EF165102 Chrysamoeba tenera
AY642697 P34.28
EF185316 Chrysosphaerella sp.
AY919800 LG34–01
AY919702 LG07–07
AY919817 LG48–06
AY642726 PG5.3
AY919757 LG21–07
AY919791 LG32–01
AY651093 Spumella JBC27
AF123292 Cyclonexis annularis
Clade B2
EF165116 Synura petersenii
DQ487199 Synura sp.
U73224 Synura glabra
EF165117 Synura petersenii
U73223 Synura petersenii
EF165129 Synura petersenii
U73226 Mallomonas splendens
U73225 Mallomonas adamas
EF165118 Mallomonas insignis
U73231 Mallomonas rasilis
M55285 Mallomonas papillosa
U73228 Mallomonas caudata
U73227 Mallomonas matvienkoae
U73232 Mallomonas striata
EU024970 Mallomonas tonsurata
M87333 Mallomonas striata
U73229 Mallomonas akrokomos
U73230 Mallomonas annulata
EF165127 Mallomonas annulata
U73222 Synura uvella
U73221 Synura sphagnicola
EU025006 Synura sphagnicola
M87336 Synura spinosa
EF165128 Synura curtispina
U73220 Synura mammillosa
U73219 Tessellaria volvocina
EF165119 Tessellaria volvocina
Clade A
AF109324 Paraphysomonas imperforata
AF109323Paraphysomonas imperforata
6445 / Paraphysomonas imperforata (99.0%)
AY665995 Paraphysomonas sp.
AY180017 CCW27
ASV_4431m
AY919816 LG47–07
AY919699 LG06–07
AY651096Spumella JBM06
AY919818 LG48–10
AJ236863Paraphysomonas sp.
11063 / Paraphysomonas sp. (99%)
AB275089 CYSGM–6
DQ310307 FV36 CilF11
9920m / Chrysophyceae–sp87
Z38025 Paraphysomonas foraminifera
AF174376 Paraphysomonas foraminifera
DQ310258 FV23 1C3
DQ310204 FV18 3A1
DQ103874 M3 18A12
DQ103873 M3 18G02
DQ103789 M1 18H01
DQ103808 M4 18F06
141 / Paraphysomonas bandaiensis (100%)
AF109322 Paraphysomonas bandaiensis
Z28335 Paraphysomonas vestita
AB022864 Paraphysomonas foraminifera
AF109325 Paraphysomonas vestita
AB168053 Monas sp.
AY642717 P1.35
DQ103782 M4 18B07
ASV_6364
101 / Chrysophyceae–sp11
ASV_3234
ASV_3873
6060 / Chrysophyceae–sp39
16 / Chrysophyceae–sp1
14071 / Chrysophyceae–sp75
11014 / Chrysophyceae–sp58
ASV_18613
AY651091 Spumella JBNZ40
AY651090 Spumella JBC29
AY919766 LG23–10
Clade F1
DQ310257 FV23 1B7
DQ310247 FV18 3B4
AF109326 Paraphysomonas butcheri
AY821968 CV1.B1.34
AY919815 LG46–06
AY642746 A1
Clade F2
22618 / Chrysophyceae–sp62
ASV_19608

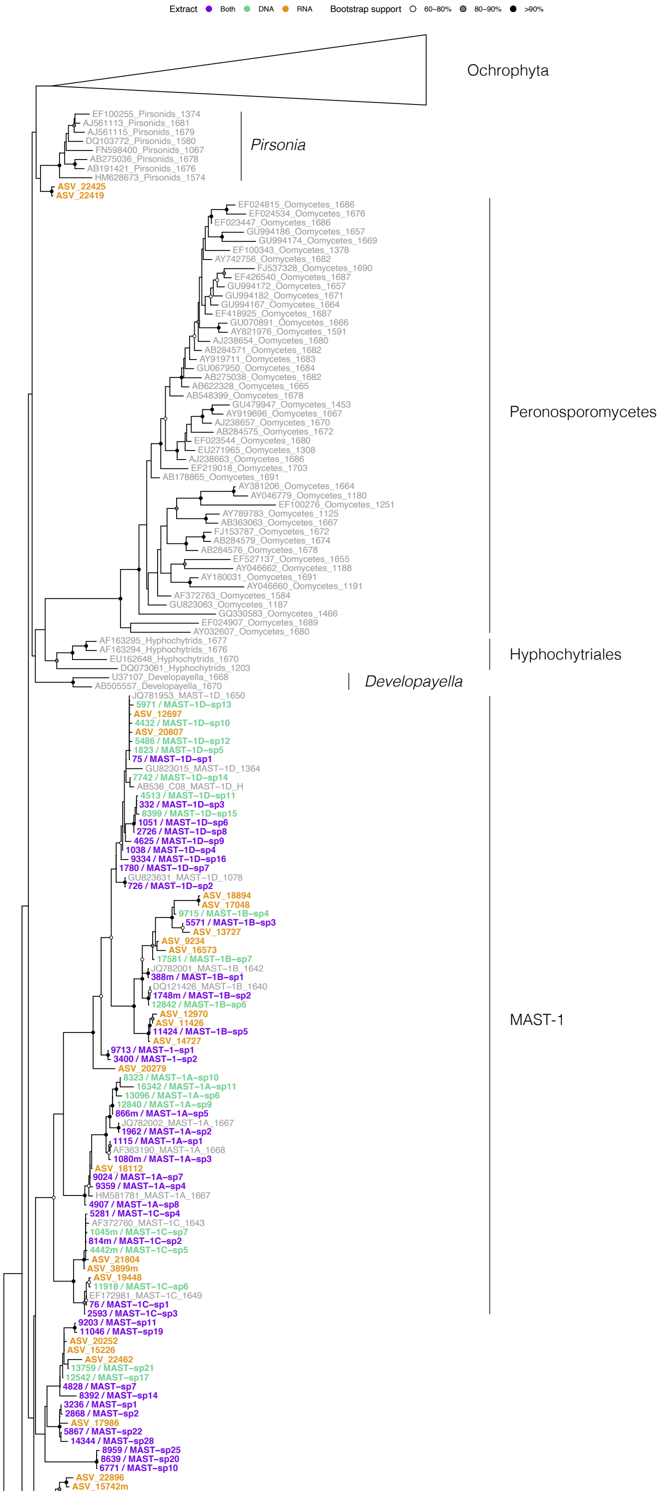
(E) Chrysophyceae (continued)

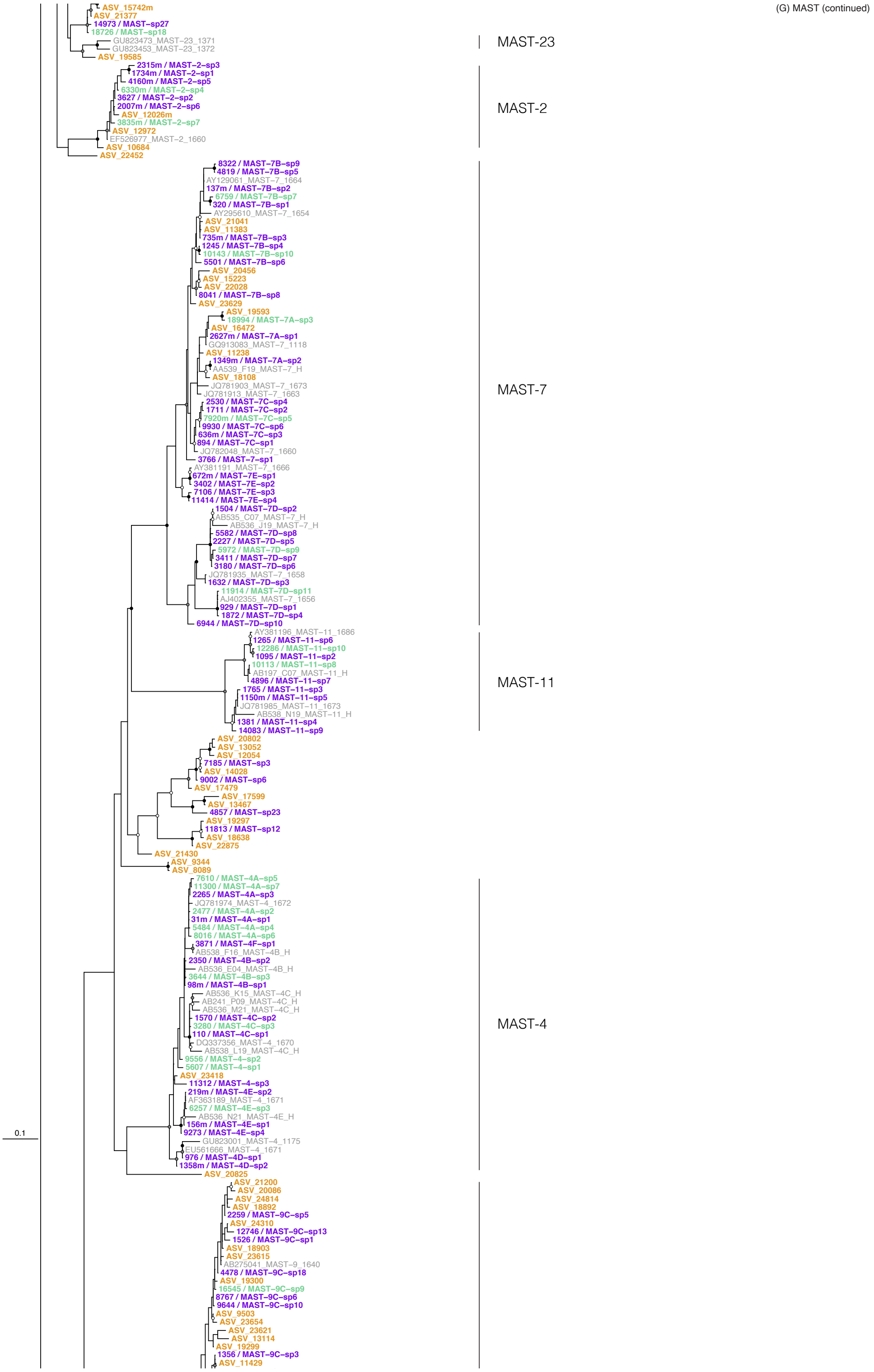


(F) Dictyochophyceae


Extract ● Both ● DNA ● RNA
Bootstrap support ○ 60–80% ● 80–90% ● >90%
0.02
2636 / Dictyochophyceae–sp3
2195 / Dictyochophyceae–sp2
3859 / Dictyochophyceae–sp11
2740 / Dictyochophyceae–sp7
13278 / Dictyochophyceae–sp18
3895 / Dictyochophyceae–sp6
ASV_17852
ASV_8772
7821 / Dictyochophyceae–sp14
4748 / Dictyochophyceae–sp10
285 / Dictyochophyceae–sp1
ASV_17867
ASV_13685
ASV_7653
22804 / Dictyochophyceae–sp21
22568 / Dictyochophyceae–sp20
8705 / Dictyochophyceae–sp16
5454 / Dictyochophyceae–sp9
5982 / Dictyochophyceae–sp12
13010m / Dictyochophyceae–sp19
2674 / Dictyochophyceae–sp5
ASV_21835
10253 / Dictyochophyceae–sp15
4449 / Dictyochophyceae–sp8
2687 / Dictyochophyceae–sp4
3763m / Dictyochophyceae–sp13
AY254857 *Florenciella parvula*
Florenciellales
U14385 *Dictyocha speculum*
Dictyochales
L37204 *Pteridomonas danica*
AB081640 *Pteridomonas danica*
U14384 *Apedinella radians*
U14387 *Pseudopedinella elastica*
HQ710561 *Pseudopedinella* sp. HSY–2011
AB081517 *Pedinella squamata*
Pedinellales
L37205 *Ciliophrys infusionum*
AB081641 *Ciliophrys infusionum*
U14388 *Rhizochromulina* cf. *marina*
Ciliophryales/
Rhizocromulinales
9475 / Dictyochophyceae–sp17
M87329 *Rhizosolenia setigera*
AF123285 *Chromulina nebulosa*
U14389 *Pelagomonas calceolata*

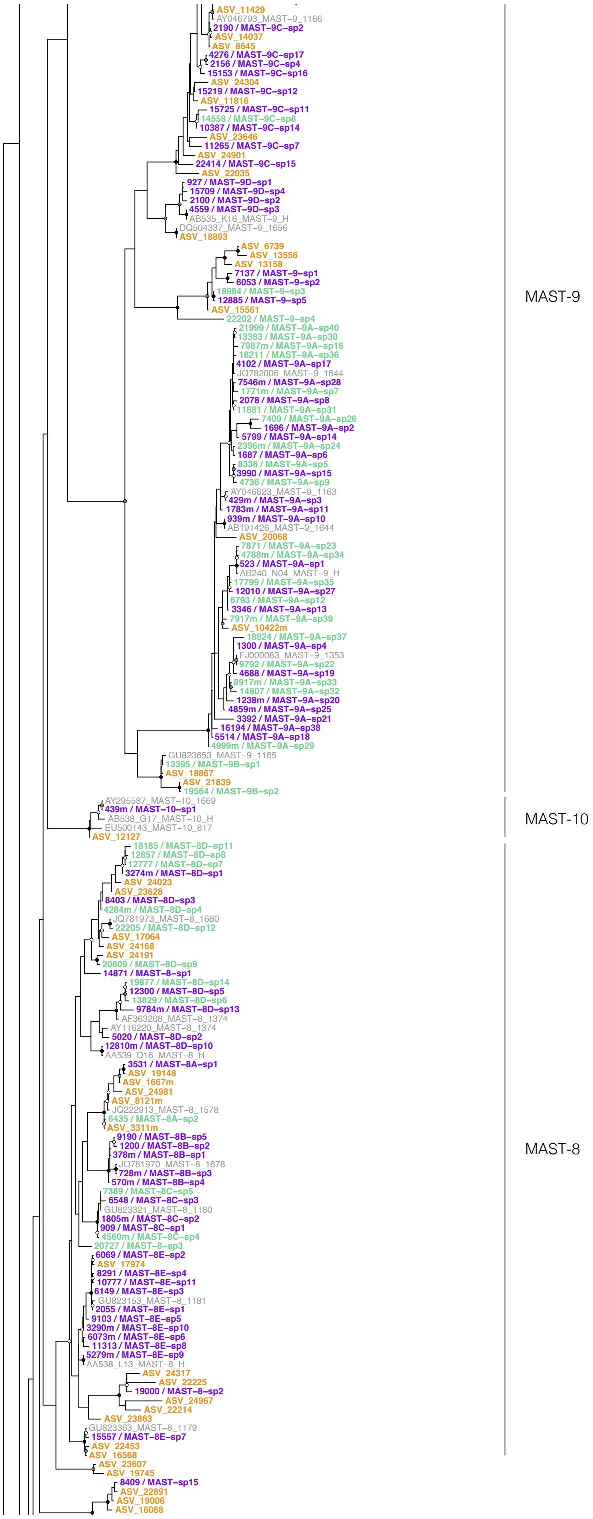
(G) MAST

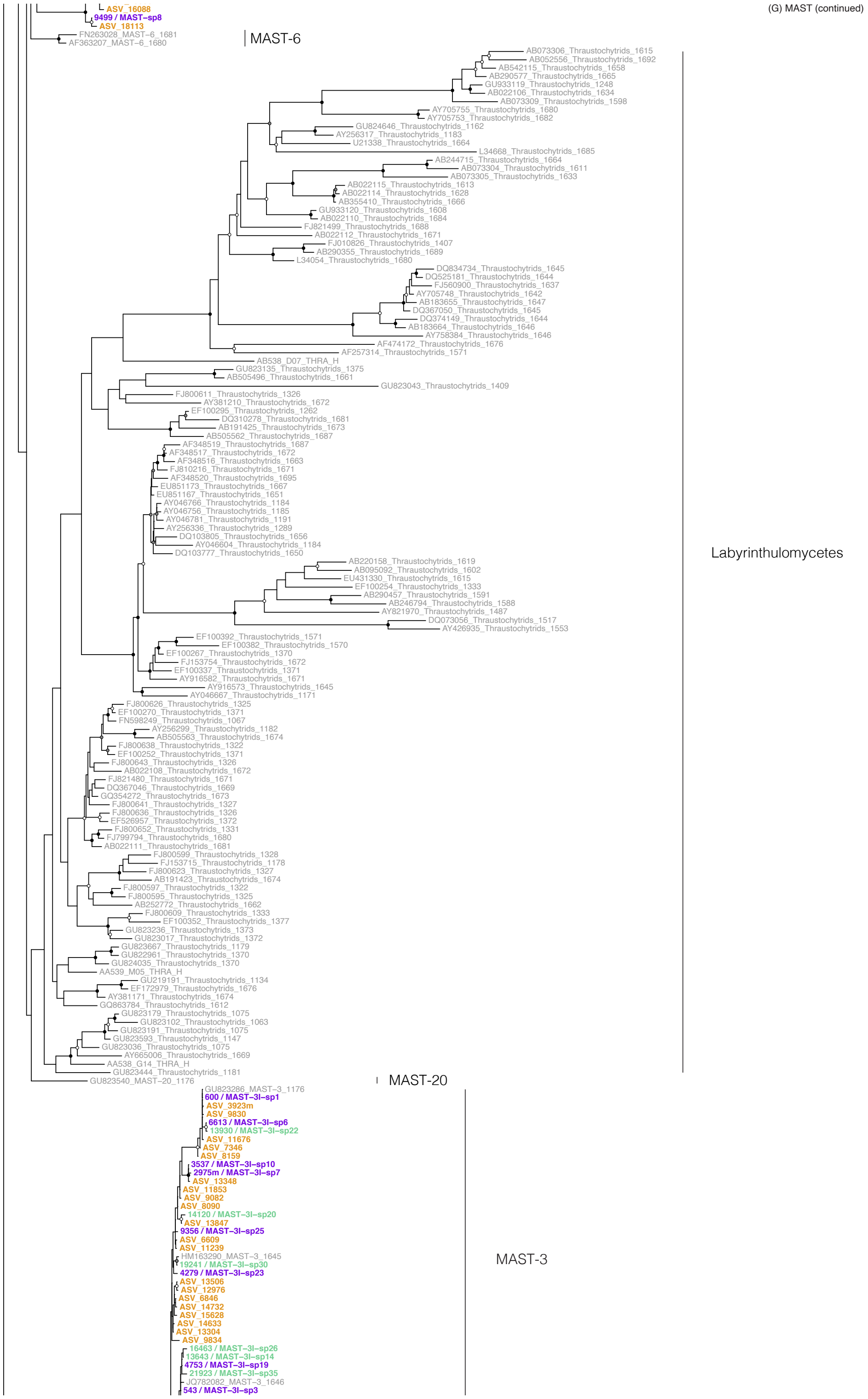

Extract ● Both ● DNA ● RNA
Bootstrap support ○ 60–80% ● 80–90% ● >90%
Ochrophyta
EF100255_Pirsonids_1374
AJ561113_Pirsonids_1681
AJ561115_Pirsonids_1679
DQ103772_Pirsonids_1580
FN598400_Pirsonids_1067
AB275036_Pirsonids_1678
AB191421_Pirsonids_1676
HM628673_Pirsonids_1574
Pirsonia
ASV_22425
ASV_22419
EF024815_Oomycetes_1686
EF024534_Oomycetes_1676
EF023447_Oomycetes_1686
GU994186_Oomycetes_1657
GU994174_Oomycetes_1669
EF100343_Oomycetes_1378
AY742756_Oomycetes_1682
FJ537328_Oomycetes_1690
EF426540_Oomycetes_1687
GU994172_Oomycetes_1657
GU994182_Oomycetes_1671
GU994167_Oomycetes_1664
EF418925_Oomycetes_1687
GU070891_Oomycetes_1666
AY821976_Oomycetes_1591
AJ238654_Oomycetes_1680
AB284571_Oomycetes_1682
AY919711_Oomycetes_1683
GU067950_Oomycetes_1684
AB275038_Oomycetes_1682
AB622328_Oomycetes_1665
AB548399_Oomycetes_1678
GU479947_Oomycetes_1453
AY919696_Oomycetes_1667
AJ238657_Oomycetes_1670
AB284575_Oomycetes_1672
EF023544_Oomycetes_1680
EU271965_Oomycetes_1308
AJ238663_Oomycetes_1686
EF219018_Oomycetes_1703
AB178865_Oomycetes_1691
AY381206_Oomycetes_1664
AY046779_Oomycetes_1180
EF100276_Oomycetes_1251
AY789783_Oomycetes_1125
AB363063_Oomycetes_1667
FJ153787_Oomycetes_1672
AB284579_Oomycetes_1674
AB284576_Oomycetes_1678
EF527137_Oomycetes_1655
AY046662_Oomycetes_1188
AY180031_Oomycetes_1691
AY046660_Oomycetes_1191
AF372763_Oomycetes_1584
GU823063_Oomycetes_1187
GQ330583_Oomycetes_1466
EF024907_Oomycetes_1689
AY032607_Oomycetes_1680
Peronosporomycetes
AF163295_Hyphochytrids_1677
AF163294_Hyphochytrids_1676
EU162648_Hyphochytrids_1670
DQ073061_Hyphochytrids_1203
Hyphochytriales
U37107_Developayella_1668
AB505557_Developayella_1670
Developayella
JQ781953_MAST−1D_1650
5971 / MAST−1D−sp13
ASV_12697
4432 / MAST−1D−sp10
ASV_20807
5486 / MAST−1D−sp12
1823 / MAST−1D−sp5
75 / MAST−1D−sp1
GU823015_MAST−1D_1364
7742 / MAST−1D−sp14
AB536_C08_MAST−1D_H
4513 / MAST−1D−sp11
332 / MAST−1D−sp3
8399 / MAST−1D−sp15
1051 / MAST−1D−sp6
2726 / MAST−1D−sp8
4625 / MAST−1D−sp9
1038 / MAST−1D−sp4
9334 / MAST−1D−sp16
1780 / MAST−1D−sp7
GU823631_MAST−1D_1078
726 / MAST−1D−sp2
ASV_18894
ASV_17048
9715 / MAST−1B−sp4
5571 / MAST−1B−sp3
ASV_13727
ASV_9234
ASV_16573
17581 / MAST−1B−sp7
JQ782001_MAST−1B_1642
388m / MAST−1B−sp1
DQ121426_MAST−1B_1640
1748m / MAST−1B−sp2
12842 / MAST−1B−sp6
ASV_12970
ASV_11426
11424 / MAST−1B−sp5
ASV_14727
9713 / MAST−1−sp1
3400 / MAST−1−sp2
ASV_20279
8323 / MAST−1A−sp10
16342 / MAST−1A−sp11
13096 / MAST−1A−sp6
12840 / MAST−1A−sp9
866m / MAST−1A−sp5
JQ782002_MAST−1A_1667
1962 / MAST−1A−sp2
1115 / MAST−1A−sp1
AF363190_MAST−1A_1668
1080m / MAST−1A−sp3
ASV_18112
9024 / MAST−1A−sp7
9359 / MAST−1A−sp4
HM581781_MAST−1A_1667
4907 / MAST−1A−sp8
5281 / MAST−1C−sp4
AF372760_MAST−1C_1643
1045m / MAST−1C−sp7
814m / MAST−1C−sp2
4442m / MAST−1C−sp5
ASV_21804
ASV_3899m
ASV_19448
11918 / MAST−1C−sp6
EF172981_MAST−1C_1649
76 / MAST−1C−sp1
2593 / MAST−1C−sp3
MAST-1
9203 / MAST−sp11
11046 / MAST−sp19
ASV_20252
ASV_15226
ASV_22462
13759 / MAST−sp21
12542 / MAST−sp17
4828 / MAST−sp7
8392 / MAST−sp14
3236 / MAST−sp1
2868 / MAST−sp2
ASV_17986
5867 / MAST−sp22
14344 / MAST−sp28
8959 / MAST−sp25
8639 / MAST−sp20
6771 / MAST−sp10
ASV_22896
ASV_15742m



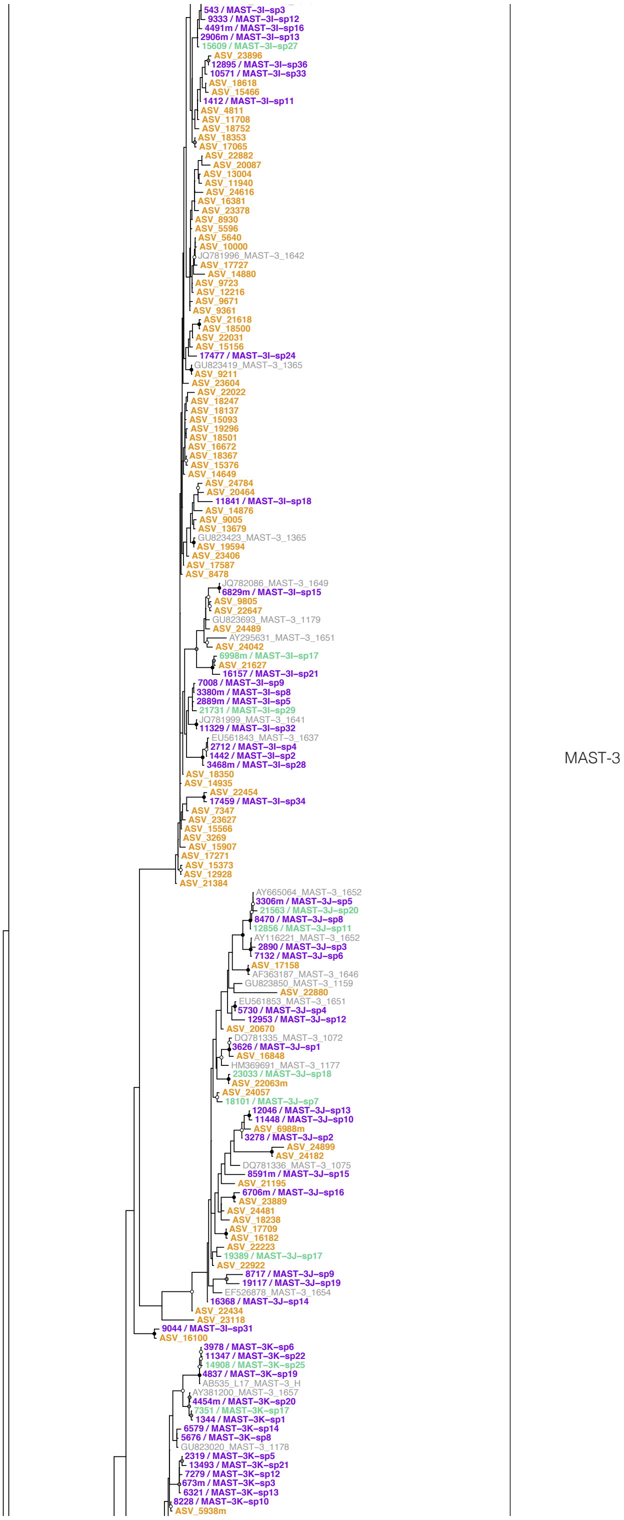
(G) MAST (continued)

(G) MAST (continued)

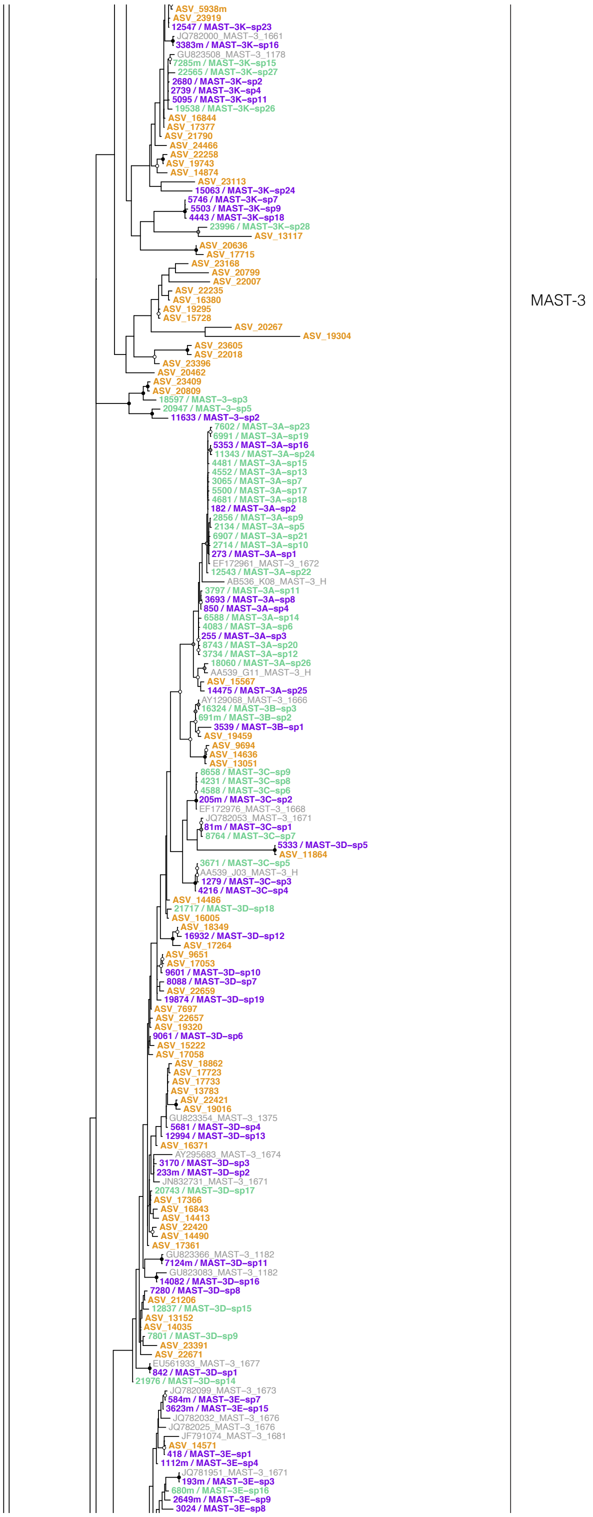


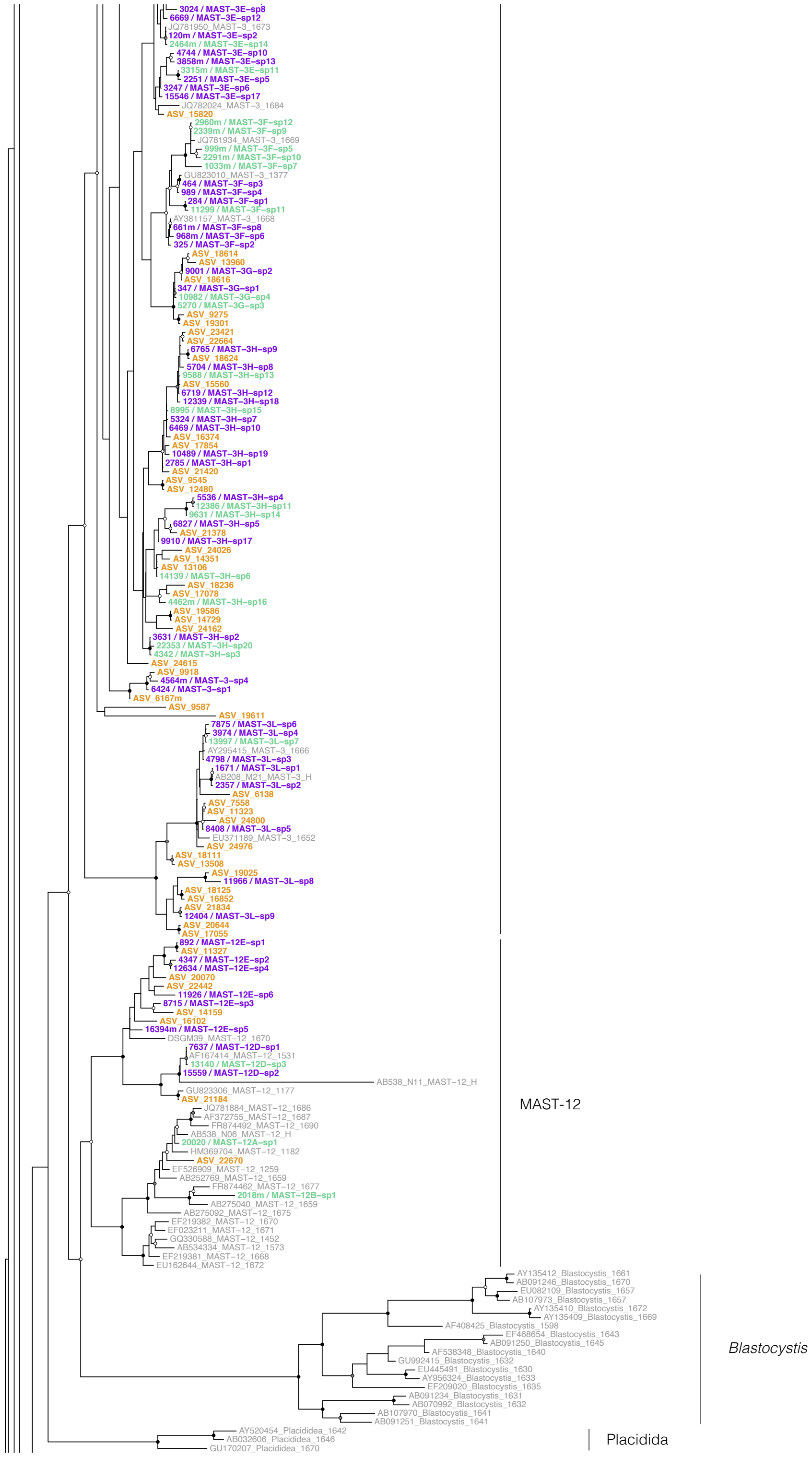

(G) MAST (continued)
ASV_16088
9499 / MAST–sp8
ASV_18113
FN263028_MAST–6_1681
AF363207_MAST–6_1680
MAST-6
AB073306_Thraustochytrids_1615
AB052556_Thraustochytrids_1692
AB542115_Thraustochytrids_1658
AB290577_Thraustochytrids_1665
GU933119_Thraustochytrids_1248
AB022106_Thraustochytrids_1634
AB073309_Thraustochytrids_1598
AY705755_Thraustochytrids_1680
AY705753_Thraustochytrids_1682
GU824646_Thraustochytrids_1162
AY256317_Thraustochytrids_1183
U21338_Thraustochytrids_1664
L34668_Thraustochytrids_1685
AB244715_Thraustochytrids_1664
AB073304_Thraustochytrids_1611
AB073305_Thraustochytrids_1633
AB022115_Thraustochytrids_1613
AB022114_Thraustochytrids_1628
AB355410_Thraustochytrids_1666
GU933120_Thraustochytrids_1608
AB022110_Thraustochytrids_1684
FJ821499_Thraustochytrids_1688
AB022112_Thraustochytrids_1671
FJ010826_Thraustochytrids_1407
AB290355_Thraustochytrids_1689
L34054_Thraustochytrids_1680
DQ834734_Thraustochytrids_1645
DQ525181_Thraustochytrids_1644
FJ560900_Thraustochytrids_1637
AY705748_Thraustochytrids_1642
AB183655_Thraustochytrids_1647
DQ367050_Thraustochytrids_1645
DQ374149_Thraustochytrids_1644
AB183664_Thraustochytrids_1646
AY758384_Thraustochytrids_1646
AF474172_Thraustochytrids_1676
AF257314_Thraustochytrids_1571
AB538_D07_THRA_H
GU823135_Thraustochytrids_1375
AB505496_Thraustochytrids_1661
GU823043_Thraustochytrids_1409
FJ800611_Thraustochytrids_1326
AY381210_Thraustochytrids_1672
EF100295_Thraustochytrids_1262
DQ310278_Thraustochytrids_1681
AB191425_Thraustochytrids_1673
AB505562_Thraustochytrids_1687
AF348519_Thraustochytrids_1687
AF348517_Thraustochytrids_1672
AF348516_Thraustochytrids_1663
FJ810216_Thraustochytrids_1671
AF348520_Thraustochytrids_1695
EU851173_Thraustochytrids_1667
EU851167_Thraustochytrids_1651
AY046766_Thraustochytrids_1184
AY046756_Thraustochytrids_1185
AY046781_Thraustochytrids_1191
AY256336_Thraustochytrids_1289
DQ103805_Thraustochytrids_1656
AY046604_Thraustochytrids_1184
DQ103777_Thraustochytrids_1650
AB220158_Thraustochytrids_1619
AB095092_Thraustochytrids_1602
EU431330_Thraustochytrids_1615
EF100254_Thraustochytrids_1333
AB290457_Thraustochytrids_1591
AB246794_Thraustochytrids_1588
AY821970_Thraustochytrids_1487
DQ073056_Thraustochytrids_1517
AY426935_Thraustochytrids_1553
Labyrinthulomycetes
EF100392_Thraustochytrids_1571
EF100382_Thraustochytrids_1570
EF100267_Thraustochytrids_1370
FJ153754_Thraustochytrids_1672
EF100337_Thraustochytrids_1371
AY916582_Thraustochytrids_1671
AY916573_Thraustochytrids_1645
AY046667_Thraustochytrids_1171
FJ800626_Thraustochytrids_1325
EF100270_Thraustochytrids_1371
FN598249_Thraustochytrids_1067
AY256299_Thraustochytrids_1182
AB505563_Thraustochytrids_1674
FJ800638_Thraustochytrids_1322
EF100252_Thraustochytrids_1371
FJ800643_Thraustochytrids_1326
AB022108_Thraustochytrids_1672
FJ821480_Thraustochytrids_1671
DQ367046_Thraustochytrids_1669
GQ354272_Thraustochytrids_1673
FJ800641_Thraustochytrids_1327
FJ800636_Thraustochytrids_1326
EF526957_Thraustochytrids_1372
FJ800652_Thraustochytrids_1331
FJ799794_Thraustochytrids_1680
AB022111_Thraustochytrids_1681
FJ800599_Thraustochytrids_1328
FJ153715_Thraustochytrids_1178
FJ800623_Thraustochytrids_1327
AB191423_Thraustochytrids_1674
FJ800597_Thraustochytrids_1322
FJ800595_Thraustochytrids_1325
AB252772_Thraustochytrids_1662
FJ800609_Thraustochytrids_1333
EF100352_Thraustochytrids_1377
GU823236_Thraustochytrids_1373
GU823017_Thraustochytrids_1372
GU823667_Thraustochytrids_1179
GU822961_Thraustochytrids_1370
GU824035_Thraustochytrids_1370
AA539_M05_THRA_H
GU219191_Thraustochytrids_1134
EF172979_Thraustochytrids_1676
AY381171_Thraustochytrids_1674
GQ863784_Thraustochytrids_1612
GU823179_Thraustochytrids_1075
GU823102_Thraustochytrids_1063
GU823191_Thraustochytrids_1075
GU823593_Thraustochytrids_1147
GU823036_Thraustochytrids_1075
AY665006_Thraustochytrids_1669
AA538_G14_THRA_H
GU823444_Thraustochytrids_1181
GU823540_MAST–20_1176
MAST-20
GU823286_MAST–3_1176
600 / MAST–3I–sp1
ASV_3923m
ASV_9830
6613 / MAST–3I–sp6
13930 / MAST–3I–sp22
ASV_11676
ASV_7346
ASV_8159
3537 / MAST–3I–sp10
2975m / MAST–3I–sp7
ASV_13348
ASV_11853
ASV_9082
ASV_8090
14120 / MAST–3I–sp20
ASV_13847
9356 / MAST–3I–sp25
ASV_6609
ASV_11239
HM163290_MAST–3_1645
19241 / MAST–3I–sp30
4279 / MAST–3I–sp23
ASV_13506
ASV_12976
ASV_6846
ASV_14732
ASV_15628
ASV_14633
ASV_13304
ASV_9834
16463 / MAST–3I–sp26
13643 / MAST–3I–sp14
4753 / MAST–3I–sp19
21923 / MAST–3I–sp35
JQ782082_MAST–3_1646
543 / MAST–3I–sp3
MAST-3

(G) MAST (continued)

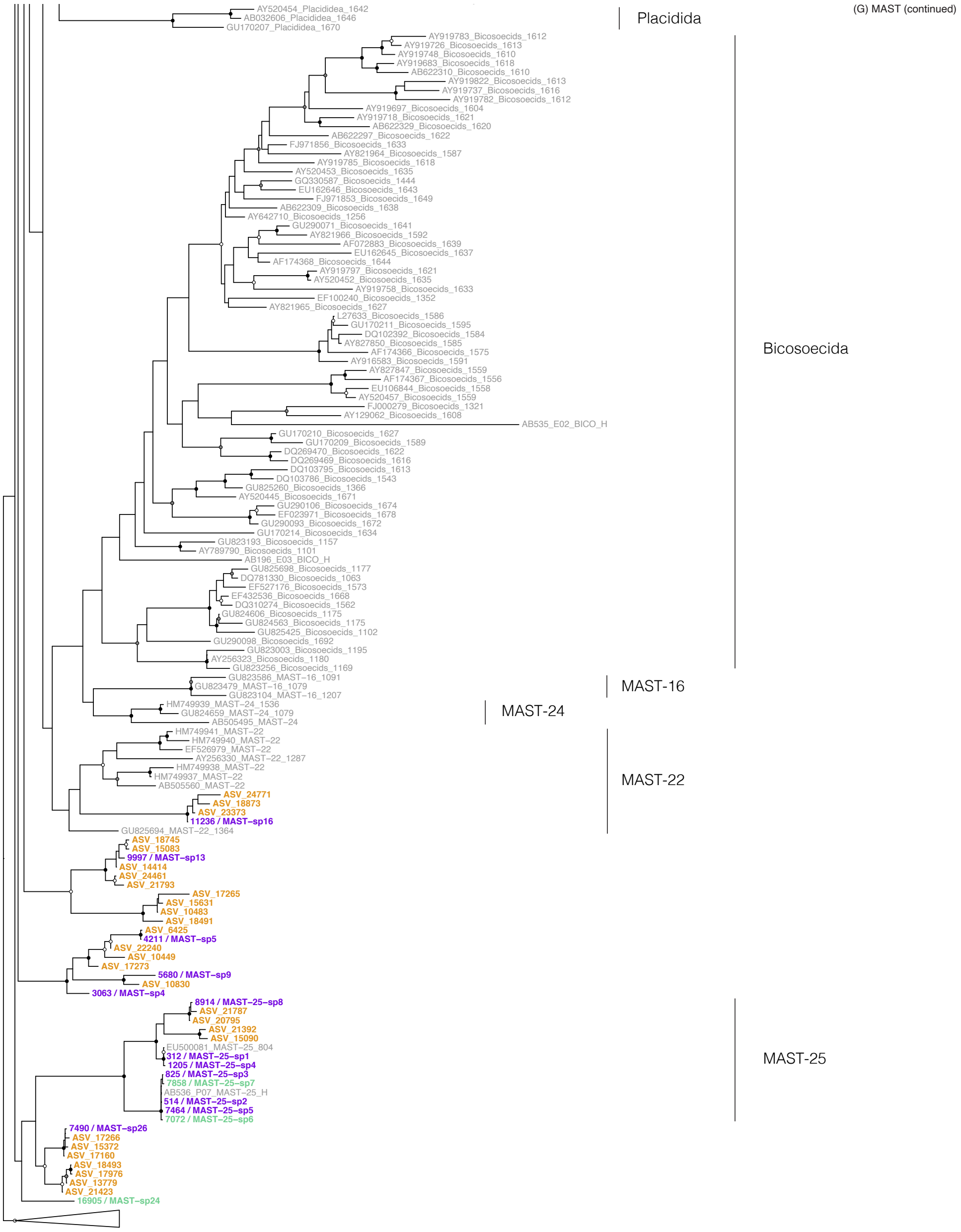


(G) MAST (continued)





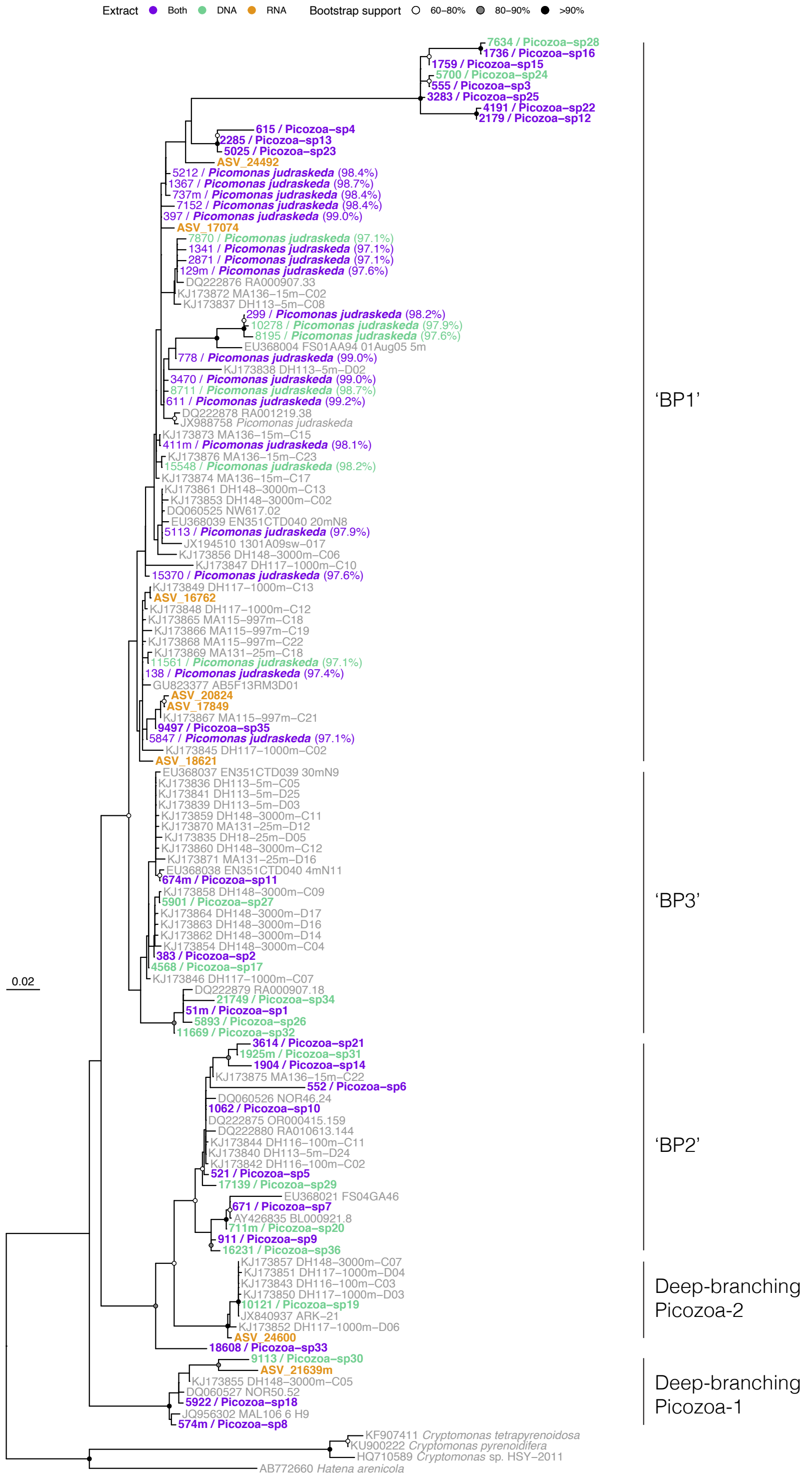
(G) MAST (continued)



(G) MAST (continued)

AY520454_Placididea_1642
AB032606_Placididea_1646
GU170207_Placididea_1670
Placidida
AY919783_Bicosoecids_1612
AY919726_Bicosoecids_1613
AY919748_Bicosoecids_1610
AY919683_Bicosoecids_1618
AB622310_Bicosoecids_1610
AY919822_Bicosoecids_1613
AY919737_Bicosoecids_1616
AY919782_Bicosoecids_1612
AY919697_Bicosoecids_1604
AY919718_Bicosoecids_1621
AB622329_Bicosoecids_1620
AB622297_Bicosoecids_1622
FJ971856_Bicosoecids_1633
AY821964_Bicosoecids_1587
AY919785_Bicosoecids_1618
AY520453_Bicosoecids_1635
GQ330587_Bicosoecids_1444
EU162646_Bicosoecids_1643
FJ971853_Bicosoecids_1649
AB622309_Bicosoecids_1638
AY642710_Bicosoecids_1256
GU290071_Bicosoecids_1641
AY821966_Bicosoecids_1592
AF072883_Bicosoecids_1639
EU162645_Bicosoecids_1637
AF174368_Bicosoecids_1644
AY919797_Bicosoecids_1621
AY520452_Bicosoecids_1635
AY919758_Bicosoecids_1633
EF100240_Bicosoecids_1352
AY821965_Bicosoecids_1627
L27633_Bicosoecids_1586
GU170211_Bicosoecids_1595
DQ102392_Bicosoecids_1584
AY827850_Bicosoecids_1585
AF174366_Bicosoecids_1575
AY916583_Bicosoecids_1591
AY827847_Bicosoecids_1559
AF174367_Bicosoecids_1556
EU106844_Bicosoecids_1558
AY520457_Bicosoecids_1559
FJ000279_Bicosoecids_1321
AY129062_Bicosoecids_1608
AB535_E02_BICO_H
GU170210_Bicosoecids_1627
GU170209_Bicosoecids_1589
DQ269470_Bicosoecids_1622
DQ269469_Bicosoecids_1616
DQ103795_Bicosoecids_1613
DQ103786_Bicosoecids_1543
GU825260_Bicosoecids_1366
AY520445_Bicosoecids_1671
GU290106_Bicosoecids_1674
EF023971_Bicosoecids_1678
GU290093_Bicosoecids_1672
GU170214_Bicosoecids_1634
GU823193_Bicosoecids_1157
AY789790_Bicosoecids_1101
AB196_E03_BICO_H
GU825698_Bicosoecids_1177
DQ781330_Bicosoecids_1063
EF527176_Bicosoecids_1573
EF432536_Bicosoecids_1668
DQ310274_Bicosoecids_1562
GU824606_Bicosoecids_1175
GU824563_Bicosoecids_1175
GU825425_Bicosoecids_1102
GU290098_Bicosoecids_1692
GU823003_Bicosoecids_1195
AY256323_Bicosoecids_1180
GU823256_Bicosoecids_1169
Bicosoecida
GU823586_MAST−16_1091
GU823479_MAST−16_1079
GU823104_MAST−16_1207
MAST-16
HM749939_MAST−24_1536
GU824659_MAST−24_1079
AB505495_MAST−24
MAST-24
HM749941_MAST−22
HM749940_MAST−22
EF526979_MAST−22
AY256330_MAST−22_1287
HM749938_MAST−22
HM749937_MAST−22
AB505560_MAST−22
ASV_24771
ASV_18873
ASV_23373
11236 / MAST−sp16
GU825694_MAST−22_1364
MAST-22
ASV_18745
ASV_15083
9997 / MAST−sp13
ASV_14414
ASV_24461
ASV_21793
ASV_17265
ASV_15631
ASV_10483
ASV_18491
ASV_6425
4211 / MAST−sp5
ASV_22240
ASV_10449
ASV_17273
5680 / MAST−sp9
ASV_10830
3063 / MAST−sp4
8914 / MAST−25−sp8
ASV_21787
ASV_20795
ASV_21392
ASV_15090
EU500081_MAST−25_804
312 / MAST−25−sp1
1205 / MAST−25−sp4
825 / MAST−25−sp3
7858 / MAST−25−sp7
AB536_P07_MAST−25_H
514 / MAST−25−sp2
7464 / MAST−25−sp5
7072 / MAST−25−sp6
MAST-25
7490 / MAST−sp26
ASV_17266
ASV_15372
ASV_17160
ASV_18493
ASV_17976
ASV_13779
ASV_21423
16905 / MAST−sp24

(H) Picozoa


Extract ● Both ● DNA ● RNA
Bootstrap support ○ 60–80% ● 80–90% ● >90%
7634 / Picozoa–sp28
1736 / Picozoa–sp16
1759 / Picozoa–sp15
5700 / Picozoa–sp24
555 / Picozoa–sp3
3283 / Picozoa–sp25
4191 / Picozoa–sp22
2179 / Picozoa–sp12
615 / Picozoa–sp4
2285 / Picozoa–sp13
5025 / Picozoa–sp23
ASV_24492
5212 / Picomonas judraskeda (98.4%)
1367 / Picomonas judraskeda (98.7%)
737m / Picomonas judraskeda (98.4%)
7152 / Picomonas judraskeda (98.4%)
397 / Picomonas judraskeda (99.0%)
ASV_17074
7870 / Picomonas judraskeda (97.1%)
1341 / Picomonas judraskeda (97.1%)
2871 / Picomonas judraskeda (97.1%)
129m / Picomonas judraskeda (97.6%)
DQ222876 RA000907.33
KJ173872 MA136–15m–C02
KJ173837 DH113–5m–C08
299 / Picomonas judraskeda (98.2%)
10278 / Picomonas judraskeda (97.9%)
8195 / Picomonas judraskeda (97.6%)
EU368004 FS01AA94 01Aug05 5m
778 / Picomonas judraskeda (99.0%)
KJ173838 DH113–5m–D02
3470 / Picomonas judraskeda (99.0%)
8711 / Picomonas judraskeda (98.7%)
611 / Picomonas judraskeda (99.2%)
DQ222878 RA001219.38
JX988758 Picomonas judraskeda
KJ173873 MA136–15m–C15
411m / Picomonas judraskeda (98.1%)
KJ173876 MA136–15m–C23
15548 / Picomonas judraskeda (98.2%)
KJ173874 MA136–15m–C17
KJ173861 DH148–3000m–C13
KJ173853 DH148–3000m–C02
DQ060525 NW617.02
EU368039 EN351CTD040 20mN8
5113 / Picomonas judraskeda (97.9%)
JX194510 1301A09sw–017
KJ173856 DH148–3000m–C06
KJ173847 DH117–1000m–C10
15370 / Picomonas judraskeda (97.6%)
KJ173849 DH117–1000m–C13
ASV_16762
KJ173848 DH117–1000m–C12
KJ173865 MA115–997m–C18
KJ173866 MA115–997m–C19
KJ173868 MA115–997m–C22
KJ173869 MA131–25m–C18
11561 / Picomonas judraskeda (97.1%)
138 / Picomonas judraskeda (97.4%)
GU823377 AB5F13RM3D01
ASV_20824
ASV_17849
KJ173867 MA115–997m–C21
9497 / Picozoa–sp35
5847 / Picomonas judraskeda (97.1%)
KJ173845 DH117–1000m–C02
ASV_18621
EU368037 EN351CTD039 30mN9
KJ173836 DH113–5m–C05
KJ173841 DH113–5m–D25
KJ173839 DH113–5m–D03
KJ173859 DH148–3000m–C11
KJ173870 MA131–25m–D12
KJ173835 DH18–25m–D05
KJ173860 DH148–3000m–C12
KJ173871 MA131–25m–D16
EU368038 EN351CTD040 4mN11
674m / Picozoa–sp11
KJ173858 DH148–3000m–C09
5901 / Picozoa–sp27
KJ173864 DH148–3000m–D17
KJ173863 DH148–3000m–D16
KJ173862 DH148–3000m–D14
KJ173854 DH148–3000m–C04
383 / Picozoa–sp2
4568 / Picozoa–sp17
KJ173846 DH117–1000m–C07
0.02
DQ222879 RA000907.18
21749 / Picozoa–sp34
51m / Picozoa–sp1
5893 / Picozoa–sp26
11669 / Picozoa–sp32
3614 / Picozoa–sp21
1925m / Picozoa–sp31
1904 / Picozoa–sp14
KJ173875 MA136–15m–C22
552 / Picozoa–sp6
DQ060526 NOR46.24
1062 / Picozoa–sp10
DQ222875 OR000415.159
DQ222880 RA010613.144
KJ173844 DH116–100m–C11
KJ173840 DH113–5m–D24
KJ173842 DH116–100m–C02
521 / Picozoa–sp5
17139 / Picozoa–sp29
EU368021 FS04GA46
671 / Picozoa–sp7
AY426835 BL000921.8
711m / Picozoa–sp20
911 / Picozoa–sp9
16231 / Picozoa–sp36
KJ173857 DH148–3000m–C07
KJ173851 DH117–1000m–D04
KJ173843 DH116–100m–C03
KJ173850 DH117–1000m–D03
10121 / Picozoa–sp19
JX840937 ARK–21
KJ173852 DH117–1000m–D06
ASV_24600
18608 / Picozoa–sp33
9113 / Picozoa–sp30
ASV_21639m
KJ173855 DH148–3000m–C05
DQ060527 NOR50.52
5922 / Picozoa–sp18
JQ956302 MAL106 6 H9
574m / Picozoa–sp8
KF907411 Cryptomonas tetrapyrenoidosa
KU900222 Cryptomonas pyrenoidifera
HQ710589 Cryptomonas sp. HSY–2011
AB772660 Hatena arenicola
'BP1'
'BP3'
'BP2'
Deep-branching Picozoa-2
Deep-branching Picozoa-1

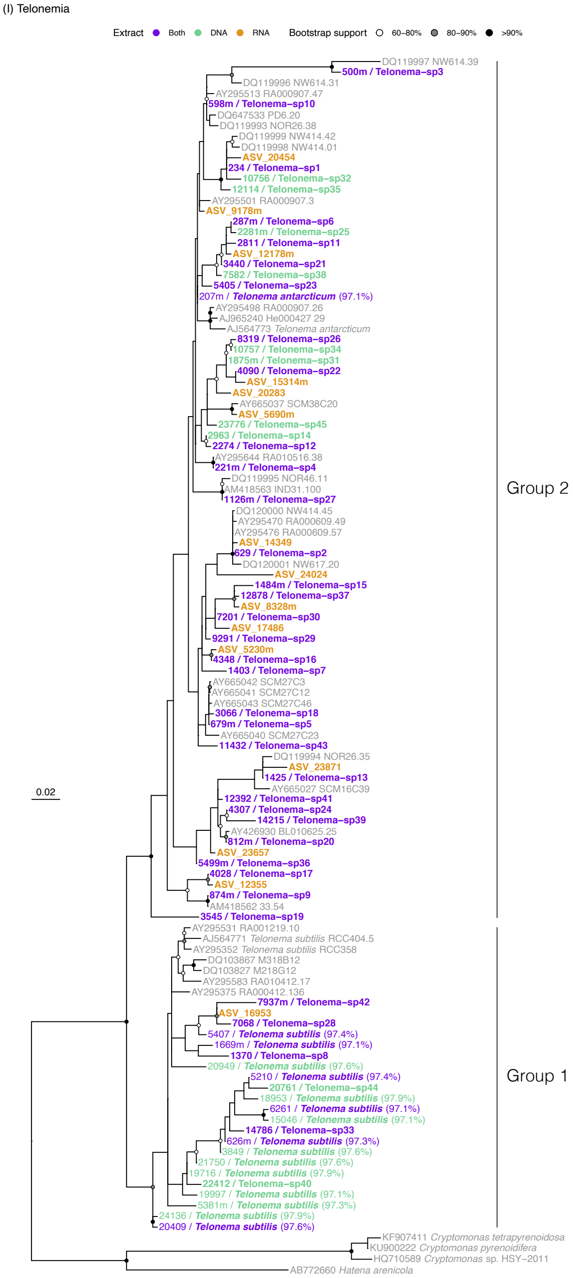


Figure S2. Phylogenetic trees with complete reference 18S rDNA sequences (see Material and Methods) and short ASVs from this study for main HF groups: (**A**) Bicosoecida. (**B**) Centrohelida. (**C**) Cercozoa. (**D**) Choanoflagellata. (**E**) Chrysophyceae. (**F**) Dictyochophyceae. (**G**) MASTs within all Stramenopiles. (**H**) Picozoa. (**I**) Telonemia. Trees were built with RAxML-ng in an all-in-one analysis (maximum likelihood tree search + non-parametric bootstrap) with 100 randomized parsimony starting trees, discrete GAMMA model of rate heterogeneity with 4 categories and 200 bootstrap replicates. Designated clades in each tree come from the original studies from which references were retrieved.

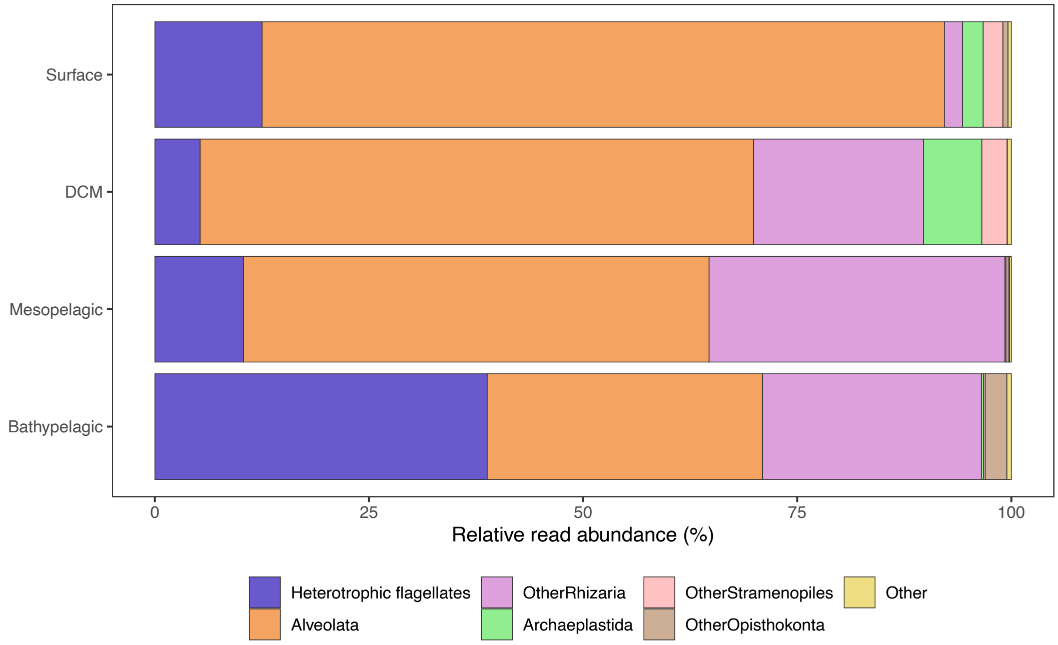


Figure S3. Contribution of broad groups to the overall picoeukaryotic diversity. Some groups do not have HFs (Alveolates, Archaeplastida), while the rest display their signal after excluding the HF groups shown in Table S2 (i.e., OtherRhizaria include all rhizarian reads except those from Cercozoa).

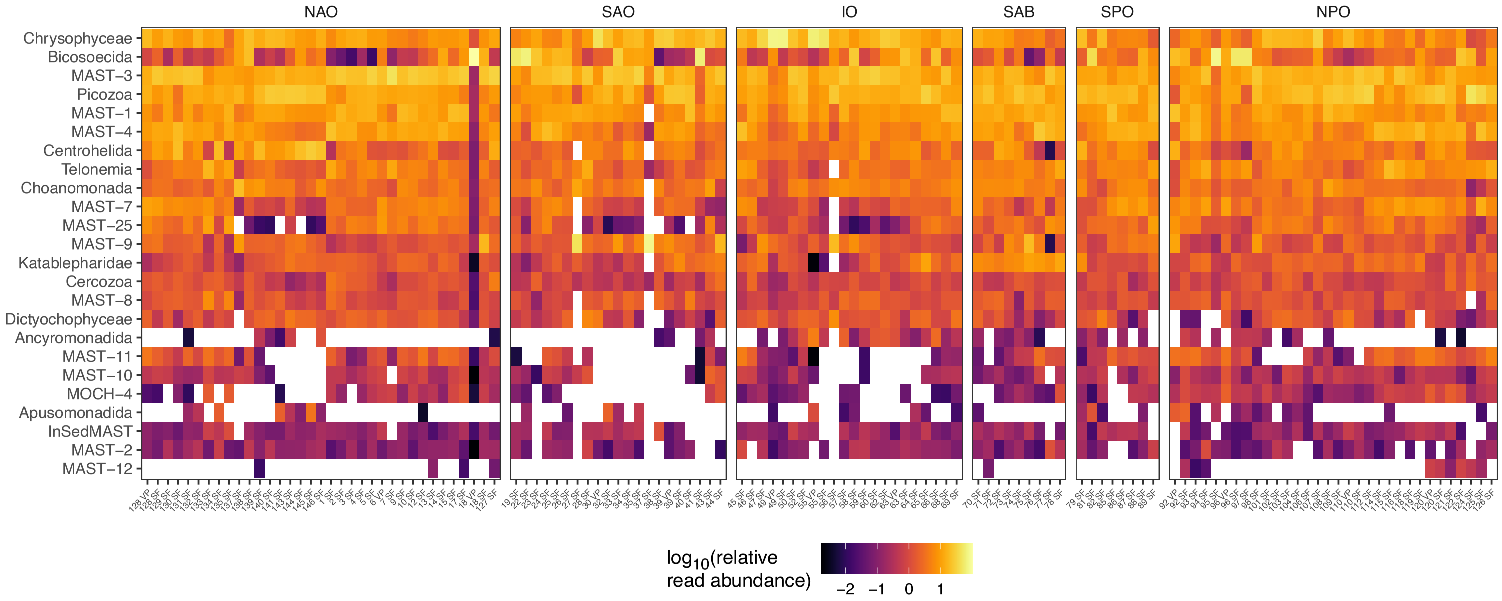


Figure S4. Contribution of HF taxonomic groups in surface samples of the global ocean. Samples are separated by main oceans (NAO: North Atlantic Ocean, SAO: South Atlantic Ocean, IO: Indic Ocean, SAB: South Australia Bight, SPO: South Pacific Ocean, NPO: North Pacific Ocean). Most samples derive from the surface data set ('SF' suffix in the name) and a few from vertical profiles ('VP' suffix). Relative read abundance in log10-scale of each taxonomic group in each sample is used.

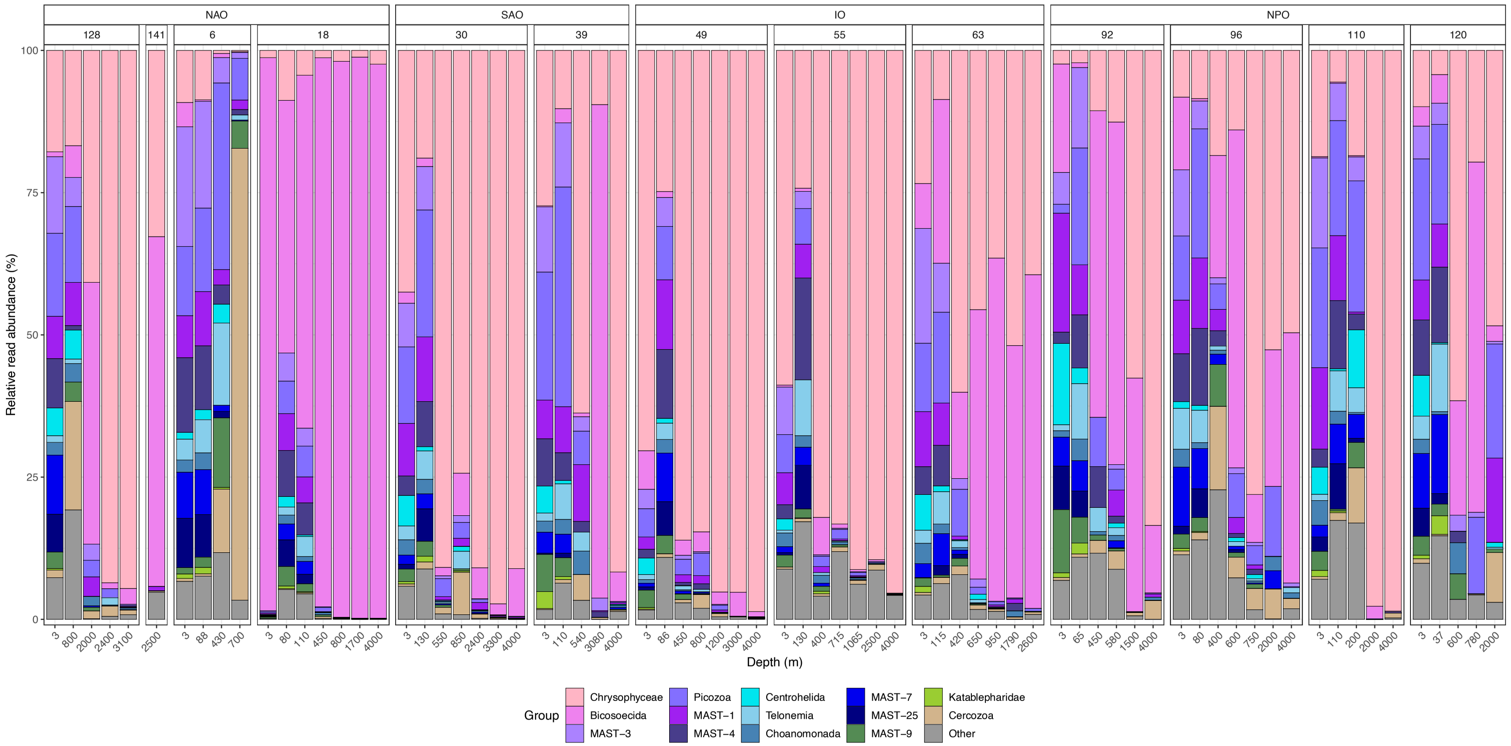


Figure S5. Contribution of HF taxonomic groups in vertical profiles of the global ocean. Samples are grouped by stations, which in turn are separated by subdivisions of main oceans as listed in Figure S4.

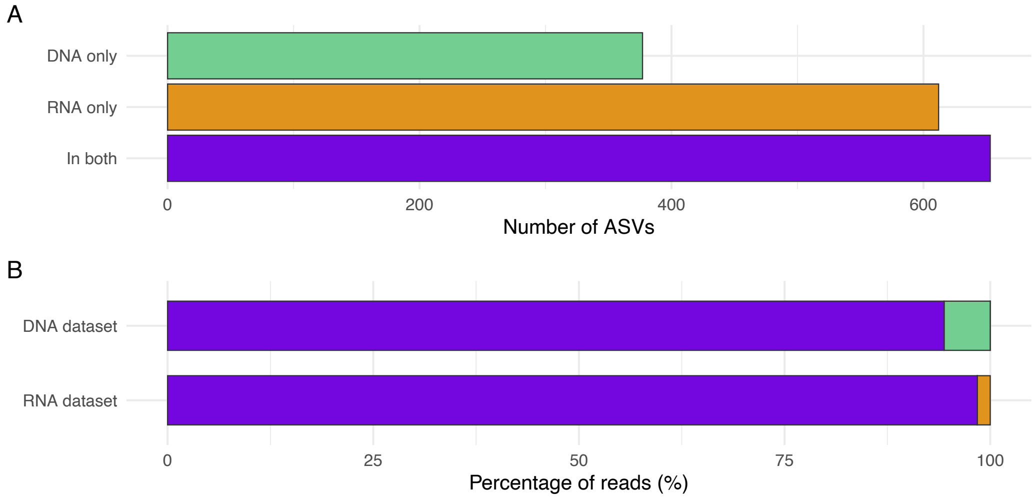


Figure S6. Report of the ASVs found in DNA and RNA data sets. (**A**) Number of ASVs found in a unique data set or shared. (**B**) Percentage of reads within each data set that derive from shared or unique ASVs.

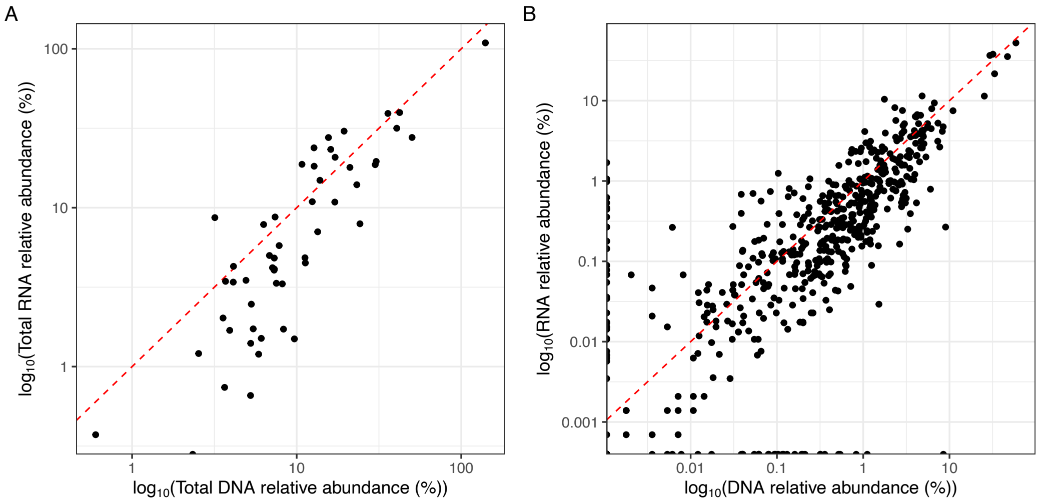


Figure S7. Scatter plots of (**A**) accumulated and (**B**) by-sample relative abundances of the 52 dominant ASVs in the 12 surface assemblages included in both DNA and RNA data sets. Only one of the 52 ASVs was missing from the RNA data set.

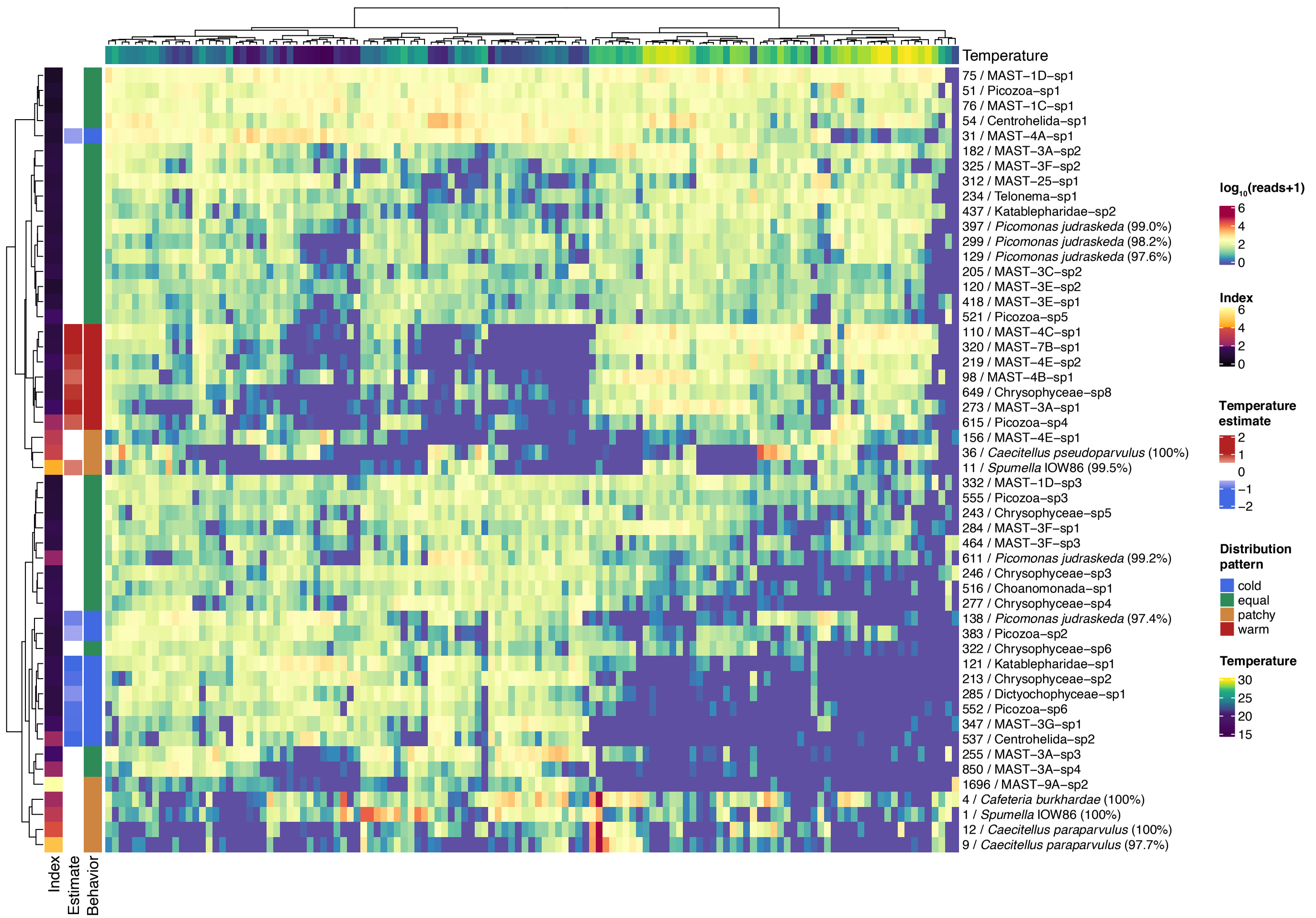


Figure S8. Heatmap displaying the distribution of the dominant 52 ASVs (vertical axis) in 127 surface samples (horizontal axis). We used an ASV table normalized by a geometric mean of pairwise ratios (Chen et al., 2018) and transformed to a log10 scale with a pseudocount of 1. For each ASV, the distribution index (SD/mean), the estimate of the differential abundance test with temperature, and the inferred distribution pattern are shown. Seawater temperature is added for each surface sample.

2.9. Supplementary tables

Table S1. List of molecular sequencing data sets used in this study and their associated information (sampling cruise, size fraction and nucleic acids analyzed, sequencing approach and sampled depth zone).

Dataset	Used in figures	Cruise	Size fraction	Nucleic acid template	Sequencing	Samples					References
						TOTAL	Surface	DCM	Mesopelagic	Bathypelagic	
V4 DNA	1 (all samples) 3 (surface samples) 4 (surface samples) 5 (surface samples)	Malaspina	0.2-3 µm	DNA	V4-18S amplicons	188	127	12	23	26	Logares et al., 2020 Giner et al., 2020
V4 RNA	2	Malaspina	0.2-3 µm	RNA	V4-18S amplicons	91	13	14	32	32	Giner et al., 2020
V9 DNA	2	Malaspina	0.2-3 µm	DNA	V9-18S amplicons	34	6	6	11	11	Obiol et al., 2020
mTags	2	Malaspina	0.2-3 µm	DNA	Metagenomes	65	9	10	26	20	Obiol et al., 2020
TARA V9	2	Tara	0.8-5 µm	DNA	V9-18S amplicons	70	31	39	-	-	de Vargas et al., 2015 Callahan, 2017
Prokaryotes	5	Malaspina	0.2-3 µm	DNA	16S amplicons	113	113	-	-	-	Logares et al., 2020

Table S2. Taxonomic groups including HF taxa considered in this study. Dinoflagellates were not included, as their known minimal size is about 5 µm.

Group	Supergroup
Mantamonas	CRuMS
Rigifilida	CRuMS
Goniomonas	Cryptista
Katablepharidae	Cryptista
Palpitomonas	Cryptista
Diplonemea	Excavata
Kinetoplastida	Excavata
Centrohelida	Haptista
Choanoflagellata	Opisthokonta
Filasterea	Opisthokonta
Ancyromonadida	Other Eukaryota
Apusomonadida	Other Eukaryota
Breviatea	Other Eukaryota
Malawimonadidae	Other Eukaryota
Picozoa	Other Eukaryota
Telonemia	Other Eukaryota
Cercozoa	Rhizaria
Bicosoecida	Stramenopiles
Cantina	Stramenopiles
Chrysophyceae	Stramenopiles
Developayella	Stramenopiles
Dictyochophyceae	Stramenopiles
MAST-1	Stramenopiles
MAST-2	Stramenopiles
MAST-3	Stramenopiles
MAST-4	Stramenopiles
MAST-6	Stramenopiles
MAST-7	Stramenopiles
MAST-8	Stramenopiles
MAST-9	Stramenopiles
MAST-10	Stramenopiles
MAST-11	Stramenopiles
MAST-12	Stramenopiles
MAST-16	Stramenopiles
MAST-20	Stramenopiles
MAST-22	Stramenopiles
MAST-23	Stramenopiles
MAST-24	Stramenopiles
MAST-25	Stramenopiles
InSedMAST	Stramenopiles
MOCH-3	Stramenopiles
MOCH-4	Stramenopiles
Opalinata	Stramenopiles
Picophagea	Stramenopiles
Placidida	Stramenopiles

Table S3. The heterotrophic flagellates catalogue. Number of HF ASVs detected in the the V4 DNA and V4 RNA data sets separated by taxonomic group and by extract (shared, only in DNA, and only in RNA).

Group	Supergroup	Number of ASVs			
		Shared	DNA	RNA	Total
Ancyromonadida	Other Eukaryota	1	1	5	7
Apusomonadida	Other Eukaryota	5	4	3	12
Bicosoecida	Stramenopiles	24	32	25	81
Cantina	Stramenopiles	0	0	2	2
Centrohelida	Haptista	16	38	8	62
Cercozoa	Rhizaria	47	35	35	117
Choanoflagellata	Opisthokonta	64	20	83	167
Chrysophyceae	Stramenopiles	70	38	57	165
Dictyochophyceae	Stramenopiles	13	8	6	27
Katablepharidae	Cryptista	4	6	0	10
MAST-1	Stramenopiles	26	17	15	58
MAST-2	Stramenopiles	5	2	3	10
MAST-3	Stramenopiles	135	73	214	422
MAST-4	Stramenopiles	13	10	1	24
MAST-7	Stramenopiles	29	6	10	45
MAST-8	Stramenopiles	28	12	18	58
MAST-9	Stramenopiles	43	26	30	99
MAST-10	Stramenopiles	1	0	1	2
MAST-11	Stramenopiles	7	2	0	9
MAST-12	Stramenopiles	8	3	7	18
MAST-25	Stramenopiles	6	2	4	12
InSedMAST	Stramenopiles	24	4	55	83
MOCH-4	Stramenopiles	2	0	7	9
Picozoa	Other Eukaryota	40	19	8	67
Telonemia	Other Eukaryota	42	19	15	76
	TOTAL	653	377	612	1642

Table S4. Results of PERMANOVA analyses for surface, vertical profiles and aphotic samples (mesopelagic + bathypelagic layers) using Bray-Curtis dissimilarity matrices and several environmental parameters.

	Surface		**Vertical profiles**		**Aphotic samples**	
	R2	**p-value**	**R2**	**p-value**	**R2**	**p-value**
Water mass	-	-	-	-	0.336	**0.001**
Depth zone	-	-	0.225	**0.001**	0.019	0.306
Temperature	0.159	**0.001**	0.021	**0.017**	0.029	0.106
Ocean subdivision	0.132	**0.001**	0.189	**0.001**	0.152	**0.001**
Conductivity	0.027	**0.001**	0.010	0.351	0.012	0.576
Oxygen (mL/L)	0.013	**0.012**	0.006	0.741	0.008	0.83
Fluorescence	0.012	**0.011**	0.008	0.582	0.052	**0.01**
Salinity	0.016	**0.002**	0.019	**0.031**	0.021	0.282

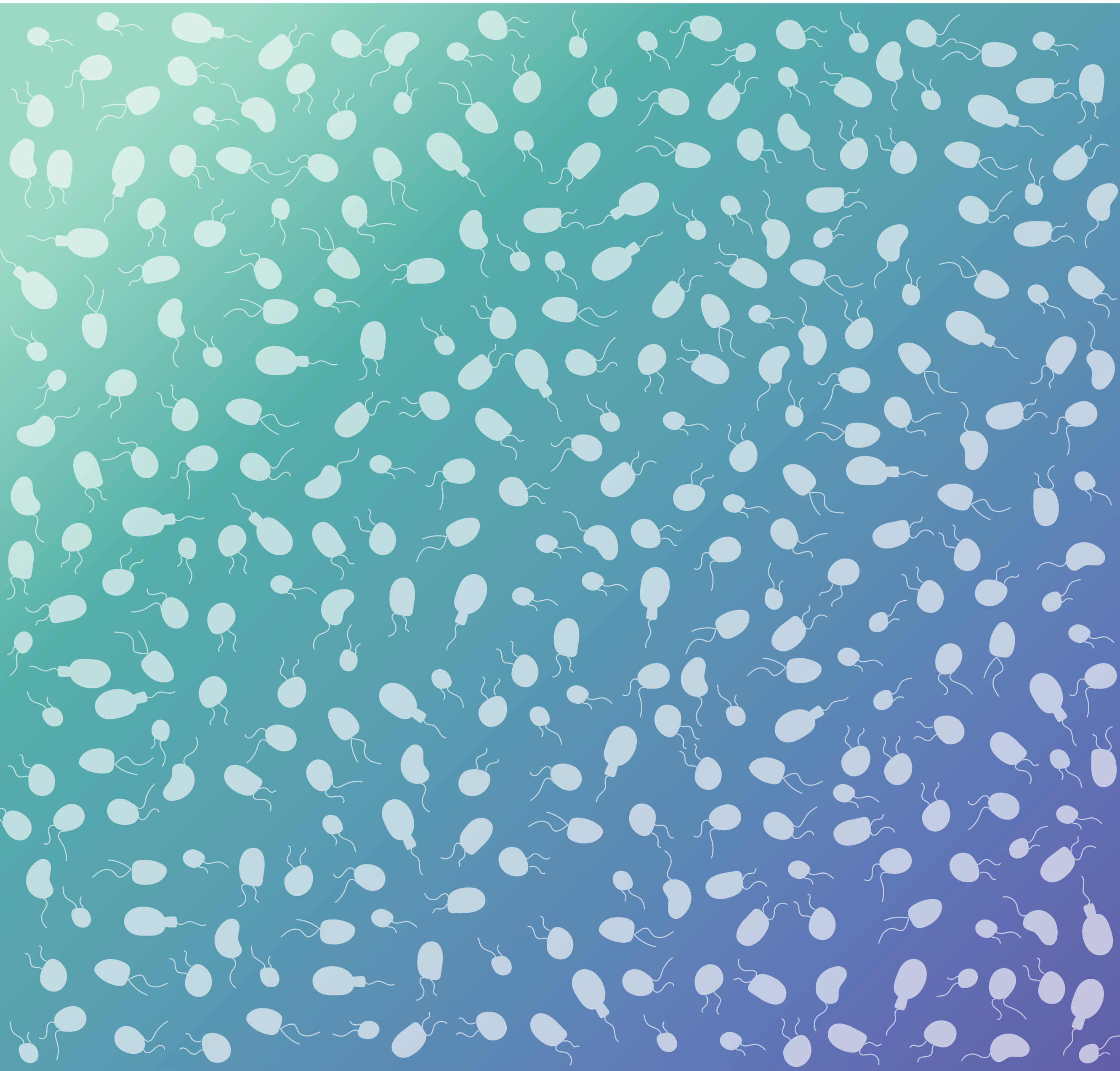
Table S5. Number of dominant HF ASVs in the water column of the global ocean. Dominance is estimated by noting the numbers of ASVs needed to explain 60%, 80%, and 95% of the reads in the four depth zones. Groups are ordered by decreasing overall read contribution in the whole data set.

Group	60% of the reads				80% of the reads				95% of the reads			
	Surf	DCM	Meso	Bathy	Surf	DCM	Meso	Bathy	Surf	DCM	Meso	Bathy
Chrysophyceae	8	2	2	2	17	7	6	3	43	29	15	9
Bicosoecida	4	3	3	1	5	4	4	2	14	5	8	5
MAST-3	11	-	-	-	21	7	-	-	63	34	4	-
Picozoa	11	4	1	-	20	8	3	-	33	21	7	3
MAST-1	3	4	-	-	10	9	1	-	21	13	5	-
MAST-4	5	4	-	-	6	7	1	-	15	8	2	-
Centrohelida	2	-	-	-	3	2	-	-	10	3	4	-
Telonemia	1	3	-	-	9	9	1	-	23	15	3	-
Choanoflagellata	1	-	-	-	5	2	-	-	31	12	4	1
MAST-7	1	3	-	-	9	7	-	-	18	13	2	-
MAST-25	1	2	-	-	4	3	-	-	4	5	1	-
MAST-9	1	-	-	-	6	3	-	-	27	8	6	-
Katablepharidae	2	-	-	-	3	1	-	-	4	3	-	-
Dictyochophyceae	1	-	-	-	1	1	-	-	7	2	-	-
Ancyromonadida	-	1	-	-	1	1	1	-	1	1	1	1
MAST-11	-	2	-	-	3	2	-	-	6	3	-	-
MAST-10	-	1	-	-	1	1	-	-	1	1	1	-
MOCH-4	-	2	-	-	1	2	1	-	1	2	2	-
Cercozoa	-	-	-	-	2	-	4	-	10	3	11	2
MAST-8	-	-	-	-	5	2	-	-	11	9	4	-
InSedMAST	-	-	-	-	-	1	-	-	2	3	5	-
Apusomonadida	-	-	-	-	-	-	-	-	3	-	-	-
MAST-2	-	-	-	-	-	-	-	-	2	-	-	-
MAST-12	-	-	-	-	-	-	-	-	-	1	3	-
TOTAL	52	31	6	3	132	79	22	5	350	194	88	21

Table S6. List of all HF ASVs found in this study together with taxonomic information and nucleotide sequences. For those found in DNA samples (**A**), mean relative abundance and SD, and prevalence are shown for each depth zone, as well as for the overall DNA and RNA data sets. ASVs are ordered by decreasing mean relative abundance. ASVs only found in RNA samples (**B**) are listed with their nucleotide sequence and taxonomic information. Available at: https://aslopubs.onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Flno.11956&file=lno11956-sup-0002-Tables.xlsx

Table S7. Comparison of the dominant 52 ASVs from Malaspina surfaces against TARA Oceans data set, noting the rank abundance position in both data sets. To do the comparison, an unambiguous link between V4 and V9 sequences (based on a long GenBank sequence nearly identical in both regions) was searched. In 7 cases, this unambiguous V4-V9 link could not be done (not comparable search). Available at: https://aslopubs.onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Flno.11956&file=lno11956-sup-0002-Tables.xlsx

CHAPTER 3



CHAPTER 3

Gene expression dynamics of natural assemblages of heterotrophic flagellates during bacterivory

Aleix Obiol, David López-Escardó, Eric D. Salomaki, Monika M. Wiśniewska, Irene Forn, Elisabet Sà, Dolors Vaqué, Martin Kolísko, Ramon Massana

Abstract

Marine heterotrophic flagellates (HF) are dominant bacterial grazers in the ocean, where they represent the trophic link between bacteria and higher trophic levels and participate in the recycling of inorganic nutrients for regenerated primary production. Studying their activity and function in the ecosystem is challenging since most of the abundant HFs in the ocean are still uncultured. In the present work, we investigated gene expression of HF natural communities in four unamended seawater incubations where bacterivory, an understudied process with global implications in biogeochemical cycles, was promoted. Most abundant species growing in the incubations belonged to the taxonomic groups MAST-4, MAST-7, Chrysophyceae and Telonemia. Gene expression dynamics were similar between experiments and could be divided into three states based on microbial counts, each state displaying distinct expression patterns. The analysis of samples where HF growth was highest revealed some highly-expressed genes that could be related to bacterivory. Using available genomic and transcriptomic references, we identified 25 species growing in our experiments, and we used them to compare the expression levels of these specific genes. Our results indicate that cysteine peptidases, together with glycoside hydrolases 3 and 20, are more expressed in heterotrophic than in phototrophic species, and thus could be used to infer the process of bacterivory in natural assemblages.

3.1. Introduction

Understanding the activity and functions of microbial communities in the ocean is fundamental to predict how marine ecosystems will change in the context of global warming (Cavicchioli et al., 2019). Marine microbes, both prokaryotes and microbial eukaryotes (protists), form the base of marine food webs and alterations in their composition and activities could directly impact biogeochemical cycles at a global scale (Hutchins & Fu, 2017). Currently, increases in surface seawater temperature are amplifying the stratification of the water column, thus hampering mixing and the delivery of nutrients from the deep ocean to upper layers (Li et al., 2009). These changes are predicted to promote smaller microorganisms in the ocean (Daufresne et al., 2009; Li et al., 2009), and uncouple bacterial production from grazing mortality (Sarmento et al., 2010). As a consequence, bacterial biomass could increase, keeping inorganic nutrients unavailable for regenerated primary production and producing an imbalance between carbon recycling and carbon export to the deep ocean. Even though bacterivory is central in marine food webs, both the players and the genes they use are still largely uncharacterized.

The use of multi-omics techniques has completely changed the field of microbial ecology, providing new approaches to study microbial diversity and functions. Given the complexity of eukaryotic genomes and the fact that the presence of an eukaryotic gene can inform little about its *in situ* function, metatranscriptomics has been the preferred approach to study the activity of microbial eukaryotes (Caron et al., 2016; Keeling & del Campo, 2017). These studies can benefit substantially from the generation of reference genomes of dominant marine species, and new genomes from uncultured protists have been recently obtained by metagenomics (Alexander et al., 2022; Delmont et al., 2022; Duncan et al., 2022) and single-cell genomics (Gawryluk et al., 2016; Labarre et al., 2021; Latorre et al., 2021; Schön et al., 2021). Recent studies using metatranscriptomics on protist communities have broadened our knowledge on different topics (Cohen et al., 2022) such as trophic strategies (Lambert et al., 2022), diel and seasonal cycles (Kolody et al., 2019; Louyakis et al., 2018), nutrient responses (Alexander et al., 2015; Muratore et al., 2022) or functional biogeography (Carradec et al., 2018). The application of these tools to study bacterivory in the ocean is becoming very promising.

Heterotrophic flagellates (HF) are main bacterial grazers in the ocean (Fenchel, 1986) and at the same time the most understudied component of the marine microbiome (del Campo et al., 2014). These very small unpigmented protists (2-5 µm), found in the photic ocean at concentrations up to 10^3 cells mL^{-1} (Jürgens & Massana, 2008), play a crucial role in the microbial loop by channeling carbon to higher trophic levels, remineralizing inorganic nutrients, and keeping bacterial abundances in balance (Pernthaler, 2005; Sherr & Sherr, 2002). HFs predate on bacteria through the process of phagocytosis, the uptake of the prey through membrane invagination and its digestion in an acidic environment (Botelho & Grinstein, 2011). Phagocytosis is an ancient evolutionarily-conserved process (Boulais et al., 2010) with deep implications in the origin and evolution of eukaryotes (Mills, 2020). It has been deeply investigated in metazoan immunity cells, given its role as a defense mechanism in human immunity (Flannagan et al., 2009), but only few studies have investigated the genes for nutrition-based phagocytosis (Bozzaro et al., 2008; Burns et al., 2015; Dayel & King, 2014; Labarre et al., 2021; Okada et al., 2005).

Therefore, it is yet unclear which functional genes involved in phagocytosis are being expressed by marine HF species during bacterivory, mainly due to the lack of model organisms and the high phylogenetic diversity of HF assemblages. Some recent studies have explored this question by differential gene expression in cultured strains (Massana et al., 2021; Prokopchuk et al., 2022). However, as many of the dominant HF species in the ocean are not available in culture (Obiol et al., 2021), we are still missing a large fraction of the molecular processes involved in bacterivory in natural HF communities. Unamended incubations have proven to be a good approach to promote a pulse of bacterivory from natural HF assemblages (Massana et al., 2006), which facilitates a subsequent study of the gene expression of their uncultured species. We recently used this approach to follow the functional dynamics of a few uncultured MAST species by combining metatanscriptomics and single cell amplified genomes (Labarre et al., 2020). Thus, the combination of unamended incubations and metatranscriptomics could allow identifying highly expressed genes by bacterivorous co-existing species.

In the present work, we performed four unamended dark incubations of surface seawater collected in the Blanes Bay Microbial Observatory (BBMO) at different seasons of the year to explore the functional dynamics of the HF community < 3μm during bacterivorous growth. Our data reveals marked changes in gene expression between the different states of the incubation, together with genes highly expressed during bacterivory that could be related to phagocytosis. Altogether, our study advances our understanding on the biological processes shaping HF communities and throws light on the characterization of bacterivory.

3.2. Materials and methods

3.2.1. Sampling at the BBMO and experimental setup

Surface seawater for unamended incubations (see **Fig. S1** for a schematic overview) was collected and pre-filtered through a 200 μm nylon mesh at the Blanes Bay Microbial Observatory (BBMO), a well-studied coastal sampling station located in the NW Mediterranean Sea (Gasol et al., 2016). In situ temperature was measured using a CTD probe. Carboys containing the seawater were covered with opaque plastic bags to avoid light penetration and transported to the laboratory in less than 2 hours. There, 50 L of seawater were gravity-filtered through 3 µm pore-size polycarbonate filters into a polycarbonate carboy (Nalgene) and incubated in the dark for 5-10 days at *in situ* seawater temperature. The carboy was gently rolled on the floor 10 times every 24 hours to promote water mixing. We sampled 2 L of seawater for RNA sequencing once a day and filtered them through 0.6 µm pore-size (47/142 mm ø) polycarbonate filters using a peristaltic pump (~10 minutes filtration time). Samples were stored at -80ºC after filtration. We also sampled 5 ml of seawater every 12-24 hours and fixed them with glutaraldehyde (1% final concentration) for microbial counts. We stained the fixed samples with 4',6-diamidino-2-phenylindole (DAPI) and filtered through 0.2 µm pore-size (25 mm ø) polycarbonate filters. We counted cell abundances by epifluorescence microscopy of bacteria and heterotrophic flagellates (under UV light), phototrophic flagellates (UV and blue lights) and *Synechococcus* (blue and green lights).

3.2.2. RNA extraction, sequencing and read analyses

We performed RNA extraction and library preparation for Illumina sequencing as detailed in Labarre et al. (2020). Briefly, we cut and vortexed the filters in tubes containing Power Soil beads (Mobio) and extracted RNA using RNeasy Mini Kit (Qiagen) followed by a DNAse treatment with Turbo DNA-free kit (Ambion). We selected a total of 21 samples for sequencing according to their RNA extraction yields. Polyadenylated transcripts were reverse transcribed to cDNA and enriched by 15 polymerase chain reaction (PCR) cycles at CNAG (https://cnag.cat/). RNASeq libraries were prepared with KAPA Stranded mRNA-Seq Illumina (Roche-KAPA Biosystems). Sequencing was carried out using Illumina platforms HiSeq2500 for "Jul17" experiment, HiSeq4000 for "Mar18" and "Nov18" experiments and NovaSeq6000 for "Sep20" experiment. Paired-end reads (2 x 100 bp) were obtained with a final sequencing depth of 15 Gbp, except for "Sep20" experiment, with 25 Gbp.

We trimmed metatranscriptomic Illumina reads for adapters and filtered them for phred scores of ≥ 20 and length ≥ 75 bp with Trimmomatic v0.38 (Bolger et al., 2014) (**Fig. S2**). In order to characterize the taxonomic dynamics of the incubations, we followed the pipeline described in Obiol et al. (2020) to extract and classify 18S rRNA fragments from the obtained reads, using version 5 of eukaryotesV4 database (https://github.com/aleixop/eukaryotesV4)

3.2.3. Assembly, annotation and quantification of transcripts

We first identified and removed ribosomal RNA fragments from quality-filtered reads using SortMeRNA v3.0.3 (Kopylova et al., 2012) with default parameters. We then co-assembled the remaining reads using rnaSPAdes v3.14.1 (Bushmanova et al., 2019) with default parameters and obtained a single assembled metatranscriptome per experiment (**Fig. S2**). For each one, we removed sequences shorter than 200 bp with vsearch v2.17.0 (Rognes et al., 2016) and kept the longest isoform of each gene. We taxonomically classified the transcripts using kaiju v1.8.2 (Menzel et al., 2016) in MEM mode with the nr+euk database and parameters -x -m 11 as in Bucchini et al. (2021) and removed transcripts associated to prokaryotic taxa. We translated the transcripts to proteins using GeneMarkS-T v5.1 (Tang et al., 2015) with minimum length of 200 bp and default parameters and removed transcripts that could not be translated to a protein. We quantified the expression of the preliminary obtained transcripts per experiment in each sample using salmon v1.8.0 (Patro et al., 2017) in mapping-based mode. We considered a gene as expressed if it had ≥2 transcripts per million (TPM) in at least one sample (Mika et al., 2021) and removed the transcripts below this threshold. With these final metatranscriptomes we did a second quantification with salmon and obtained the expression profiles for each sample. We functionally annotated the predicted proteins using eggNOG-mapper v2.1.2 (Cantalapiedra et al., 2021; Huerta-Cepas et al., 2019). We performed most functional analyses using KEGG ortholog (KO) annotations (Kanehisa et al., 2021) and its higher-order associated classifications (BRITE and Pathway). To do so, we generated an expression profile KO table by pooling the TPM values from transcripts associated to each individual KO.

3.2.4. Detection in the experiments of species with known genomic data

We built a reference protein database of 1038 eukaryotic genomes and transcriptomes by combining EukProt database version 3 (Richter et al., 2022) and 50 single amplified genomes prepared from BBMO samples (López-Escardó et al. in prep). In order to detect the presence of these species in our experiments, we aligned the metatranscriptomes to this protein database using DIAMOND blastp v2.0.14 (Buchfink et al., 2021) in sensitive mode. We kept the top scoring hits for each transcript after selecting alignments with >90% identity and a minimum of 50 aminoacids. We removed transcripts having more than one top hit (same e-value) with different reference genomes (1.6% of the cases). With this, we identified 51 species that had a reasonable similarity in the experiments (at least 100 transcripts found per experiment). Then, we verified if the signal detected derived from the same exact species (or a closely related one) by mapping the retrieved transcripts to the coding sequences (CDS) of the reference species at the nucleotide level using blastn v2.7.1 (Altschul et al., 1990). We considered a species present in the experiments when the transcripts and the reference CDS had a median identity >99%. We selected 25 species, annotated their reference genomes/transcriptomes using eggNOG-mapper and quantified their gene expression by mapping the unassembled metatranscriptomic reads to their CDS sequences using salmon. We normalized the obtained read counts using the effective length of each mapped CDS sequence (i.e., we divided mapped read counts by the effective length reported by salmon) and converted them to integer counts ranging from 0 to 10^6 using a pseudo-count as in Salazar et. al (2019). Then, we corrected the obtained expression profiles using edgeR's (Robinson et al., 2009) TMM transformation and converted them to pseudocounts per million.

3.2.5. Statistical analyses

We performed all general analyses using R v4.1.1 (R Core Team, 2021) and packages tidyverse v1.3.1 (Wickham et al., 2019) and vegan v2.5.7. We divided the samples of each experiment into "lag", "growth" and "decline" states according to their placement in the growth curve assessed by microscopy. We validated this classification using plotPCA function in package DESeq2 v1.32.0 (Love et al., 2014) using transcript counts obtained by salmon followed by a variance stabilizing transformation (**Fig. S3**). From the list of 359 highly expressed genes at the growth state in the experiments (**Table S1**), we identified 104 housekeeping genes, later used later for normalizing gene expresion functions, by choosing ribosomal proteins and other genes generally used in the literature (Alexander et al., 2012; Li et al., 2022).

3.3. Results

3.3.1. Growth and taxonomic dynamics

We conducted a total of four unamended incubation experiments between July 2017 and September 2020 following similar experimental procedures (**Table 1**; **Fig. S1**). For each experiment, we gravity-filtered 50 L of surface seawater by 3 µm pore-size, incubated the resulting community (bacteria plus picoeukaryotes) in the dark for 5-10 days at *in situ* seawater temperature, and followed the growth dynamics by epifluorescence microscopy (**Fig. 1A**). In all cases, there was an initial peak of bacteria followed by a peak of heterotrophic flagellates (HF) and a continuous decrease of photosynthetic populations. Initial bacterial abundances were around 7-9 x 10^5 cells ml^{-1} and peaked to ~2 x 10^6 cells ml^{-1} at 4-5 days in the first tree experiments, while in "Sep20" the bacterial peak was earlier and lower. The initial peak of bacteria was followed by a peak of HF, which increased from 5-10 x 10^2 cells ml^{-1} to 7-13 x 10^3 cells ml^{-1}. In the absence of light, the phototrophic flagellates (PF) presented a steady decrease from the initial counts of 1-3 x 10^3 cells ml^{-1}, with the exception of "Mar18" experiment, where initial cell abundances were one order magnitude higher due to a bloom of very small cells (1-2 µm). *Synechococcus* cell abundances steadily decreased from initial abundances about 2 x 10^4 cells ml^{-1} in the first three experiments and twice these initial values (and more marked decreases) in "Sep20". Following the HF dynamics in **Fig. 1A** we classified the incubation into three states ("lag", "growth" and "decline").

We obtained metatranscriptomic reads from extracted RNA collected at different time points of the incubations (white dots in **Fig. 1A**). Prior to assembly, we extracted reads containing the V4 region of the 18S rRNA gene to perform a taxonomic profiling of the eukaryotic communities (**Fig. 1B**). Taxonomic groups with chloroplast-harboring species – Prymnesiophyceae, Dictyochophyceae, MOCH-2, Pelagophyceae, Chlorophyta and Dinoflagellata – showed a clear decreasing trend along the experiments (bottom panels in **Fig. 1B**). These accounted for 43-70% of relative read abundance at the beginning of the incubations and were nearly absent towards the end. Dictyochophyceae in "Nov18" was the only exception to this trend, as its relative abundance increased from 1% at around day 4 to 22% at the final time. Heterotrophic protists (upper panels in **Fig. 1B**) initially represented 17-42% of the relative read abundance and explained most of the read signal during the "growth" state (83-91% total relative read abundance). The most abundant groups considering the 4 experiments were MAST-1 and MAST-7. Chrysophyceae and Telonemia were also very abundant in "Mar18" and "Nov18" experiments (the latter also in "Sep20"). In terms of the overall development of main taxonomic groups, experiments grouped by pairs: "Jul17" and "Sep20" were closer, as well as "Mar18" and "Nov18".

Table 1. Overview of the experiments. General information of the experiments and statistics of the obtained metatranscriptomes.

Experiment	Sampling date	Temperature (°C)	Num. samples	Num. transcripts (10^3)	Size (Mbp)	N50 (bp)	KEGG annotated (%)
Jul17	4/7/17	24	6	225	162	1438	32.9
Mar18	6/3/18	14	5	294	187	1112	31.4
Nov18	5/11/18	19	6	350	228	1216	30.6
Sep20	15/9/20	24	4	410	201	799	22.4

3.3.2. General functional dynamics

We generated four *de novo* metatranscriptomes by coassembling reads and curating the transcripts sets (see **Fig. S2** for a schematic overview of the process) and build expression profile (TPM) tables by mapping these reads back to each metatranscriptome. After removing prokaryotic and low-expression transcripts the final data sets contained 2-4 x 10^5 transcripts with a N50 ranging from 799 to 1438 bp (**Table 1**). Using KEGG database, we could functionally annotate approximately a third of the transcripts (**Table 1**), which represented around half of total TPM in each experiment (**Fig. S4**). We kept these functionally annotated transcripts for further analyses. A simple PCA plot with the normalized counts for these transcripts roughly validated the three states used here (**Fig. S3**).

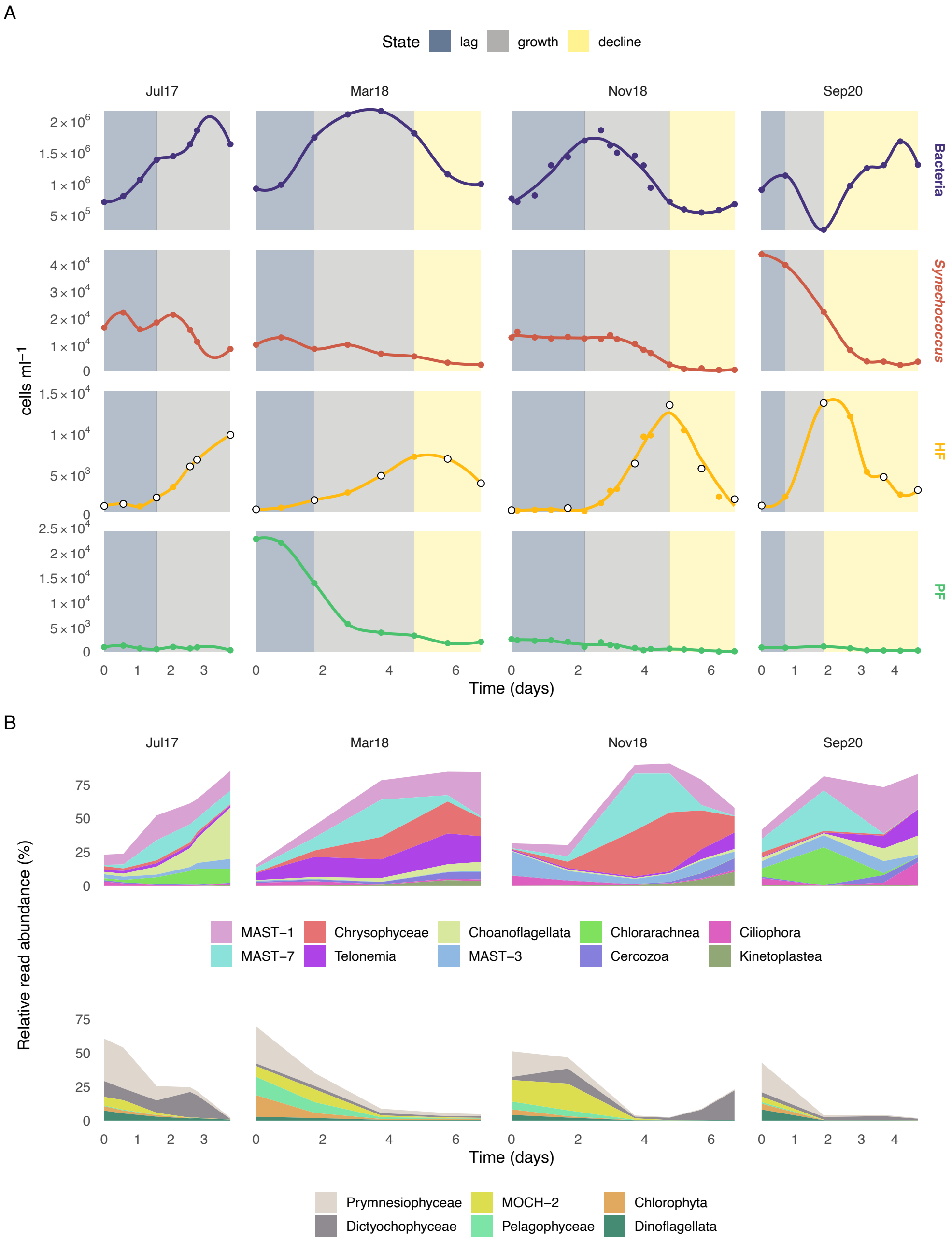


Figure 1. Cell counts and taxonomic dynamics during the incubations. (**A**) Cell counts of bacteria, *Synecococcus*, heterotrophic flagellates (HF) and phototrophic flagellates (PF) conducted by epifluorescence microscopy with DAPI staining. The background of the plots is colored by the different incubation states of the HF community. White dots in HF curves represent time points from which we obtained metatranscriptomic data. (**B**) Relative read abundance of the main taxonomic eukaryotic groups during the incubations as seen by 18S-V4 mTags. Groups are divided into 2 plots by their overall dynamics: increasing (upper panels) or decreasing (bottom panels) their relative read abundance.

The most expressed KEGG orthologs (KOs) considering all experiments (**Fig. 2A**) were associated to proteins involved in cytoskeleton structure, such as actin, tubulin and centrin. These showed relatively similar levels of expression along incubation states except for centrins, which displayed a higher expression during “decline” state. Ribosomal proteins and elongation factors, involved in protein synthesis, were also highly expressed, with rather constant expression levels along time. Other highly-expressed KOs were calmodulin and ubiquitin, proteins related to signal transduction, which followed the same pattern; and cathepsins L and X, cysteine peptidases that were highly expressed in the “growth” state. Among the highly expressed KOs there were also 2 photosynthesis-related proteins (chlorophyll a/b binding proteins) that exhibited high TPM numbers in “lag” state and a dramatic decrease at the “growth” and “decline” states. The experiments exhibited similar overall dynamics when comparing their gene level KOs TPMs in a non-metric multidimensional scaling using Bray-Curtis dissimilarities (**Fig. 2B**). Thus, samples seemed to be organized by time of the incubation rather than by experiment. In terms of states of the incubation, “lag” was clearly differentiated from the rest, while “growth” and “decline” states displayed a clear overlap.

We then looked at the higher-level annotation of each KO represented by KEGG BRITE categories to report the expression dynamics of the main functions and proteins along the states and the experiments (**Fig. 2C**). The structuring between states seen in **Fig. 2B** working at the gene level was also apparent when using these broader categories (**Fig. 2C**), as “growth” and “decline” samples clustered together and “lag” samples formed a separate cluster. The decrease along incubation time in the expression of photosynthesis-related proteins was very apparent (**Fig. 2C**), with maximum TPM values in “lag” samples and virtually no expression in “decline” samples. Apart from this clear pattern, the dynamics of the remaining categories were less marked and could be roughly divided into 3 different trends: (1) functions with genes exhibiting high expression through all incubation states, (2) those showing higher expression in both “growth” and “decline” states and (3) those more expressed in the “growth” state. The first group included constitutive processes of the cell, such as spliceosome, translation factors or chaperones. Also, membrane trafficking processes and transporters had a comparable expression across incubation states. In the second category, ribosome was the most expressed category, followed by ubiquitin, GTP-binding and transcription- and replication-related proteins. The third group contained functions and proteins with highest expression in the “growth” state, such as peptidases, GPI-anchored proteins and glycosyltransferases.

3.3.3. Most expressed genes during “growth” state

Based on the dynamics of bacteria and HF in all incubations, it could be deduced that the “growth” state was the period when the bacterivory process had the highest relative importance. So, we focused on the most expressed genes (KOs) during that state. We obtained

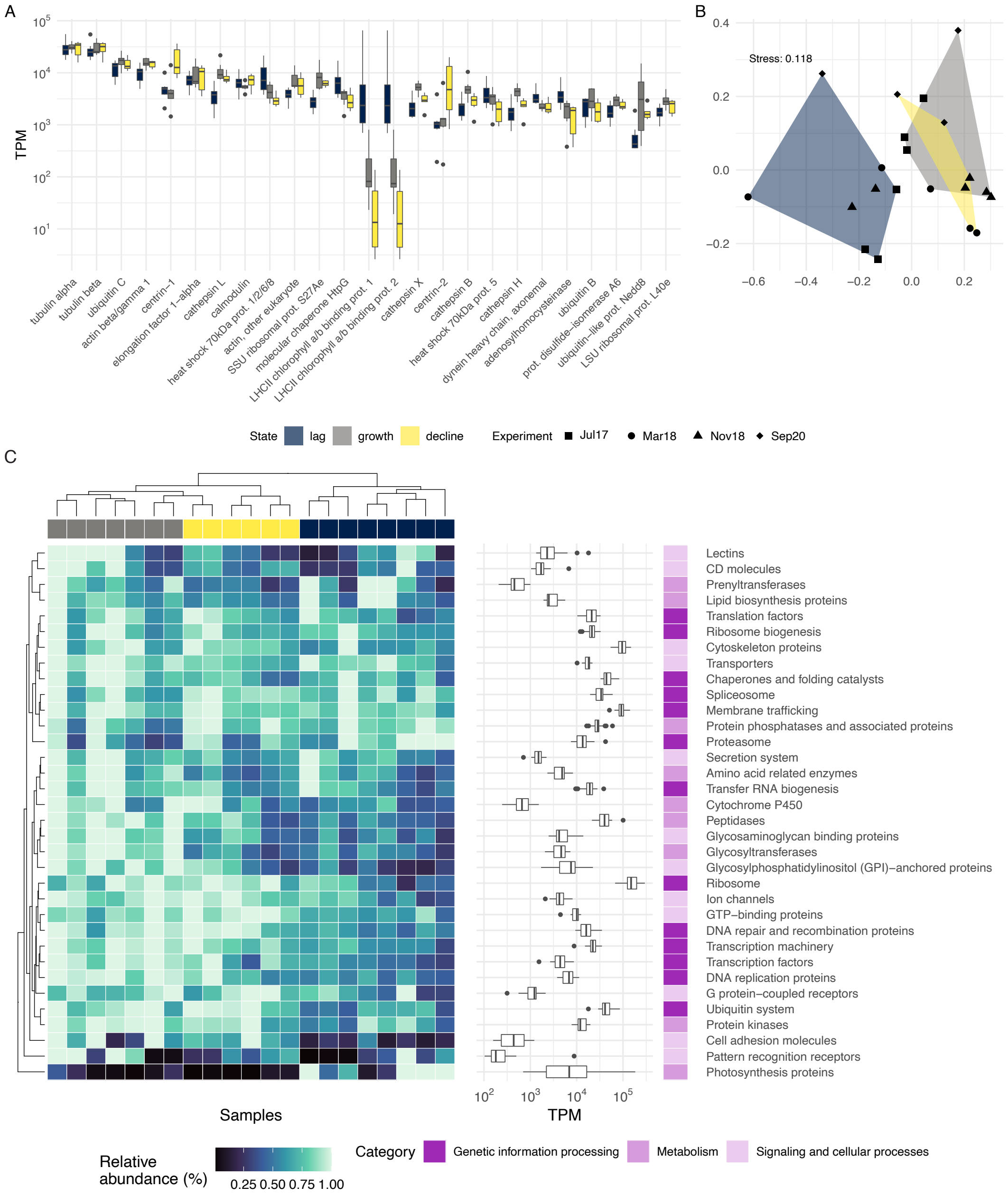


Figure 2. General functional dynamics of the 4 experiments. (**A**) Expression values per sample of the 25 most expressed KOs (KEGG orthologs) in the incubations represented by boxplots colored by incubation state. (**B**) Non-metric multidimensional scaling (NMDS) plot using Bray-Curtis dissimilarities between the expression of KOs in the different samples. Samples are grouped by the incubation state they belong to. (**C**) Heatmap displaying the expression of main categories as represented by KEGG BRITE classifications. Values shown are computed by scaling TPM values to a 0-1 scale per category and experiment (i.e., TPM values belonging to a category and experiment are divided by their maximum value). Boxplots display the actual TPM values per sample of each category.

a list of 359 highly-expressed KOs (those with at least >500 TPM on average in one of the experiments), which we then grouped into a custom system of 24 categories partially based on KEGG BRITE and Pathway hierarchies, assigning each KO to a single category (**Table S1**) to avoid having duplicated signal. These general functions presented comparable gene expression levels when putting together all "growth" samples from the four incubations (**Fig. 3A**). Some of the most expressed categories were related to processes generally considered as constitutive or housekeeping, as already seen in general dynamics (**Fig. 2**): ribosomal and cytoskeleton proteins, protein processing, translation, replication or transcription. However, other categories related to metabolism also emerged, such as peptidases, CAZy enzymes or other hydrolases, as well as translocases (mainly including proton pumps), and membrane trafficking proteins (**Fig. 3A**).

Following the data revealed in **Fig. 3A** and past evidence as their role in bacterivory through phagocytosis (see Discussion for further details), we analyzed in detail the KOs of three of the previous functional categories, namely peptidases, translocases (proton pumps) and CAZy enzymes (**Fig. 3B**). We first assessed the overlap in the functional annotation of transcripts so as to avoid having different KOs with the same transcript associated. This was particularly critical within peptidases, in which the 20 KOs were grouped in 6 broader categories (**Fig. S5**), and also occurred once with CAZy enzymes. We then computed the average FC (Fold Change) of the 104 housekeeping genes (HK; **Table S1**) between "lag" and "growth" states per experiment after removing outliers (as calculated in boxplots) and used this value as baseline to compare the expression FC of the rest of the genes (calculated again by dividing their averaged expression values in "lag" and "growth" states). All peptidase genes had a higher expression in the "growth" state than HK genes in at least 3 of the 4 experiments (**Fig. 3B**), suggesting a putative upregulation during growth by bacterivory, with an average FC >3 (i.e., their expression increased more than 3-fold on average from "lag" to "growth" states). Among these, cysteine peptidases (including 7 cathepsin types, see **Fig. S5**) were by far the most expressed in log phase (**Fig. 3B**). For translocases, genes related to V-type ATPase displayed different patterns, with two subunits (a and 16kDa proteolipid) being more expressed than HK genes in 3 experiments and subunit B in only one. We also detected this latter trend with inorganic pyrophosphatase, which was one of the most expressed genes related to proton pumps in the incubations (**Fig. 3B**). For CAZy enzymes, two glycoside hydrolases (GH; chitinase and hexosaminidase) and two glycosyltransferases (GT; dolichyl–diphosphooligosaccharide protein glycosyltransferase and reversibly glycosylated polypeptide UDP–arabinopyranose mutase) were more expressed than HK genes in all the experiments. Carboxylesterase 2 and GH7 genes showed the opposite trend, with a lower FC than HK in all cases. GT66, AA13 and CE10 were the most expressed genes in this category (**Fig. 3B**).

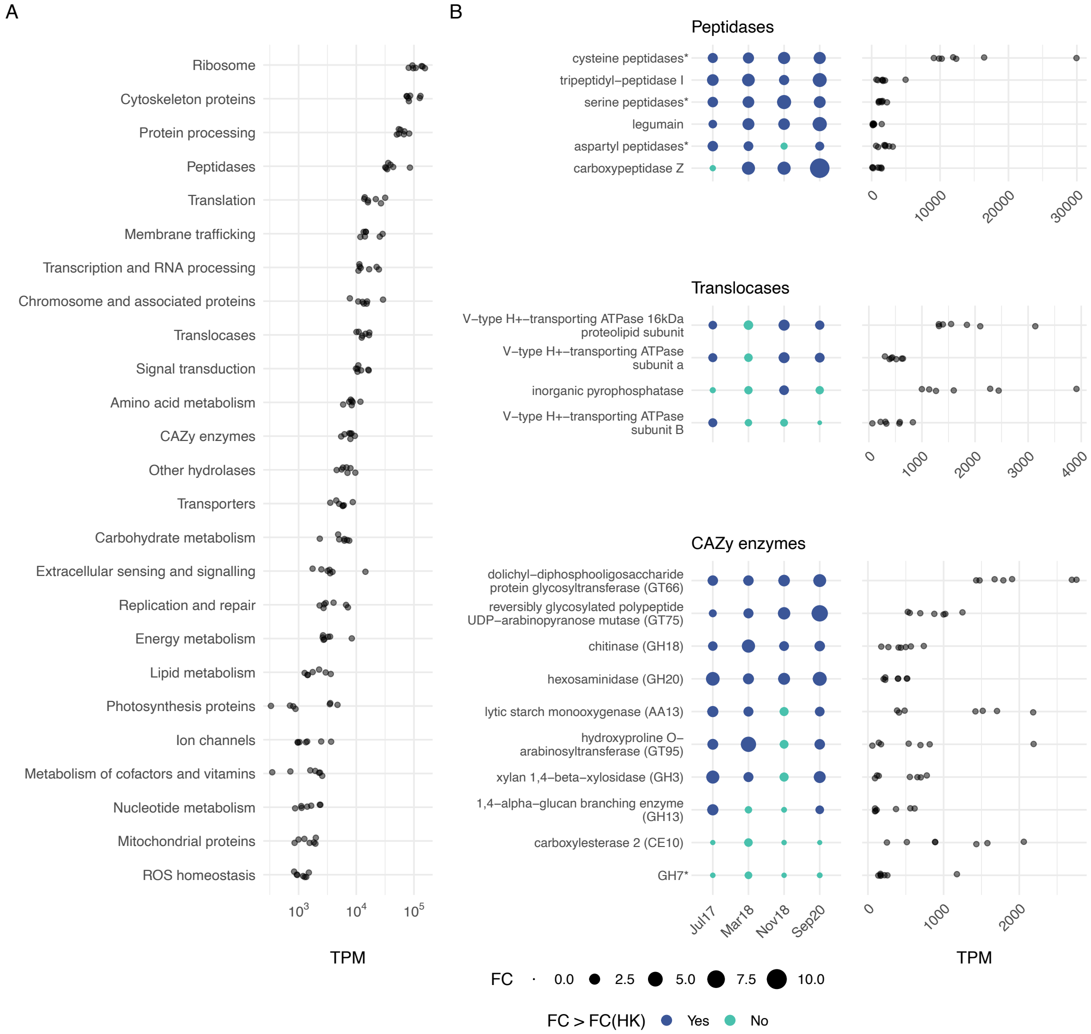


Figure 3. Most expressed functions and genes in growth state. (**A**) Expression values (TPM) of the 359 most expressed genes in "growth" samples pooled into custom categories (see **Table S1** for further details). We created these categories partially based on KEGG BRITE and Pathway classifications to avoid having KOs associated to more than one category (as happens with KEGG BRITE). (**B**) Fold change (FC) between "lag" and "growth" incubation states of genes (KOs) annotated as peptidases, translocases (proton pumps) and CAZy enzymes from the 359 highly-expressed genes in "growth" state and their expression values. Dots representing fold change are colored differently according to their comparison to the average FC of housekeeping genes. In some cases (marked with an asterisk), KOs displaying overlap in functional annotations (i.e., different KOs associated to the same transcript) needed to be grouped into broader sets (see **Fig. S5** for further details). 'Cysteine peptidases' includes cathepsins B, F, H, K, L, O, X, as well as KDEL-tailed endopeptidase and xylem cysteine peptidase; 'aspartyl peptidases' includes cathepsins D and E, phytepsin and saccharopepsin; 'serine peptidases' includes cathepsin A, serine carboxypeptidase-like clades I and II and vitellogenic carboxypeptidase-like protein. For CAZy enzymes, 'GH7' groups cellulose 1,4-beta-cellobiosidase and cellulase.

3.3.4. Gene expression at the species level

After looking at the expression dynamics at the community level, we investigated which species with genomic/transcriptomic data available were present in the incubations, which would allow a detailed analysis of their gene expression in our metatranscriptomes. By mapping the transcripts to an exhaustive custom protein database of eukaryotic species (see Methods for details), we obtained a preliminary list of 51 candidate taxa (**Fig. S6**). The comparison between the species coding sequences and their associated transcripts revealed cases of virtually identical sequences (green bars in **Fig. S6**) but also cases of transcripts having median identity ranging from 90 to 95% (red bars in **Fig. S6**) that could not be considered to derive from that species but from a highly related one. After filtering these cases, we obtained a final list of 25 species present in our incubations (**Table 2**). In terms of taxonomic diversity, the list contained 12 Stramenopiles (several MASTs, Ochrophyta and Bicosoecida), 6 Archaeplastida (Chlorophyta and Picozoa), 5 Haptista (Prymnesiophyceae), 1 cercozoan and 1 choanoflagellate (**Table 2**). The re-quantification of the metatranscriptomic reads against the complete species proteome provided a picture of the dynamics of the 25 species in the different incubations (**Fig. 4**, see **Fig. S7** for the full display with all the species). In general, heterotrophic species tended to increase their expression towards the middle and end of the incubation, while phototrophic species were highly expressed at the beginning and decreased along incubation time. In the case of species labelled as mixotrophic, a mix of the above-mentioned trends was seen, as well as some species displaying a rather steady expression (**Fig. S7**).

Taking advantage of the different trophic modes represented by the 25 species (12 heterotrophs, 8 mixotrophs and 5 phototrophs), we analyzed whether the expression of the genes identified as putatively relevant for bacterivory varied between nutritional strategies. Thus, if these genes participated actively in bacterivory, we would expect them to be more expressed in phagotrophs than in phototrophs. Within peptidases, cysteine peptidases were the ones with the highest relative expression values, with approximately an order of magnitude higher than the rest (**Fig. 5**). Relative gene expression also differed depending on trophic mode, with heterotrophic species always displaying slightly higher values (around 2% of gene expression) than mixotrophic species and markedly higher than phototrophic ones (<0.5%). Despite exhibiting lower expression levels, the remaining peptidase genes followed similar trends among trophic modes. In translocases, inorganic pyrophosphatase had the highest levels of relative expression and was similarly expressed in both heterotrophs and mixotrophs and a bit lower in phototrophs, a pattern slightly visible in the V-ATPase subunits (**Fig. 5**). CAZy enzymes showed different trends. GH13 seemed to be more expressed in phototrophs, CE10 in mixotrophs and GT66, GH20, AA13, GT95 and GH3 in heterotrophs. As revealed by functional annotation results, genes encoding enzymes CE10 and GH7 were not present in genomes/

transcriptomes from the selected phototrophic species, and AA13 had a single sequence annotated in *Pycnococcus provasolii*. Therefore, these did not display any relative expression value for phototrophs (**Fig. 5**).

Table 2. Species with genomic data well represented in the metatranscriptomes. The 25 species with strong signal in the experiments are displayed with general taxonomy, genome completeness (BUSCO), and quantification information in the incubations.

Species	Supergroup	Group	Source	Trophic mode	BUSCO (%)	Experiments present	Max reads recovered per sample (%)
Bathycoccus prasinos	Archaeplastida	Mamiellophyceae	genome	Phototroph	75.3	2	2.74
Micromonas-sp1	Archaeplastida	Mamiellophyceae	single-cell genome	Phototroph	27.1	2	0.69
Ostreococcus lucimarinus	Archaeplastida	Mamiellophyceae	genome	Phototroph	78.1	1	0.62
Picozoa sp. COSAG01	Archaeplastida	Picozoa	single-cell genome	Heterotroph	21.6	1	0.02
Picozoa sp. COSAG02	Archaeplastida	Picozoa	single-cell genome	Heterotroph	32.1	1	0.03
Pycnococcus provasolii	Archaeplastida	Pycnococcaceae	transcriptome	Phototroph	55.3	1	0.01
Chrysochromulina rotalis	Haptista	Prymnesiophyceae	transcriptome	Mixotroph	53.7	1	0.11
Dicrateria rotunda	Haptista	Prymnesiophyceae	transcriptome	Mixotroph	36.1	1	0.28
Emiliania huxleyi	Haptista	Prymnesiophyceae	genome	Mixotroph	56.1	3	0.10
Isochrysidales sp. CCMP1244	Haptista	Prymnesiophyceae	transcriptome	Mixotroph	57.2	3	0.07
Phaeocystis cordata	Haptista	Prymnesiophyceae	transcriptome	Mixotroph	53.4	3	1.00
Acanthoecidae sp. 10tr	Opisthokonta	Choanoflagellata	transcriptome	Heterotroph	78.8	3	0.51
Mataza sp. D1	Rhizaria	Cercozoa	transcriptome	Heterotroph	77.7	2	0.10
Cafeteria burkhardae	Stramenopiles	Bicosoecida	genome	Heterotroph	67.1	1	0.10
Triparma eleuthera	Stramenopiles	Bolidophyceae	transcriptome	Mixotroph	53	3	0.19
Triparma laevis	Stramenopiles	Bolidophyceae	transcriptome	Mixotroph	40.8	2	0.23
ChrysophyceaeNA-sp1	Stramenopiles	Chrysophyceae	single-cell genome	Heterotroph	21.2	1	0.13
Leptocylindrus hargravesii	Stramenopiles	Diatomeae	transcriptome	Phototroph	65.1	1	0.05
Rhizochromulina sp. CCMP1243	Stramenopiles	Dictyochophyceae	transcriptome	Mixotroph	67.9	2	1.90
MAST-1C-sp1	Stramenopiles	MAST-1	single-cell genome	Heterotroph	3.9	2	0.08
MAST-1D-sp2	Stramenopiles	MAST-1	single-cell genome	Heterotroph	11.4	1	0.02
MAST-3C-sp2	Stramenopiles	MAST-3	single-cell genome	Heterotroph	31.4	1	0.04
MAST-4A-sp1	Stramenopiles	MAST-4	single-cell genome	Heterotroph	73.8	4	0.28
MAST-4E-sp1	Stramenopiles	MAST-4	single-cell genome	Heterotroph	57.3	2	1.56
MAST-8B-sp1	Stramenopiles	MAST-8	single-cell genome	Heterotroph	18.4	2	0.05

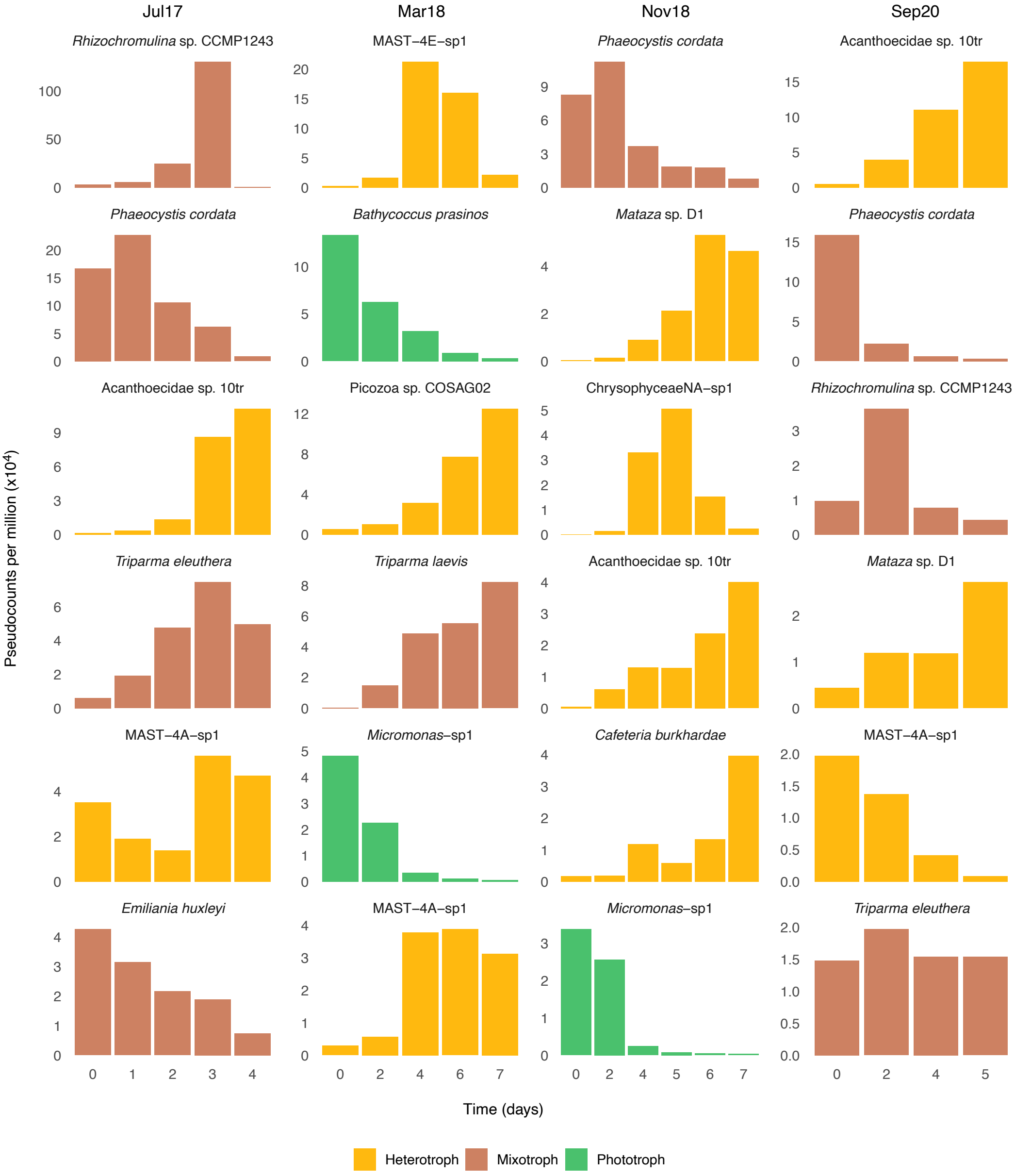


Figure 4. Expression dynamics of some species with genomic data found in the metatranscriptomes. See the full list of the 25 detected species in **Table 2** and the display of their expression dynamics in **Fig. S6**. Values represent pseudocounts per million, obtained after correcting the abundance profiles by gene length and sequencing depth (see Material and Methods for details).

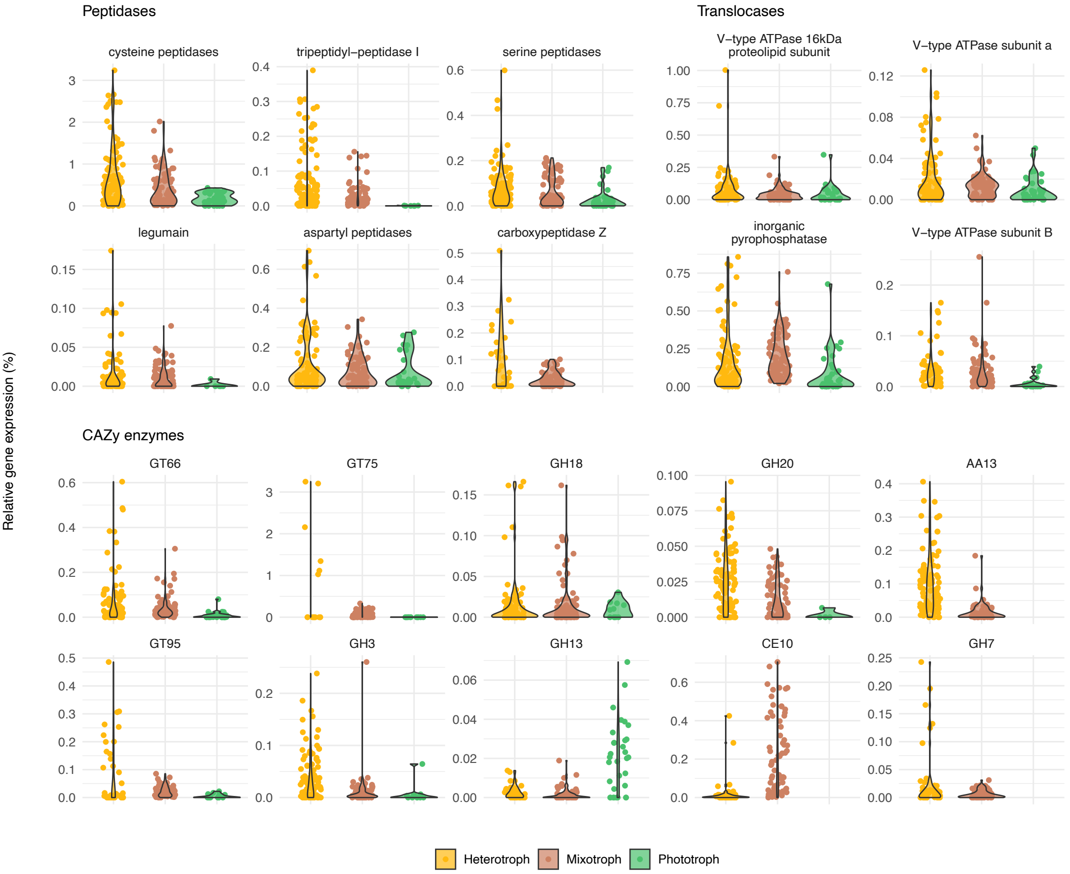


Figure 5. Expression of selected genes in species with different trophic modes. Points represent the relative expression of the gene in a single species and sample. Values were computed by dividing the expression of the selected gene by the total expression for each species and sample. Values are separated by the trophic mode of the species they come from (**Table 2**).

3.4. Discussion

Marine heterotrophic flagellates (HF) remain largely undersampled, and very few ecologically relevant species are available in culture (del Campo et al., 2014). As a consequence, key processes in global biogeochemical cycles, such as bacterivory, still need to be well characterized. In this study, we performed 4 unamended incubations of coastal seawater in the dark to promote the growth of natural HF assemblages and assess their gene expression during bacterivory.

3.4.1. Circumventing the lack of cultured representative HF species

The use of unamended incubations allowed a more than 10-fold increase in cell densities of natural HF assemblages. With this, we could obtain good quality metatranscriptomic samples from ecologically relevant microorganisms. Even after performing a polyA selection towards eukaryotic messenger RNA, the metatranscriptomes contained a remarkable signal of ribosomal RNA (around 3% of reads on average matched the V4 region of the 18S rDNA) due to its huge abundance in the cell (Cui et al., 2010). We took advantage of this to assess the general taxonomic dynamics in the incubations, an approach supported by a previous study where we reported that the relative abundance of this rRNA signal was well correlated with the FISH counts of the target cells (Labarre et al., 2020). The most represented taxa belonged to MAST clades, Chrysophyceae and Telonemia, groups that have been identified to be highly abundant and widespread in the surface ocean (Obiol et al., 2021). Despite the general dominance of these taxonomic groups, differences emerged between the individual species growing in the experiments, thus highlighting the seasonality of protist communities in BBMO (Giner et al., 2019) and the large diversity of natural assemblages.

Although most of the groups present in the incubations are known to be poorly represented in public databases, we tried to map the unassembled metatranscriptomic reads to an exhaustive database of microbial eukaryotes to see if we could taxonomically bin them, as performed in other studies (Alexander et al., 2015; Metegnier et al., 2020). As expected, however, less than 15% of reads on average mapped to the reference data set with an identity >90% (**Fig. S8**). Thus, we only kept this approach to answer specific questions on some bacterivory-related genes (see last section) and opted to perform a *de novo* assembly for each of the experiments to analyze the expression patterns of the whole community. With this method, we could functionally annotate half of our transcripts using different databases, while the other half remained completely unknown. This issue was also reported in a global eukaryotic metatranscriptomic survey (Carradec et al., 2018), and agrees with the estimation that currently 40-60% of microbial predicted genes cannot be assigned to a known function (Vanni et al., 2022).

3.4.2. Functional insights into HF gene expression

The general functional dynamics in the 4 experiments showed seemingly similar patterns of gene expression. This is relevant, as the incubations were conducted during different times of the year, and due to species succession following seasonality trends (Giner et al., 2019; Lambert et al., 2018) initial assemblages were certainly different. Therefore, the similar gene expression patterns could reflect some kind of functional redundancy, as different sets of co-existing taxa could be performing the same function in the ecosystem (Louca et al., 2018).

For a better characterization of the species dynamics in each incubation, a more thorough analysis with amplicon sequencing could have been done, but this was outside the scope of this paper.

The most visible pattern in functional dynamics was the expression decrease of photosynthesis-related genes to virtually zero, in agreement with the decrease of phototrophic flagellates detected by microscopy and the fact that incubations were performed in the dark. Despite representing an obvious outcome of the experimental setup, this transition from phototrophy to heterotrophy in all the incubations highlights the reliability of our data to analyze the expression of bacterivory-related genes. Many of the highly-expressed transcripts were related to constitutive processes of eukaryotic cells, such as tubulin, actin, ubiquitin or ribosomal proteins. Given their house-keeping role, these had high expression values regardless of the state of the incubations.

The experimental setup presented in this study was different from a conventional RNA-seq experiment with separated control and treatment samples. Rather, in our samples many species were growing with putatively different dynamics, thus forming a complex system to analyze. Following the evidence given by microscopic counts, we tried to reduce this complexity by grouping samples into 3 states (“lag”, “growth” and “decline”) according to the HFs growth curve. This separation was also supported by metatranscriptomic data, so we could treat samples separately: “lag” samples were representative of the initial state of the community, still dominated by photosynthesis and with a basal bacterivory activity, while “growth” samples was the state with highest bacterivory. The case of “decline” samples was less clear, as the HF counts decreased but we also identified a few species growing at that phase. This likely explains why the expression patterns of “decline” samples behave similarly to “growth” samples.

3.4.3. Highly expressed genes during bacterivory

Following the separation on states, we focused on the “growth” samples to analyze which genes had highest expression values during bacterivory. Some of the most expressed categories belonged to constitutive processes, with genes involved in several cellular functions (like actin or tubulin), and we did not analyze them any further. Instead, we focused on another set of highly-expressed genes that belonged to peptidases, translocases and CAZy enzymes, as these could be promising targets in the study of bacterivory.

The majority of the highly-expressed peptidase genes belonged to cathepsins. According to the overlap displayed in functional annotations, these clustered into separate groups by their catalytic mechanism (cysteine, serine and aspartyl peptidases). An exhaustive analysis should be performed to identify the phylogenetic relationships within these groups, as most of them were described in humans or model organisms (Berti & Storer, 1995; Ritonja et al., 1988) and little is known for uncultured protists species. Cathepsins are mainly found in the lysosome, where they act as digestive enzymes degrading proteins in acidic conditions (Turk et al., 2012). Their high gene expression during bacterivory was consistent with the fact that more than 60% of bacterial dry weight is composed by proteins (Simon & Azam, 1989). Some peptidases have been previously found in phagosomes, such as cathepsin L in metazoan macrophages (Szulc-Dąbrowska et al., 2020) or cathepsin D in the amoeba *Dictyostelium discoideum* (Gotthardt et al., 2002). A recent paper working with the bicosoecid flagellate *Cafeteria burkhardae* detected an abundant cysteine peptidase gene being differentially expressed during bacterivory (Massana et al., 2021), and other studies have reported their presence in mixotrophic algae (McKie-Krisberg et al., 2018), mixotrophic dinoflagellates (Cohen et al., 2021) and MAST groups (Labarre et al., 2020). Most of the highly-expressed peptidase genes identified here also displayed a higher fold change than house-keeping genes when comparing “lag” and “growth” states, a result that could support their role in bacterivory-related cellular processes in HFs.

Within translocases, V-type ATPases and inorganic pyrophosphatase genes were among those highly-expressed in “growth” samples. V-type ATPase proton pumps are involved in the acidification of the lysosome, among other organelles (Nelson et al., 2000), and it has been hypothesized that pyrophosphatase could play a similar role in protists not belonging to Opisthokonta and Amoebozoa supergroups (Baltscheffsky et al., 1999; Labarre et al., 2021), where this protein is absent. The gene for this proton pump was more expressed than the V-type ATPase in the above-mentioned *C. burkhardae* study, thus supporting this view (Massana et al., 2021). Despite being highly expressed, we did not find a clear pattern of higher expression of the pyrophosphatase proton pump between “growth” and “lag” states, whereas V-ATPases-related genes showed contrasting trends depending on the targeted subunit. This could be explained by the complexity of the dynamics in the incubations and the fact that these proton pumps may have active roles in other cellular processes.

Carbohydrate-active (CAZy) enzymes are related to carbohydrate metabolism (Lombard et al., 2014), including glycoside hydrolases (GH) and glycosyltransferases (GT). Glycosyltransferases create glycosidic bonds (Bourne & Henrissat, 2001), although the reactions they catalyze can be reversible (Zhang et al., 2006). GT66 gene, involved in the N-glycosylation of proteins (Henrissat et al., 2008) – a highly-conserved metabolic process obligatory for viability in eukaryotes (Kukuruzinska & Lennon, 1998) –, as well as GT75 gene, which can be associated to cell wall metabolism (Popper et al., 2011), were overexpressed in the “growth” state of the incubations compared to HK genes. Regarding GT66, this could represent an increase in

N-glycans, the functions of which are well-studied in humans but mostly unknown for protists (Hykollari et al., 2017), while the high expression of GT75 could be due to either a biosynthetic (coming from a remaining photosynthetic signal) or degradative activity in "growth" samples. In the case of glycoside hydrolases, they are responsible for carbohydrate hydrolysis (Bourne & Henrissat, 2001), and they could be involved in bacterivory as digestive enzymes. It was recently reported that GHs account on average for 3% of predicted genes in four MAST-4 species (Latorre et al., 2021), which are among the most abundant HFs in the ocean (Obiol et al., 2021). A gene encoding a chitinase belonging to GH18 was the most expressed GH in "growth" samples. Chitin is mostly found in metazoans, fungi and diatoms (Cheng et al., 2021), and the presence of chitinase in picosized HFs indicates that these enzymes may have other physiological functions that are still unknown (Taira et al., 2018). Another highly-expressed GH gene in our incubations was hexosaminidase (GH20). Hexosaminidases are abundant components found in phagosomes of the parasite *Entamoeba histolytica* (Okada et al., 2005) and several GH20 genes were found in the genomes of MAST-4 and -3 species (Seeleuthner et al., 2018). This, together with their high expression reported here, suggests that GH20 enzymes are also present in phagosomes of HF.

3.4.4. Expression of bacterivory genes at the species level

From the subset of more than 1000 species having genomic/transcriptomic data, we found 25 of them in the incubations, as we detected them in the metatranscriptomes. Ten of them had a partial genome obtained through single-cell genomics (SCG), highlighting both the number of uncultured taxa growing in the incubations and the power of SCG to access the genomes of environmental taxa. As expected, all the species identified had a picoplanktonic size (≤ 3 µm), with the exception of the diatom *Leptocylindrus hargravesii*, which belongs to the microplankton (20-200 µm). Technically this species should not be detected in our data sets, as we performed the incubations with seawater prefiltered by 3µm, and a possible explanation could be that we are targeting gametes produced for sexual reproduction (Nanjappa et al., 2017) that are passing through filters. When assessing the presence of given species in our metatranscriptomes, we noted several cases of transcripts with a median identity at the nucleotide level of 90-98% to a species in the database, indicating the presence of a close relative, but not the reference species (**Fig. S6**). If we take as an example MAST-4A and MAST-4B species available in EukProt, which are very close phylogenetically (2 bp difference at the V4 rDNA amplicon), and compare their coding sequences, the median identity we obtain is around 85% (**Fig. S9**). So, it could be inferred that 25 of the reference species, together with additional closely related ones, were growing in the incubations.

With the expression data retrieved from the 25 species, which represented different trophic modes, we could test whether the highly expressed genes in "growth" samples were more expressed in bacterivorous species compared to phototrophic ones, and thus could be related to the process of bacterivory. Given that phototrophs may have been in suboptimal conditions under dark conditions, one could argue that this analysis could be biased towards heterotrophic modes. However, if this happened, we would have detected strong differences in relative expression for phototrophic species between initial samples, where PF abundance was highest and the absence of light was just starting, and the rest of the incubations (i.e., we would have detected 2 separated clusters of points in Fig. 5, forming a taller violin plot).

Our analysis revealed that the set of peptidase genes were generally more expressed in species with a heterotrophic mode of nutrition, with cysteine peptidases representing up to 3% of the total gene expression of some species. This suggests that, although these could participate in other cellular processes, they have a key role in bacterivory as digestive enzymes. Proton pumps (translocases) did not display these marked differences between trophic modes, suggesting that they actively participate in other cellular processes apart from bacterivory. For CAZy enzymes, GH20 and GH3 were the genes displaying the most marked differences of relative expression between heterotrophic species and mixotrophic/phototrophic ones. Thus, these could be key players in bacterivory, with a role in the digestion of ingested bacteria in the phagolysosome. Although the transcripts associated to the 25 identified species only represent 2.5% of the total signal in the 4 metatranscriptomes (data not shown), this analysis represents a proof-of-concept of what can be achieved with more reference genomes from species that refuse culturing.

Altogether, our study gives new insights in HF functional dynamics and more specifically the expression of genes related to bacterivory. Our results indicate that cysteine peptidases and glycoside hydrolases 3 and 20 may be key players in this process. A further analysis of these genes and their expression in more targeted studies could reveal the intricate mechanisms of this understudied but crucial process in marine ecosystems.

3.5. Acknowledgements

This research was supported by the Spanish Ministry of Science and Innovation projects ALLFLAGS (CTM2016-75083-R), DIVAS (PID2019-108457RB-I00) and the "Severo Ochoa Centre of Excellence" accreditation (CEX2019-000928-S), and the European Union project SINGEK (H2020-MSCA-ITN-2015-675752). AO was supported by a Spanish FPI grant. Bioinformatic analyses were performed at the Marbits (ICM-CSIC; https://marbits.icm.csic.es), FinisTerrae II (CESGA; https://www.cesga.es/) and Mole platforms (https://www.paru.cas.cz/).

3.6. Data availability statement

Raw data for “Jul17” experiment were already published at the NCBI BioSample database with accession number SAMN11783926 (Labarre et al., 2020). For the rest of the experiments, raw data will be deposited at NCBI after acceptance for publication. Assemblies, quantification and functional annotation tables will be uploaded at figshare and all code used for data processing and analyses will be available at GitHub.

3.7. References

Alexander, H., Hu, S. K., Krinos, A. I., Pachiadaki, M., Tully, B. J., Neely, C. J., & Reiter, T. (2022). Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton (p. 2021.07.25.453713). *bioRxiv*. https://doi.org/10.1101/2021.07.25.453713

Alexander, H., Jenkins, B. D., Rynearson, T. A., & Dyhrman, S. T. (2015). Metatranscriptome analyses indicate resource partitioning between diatoms in the field. *Proceedings of the National Academy of Sciences*, 112(17), E2182–E2190. https://doi.org/10.1073/pnas.1421993112

Alexander, H., Jenkins, B., Rynearson, T., Saito, M., Mercier, M., & Dyhrman, S. (2012). Identifying reference genes with stable expression from high throughput sequence data. *Frontiers in Microbiology*, 3. https://www.frontiersin.org/article/10.3389/fmicb.2012.00385

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. https://doi.org/10.1016/S0022-2836(05)80360-2

Baltscheffsky, M., Schultz, A., & Baltscheffsky, H. (1999). H+-PPases: A tightly membrane-bound family. *FEBS Letters*, 457(3), 527–533. https://doi.org/10.1016/S0014-5793(99)90617-8

Berti, P. J., & Storer, A. C. (1995). Alignment/Phylogeny of the Papain Superfamily of Cysteine Proteases. *Journal of Molecular Biology*, 246(2), 273–283. https://doi.org/10.1006/jmbi.1994.0083

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

Botelho, R. J., & Grinstein, S. (2011). Phagocytosis. *Current Biology*, 21(14), R533–R538. https://doi.org/10.1016/j.cub.2011.05.053

Boulais, J., Trost, M., Landry, C. R., Dieckmann, R., Levy, E. D., Soldati, T., Michnick, S. W., Thibault, P., & Desjardins, M. (2010). Molecular characterization of the evolution of phagosomes. *Molecular Systems Biology*, 6(1), 423. https://doi.org/10.1038/msb.2010.80

Bourne, Y., & Henrissat, B. (2001). Glycoside hydrolases and glycosyltransferases: Families and functional modules. *Current Opinion in Structural Biology*, 11(5), 593–600. https://doi.org/10.1016/S0959-440X(00)00253-0

Bozzaro, S., Bucci, C., & Steinert, M. (2008). Phagocytosis and Host–Pathogen Interactions in Dictyostelium with a Look at Macrophages. In *International Review of Cell and Molecular Biology* (Vol. 271, pp. 253–300). Academic Press. https://doi.org/10.1016/S1937-6448(08)01206-9

Bucchini, F., Del Cortona, A., Kreft, Ł., Botzki, A., Van Bel, M., & Vandepoele, K. (2021). TRAPID 2.0: A web application for taxonomic and functional analysis of de novo transcriptomes. *Nucleic Acids Research*, gkab565. https://doi.org/10.1093/nar/gkab565

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4), 366–368. https://doi.org/10.1038/s41592-021-01101-x

Burns, J. A., Paasch, A., Narechania, A., & Kim, E. (2015). Comparative Genomics of a Bacterivorous Green Alga Reveals Evolutionary Causalities and Consequences of Phago-Mixotrophic Mode of Nutrition. *Genome Biology and Evolution*, 7(11), 3047–3061. https://doi.org/10.1093/gbe/evv144

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

Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12), 5825–5829. https://doi.org/10.1093/molbev/msab293

Caron, D. A., Alexander, H., Allen, A. E., Archibald, J. M., Armbrust, E. V., Bachy, C., Bell, C. J., Bharti, A., Dyhrman, S. T., Guida, S. M., Heidelberg, K. B., Kaye, J. Z., Metzner, J., Smith, S. R., & Worden, A. Z. (2016). Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology*, 15(1), 6–20. https://doi.org/10.1038/nrmicro.2016.160

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., … Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, 9(1), 373. https://doi.org/10.1038/s41467-017-02342-1

Cavicchioli, R., Ripple, W. J., Timmis, K. N., Azam, F., Bakken, L. R., Baylis, M., Behrenfeld, M. J., Boetius, A., Boyd, P. W., Classen, A. T., Crowther, T. W., Danovaro, R., Foreman, C. M., Huisman, J., Hutchins, D. A., Jansson, J. K., Karl, D. M., Koskella, B., Mark Welch, D. B., … Webster, N. S. (2019). Scientists' warning to humanity: Microorganisms and climate change. *Nature Reviews Microbiology*, 17(September). https://doi.org/10.1038/s41579-019-0222-5

Cheng, H., Shao, Z., Lu, C., & Duan, D. (2021). Genome-wide identification of chitinase genes in Thalassiosira pseudonana and analysis of their expression under abiotic stresses. *BMC Plant Biology*, 21(1), 87. https://doi.org/10.1186/s12870-021-02849-2

Cohen, N. R., Alexander, H., Krinos, A. I., Hu, S. K., & Lampe, R. H. (2022). Marine Microeukaryote Metatranscriptomics: Sample Processing and Bioinformatic Workflow Recommendations for Ecological Applications. *Frontiers in Marine Science*, 9. https://www.frontiersin.org/article/10.3389/fmars.2022.867007

Cohen, N. R., McIlvin, M. R., Moran, D. M., Held, N. A., Saunders, J. K., Hawco, N. J., Brosnahan, M., DiTullio, G. R., Lamborg, C., McCrow, J. P., Dupont, C. L., Allen, A. E., & Saito, M. A. (2021). Dinoflagellates alter their carbon and nutrient metabolic strategies across environmental gradients in the central Pacific Ocean. *Nature Microbiology*, 6(2), 173–186. https://doi.org/10.1038/s41564-020-00814-7

Cui, P., Lin, Q., Ding, F., Xin, C., Gong, W., Zhang, L., Geng, J., Zhang, B., Yu, X., Yang, J., Hu, S., & Yu, J. (2010). A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. *Genomics*, 96(5), 259–265. https://doi.org/10.1016/j.ygeno.2010.07.010

Daufresne, M., Lengfellner, K., & Sommer, U. (2009). Global warming benefits the small in aquatic ecosystems. *Proceedings of the National Academy of Sciences*, 106(31), 12788–12793. https://doi.org/10.1073/pnas.0902080106

Dayel, M. J., & King, N. (2014). Prey Capture and Phagocytosis in the Choanoflagellate Salpingoeca rosetta. *PLoS ONE*, 9(5), e95577. https://doi.org/10.1371/journal.pone.0095577

del Campo, J., Sieracki, M. E., Molestina, R., Keeling, P. J., Massana, R., & Ruiz-Trillo, I. (2014). The others: Our biased perspective of eukaryotic genomes. *Trends in Ecology & Evolution*, 29(5), 252–259. https://doi.org/10.1016/j.tree.2014.03.006

Delmont, T. O., Gaia, M., Hinsinger, D. D., Frémont, P., Vanni, C., Fernandez-Guerra, A., Eren, A. M., Kourlaiev, A., d'Agata, L., Clayssen, Q., Villar, E., Labadie, K., Cruaud, C., Poulain, J., Da Silva, C., Wessner, M., Noel, B., Aury, J.-M., Sunagawa, S., … Jaillon, O. (2022). Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics*, 100123. https://doi.org/10.1016/j.xgen.2022.100123

Duncan, A., Barry, K., Daum, C., Eloe-Fadrosh, E., Roux, S., Schmidt, K., Tringe, S. G., Valentin, K. U., Varghese, N., Salamov, A., Grigoriev, I. V., Leggett, R. M., Moulton, V., & Mock, T. (2022). Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans. *Microbiome*, 10(1), 67. https://doi.org/10.1186/s40168-022-01254-7

Fenchel, T. (1986). The Ecology of Heterotrophic Microflagellates. In K. C. Marshall (Ed.), *Advances in Microbial Ecology* (pp. 57–97). Springer US. https://doi.org/10.1007/978-1-4757-0611-6_2

Flannagan, R. S., Cosío, G., & Grinstein, S. (2009). Antimicrobial mechanisms of phagocytes and bacterial evasion strategies. *Nature Reviews Microbiology*, 7(5), 355–366. https://doi.org/10.1038/nrmicro2128

Gasol, J. M., Cardelús, C., Morán, X. A. G., Balagué, V., Forn, I., Marrasé, C., Massana, R., Pedrós-Alió, C., Sala, M. M., Simó, R., Vaqué, D., & Estrada, M. (2016). Seasonal patterns in phytoplankton photosynthetic parameters and primary production at a coastal NW Mediterranean site. *Scientia Marina*, 80(S1), 63–77. https://doi.org/10.3989/scimar.04480.06E

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

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

Gotthardt, D., Warnatz, H. J., Henschel, O., Brückert, F., Schleicher, M., & Soldati, T. (2002). High-Resolution Dissection of Phagosome Maturation Reveals Distinct Membrane Trafficking Phases. *Molecular Biology of the Cell*, 13(10), 3508–3520. https://doi.org/10.1091/mbc.e02-04-0206

Henrissat, B., Sulzenbacher, G., & Bourne, Y. (2008). Glycosyltransferases, glycoside hydrolases: Surprise, surprise! *Current Opinion in Structural Biology*, 18(5), 527–533. https://doi.org/10.1016/j.sbi.2008.09.003

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). eggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, 47(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Hutchins, D. A., & Fu, F. (2017). Microorganisms and ocean global change. *Nature Microbiology*, 2(6), 17058. https://doi.org/10.1038/nmicrobiol.2017.58

Hykollari, A., Paschinger, K., Eckmair, B., & Wilson, I. B. H. (2017). Analysis of Invertebrate and Protist N-Glycans. In G. Lauc & M. Wuhrer (Eds.), *High-Throughput Glycomics and Glycoproteomics: Methods and Protocols* (pp. 167–184). Springer. https://doi.org/10.1007/978-1-4939-6493-2_13

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). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470281840.ch11

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., & Tanabe, M. (2021). KEGG: Integrating viruses and cellular organisms. *Nucleic Acids Research*, 49(D1), D545–D551. https://doi.org/10.1093/nar/gkaa970

Keeling, P. J., & del Campo, J. (2017). Marine Protists Are Not Just Big Bacteria. *Current Biology*, 27(11), R541–R549. https://doi.org/10.1016/j.cub.2017.03.075

Kolody, B. C., McCrow, J. P., Allen, L. Z., Aylward, F. O., Fontanez, K. M., Moustafa, A., Moniruzzaman, M., Chavez, F. P., Scholin, C. A., Allen, E. E., Worden, A. Z., Delong, E. F., & Allen, A. E. (2019). Diel transcriptional response of a California Current plankton microbiome to light, low iron, and enduring viral infection. *The ISME Journal*. https://doi.org/10.1038/s41396-019-0472-2

Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, 28(24), 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

Kukuruzinska, M. A., & Lennon, K. (1998). Protein N-Glycosylation: Molecular Genetics and Functional Significance. *Critical Reviews in Oral Biology & Medicine*, 9(4), 415–448. https://doi.org/10.1177/10454411980090040301

Labarre, A., López-Escardó, D., Latorre, F., Leonard, G., Bucchini, F., Obiol, A., Cruaud, C., Sieracki, M. E., Jaillon, O., Wincker, P., Vandepoele, K., Logares, R., & Massana, R. (2021). Comparative genomics reveals new functional insights in uncultured MAST species. *The ISME Journal*, 15(6), 1767–1781. https://doi.org/10.1038/s41396-020-00885-8

Labarre, A., Obiol, A., Wilken, S., Forn, I., & Massana, R. (2020). Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates. *Limnology and Oceanography*, 65(S1), lno.11379. https://doi.org/10.1002/lno.11379

Lambert, B. S., Groussman, R. D., Schatz, M. J., Coesel, S. N., Durham, B. P., Alverson, A. J., White, A. E., & Armbrust, E. V. (2022). The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. *Proceedings of the National Academy of Sciences*, 119(7). https://doi.org/10.1073/pnas.2100916119

Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., & Galand, P. E. (2018). Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *The ISME Journal*, submitted. https://doi.org/10.1038/s41396-018-0281-z

Latorre, F., Deutschmann, I. M., Labarre, A., Obiol, A., Krabberød, A. K., Pelletier, E., Sieracki, M. E., Cruaud, C., Jaillon, O., Massana, R., & Logares, R. (2021). Niche adaptation promoted the evolutionary diversification of tiny ocean predators. *Proceedings of the National Academy of Sciences*, 118(25). https://doi.org/10.1073/pnas.2020955118

Li, W. K. W., McLaughlin, F. A., Lovejoy, C., & Carmack, E. C. (2009). Smallest Algae Thrive As the Arctic Ocean Freshens. *Science*, 326(5952), 539–539. https://doi.org/10.1126/science.1179798

Li, Z., Zhang, Y., Li, W., Irwin, A. J., & Finkel, Z. V. (2022). Conservation and architecture of housekeeping genes in the model marine diatom Thalassiosira pseudonana. *New Phytologist*, 234(4), 1363–1376. https://doi.org/10.1111/nph.18039

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, 42(D1), D490–D495. https://doi.org/10.1093/nar/gkt1178

Louca, S., Polz, M. F., Mazel, F., Albright, M. B. N., Huber, J. A., O'Connor, M. I., Ackermann, M., Hahn, A. S., Srivastava, D. S., Crowe, S. A., Doebeli, M., & Parfrey, L. W. (2018). Function and functional redundancy in microbial systems. *Nature Ecology & Evolution*, 2(6), 936–943. https://doi.org/10.1038/s41559-018-0519-1

Louyakis, A. S., Gourlé, H., Casaburi, G., Bonjawo, R. M. E., Duscher, A. A., & Foster, J. S. (2018). A year in the life of a thrombolite: Comparative metatranscriptomics reveals dynamic metabolic changes over diel and seasonal cycles. *Environmental Microbiology*, 20(2), 842–861. https://doi.org/10.1111/1462-2920.14029

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Massana, R., Guillou, L., Terrado, R., Forn, I., & Pedrós-Alió, C. (2006). Growth of uncultured heterotrophic flagellates in unamended seawater incubations. *Aquatic Microbial Ecology*, 45(2), 171–180. https://doi.org/10.3354/ame045171

Massana, R., Labarre, A., López-Escardó, D., Obiol, A., Bucchini, F., Hackl, T., Fischer, M. G., Vandepoele, K., Tikhonenkov, D. V., Husnik, F., & Keeling, P. J. (2021). Gene expression during bacterivorous growth of a widespread marine heterotrophic flagellate. *The ISME Journal*, 15(1), 154–167. https://doi.org/10.1038/s41396-020-00770-4

McKie-Krisberg, Z. M., Sanders, R. W., & Gast, R. J. (2018). Evaluation of Mixotrophy-Associated Gene Expression in Two Species of Polar Marine Algae. *Frontiers in Marine Science*, 5. https://doi.org/10.3389/fmars.2018.00273

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature Communications*, 7(1), 11257. https://doi.org/10.1038/ncomms11257

Metegnier, G., Paulino, S., Ramond, P., Siano, R., Sourisseau, M., Destombe, C., & Le Gac, M. (2020). Species specific gene expression dynamics during harmful algal blooms. *Scientific Reports*, 10(1), 6182. https://doi.org/10.1038/s41598-020-63326-8

Mika, K., Marinić, M., Singh, M., Muter, J., Brosens, J. J., & Lynch, V. J. (2021). Evolutionary transcriptomics implicates new genes and pathways in human pregnancy and adverse pregnancy outcomes. *ELife*, 10, e69584. https://doi.org/10.7554/eLife.69584

Mills, D. B. (2020). The origin of phagocytosis in Earth history. *Interface Focus*, 10(4), 20200019. https://doi.org/10.1098/rsfs.2020.0019

Muratore, D., Boysen, A. K., Harke, M. J., Becker, K. W., Casey, J. R., Coesel, S. N., Mende, D. R., Wilson, S. T., Aylward, F. O., Eppley, J. M., Vislova, A., Peng, S., Rodriguez-Gonzalez, R. A., Beckett, S. J., Virginia Armbrust, E., DeLong, E. F., Karl, D. M., White, A. E., Zehr, J. P., … Weitz, J. S. (2022). Complex marine microbial communities partition metabolism of scarce resources over the diel cycle. *Nature Ecology & Evolution*, 6(2), 218–229. https://doi.org/10.1038/s41559-021-01606-w

Nanjappa, D., Sanges, R., Ferrante, M. I., & Zingone, A. (2017). Diatom flagellar genes and their expression during sexual reproduction in Leptocylindrus danicus. *BMC Genomics*, 18(1), 813. https://doi.org/10.1186/s12864-017-4210-8

Nelson, N., Perzov, N., Cohen, A., Hagai, K., Padler, V., & Nelson, H. (2000). The cellular biology of proton-motive force generation by V-ATPases. *Journal of Experimental Biology*, 203(1), 89–95. https://doi.org/10.1242/jeb.203.1.89

Obiol, A., Giner, C. R., Sánchez, P., Duarte, C. M., Acinas, S. G., & Massana, R. (2020). A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Molecular Ecology Resources*, 20(3), 718–731. https://doi.org/10.1111/1755-0998.13147

Obiol, A., Muhovic, I., & Massana, R. (2021). Oceanic heterotrophic flagellates are dominated by a few widespread taxa. *Limnology and Oceanography*, 66(12), 4240–4253. https://doi.org/10.1002/lno.11956

Okada, M., Huston, C. D., Mann, B. J., Petri, W. A., Kita, K., & Nozaki, T. (2005). Proteomic Analysis of Phagocytosis in the Enteric Protozoan Parasite Entamoeba histolytica. *Eukaryotic Cell*, 4(4), 827–831. https://doi.org/10.1128/EC.4.4.827-831.2005

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4), 417–419. https://doi.org/10.1038/nmeth.4197

Pernthaler, J. (2005). Predation on prokaryotes in the water column and its ecological implications. *Nature Reviews Microbiology*, 3(7), 537–546. https://doi.org/10.1038/nrmicro1180

Popper, Z. A., Michel, G., Hervé, C., Domozych, D. S., Willats, W. G. T., Tuohy, M. G., Kloareg, B., & Stengel, D. B. (2011). Evolution and Diversity of Plant Cell Walls: From Algae to Flowering Plants. *Annual Review of Plant Biology*, 62(1), 567–590. https://doi.org/10.1146/annurev-arplant-042110-103809

Prokopchuk, G., Korytář, T., Juricová, V., Majstorović, J., Horák, A., Šimek, K., & Lukeš, J. (2022). Trophic flexibility of marine diplonemids—Switching from osmotrophy to bacterivory. *The ISME Journal*, 1–11. https://doi.org/10.1038/s41396-022-01192-0

R Core Team. (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

Richter, D. J., Berney, C., Strassert, J. F. H., Poh, Y.-P., Herman, E. K., Muñoz-Gómez, S. A., Wideman, J. G., Burki, F., & Vargas, C. de. (2022). EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. In *BioRxiv* (p. 2020.06.30.180687). https://doi.org/10.1101/2020.06.30.180687

Ritonja, A., Popović, T., Kotnik, M., Machleidt, W., & Turk, V. (1988). Amino acid sequences of the human kidney cathepsins H and L. *FEBS Letters*, 228(2), 341–345. https://doi.org/10.1016/0014-5793(88)80028-0

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, 4, e2584. https://doi.org/10.7717/peerj.2584

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H. J., Cuenca, M., Field, C. M., Coelho, L. P., Cruaud, C., Engelen, S., Gregory, A. C., Labadie, K., Marec, C., Pelletier, E., Royo-Llonch, M., Roux, S., Sánchez, P., Uehara, H., Zayed, A. A., … Wincker, P. (2019). Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell*, 179(5), 1068-1083.e21. https://doi.org/10.1016/j.cell.2019.10.014

Sarmento, H., Montoya, J. M., Vázquez-Domínguez, E., Vaqué, D., & Gasol, J. M. (2010). Warming effects on marine microbial food web processes: How far can we go when it comes to predictions? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1549), 2137–2149. https://doi.org/10.1098/rstb.2010.0045

Schön, M. E., Zlatogursky, V. V., Singh, R. P., Poirier, C., Wilken, S., Mathur, V., Strassert, J. F. H., Pinhassi, J., Worden, A. Z., Keeling, P. J., Ettema, T. J. G., Wideman, J. G., & Burki, F. (2021). Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. *Nature Communications*, 12(1), 6651. https://doi.org/10.1038/s41467-021-26918-0

Seeleuthner, Y., Mondy, S., Lombard, V., Carradec, Q., Pelletier, E., Wessner, M., Leconte, J., Mangot, J.-F., Poulain, J., Labadie, K., Logares, R., Sunagawa, S., de Berardinis, V., Salanoubat, M., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Pesant, S., … Wincker, P. (2018). Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nature Communications*, 9(1), 310. https://doi.org/10.1038/s41467-017-02235-3

Sherr, E. B., & Sherr, B. F. (2002). Significance of predation by protists in aquatic microbial food webs. *Antonie van Leeuwenhoek*, 81(1), 293–308. https://doi.org/10.1023/A:1020591307260

Simon, M., & Azam, F. (1989). Protein content and protein synthesis rates of planktonic marine bacteria. *Marine Ecology Progress Series*, 51, 201–213. https://doi.org/10.3354/meps051201

Szulc-Dąbrowska, L., Bossowska-Nowicka, M., Struzik, J., & Toka, F. N. (2020). Cathepsins in Bacteria-Macrophage Interaction: Defenders or Victims of Circumstance? *Frontiers in Cellular and Infection Microbiology*, 10, 601072. https://doi.org/10.3389/fcimb.2020.601072

Taira, T., Gushiken, C., Sugata, K., Ohnuma, T., & Fukamizo, T. (2018). Unique GH18 chitinase from Euglena gracilis: Full-length cDNA cloning and characterization of its catalytic domain. *Bioscience, Biotechnology, and Biochemistry*, 82(7), 1090–1100. https://doi.org/10.1080/09168451.2018.1459463

Tang, S., Lomsadze, A., & Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. *Nucleic Acids Research*, 43(12), e78. https://doi.org/10.1093/nar/gkv227

Turk, V., Stoka, V., Vasiljeva, O., Renko, M., Sun, T., Turk, B., & Turk, D. (2012). Cysteine cathepsins: From structure, function and regulation to new frontiers. *Biochimica et Biophysica Acta. Proteins and Proteomics*, 1824(1), 68–88. https://doi.org/10.1016/j.bbapap.2011.10.002

Vanni, C., Schechter, M. S., Acinas, S. G., Barberán, A., Buttigieg, P. L., Casamayor, E. O., Delmont, T. O., Duarte, C. M., Eren, A. M., Finn, R. D., Kottmann, R., Mitchell, A., Sánchez, P., Siren, K., Steinegger, M., Gloeckner, F. O., & Fernàndez-Guerra, A. (2022). Unifying the known and unknown microbial coding sequence space. *ELife*, 11, e67667. https://doi.org/10.7554/eLife.67667

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. https://doi.org/10.21105/joss.01686

Zhang, C., Griffith, B. R., Fu, Q., Albermann, C., Fu, X., Lee, I.-K., Li, L., & Thorson, J. S. (2006). Exploiting the Reversibility of Natural Product Glycosyltransferase-Catalyzed Reactions. *Science*, 313(5791), 1291–1294. https://doi.org/10.1126/science.1130028

3.8. Supplementary figures

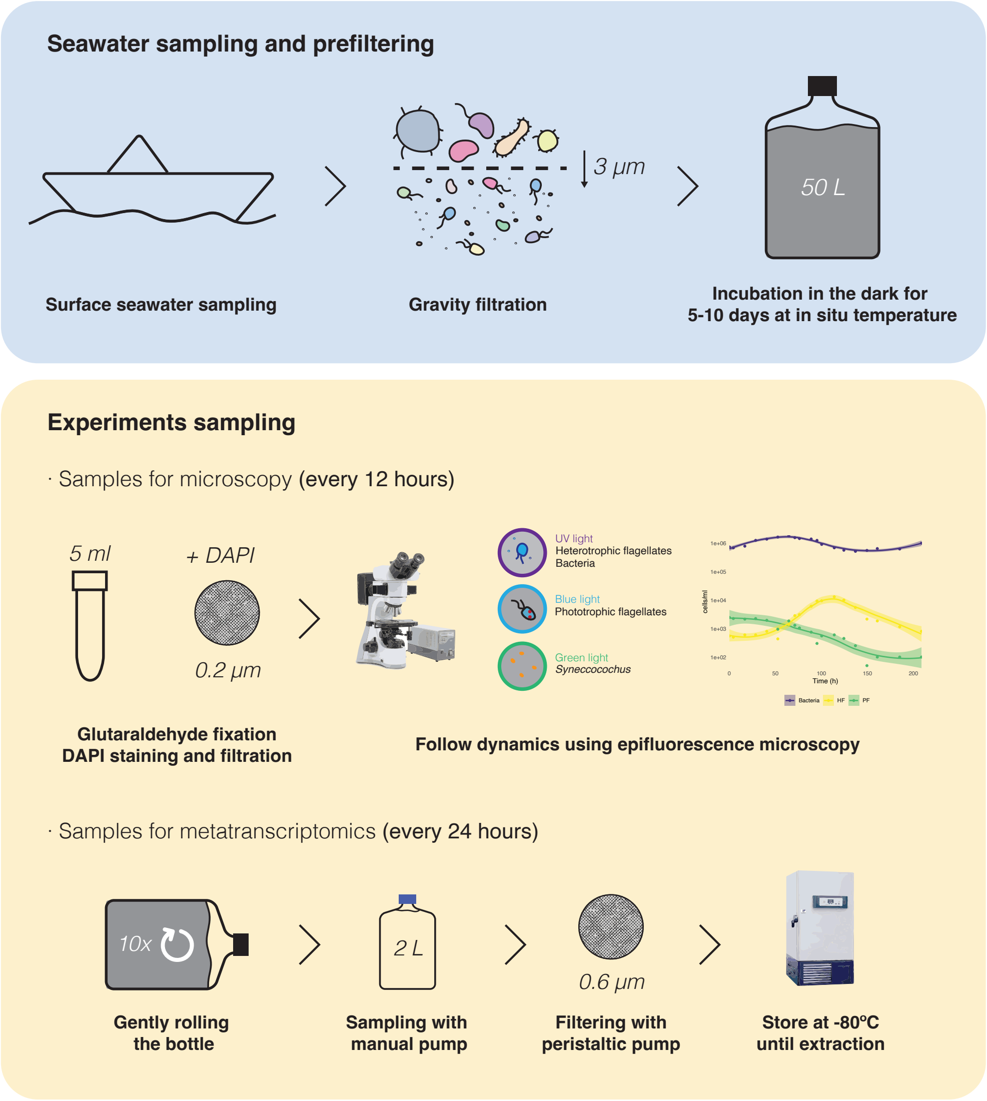


Figure S1. Overview of the experimental setup.

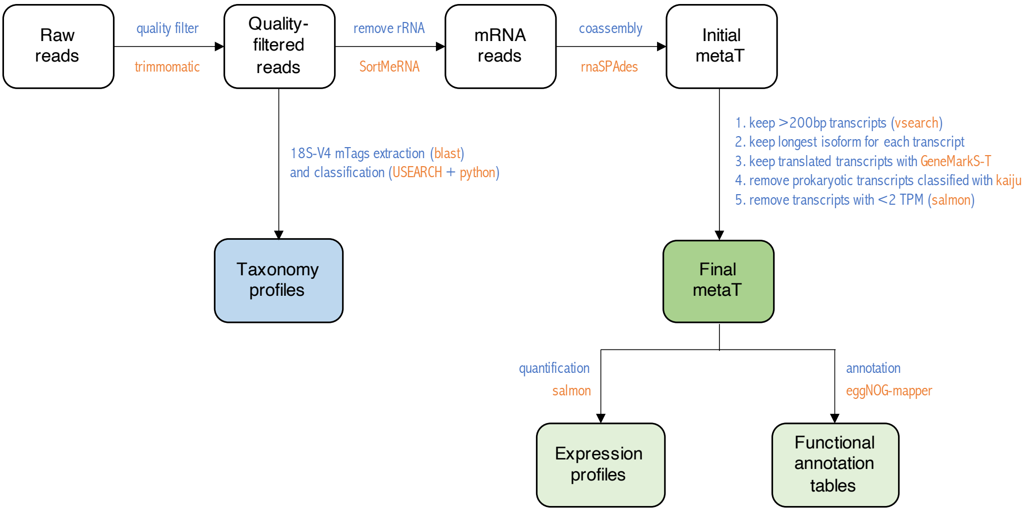

Raw reads
quality filter
trimmomatic
Quality-filtered reads
remove rRNA
SortMeRNA
mRNA reads
coassembly
rnaSPAdes
Initial metaT
18S-V4 mTags extraction (blast) and classification (USEARCH + python)
1. keep >200bp transcripts (vsearch)
2. keep longest isoform for each transcript
3. keep translated transcripts with GeneMarkS-T
4. remove prokaryotic transcripts classified with kaiju
5. remove transcripts with <2 TPM (salmon)
Taxonomy profiles
Final metaT
quantification
salmon
annotation
eggNOG-mapper
Expression profiles
Functional annotation tables

Figure S2. Overview of the bioinformatic processing of the metatranscriptomic reads.

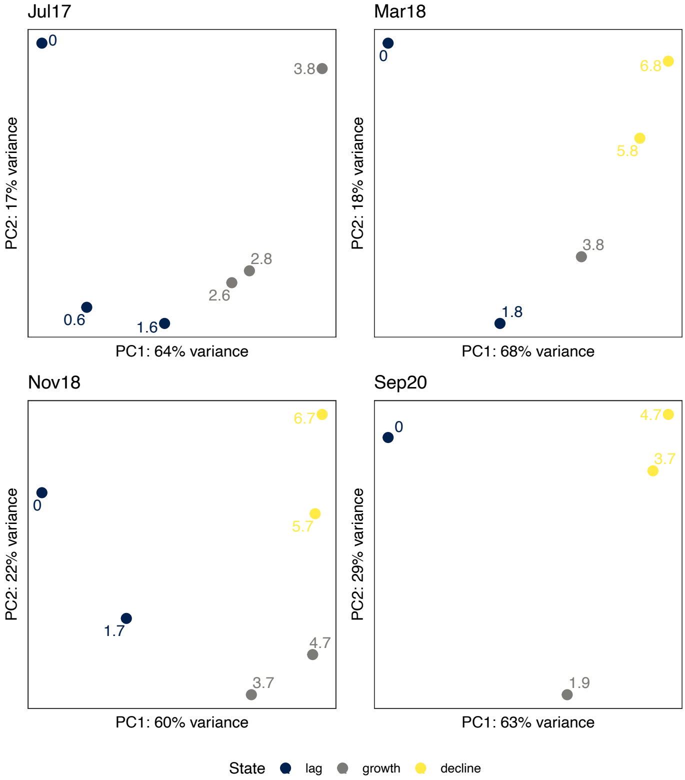


Figure S3. PCA plots used for validation of the different incubation states. Each point represents a sample, and the value next to it represents the time of incubation expressed as days.

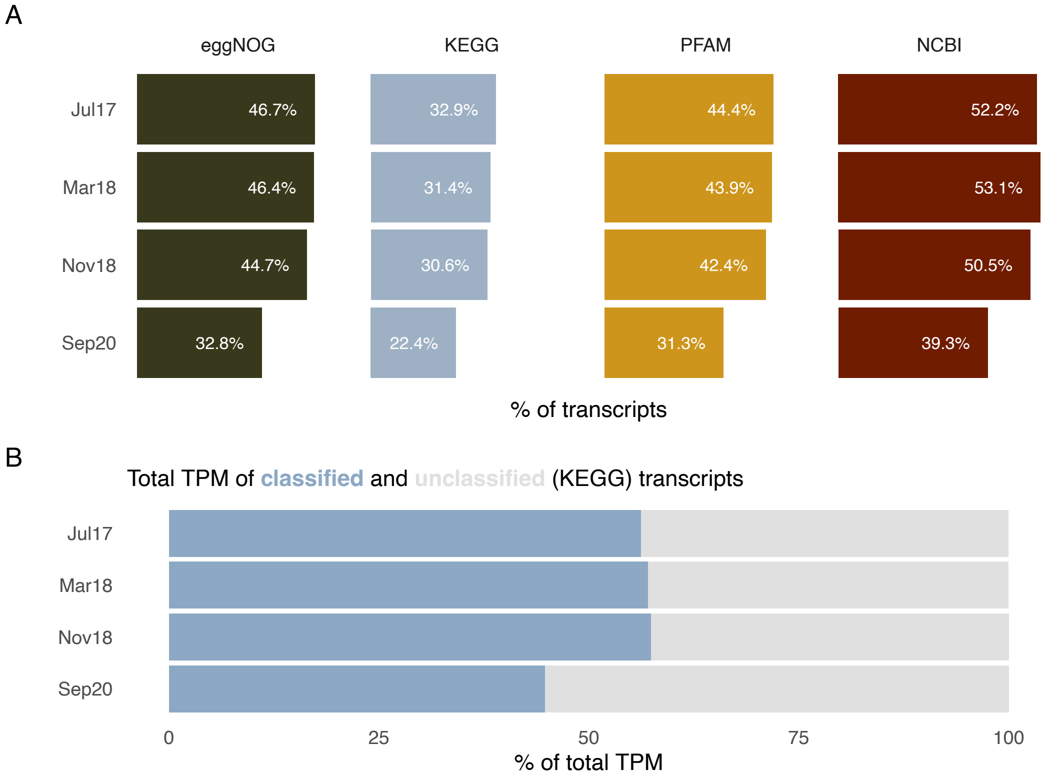


Figure S4. (**A**) Percentage of transcripts annotated with different databases. (**B**) Percentage of total TPM explained by the transcripts annotated with KEGG.

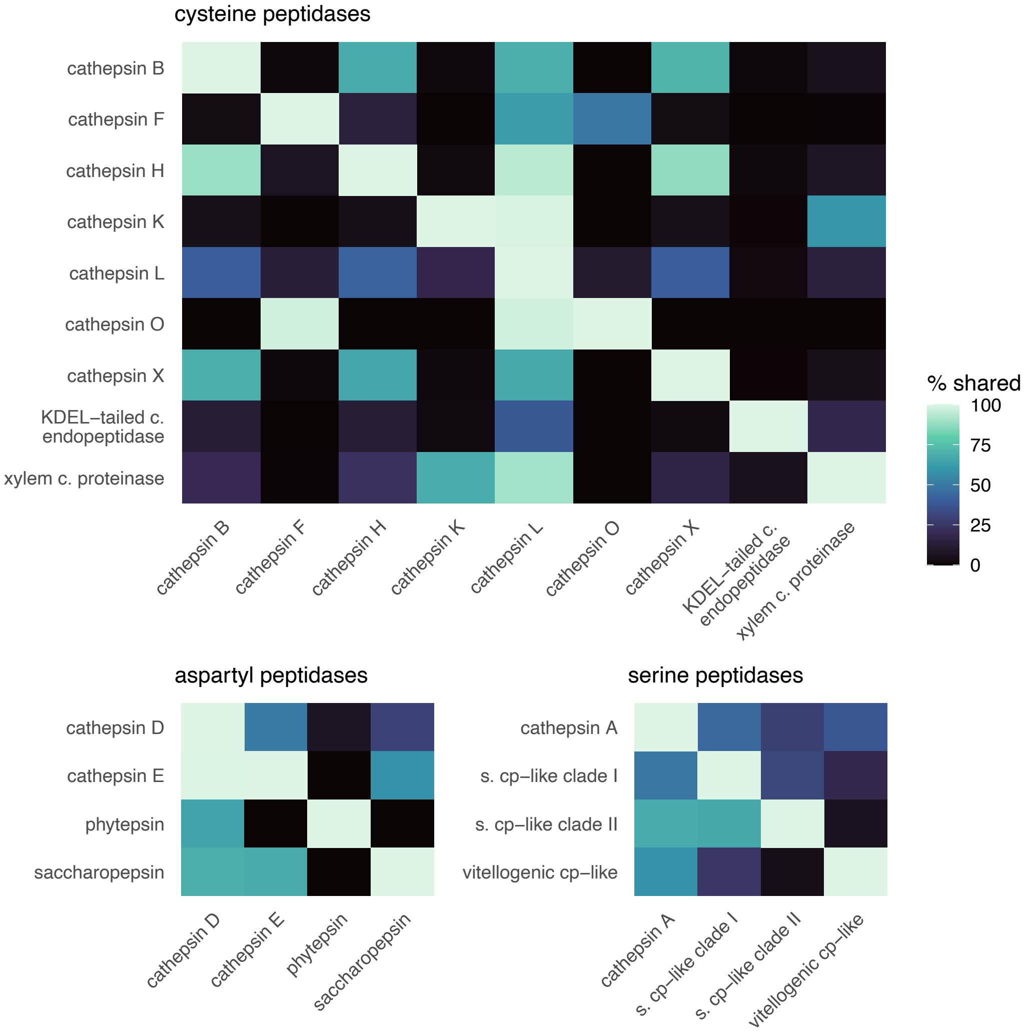
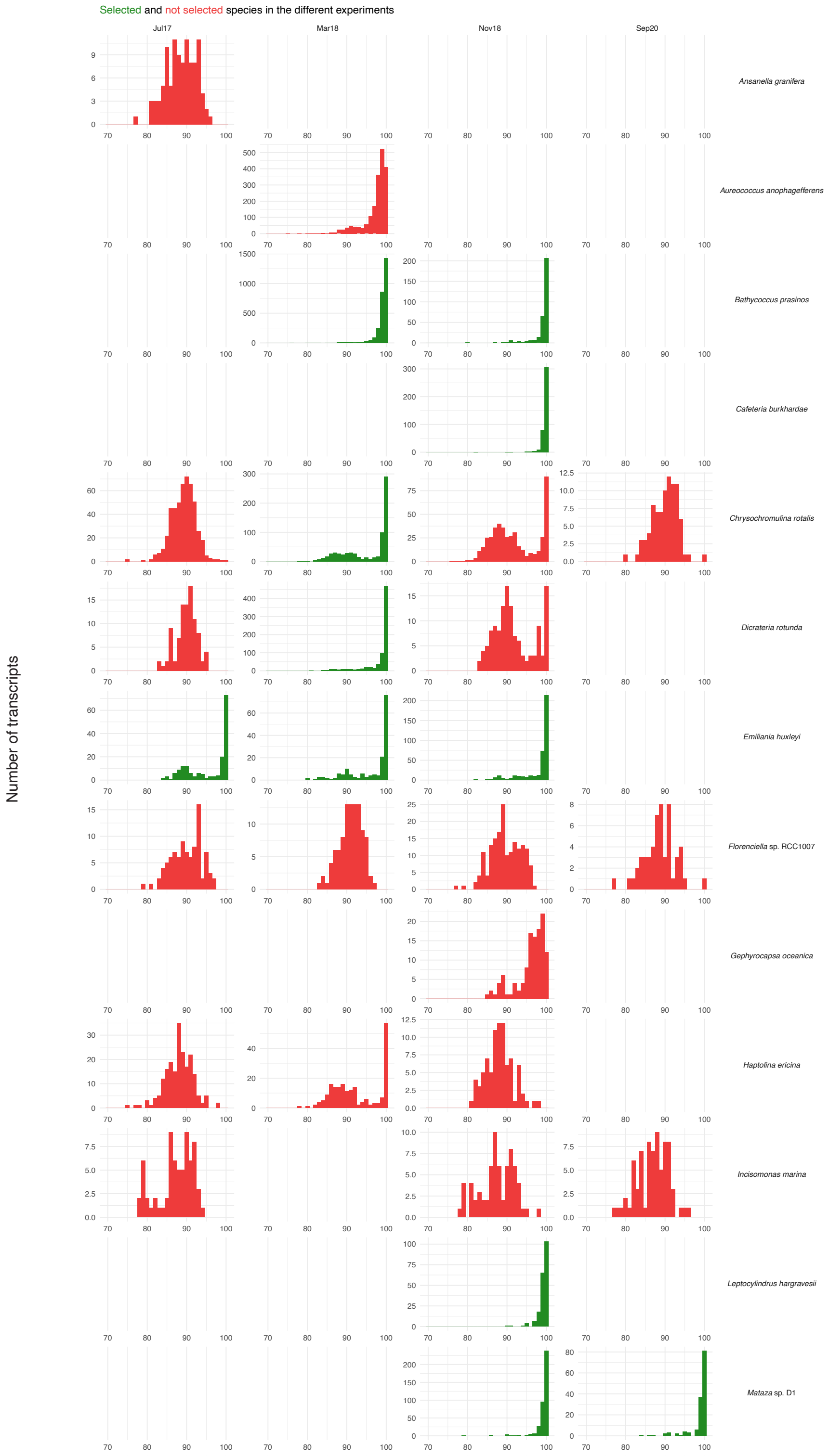
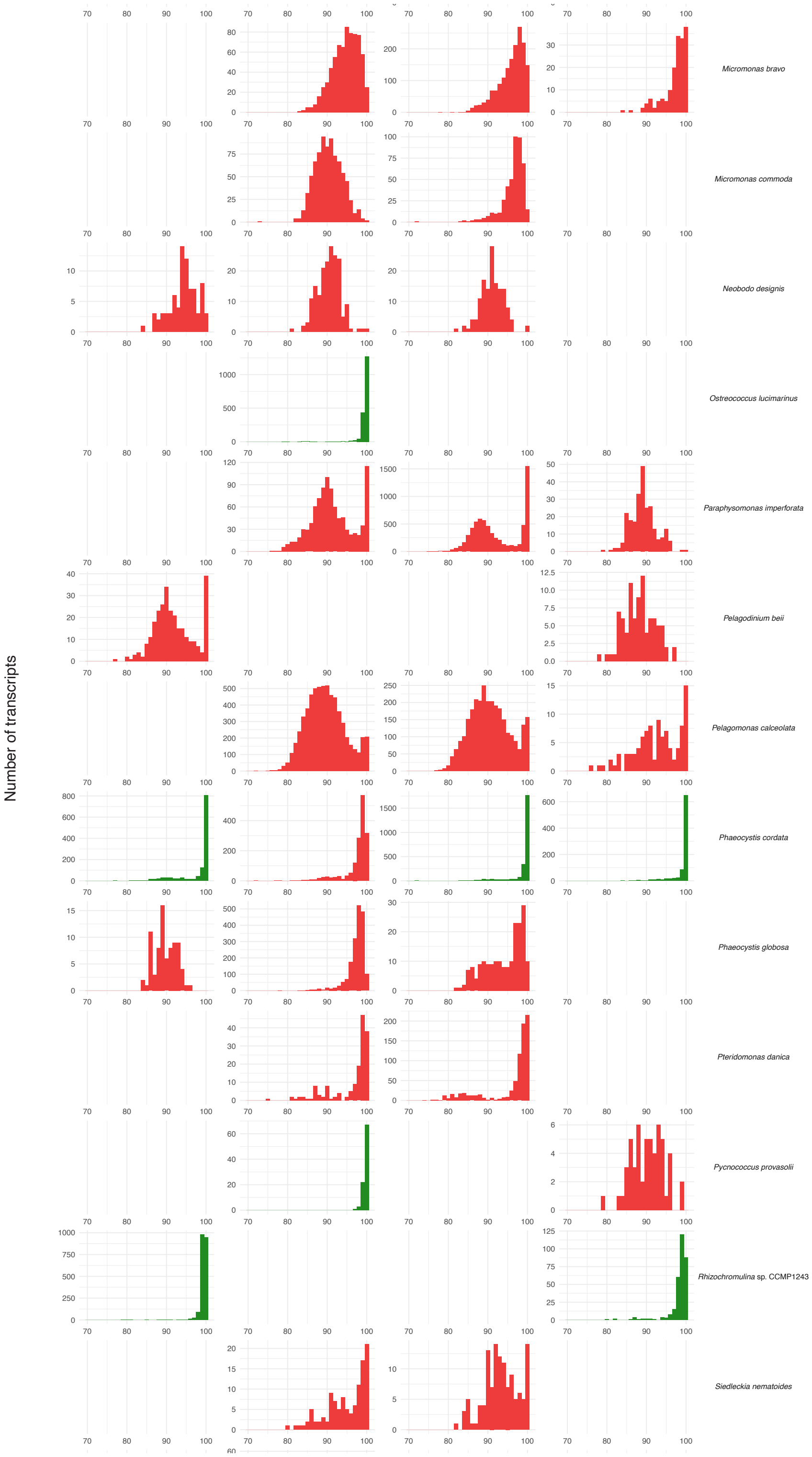
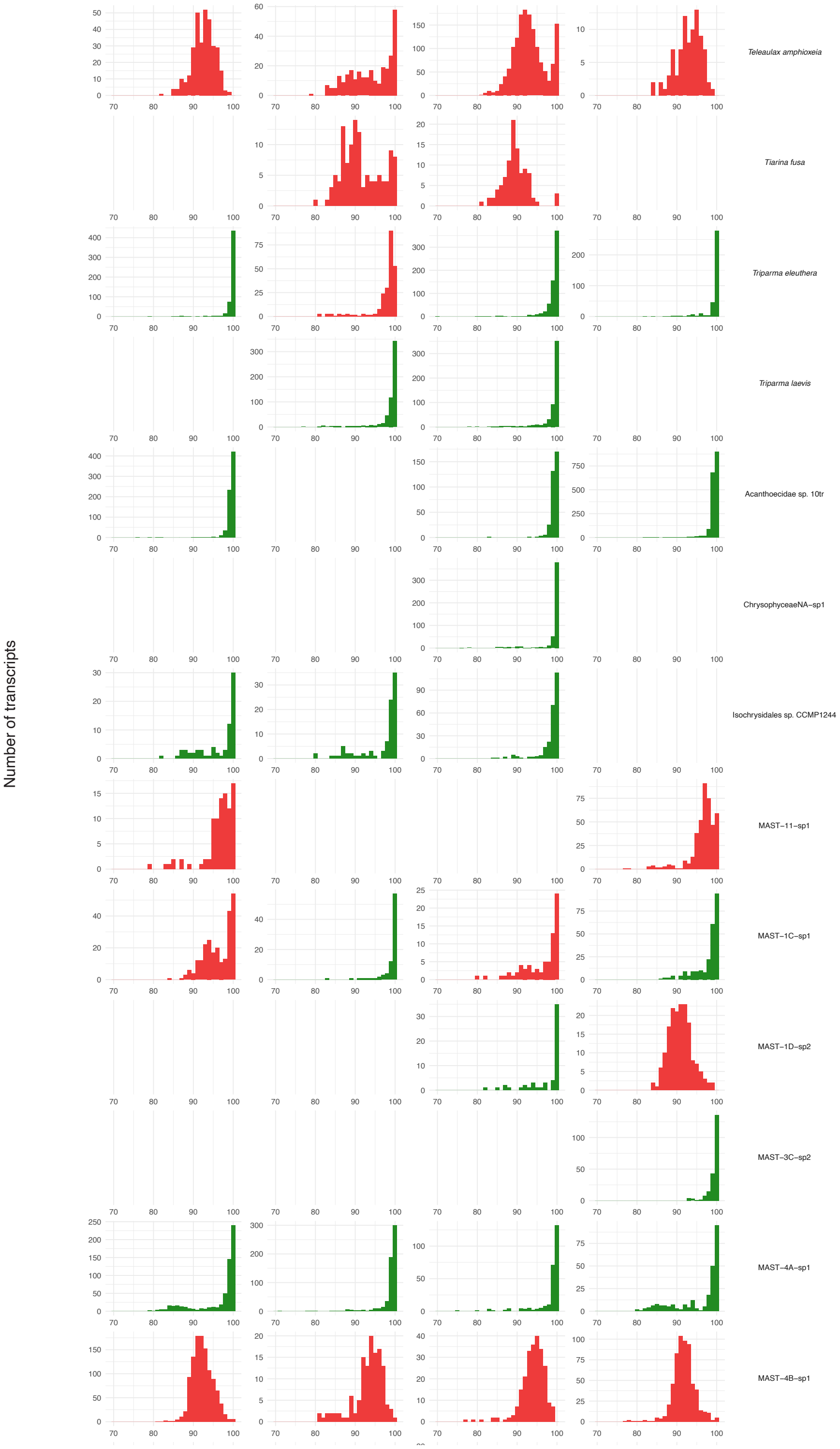


Figure S5. Overlap in functional KEGG annotations in three groups of peptidases (cysteine, aspartyl and serine peptidases) using the whole metatranscriptomic data set. Values represent the percentage of overlap for each KO pair computed as shared transcripts (i.e. transcripts annotated with both KOs) divided by the total transcripts of the KO displayed in the y-axis. Abbreviations: c (cysteine); s (serine); cp (carboxypeptidase).


Selected and not selected species in the different experiments
Jul17
Mar18
Nov18
Sep20
Number of transcripts
70
80
90
100
Ansanella granifera
Aureococcus anophagefferens
Bathycoccus prasinos
Cafeteria burkhardae
Chrysochromulina rotalis
Dicrateria rotunda
Emiliania huxleyi
Florenciella sp. RCC1007
Gephyrocapsa oceanica
Haptolina ericina
Incisomonas marina
Leptocylindrus hargravesii
Mataza sp. D1


Number of transcripts
70
80
90
100
Micromonas bravo
Micromonas commoda
Neobodo designis
Ostreococcus lucimarinus
Paraphysomonas imperforata
Pelagodinium beii
Pelagomonas calceolata
Phaeocystis cordata
Phaeocystis globosa
Pteridomonas danica
Pycnococcus provasolii
Rhizochromulina sp. CCMP1243
Siedleckia nematoides


Number of transcripts
Teleaulax amphioxeia
Tiarina fusa
Triparma eleuthera
Triparma laevis
Acanthoecidae sp. 10tr
ChrysophyceaeNA−sp1
Isochrysidales sp. CCMP1244
MAST−11−sp1
MAST−1C−sp1
MAST−1D−sp2
MAST−3C−sp2
MAST−4A−sp1
MAST−4B−sp1
70
80
90
100

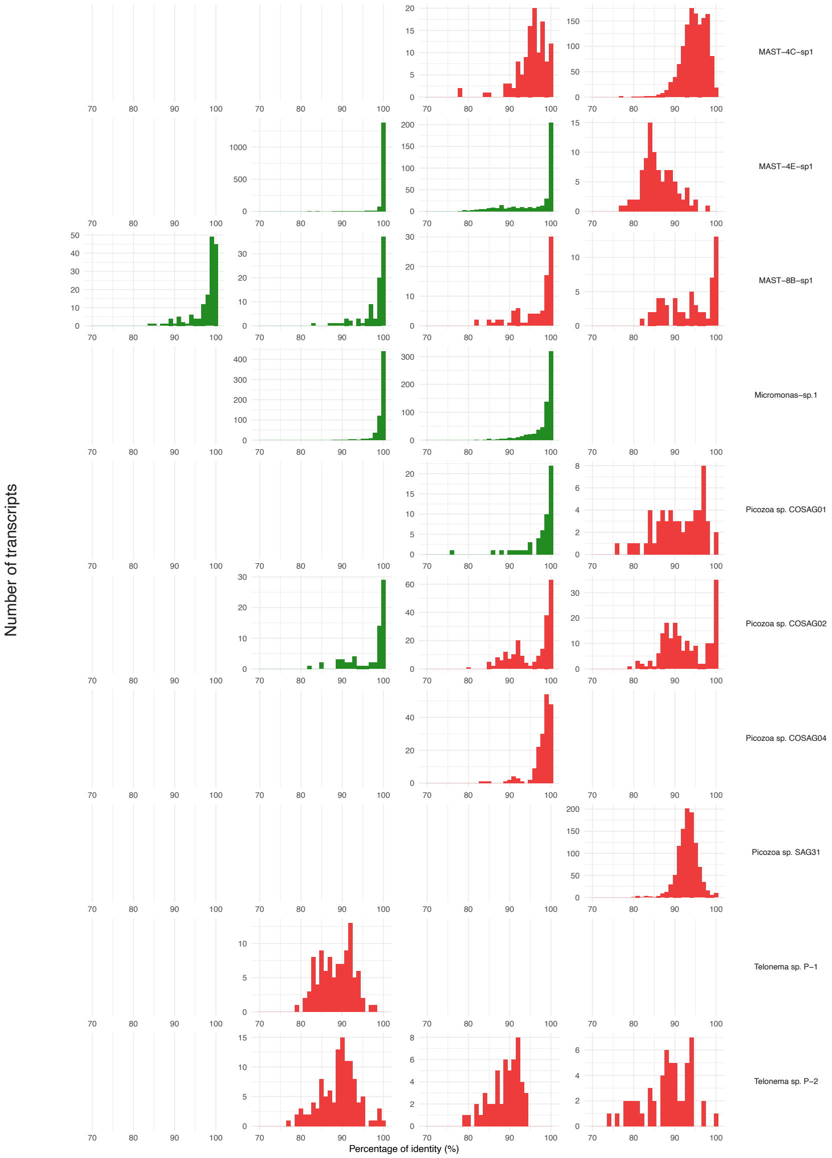
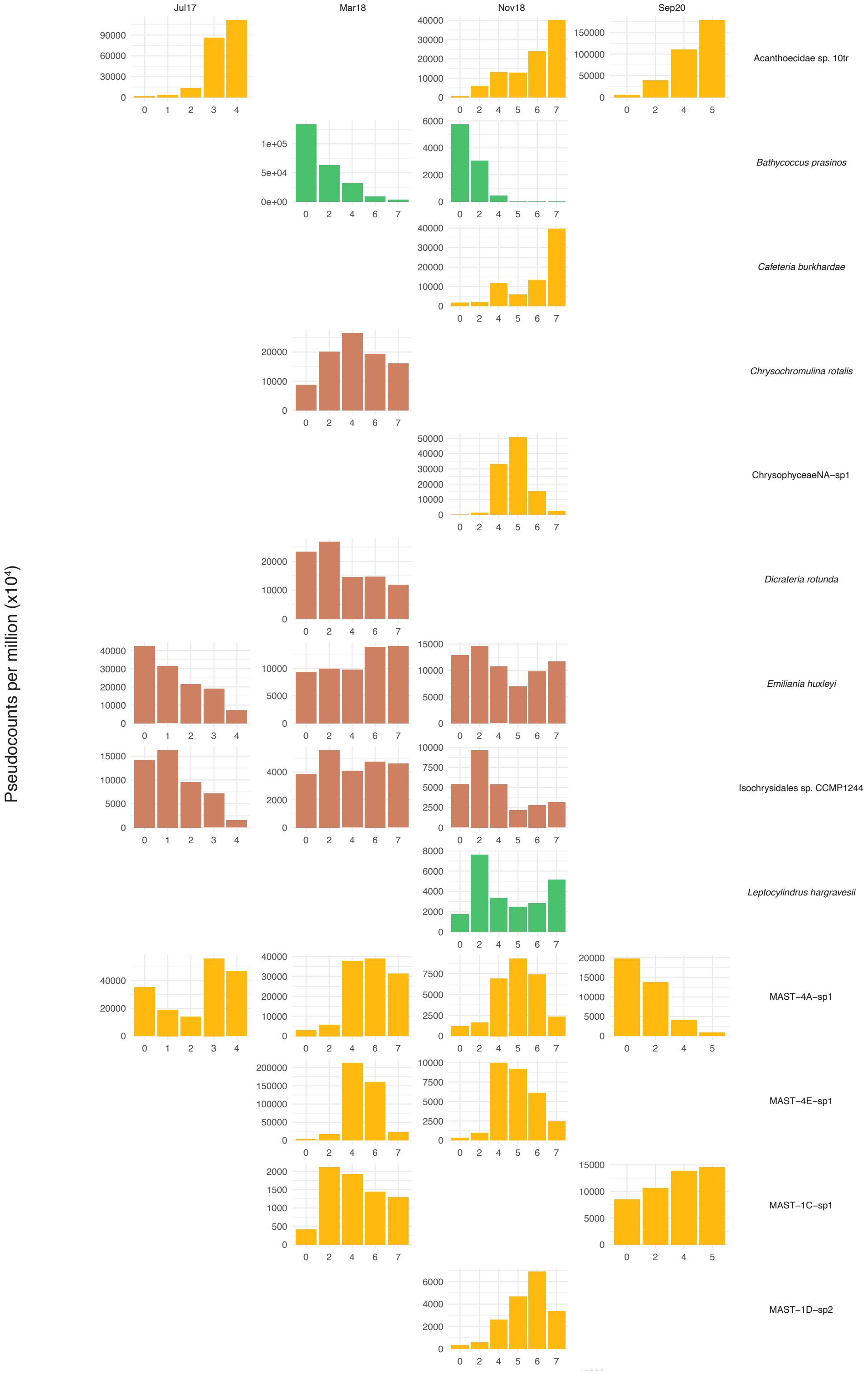


Figure S6. Nucleotide identity of transcripts associated to 51 represented species in the metatranscriptomes. We selected species having a median identity higher than 99% (green bars), while we discarded the rest (red bars).


Jul17
Mar18
Nov18
Sep20
Acanthoecidae sp. 10tr
Bathycoccus prasinos
Cafeteria burkhardae
Chrysochromulina rotalis
ChrysophyceaeNA−sp1
Dicrateria rotunda
Emiliania huxleyi
Isochrysidales sp. CCMP1244
Leptocylindrus hargravesii
MAST−4A−sp1
MAST−4E−sp1
MAST−1C−sp1
MAST−1D−sp2
Pseudocounts per million ($x10^4$)

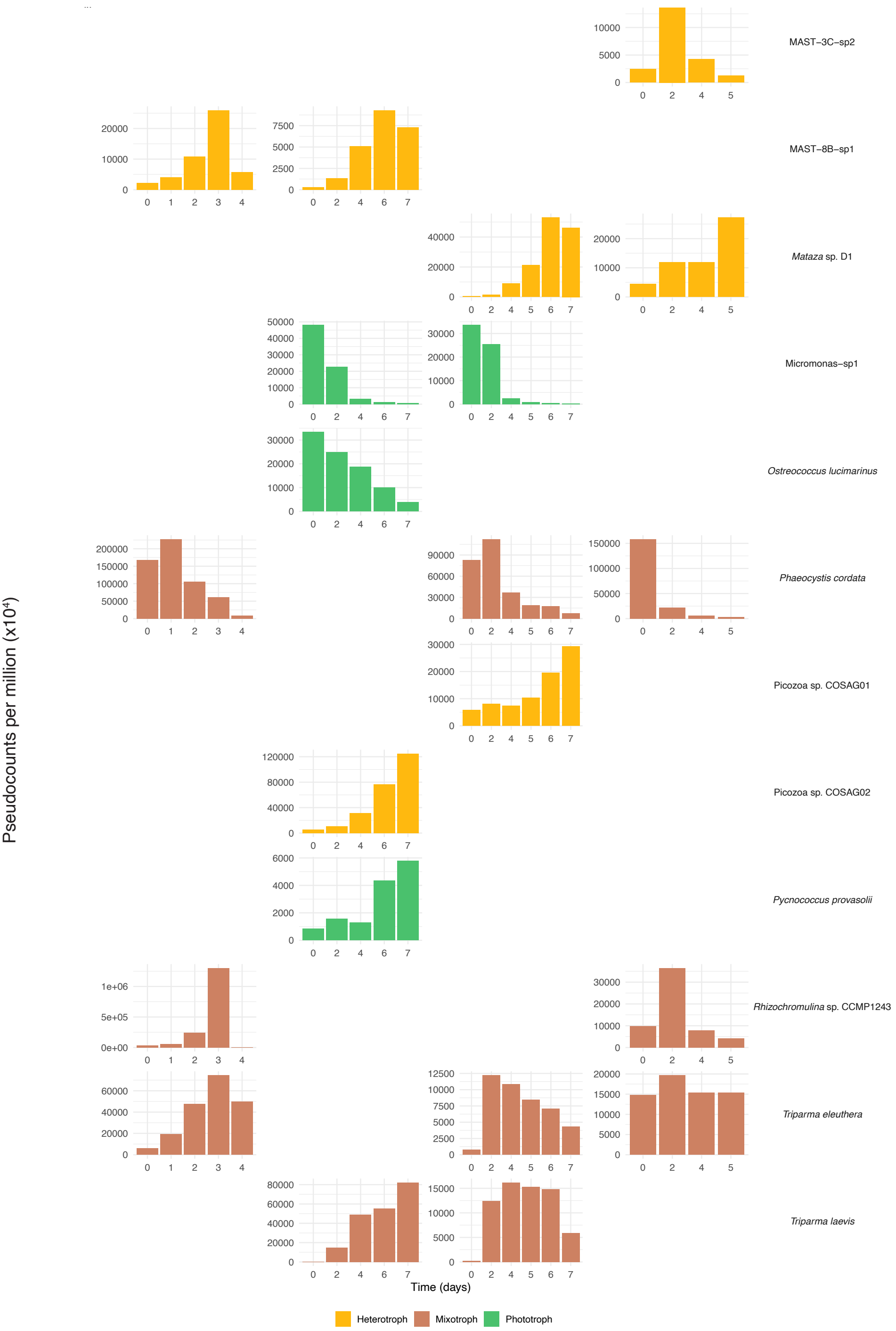


Figure S7. Expression dynamics of the 25 species with genomic data found in the metatranscriptomes. Abundance values represent pseudocounts per million, obtained after correcting the abundance profiles by gene lengths and sequencing depth (see Material and Methods for details). Note that some species appear in several experiments.

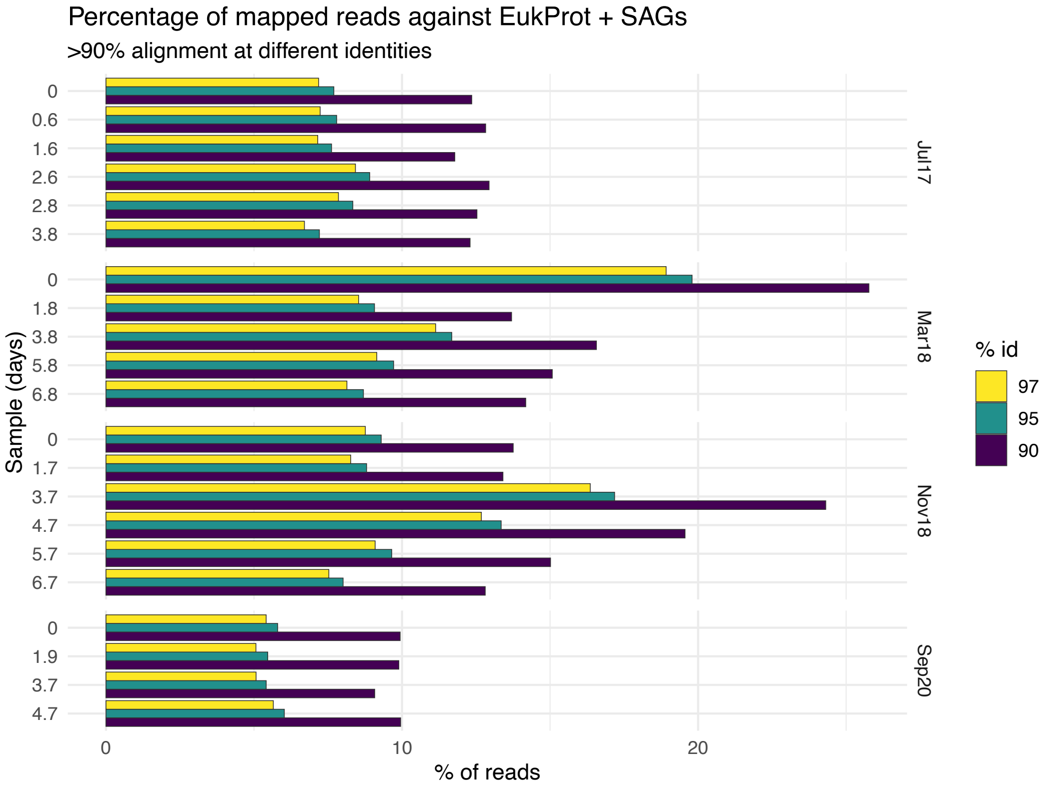


Figure S8. Per-sample summary of the mapping of the unassembled metatranscriptomic reads to the database EukProt+SAGs using DIAMOND blastx. Only alignments with >90% query coverage are considered.

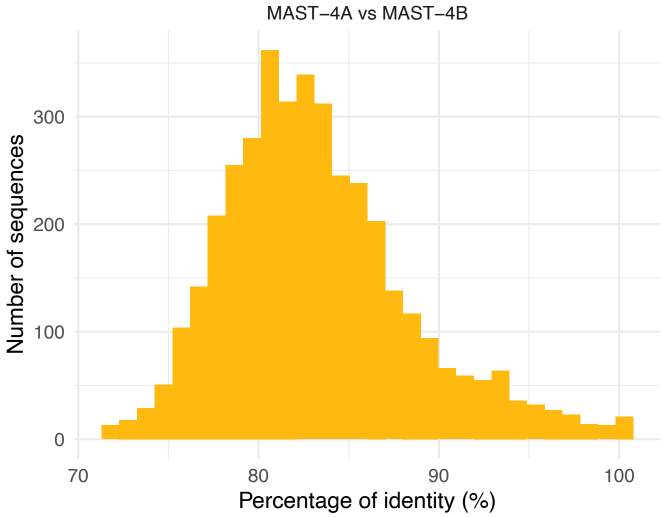


Figure S9. Histogram of the CDS identity between the genes of MAST-4A and MAST-4B species available in EukProt v3. Percentage identities were calculated using blastn.

3.9. Supplementary tables

Table S1. List of the 359 most expressed KEGG Orthologs (KO) in "growth" samples and general gene expression statistics associated.

KEGG ko	KO description	Main category	Mean TPM				Number of transcripts	House keeping
			Jul17	Mar18	Nov18	Sep20		
ko:K00933	creatine kinase	Amino acid metabolism	1308	1422	1660	678	652	No
ko:K00799	glutathione S-transferase	Amino acid metabolism	591	552	693	392	641	No
ko:K19199	histone-lysine N-methyltransferase SETD3	Amino acid metabolism	375	193	248	600	227	No
ko:K01802	peptidylprolyl isomerase	Amino acid metabolism	2571	2549	3268	1859	1833	No
ko:K08286	protein-serine/threonine kinase	Amino acid metabolism	247	69	101	573	275	No
ko:K00789	S-adenosylmethionine synthetase	Amino acid metabolism	2427	2005	1168	556	1244	Yes
ko:K01008	selenide, water dikinase	Amino acid metabolism	357	371	343	780	214	No
ko:K00474	trimethyllysine dioxygenase	Amino acid metabolism	59	161	850	46	95	No
ko:K00453	tryptophan 2,3-dioxygenase	Amino acid metabolism	229	684	1764	433	352	No
ko:K01674	carbonic anhydrase	Carbohydrate metabolism	78	47	12	514	62	No
ko:K01623	fructose-bisphosphate aldolase, class I	Carbohydrate metabolism	329	520	265	190	507	No
ko:K01624	fructose-bisphosphate aldolase, class II	Carbohydrate metabolism	526	1284	300	293	607	No
ko:K10046	GDP-D-mannose 3', 5'-epimerase	Carbohydrate metabolism	502	216	216	1047	247	No
ko:K00134	glyceraldehyde 3-phosphate dehydrogenase	Carbohydrate metabolism	2098	1987	621	675	1344	Yes
ko:K00015	glyoxylate reductase	Carbohydrate metabolism	65	139	839	43	74	No
ko:K00844	hexokinase	Carbohydrate metabolism	394	231	509	339	161	No
ko:K01000	phospho-N-acetylmuramoyl-pentapeptide-transferase	Carbohydrate metabolism	65	806	374	41	141	No
ko:K01807	ribose 5-phosphate isomerase A	Carbohydrate metabolism	308	898	2237	1406	307	No
ko:K00615	transketolase	Carbohydrate metabolism	1010	820	270	303	616	No
ko:K00700	1,4-alpha-glucan branching enzyme	CAZy enzymes	516	94	109	94	103	No
ko:K03927	carboxylesterase 2	CAZy enzymes	1509	1433	701	251	777	No
ko:K19357	cellulase	CAZy enzymes	39	57	36	537	162	No
ko:K01225	cellulose 1,4-beta-cellobiosidase	CAZy enzymes	135	108	177	638	441	No
ko:K01183	chitinase	CAZy enzymes	501	738	223	405	199	No
ko:K07151	dolichyl-diphosphooligosaccharide---protein glycosyltransferase	CAZy enzymes	1646	1430	2725	1905	780	No
ko:K12373	hexosaminidase	CAZy enzymes	438	229	218	513	366	No
ko:K20782	hydroxyproline O-arabinosyltransferase	CAZy enzymes	683	2189	101	173	109	No
ko:K00516	lytic starch monooxygenase	CAZy enzymes	1548	485	398	2184	1092	No
ko:K13379	reversibly glycosylated polypeptide / UDP-arabinopyranose mutase	CAZy enzymes	590	999	1063	1023	192	No
ko:K15920	xylan 1,4-beta-xylosidase	CAZy enzymes	710	143	107	554	357	No
ko:K16465	centrin-1	Chromosome and associated proteins	4153	14016	3608	1448	1788	Yes
ko:K10840	centrin-2	Chromosome and associated proteins	1214	6372	1103	173	355	Yes
ko:K16466	centrin-3	Chromosome and associated proteins	585	1435	504	379	433	Yes
ko:K11273	chromosome transmission fidelity protein 1	Chromosome and associated proteins	440	1765	1960	364	324	No
ko:K11275	histone H1/5	Chromosome and associated proteins	624	414	891	1349	209	No
ko:K11251	histone H2A	Chromosome and associated proteins	1289	886	1044	503	561	No
ko:K11252	histone H2B	Chromosome and associated proteins	511	387	816	304	293	No
ko:K11253	histone H3	Chromosome and associated proteins	1702	782	1079	1092	635	No
ko:K11254	histone H4	Chromosome and associated proteins	586	388	746	277	342	Yes
ko:K11493	regulator of chromosome condensation	Chromosome and associated proteins	160	1156	167	228	198	No
ko:K06674	structural maintenance of chromosome 2	Chromosome and associated proteins	682	1143	967	386	473	No
ko:K11649	SWI/SNF related-matrix-associated actin-dependent regulator of chromatin subfamily C	Chromosome and associated proteins	442	353	2389	1238	174	No
ko:K05692	actin beta/gamma 1	Cytoskeleton proteins	16051	11016	19032	13560	2503	Yes
ko:K10355	actin, other eukaryote	Cytoskeleton proteins	5345	6475	13184	5554	1425	Yes
ko:K10408	dynein heavy chain, axonemal	Cytoskeleton proteins	2073	2530	2160	2996	6302	Yes
ko:K10418	dynein light chain LC8-type	Cytoskeleton proteins	936	873	1636	585	712	No
ko:K04437	filamin	Cytoskeleton proteins	93	27	79	682	116	No
ko:K10394	kinesin family member 3A	Cytoskeleton proteins	477	1796	2020	395	407	No
ko:K07611	lamin B	Cytoskeleton proteins	150	903	1426	1141	158	No
ko:K06990	MEMO1 family protein	Cytoskeleton proteins	482	430	748	796	190	No
ko:K10388	plectin	Cytoskeleton proteins	56	865	1821	87	112	No
ko:K17338	receptor expression-enhancing protein 1/2/3/4	Cytoskeleton proteins	336	400	329	808	181	No
ko:K07374	tubulin alpha	Cytoskeleton proteins	26927	30408	39614	30915	4475	Yes
ko:K07375	tubulin beta	Cytoskeleton proteins	23822	30159	45703	24251	3962	Yes
ko:K08738	cytochrome c	Energy metabolism	849	1027	1384	6393	536	Yes
ko:K18561	FAD-dependent fumarate reductase	Energy metabolism	581	337	142	141	505	No
ko:K00026	malate dehydrogenase	Energy metabolism	699	533	325	530	559	No
ko:K03955	NADH dehydrogenase (ubiquinone) 1 alpha/beta subcomplex 1	Energy metabolism	223	338	907	146	297	No
ko:K00236	succinate dehydrogenase (ubiquinone) cytochrome b560 subunit	Energy metabolism	56	122	271	522	141	No
ko:K00413	ubiquinol-cytochrome c reductase cytochrome c1 subunit	Energy metabolism	281	357	350	630	286	No
ko:K10071	C-type lectin domain family 2 member B	Extracellular sensing and signalling	1383	0	0	1	3	No
ko:K10072	C-type lectin domain family 2 member D	Extracellular sensing and signalling	1383	0	0	1	3	No
ko:K06479	CD48 antigen	Extracellular sensing and signalling	59	139	838	38	70	No

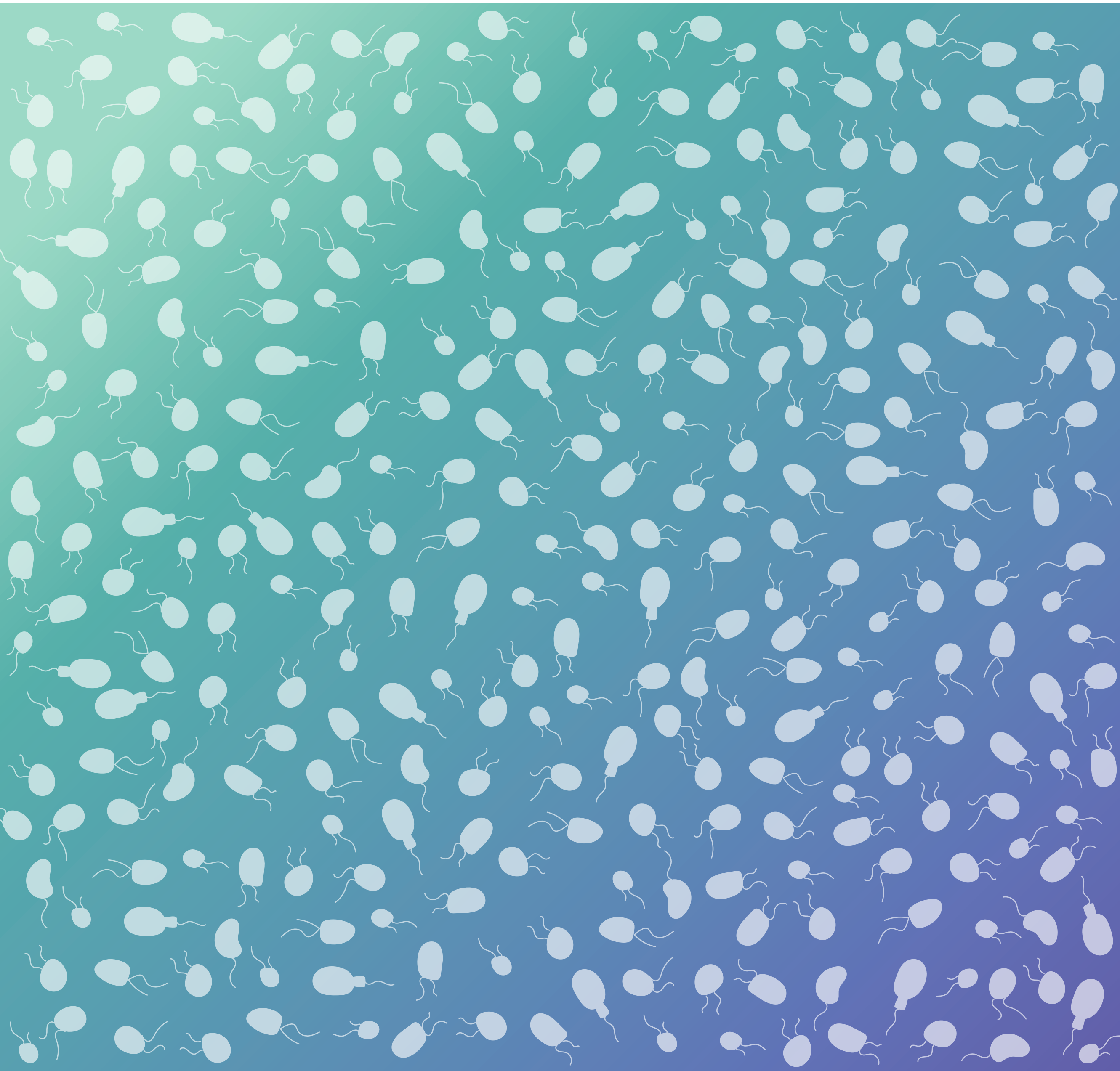
ko:K04000	complement component 9	Extracellular sensing and signalling	10	576	5	14	13	No
ko:K10104	ficolin	Extracellular sensing and signalling	832	22	20	114	23	No
ko:K06468	low affinity immunoglobulin epsilon Fc receptor	Extracellular sensing and signalling	1386	4	0	0	8	No
ko:K04324	Mas-related G-protein coupled receptor MRG	Extracellular sensing and signalling	304	872	574	215	192	No
ko:K17284	perilipin-2	Extracellular sensing and signalling	150	900	1423	1141	156	No
ko:K06572	plexin C	Extracellular sensing and signalling	23	1	8	624	39	No
ko:K16669	protocadherin Fat 4	Extracellular sensing and signalling	306	368	196	636	307	No
ko:K17086	transmembrane 9 superfamily member 2/4	Extracellular sensing and signalling	379	641	460	656	584	No
ko:K04832	acid-sensing ion channel 5	Ion channels	230	685	1766	442	357	No
ko:K04805	nicotinic acetylcholine receptor alpha-3	Ion channels	474	421	746	771	170	No
ko:K15040	voltage-dependent anion channel protein 2	Ion channels	287	235	554	195	181	No
ko:K08242	24-methylenesterol C-methyltransferase	Lipid metabolism	268	106	62	581	73	No
ko:K11262	acetyl-CoA carboxylase / biotin carboxylase 1	Lipid metabolism	529	317	315	582	753	No
ko:K21737	acyl-lipid Delta6-acetylenase / acyl-lipid (9-3)-desaturase	Lipid metabolism	239	156	326	604	259	No
ko:K15013	long-chain-fatty-acid--CoA ligase ACSBG	Lipid metabolism	352	329	260	619	382	No
ko:K05288	phosphatidylinositol glycan, class O	Lipid metabolism	377	519	1235	1226	262	No
ko:K07937	ADP-ribosylation factor 1	Membrane trafficking	1244	1117	2027	693	974	No
ko:K07939	ADP-ribosylation factor 4	Membrane trafficking	539	37	24	29	66	No
ko:K17593	ankyrin repeat domain-containing protein 42	Membrane trafficking	490	456	759	775	190	No
ko:K21442	ankyrin repeat domain-containing protein 54	Membrane trafficking	822	1426	1857	760	287	No
ko:K07977	Arf/Sar family, other	Membrane trafficking	1483	883	2054	659	810	No
ko:K21997	autophagy-related protein 19/34	Membrane trafficking	67	67	819	24	89	No
ko:K19933	calcium-dependent secretion activator	Membrane trafficking	226	379	378	507	140	No
ko:K05765	cofilin	Membrane trafficking	855	486	1024	217	328	No
ko:K17697	dedicator of cytokinesis protein 4	Membrane trafficking	412	339	2384	1228	150	No
ko:K17065	dynamin 1-like protein	Membrane trafficking	331	590	939	333	256	No
ko:K20365	endoplasmic reticulum-Golgi intermediate compartment protein 1	Membrane trafficking	365	654	1864	325	412	No
ko:K10949	ER lumen protein retaining receptor	Membrane trafficking	1056	441	311	471	488	No
ko:K08341	GABA(A) receptor-associated protein	Membrane trafficking	469	669	373	236	491	No
ko:K18061	myotubularin-related protein 5/13	Membrane trafficking	560	1286	1908	192	193	No
ko:K18739	protein bicaudal D	Membrane trafficking	605	1105	1079	1021	200	No
ko:K20068	RalBP1-associated Eps domain-containing protein	Membrane trafficking	326	326	237	1019	139	No
ko:K08267	rapamycin-insensitive companion of mTOR	Membrane trafficking	62	152	846	89	94	No
ko:K08053	Ras GTPase-activating protein 2	Membrane trafficking	540	1284	1888	139	177	No
ko:K04392	Ras-related C3 botulinum toxin substrate 1	Membrane trafficking	651	349	643	421	345	No
ko:K07861	Ras-related C3 botulinum toxin substrate 3	Membrane trafficking	507	26	84	31	108	No
ko:K07874	Ras-related protein Rab-1A	Membrane trafficking	663	610	752	261	629	No
ko:K07901	Ras-related protein Rab-8A	Membrane trafficking	241	296	388	704	318	No
ko:K20168	TBC1 domain family member 15	Membrane trafficking	106	162	626	103	163	No
ko:K20301	trafficking protein particle complex subunit 2	Membrane trafficking	66	220	753	74	103	No
ko:K20184	vacuolar protein sorting-associated protein 41	Membrane trafficking	915	633	2121	572	277	No
ko:K21248	vacuole membrane protein 1	Membrane trafficking	275	513	504	116	159	No
ko:K20241	WD repeat-containing protein 44	Membrane trafficking	159	121	420	774	133	No
ko:K01662	1-deoxy-D-xylulose-5-phosphate synthase	Metabolism of cofactors and vitamins	916	572	199	236	446	No
ko:K03644	lipoyl synthase	Metabolism of cofactors and vitamins	230	1176	2184	378	343	No
ko:K03403	magnesium chelatase subunit H	Metabolism of cofactors and vitamins	150	511	78	106	346	No
ko:K18914	adrenodoxin-NADP+ reductase	Mitochondrial proteins	298	665	779	384	193	No
ko:K01412	mitochondrial-processing peptidase subunit alpha	Mitochondrial proteins	554	1237	1006	667	405	No
ko:K17785	mitofilin	Mitochondrial proteins	185	90	111	502	127	No
ko:K13998	dihydrofolate reductase / thymidylate synthase	Nucleotide metabolism	312	527	1811	219	292	No
ko:K00940	nucleoside-diphosphate kinase	Nucleotide metabolism	491	444	499	627	417	No
ko:K00390	phosphoadenosine phosphosulfate reductase	Nucleotide metabolism	589	157	49	25	148	No
ko:K22390	acid phosphatase type 7	Other hydrolases	352	459	464	528	580	No
ko:K01054	acylglycerol lipase	Other hydrolases	425	349	2408	1239	202	No
ko:K01509	adenosinetriphosphatase	Other hydrolases	242	102	133	502	221	No
ko:K01251	adenosylhomocysteinase	Other hydrolases	2258	2872	1708	379	1243	No
ko:K07407	alpha-galactosidase	Other hydrolases	813	280	238	1372	608	No
ko:K01135	arylsulfatase B	Other hydrolases	345	64	30	679	233	No
ko:K12375	arylsulfatase I/J	Other hydrolases	441	70	43	810	277	No
ko:K01192	beta-mannosidase	Other hydrolases	662	887	326	716	674	No
ko:K01137	N-acetylglucosamine-6-sulfatase	Other hydrolases	448	101	97	1167	324	No
ko:K19882	O-palmitoleoyl-L-serine hydrolase	Other hydrolases	409	112	98	876	363	No
ko:K01074	palmitoyl-protein thioesterase	Other hydrolases	554	271	127	1365	236	No
ko:K13022	carboxypeptidase Z	Peptidases	158	900	1423	1141	158	No
ko:K13289	cathepsin A (carboxypeptidase C)	Peptidases	554	978	1244	1143	671	No
ko:K01363	cathepsin B	Peptidases	5117	5008	3065	10417	1983	No
ko:K01379	cathepsin D	Peptidases	2326	1479	620	1779	927	No
ko:K01382	cathepsin E	Peptidases	878	680	402	543	469	No
ko:K01373	cathepsin F	Peptidases	424	894	2121	888	708	No
ko:K01366	cathepsin H	Peptidases	4808	4277	2886	6276	1525	No
ko:K01371	cathepsin K	Peptidases	1592	901	852	11633	603	No
ko:K01365	cathepsin L	Peptidases	8231	7741	11669	21683	3422	No
ko:K01374	cathepsin O	Peptidases	230	685	1766	442	357	No
ko:K08568	cathepsin X	Peptidases	5857	5218	3771	7061	2026	No
ko:K16292	KDEL-tailed cysteine endopeptidase	Peptidases	59	286	1922	52	158	No
ko:K01369	legumain	Peptidases	209	362	242	1473	223	No
ko:K08245	phytepsin	Peptidases	981	164	28	482	105	No

ko:K01381	saccharopepsin	Peptidases	893	976	428	633	403	No
ko:K16296	serine carboxypeptidase-like clade I	Peptidases	551	678	983	1018	601	No
ko:K16297	serine carboxypeptidase-like clade II	Peptidases	268	406	654	632	284	No
ko:K01279	tripeptidyl-peptidase I	Peptidases	1719	1606	797	4960	1026	No
ko:K09645	vitellogenic carboxypeptidase-like protein	Peptidases	431	547	601	452	434	No
ko:K16290	xylem cysteine proteinase	Peptidases	1765	589	501	12417	523	No
ko:K02641	ferredoxin--NADP+ reductase	Photosynthesis proteins	641	1055	184	284	513	No
ko:K08907	light-harvesting complex I chlorophyll a/b binding protein 1	Photosynthesis proteins	968	622	152	184	764	No
ko:K08908	light-harvesting complex I chlorophyll a/b binding protein 2	Photosynthesis proteins	196	559	135	33	272	No
ko:K08912	light-harvesting complex II chlorophyll a/b binding protein 1	Photosynthesis proteins	173	806	53	55	314	No
ko:K08913	light-harvesting complex II chlorophyll a/b binding protein 2	Photosynthesis proteins	173	806	46	55	302	No
ko:K01602	ribulose-bisphosphate carboxylase small chain	Photosynthesis proteins	458	879	39	98	125	No
ko:K04532	amyloid beta precursor protein binding protein 1	Protein processing	741	682	747	1020	269	No
ko:K08057	calreticulin	Protein processing	2632	924	995	4606	778	No
ko:K10098	calreticulin 3	Protein processing	633	355	565	4010	339	No
ko:K12581	CCR4-NOT transcription complex subunit 7/8	Protein processing	447	163	186	903	207	No
ko:K03798	cell division protease FtsH	Protein processing	276	523	120	196	434	No
ko:K04077	chaperonin GroEL	Protein processing	344	547	444	180	389	No
ko:K17495	CUB and sushi domain-containing protein	Protein processing	524	330	238	478	390	No
ko:K17822	DCN1-like protein 1/2	Protein processing	502	462	831	786	238	No
ko:K09503	DnaJ homolog subfamily A member 2	Protein processing	968	935	1230	1045	587	No
ko:K22377	E3 ubiquitin-protein ligase listerin	Protein processing	629	837	1954	1400	175	Yes
ko:K15708	E3 ubiquitin-protein ligase RNF180	Protein processing	584	597	1002	271	186	Yes
ko:K09577	FK506-binding protein 14	Protein processing	122	312	178	619	126	No
ko:K03773	FKBP-type peptidyl-prolyl cis-trans isomerase FKB	Protein processing	312	267	521	356	270	No
ko:K03283	heat shock 70kDa protein 1/2/6/8	Protein processing	5440	5951	3447	2554	2810	No
ko:K09490	heat shock 70kDa protein 5	Protein processing	3633	5335	1437	3483	1592	No
ko:K09487	heat shock protein 90kDa beta	Protein processing	1434	1311	1170	670	765	Yes
ko:K04043	molecular chaperone DnaK	Protein processing	230	347	464	586	261	Yes
ko:K04079	molecular chaperone HtpG	Protein processing	4380	4325	3272	1474	2206	Yes
ko:K03626	nascent polypeptide-associated complex subunit alpha	Protein processing	291	948	554	213	335	No
ko:K01527	nascent polypeptide-associated complex subunit beta	Protein processing	550	960	1594	214	319	No
ko:K17987	next to BRCA1 gene 1 protein	Protein processing	561	246	924	97	272	No
ko:K03768	peptidyl-prolyl cis-trans isomerase B (cyclophilin B)	Protein processing	537	435	175	124	215	Yes
ko:K09565	peptidyl-prolyl isomerase F (cyclophilin D)	Protein processing	1025	690	615	736	419	No
ko:K13050	proprotein convertase subtilisin/kexin type 9	Protein processing	2	30	760	11	11	No
ko:K09580	protein disulfide-isomerase A1	Protein processing	1416	1054	1367	2933	695	No
ko:K09584	protein disulfide-isomerase A6	Protein processing	3580	2977	2255	2373	1096	No
ko:K10956	protein transport protein SEC61 subunit alpha	Protein processing	438	1032	530	619	408	No
ko:K03094	S-phase kinase-associated protein 1	Protein processing	606	379	343	206	394	No
ko:K10523	speckle-type POZ protein	Protein processing	334	543	1629	256	317	No
ko:K09499	T-complex protein 1 subunit eta	Protein processing	717	165	152	98	265	No
ko:K09500	T-complex protein 1 subunit theta	Protein processing	503	134	325	135	317	No
ko:K03671	thioredoxin 1	Protein processing	679	599	658	505	679	No
ko:K13984	thioredoxin domain-containing protein 5	Protein processing	2100	2634	1731	2156	857	No
ko:K13525	transitional endoplasmic reticulum ATPase	Protein processing	798	565	428	331	789	No
ko:K04551	ubiquitin B	Protein processing	3450	1761	4943	1164	467	Yes
ko:K08770	ubiquitin C	Protein processing	20950	9756	16680	14436	2543	Yes
ko:K06689	ubiquitin-conjugating enzyme E2 D	Protein processing	667	478	1550	325	468	No
ko:K12158	ubiquitin-like protein Nedd8	Protein processing	9572	871	545	3080	214	No
ko:K06630	14-3-3 protein epsilon	Replication and repair	984	941	1544	538	898	No
ko:K10886	DNA-repair protein XRCC4	Replication and repair	632	1154	1971	356	316	No
ko:K15076	elongin-A	Replication and repair	137	434	1461	81	156	No
ko:K04802	proliferating cell nuclear antigen	Replication and repair	403	553	355	739	338	No
ko:K10755	replication factor C subunit 2/4	Replication and repair	658	958	1579	601	275	No
ko:K11131	H/ACA ribonucleoprotein complex subunit 4	Ribosome	218	957	1931	157	248	No
ko:K20221	importin-4	Ribosome	292	538	985	266	150	No
ko:K02865	large subunit ribosomal protein L10Ae	Ribosome	1057	1937	1437	1374	508	Yes
ko:K02866	large subunit ribosomal protein L10e	Ribosome	2076	2303	3232	1382	731	Yes
ko:K02868	large subunit ribosomal protein L11e	Ribosome	715	1504	2269	425	556	Yes
ko:K02870	large subunit ribosomal protein L12e	Ribosome	1117	1305	1722	527	523	Yes
ko:K02872	large subunit ribosomal protein L13Ae	Ribosome	1254	2137	1984	523	537	Yes
ko:K02873	large subunit ribosomal protein L13e	Ribosome	1820	1990	2391	1014	562	Yes
ko:K02875	large subunit ribosomal protein L14e	Ribosome	751	1018	1649	898	428	Yes
ko:K02877	large subunit ribosomal protein L15e	Ribosome	1371	2046	2624	1083	588	Yes
ko:K02880	large subunit ribosomal protein L17e	Ribosome	984	925	2255	293	475	Yes
ko:K02882	large subunit ribosomal protein L18Ae	Ribosome	622	1556	1121	1165	437	Yes
ko:K02883	large subunit ribosomal protein L18e	Ribosome	1145	1922	2252	1211	622	Yes
ko:K02885	large subunit ribosomal protein L19e	Ribosome	1533	1893	2243	888	604	Yes
ko:K02889	large subunit ribosomal protein L21e	Ribosome	906	1954	2553	749	547	Yes
ko:K02891	large subunit ribosomal protein L22e	Ribosome	1333	1148	1766	728	423	Yes
ko:K02893	large subunit ribosomal protein L23Ae	Ribosome	1020	1442	497	247	367	Yes
ko:K02894	large subunit ribosomal protein L23e	Ribosome	655	1685	2078	1220	558	Yes
ko:K02896	large subunit ribosomal protein L24e	Ribosome	1022	1012	586	547	514	Yes

ko:K02898	large subunit ribosomal protein L26e	Ribosome	1281	1251	959	1918	537	Yes
ko:K02900	large subunit ribosomal protein L27Ae	Ribosome	1234	2494	2616	943	630	Yes
ko:K02901	large subunit ribosomal protein L27e	Ribosome	1165	1565	1624	1309	490	Yes
ko:K02903	large subunit ribosomal protein L28e	Ribosome	571	1536	1712	623	304	Yes
ko:K02905	large subunit ribosomal protein L29e	Ribosome	516	436	247	1134	239	Yes
ko:K02908	large subunit ribosomal protein L30e	Ribosome	1252	1590	3067	434	511	Yes
ko:K02910	large subunit ribosomal protein L31e	Ribosome	861	1073	695	680	506	Yes
ko:K02912	large subunit ribosomal protein L32e	Ribosome	1251	1174	1057	1365	516	Yes
ko:K02915	large subunit ribosomal protein L34e	Ribosome	1474	1622	2617	573	533	Yes
ko:K02917	large subunit ribosomal protein L35Ae	Ribosome	1149	1377	1086	798	455	Yes
ko:K02918	large subunit ribosomal protein L35e	Ribosome	1323	1836	2446	730	474	Yes
ko:K02920	large subunit ribosomal protein L36e	Ribosome	1204	944	1471	439	479	Yes
ko:K02921	large subunit ribosomal protein L37Ae	Ribosome	777	987	867	1094	453	Yes
ko:K02922	large subunit ribosomal protein L37e	Ribosome	829	1331	1181	626	418	Yes
ko:K02923	large subunit ribosomal protein L38e	Ribosome	804	1219	1233	340	342	Yes
ko:K02925	large subunit ribosomal protein L3e	Ribosome	1202	2074	1077	444	666	Yes
ko:K02927	large subunit ribosomal protein L40e	Ribosome	2661	2603	4325	1881	730	Yes
ko:K02929	large subunit ribosomal protein L44e	Ribosome	1228	1464	1768	1160	476	Yes
ko:K02930	large subunit ribosomal protein L4e	Ribosome	1176	1707	1152	398	586	Yes
ko:K02932	large subunit ribosomal protein L5e	Ribosome	1306	2159	2508	378	625	Yes
ko:K02934	large subunit ribosomal protein L6e	Ribosome	1033	1442	2339	864	476	Yes
ko:K02936	large subunit ribosomal protein L7Ae	Ribosome	991	2028	2901	1590	643	Yes
ko:K02937	large subunit ribosomal protein L7e	Ribosome	1245	1664	721	1462	556	Yes
ko:K02938	large subunit ribosomal protein L8e	Ribosome	1727	3294	4262	1361	759	Yes
ko:K02940	large subunit ribosomal protein L9e	Ribosome	1270	1758	1366	1389	589	Yes
ko:K02941	large subunit ribosomal protein LP0	Ribosome	963	1483	2104	435	537	Yes
ko:K02942	large subunit ribosomal protein LP1	Ribosome	881	1336	1536	552	596	Yes
ko:K02943	large subunit ribosomal protein LP2	Ribosome	705	1728	2219	1041	655	Yes
ko:K14558	periodic tryptophan protein 2	Ribosome	58	59	60	556	79	No
ko:K14855	ribosome assembly protein 4	Ribosome	95	56	74	654	78	No
ko:K14842	ribosome biogenesis protein NSA2	Ribosome	462	1376	252	69	212	Yes
ko:K02946	small subunit ribosomal protein S10	Ribosome	134	1117	144	91	140	Yes
ko:K02947	small subunit ribosomal protein S10e	Ribosome	679	1731	2105	194	407	Yes
ko:K02949	small subunit ribosomal protein S11e	Ribosome	1739	2219	2012	749	510	Yes
ko:K02951	small subunit ribosomal protein S12e	Ribosome	1075	1180	1617	1193	479	Yes
ko:K02953	small subunit ribosomal protein S13e	Ribosome	1832	1754	1107	1128	510	Yes
ko:K02955	small subunit ribosomal protein S14e	Ribosome	964	1510	1490	720	618	Yes
ko:K02957	small subunit ribosomal protein S15Ae	Ribosome	910	1588	1013	764	476	Yes
ko:K02958	small subunit ribosomal protein S15e	Ribosome	1044	1756	2466	348	571	Yes
ko:K02960	small subunit ribosomal protein S16e	Ribosome	614	1122	1422	1231	479	Yes
ko:K02962	small subunit ribosomal protein S17e	Ribosome	1056	1458	1202	452	489	Yes
ko:K02964	small subunit ribosomal protein S18e	Ribosome	998	2304	1580	1925	578	Yes
ko:K02966	small subunit ribosomal protein S19e	Ribosome	777	1477	2002	1161	446	Yes
ko:K02969	small subunit ribosomal protein S20e	Ribosome	807	2014	1405	631	453	Yes
ko:K02971	small subunit ribosomal protein S21e	Ribosome	704	892	1272	290	450	Yes
ko:K02973	small subunit ribosomal protein S23e	Ribosome	1346	1892	1693	1647	636	Yes
ko:K02974	small subunit ribosomal protein S24e	Ribosome	1470	1647	2345	1049	576	Yes
ko:K02975	small subunit ribosomal protein S25e	Ribosome	984	1542	2169	404	540	Yes
ko:K02976	small subunit ribosomal protein S26e	Ribosome	1925	1512	1348	1189	526	Yes
ko:K02977	small subunit ribosomal protein S27Ae	Ribosome	12182	5158	4875	9245	1172	Yes
ko:K02978	small subunit ribosomal protein S27e	Ribosome	1011	1725	2098	348	444	Yes
ko:K02979	small subunit ribosomal protein S28e	Ribosome	741	1022	776	626	440	Yes
ko:K02980	small subunit ribosomal protein S29e	Ribosome	213	611	323	201	184	Yes
ko:K02981	small subunit ribosomal protein S2e	Ribosome	918	1962	2685	331	701	Yes
ko:K02983	small subunit ribosomal protein S30e	Ribosome	794	917	1312	464	349	Yes
ko:K02984	small subunit ribosomal protein S3Ae	Ribosome	1110	3248	2892	1836	703	Yes
ko:K02985	small subunit ribosomal protein S3e	Ribosome	947	1249	1588	974	530	Yes
ko:K02987	small subunit ribosomal protein S4e	Ribosome	2253	1844	2463	2543	761	Yes
ko:K02989	small subunit ribosomal protein S5e	Ribosome	1041	2169	1460	966	620	Yes
ko:K02991	small subunit ribosomal protein S6e	Ribosome	1040	1927	2017	1550	594	Yes
ko:K02993	small subunit ribosomal protein S7e	Ribosome	1469	1757	2528	432	501	Yes
ko:K02995	small subunit ribosomal protein S8e	Ribosome	1444	2037	1403	642	653	Yes
ko:K02996	small subunit ribosomal protein S9	Ribosome	197	430	915	1144	177	Yes
ko:K02997	small subunit ribosomal protein S9e	Ribosome	993	2044	2342	904	595	Yes
ko:K02998	small subunit ribosomal protein SAe	Ribosome	1046	1955	2064	803	586	Yes
ko:K14566	U3 small nucleolar RNA-associated protein 24	Ribosome	341	422	730	900	304	No
ko:K00428	cytochrome c peroxidase	ROS homeostasis	275	322	587	903	384	No
ko:K17609	nucleoredoxin	ROS homeostasis	451	590	622	281	729	No
ko:K07305	peptide-methionine (R)-S-oxide reductase	ROS homeostasis	179	594	69	92	182	No
ko:K08272	calcium binding protein 39	Signal transduction	488	358	2404	1283	190	No
ko:K13412	calcium-dependent protein kinase	Signal transduction	502	441	493	477	855	No
ko:K04515	calcium/calmodulin-dependent protein kinase (CaM kinase) II	Signal transduction	292	56	266	596	266	No
ko:K08794	calcium/calmodulin-dependent protein kinase I	Signal transduction	236	216	532	323	523	No
ko:K02183	calmodulin	Signal transduction	5375	4845	6544	3846	2794	Yes
ko:K04739	cAMP-dependent protein kinase regulator	Signal transduction	560	194	450	141	516	No
ko:K07376	cGMP-dependent protein kinase 1	Signal transduction	405	656	796	627	809	No
ko:K08819	cyclin-dependent kinase 12/13	Signal transduction	573	587	441	1266	269	Yes
ko:K00927	phosphoglycerate kinase	Signal transduction	643	978	293	196	649	No
ko:K04345	protein kinase A	Signal transduction	857	906	2396	1228	826	No

ko:K18449	prune homolog 2	Signal transduction	337	327	251	1016	147	No
ko:K14498	serine/threonine-protein kinase SRK2	Signal transduction	150	902	1423	1145	170	No
ko:K14807	ATP-dependent RNA helicase DDX51/DBP6	Transcription and RNA processing	321	892	585	241	226	No
ko:K14780	ATP-dependent RNA helicase DHX37/DHR1	Transcription and RNA processing	459	225	521	334	168	No
ko:K17679	ATP-dependent RNA helicase MSS116, mitochondrial	Transcription and RNA processing	46	71	595	34	72	No
ko:K17675	ATP-dependent RNA helicase SUPV3L1/SUV3	Transcription and RNA processing	502	717	1294	679	198	No
ko:K12812	ATP-dependent RNA helicase UAP56/SUB2	Transcription and RNA processing	430	567	710	198	485	No
ko:K20099	ATP-dependent RNA helicase YTHDC2	Transcription and RNA processing	316	878	577	216	203	No
ko:K03010	DNA-directed RNA polymerase II subunit RPB2	Transcription and RNA processing	452	508	622	310	319	No
ko:K11592	endoribonuclease Dicer	Transcription and RNA processing	796	682	2026	200	204	No
ko:K14753	guanine nucleotide-binding protein subunit beta-2-like 1 protein	Transcription and RNA processing	1211	2681	2210	1683	661	No
ko:K09228	KRAB domain-containing zinc finger protein	Transcription and RNA processing	324	1126	2775	446	328	No
ko:K05925	mRNA m6A methyltransferase	Transcription and RNA processing	63	63	768	18	71	No
ko:K09329	paired mesoderm homeobox protein	Transcription and RNA processing	474	421	746	771	168	No
ko:K13199	plasminogen activator inhibitor 1 RNA-binding protein	Transcription and RNA processing	1790	725	641	411	218	No
ko:K13126	polyadenylate-binding protein	Transcription and RNA processing	543	955	557	319	521	No
ko:K15542	polyadenylation factor subunit 2	Transcription and RNA processing	422	646	2047	991	266	No
ko:K12867	pre-mRNA-splicing factor SYF1	Transcription and RNA processing	261	499	514	570	225	No
ko:K18749	protein LSM14	Transcription and RNA processing	212	1436	1875	768	312	No
ko:K15219	RNA polymerase I-specific transcription initiation factor RRN7	Transcription and RNA processing	992	356	125	72	105	Yes
ko:K06269	serine/threonine-protein phosphatase PP1 catalytic subunit	Transcription and RNA processing	716	665	551	455	599	No
ko:K09422	transcription factor MYB, plant	Transcription and RNA processing	206	1464	1969	355	293	No
ko:K03124	transcription initiation factor TFIIB	Transcription and RNA processing	399	710	1130	1465	249	No
ko:K03132	transcription initiation factor TFIID subunit 7	Transcription and RNA processing	491	424	754	790	178	No
ko:K01872	alanyl-tRNA synthetase	Translation	286	227	537	215	286	No
ko:K06927	diphthine-ammonia ligase	Translation	349	555	724	6212	217	No
ko:K03231	elongation factor 1-alpha	Translation	6392	10694	17186	5220	2143	Yes
ko:K03232	elongation factor 1-beta	Translation	302	539	559	212	221	No
ko:K03233	elongation factor 1-gamma	Translation	235	201	533	156	162	No
ko:K03234	elongation factor 2	Translation	1340	2019	1397	746	1004	No
ko:K03235	elongation factor 3	Translation	1417	1633	1142	543	1400	No
ko:K07936	GTP-binding nuclear protein Ran	Translation	343	434	502	181	323	No
ko:K01875	seryl-tRNA synthetase	Translation	282	448	567	230	283	No
ko:K01868	threonyl-tRNA synthetase	Translation	1136	1374	953	483	692	No
ko:K03236	translation initiation factor 1A	Translation	246	124	1265	112	200	No
ko:K03257	translation initiation factor 4A	Translation	730	1050	1010	377	598	No
ko:K03263	translation initiation factor 5A	Translation	853	1425	1316	562	562	No
ko:K01866	tyrosyl-tRNA synthetase	Translation	267	465	548	307	311	No
ko:K01873	valyl-tRNA synthetase	Translation	385	595	1013	351	399	No
ko:K05850	Ca2+ transporting ATPase, plasma membrane	Translocases	222	322	285	890	350	No
ko:K01537	Ca2+-transporting ATPase	Translocases	413	385	331	1031	636	No
ko:K02132	F-type H+-transporting ATPase subunit alpha	Translocases	650	760	917	214	361	No
ko:K02133	F-type H+-transporting ATPase subunit beta	Translocases	1288	1163	3020	709	420	No
ko:K02134	F-type H+-transporting ATPase subunit delta	Translocases	128	361	578	457	178	No
ko:K02114	F-type H+-transporting ATPase subunit epsilon	Translocases	548	190	67	475	132	No
ko:K02136	F-type H+-transporting ATPase subunit gamma	Translocases	146	300	547	134	190	No
ko:K02112	F-type H+/Na+-transporting ATPase subunit beta	Translocases	581	294	105	472	160	No
ko:K00323	H+-translocating NAD(P) transhydrogenase	Translocases	249	456	530	176	394	No
ko:K01535	H+-transporting ATPase	Translocases	390	529	331	586	385	No
ko:K01541	H+/K+-exchanging ATPase	Translocases	781	283	136	1747	197	No
ko:K01542	H+/K+-exchanging ATPase alpha polypeptide	Translocases	780	283	136	1743	192	No
ko:K01507	inorganic pyrophosphatase	Translocases	1287	2445	3098	1136	1592	No
ko:K01544	non-gastric H+/K+-exchanging ATPase	Translocases	780	292	136	1747	200	No
ko:K01530	phospholipid-translocating ATPase	Translocases	285	119	277	599	392	No
ko:K14802	phospholipid-transporting ATPase	Translocases	346	191	278	609	430	No
ko:K01539	sodium/potassium-transporting ATPase subunit alpha	Translocases	880	568	462	1911	330	No
ko:K02155	V-type H+-transporting ATPase 16kDa proteolipid subunit	Translocases	1496	1549	2617	1394	747	No
ko:K02154	V-type H+-transporting ATPase subunit a	Translocases	417	303	631	515	429	No
ko:K02147	V-type H+-transporting ATPase subunit B	Translocases	662	220	321	66	312	No
ko:K03320	ammonium transporter, Amt family	Transporters	657	307	77	160	467	No
ko:K13749	solute carrier family 24 (sodium/potassium/calcium exchanger), member 1	Transporters	241	341	124	702	251	No
ko:K13750	solute carrier family 24 (sodium/potassium/calcium exchanger), member 2	Transporters	245	345	139	702	271	No
ko:K05863	solute carrier family 25 (mitochondrial adenine nucleotide translocator), member 4/5/6/31	Transporters	1828	4924	2097	1141	798	No
ko:K15102	solute carrier family 25 (mitochondrial phosphate transporter), member 3	Transporters	516	701	603	276	393	No
ko:K15276	solute carrier family 35 (adenosine 3'-phospho 5'-phosphosulfate transporter), member B2	Transporters	1078	1575	2370	1019	407	No
ko:K15377	solute carrier family 44 (choline transporter-like protein), member 2/4/5	Transporters	247	502	538	497	393	No

GENERAL DISCUSSION



General discussion

It was 40 years ago when Fenchel published a series of studies conducted on HF isolates where he proposed that these microorganisms were responsible for maintaining bacterioplankton abundances in check (Fenchel, 1982). This represented a turning point for the consideration of the ecological significance of HFs in the ecosystem, and subsequent studies targeting these microbes revealed their massive functional and taxonomic diversity (Jürgens & Massana, 2008). However, studying HF assemblages presents several challenges, and this initial interest in assessing them diminished, to the extent of making these microorganisms one of the most understudied components of the marine microbiome. This thesis represents a return to the study of marine HFs, and highlights again their key role in the ocean. Overall, we aimed at improving our understanding of the biogeography, functional role and ecology of HFs in the ocean. To do so, we combined several molecular approaches, namely metagenomics (Chapter 1), metabarcoding (Chapter 1 and Chapter 2) and metatranscriptomics (Chapter 3), as well as epifluorescence microscopy (Chapter 3). Two of the studies consisted on re-analyses of already published data sets (part of Chapter 1 and the whole Chapter 2), and the other was obtained during the development of this thesis (Chapter 3). In Chapters 1 and 2, the scale of the data was global, with a focus on horizontal and vertical distributions of marine eukaryotes and HFs in the open ocean, while in Chapter 3 the scale was local (the BBMO station), with a focus on temporal dynamics of gene expression in slightly manipulated natural assemblages. So, we started studying the overall picture of diversity and biogeography of HFs and we ended investigating some of their specific biological processes.

Main results and the importance of the technical approach

In Chapter 1 we investigated microbial eukaryotic diversity in the water column of the ocean by means of metagenomics (metaG), an approach that served two main purposes. First, we could assess this diversity without the need of using PCR amplification of a marker gene with specific primers; and second, we could compare the results with the ones obtained by metabarcoding. Thanks to this amplification-free approach, we detected taxonomic groups that were neglected in metabarcoding surveys due to technical biases, such as Diplonemea, Kinetoplastida or Amoebozoa, while the majority of groups displayed a similar relative read abundance profile regardless of the sequencing approach. We identified that the community structure was clearly influenced by depth region (photic or aphotic) and by the cell size (pico- or nanoplanktonic fraction), as most microbial eukaryotic groups were preferentially present in one of the four niches created by these two factors. Despite being extremely informative, this metaG approach lacks taxonomic resolution given that it relies on 100 bp long fragments, and consequently we could only report the structure of protist communities at the group (generally a formal Class) level. Thus, in order to go deeper into the biogeography of HF species, this is not a valid approach. Given that the analysis in Chapter 1 revealed that most of the taxonomic groups containing HFs were not highly affected by putative biases of PCR amplification, we embraced the use of metabarcoding in Chapter 2.

In Chapter 2 we determined the most abundant species of HF in the ocean, together with their patterns of distribution. We did so by using global metabarcoding data sets released by both Malaspina and TARA oceanographic surveys. We detected a core of HF taxonomic groups in the photic layer of the ocean with rather constant abundances at the horizontal scale, comprised by MAST clades (mostly -1, -3 and -4), Chrysophyceae (uncultured clades G, H, and I) and Picozoa. At the vertical scale, we identified marked changes in HF diversity between depth zones. The core groups of the photic zone were poorly detected in the deep ocean, which was mainly populated by other Chrysophyceae, Bicosoecida and Diplonemea groups. In Chapter 2, the approach used allowed us going beyond the group level and assess the distribution and abundance of HF species, being considered as individuals having identical V4 region of the 18S rRNA gene. Indeed, we identified species differing by only 1 bp in this marker region that followed completely contrasted ecological distributions. Thanks to this we obtained a list of 52 dominant species in the photic zone of the ocean, the majority of which still remain uncultured. As already discussed in Chapter 2, we cannot omit that microbiome data yielded by metabarcoding are compositional (Gloor et al., 2017), meaning that they quantitatively represent a portion of a some whole, thus always yielding relative rather than absolute information. Although in our study we tried to minimize this compositional effect by, when possible, using software developed to this aim, we also used relative abundances to determine which species where the most abundant in the environment. We argued that the obtained results are a good approximation given that HF tend to harbor a low 18S rDNA copy number (Zhu et al., 2005), but it would be extremely useful to perform a similar study using FISH microscopy to validate our results. We already did this with three MAST lineages using automated microscopy (Mangot et al., 2018), and a similar analysis could be performed with new probes targeting the most abundant taxa we report here. In fact, little effort has been done so far to target single HF species by FISH (Piwosz et al., 2021), mainly because the knowledge on which specific taxa are the most abundant in the ocean was still missing. Apart from this, the use of already established long-read sequencing techniques such as Nanopore or PacBio (Jamy et al., 2020) coupled with our short reads, could increase even more the taxonomic resolution at which HF can be studied.

With the information retrieved in the first 2 chapters, we obtained a clearer picture on HF biogeography as well as a list of HF taxa that are putatively the most ecologically relevant in marine ecosystems. With this, we could jump from analyzing HF communities' taxonomic diversity and distribution to investigating their activity. Thus, in Chapter 3 we analyzed the patterns of gene expression of HF assemblages during bacterivory, the process by which these microorganisms feed and a key ecological function in global biogeochemical cycles. Ideally, the best approach to study this process would have been to perform independent transcriptomic experiments for each of the 52 abundant HF species, followed by differential expression analyses as in Massana et al. (2021). Additionally, ecophysiology studies could be carried out to assess other aspects of the ecology of these protists. However, the majority of the 52 species are not available in culture, so we had to use a different approach. This consisted in promoting bacterivory in slightly modified natural communities of HFs using unamended incubations of surface seawater. In these incubations, the initial concentration of HF cells was increased by approximately 10-fold, thus yielding a signal strong enough to perform metatranscriptomic (metaT) sequencing of samples collected at different time points. This approach nonetheless involved some challenges. First of all, it was difficult to define clear states during the incubations, as there were hundreds of species growing in each incubation with a potentially different pattern each; second, with this complexity we had to think of different ways than standard RNA-seq experiments, which usually comprise separated treatment and control samples to determine overexpressed genes during bacterivory; and third, we could only work with part of the metatranscriptomes as more than half of the transcripts did not have any functional annotation. Considering all these limitations, we contrived to obtain remarkable results on the gene expression of HFs. We could divide samples into three states according to microscopic counts and this separation was supported by metatranscriptomic data. We assembled metatranscriptomes to assess the general functional dynamics of HF assemblages during the experiments. Gene expression patterns were rather similar between experiments, with a high expression of genes related to constitutive cellular processes, like actin, tubulin or ribosomal proteins. We then obtained a list of candidate genes highly expressed during bacterivory that could be key in this biological process, which were functionally annotated as peptidases, translocases and CAZy enzymes. We finally took advantage of public databases to extract the signal of specific species in the incubations, and compared the relative expression of these bacterivory-related genes between species with different trophic modes. Although the amount of data retrieved using this approach was rather low compared to the overall data set, we could identify groups of genes more expressed in heterotrophic taxa than mixotrophic or phototrophic ones. These were related to cysteine peptidases, as well as some glycoside hydrolases (GH3 and GH20), and represent clear targets to be used in future explorations.

Heterotrophic flagellates: advances and future work

With this thesis, we make a step forward to better understand the ecology of HFs by building a solid foundation on which to perform renewed research on the functional role of this functional group. Our results in Chapter 1 give a different view of the community structure of pico- and nanoeukaryotes in the water column, and highlight the biases that should be considered by researchers working with metabarcoding of some eukaryotic groups. Chapter 2, together with Chapter 1, provides a clear picture on the distribution of HFs in the ocean, specifically in the tropical and subtropical zones. This kind of study was already performed with several other eukaryotic groups (see Chapter 2) but was still missing for HFs. With it, we present a guide that can be used for further exploration on their biogeography. Recently, some community efforts have been conducted to compile the most comprehensive data set of global eukaryotic metabarcoding surveys to date, which have crystallized into tools like meta-pr2 (Vaulot et al., 2022) or the EukBank project (Berney et al., 2017). A clear next step on HF biogeography research would be to gather the results obtained in this thesis and extend them with all the possible data available. As mentioned in the previous section, it is now time to use FISH microscopy to target the most abundant but yet uncultured taxa in the ocean, a challenge that could yield key information on the ecology of these microorganisms. Additionally, we need to bear in mind that samples in these biogeography studies tend to represent a single point in time, therefore neglecting the temporal dimension. Thus, besides maximizing the breadth of sampling, it is also needed to add the temporal scale to the spatial results. To the best of our knowledge, this work has not yet been done for HFs, despite several data sets being already available, such as the BBMO or SOLA time series in the Mediterranean sea (Giner et al., 2019; Lambert et al., 2018) or SPOT in the Pacific Ocean (Yeh & Fuhrman, 2022), just to name a few.

The catalogue of the most abundant HF taxa in the surface ocean presented in Chapter 2 is a central resource pointing at the key species that future research on HFs should focus on. The lack of model HF species is a main issue in the study of HF ecology, and a renewed culturing effort targeting these species could be the starting point to tackle it. As some of these abundant HF taxa may be extremely difficult to grow in culture, single-cell sorting could be a good trade-off to obtain new reference genomes. In fact, apart from the 3 species from the catalogue that are already available in culture, we have SAGs for another 10 species (Labarre et al., 2021), and retrieving single amplified genomes from the rest of them should be feasible. In Chapter 3 we had to use innovative solutions to access the gene expression of relevant uncultured HF species. This work represents one of the first studies assessing gene expression dynamics of natural HF assemblages where bacterivory has been promoted. With this, we could obtain candidate genes related to bacterivory, such as cysteine peptidases or glycoside hydrolases 3 and 20, that should be further investigated in more targeted studies.

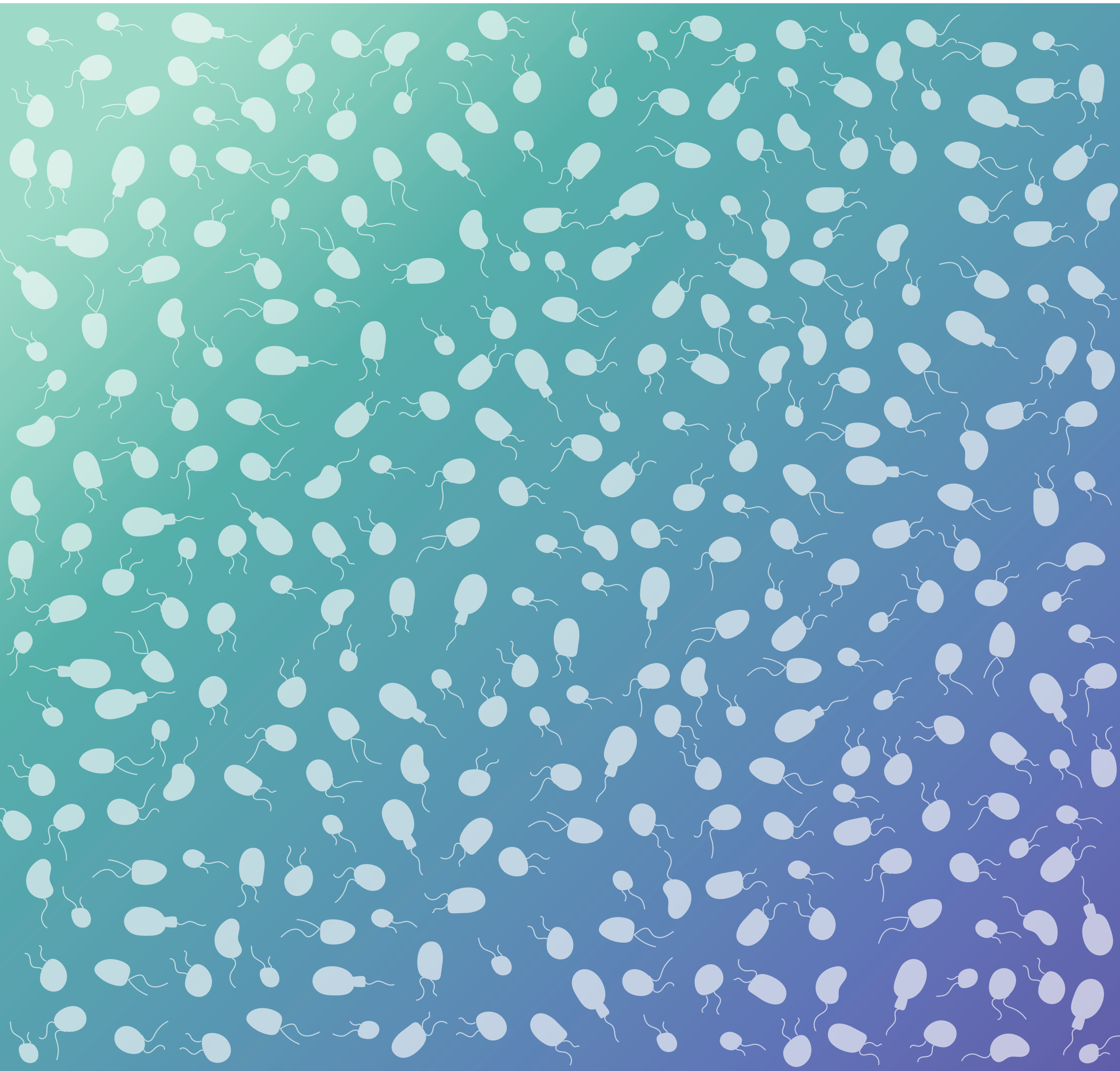
Generally, these genes are well studied in model organisms, but few information is available for uncultured lineages. Thus, good phylogenies spanning the complete eukaryotic diversity should be built in order to better characterize them, and new transcriptomic experiments with distantly-related species could be performed. With this, good references for these key bacterivory genes could be obtained, opening the way to new studies focused on better understanding the still poorly known biological process of phagocytosis.

The need for open data and code

The amount of data obtained by studies based on metabarcoding and omics tools is rapidly growing, boosted by the advances in sequencing technologies and the reduction of its costs (Goodwin et al., 2016). As it happens, in this thesis, and particularly in Chapter 3, we generated hundreds of gigabytes of data. This data increase is clearly positive, as it increments the amount of information available to draw more robust conclusions, but it is accompanied by some challenges, namely the increasing complexity of the analyses required to process them and the difficulty of data sharing with other teams interested in reproducing the results or making new analyses (Ravel & Wommack, 2014). Reproducibility is defined as the ability to recompute data analytic results given a data set and knowledge of its analysis. Together with replicability – the chances other researchers will achieve a consistent result –, they are foundational characteristics of successful scientific research (Leek & Peng, 2015). In fact, if a result cannot be reproduced, it is difficult to have confidence that it can be replicated or generalized (Schloss, 2018). Some of the main threats to reproducibility are data/code not being publicly available and the unavailability of specific software. It is crucial that researchers make raw data as well as comprehensive metadata accessible, an obvious requirement that is too often not met (Stodden et al., 2018). There are well-established databases for storing all types of sequencing data, such as figshare (figshare.com) or Zenodo (zenodo.org) that should be used to overcome this problem. The issue is even worse for code since few studies make it available. And those who do, they often do not make an effort to make it easily reusable. Repositories such as GitHub (github.com) or GitLab (gitlab.com) facilitate publishing analytical workflows in a well-documented manner, and this should be mandatory for all published bioinformatic research. In the case of software, all scientific papers should include the version used in the Methods section and, whenever possible, use open-source software. This kind of software is transparent (everyone can see its underlying code) and their development and maintenance are often community-driven, enabling all users to report detected issues or submit questions. In this thesis, I tried to make all the obtained results reproducible by making data and code publicly available, but there is still room for improvement. In Chapter 1, raw metagenomes could not be published, and we used commercial software (USEARCH) in our analyses due to the open-source version not having the capabilities we needed. Regarding general code, for all 3 chapters we could have used workflow management systems such as snakemake (snakemake.github.io) or NextFlow (nextflow.io) in order to make analyses more easily reproducible.

Open data and code not only serve as means for reproducibility, they can also open the door to new opportunities in research and are a tool for advancing the democratization of scientific knowledge. The public availability of these resources gives new opportunities for research groups that cannot afford large sequencing projects or oceanographic expeditions, for example. In fact, the amount of information contained in these data sets make them completely reusable and new results can be obtained from them, as we show in Chapter 2. We as scientists should consider if new questions can be answered with already available data and evaluate whether the collection of new samples or new sequencing batches are really required. This could not only reduce economic costs, but it also could avoid an unnecessary waste of resources. In a recently published perspective by Graves et al. (2022) on how to overcome the inequality of science, which is a persistent problem in the field at many levels (such as gender, race or social class), the authors argue that “the process of how we conduct research is just as important as the results of research”. I would say that the process is of even greater importance than the outcome, and this should be the guideline for a better science and, extensively, a better world.

CONCLUSIONS



Conclusions

1. Metagenomics tags (mTags) of the 18S-V4 rDNA are a powerful resource to assess the distribution of eukaryotic taxonomic groups in the water column. Given its low taxonomic resolution, these are best used as a by-product of functional studies rather than the ultimate approach for diversity surveys. For this, metabarcoding is still the way to go provided that putative amplification biases are considered.

2. PCR amplification with widely-used universal eukaryotic primers of the V4 region of the 18S rDNA is not suitable for groups like Diplonemea, Kinetoplastida, Prymnesiophyceae, Amoebozoa and Fungi. For Prymnesiophyceae, this is due to a critical mismatch (that can be easily solved), while for the rest of taxa a longer V4 insert is preventing a correct amplification. The majority of HF groups do not face strong technical biases derived from PCR amplification, so metabarcoding is a suitable approach to study their diversity and distribution.

3. The community structure obtained through mTags reveals a clear separation between pico- (0.2-3 µm) and nanoplanktonic (3-20 µm) fractions, as well as between photic (0-200 m) and aphotic (> 200 m) regions of the water column. Eukaryotic taxonomic groups tended to preferentially occupy one of the 4 niches originated from these 2 dimensions.

4. Using assembled eukaryotic rDNA operon sequences from metagenomes expands the resolution obtained by mTags and allows improving phylogenetic reconstructions of particular groups, such as the highly-diverse Diplonemea.

5. Temperature emerges as one of the main environmental factors shaping HF communities in the surface ocean. Along the water column, increasing depth results in a clear drop of HF diversity.

6. A few dozens of widespread taxa, mostly affiliating to MAST clades (mainly -1, -3, and -4), Picozoa, Bicosoecida and Chrysophyceae (uncultured clades G, H, and I), seem to dominate surface HF assemblages. The majority of these dominant HFs are present at relatively constant abundances, while others are influenced by temperature or display a patchy distribution.

7. In the deep ocean, only a handful of taxa belonging to Bicosoecida (species *C. burkhardae*, and *C. paraparvulus*) and Chrysophyceae (genus *Spumella* in clade C), together with Diplonemea and Kinetoplastida, explained most of the HF signal.

8. While in the deep ocean the majority of dominant HFs detected seem to be already in culture, no cultured representative exists for the dominant surface species besides a few patchy species, highlighting the current bias that exists in protistan knowledge.

9. Co-occurrence networks between HFs and prokaryotic taxa at the surface ocean reveal two main clusters influenced by temperature that do not seem to show specific patterns of interaction. However, some correlations emerge outside these thermal groups that could represent new prey-predator interactions.

10. The functional dynamics displayed by the 4 different unamended incubations followed similar patterns, with marked differences between the "lag" state and the other two stages of the incubation, "growth" and "decline".

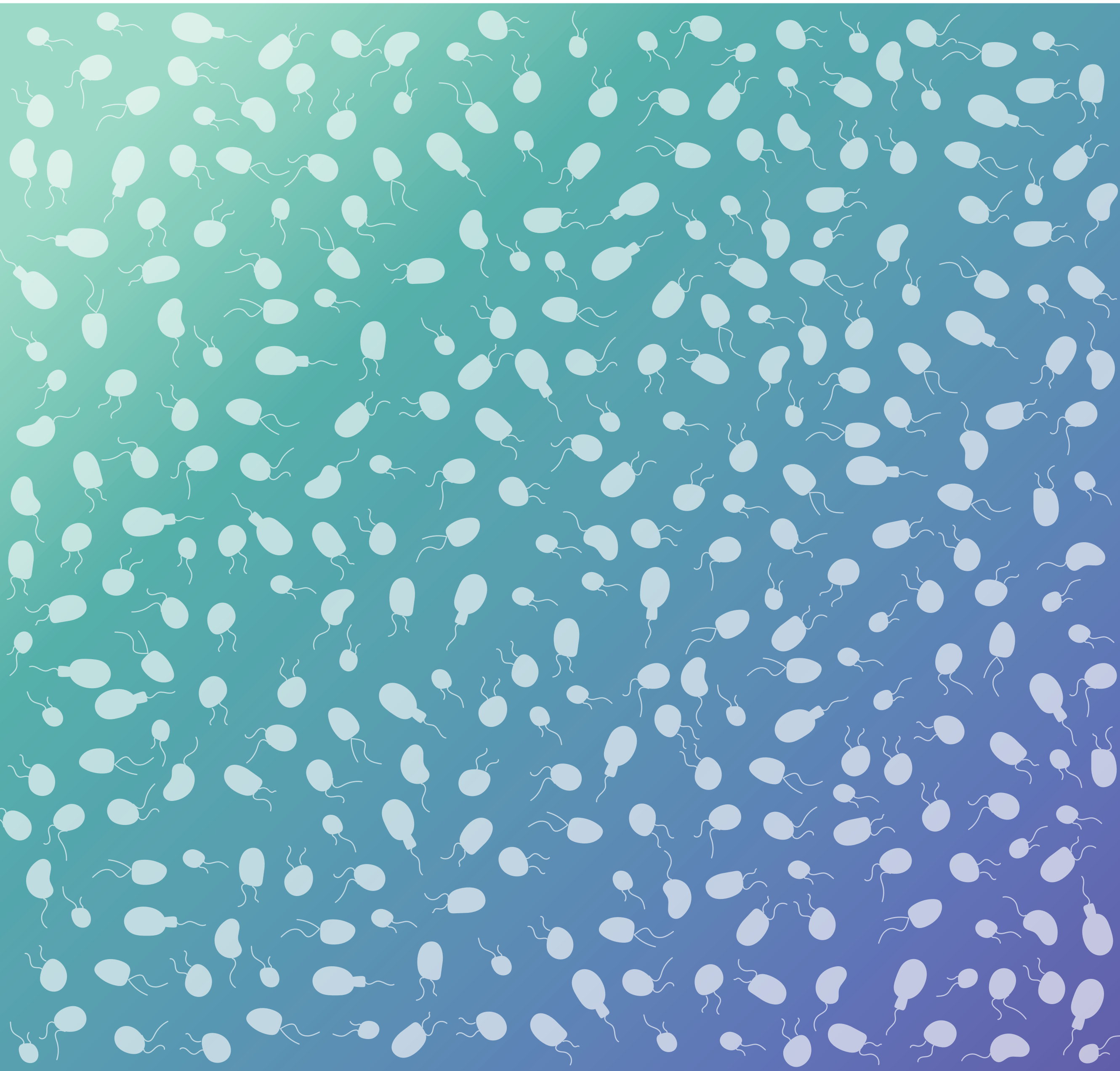
11. Genes related to photosynthesis rapidly decreased their expression in the incubations due to the absence of light.

12. Overall, the most expressed genes during the incubations were related to constitutive processes of the eukaryotic cell, namely actin, tubulin, ubiquitin and genes encoding for ribosomal proteins.

13. In the "growth" state of the incubations, genes related to constitutive processes were also highly expressed, together with genes related to peptidases, translocases and CAZy enzymes, which could be related to bacterivory.

14. Cysteine peptidases (mostly cathepsin L), and glycoside hydrolases (mostly 3 and 20 in CAZy classification) showed higher levels of relative expression in heterotrophic species compared to mixotrophic and phototrophic ones. Thus, they are clear targets for further exploration in the study of bacterivory.

GENERAL REFERENCES



General 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. https://doi.org/10.1128/AEM.71.12.8966-8969.2005

Acinas, S. G., Sebastián, M., & Ferrera, I. (2022). Towards a Global Perspective of the Marine Microbiome. In L. J. Stal & M. S. Cretoiu (Eds.), *The Marine Microbiome* (pp. 357–394). Springer International Publishing. https://doi.org/10.1007/978-3-030-90383-1_8

Adl, S. M., Bass, D., Lane, C. E., Lukeš, J., Schoch, C. L., Smirnov, A., Agatha, S., Berney, C., Brown, M. W., Burki, F., Cárdenas, P., Čepička, I., Chistyakova, L., Campo, J., Dunthorn, M., Edvardsen, B., Eglit, Y., Guillou, L., Hampl, V., … Zhang, Q. (2019). Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *Journal of Eukaryotic Microbiology*, 66(1), 4–119. https://doi.org/10.1111/jeu.12691

Alexander, H., Hu, S. K., Krinos, A. I., Pachiadaki, M., Tully, B. J., Neely, C. J., & Reiter, T. (2022). Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton (p. 2021.07.25.453713). *bioRxiv*. https://doi.org/10.1101/2021.07.25.453713

Amann, R. I., Ludwig, W., & Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews*, 59(1), 143–169. https://doi.org/10.1128/mr.59.1.143-169.1995

Andersen, R. A. (2004). Biology and systematics of heterokont and haptophyte algae. *American Journal of Botany*, 91(10), 1508–1522. https://doi.org/10.3732/ajb.91.10.1508

Arístegui, J., Gasol, J. M., Duarte, C. M., & Herndld, G. J. (2009). Microbial oceanography of the dark ocean's pelagic realm. *Limnology and Oceanography*, 54(5), 1501–1529. https://doi.org/10.4319/lo.2009.54.5.1501

Arndt, H., Dietrich, D., Auer, B., Cleven, E.-J., Gräfenhan, T., Weitere, M., & Mylnikov, A. (2000). Functional diversity of heterotrophic flagellates in aquatic ecosystems. In *The flagellates: Unity, diversity and evolution* (pp. 240–268).

Azam, F., Fenchel, T., Field, J., Gray, J., Meyer-Reil, L., & Thingstad, F. (1983). The Ecological Role of Water-Column Microbes in the Sea. *Marine Ecology Progress Series*, 10, 257–263. https://doi.org/10.3354/meps010257

Azuma, T., Pánek, T., Tice, A. K., Kayama, M., Kobayashi, M., Miyashita, H., Suzaki, T., Yabuki, A., Brown, M. W., & Kamikawa, R. (2022). An Enigmatic Stramenopile Sheds Light on Early Evolution in Ochrophyta Plastid Organellogenesis. *Molecular Biology and Evolution*, 39(4), msac065. https://doi.org/10.1093/molbev/msac065

Bachy, C., Hehenberger, E., Ling, Y.-C., Needham, D. M., Strauss, J., Wilken, S., & Worden, A. Z. (2022). Marine Protists: A Hitchhiker's Guide to their Role in the Marine Microbiome. In L. J. Stal & M. S. Cretoiu (Eds.), *The Marine Microbiome* (pp. 159–241). Springer International Publishing. https://doi.org/10.1007/978-3-030-90383-1_4

Baldauf, S. L. (2003). The Deep Roots of Eukaryotes. *Science*, 300(5626), 1703–1706. https://doi.org/10.1126/science.1085544

Bar-On, Y. M., & Milo, R. (2019). The Biomass Composition of the Oceans: A Blueprint of Our Blue Planet. *Cell*, 179(7), 1451–1454. https://doi.org/10.1016/j.cell.2019.11.018

Béjà, O. (2000). Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. *Science*, 289(5486), 1902–1906. https://doi.org/10.1126/science.289.5486.1902

Berney, C., Ciuprina, A., Bender, S., Brodie, J., Edgcomb, V., Kim, E., Rajan, J., Parfrey, L. W., Adl, S., Audic, S., Bass, D., Caron, D. A., Cochrane, G., Czech, L., Dunthorn, M., Geisen, S., Glöckner, F. O., Mahé, F., Quast, C., … de Vargas, C. (2017). UniEuk: Time to Speak a Common Language in Protistology! *Journal of Eukaryotic Microbiology*, 64(3), 407–411. https://doi.org/10.1111/jeu.12414

Blaxter, M., Archibald, J. M., Childers, A. K., Coddington, J. A., Crandall, K. A., Palma, F. D., Durbin, R., Edwards, S. V., Graves, J. A. M., Hackett, K. J., Hall, N., Jarvis, E. D., Johnson, R. N., Karlsson, E. K., Kress, W. J., Kuraku, S., Lawniczak, M. K. N., Lindblad-Toh, K., Lopez, J. V., … Lewin, H. A. (2022). Why sequence all eukaryotes? *Proceedings of the National Academy of Sciences*, 119(4), e2115636118. https://doi.org/10.1073/pnas.2115636118

Bochdansky, A. B., Clouse, M. A., & Herndl, G. J. (2017). Eukaryotic microbes, principally fungi and labyrinthulomycetes, dominate biomass on bathypelagic marine snow. *The ISME Journal*, 11(2), 362–373. https://doi.org/10.1038/ismej.2016.113

Boenigk, J., & Arndt, H. (2002). Bacterivory by heterotrophic flagellates: Community structure and feeding strategies. *Antonie van Leeuwenhoek*, 81(1), 465–480. https://doi.org/10.1023/A:1020509305868

Boenigk, J., Matz, C., Jürgens, K., & Arndt, H. (2001). Confusing Selective Feeding with Differential Digestion in Bacterivorous Nanoflagellates. *Journal of Eukaryotic Microbiology*, 48(4), 425–432. https://doi.org/10.1111/j.1550-7408.2001.tb00175.x

Boulais, J., Trost, M., Landry, C. R., Dieckmann, R., Levy, E. D., Soldati, T., Michnick, S. W., Thibault, P., & Desjardins, M. (2010). Molecular characterization of the evolution of phagosomes. *Molecular Systems Biology*, 6(1), 423. https://doi.org/10.1038/msb.2010.80

Bozzaro, S., Bucci, C., & Steinert, M. (2008). Phagocytosis and Host–Pathogen Interactions in Dictyostelium with a Look at Macrophages. In *International Review of Cell and Molecular Biology* (Vol. 271, pp. 253–300). Academic Press. https://doi.org/10.1016/S1937-6448(08)01206-9

Brown, M. W., Heiss, A. A., Kamikawa, R., Inagaki, Y., Yabuki, A., Tice, A. K., Shiratori, T., Ishida, K. I., Hashimoto, T., Simpson, A. G. B., & Roger, A. J. (2018). Phylogenomics Places Orphan Protistan Lineages in a Novel Eukaryotic Super-Group. *Genome Biology and Evolution*, 10(2), 427–433. https://doi.org/10.1093/gbe/evy014

Brown, M. W., Sharpe, S. C., Silberman, J. D., Heiss, A. A., Lang, B. F., Simpson, A. G. B., & Roger, A. J. (2013). Phylogenomics demonstrates that breviate flagellates are related to opisthokonts and apusomonads. *Proceedings of the Royal Society B: Biological Sciences*, 280(1769), 20131755. https://doi.org/10.1098/rspb.2013.1755

Burki, F., & Keeling, P. J. (2014). Rhizaria. *Current Biology*, 24(3), R103–R107. https://doi.org/10.1016/j.cub.2013.12.025

Burki, F., Roger, A. J., Brown, M. W., & Simpson, A. G. B. (2020). The New Tree of Eukaryotes. *Trends in Ecology & Evolution*, 35(1), 43–55. https://doi.org/10.1016/j.tree.2019.08.008

Burki, F., Sandin, M. M., & Jamy, M. (2021). Diversity and ecology of protists revealed by metabarcoding. *Current Biology*, 31(19), R1267–R1280. https://doi.org/10.1016/j.cub.2021.07.066

Burki, F., Shalchian-Tabrizi, K., Minge, M., Skjæveland, Å., Nikolaev, S. I., Jakobsen, K. S., & Pawlowski, J. (2007). Phylogenomics Reshuffles the Eukaryotic Supergroups. *PLoS ONE*, 2(8), e790. https://doi.org/10.1371/journal.pone.0000790

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. https://doi.org/10.1146/annurev-marine-120709-142802

Caron, D. A., Davis, P. G., & Sieburth, J. M. (1989). Factors responsible for the differences in cultural estimates and direct microscopical counts of populations of bacterivorous nanoflagellates. *Microbial Ecology*, 18(2), 89–104. https://doi.org/10.1007/BF02030118

Caron, D. A., Goldman, J. C., & Dennett, M. R. (1986). Effect of Temperature on Growth, Respiration, and Nutrient Regeneration by an Omnivorous Microflagellate. *Applied and Environmental Microbiology*, 52(6), 1340 LP – 1347.

Caron, D. A., Goldman, J. C., & Dennett, M. R. (1990). Carbon utilization by the omnivorous flagellate Paraphysomonas imperforata. *Limnology and Oceanography*, 35(1), 192–201. https://doi.org/10.4319/lo.1990.35.1.0192

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Méheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., … Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, 9(1), 373. https://doi.org/10.1038/s41467-017-02342-1

Cavalier-Smith, T., & Chao, E. E. (2012). Oxnerella micra sp. N. (Oxnerellidae fam. N.), a Tiny Naked Centrohelid, and the Diversity and Evolution of Heliozoa. *Protist*, 163(4), 574–601. https://doi.org/10.1016/j.protis.2011.12.005

Cavalier-Smith, T., & Scoble, J. M. (2013). Phylogeny of Heterokonta: Incisomonas marina, a uniciliate gliding opalozoan related to Solenicola (Nanomonadea), and evidence that Actinophryida evolved from raphidophytes. *European Journal of Protistology*, 49(3), 328–353. https://doi.org/10.1016/j.ejop.2012.09.002

Charette, M. A., & Smith, W. H. F. (2010). The Volume of Earth’s Ocean. *Oceanography*, 23(2), 112–114.

Chrzanowski, T. H., & Foster, B. L. L. (2014). Prey element stoichiometry controls ecological fitness of the flagellate Ochromonas danica. *Aquatic Microbial Ecology*, 71(3), 257–269. https://doi.org/10.3354/ame01680

Chrzanowski, T. H., & Šimek, K. (1990). Prey-size selection by freshwater flagellated protozoa. *Limnology and Oceanography*, 35(7), 1429–1436. https://doi.org/10.4319/lo.1990.35.7.1429

Corliss, J. O. (1984). The kingdom PROTISTA and its 45 phyla. *Biosystems*, 17(2), 87–126. https://doi.org/10.1016/0303-2647(84)90003-0

Curtis, B. A., Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M. C., Ball, S. G., Gile, G. H., Hirakawa, Y., Hopkins, J. F., Kuo, A., Rensing, S. A., Schmutz, J., Symeonidi, A., Elias, M., Eveleigh, R. J. M., Herman, E. K., Klute, M. J., … Archibald, J. M. (2012). Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature*, 492(7427), 59–65. https://doi.org/10.1038/nature11681

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.-M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., … Velayoudon, D. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237), 1261605–1261605. https://doi.org/10.1126/science.1261605 del Campo, J., Balagué, V., Forn, I., Lekunberri, I., & Massana, R. (2013). Culturing Bias in Marine Heterotrophic Flagellates Analyzed Through Seawater Enrichment Incubations. *Microbial Ecology*, 66(3), 489–499. https://doi.org/10.1007/s00248-013-0251-y

del Campo, J., Bass, D., & Keeling, P. J. (2020). The eukaryome: Diversity and role of microeukaryotic organisms associated with animal hosts. *Functional Ecology*, 34(10), 2045–2054. https://doi.org/10.1111/1365-2435.13490

del Campo, J., Heger, T. J., Rodríguez-Martínez, R., Worden, A. Z., Richards, T. A., Massana, R., & Keeling, P. J. (2019). Assessing the Diversity and Distribution of Apicomplexans in Host and Free-Living Environments Using High-Throughput Amplicon Data and a Phylogenetically Informed Reference Framework. *Frontiers in Microbiology*, 10(OCT), 1–15. https://doi.org/10.3389/fmicb.2019.02373

del Campo, J., Not, F., Forn, I., Sieracki, M. E., & Massana, R. (2013). Taming the smallest predators of the oceans. *The ISME Journal*, 7(2), 351–358. https://doi.org/10.1038/ismej.2012.85

del Campo, J., Sieracki, M. E., Molestina, R., Keeling, P. J., Massana, R., & Ruiz-Trillo, I. (2014). The others: Our biased perspective of eukaryotic genomes. *Trends in Ecology & Evolution*, 29(5), 252–259. https://doi.org/10.1016/j.tree.2014.03.006

Delmont, T. O., Gaia, M., Hinsinger, D. D., Frémont, P., Vanni, C., Fernandez-Guerra, A., Eren, A. M., Kourlaiev, A., d'Agata, L., Clayssen, Q., Villar, E., Labadie, K., Cruaud, C., Poulain, J., Da Silva, C., Wessner, M., Noel, B., Aury, J.-M., Sunagawa, S., … Jaillon, O. (2022). Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics*, 100123. https://doi.org/10.1016/j.xgen.2022.100123

DeLong, E. F. (1992). Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences*, 89(12), 5685–5689. https://doi.org/10.1073/pnas.89.12.5685

Díez, 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. https://doi.org/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. https://doi.org/10.1002/lob.10008

Duncan, A., Barry, K., Daum, C., Eloe-Fadrosh, E., Roux, S., Schmidt, K., Tringe, S. G., Valentin, K. U., Varghese, N., Salamov, A., Grigoriev, I. V., Leggett, R. M., Moulton, V., & Mock, T. (2022). Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans. *Microbiome*, 10(1), 67. https://doi.org/10.1186/s40168-022-01254-7

Eccleston-Parry, J. D., & Leadbeater, B. S. C. (1994). A comparison of the growth kinetics of six marine heterotrophic nanoflagellates fed with one bacterial species. *Marine Ecology Progress Series*, 105(1/2), 167–177.

Eme, L., Spang, A., Lombard, J., Stairs, C. W., & Ettema, T. J. G. (2017). Archaea and the origin of eukaryotes. *Nature Reviews Microbiology*, 15(12), 711–723. https://doi.org/10.1038/nrmicro.2017.133

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

Falkowski, P., Fenchel, T., & Delong, E. F. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, 320(5879), 1034–1039. https://doi.org/10.1126/science.1153213

Fenchel, T. (1982a). Ecology of Heterotrophic Microflagellates. I. Some Important Forms and Their Functional Morphology. *Marine Ecology Progress Series*, 8, 211–223. https://doi.org/10.3354/meps008211

Fenchel, T. (1982b). Ecology of Heterotrophic Microflagellates. II. Bioenergetics and Growth. *Marine Ecology Progress Series*, 8, 225–231. https://doi.org/10.3354/meps008225

Fenchel, T. (1982c). Ecology of Heterotrophic Microflagellates. III. Adaptations to Heterogeneous Environments. *Marine Ecology Progress Series*, 9, 25–33. https://doi.org/10.3354/meps009025

Fenchel, T. (1982d). Ecology of Heterotrophic Microflagellates. IV Quantitative Occurrence and Importance as Bacterial Consumers. *Marine Ecology Progress Series*, 9(1977), 35–42. https://doi.org/10.3354/meps009035

Fenchel, T. (1986). The Ecology of Heterotrophic Microflagellates. In K. C. Marshall (Ed.), *Advances in Microbial Ecology* (pp. 57–97). Springer US. https://doi.org/10.1007/978-1-4757-0611-6_2

Fenchel, T. (1987). Ecology of Protozoa. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-06817-5

Fiehn, O. (2002). Metabolomics—The link between genotypes and phenotypes. In C. Town (Ed.), *Functional Genomics* (pp. 155–171). Springer Netherlands. https://doi.org/10.1007/978-94-010-0448-0_11

Flannagan, R. S., Jaumouillé, V., & Grinstein, S. (2012). The Cell Biology of Phagocytosis. *Annual Review of Pathology: Mechanisms of Disease*, 7(1), 61–98. https://doi.org/10.1146/annurev-pathol-011811-132445

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

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

Fuhrman, J. A., McCallum, K., & Davis, A. A. (1992). Novel major archaebacterial group from marine plankton. *Nature*, 356(6365), 148–149. https://doi.org/10.1038/356148a0

Fuhrman, J. A., & Noble, R. T. (1995). Viruses and protists cause similar bacterial mortality in coastal seawater. *Limnology and Oceanography*, 40(7), 1236–1242. https://doi.org/10.4319/lo.1995.40.7.1236

Gabaldón, T. (2021). Origin and Early Evolution of the Eukaryotic Cell. *Annual Review of Microbiology*, 75(1), 631–647. https://doi.org/10.1146/annurev-micro-090817-062213

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

Gawryluk, R. M. R., Tikhonenkov, D. V., Hehenberger, E., Husnik, F., Mylnikov, A. P., & Keeling, P. J. (2019). Non-photosynthetic predators are sister to red algae. *Nature*, 572(7768), 240–243. https://doi.org/10.1038/s41586-019-1398-6

Geider, R., & Leadbeater, B. (1988). Kinetics and energetics of growth of the marine choanoflagellate Stephanoeca diplocostata. *Marine Ecology Progress Series*, 47, 169–177. https://doi.org/10.3354/meps047169

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

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. https://doi.org/10.1038/s41396-019-0506-9

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., & Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, 345(6270), 60–63. https://doi.org/10.1038/345060a0Glücksman, E., Bell, T., Griffiths, R. I., & Bass, D. (2010). Closely related protist strains have different grazing impacts on natural bacterial communities. *Environmental Microbiology*, 12(12), 3105–3113. https://doi.org/10.1111/j.1462-2920.2010.02283.x

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Data sets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8(NOV), 1–6. https://doi.org/10.3389/fmicb.2017.02224

Goldman, J. C., & Caron, D. A. (1985). Experimental studies on an omnivorous microflagellate: Implications for grazing and nutrient regeneration in the marine microbial food chain. *Deep Sea Research Part A. Oceanographic Research Papers*, 32(8), 899–915. https://doi.org/10.1016/0198-0149(85)90035-4

Gómez, F., Moreira, D., Benzerara, K., & López-García, P. (2011). Solenicola setigera is the first characterized member of the abundant and cosmopolitan uncultured marine stramenopile group MAST-3. *Environmental Microbiology*, 13(1), 193–202. https://doi.org/10.1111/j.1462-2920.2010.02320.x

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. https://doi.org/10.1038/nrg.2016.49

Graves, J. L., Kearney, M., Barabino, G., & Malcom, S. (2022). Inequality in science and the case for a new agenda. *Proceedings of the National Academy of Sciences*, 119(10), e2117831119. https://doi.org/10.1073/pnas.2117831119

Griessmann, K. (1914). Über marine Flagellaten. *Archiv für Protistenkunde*, 32, 1–78.

Grossart, H. P., Van den Wyngaert, S., Kagami, M., Wurzbacher, C., Cunliffe, M., & Rojas-Jimenez, K. (2019). Fungi in aquatic ecosystems. *Nature Reviews Microbiology*. https://doi.org/10.1038/s41579-019-0175-8

Guillou, L., Viprey, M., Chambouvet, A., Welsh, R. M., Kirkham, A. R., Massana, R., Scanlan, D. J., & Worden, A. Z. (2008). Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environmental Microbiology*, 10(12), 3349–3365. https://doi.org/10.1111/j.1462-2920.2008.01731.x

Haas, L. W., & Webb, K. L. (1979). Nutritional mode of several non-pigmented microflagellates from the York River estuary, Virginia. *Journal of Experimental Marine Biology and Ecology*, 39(2), 125–134. https://doi.org/10.1016/0022-0981(79)90009-1

Haeckel, E. H. P. A. (1887). *Report on the Radiolaria collected by H.M.S. Challenger during the years 1873-76*. http://archive.org/details/reportonradiolar00haecrich

Hahn, M. (2002). Grazing of protozoa and its effect on populations of aquatic bacteria. *FEMS Microbiology Ecology*, 35. https://doi.org/10.1016/s0168-6496(00)00098-2

Jamy, M., Foster, R., Barbera, P., Czech, L., Kozlov, A., Stamatakis, A., Bending, G., Hilton, S., Bass, D., & Burki, F. (2020). Long-read metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular Ecology Resources*, 20(2), 429–443. https://doi.org/10.1111/1755-0998.13117

Jeong, H. J., Yoo, Y. D., Kim, J. S., Seong, K. A., Kang, N. S., & Kim, T. H. (2010). Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs. *Ocean Science Journal*, 45(2), 65–91. https://doi.org/10.1007/s12601-010-0007-2

Jeuck, A., & Arndt, H. (2013). A Short Guide to Common Heterotrophic Flagellates of Freshwater Habitats Based on the Morphology of Living Organisms. *Protist*, 164(6), 842–860. https://doi.org/10.1016/j.protis.2013.08.003

Jürgens, K., & DeMott, W. R. (1995). Behavioral flexibility in prey selection by bacterivorous nanoflagellates. *Limnology and Oceanography*, 40(8), 1503–1507. https://doi.org/10.4319/lo.1995.40.8.1503

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). John Wiley & Sons, Inc. https://doi.org/10.1002/9780470281840.ch11

Jürgens, K., & Matz, C. (2002). Predation as a shaping force for the phenotypic and genotypic composition of planktonic bacteria. *Antonie van Leeuwenhoek*, 81(1), 413–434. https://doi.org/10.1023/A:1020505204959

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., Vargas, C. D., Raes, J., Sullivan, M., Arendt, D., Benzoni, F., Claverie, J.-M., Follows, M., Gorsky, G., Hingamp, P., Iudicone, D., Jaillon, O., Kandels-Lewis, S., Krzic, U., Not, F., Ogata, H., … Consortium, the T. O. (2011). A Holistic Approach to Marine Eco-Systems Biology. *PLOS Biology*, 9(10), e1001177. https://doi.org/10.1371/journal.pbio.1001177

Keeling, P. J., & Burki, F. (2019). Progress towards the Tree of Eukaryotes. *Current Biology*, 29(16), R808–R817. https://doi.org/10.1016/j.cub.2019.07.031

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., … Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biology*, 12(6). https://doi.org/10.1371/journal.pbio.1001889

Keeling, P. J., & del Campo, J. (2017). Marine Protists Are Not Just Big Bacteria. *Current Biology*, 27(11), R541–R549. https://doi.org/10.1016/j.cub.2017.03.075

Kleiner, M. (2019). Metaproteomics: Much More than Measuring Gene Expression in Microbial Communities. *MSystems*, 4(3), e00115-19. https://doi.org/10.1128/mSystems.00115-19

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. https://doi.org/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. https://doi.org/10.1016/j.ejop.2015.04.002

Kudryavtsev, A., Pawlowski, J., & Smirnov, A. (2018). More amoebae from the deep-sea: Two new marine species of Vexillifera (Amoebozoa, Dactylopodida) with notes on taxonomy of the genus. *European Journal of Protistology*, 66, 9–25. https://doi.org/10.1016/j.ejop.2018.07.001

Kudryavtsev, A., Voytinsky, F., & Volkova, E. (2022). Coronamoeba villafranca gen. Nov. Sp. Nov. (Amoebozoa, Dermamoebida) challenges the correlation of morphology and phylogeny in Amoebozoa. *Scientific Reports*, 12(1), 12541. https://doi.org/10.1038/s41598-022-16721-2

Labarre, A., López-Escardó, D., Latorre, F., Leonard, G., Bucchini, F., Obiol, A., Cruaud, C., Sieracki, M. E., Jaillon, O., Wincker, P., Vandepoele, K., Logares, R., & Massana, R. (2021). Comparative genomics reveals new functional insights in uncultured MAST species. *The ISME Journal*, 15(6), 1767–1781. https://doi.org/10.1038/s41396-020-00885-8

Labarre, A., Obiol, A., Wilken, S., Forn, I., & Massana, R. (2020). Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates. *Limnology and Oceanography*, 65(S1), lno.11379. https://doi.org/10.1002/lno.11379

Lambert, B. S., Groussman, R. D., Schatz, M. J., Coesel, S. N., Durham, B. P., Alverson, A. J., White, A. E., & Armbrust, E. V. (2022). The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. *Proceedings of the National Academy of Sciences*, 119(7). https://doi.org/10.1073/pnas.2100916119

Lambert, S., Tragin, M., Lozano, J.-C., Ghiglione, J.-F., Vaulot, D., Bouget, F.-Y., & Galand, P. E. (2018). Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. *The ISME Journal*, submitted. https://doi.org/10.1038/s41396-018-0281-z

Latorre, F., Deutschmann, I. M., Labarre, A., Obiol, A., Krabberød, A. K., Pelletier, E., Sieracki, M. E., Cruaud, C., Jaillon, O., Massana, R., & Logares, R. (2021). Niche adaptation promoted the evolutionary diversification of tiny ocean predators. *Proceedings of the National Academy of Sciences*, 118(25). https://doi.org/10.1073/pnas.2020955118

Lax, G., Eglit, Y., Eme, L., Bertrand, E. M., Roger, A. J., & Simpson, A. G. B. (2018). Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature*, 564(7736), 410–414. https://doi.org/10.1038/s41586-018-0708-8

Leek, J. T., & Peng, R. D. (2015). Reproducible research can still be wrong: Adopting a prevention approach. *Proceedings of the National Academy of Sciences*, 112(6), 1645–1646. https://doi.org/10.1073/pnas.1421412111

Leeuwenhoek, A. V. (1677). Observations, communicated to the publisher by Mr. Antony van Leewenhoeck, in a dutch letter of the 9th Octob. 1676. here English'd: Concerning little animals by him observed in rain-well-sea- and snow water; as also in water wherein pepper had lain infused. *Philosophical Transactions of the Royal Society of London*, 12(133), 821–831. https://doi.org/10.1098/rstl.1677.0003

Lohmann, H. (1911). Über das Nannoplankton und die Zentrifugierung kleinster Wasserproben zur Gewinnung desselben in lebendem Zustande. *Internationale Revue der gesamten Hydrobiologie und Hydrographie*, 4(1–2), 1–38. https://doi.org/10.1002/iroh.19110040102

Lønborg, C., Carreira, C., Jickells, T., & Álvarez-Salgado, X. A. (2020). Impacts of Global Change on Ocean Dissolved Organic Carbon (DOC) Cycling. *Frontiers in Marine Science*, 7. https://www.frontiersin.org/article/10.3389/fmars.2020.00466

López-García, P., Eme, L., & Moreira, D. (2017). Symbiosis in eukaryotic evolution. *Journal of Theoretical Biology*, 434, 20–33. https://doi.org/10.1016/j.jtbi.2017.02.031

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. https://doi.org/10.1038/35054537

Mangot, J., Forn, I., Obiol, A., & Massana, R. (2018). Constant abundances of ubiquitous uncultured protists in the open sea assessed by automated microscopy. *Environmental Microbiology*, 20(10), 3876–3889. https://doi.org/10.1111/1462-2920.14408

Mann, D. G. (1999). The species concept in diatoms. *Phycologia*, 38(6), 437–495. https://doi.org/10.2216/i0031-8884-38-6-437.1

Marin, B., M. Nowack, E. C., & Melkonian, M. (2005). A Plastid in the Making: Evidence for a Second Primary Endosymbiosis. *Protist*, 156(4), 425–432. https://doi.org/10.1016/j.protis.2005.09.001

Massana, R. (2011). Eukaryotic Picoplankton in Surface Oceans. *Annual Review of Microbiology*, 65(1), 91–110. https://doi.org/10.1146/annurev-micro-090110-102903

Massana, R. (2020). Petits depredadors marins. *Treballs de la Societat Catalana de Biologia*, 70, 12–18.

Massana, R., Castresana, J., Balagué, V., Guillou, L., Romari, K., Valentin, K., & Pedro, C. (2004). Phylogenetic and Ecological Analysis of Novel Marine Stramenopiles. *Applied and Environmental Microbiology*, 70(6), 3528–3534. https://doi.org/10.1128/AEM.70.6.3528

Massana, R., del Campo, J., Sieracki, M. E., Audic, S., & Logares, R. (2014). Exploring the uncultured microeukaryote majority in the oceans: Reevaluation of ribogroups within stramenopiles. *The ISME Journal*, 8(4), 854–866. https://doi.org/10.1038/ismej.2013.204

Massana, R., Labarre, A., López-Escardó, D., Obiol, A., Bucchini, F., Hackl, T., Fischer, M. G., Vandepoele, K., Tikhonenkov, D. V., Husnik, F., & Keeling, P. J. (2021). Gene expression during bacterivorous growth of a widespread marine heterotrophic flagellate. *The ISME Journal*, 15(1), 154–167. https://doi.org/10.1038/s41396-020-00770-4

Massana, R., & Logares, R. (2013). Eukaryotic versus prokaryotic marine picoplankton ecology. *Environmental Microbiology*, 15(5), 1254–1261. https://doi.org/10.1111/1462-2920.12043

Massana, R., & López-Escardó, D. (2022). Metagenome assembled genomes are for eukaryotes too. *Cell Genomics*, 2(5), 100130. https://doi.org/10.1016/j.xgen.2022.100130

Massana, R., Unrein, F., Rodríguez-Martínez, R., Forn, I., Lefort, T., Pinhassi, J., & Not, F. (2009). Grazing rates and functional diversity of uncultured heterotrophic flagellates. *The ISME Journal*, 3(5), 588–595. https://doi.org/10.1038/ismej.2008.130

Mast, S. O. (1947). The food-vacuole in paramecium. *The Biological Bulletin*, 92(1), 31–72. https://doi.org/10.2307/1537967

McNichol, J., Berube, P. M., Biller, S. J., Fuhrman, J. A., & Gilbert, J. A. (2021). Evaluating and Improving Small Subunit rRNA PCR Primer Coverage for Bacteria, Archaea, and Eukaryotes Using Metagenomes from Global Ocean Surveys. *MSystems*, 0(0), e00565-21. https://doi.org/10.1128/mSystems.00565-21

Mills, D. B. (2020). The origin of phagocytosis in Earth history. *Interface Focus*, 10(4), 20200019. https://doi.org/10.1098/rsfs.2020.0019

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. https://doi.org/10.1038/35054541

Not, F., Valentin, K., Romari, K., Lovejoy, C., Massana, R., Töbe, K., Vaulot, D., & Medlin, L. K. (2007). Picobiliphytes: A Marine Picoplanktonic Algal Group with Unknown Affinities to Other Eukaryotes. *Science*, 315(5809), 253–255. https://doi.org/10.1126/science.1136264

Okamoto, N., Chantangsi, C., Horák, A., Leander, B. S., & Keeling, P. J. (2009). Molecular Phylogeny and Description of the Novel Katablepharid Roombia truncata gen. Et sp. Nov., and Establishment of the Hacrobia Taxon nov. *PLoS ONE*, 4(9), e7080. https://doi.org/10.1371/journal.pone.0007080

Pedrós-Alió, C. (2006). Marine microbial diversity: Can it be determined? *Trends in Microbiology*, 14(6), 257–263. https://doi.org/10.1016/j.tim.2006.04.007

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

Pernthaler, J. (2005). Predation on prokaryotes in the water column and its ecological implications. *Nature Reviews Microbiology*, 3(7), 537–546. https://doi.org/10.1038/nrmicro1180

Piwosz, K., Mukherjee, I., Salcher, M. M., Grujčić, V., & Šimek, K. (2021). CARD-FISH in the Sequencing Era: Opening a New Universe of Protistan Ecology. *Frontiers in Microbiology*, 12. https://doi.org/10.3389/fmicb.2021.640066

Pomeroy, L. R. (1974). The Ocean's Food Web, A Changing Paradigm. *BioScience*, 24(9), 499–504. https://doi.org/10.2307/1296885

Prokopchuk, G., Korytář, T., Juricová, V., Majstorović, J., Horák, A., Šimek, K., & Lukeš, J. (2022). Trophic flexibility of marine diplonemids—Switching from osmotrophy to bacterivory. *The ISME Journal*, 1–11. https://doi.org/10.1038/s41396-022-01192-0

Ravel, J., & Wommack, K. (2014). All hail reproducibility in microbiome research. *Microbiome*, 2(1), 8. https://doi.org/10.1186/2049-2618-2-8

Richards, T. A., Massana, R., Pagliara, S., & Hall, N. (2019). Single cell ecology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1786), 20190076. https://doi.org/10.1098/rstb.2019.0076

Richter, D. J., Berney, C., Strassert, J. F. H., Poh, Y.-P., Herman, E. K., Muñoz-Gómez, S. A., Wideman, J. G., Burki, F., & Vargas, C. de. (2022). EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. In *BioRxiv* (p. 2020.06.30.180687). https://doi.org/10.1101/2020.06.30.180687

Rodríguez-Martínez, R., Vaqué, D., Forn, I., & Massana, R. (2022). Dominant marine heterotrophic flagellates are adapted to natural planktonic bacterial abundances. *Environmental Microbiology*, n/a(n/a). https://doi.org/10.1111/1462-2920.15911

Rogerson, A., Anderson, O. R., & Vogel, C. (2003). Are planktonic naked amoebae predominately floc associated or free in the water column? *Journal of Plankton Research*, 25(11), 1359–1365. https://doi.org/10.1093/plankt/fbg102

Rosales, C., & Uribe-Querol, E. (2017). Phagocytosis: A Fundamental Process in Immunity. *BioMed Research International*, 2017, 9042851. https://doi.org/10.1155/2017/9042851

Sanders, R. W., Caron, D. A., & Berninger, U.-G. (1992). Relationships between bacteria and heterotrophic nanoplankton in marine and fresh waters: An inter-ecosystem comparison. *Marine Ecology Progress Series*, 86(1), 1–14.

Schloss, P. D. (2018). Identifying and Overcoming Threats to Reproducibility, Replicability, Robustness, and Generalizability in Microbiome Research. *MBio*, 9(3), e00525-18. https://doi.org/10.1128/mBio.00525-18

Schoenle, A., Hohlfeld, M., Rosse, M., Filz, P., Wylezich, C., Nitsche, F., & Arndt, H. (2020). Global comparison of bicosoecid Cafeteria-like flagellates from the deep ocean and surface waters, with reorganization of the family Cafeteriaceae. *European Journal of Protistology*, 73, 125665. https://doi.org/10.1016/j.ejop.2019.125665

Schön, M. E., Zlatogursky, V. V., Singh, R. P., Poirier, C., Wilken, S., Mathur, V., Strassert, J. F. H., Pinhassi, J., Worden, A. Z., Keeling, P. J., Ettema, T. J. G., Wideman, J. G., & Burki, F. (2021). Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. *Nature Communications*, 12(1), 6651. https://doi.org/10.1038/s41467-021-26918-0

Sebé-Pedrós, A., Degnan, B. M., & Ruiz-Trillo, I. (2017). The origin of Metazoa: A unicellular perspective. *Nature Reviews Genetics*, 18(8), 498–512. https://doi.org/10.1038/nrg.2017.21

Seeleuthner, Y., Mondy, S., Lombard, V., Carradec, Q., Pelletier, E., Wessner, M., Leconte, J., Mangot, J.-F., Poulain, J., Labadie, K., Logares, R., Sunagawa, S., de Berardinis, V., Salanoubat, M., Dimier, C., Kandels-Lewis, S., Picheral, M., Searson, S., Pesant, S., … Wincker, P. (2018). Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nature Communications*, 9(1), 310. https://doi.org/10.1038/s41467-017-02235-3

Shalchian-Tabrizi, K., Eikrem, W., Klaveness, D., Vaulot, D., Minge, M. a, Le Gall, F., Romari, K., Throndsen, J., Botnen, A., Massana, R., Thomsen, H. a, & Jakobsen, K. s. (2006). Telonemia, a new protist phylum with affinity to chromist lineages. *Proceedings of the Royal Society B: Biological Sciences*, 273(1595), 1833–1842. https://doi.org/10.1098/rspb.2006.3515

Sherr, E. B., & Sherr, B. F. (2002). Significance of predation by protists in aquatic microbial food webs. *Antonie van Leeuwenhoek*, 81(1), 293–308. https://doi.org/10.1023/A:1020591307260

Shiratori, T., Thakur, R., & Ishida, K. (2017). Pseudophyllomitus vesiculosus (Larsen and Patterson 1990) Lee, 2002, a Poorly Studied Phagotrophic Biflagellate is the First Characterized Member of Stramenopile Environmental Clade MAST-6. *Protist*, 168(4), 439–451. https://doi.org/10.1016/j.protis.2017.06.004

Sibbald, S. J., & Archibald, J. M. (2020). Genomic Insights into Plastid Evolution. *Genome Biology and Evolution*, 12(7), 978–990. https://doi.org/10.1093/gbe/evaa096

Šimek, K., Grujčić, V., Hahn, M. W., Horňák, K., Jezberová, J., Kasalický, V., Nedoma, J., Salcher, M. M., & Shabarova, T. (2018). Bacterial prey food characteristics modulate community growth response of freshwater bacterivorous flagellates. *Limnology and Oceanography*, 63(1), 484–502. https://doi.org/10.1002/lno.10759

Šimek, K., Kasalický, V., Jezbera, J., Horňák, K., Nedoma, J., Hahn, M. W., Bass, D., Jost, S., & Boenigk, J. (2013). Differential freshwater flagellate community response to bacterial food quality with a focus on Limnohabitans bacteria. *The ISME Journal*, 7(8), 1519–1530. https://doi.org/10.1038/ismej.2013.57

Simpson, A. G. B., & Eglit, Y. (2016). Protist Diversification. In *Encyclopedia of Evolutionary Biology* (Vol. 3, pp. 344–360). Elsevier. https://doi.org/10.1016/B978-0-12-800049-6.00247-X

Smetacek, V. (1999). Revolution in the ocean. *Nature*, 401(6754), 647–647. https://doi.org/10.1038/44281

Sorokin, Y. (1978). Decomposition of organic matter and nutrient regeneration. In *Marine Ecology, IV: Dynamics* (O. Kinne, pp. 501–516). John Wiley.

Sorokin, Y. (1979). Zooflagellates as a component of the community of eutrophic and oligotrophic waters in the Pacific Ocean. *Oceanology*, 19, 316–319.

Staley, J. T., & Konopka, A. (1985). Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annual Review of Microbiology*, 39, 321–346. https://doi.org/10.1146/annurev.mi.39.100185.001541

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11), 2584–2589. https://doi.org/10.1073/pnas.1708290115

Stoecker, D. K., Hansen, P. J., Caron, D. A., & Mitra, A. (2017). Mixotrophy in the Marine Plankton. *Annual Review of Marine Science*, 9(1), 311–335. https://doi.org/10.1146/annurev-marine-010816-060617

Stoecker, D. K., Johnson, M. D., Vargas, C. de, & Not, F. (2009). Acquired phototrophy in aquatic protists. *Aquatic Microbial Ecology*, 57(3), 279–310. https://doi.org/10.3354/ame01340

Strassert, J. F. H., Jamy, M., Mylnikov, A. P., Tikhonenkov, D. V., & Burki, F. (2019). New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life. *Molecular Biology and Evolution*, 36(4), 757–765. https://doi.org/10.1093/molbev/msz012

Taylor, F. J. R. (2003). The collapse of the two-kingdom system, the rise of protistology and the founding of the International Society for Evolutionary Protistology (ISEP). In I*nternational Journal of Systematic and Evolutionary Microbiology* (Vol. 53, Issue 6, pp. 1707–1714). Microbiology Society. https://doi.org/10.1099/ijs.0.02587-0

Taylor, F. J. R., Hoppenrath, M., & Saldarriaga, J. F. (2008). Dinoflagellate diversity and distribution. *Biodiversity and Conservation*, 17(2), 407–418. https://doi.org/10.1007/s10531-007-9258-3

Tice, A. K., Žihala, D., Pánek, T., Jones, R. E., Salomaki, E. D., Nenarokov, S., Burki, F., Eliáš, M., Eme, L., Roger, A. J., Rokas, A., Shen, X.-X., Strassert, J. F. H., Kolísko, M., & Brown, M. W. (2021). PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. *PLOS Biology*, 19(8), e3001365. https://doi.org/10.1371/journal.pbio.3001365

Tikhonenkov, D. V., Jamy, M., Borodina, A. S., Belyaev, A. O., Zagumyonnyi, D. G., Prokina, K. I., Mylnikov, A. P., Burki, F., & Karpov, S. A. (2022). On the origin of TSAR: Morphology, diversity and phylogeny of Telonemia. *Open Biology*, 12(3), 210325. https://doi.org/10.1098/rsob.210325

Torruella, G., Derelle, R., Paps, J., Lang, B. F., Roger, A. J., Shalchian-Tabrizi, K., & Ruiz-Trillo, I. (2012). Phylogenetic Relationships within the Opisthokonta Based on Phylogenomic Analyses of Conserved Single-Copy Protein Domains. *Molecular Biology and Evolution*, 29(2), 531–544. https://doi.org/10.1093/molbev/msr185

Unrein, F., Gasol, J. M., Not, F., Forn, I., & Massana, R. (2014). Mixotrophic haptophytes are key bacterial grazers in oligotrophic coastal waters. *The ISME Journal*, 8(1), 164–176. https://doi.org/10.1038/ismej.2013.132

Vanni, C., Schechter, M. S., Acinas, S. G., Barberán, A., Buttigieg, P. L., Casamayor, E. O., Delmont, T. O., Duarte, C. M., Eren, A. M., Finn, R. D., Kottmann, R., Mitchell, A., Sánchez, P., Siren, K., Steinegger, M., Gloeckner, F. O., & Fernàndez-Guerra, A. (2022). Unifying the known and unknown microbial coding sequence space. *ELife*, 11, e67667. https://doi.org/10.7554/eLife.67667

Vaulot, D., Sim, C. W. H., Ong, D., Teo, B., Biwer, C., Jamy, M., & Lopes dos Santos, A. (2022). metaPR2: A database of eukaryotic 18S rRNA metabarcodes with an emphasis on protists. *Molecular Ecology Resources*, n/a(n/a). https://doi.org/10.1111/1755-0998.13674

Venter, J. C. (2004). Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, 304(5667), 66–74. https://doi.org/10.1126/science.1093857

Vorobev, A., Dupouy, M., Carradec, Q., Delmont, T. O., Annamalé, A., Wincker, P., & Pelletier, E. (2020). Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *Genome Research*, 30(4), 647–659. https://doi.org/10.1101/gr.253070.119

Weber, F., Mylnikov, A. P., Jürgens, K., & Wylezich, C. (2017). Culturing Heterotrophic Protists from the Baltic Sea: Mostly the "Usual Suspects" but a Few Novelties as Well. The *Journal of Eukaryotic Microbiology*, 64(2), 153–163. https://doi.org/10.1111/jeu.12347

Williams, P. J. le B., & Ducklow, H. W. (2019). The microbial loop concept: A history, 1930–1974. *Journal of Marine Research*, 77(2), 23–81. https://doi.org/10.1357/002224019828474359

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. https://doi.org/10.1126/science.1257594

Yeh, Y.-C., & Fuhrman, J. A. (2022). Contrasting diversity patterns of prokaryotes and protists over time and depth at the San-Pedro Ocean Time series. *ISME Communications*, 2(1), 1–12. https://doi.org/10.1038/s43705-022-00121-8

Zehr, J. P., & Kudela, R. M. (2011). Nitrogen Cycle of the Open Ocean: From Genes to Ecosystems. *Annual Review of Marine Science*, 3(1), 197–225. https://doi.org/10.1146/annurev-marine-120709-142819

Zhu, F., Massana, R., Not, F., Marie, D., & Vaulot, D. (2005). Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiology Ecology*, 52(1), 79–92. https://doi.org/10.1016/j.femsec.2004.10.006

Zobell, C. E. (1946). Marine microbiology. A monograph on hydrobacteriology. *Chronica Botanica*. https://www.cabdirect.org/cabdirect/abstract/19471100771

Zubkov, M. V., & Tarran, G. A. (2008). High bacterivory by the smallest phytoplankton in the North Atlantic Ocean. *Nature*, 455(7210), 224–226. https://doi.org/10.1038/nature07236

Zuñiga, C., Zaramela, L., & Zengler, K. (2017). Elucidation of complexity and prediction of interactions in microbial communities. *Microbial Biotechnology*, 10(6), 1500–1522. https://doi.org/10.1111/1751-7915.12855

Other works by the author

Canals, O., Obiol, A., Muhovic, I., Vaqué, D., & Massana, R. (2020). Ciliate diversity and distribution across horizontal and vertical scales in the open ocean. *Molecular Ecology*, 29(15), 2824–2839. https://doi.org/10.1111/mec.15528

Jamy, M., Biwer, C., Vaulot, D., Obiol, A., Jing, H., Peura, S., Massana, R., & Burki, F. (2022). Global patterns and rates of habitat transitions across the eukaryotic tree of life. *Nature Ecology & Evolution*. https://doi.org/10.1038/s41559-022-01838-4

Labarre, A., López-Escardó, D., Latorre, F., Leonard, G., Bucchini, F., Obiol, A., Cruaud, C., Sieracki, M. E., Jaillon, O., Wincker, P., Vandepoele, K., Logares, R., & Massana, R. (2021). Comparative genomics reveals new functional insights in uncultured MAST species. *The ISME Journal*, 15(6), 1767–1781. https://doi.org/10.1038/s41396-020-00885-8

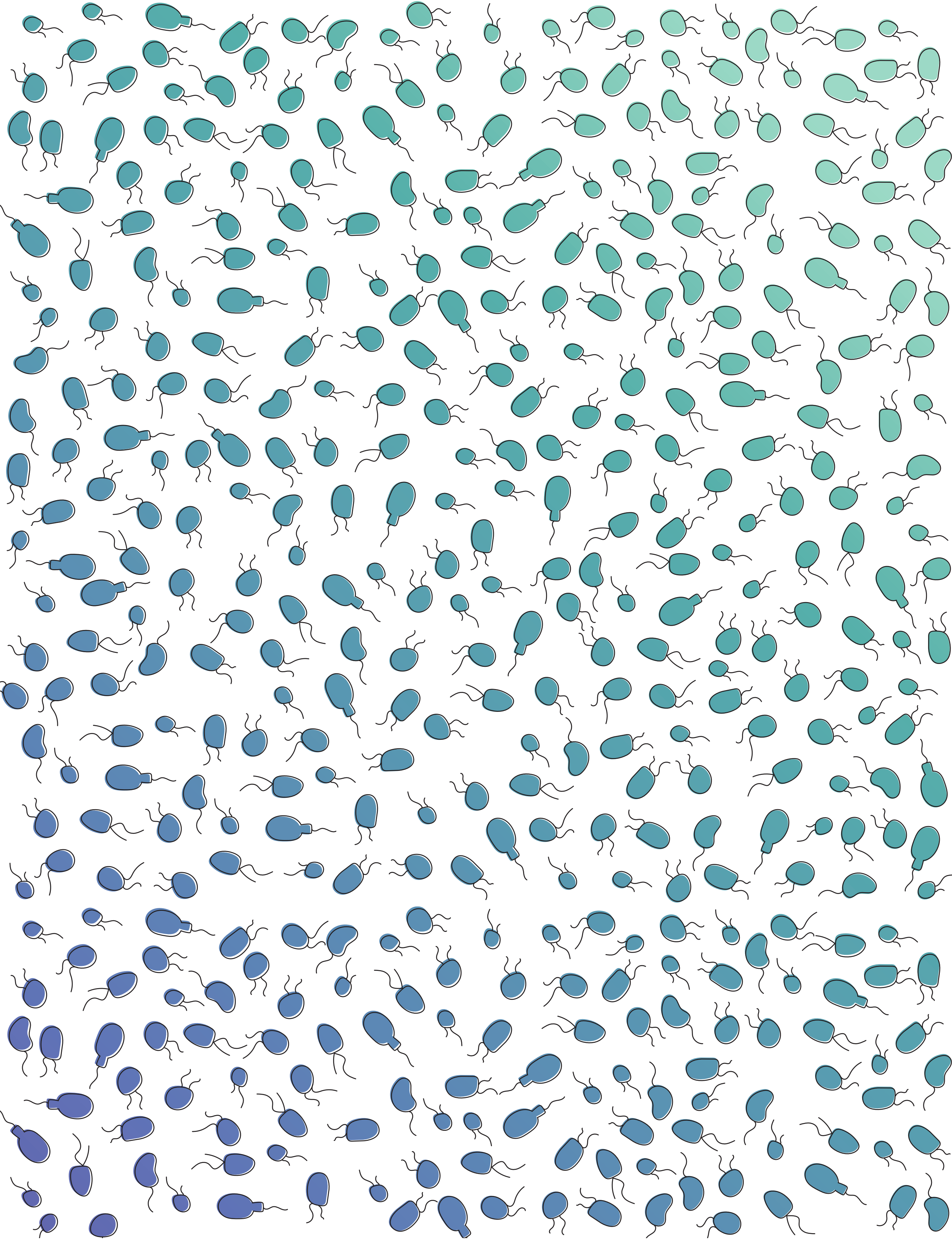
Labarre, A., Obiol, A., Wilken, S., Forn, I., & Massana, R. (2020). Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates. *Limnology and Oceanography*, 65(S1), lno.11379. https://doi.org/10.1002/lno.11379

Latorre, F., Deutschmann, I. M., Labarre, A., Obiol, A., Krabberød, A. K., Pelletier, E., Sieracki, M. E., Cruaud, C., Jaillon, O., Massana, R., & Logares, R. (2021). Niche adaptation promoted the evolutionary diversification of tiny ocean predators. *Proceedings of the National Academy of Sciences*, 118(25). https://doi.org/10.1073/pnas.2020955118

Mangot, J., Forn, I., Obiol, A., & Massana, R. (2018). Constant abundances of ubiquitous uncultured protists in the open sea assessed by automated microscopy. *Environmental Microbiology*, 20(10), 3876–3889. https://doi.org/10.1111/1462-2920.14408

Massana, R., Labarre, A., López-Escardó, D., Obiol, A., Bucchini, F., Hackl, T., Fischer, M. G., Vandepoele, K., Tikhonenkov, D. V., Husnik, F., & Keeling, P. J. (2021). Gene expression during bacterivorous growth of a widespread marine heterotrophic flagellate. *The ISME Journal*, 15(1), 154–167. https://doi.org/10.1038/s41396-020-00770-4

Steiner, P. A., Geijo, J., Fadeev, E., Obiol, A., Sintes, E., Rattei, T., & Herndl, G. J. (2020). Functional Seasonality of Free-Living and Particle-Associated Prokaryotic Communities in the Coastal Adriatic Sea. *Frontiers in Microbiology*, 11. https://doi.org/10.3389/fmicb.2020.584222



Institut de Ciències del Mar

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH