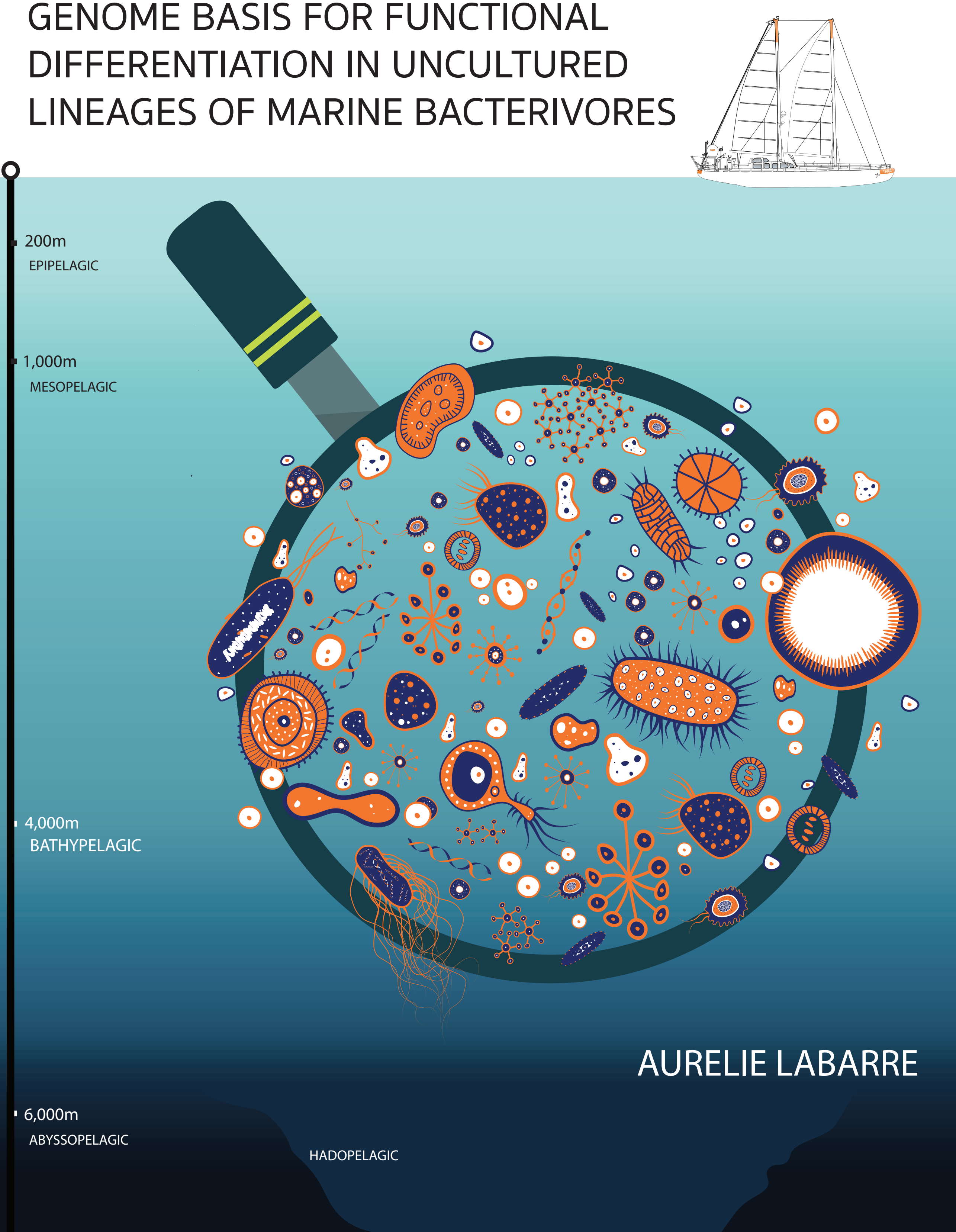
GENOME BASIS FOR FUNCTIONAL DIFFERENTIATION IN UNCULTURED LINEAGES OF MARINE BACTERIVORES



AURELIE LABARRE

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Institut
de Ciències
del Mar

GENOME BASIS FOR FUNCTIONAL DIFFERENTIATION IN UNCULTURED LINEAGES OF MARINE BACTERIVORES

International PhD degree in Marine Sciences

Doctoral thesis to apply for the title of Doctor in Marine Sciences from the department of Civil and Environmental Engineering from the Universitat Politècnica de Catalunya under the direction of Dr. Ramon Massana (ICM-CSIC)

Institut de Ciències del Mar (ICM) Spanish National Research Council (CSIC)

Universitat Politècnica de Catalunya (UPC)

Aurélie Labarre

Ramon Massana (Advisor)

November 26th, 2020

UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Institut de Ciències del Mar

CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

TABLE OF CONTENTS

SUMMARY

In the vast network of the ocean, microbes are abundant and unevenly distributed. As an important microbial component, the protists play a key role in global biogeochemical cycles and contribute to the recycling of nutrients necessary to sustain life on Earth. These unicellular eukaryotes exist and function as primary producers (drivers of photosynthesis), decomposers, parasites or as trophic linkers in aquatic food webs. Phagotrophic species, which acquire nutrition through feeding on other organisms, are commonly understudied due to the difficulty in culturing them. The recent characterization of their genomic and metabolic diversity starts to unveil their great ecological relevance in the oceans. In this dissertation, we focused on heterotrophic flagellates, the main bacterial grazers in marine systems, and especially on the MArine STramenopile (MAST) lineages that display numerous uncultured and, therefore, undefined species. The aim was to elucidate their ecological importance in marine food webs by understanding their presumed trophic strategy: phagocytosis, a process only well characterized in animals as an immune system response.

We first attempted to provide new reference genomes of MAST species using single cell genomic sequencing and a co-assembly approach. Thus, we assembled 15 draft genomes from different MAST lineages, and predicted their gene repertoire with the objective to characterize specific genes related to their trophic strategy. Our comparative genomics analysis indicated that all MAST species were phagotrophs. We then targeted peptidases involved in prey digestion as well as proton pumps for vacuole acidification, but we did not find preferential genes specific for phagocytosis. In addition, this study revealed the relevant presence of rhodopsin proteins that may contribute in the acidification of the phagolysosome.

In the second paper, we did a functional study of MASTs using metatranscriptomics in order to gain access to their gene expression within the natural environment. To do so, we started a grazing experiment with a natural sample from the Mediterranean Sea: in a controlled microcosm in the dark, we followed the cell growth of a natural community where we aimed to enrich for heterotrophic flagellates and therefore phagocytosis. We showed an increase in the relative abundance of heterotrophs, as compared with phototrophs, when phagocytosis occurred. Using the previously established reference genome collection of a few MASTs, we were able to target the MAST reads in the metatranscriptome and analyze the expression profile of genes involved in phagocytosis for a couple of MAST-4 species. Cathepsins and other digestive enzymes were highly expressed when bacterial consumption was observed.

Finally, a similar experiment was conducted with a cultured organism, *Cafeteria burkharda*e, a cosmopolitan heterotrophic flagellate that proved to be a good model to study bacterivory within the Stramenopiles. Results demonstrated distinct expression profiles depending on the growth phase of this species. Upregulated genes at the Exponential phase were related to DNA duplication, transcription, translation, and phagocytosis, whereas upregulated genes in the Stationary phase were involved in signal transduction, cell adhesion and lipid metabolism. Phagocytosis genes, like peptidases and proton pumps, were highly expressed and could be used to target this ecologically relevant process in marine ecosystems.

This thesis contributes to the understanding of the community of marine bacterial grazers, which include the smallest phagotrophs in the ocean, with a focus on their functional behavior within the natural and complex protistan assemblage.

RESUMEN

En la vasta y compleja red del océano, los microbios son abundantes y están desigualmente distribuidos. Como uno de los componentes microbianos importantes, los protistas juegan un papel clave en los ciclos biogeoquímicos globales y contribuyen al reciclado de nutrientes necesarios para mantener la vida en la Tierra. Estos eucariotas unicelulares funcionan como productores primarios (realizando la fotosíntesis), descomponedores, parásitos o conectores tróficos en las redes tróficas acuáticas. Las especies fagotróficas, que adquieren nutrición al alimentarse de otros organismos, han sido poco estudiadas debido a la dificultad de cultivarlas. Sin embargo, la reciente caracterización de su diversidad genómica y metabólica comienza a desvelar su gran relevancia ecológica en los océanos. En esta tesis doctoral, me he centrado en los flagelados heterótrofos, considerados los principales depredadores de bacterias en los sistemas marinos, y especialmente en los linajes MArine STramenopiles (MAST) que muestran numerosas especies no cultivadas y, por lo tanto, indefinidas. El objetivo es dilucidar su importancia ecológica en las redes tróficas marinas mediante la comprensión de su estrategia trófica: la fagocitosis, un proceso bien caracterizado únicamente en animales como una respuesta del sistema inmunológico.

Primero intentamos proporcionar nuevos genomas de referencia de especies MAST utilizando secuenciación genómica de una sola célula (“single cell genomics”) y un enfoque de ensamblaje conjunto. En el primer capítulo preparamos 15 genomas parciales de diferentes linajes MAST y predecimos su repertorio de genes con el objetivo de caracterizar genes específicos relacionados con su estrategia trófica. Nuestro análisis de genómica comparativa indicó que todas las especies de MAST eran fagótrofas. Después nos focalizamos en las peptidasas involucradas en la digestión de las presas, así como en las bombas de protones necesarias para la acidificación de las vacuolas, pero no encontramos genes preferenciales específicos para la mencionada fagocitosis. Asimismo, este estudio reveló la presencia relevante de proteínas de rodopsina que pueden contribuir a la acidificación del fagolisosoma.

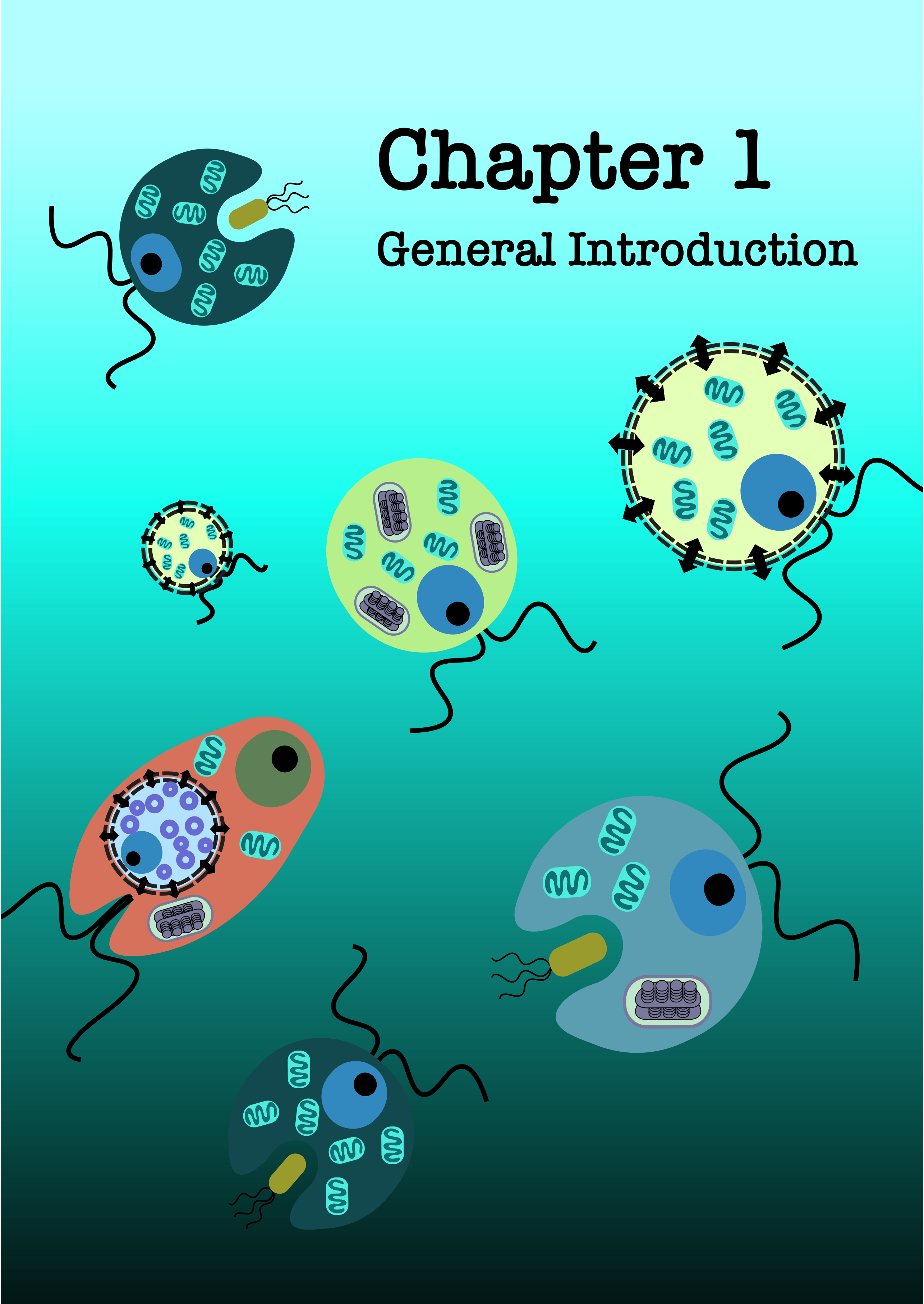
En el segundo artículo realizamos un estudio funcional de los MAST utilizando metatranscriptómica para poder acceder a su expresión génica dentro del entorno natural. Para ello, iniciamos un experimento de bacterivoría con una muestra natural del Mediterráneo: en un enriquecimiento controlado en la oscuridad, seguimos el crecimiento celular de una comunidad natural donde pretendíamos incrementar la abundancia de flagelados heterotróficos y, por tanto, de fagocitosis. Mostramos un aumento en la abundancia relativa de heterótrofos, en comparación con los fotótrofos, cuando ocurrió la fagocitosis. Utilizando la colección de genomas de referencia previamente establecida de algunos MAST, pudimos extraer las secuencias de MAST en el metatranscriptoma y analizar la expresión de genes involucrados en la fagocitosis para un par de especies de MAST-4. Las catepsinas y otras enzimas digestivas fueron altamente expresadas durante el consumo bacteriano.

Finalmente, se llevó a cabo un experimento similar con un organismo cultivado, *Cafeteria burkhardae*, un flagelado heterotrófico cosmopolita que demostró ser un buen modelo para estudiar la bacterivoría dentro de los Estramenópilos. Los resultados mostraron distintos perfiles de expresión génica dependiendo de la fase de crecimiento. Los genes regulados al alza en la fase exponencial estaban relacionados con la duplicación, transcripción, traducción y fagocitosis, mientras que los genes regulados al alza en la fase estacionaria estaban involucrados en la transducción de señales, la adhesión celular y el metabolismo lipídico. Los genes de fagocitosis, como las peptidasas y las bombas de protones, estaban altamente expresados y podrían usarse para abordar este proceso de importancia ecológica para los ecosistemas marinos.

Esta tesis doctoral contribuye a la comprensión de la comunidad de bacterívoros marinos, que incluyen los fagótrofos más pequeños del océano, con un enfoque en su comportamiento funcional dentro de la comunidad compleja de protistas marinos.

Chapter 1

General Introduction



CHAPTER 1. GENERAL INTRODUCTION

1.1- MARINE MICROBIAL ECOLOGY

The oceans form the largest ecosystem on Earth encompassing a range of habitats separated into the photic zone (up to 200 meters depth) and the aphotic zone subdivided in several layers: the mesopelagic (200 to 1000 m, where dim light still penetrates), the bathypelagic (1000 – 4000 m) and the abyssopelagic zone (below 4000 m). Biodiversity in the ocean is considerable, and besides the obvious animal diversity, it includes the existence of marine microorganisms as well. These organisms are exceedingly small (too small to be observed by the unaided naked eye) and constitute the hidden majority of living organisms, with up to a million of them living in just one milliliter of seawater. Despite their microscopic size, marine microbes encompass a complexity and a diversity that rivals any other life on Earth - including Bacteria, Archaea, and Eukaryota (along with their associated viruses). Collectively, they account for more than 98% of the biomass in the ocean (Bar-On and Milo, 2018). Marine microbes are fundamental to all biological and ecological processes in the ocean. They catalyze the metabolic reactions responsible for the biogeochemical cycling of carbon, nitrogen, phosphorus and sulfur. Generating oxygen but also sequestering CO_2, microbes allowed life to develop and to sustain (Worden et al. 2015). The millions of different microorganisms known today have evolved and continue to evolve in the ocean and, despite continuous discoveries, even more remain to be discovered. In order to understand the functioning of our oceans, we need to consider the contribution of marine microbes, especially within plankton communities for which our current knowledge is relatively incomplete.

1.1.1- UNICELLULAR EUKARYOTIC MICROORGANISMS: THE PROTISTS

Protists are ubiquitous components of terrestrial and aquatic environments (Finlay et al. 2002), where they represent a heterogeneous collection of mostly unicellular microscopic eukaryotic organisms. They span three orders of magnitude in size, forming the picoeukaryotes (from 0.2-2 μm), the nanoeukaryotes (from 2-20 μm) and the microeukaryotes (from 20-200 μm). Marine picoeukaryotes are found in all major algal groups (e.g., green algae, Haptophytes, and Stramenopiles) and include many heterotrophic lineages as well. Nanoeukaryotes, include many species of flagellated taxa, together with smaller non-flagellated green algae, diatoms, the smallest dinoflagellates and ciliates (Sherr and Sherr, 2009). Microeukaryotes cover the larger-sized plankton and include mainly diatoms, dinoflagellates, ciliates and radiolarians (Figure 1) (Caron et al. 2012, Massana, 2015).

Protists represent countless morphological variations; most are unicellular, but others group forming filaments, chains, colonies, or coenobia (a specific type of colony). Whilst a few species move by floating, many of them are capable of motility using striking features such as flagella and cilia as their locomotory organelle (these organelles give the name to conspicuous groups - i.e. flagellates and ciliates respectively). Unicellular eukaryotes are not only highly diverse in species richness, but also exhibit a variety of ecological and physiological characteristics. Many protists are phototrophs, producing new biomass from inorganic resources (carbon dioxide and mineral nutrients) via photosynthesis, such as diatoms and dinoflagellates. Others are heterotrophs and rely on other microbes for nutritional intake (using fixed organic carbon sources as substrates). Heterotrophy may occur as phagotrophy, which is essentially the engulfment of particulate food, but also as osmotrophy - taking up dissolved organic matter from the medium as Fungi do (Richards et al. 2012).

Parasitism represents a third type of heterotrophy. Finally, other organisms have the capacity to combine two nutrition modes; they can feed on other microorganisms whilst also fixing carbon photosynthetically. They are known as mixotrophs. In fact, many marine algal groups exhibit this strategy (Flynn et al. 2019).

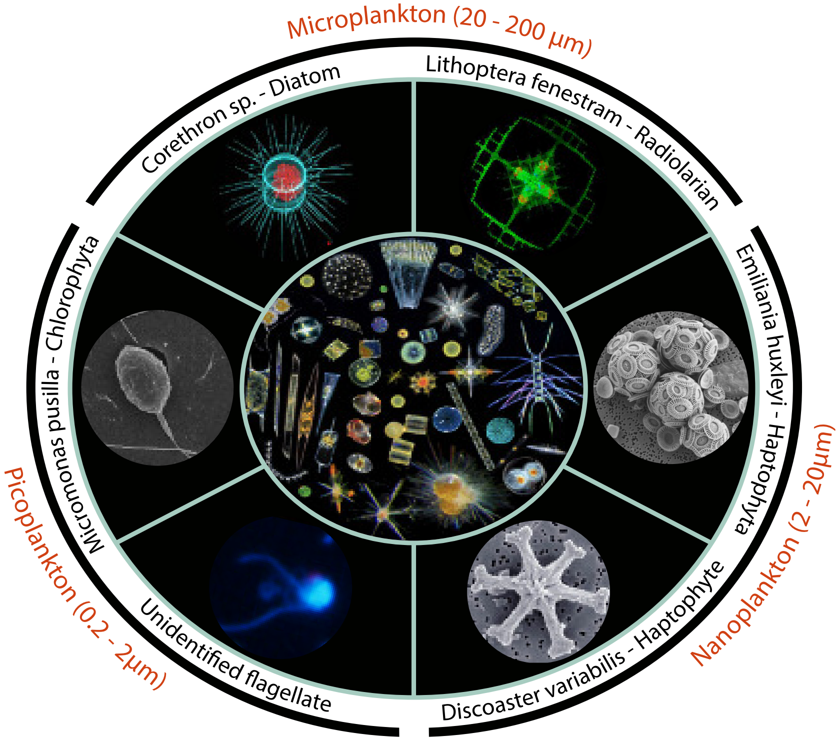

Microplankton (20 - 200 µm)
Corethron sp. - Diatom
Lithoptera fenestram - Radiolarian
Emiliania huxleyi - Haptophyta
Nanoplankton (2 - 20µm)
Discoaster variabilis - Haptophyte
Unidentified flagellate
Picoplankton (0.2 - 2µm)
Micromonas pusilla - Chlorophyta

Figure 1 - Diversity of single-celled eukaryotes. Most of the unicellular species are microscopic. The smallest, known as picoeukaryotes, are up to 2 µm in size whereas the larger microplankton can reach up to 200 µm. In between, the nanoplankton (2-20 µm). (Images courtesy from Sebastien Colin, Michel Flores, Ramon Massana and Christian Sardet).

The term protist has for a long time been a problematic taxonomic unit (Adl et al. 2005, 2007). Indeed, no single or unique feature sets the protists apart as a group or kingdom, and the reason is because they are defined by exclusion; they include all eukaryotic life that is not part of the traditional plant, animal, or fungal domains. Nowadays, protists are known to be a polyphyletic group of organisms (polyphyletic refers to organisms descending from different ancestor) that exhibit representatives in most eukaryotic lineages (Adl et al. 2012, 2018), which are combined together into eight main supergroups (Figure 2).

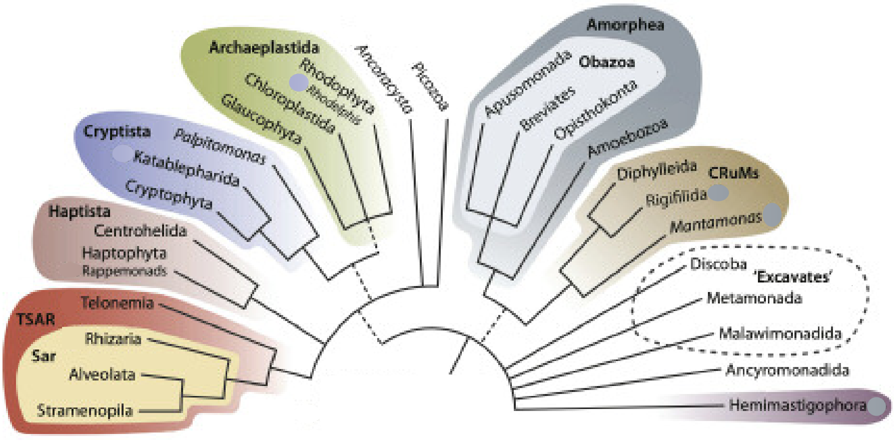


Figure 2 – Eukaryote tree of life. Phylogenetic tree representing the major groups of eukaryotes differentiated by colors. Dashed lines reflect uncertainties about the monophyly of certain groups. Figure adapted from Burki et al. (2019).

'Obazoa' groups the opisthokonta, multicellular animals (Metazoan), Fungi plus Choanoflagellates, together with two lineages of heterotrophic flagellates: the Breviates and the Apusomonada. 'Archaeplastida' unites taxa that have retained green pigments (Chlorophytes and Prasinophytes) or red pigments (Rhodophytes) from the primary endosymbiosis with a cyanobacteria (Falkowski et al. 2004). The group called Cryptista (cryptomonads, katablepharids, and palpitomonads) appears to have a phylogenetic connection with archaeplastids (Burki et al. 2016). The clade

Haptista (Haptophytes and Centrohelids) includes microbes that are crucial for the marine system, illustrated by the famous calcifying coccolithophorid *Emiliania huxleyi* (Haptophyte). The very large clade SAR cluster together the 'Alveolates', the 'Stramenopiles' and 'Rhizarians'. These clades include numerous taxa present in marine ecosystems and comprise about half of all eukaryote species (del Campo et al. 2014). 'Excavates' contains numerous heterotrophic predators, photosynthetic species and parasiites represented by the Discoba, Metamonada and Malawimonadida; however the clade is not resolved and possibly paraphyletic (Burki et al. 2019).

'CRuMs' is a novel described supergroup including previously orphans taxa with different morphologies. These extremely diverse eukaryotic supergroups are assumed to be descended from the ancestral diversification and radiation of the earliest eukaryotic organism LECA (last eukaryotic common ancestor), which first appeared around 1–1.5 Gy ago (O'Malley et al. 2019).

1.1.2- GLOBAL IMPORTANCE AND ECOLOGICAL SIGNIFICANCE

Organisms within a community are bound by a network of interactions. In marine pelagic ecosystems, the most important are the trophic interactions linking photosynthesis and biological productivity to global nutrient cycles - the food chain. Food chains delineate one of the pathways to transfer energy and matter through various trophic levels, impacting on the world's food production, climate and the global carbon cycle. Myriads of food chains within an ecosystem form a food web.

Made of interconnected food chains, the bases of aquatic food webs are formed by the primary producers via photosynthesis (Stoecker et al. 2009). Along with prokaryotic cyanobacteria, eukaryotic phytoplankton such as green algae, haptophytes, diatoms and dinoflagellates are the most common primary producers (Worden et al. 2004). The next trophic levels are heterotrophic consumers that feed on primary producers. When algal cells are relatively large, microzooplankton like dinoflagellates, ciliates and radiolarians can be important consumers of primary production, while pico- and nanosized flagellates appear to be the main grazers of smaller phytoplankters (Calbet, 2008). Responsible for grazing the majority of global primary production (Calbet, 2008), predation by protists is a major mediator of nutrient recycling; more than 90% of organic matter mineralization and nutrient recycling is achieved by microbes smaller than 100 µm. In the pelagic system, microbes also contribute substantially to carbon flux that is transfered down into the twilight and deep zones of the ocean.

Finally, parallel to the carbon export through food webs or down in the ocean by the biological pump, a fraction of carbon fixed by phytoplankton is released as dissolved organic matter (DOM) and recycled via the microbial loop (Pernthaler and Posch, 2009). As a basic resource, DOM is used by bacteria and archaea that are then grazed by hetero- and mixotrophic protists, contributing both to trophic flows and nutrient remineralization (Worden et al. 2015). Thus, heterotrophic phagotrophic metabolism grazing on bacteria represents an important fraction of the ocean’s functioning.

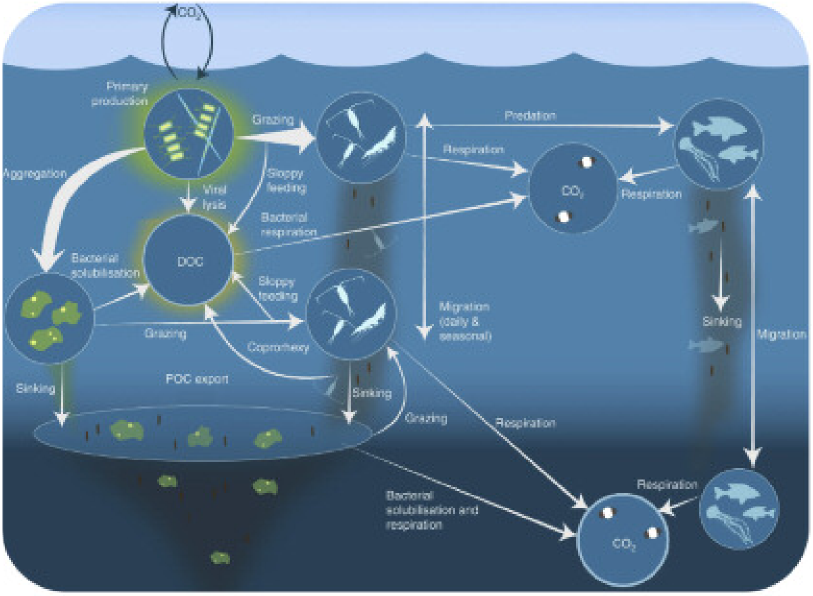


Figure 3 – Conceptual biological processes in the marine food web. During primary production, phytoplankton convert CO2 from the atmosphere into particulate organic carbon (POC) . Phytoplankton are in turn preyed upon by higher trophic levels thereby forming the base of marine food webs. Adapted from Cavan et al. 2019.

1.1.3- PHAGOTROPHIC FLAGELLATES: MARINE STRAMENOPILES (MAST)

An important component of free-living protists is the Heterotrophic Flagellates (HF). Heterotrophic flagellates are unpigmented cells characterized by the possession of one or more flagella, which are long, tapering, hair-like appendages that function as organelles of locomotion, substrate attachment, or for feeding. They are a very heterogeneous group including organisms smaller than 2 µm up to larger than 15 µm (Arndt et al. 2000). Very abundant in the ocean and routinely enumerated (Christaki et al. 2011), they are found from the pelagial areas to the deep sea (Gooday et al. 2020). A common example would be *Cafeteria burkhardae*, described in several abyssals as well as in global analyses of planktonic communities (de Vargas et al. 2015; Schoenle et al. 2020, Chapter 4). Known HF species belong to multiple taxonomic groups such as the choanoflagellates, chrysophytes, kinetoplastids, diplomonads, and bicoecids, but the true extent of the species composition in natural assemblages is poorly determined. Together, HFs are known as the most important bacterial grazers, responsible for more than 60% of bacterial mortality (Sherr and Sherr, 2002; Calbet and Landry, 2004), with some particular groups having preferred species as prey (Verity, 1991; Matz et al. 2002). By grazing on bacteria and also on small phytoplankton, HFs release essential elements necessary for the growth of other phytoplankton (Sherr & Sherr 2002).

Despite their crucial role in marine habitats (Pernthaler, 2005), our understanding of the species forming the small-sized heterotrophic flagellates (2-5 µm in size) is still limited, mainly due to methodological limitations. Many obligate phagotrophic flagellates belong to the Stramenopile lineage. Stramenopiles are part of the SAR supergroup that also includes Alveolata, and Rhizaria (Burki 2014, Grattepanche et al. 2018); they are one of the major established eukaryotic assemblages (Cavalier-Smith, 1986). Stramenopiles encompass a very large diversity of organisms, from large multicellular to tiny unicellular species, and they are present in every kind of environment (e.g. marine, freshwater and terrestrial). Their unifying feature is the presence of two distinct flagella, one anteriorly-directed flagellum with tripartite hairs (mastigonemes) and another smooth posterior flagellum used to propel and lead the swimming direction. Grouping numerous photosynthetic taxa into the monophyletic cluster Ochrophyta (Grattepanche et al. 2018) (with the exception of some heterotrophic taxa such as Paraphysomonas), the Stramenopile radiation contains many non-photosynthetic (heterotrophic) lineages (Yubuki et al. 2010) in several clades that branch before the stem lineage divergence of Ochrophytes.

A large component of the stramenopile radiation are the uncultured MArine STramenopiles (MASTs). Identified in abundance in surface marine waters (Massana et al. 2004, 2006), some MAST clades have been shown to be free-living bacterivorous HFs (Massana et al. 2009). Widely distributed, they account for a large fraction (up to 35%) of the HFs in diverse geographic regions (Rodríguez-Martínez et al. 2009). Placed in different phylogenetic regions across the Stramenopiles, eighteen MAST clades have been currently identified and labeled (Massana et al. 2014) (Figure 4).

Because these clades are placed in different phylogenetic positions of the Stramenopile radiation, which include phototrophs, mixotrophs, osmotrophs, phagotrophs and parasites (Andersen 2004, Derelle et al. 2016), the cellular identity and general trophic mode of the MAST clades is still unclear. Partial data exist for some clades; e.g. MAST-3 contains parasites (Gómez et al. 2011), and MAST-1 and MAST-4 contain active bacterivores (Massana et al. 2009), but this essential knowledge is still unknown for many of the other existing MAST clades. As they diverged before the clade of photosynthetic Ochrophyta, the new MAST lineages are expected to be key to understand the early evolutionary history of Stramenopiles.

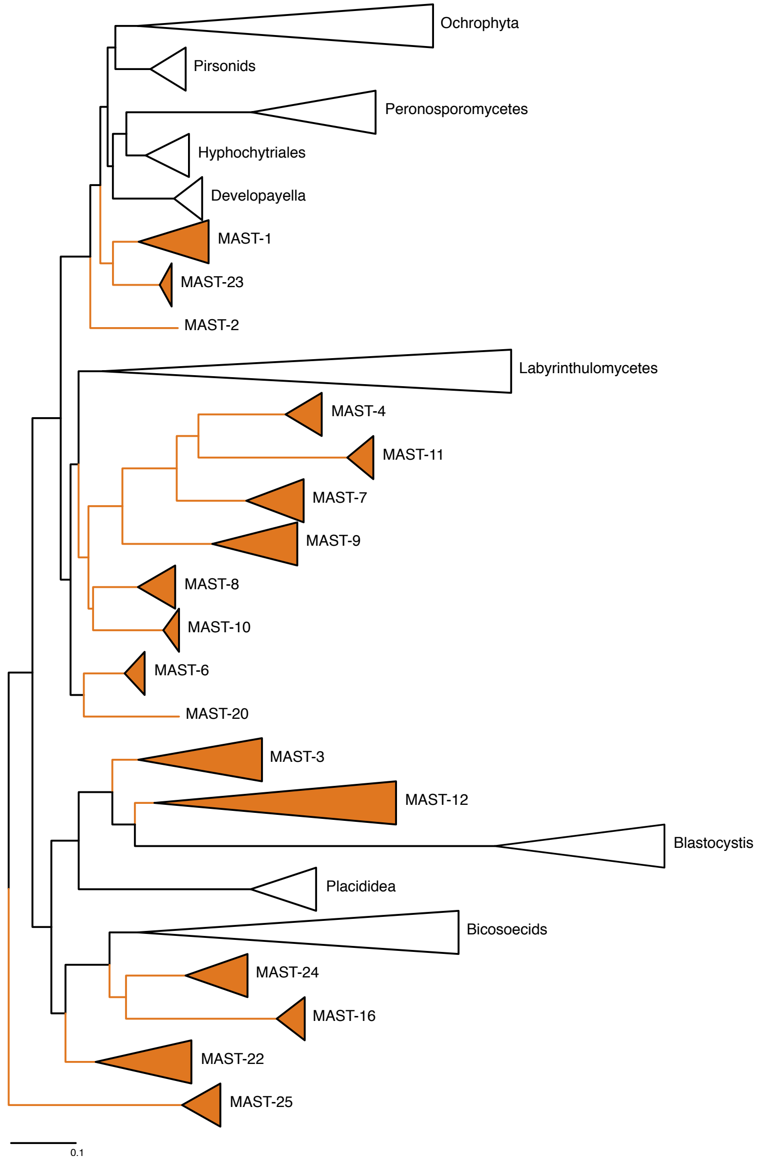


Figure 4 – Schematic phylogenetic tree of MAST clades. Representation of the 18 lineages of Marine Stramenopiles inferred from 18S rDNA sequences. Adapted from Massana et al. (2014).

1.1.4- GAPS IN OUR KNOWLEDGE ABOUT EUKARYOTIC DIVERSITY AND FUNCTION

Protist constitute the majority of phylogenetic lineages in the eukaryote tree of life and yet, our current knowledge of protistan diversity remains surprisingly limited. Indeed, the protistan component of biological communities across a broad scale remains relatively unexplored (Caron et al. 2009; Pawlowski et al. 2012). Part of this originates from the fact that in many cases species identification relies on morphology, and therefore traditional research has focused mostly on animal, plant, and fungal model species (del Campo et al. 2014). Many protists have cell sizes from 2 to 5 µm (e.g. heterotrophic nanoflagellates, small algae or amoeba), and lack distinct morphological features to allow taxonomic identification. Thus, described protist species represent a small fraction of what has been evaluated (close to 150,000 eukaryotic species have been estimated (de Vargas et al. 2015)). This major gap in eukaryotic diversity exists even more due to environmental sampling being limited to a very few geographic regions (del Campo et al. 2018). In fact, insights into the diversity and function of microorganisms have mainly been based on studies from prokaryotic communities (Keeling and del Campo, 2017). Comparable research on microbial eukaryotes lags behind, and protists often remain overlooked in biodiversity surveys.

Another major reason to explain the poor knowledge on protists is the inability to culture many of the existing species. A culture makes it possible to extract high amounts of DNA, allowing direct genomic sequencing, and an access to specific genomic regions. It also allows proper ecophysiological characterizations of the species. Often, free-living protists are small (pico- and nanoeukaryotes), may be rare, and we do not know their growth requirements. Hence it is very difficult to isolate them and consequently to sequence their genomes by conventional approaches that require large amounts of DNA. Therefore, preferred studies of cultured organisms create a gap in protist genomics. Moreover, as it is easier to culture phototrophic species, the eukaryote groups that are well studied are mainly autotrophs and therefore outstrips our understanding of heterotrophs, which perhaps represent the most abundant forms of microbial eukaryotes (del Campo et al. 2014).

From an evolutionary perspective, missing species can be problematic when specific questions such as defining the origin of eukaryotes using phylogenies are addressed. To infer the position of lineages that are deeply rooted in the eukaryotic tree, phylogenomic analyses with multiple concatenated gene alignments are needed (Lax et al. 2018; Strassert et al. 2019), but missing taxa can produce ambiguous and unstable topologies. The best approach to make progress in finding the position for the true root in the eukaryotic tree is to generate more genomes covering environmental protists (Sibbald and Archibald, 2017).

Studying the ecology of microbial eukaryotes requires molecular tools that complement morphological observations. DNA-based taxonomy made a major breakthrough in marine microbial diversity at the dawn of the 21st century. The basis of this method consists in extracting DNA from a natural community and amplifying one or multiple genes (i.e. genetic barcoding with PCR). This method allowed the characterization of numerous uncultured and unappreciated organisms from several lineages. A few examples are the bacterivorous MASTs (Marine Stramenopiles) (Massana et al. 2004, Not et al. 2009), the parasitic MALVs (Marine Alveolates) (López-García et al. 2001) and the recent discovery of diplonemids (marine heterotrophs) from Discicristata for which very few species had been described (Gawryluk et al. 2016; Tashyreva et al. 2018).

1.2- THE RISE OF GENOMICS

A genome is an organism’s complete set of genetic instructions necessary for that organism to grow and function. In extant eukaryotic organisms, the genome is most often linear and stored in long molecules of DNA (deoxyribonucleic acid) in a double helix structure. Embedded in Nucleosome-complex, DNA and the proteins histones are packed together to form chromosomes. A major feature that distinguishes the genomes of eukaryotes is the division of genes into protein-coding exons and non-coding introns, and the presence of often large quantities of repetitive non-genic DNA. In molecular terms, a gene can be defined as a segment of DNA that is expressed to yield a functional product, being a protein or a regulatory RNA molecule. Genomes of eukaryotic cells contain not only functional genes but also large amounts of DNA sequences that do not code for proteins or regulatory RNAs, defining the dynamic picture of the eukaryotic genome (Parfrey et al. 2008).

Genome sequencing is the process of determing the nucleic acid sequence – the exact order of the four bases (Adenine, Guanine, Thymine and Cytosine). Sequencing technologies fragment the genome prior to sequencing, and each sequenced fragment produces a ‘read’. The complete genome has to be deduced from these short reads by a series of overlapping steps, known as *de novo* genome assembly.
The first DNA fragment sequenced , from the yeast Saccharomyces cerevisiae, happened in 1965 (Figure 5). This was followed by several short regions of various phages and the first whole genome of a virus, namely bacteriophage ΦX174 (Sanger et al. 1977). Starting from the 1980s, Sanger-based shotgun sequencing flourished, and projects to sequence model organisms such as the bacterium *Escherichia coli* or *Caenorhabditis elegans* began all around the world (Figure 5). Rapidly, came the new generation sequencing (NGS) that relied on library preparation using native or amplified DNA. Based on considerable advances in technology, NGS allowed the assembly of draft genomes for most eukaryotic model species (Figure 5) and opened up new parallel areas, such as RNA-seq (high-throughput RNA sequencing) or ChIPseq (chromatin immunoprecipitation). At this time, follow-ups from the human genome sequencing, released in 2001, were becoming popular (Levy et al. 2007, Wheeler et al. 2008) and larger scale projects (e.g. Trust UK10K in 2010 and the All

of Us in 2015) yielded thousands of new sequenced eukaryotic genomes. From the 2010s, a third-generation sequencing (TGS) was born allowing the sequencing of single DNA molecules without amplification. This technology produces longer reads and provides a more uniform coverage of the genome; a great advantage to detect overlaps between reads and therefore generate better-quality assemblies, including the proper sequencing of repeated regions that were missing from NGS-based assemblies.

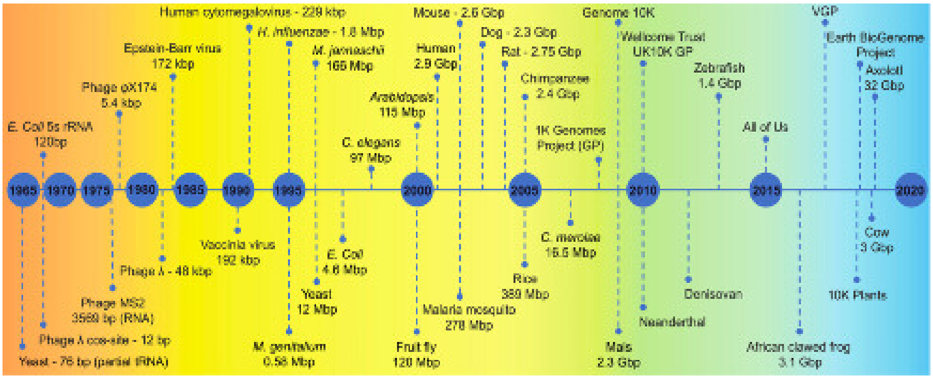


Figure 5. Timeline representation of genomics events. The graph shows the main areas in the history of sequencing. In orange are the first sequencing attempts, yellow represents Sanger-based shotgun sequencing, green NGS (Next Generation Sequencing) and blue TGS (third Generation Sequencing). Image from Giani et al. (2020).

1.2.1- PROTIST GENOMICS

To date, the genomic revolution has been limited on a subset of eukaryotes, as sequencing efforts have overwhelmingly focused on plant lineages, opisthokonts (animals and fungi) and their parasites (Dawson and Fritz-Laylin 2009). Therefore, the immense diversity of microbial free-living protists is not reflected in the currently available genome projects. Some of the reasons were discussed in section 1.1.4. In addition, unicellular eukaryotes were overlooked because of the assumption that their genomes were extremely large. However, the size of protist genomes can be close to large-sized bacterial genomes (~10 Mb), and often is between 50 and 100 Mb. Dinoflagellates, with genome sizes estimated from 3000 to 215,000 Mb, are the exception (Hackett et al. 2004). One of the critical challenges of protist genomics concerns their growing in the laboratory. It is very difficult to obtain a pure culture (especially for heterotrophs), which is an obstacle for genome sequencing. Today, protist genomes account for a small part of all eukaryotic studies (Figure 6). Richter et al. (2020) have combined every genome and transcriptome project into a single site, the EukProt database that contains around 742 eukaryotic species.

A few protists are established as model organisms (e.g: taxa that can be studied and manipulated in controlled conditions to answer defined experimental questions). Examples are the chlorophytes *Acetabularia* and *Chlamydomona*s, the ciliates *Paramecium* or *Tetrahymena*, and the amoebozoan *Dictyostelium* (Kuspa et al. 2001). Another fairly well-developed model is the apicomplexan *Plasmodium*. Most genomes that have been sequenced were chosen for special interests such as the parasite diplomonad *Spironucleus vortens* (18 Mb), a close relative to the well know parasite *Giardia intestinalis* that significantly impacts human health.

Recently, a strong interest in marine protists has encouraged their development as model organisms as well (Waller et al. 2018; Collier and Rest, 2019). This has allowed the expansion of genetically tractable models using some marine protists species, promoting the power of genetic approaches for studying marine microbial processes. The rhizarian *Vampyrellid trophozoites,* for example, has received great attention for its particular feeding behavior; it perforates the cell walls of its chlorophyte prey and extracts the prey content by phagocytosis (More et al. 2019). *Thalassiosira pseudonana* is another example of a marine unicellular model organism. This photosynthetic algae was the first genome of a diatom to be assembled (Armbrust et al. 2004). As a model of heterotrophy, the dinoflagellate *Oxyrrhis marina* has also been sequenced*.* Collective work has, therefore, mostly been applied to diatoms (Bacillariophyceae) and “core” dinoflagellates (Dinophyceae) due to their abundance, diversity, and ecological importance. However, most major lineages of eukaryotes, including some that are very important ecologically as well, are still lacking representative model organisms, including many taxonomic classes within Euglenozoa, the Stramenopiles and Haptophytes (Collier and Rest, 2019).

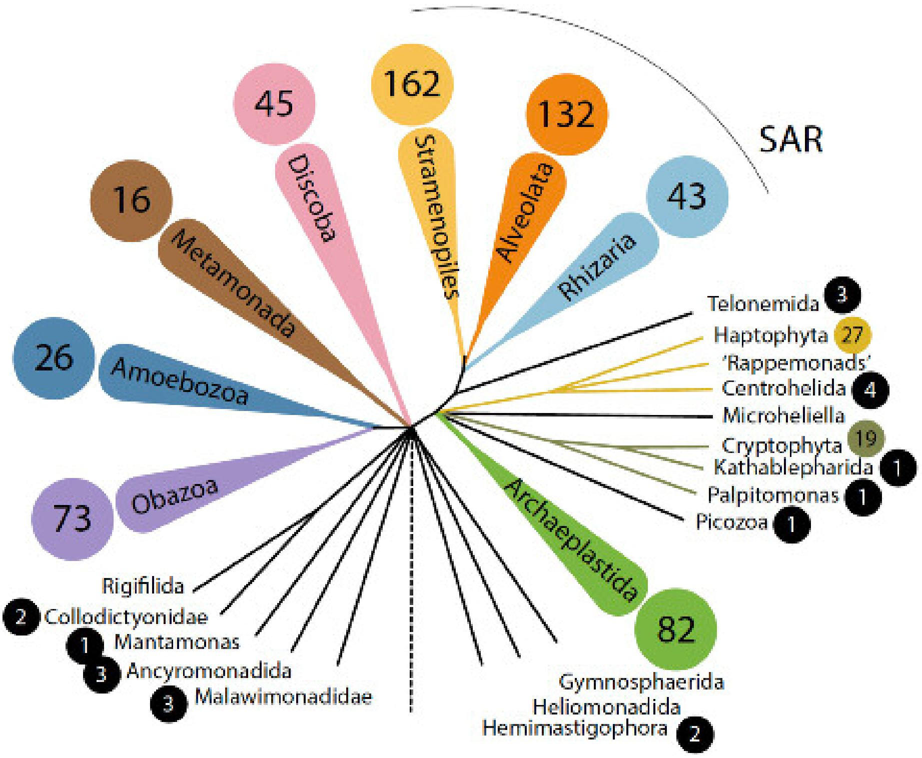


Figure 6 – Representation of protist genomes in the Eukaryotic tree of life. This tree is based on a consensus of recent studies adapted from Simpson et al. 2017. Protists have sequenced representatives in all major evolutionary lineages, where the numbers represent the corresponding available genomes based on Richter et al. (2020).

1.2.2- ANSWERS FROM GENOMICS

Over the past decade, the extent of new genomic information has generated major progress in the understanding of ancestral eukaryotic features, as well as eukaryotic diversification. The incredible potential of the eukaryotic conserved genes has allowed multi-gene phylogenomic studies to help refining the eukaryotic tree of life (Keeling et al. 2005; Burki et al. 2019) and to gain insights into deeply phylogenetically divergent lineages (Adl et al. 2018). Genomics was a starting point for the essential question of eukaryotic evolution, and especially the acquisition of organelles through endosymbiosis and their evolution (Archibald, 2015). Clarifying the phylogenetic relationships allowed us to address general evolutionary concepts such as the emergence of multicellularity from protist lineages (King, 2004). For example, some of the hundreds of gene families initially identified to be animal-specific have been recently also discovered in the common ancestor of metazoans and choanoflagellates (Richter et al. 2018).

Based on gene content, critical advancement has also occured in the understanding of the eukaryotic cell's functioning along with more general ecological concepts like their trophic strategies (Burns et al., 2015, 2018). Several experiments with various protist species (considered as model organisms (Li and Montagnes, 2015) for genomics projects (Montagnes et al. 2012)) have been used to study predation. This includes experiments with cultured heterotrophs (Lee et al. 2014) and directly with natural assemblages to reach uncultured HFs as well. Weber et al. (2012) used experimental work stimulating growth and ingestion in an "unamended" incubation. The potential of microcosm experiments to investigate general concepts in community ecology and evolutionary biology has recently been reviewed (Altermatt et al. 2015). Much progress has been achieved in the understanding of photosynthesis. Fundamental insights into the structure, function, and regulation of the photosynthetic apparatus came from studies with the unicellular algae *Chlamydomonas reinhardtii* (Dent et al. 2005). Moreover, genomic approaches proved to be powerful for identifying other specific eukaryotic features such as meiosis (a stage of sexual reproduction), which evolved early in eukaryotes (Speijer et al. 2015) and for which a full set of responsible genes was proposed, mostly defined from Opistokonta and plants. Using the proposed meiosis-toolkit, it has been possible to perform large-scale comparative analysis and search for these genes in other protists like diatoms (Ramesh et al. 2005, Malik et al. 2008, Patil et al. 2015, Hofstatter et al. 2020). Similarly, genomics supported the discovery of flagellum specific proteins (flagellum toolkit) in Opistokonta (Torruella et al. 2015) or Fungi (Leonard et al. 2018), helping to elucidate the evolutionary history of this ancestral feature specific of a given lifestyle.

1.3- SEQUENCING DATA FROM MARINE PROTISTS ASSEMBLAGES

Genomic regions can also be sequenced directly from an environmental sample, i.e. without cultivation; this is also referred to as 'environmental' or 'community' genomics. Before conducting metagenomic studies, it was mandatory to determine the community composition of natural communities, and this was done by amplifying and sequencing a single taxonomic marker gene. This marker gene had to be conserved across all species within the community and with enough variability to distinguish between the existent taxa. For eukaryotic diversity, the most used marker gene is the small subunit ribosomal DNA (SSU rDNA) (Stoeck et al. 2010). Curated specific SSU rDNA sequences are combined in comprehensive databases for protist identification. One of the first gene-catalogue databases was the SILVA database (Quast et al. 2012), which groups ribosomal RNA sequence data from the three domains of life (Bacteria, Archaea and Eukarya). Later, a eukaryotic specific SSU rDNA database was created, the continually updated PR2 database (Guillou et al. 2012). This was followed by the 'EukRep' project. Seminal diversity surveys used Sanger sequencing of the whole SSU rDNA gene and presented the sequences one by one. The advancement in sequencing technologies, enabling millions of DNA molecules to be sequenced on a daily basis, allowed massively parallel sequencing - offering larger throughput (gigabases of reads) than the conventional Sanger sequencing approach (Metzker, 2010). This had two

implications for microbial diversity surveys. The first was that the short reads required to focus in just a region of the complete gene, being the hypervariable V4 and V9 regions the most popular. The second is that sequences had to be grouped before being reported. Unsupervised bioinformatics methods have helped to detect and cluster together markers that are highly similar (Mahé et al. 2015), which can classify sequences at the species level (DNA based-taxonomy) and are called Operational Taxonomic Units (OTUs). Recently, amplicon sequence variant (ASVs) have replaced OTUs (Callahan et al. 2017), demonstrating specificity and sensitivity to better discriminate ecological pattern (Needham et al. 2017).
Suddenly, the rare biosphere has become accessible (Sogin et al. 2006, Kilias et al. 2014) and environment samples from the world's oceans (de Vargas et al. 2015; Giner et al. 2020) are now accessible. Moreover, the most critical output of NGS tools is that genomics of natural protists become feasible (see later).

Large-scale sequencing approaches present their own set of *in silico* and computational/bioinformatics challenges. As DNA sequencing has greatly accelerated the rate of data generation, new difficulties have emerged at the stages of data processing, analysis, and interpretation (Ward et al. 2013). A first difficulty is the storage of large amounts of data. Significant efforts have been made to reduce the sequencing data sets that are produced in text formats (FASTQ and FASTA) by converting them into binary (Sequence Read Archive, BAM, CRAM, etc). A second challenge is the choice of the optimal sequencing platform, as each one offers distinct trade-offs in speed, throughput, read lengths, error rates and bias. Finally, assembly is one of the critical steps in the environmental samples analysis. In the case of species of interest that are significantly underrepresented in existing databases, longer reads are essential for the *de novo* assembly.

1.3.1- METAGENOMICS AND METATRANSCRIPTOMICS

Metagenomics (DNA-based) uses a similar approach as genomics but differs mainly in the nature of the samples. Genomics focuses on a single organism, whereas metagenomics is an approach to explore the whole microbial community in its natural habitat. The principal goal of metagenomics is to sequence the genomes of untargeted cells in a community in order to elucidate community composition and function. Thus, the entire DNA content of all cells from the community is extracted directly from environmental samples, without isolating or identifying individual organisms and regardless of the abundance of microbial entities. The starting material for metagenomics is a community DNA extract that includes bacterial, archaeal, eukaryotic, and viral species and at different abundances. Once the whole DNA is extracted (in sufficient quantity and quality), library construction is the next important step. Several library construction methods have been developed but they generally comprise three steps: random DNA fragmentation into smaller molecules, repairing and end-polishing of fragmented DNA, and ligation of specific adaptors at the two ends (van Dijk et al. 2014). Compared to the first-generation sequencing, NGS can generate several hundred thousand to millions of sequencing reads in parallel. Several next-generation sequencing platforms have been introduced, including Pyrosequencing (Roche 454), Illumina, Applied Biosystems SOLiD, and Ion Torrent. More recently, the PACBIO is capable of generating very long reads without the need to clone the fragments to amplify the signal. All next-generation sequencing utilizes optical sensors that detect luminescent signals, which are produced during incorporation in the sequence of bases with fluorescent tagged dinucleotides. Nanopore technology works in a different way and use a synthetic membrane bathed in an electrophysiological solution. An ionic current drives DNA strands through the Nanopore, where nucleobases cause a disruption in the current. This change allows sequences to be read out.

Whilst metagenomic studies indicate the genomic content and identification of microbes present within a community using DNA, metatranscriptomics focuses on the genes expressed by sequencing the community mRNA. It is able to distinguish the active from inactive members, and supports investigation of the whole gene expression profile within a community. In this sense, metatranscriptomics complements metagenomic information. Metatranscriptomics typically starts with the isolation of mRNA, which can be selected by synthesizing cDNA with random hexamers and using oligo-d(T) primers that take advantage of the poly-A tail characterizing eukaryotic mRNA. The use of random hexamers in the reverse transcription allows the detection of novel taxa, which would be missed when using designed primers towards known conserved regions. What is produced is an enriched population of mRNAs representative of transcriptionally active genes. Once fractioned, the cDNA is subsequently ligated with a DNA adaptor (sequencing adapters) to the 3′ end. Although current metatranscriptomic techniques are promising, there are still several drawbacks that can limit their application. For example, mRNA is unstable and has a short 'life'; experimental design is therefore challenging as the collection of sufficient material for sequencing needs to be as fast as possible to minimize mRNA losses. In addition, a large part of the harvested RNA comes from ribosomal RNA (rRNA) which can dramatically reduce the coverage of any mRNA retrieved.

Metagenomic and metatranscriptomic reads are a challenge to analyze and can be difficult to assemble, especially for protist genomes (Keeling et al. 2014) but they were one of the first innovations that gave access to the functional gene composition of microbial communities. Meta-omics have played a role in the discovery of novel genes; providing more complete descriptions than phylogenetic surveys that were supported by the diversity of only one gene, as the SSU rDNA. Also, it participated in the discovery of novel metabolic pathways.

1.3.2- SINGLE CELL GENOMICS

The Single Cell Genomic (SCG) approach provides *de novo* genomic sequence data with a single cell as input, compared to metagenomics that uses the whole community. This method provides a unique opportunity to analyze whole genome information, and possible interactions, at the resolution of an individual cell without the need for cultivation (Yilmaz and Singh, 2012). Therefore, SGC had a profound impact on our understanding of new eukaryotic lineages that were not accessible in the past due to their unculturable nature.

SGC technology starts with cell isolation. There are various approaches (Figure 7) for isolating single cells from a suspension: Manual isolation - either using specialized pipettes or micromanipulation equipment, Microfluidic technologies – supporting the manipulation of small volumes of fluids on a microscopic level built onto microchips, and Fluorescent activated cell sorting (FACs) - allowing the separation of live heterogeneous mixtures (natural communities) into sub-population of cells, employing a flow cytometer. In FACS, the stream of single cells is pushed through a nozzle creating droplets where a flow cytometer excites the cell-bound fluorophores (or intrinsic fluorescence) causing light scattering and fluorescent emissions. The fluorescent colors (i.e. the different wavelengths produced) and scattering properties of the droplets are recorded and converted into an electronic pulse that assigns a charge to the droplets. Based on their charge, each droplet is either selected or falls into a waste chamber. Followed by cell lysis of single cells, and due to the low number of nucleic acid molecules, whole genome amplification (WGA) is a prerequisite. Typically, this is done by multiple displacement amplification (MDA), which uses random primers and Phi29 DNA polymerase. Amplified DNA is then screened to target the SSU rDNA gene allowing to chose particular single cells for library preparation and genome sequencing.

Although SCG has mostly been used to investigate cancer and other diseases in human cells (Kamies et al. 2020), this technology has successfully provided a few protist genomes (Yoon et al. 2011; Roy et al. 2014; Gawryluk et al. 2016, Mangot et al. 2017, López-Escardó et al. 2017, Strassert et al. 2018) allowing access to the genetic information of uncultivated microbial eukaryotes. In addition, single cell genomics has also allowed to study mitochondrial genomes of uncultured heterotrophic flagellates (Wideman et al. 2020), suggesting another path for taxonomy identification. However, errors are often introduced during MDA amplification (Pinard et al. 2006; Podar et al. 2009), such as base mis-incorporation, insertions, deletions or the formation of chimeras. Another common type of flaw is the "preferential amplification" of some regions over others, leading to non-uniform sequencing read depth (Yilmaz and Singh 2012). This influences genome recovery, which can be rather low.

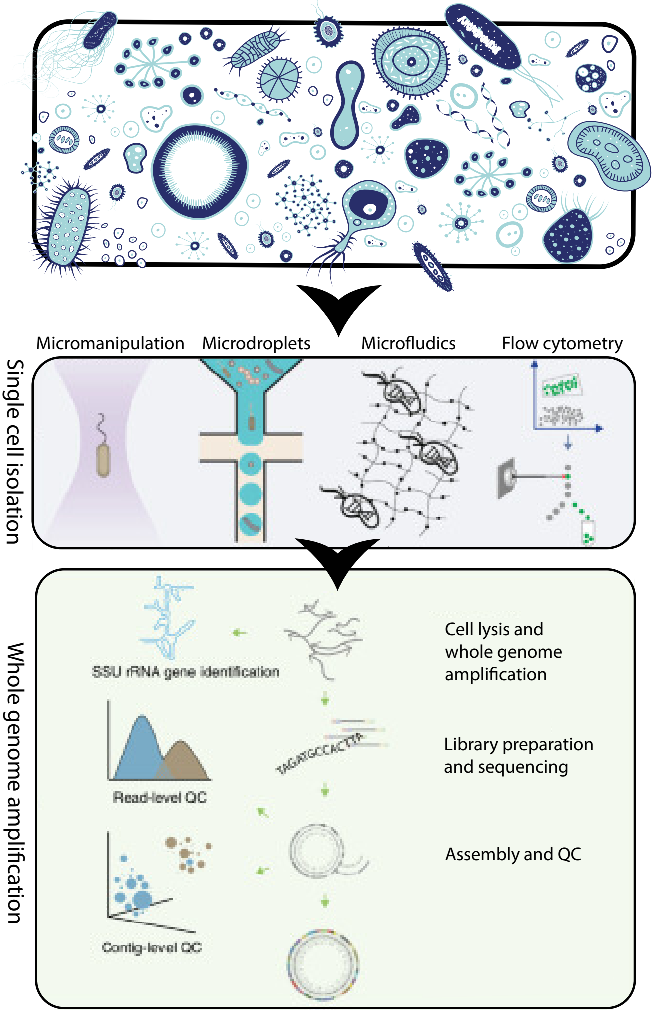


Figure 7 – Steps in single cell genomics. The first step is to isolate individual cells (blue panel). Technologies include microfluidics, micromanipulation or fluorescence-activated cell sorting (FACS). In the second step, whole genome amplification is necessary prior the library preparation, sequencing and assembling (green panel). Modified from Woyke et al. (2017).

1.3.3- GLOBAL OCEANOGRAPHIC SURVEYS: A COMPREHENSIVE PLANKTON SAMPLING

A large fraction of the microbial diversity in the pelagic system still remains unknown, and consequently our knowledge on the global functioning of the oceans is limited. Large-scale ocean explorations have largely participated in the first steps towards understanding the role of the ocean in global biogeochemical cycles and revealing the ocean's invisible abundances. Hence, "bulk-sampling" approaches aim to target complete communities of organisms in their natural environment. An initial expedition was conducted by the *Sorcerer II expeditions* (2003-2010) (Rusch, 2007), and made possible the first large collection of samples, yielding 7.7 million sequencing reads from the North Atlantic to the South Pacific. Two European global oceanographic surveys followed: the *Tara Oceans Expedition* (2009-2013) and the *Malaspina expedition* (2010-2011) (Laursen, 2011). Both studying the biology, chemistry and physics of the oceans from the surface layer to deep waters.

The *Tara Oceans Expedition* was a French non-profit effort that occured from September 2009 to November 2012. During these three years the expedition carried out global surveys to attempt the first global study of marine plankton (protists, bacteria, viruses and small metazoans). Collecting a wide variety of planktonic microbial organisms in different size fractions (about 35,000 samples) they aimed to provide an extensive biodiversity picture of surface (0-200 m) and mesopelagic layers (200-1000 m) (Alberti et al. 2017) applying multi-disciplinary methods. Plankton assemblages were collected at discrete depths using advanced techniques for offshore sampling. Essentially, devices were plankton nets, a high-volume peristaltic pump for water filtration, and a Rosette vertical sampling system (including Niskin bottles) to assess the structure and functions of an entire ecological system at specified depths. In this context, the expedition allowed the sampling in 210 different locations (Figure 8) and made possible the study of non-model organisms.

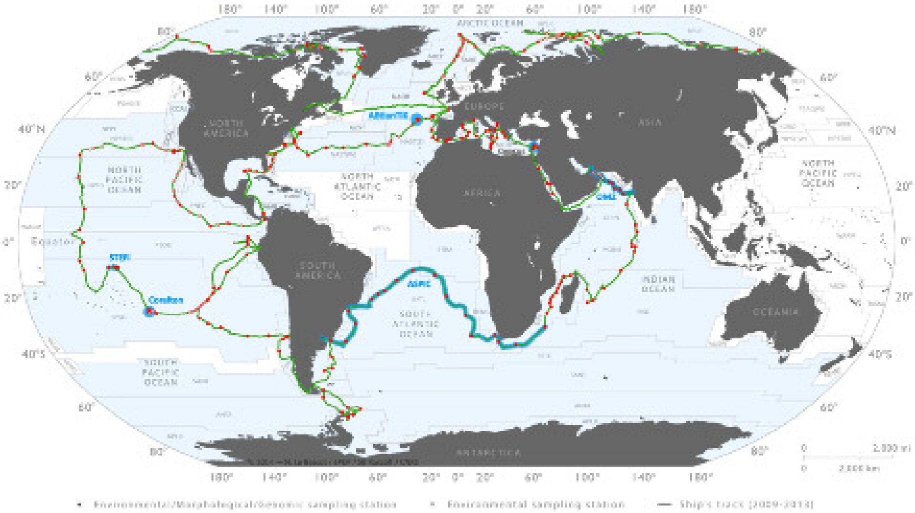


Figure 8 - Sampling road of the TARA Oceans Expedition. The green line shows the cruise track and the red dots are the sampled stations in contrasting ecosystems of the world oceans. From Pesant et al. (2015).

Malaspina was led by the Spanish National Research Council from December 2010 to July 2011. Aboard the ship *Hesperides*, samples were collected across the Atlantic Ocean towards the Pacific Ocean and the Indian Ocean (Figure 9). Water samples were collected in more than 180 stations at 7 different depths per station, from the photic zone to the bathypelagic layer (up to 4000 m). Mainly, the expedition involved measuring temperature, salinity and nutrient concentration in the different ocean regions, studying the exchange of gases, and determining the fate of CO_2 absorbed by the sea. The expedition also explored the diversity and metabolism of phytoplankton and zooplankton at every depth, with a stronger focus on the smallest microbial fractions, where samples for both biodiversity and metagenomics were collected. Together, these comprehensive methodologies and large investigations have revealed an unexpected range of novel protist biodiversity previously undescribed in marine environments.

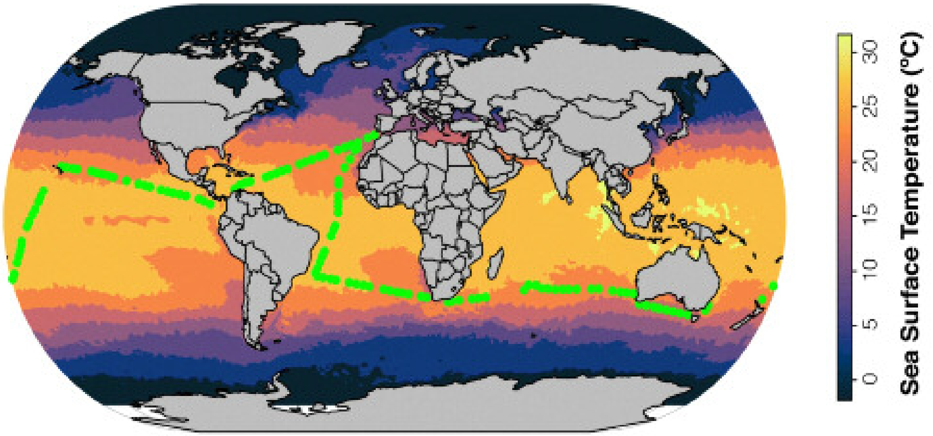


Figure 9 – Representation of the Malaspina expedition. Stations sampled in the tropical and subtropical ocean are shown in green. From Logares et al. (2020).

1.4- AN UNDERSTUDIED FUNCTION IN PROTISTS: PHAGOCYTOSIS

Phagocytosis refers to the engulfment of a particle by a single cell through invagination of its cell membrane. It is one of the oldest eukaryotic interactions as it may be related to the origin of the eukaryotic cell, which could have arisen by the acquisition of the mitochondrion by another prokaryote by phagocytosis (Cavalier-Smith, 2002; Yutin et al. 2009). It appeared very early in evolution and remained a conserved function from unicellular protists to animals. Thought to be an exclusive feature of eukaryotes, a recent study has observed phagocytic behaviors in bacteria (Shiratori et al. 2019). More studied as a process associated with the function of the immune system or as a system to maintain homeostasis (clean debris and dead cells) in metazoans, phagocytosis is also fundamental for nutrition in unicellular organisms. This type of phagocytosis is carried out by protists and is much less understood.

1.4.1- PHAGOCYTOSIS: A FUNDAMENTAL PROCESS

As a broad concept, phagocytosis consists of ingestion of large particles (>0.5 μm in diameter) into membrane-bound vesicles called phagosomes (Boulais et al. 2010). Phagocytosis is therefore initiated by a physical contact at the surface of a cell, where receptors recognize ligands exposed by the prey particle. There exist different types of receptors (Freeman and Grinstein, 2014) but it is often unclear how they are activated. The internalization of particles requires a dramatic change of the cell shape. Once a particle is received, a signaling pathway induces remodeling of the actin cytoskeleton that allows extension of membrane protrusions (actin polymerization) to guide the membrane around the particle (Levin et al. 2016). Therefore, a nascent phagosome (membrane-bound phagocytic vacuole) is formed, which later matures upon fusion and interaction with cytosolic organelles, e.g. endomembrane compartments including endosomes and lysosomes (Niedergang and Grinstein, 2018a; 2018b). Phagosome maturation follows three stages: 1) the early phagosome, 2) the late phagosome, and 3) the phagolysosome. The first stage consists of consecutive fusions of the phagosome with plasma membrane and early endosome membranes merged right before the phagosome closes itself. The early phagosome then fuses with late endosomes and strongly acidifies its lumen (to a pH ~4.5) by the acquisition of proton pumps, like the vacuolar adenosine triphosphate V-ATPases, to become a late phagosome. Finally, the phagolysosome is formed by the fusion of the late phagosome with lysosomes, which provide the degradative components and properties that allow the final digestion of the ingested particle in the robust environment of the phagolysosome.

1.4.2- CORE SET OF PROTEINS INVOLVED IN PHAGOCYTOSIS

As described above, the process of phagosome formation and maturation corresponds to a succession of events involving distinct proteins responsible for continuous action (Figure 10). The phagocytosis life strategy is a complex functional process that imply multiple genes. First, once a particle is captured, actin polymerization allows membrane protrusions and pseudopodia extension. Actin nucleation is mediated by an assembly factor, the Arp2/3 protein complex. The Arp2/3 complex consists of 7 distinct subunits that are activated either by the Wiskott-Aldrich syndrome protein WASp/N-WASp or Scar/WAVE nucleation promoting factors (Rohatgi et al. 1999), themselves activated via Rho-family GTPases such as Cdc42 or Rac. Together, they lead to actin reorganization and drive the polymerization of actin into branched filamentous networks (Kinchen and Ravichandran, 2008). Depolymerization of actin filaments at the base of the nascent phagosome is then helping the membrane to surrender the uptake particle. This action is controlled by phosphatidylinositol 3-kinase (PI3K) which recruit proteins to inactivate Cdc42 and Rac and therefore reduce Arp2/3 activity. Then it comes the important recruitment of Rab5, a GTPase that is crucial to promote the first fusion events forming the early phagosome. As phagocytosis depends on cytoskeleton remodeling, actin and tubulin are part of the conserved proteins in the generation of the phagocytic cup. In addition, actin-binding proteins such as gelsolin, profilin, cofilin, formin, and coronin, all present in eukaryotes (Yutin et al. 2009), also participate in the core set of proteins in phagocytosis. The early phagosome becomes a little acidic (pH 6.1–6.5) by the action of V-ATPase protein complex accumulating on its membrane. These V-ATPase proton pumps transport protons (H+) into the lumen of the phagosome using cytosolic ATP as the energy source (Marshansky and Futai, 2008). One of the essential Rab5 effectors is the PI3K (Roberts et al. 2000; Vieira et al. 2003) which helps towards the formation of the late phagosome. The latter is defined by the presence of Rab7 proteins, ending the Rab5 activation. With the help of microtubules, Rab7 promotes the contact of the current phagosome with lysosomes. A gradual fusion takes place to become a phagolysosome where targeted particles are degraded (Figure 10). The phagolysosome is strongly acidified due to the accumulation of V-ATPase, and these acidic conditions allow the activity of degradative enzymes such as proteases, cathepsins, hydrolases and lipases already present from the fusion with lysosomes (Kinchen and Ravichandran, 2008).

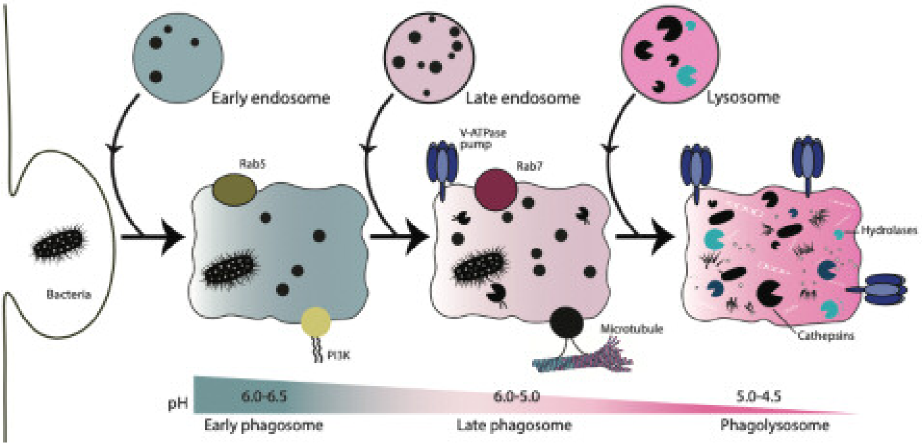


Figure 10 – Phagosome maturation. A newly formed phagosome quickly evolves by a series of fusion events with endosomes and lysosomes. The early phagosome is marked by the presence of GTPase Rab5 while GTPase Rab7 is unique to the late phagosome. The latter becomes acidic by the accumulation of V-ATPases and finally forms the phagolysosome by fusing with the lysosome.

1.4.3 – PHAGOCYTOSIS IN PROTISTS

In protists like heterotrophic flagellates, phagocytosis mainly serves in the food uptake (Cosson and Soldati, 2008). Effective protistan grazing and growth on bacteria relies on the success of two successive steps, ingestion and digestion. Ingestion starts with the contact of the food particle, but the exact mechanism by which phagocytosis is initiated is relatively unknown. Phagotrophic protists have developed a variety of feeding strategies to acquire food particles (Jürgens and Massana 2008; Montagnes et al. 2008). In a "filter-feeding" strategy, protists such as small ciliates and choanoflagellates (Simek et al. 2004) transport water through a filter formed by cilia or pseudopodia tentacles that strain prey particles from the water. In a "direct interception" (also named raptorial feeding) approach, preys are drawn towards the flagellate by a feeding current created by the beat of one flagellum. Captured by the flagella, the particle is brought to the cell surface, waiting to be phagocytized. Both methods require protist mobility. A third mechanism, known as "diffusion feeding" (found in heliozoans), depends on prey mobility, as the predator use their axopods – an arm-like structure - into which the prey collides.

A few studies have suggested that the participation of some proteins like lectins (which bind carbohydrates) may play a role in prey recognition and attachment (Roberts et al. 2006; Wootton et al. 2007). Also, Ca^{2+}/Calmodulin have been suggested to act as regulators for the formation of the food vacuole (Gonda et al. 2000). However, in most cases protists don't make a distinction between their ingested food. Experiments have shown that the ciliate *Tetrahymena* can ingest various particles including, latex beads, carbon nanotubes, bacteriophages and bacteria (Batz and Wunderlich, 1976; Nilsson, 1977; Maicher and Tiedtke, 1999; Hennemuth et al. 2008; Chan et al. 2013). Nevertheless, some heterotrophic flagellates have proved to have prey preferences (Matz et al. 2002).

The degradative capacity of bacterivorous protists to digest prey is acquired through phagosome maturation (Figure 10), a succession of membrane fusion events with endocytic components (Pauwels et al. 2017). Main fusion events are with acidosomes (considered as a late endosome) to reduce the vacuole pH, and with lysosomes (to provide the enzymes for digestion). Prey digestion in protists has mainly been studied with pulse-chase experiments (adding fluorescent prey to a protist culture) (Sherr et al. 1988; Dolan and Simek, 1998; Jacobs et al. 2006). This allowed the identification of the vacuole formation and the digestion of the bacterial prey (Thurman et al. 2010). A rapid acidification (pH from ~7 to 3) has been observed within the first 5 minutes of the vacuole formation (Fok et al. 1982), partly achieved by the action of V-ATP-ases (Yates et al. 2005). The acidified phagosomes gain their digestive enzymes via fusions with lysosomes, including proteases, lipases, phosphatases and glycosidases. The phagosome maturation is however not as simple as a succession of events. Rogers and Foster (2008) suggested that the maturation process occurs over many parallel pathways.

AIMS AND OUTLINE

The overall objective of this thesis is to provide new understanding on the functional ecology of marine protists, and more specifically of heterotrophic flagellates that have often been neglected. Our objective was therefore to gain new insights into uncultured lineages, the MArine STramenopiles (MASTs) that are divided in several lineages with potential ecological differentiation. Towards this idea, we combined different approaches in order to get access to the genomic content and functional capacity of several MAST species. Especially, we focused on their ability to ingest prey by phagocytosis to sustain their nutritional needs. In addition, we benefited from genomic and expression data from the cultured *Cafeteria burkhardae,* a model heterotrophic flagellate within the Stramenopile clade and therefore evolutionary close to the MASTs. This dissertation contains three chapters that provide new knowledge on heterotrophic flagellates by the study of the widespread and abundant uncultured MASTs and *Cafeteria burkhardae*. Each chapter is structured as a scientific paper, already published or submitted for review to a journal (Chapter 1 being the introduction).

Chapter 2: *Comparative genomics reveals the basic trophic lifestyle of uncultured MAST species*

Single cell genomics provided the opportunity to reach the gene content of uncultured protists such as the MASTs. We used samples collected from several oceans and obtained the draft genomes of 15 different MAST lineages via co-assembly. In this chapter we focused on the metabolic traits characterizing the trophic lifestyle of the MASTs, with special emphasis on the digestion step within their acidified vacuoles.

- Provide reference genomes from uncultured lineages.
- Use comparative genomics to identify subset of genes involved in phagocytosis.

- Focus on the digestion within acidified vacuoles, with the potential contribution of rhodopsin proteins, digestive enzymes like peptidases and proton pumps.

Chapter 3: *Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates*

Being uncultured, none of the studies on MASTs have revealed their gene expression during the important ecological behavior of bacterivory. To by-pass this unculturable feature, we set up an unamended experiment in a controlled conditions to follow the MASTs growing dynamics and their gene expression using metatranscriptomics mapped towards the single cell genomes.

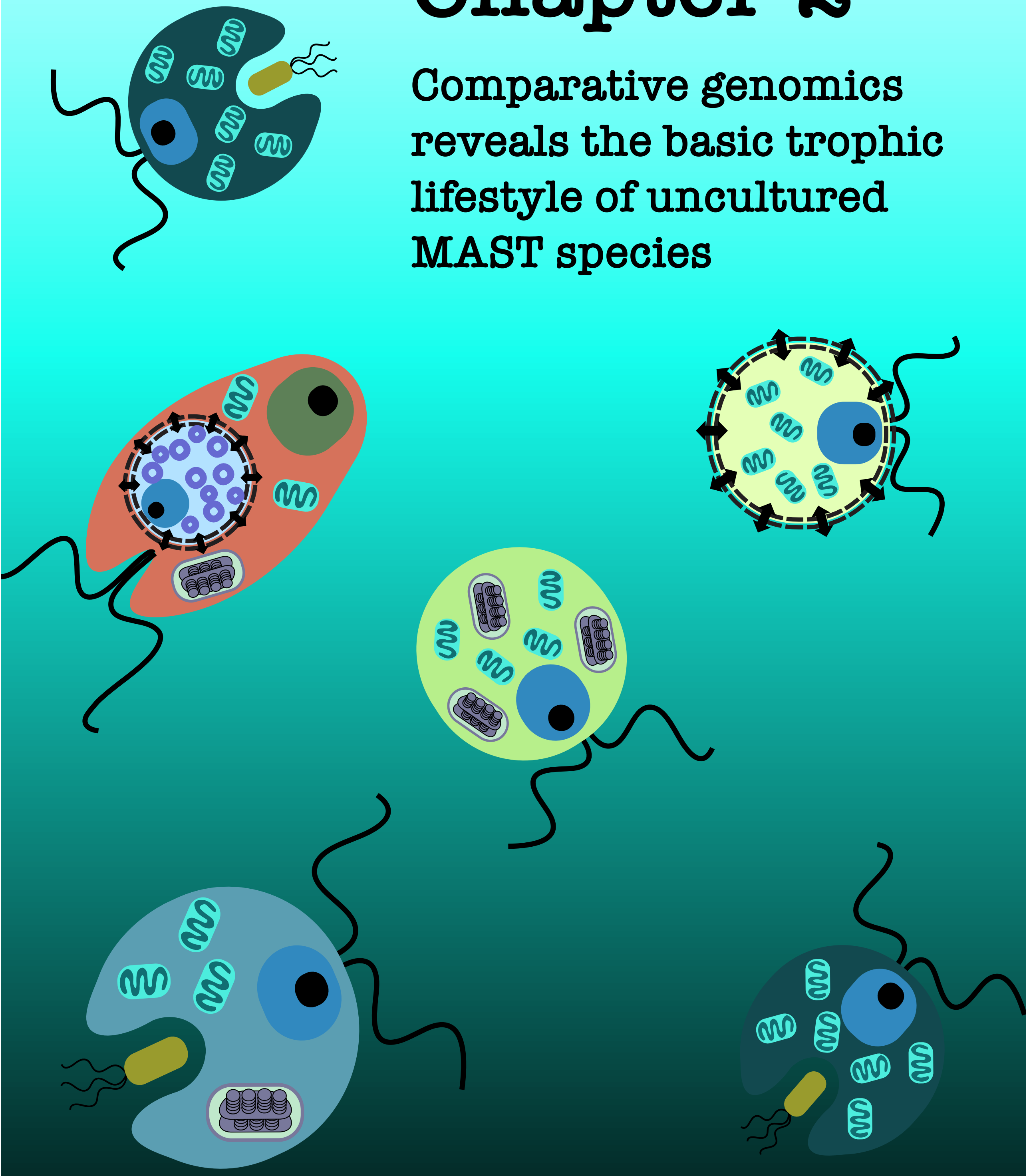
- Develop an original method to access the gene expression profile of uncultured MASTs in their natural habitats.
- Compare the growth response of phototrophic and heterotrophic protists and investigate their species composition.
- Obtain an overview of the genes expressed during bacterivory for a subset of MAST species for which we had the reference genome (obtained in Chapter 2).

Chapter 4: *Gene expression during bacterivorous growth of a widespread marine heterotrophic flagellate*

Our knowledge of phagocytosis derives from animals and their immunity system. In this chapter we use the opportunity to study this fundamental process in the cultured free-living heterotrophic flagellate *Cafeteria burkhardae*. Differential gene expression analysis demonstrated the different genes used during the active growth by bacterivory and during the stationary phase.

- Beneficial use of a cultured and cosmopolitan heterotrophic flagellate in a controlled environment.

- Elucidate an ecological process poorly understood in protists: the phagocytosis.
- Detection of the highly expressed genes during active phagocytosis and under starvation.



Chapter 2

Comparative genomics reveals the basic trophic lifestyle of uncultured MAST species

Under review in The ISME Journal: Multidisciplinary Journal of Microbial Ecology

Aurelie Labarre, David López-Escardó , Francisco Latorre, Guy Leonard, Francois, Bucchini, Aleix Obiol, Corinne Cruaud, Michael E. Sieracki, Olivier Jaillon, Patrick Wincker, Klaas Vandepoele, Ramiro Logares & Ramon Massana

ABSTRACT

Heterotrophic lineages of Stramenopiles exhibit enormous diversity in morphology, lifestyle, and habitat. Among them, the MASTs represent numerous independent lineages that are only known from environmental sequences retrieved from marine habitats. The core energy metabolism characterizing these unicellular eukaryotes is poorly understood. Here we used single cell genomics to retrieve, annotate and compare the genomes of 15 MAST species, obtained by co-assembling sequences from 140 individual cells sampled from the marine surface plankton. Functional annotations from their gene repertoires is compatible with all of them being phagocytotic. Subsets of genes used in phagocytosis, like proton pumps for vacuole acidification and peptidases for prey digestion, did not reveal particular trends in MAST genomes as compared with non-phagocytotic Stramenopiles, except a remarkable presence of V-PPases and rhodopsin genes. Our results support the idea that MASTs may be capable of using sunlight to facilitate phagocytosis, with rhodopsins potentially contributing to vacuole acidification. Our analysis reflects the complexity of phagocytosis machinery in microbial eukaryotes, which contrasts with the well-defined set of genes for photosynthesis. This new genomic data provides the essential framework to study ecophysiology of uncultured species and to gain better understanding of the function of rhodopsins and related carotenoids in Stramenopiles.

INTRODUCTION

Oceans are the largest habitats on Earth, and living biomass in these systems is dominated by planktonic microbes [1]. Together, they introduce heterogeneity into the ocean, govern trophic interactions, and drive energy and nutrient flows [2]. Depending on the way microbes acquire energy and food, they stand along a trophic spectrum between phototrophs, which synthesize organic matter using solar energy, and heterotrophs, which live at the expense of acquired organic matter. The study of trophic strategies is of primary interest to understand the ecological role and behavior of microbial species. This basic information is not always easy to access, especially because as seen in molecular surveys, the vast majority of microbial diversity has not been cultured and therefore remains uncharacterized [3]. Within marine microbial eukaryotes, an important component of this unknown diversity are the Marine Stramenopiles (MASTs) lineages [4, 5]. They are divided into 18 different phylogenetic clades [6] placed in different positions of the Stramenopile radiation that include phototrophs, phagotrophs, mixotrophs, osmotrophs, and parasites [7, 8]. A clear assignment of the trophic strategy of MASTs is also challenging because of their small size and lack of recognizable morphological features. Partial data exists for a few clades, some MAST-3 are parasites (for example, the diatom parasite *Solenicola setigera* belongs to this clade [9]), MAST-1 and MAST-4 are active bacterivores [10], but this elementary knowledge is still unknown for many other MAST lineages.

Genomics is increasingly contributing to our understanding of the global ocean, expanding our knowledge on marine microbial life and their metabolic potential. Sequencing the genome of a given microbial species may provide strong evidences about its ecological function and may identify unique features defining ecological niches. This requires a certain amount of DNA for sequencing, typically extracted from a high-biomass culture if the taxa was cultured. Nowadays, Single Cell Genomics (SCG) has become a widely used approach to access the genomes of uncultured microbial species [11, 12]. SCG methods are currently powered by multiple displacement amplification (MDA), which amplifies the minute DNA amounts of a single cell, and has proved to provide useful genomic data of uncultured marine protists [13] including the MAST-4 [14]. Nevertheless, the quality of SCG assemblies is lower than what is obtained by standard genomics, as the MDA may cause uneven coverage depth, chimeric sequences, and increased contamination [15] leading to incomplete genome reconstructions. A computational solution to circumvent MDA drawbacks is the combination of sequencing reads of several single cells into a co-assembly, which improves genome completeness [16,17].

We investigated the molecular functioning of unicellular heterotrophic organisms that satisfy their food needs by eating other organisms via phagocytosis. This mechanism is a distinct form of endocytosis that incorporates particles >0.45 μm in diameter through the formation of membrane-bound vesicles called phagosomes. After maturation, phagosomes fuse with lysosomes and become a final phagolysosome where prey cells are degraded [18, 19]. Lysosomes are important organelles that can contain more than 50 degradative enzymes (targeting proteins, carbohydrates or nucleic acids) commonly named acid hydrolases as they are activated at acidic conditions (i.e., pH <5). To maintain the acidic medium and keep control over the digestive enzymes, phagolysosomes accumulate H^+ ions by the action of the vacuolar-type H^+-translocating ATPase (V-ATPase) [20]. Other proton pumps like the vacuolar-type H^+-translocating pyrophosphatase (V-PPase) can also participate to acidification [21]. The two proton pumps obtain their energy by hydrolyzing phosphate bonds, in ATP or inorganic pyrophosphate respectively [22], and represent distinct classes of ion translocases with no sequence homology. Functional related genes that are gaining momentum in marine microbial ecology are the rhodopsins. Microbial type-I rhodopsins are photoactive proteins containing a retinal chromophore that work as light-driven proton pumps or photoreceptors [23, 24]. They are widely present in marine microbes [25, 26] and have been found highly expressed in a growing MAST-4A population [27]. It has been suggested that besides energy processing, rhodopsins can participate in food vacuole acidification in eukaryotic phagotrophs [28].

In this study, we have analyzed the genomes of 140 single cells retrieved during the *Tara Oceans* expedition as well as at the Blanes Bay Microbial Observatory. These cells affiliate within seven MAST clades highly represented in marine molecular surveys [6]. The 140 SAGs have been further co-assembled into 15 genomes of relatively high quality and subsequently analyzed by comparative genomics together with other well-characterized Stramenopiles. We first focused on assigning a trophic function to these uncultured clades by comparative genomics, and then analyzed the enrichment of the degradative enzymes peptidases according to trophic function. We also considered in detail the presence and diversity of proton pumps and microbial rhodopsins in MASTs to further understand the potential physiological cell capabilities and the role of light in phagolysosome acidification.

MATERIAL AND METHODS

Single Amplified Genome (SAG) sequencing, assembly, and co-assembly

Epipelagic microbial communities sampled during the *Tara Oceans* expedition were used for flow cytometry cell sorting at the Single Cell Sorting Center in Bigelow (scgc.bigelow.org) based on size and the presence or absence of pigments. Whole genome amplification from single cells was done with MDA, and SAGs were taxonomically classified by sequencing their 18S rDNA amplified with universal eukaryotic primers. Details of the methods used and a complete list of taxa ID for all SAGs collected in Tara are presented in Sieracki et al. [29]. Seventy-four of the SAGs used here have been sequenced and analyzed previously [16, 17, 30], while 50 SAGs are new from this study (Table S1). We did a single cell sorting effort at the Blanes Bay Microbial Observatory (BBMO) in May 2018 using similar protocols that provided 16 additional SAGs. Sequencing libraries for cells collected in Tara were prepared as described before [17], while we used the KAPA or Nextera preparation kits in BBMO cells. SAGs were paired-end sequenced (reads of 110 bp in Tara and 250 bp in BBMO) in different Illumina platforms and sequencing services (Table S1).

After adapter trimming and cleaning of the raw reads using Trimmomatic v. 0.32 [31] (reads with a Phred score <20 and <100 bp were discarded), we performed a digital kmer-based normalisation with BBNorm (sourceforge.net/projects/bbmap/) that reduces the average error rate and allows down-sampling of reads for a better coverage distribution (a critical issue with MDA products). An initial *de novo* assembly using the *de Bruijn* graph assembler SPAdes [32], combining information from 21, 33 and 55 k-mer sizes, was generated for every individual SAG read set. Based on previous work [16, 30], we followed a co-assembly strategy using stringent criteria: only SAGs with nearly identical 18S rDNA, very similar GC content, and tetranucleotide homogeneity verified with the ESOM tool (http://databionic-esom.sourceforge.net) were eligible for co-assembly, which was done with SPAdes including the "single cell" option. We identified (and later removed) prokaryotic contamination in the assembled scaffolds with the default parameters of EukRep [33] and Blobtools [34]. In one of the sequencing batches, cross-contamination between SAGs in the same Illumina lane occurred due to HiSeq reagents problems. We computed the average nucleotide identity [35] between contigs in all pairs of individual SAGs, identified problematic contigs (those that share regions with similarity >99% in fragments longer than 300bp), and removed those from the SAG where they had the lowest k-mer read coverage. In the final co-assemblies, contigs shorter than 1 kb were removed, and genome statistics were computed with QUAST [36]. Genome completeness was determined by the presence of 248 universal, single-copy core eukaryotic genes (CEGs) with CEGMA [37] or the presence of 303 single copy Eukaryotic orthologous genes with BUSCO v3 [38].

Gene predictions, gene family inference and functional annotation

Gene predictions from the co-assembled genomes started by using the CEGMA and BUSCO retrieved genes to train SNAP (http://korflab.ucdavis.edu/software.html), which generates a set of *ab initio* gene models. In parallel, GENEMARK-ES [39] was run to obtain another set of predicted genes. Both sets were then used as input for the MAKER [40] pipeline, developed to combine multiple sources of information into a final set of gene annotations. The new predicted models were then used in a second run of MAKER, with default settings, to train the program AUGUSTUS [41], finally providing transcripts and protein predictions for each co-assembled genome. The pipeline used can be found on Github: (https://github.com/guyleonard/gene_prediction_pipeline).

Predicted coding sequences (CDS) from the co-assembled MAST genomes were loaded into a custom instance of the PLAZA framework [42] together with the CDS of other Stramenopiles and non-Stramenopile model species (Fig. S1). Based on an 'all-against-all' protein sequence similarity search done with DIAMOND v. 0.9.18 [43] ('more sensitive' mode with a maximum e-value cutoff of 10^{-5} and retaining up to 2,500 hits), orthologous gene families were delineated with OrthoFinder v. 2.3.3 [44] (default parameters). Functional annotation of all CDS was performed using InterProScan v. 5.39-77.0 [45], including mapping InterPro entries to GO annotations. For the model organisms in the database (Fig. S1), GO annotations were retrieved from the GO website. Finally, functional enrichment analyses were performed to assign informative InterPro and GO terms to each orthologous gene family. The enrichment analysis used the hypergeometric distribution with a maximum Bonferroni corrected p-value cutoff of 0.05, and all coding genes from the organisms included in the gene family as background frequency. Enriched functional annotations were retained when present in at least half of the genes in the family.

Comparative genomics analysis

We used a computational model designed to predict, using genomic data, if an organism has the ability to be phagocytotic (able to capture prey), photosynthetic (able to fix inorganic carbon), or prototrophic (self-sufficient producer of essential amino acids or vitamins) [46]. The model is based on clusters of shared proteins among a large diversity of eukaryotic genomes and on an evaluation of their enrichment in organisms adopting different lifestyles. The presence of specific proteins in the query genomes, detected by a search with HMM models, is used to predict the lifestyle of unknown organisms.

On a second level, we used the number of copies for each orthologous gene family (or orthologous group, OG) in every species to identify broad patterns within the 30 Stramenopile species. OGs found in only one species were discarded, and the number of genes per OGs were normalized to percentages in each genome. Based on the OG table, genomes were compared using Bray-Curtis dissimilarities and analyzed by NMDS (non-metric multidimensional scaling) with the R package vegan v2.5-6 [47]. The grouping of species based on trophic lifestyle was tested by a PERMANOVA analysis using vegan's function *adonis2()*. A multi-level pattern analysis to identify OGs that characterize a given trophic mode (Indicator Value (IndVal) > 0.7 and p value < 0.05) was performed using the function *multipatt()* implemented in the R package indicspecies v1.7.9 [48]. A heatmap displaying OGs annotated as peptidases and proteases was created with R package pheatmap v1.0.12 [49], using Ward's method for hierarchical clustering with log$_{10}$ -transformed OGs gene counts (with a pseudocount of 1).

Homology searches and phylogenetic analyses for specific proteins

Protein sequences from three gene families of proton pumps were retrieved from public databases. Reference sequences for V-ATPases were extracted from Mulkidjanian et al. [50], while for V-PPases we used the phylogenetic tree in Goodenough et al. [21]. Rhodopsin reference sequences were collected from several articles [28, 51, 52], and the MicRhoDE project [53]. Using these reference datasets, homologous MAST sequences were identified by sequence similarity using BLAST v.2.2.28 (maximum e-value threshold of 10^{-5}). The selected contigs have been checked to discard potential bacterial contamination. Homology searches using Pfam domains were conducted against the key enzymes involved in retinal formation: GGPP synthase (PF00348), Phytoene synthase (PF00484.18), Phytoene dehydrogenase (PF01493.23), Lycopene cyclase (PF05834), and β-carotene 15,15'-dioxygenase (PF15461.5). Selected sequences were aligned with MAFFT v7.470 [54] (--globalpair) and trimmed with TRIMAL v1.4 [55] (-automated option) to obtain a curated subset for phylogenetic analyses. Phylogenetic trees were constructed with the Maximum Likelihood method using the LG+F+R6 substitution model in IQ-TREE [56] and topology support was determined with 1000 bootstrap replicates.

Data availability

Data have been deposited in Figshare under the project number 10.6084/m9.figshare.c.5008046, including genome co-assemblies, CDS predictions, phylogenetic analyses, and scripts used in our analyses. Sequencing reads have been deposited at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under the BioProject. Individual SAGs, co-assembled contigs, predicted genes and proteins can also be explored through an in-house developed web repository (sag.icm.csic.es).

RESULTS

A new set of MAST genomes

Unicellular eukaryotic microorganisms were single cell sorted from planktonic assemblages in the Adriatic Sea and the Indian ocean during the *Tara Ocean* expedition, and in Spring 2018 from the BBMO (Fig. 1A). Based on their 18S rDNA signature, 140 cells from the unpigmented sort that affiliated to Marine Stramenopile lineages (MASTs) were selected for genome sequencing. Essential sampling and sequencing information regarding these Single Amplified Genomes (SAGs) is listed in Table S1. SAGs with similar tetranucleotide frequency and very high nucleotide similarity (fulfilling the criteria explained in M&M) were considered to be from the same species and combined into a co-assembly, thus yielding improved genomes of 15 MAST species. The individual SAGs used in each co-assembly often derived from different marine locations (Fig. 1B).

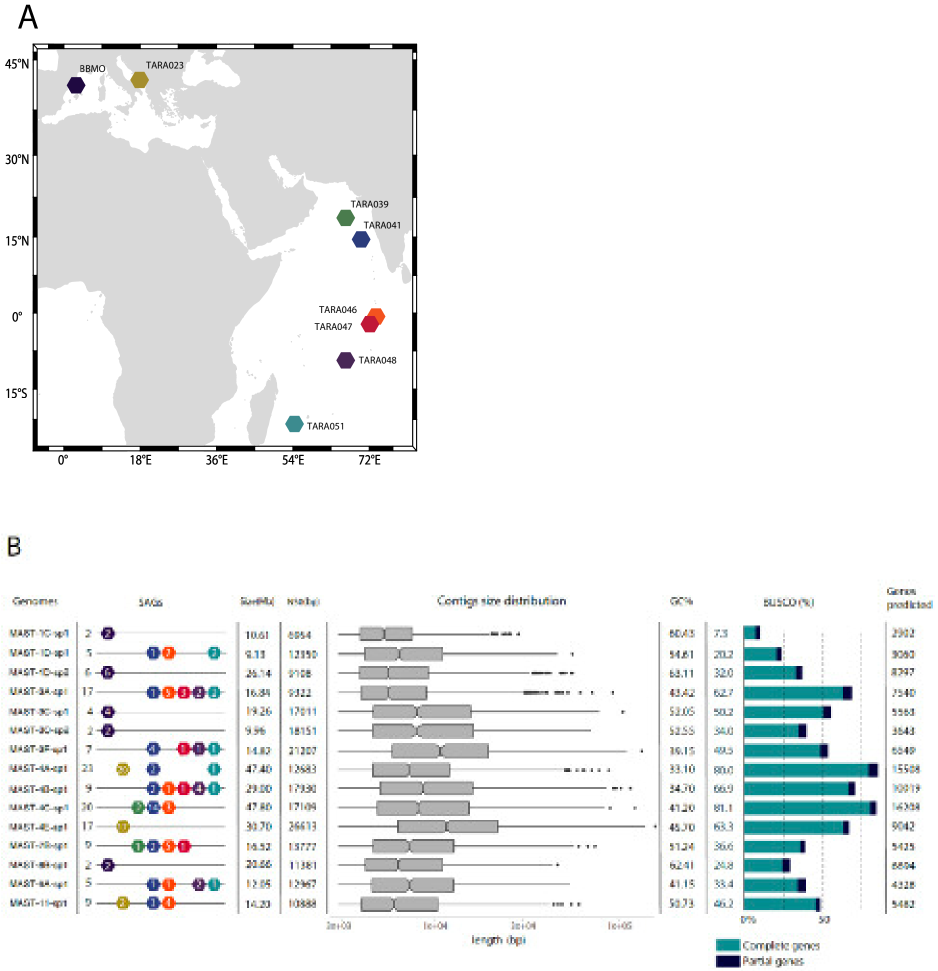


Figure 1. Genomic characteristics of 15 MAST species obtained by co-assembling individual SAGs. (A) Location of marine sites where microbial communities were sampled. (B) Genome parameters of the 15 co-assembled species: number of individual SAGs assembled and their distribution across sampling sites; assembled genome size; N50 assembly statistics and size distribution of contigs; GC content; genome completeness as the percentage of BUSCO complete (light blue) or fragmented (dark blue) gene models; number of predicted genes.

Taking into account contigs ≥ 1kb, we obtained genome sizes ranging from 9.13 to 47.80 Mb, each one with a characteristic GC content. Assembly quality assessments were carried out via the N50, the size distribution of contigs, and the genome completeness. The later, based on the percentage of conserved single copy orthologous genes present in the final co-assembly, averaged 46% across genomes, ranging from values as high as 80% in MAST-4A-sp1 and MAST-4C-sp1 to values as low as 7% in MAST-1C-sp1 (Fig. 1). As expected, genomes with higher completeness also recovered more genes: 15,508 genes were predicted in MAST-4A-sp1, 16,260 in MAST-4C-sp1, and 2902 in MAST-1C-sp1. Thus, there was a clear correlation between genome size and both the BUSCO completeness and the number of predicted genes. Overall, co-assembled genomes provide reasonable gene completeness and represent a very promising resource to reveal the genes and the metabolic potential of uncultured Marine Stramenopiles.

Predicting the lifestyle of MAST species from genomics

We investigated the trophic lifestyle of the 15 MAST species using a recently published comparative genomics model [46]. Specifically, the training-based model interrogates the genomes of unknown species for the presence of genes predictive of phagotrophic, photosynthetic or prototrophic lifestyles (Fig. 2). The model clearly predicted that none of the MAST species was photosynthetic: all of them were outside the photosynthetic PCA cluster, with 73% of the variation explained by the first principal component (Fig. 2A), and virtually zero prediction probabilities of being photosynthetic (Fig. 2C). Based on the set of genes defining phagotrophy, the majority of MAST species were placed with phagocytotic genomes (the first principal component explained 73% of the divergence) and within the 95% confidence ellipse in the PCA plot (Fig. 2B). The prediction probability for phagotrophy was above 80% in most cases, but it was very low in 4 of them, MAST-1C-sp1, MAST-1D-sp1, MAST-3C-sp2, and MAST-9A-sp1, precisely the ones that had the lowest number of predicted genes. At first sight, MAST species do not seem to perform prototrophy, being outside the prototrophic PCA cluster (Fig. S2). However, the species with most predicted genes (several MAST-4 and MAST-3A-sp1) display a moderate prediction probability to present this capacity (Fig. 2C).

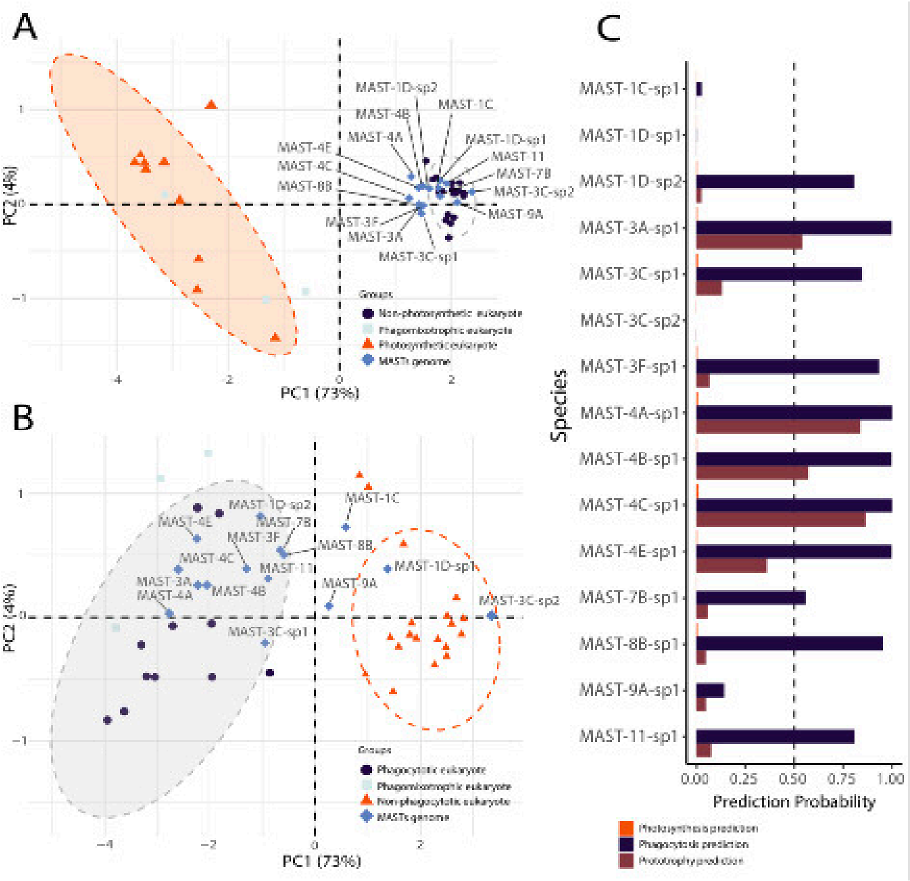


Figure 2. Lifestyle prediction of MAST species using a comparative genomics model [46]. (A) Plot of two first principal components (PC1 and PC2) placing genomes based on their genes associated to GO categories defining the photosynthetic lifestyle. (B) PCA plot placing genomes based on their genes associated to GO categories defining a phagocytotic lifestyle. (C) Prediction probabilities for MAST species to the three lifestyles. Dashed line ellipses in A and B illustrate 95% confidence assessments of the groupings based on photosynthetic and phagocytotic predictions.

Furthermore, while the previous analysis relied on preselected group of genes, we also performed a direct comparison of the 15 MAST species against a selection of other Stramenopiles with known lifestyle (Fig. S1) using the number of genes in inferred orthologous groups (OGs) within each genome. The corresponding NMDS test revealed that the species grouped according to the defined trophic strategies: a tight photosynthetic cluster, an intermixed osmotrophic cluster, and a loose group including *Cafeteria burkhardae* and MAST species (Fig. 3).

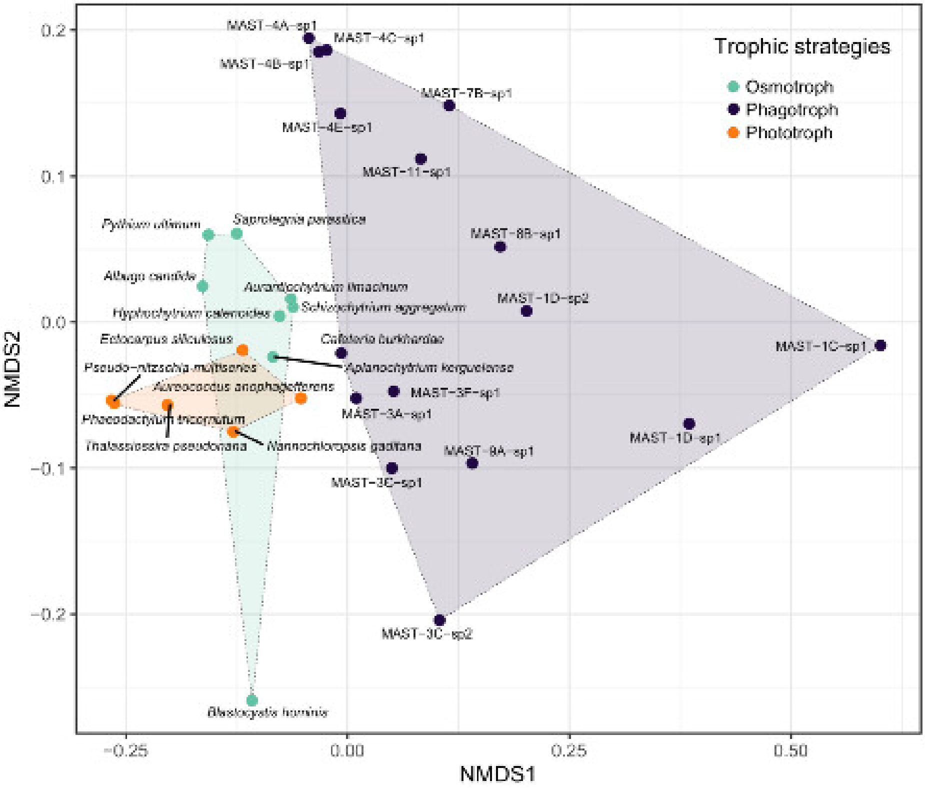


Figure 3. NMDS plot relating the 30 Stramenopiles genomes based on their Bray-Curtis dissimilarity calculated from the relative abundance of genes per genome within defined orthologous groups. The species are colored and grouped with a shadowed area according to their trophic lifestyle.

A PERMANOVA analysis showed that 22% of the variance in the plot (p<0.001) was explained by the trophic mode, and this justified the use of the indicator value (IndVal) statistic to this dataset. Among the 28 OGs indicators of the phagocytosis trophic mode (Table 1), we identified many digestive enzymes (peptidases, glycosidases, lipases), and other genes related to cell growth and responses to the environment. A larger number of OGs characterized osmotrophs (Table S2) and phototrophs (Table S3), 133 and 744 OGs, respectively. In particular, phototrophs displayed many genes encoding for photosystem and other plastidic proteins.

Ortholog groups	IndVal	p.value	InterPro	Description	GO Term	General function
ORTHO03S000834	0.91	0.01	IPR011040	Sialidase	GO:0004553	Digestive enzyme
ORTHO03S000616	0.89	0.01	IPR004302	Cellulose/chitin-binding protein	--	Cell interactions
ORTHO03S000329	0.88	0.01	IPR004963	Pectinacetylesterase/NOTUM	GO:0016787	Digestive enzyme
ORTHO03S004730	0.87	0.01	IPR004981	Tryptophan 2,3-dioxygenase	GO:0019441	Digestive enzyme
ORTHO03S002955	0.83	0.01	IPR033396	Domain of unknown function DUF5107	--	Unknown function
ORTHO03S001168	0.83	0.01	IPR001577	Peptidase M8, leishmanolysin	GO:0008233	Digestive enzyme
ORTHO03S004520	0.83	0.01	IPR006201	Neurotransmitter-gated ion-channel	GO:0034220	Membrane transport
ORTHO03S000334	0.82	0.03	IPR000884	Thrombospondin type-1 (TSP1) repeat	--	Cell interactions
ORTHO03S004517	0.79	0.01	IPR004911	Gamma interferon inducible lysosomal thiol reductase	--	Vacuolization
ORTHO03S004519	0.79	0.01	IPR016201	PSI domain	--	Cell adhesion
ORTHO03S005547	0.79	0.01	IPR002477	Peptidoglycan binding domain	--	Digestive enzyme
ORTHO03S002888	0.77	0.02	IPR011040	Sialidase	GO:0004553	Digestive enzyme
ORTHO03S003756	0.76	0.03	IPR021345	Protein of unknown function DUF2961	--	Unknown function
ORTHO03S004503	0.75	0.02	IPR012338	Beta-lactamase/transpeptidase-like	GO:0005576	Digestive enzyme
ORTHO03S004518	0.75	0.01	IPR029787	Nucleotide cyclase	GO:0007165	Signal transduction
ORTHO03S004748	0.75	0.02	IPR036452	Ribonucleoside hydrolase	GO:0016614	Digestive enzyme
ORTHO03S005894	0.75	0.01	IPR008139	Saposin B type domain	--	Digestive enzyme
ORTHO03S004453	0.72	0.05	IPR017920	COMM domain	--	Regulation
ORTHO03S003676	0.72	0.03	IPR004007	Dihydroxyacetone kinase, subunit L	GO:0004371	Signal transduction
ORTHO03S005231	0.72	0.03	IPR004785	Ribose 5-phosphate isomerase B	GO:0005975	Sugar metabolism
ORTHO03S003865	0.72	0.04	IPR005524	Predicted permease DUF318	--	Membrane transport
ORTHO03S005235	0.71	0.02	IPR028730	Zinc finger FYVE domain-containing protein 26	GO:0061640	Cell division
ORTHO03S005554	0.71	0.03	IPR029723	Integral membrane protein GPR137	--	Transmembrane protein
ORTHO03S005577	0.71	0.01	IPR009613	Lipase maturation factor	--	Lipid metabolism
ORTHO03S005836	0.71	0.01	IPR001124	Lipid-binding serum glycoprotein	GO:0008289	Lipid metabolism
ORTHO03S005884	0.71	0.02	IPR002889	Carbohydrate-binding WSC	--	Cell interactions
ORTHO03S005895	0.71	0.02	IPR008139	Saposin B type domain	--	Digestive enzyme
ORTHO03S005965	0.71	0.01	IPR011124	Zinc finger, CW-type	GO:0046872	Regulation

Table 1. List of orthologous groups defining the phagotrophic lifestyle within the dataset of 30 stramenopile genomes. These OGs are first selected by the IndVal test (phagotrophs versus other genomes) and kept when their IPR identification was not found in the lists of OGs characterizing other lifestyles. The InterPro domain annotating each of the 28 OGs is shown, together with its description and a general function. When available the corresponding GO term identifier is also provided.

We then focused on a given group of digestive enzymes, the peptidases, and explored how frequent they were among the complete set of Stramenopile genomes. For this, we selected the 295 OGs that were functionally annotated as peptidases or proteases and studied their distribution in the 30 genomes, both at OGs level (Fig. S3) or after grouping OGs in 71 peptidase families (Fig. 4). These digestive enzymes were present in all species of phototrophs, osmotrophs and phagotrophs in roughly similar gene copy numbers, around 250 genes on average per genome. Therefore, the number of peptidases genes could not be used as indicators of phagotrophic lifestyle. In the OGs heatmap (Fig. S3), the genomes clearly grouped by lifestyle (except *Blastocytis hominis* that appeared with phagotrophs) and some clusters accumulated OGs with IndVal scores, so seemed indicative of given lifestyles. However, in the heatmap constructed with peptidase families (Fig. 4), the grouping of genomes per lifestyle was less clear and a poor correlation of peptidase types and trophic mode was observed.

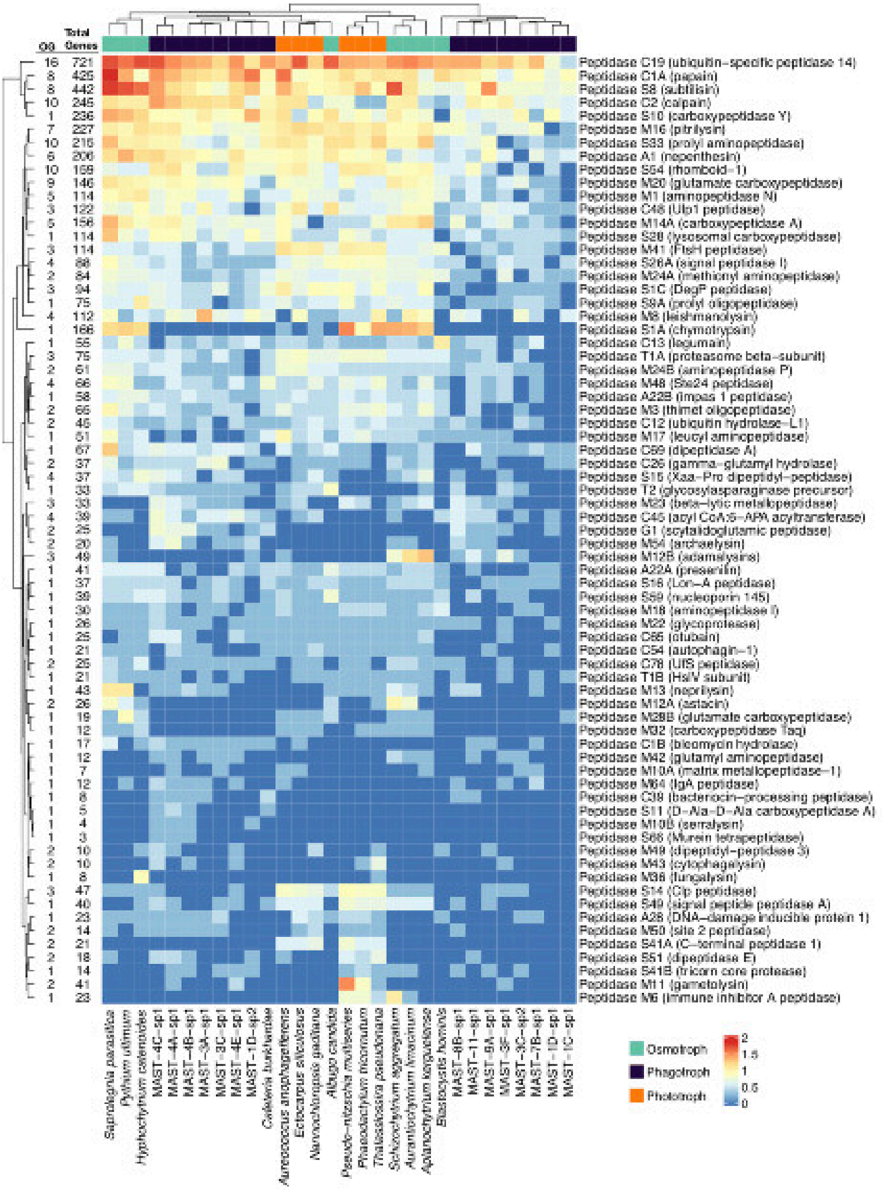


Figure 4. Distribution and abundance (log-transformed number of genes) of peptidase families in the 30 Stramenopile genomes. Each peptidase family follows the MEROPS classification (type enzyme in parenthesis) and may represent several OGs (number of OGs per peptidase in the first column at the left of the heatmap) including many genes (overall number in the second column).

Canonical proton pumps in their role of vacuole acidification

Vacuole acidification, a necessary step for the function of acidic digestive enzymes in mature phagosomes, is achieved by the action of the proton pump V-ATPase, and perhaps the V-PPase. We investigated the presence and the sequence homology of both genes in uncultured MASTs, other Stramenopiles, and several other eukaryotes by phylogeny (Fig. 5). We first looked for the presence of the subunits A and B of the V-ATPase complex, which are homologous to the two subunits of the F-ATPase (Fig. S4). As expected, they were found in all complete genomes but were undetected in about half of the MAST species, most likely due to genome incompleteness.

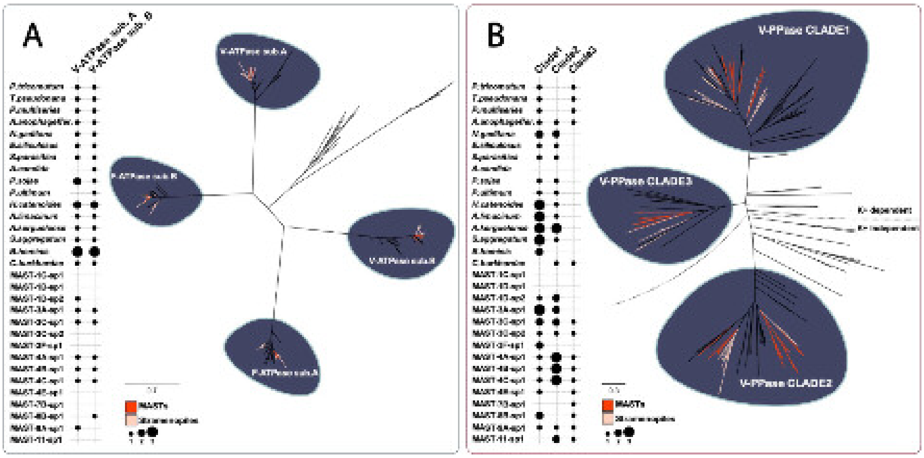


Figure 5. Phylogenetic representation of two distinct proton pumps across stramenopile genomes: V-ATPases (A) and V-PPases (B). The trees are based on 185 and 184 protein sequences respectively. The MAST are represented in orange, other stramenopile species in white, and selected eukaryotic and prokaryotic species in black. Insets show the distribution and number of genes per genome in the different clades.

With respect to V-PPase, these were distributed in the three described clades: clade 1 homologous to the prokaryotic K^+ dependent H^+-PPases; clade 2 homologous to the prokaryotic K^+ independent H^+-PPases; and clade 3 related to the prokaryotic K^+ dependent Na^+ PPases (Fig. S5). Despite genome incompleteness, MASTs species show a remarkably high number of V-PPase genes, three on average, often within the three separate clades. Among them, MAST-4A-sp1, MAST-4B-sp1 and MAST-4C-sp1 contain a particular duplication of the Clade 2 ancient to the divergence of the three species (Fig. S5). It is particularly interesting that the presence of clade 3 V-PPase was detected in MAST species, as this paralog is less frequent in other eukaryotic genomes. Thus, in the Stramenopile set studied here, Oomycetes, Labyrinthulomycetes, and the multicellular brown algae *Ectocarpus* appear to have lost clade 3, which is retained only in some diatoms and *C. burkhardae*. Finally, only two MAST species lacked V-PPase genes (MAST-1C-sp1 and MAST-1D-sp1), and this may likely be due to genome incompleteness.

Rhodopsins and genes for retinal biosynthesis

Rhodopsins are transmembrane proteins that together with a retinal pigment use light energy for proton translocation. Sequence similarity searches confirmed the presence of rhodopsin-like proteins in 11 of the 15 MAST genomes, typically found in multiple individual SAGs (Fig. S6). We carried out a phylogenetic analysis of the full range of microbial type I rhodopsins including also eukaryotic and viral sequences. The new MAST rhodopsin proteins classified into distinct phylogenetic branches (Fig. 6). Some affiliated with the xanthorhodopsins, which are present in marine haptophytes, dinoflagellates, and diatoms. Xanthorhodopsins pump ions across cell membranes and contain carotenoid accessory pigments as a light harvesting mechanism. With the exception of MAST-3F-sp1, in which only 1 of 9 cells contained xanthorhodopsin (Fig. 6), this gene was found in several cells of MAST-4A-sp1, MAST-4C-sp1, MAST-7B-sp1, and MAST-9A-sp1. This strongly supports the idea that these rhodopsins truly belong to MAST species and are not a product of contamination.

A second clade revealed the presence in MAST species of the recently identified MerMAIDs rhodopsins. These light gated ion channel rhodopsins seem specific of marine microbes and were present in MAST-4E-sp1 (in several cells and featuring two distinct copies), as well as in a MAST-7B-sp1 cell with moderate bootstrap support (82%). The amino acid sequences of MAST MerMAIDs aligned very well with the original reports and revealed a well conserved structure (Fig. S7). Similar to other microbial rhodopsins, it features seven transmembrane helices and the lysine Schiff base in the seventh helix where the retinal chromophore typically attaches (Fig. S7). The sequence from MAST-7B-G22 lacks part of the protein but still shows the retinal-binding lysine. The remaining MAST rhodopsins were included in a large bacteriorhodopsins-like clade. Those from MAST-8B-sp1 and MAST-3F-sp1 were closer to halorhodopsins (chloride pumps) and sensory rhodopsins generally limited to halophilic Archaea, as well as to xenorhodopsins (inward H+ directed proton pumps). Those from MAST-1C-sp1, MAST-1D-sp2, MAST-3A-sp1 and MAST-3C-sp2 were closer to a large clade including fungal and bacterial rhodopsins. Our phylogenetic tree also shows that some species, i.e. MAST-3F-sp1 and MAST-7B-sp1, encode microbial rhodopsins from different clades, having putatively different functions. Overall, our data demonstrate that most of the MAST species studied here contain rhodopsins and reveal an important heterogeneity of this gene.

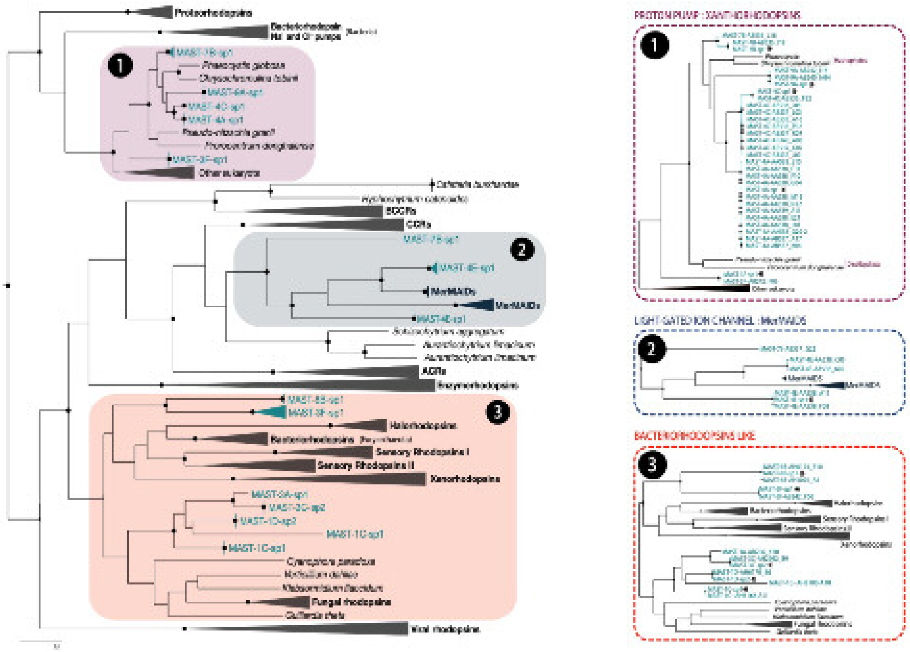


Figure 6. Phylogenetic tree of microbial type I rhodopsins based on 207 protein sequences, including the new MASTs, showing the recognized groups and their prevalent function. Black dots indicate bootstrap support >80% over 1000 replicates. Stars highlight sequences recovered from co-assemblies.

In addition to rhodopsins, we searched for the genes encoding the retinal biosynthetic pathway (Fig. 7 and Fig. S6). This pathway starts with the enzyme GGPP synthase (crtE), the last enzyme involved in Isoprenoid biosynthesis, which produces geranyl$_2$-PP. The next step involves the synthesis of phytoene from two geranyl$_2$-PP, carried out by phytoene synthase (crtB), followed by a sequential desaturation and isomerization via phytoene desaturase (crtI) to synthetize lycopene. The enzymes crtE, crtB and crtI are present in most of the studied MAST species and in many of the individual SAGs (Fig. 7, Fig. S6).

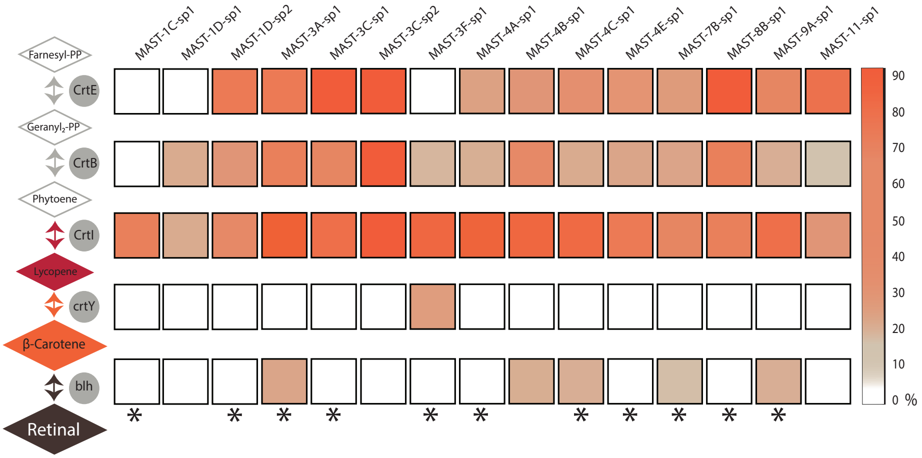


Fig. 7 Presence of enzymes needed for retinal biosynthesis in MAST genomes: GGPP synthase (crtE), phytoene synthase (crtB), phytoene dehydrogenase (crtI), lycopene cyclase (crtY), and β-carotene 15,15'-dioxygenase (blh). The heatmap represents the proportion of individual SAGs within each species having the corresponding gene. Stars indicate species containing rhodopsins.

Synthesis of β-carotene is then catalyzed by the lycopene cyclase (crtY). The key and final step is the oxidative cleavage of β-carotene into retinal by the enzyme β-carotene 15,15'-dioxygenase (blh). This crucial step was detected in only a few MASTs, and the previous step in a single one, which suggests that this pathway is not functional in MASTs. The gene RPE65 (Retinal pigment epithelium-specific 65 kDa protein), which encodes a protein for the regeneration of the 11-cis-retinal chromophore of rhodopsin in vertebrates, has been detected (Fig. S6).

DISCUSSION

Obtaining reliable genomes of uncultured organisms by Single Cell Genomics

In marine ecosystems, unicellular planktonic microbes typically have distinct trophic strategies placed in a trophic continuum mostly defined by energy transfer, from pure photosynthesis to prey uptake heterotrophy [57]. An important component of the marine plankton, the picoeukaryotes, are widespread, widely diverse, and include multiple metabolic types [58, 59]. To date, the vast majority of heterotrophic picoeukaryotes cannot be cultured by traditional techniques, and this prevents the understanding of their functional traits, as both ecophysiological and genomic studies are not possible. Single Cell Genomics (SCG) has proved to be reliable to recover genomic data from uncultured picoeukaryotes [14, 16, 17], to elucidate viral infections [13, 60] or phagotrophic interactions [61], and to highlight new evolutionary insights within animal multicellularity [62]. Here, we used SCG to obtain genome sequences and infer metabolic capacities of previously inaccessible Marine Stramenopiles.

The new genomes of 15 MAST species, obtained by a co-assembly strategy [16], showed a completeness often above 50%, higher to what is generally observed using single cells [63]. From these, we recovered a large number of predicted proteins per genome, the number of which generally correlates with genome size and completeness. While this represents a valuable culture-independent genomic resource, we cannot ignore the technical limitations of SCG. The necessary step of whole genome amplification by MDA is well known to produce a patchy recovery of the original genome, which leads to fragmented and incomplete sequenced genomes that may affect subsequent analysis [12]. This can be partially alleviated (but not completely) by co-assembling multiple cells. Thus, a gene not detected could be because it was absent in the genome or because it was lost during SAG generation and assembly. Nonetheless, we successfully provide genomic data from 15 uncharted branches of the Stramenopile radiation, enabling us accessing to metabolic features and new physiological capabilities of MAST species.

Predicting a general lifestyle for uncultured MASTs by comparative genomics

The placement of the MASTs at the base of the Stramenopiles [6, 8], a phylogenetic region with a large diversity in life-strategies including phagotrophy, osmotrophy and parasitism, implies that the trophic roles of MAST species are not necessarily known. Here we investigated the putative lifestyle of a phylogenetically varied set of MAST species using a recently published model based on comparative genomics [46]. As expected, the model showed evidence that MASTs do not have the proteins necessary for photosynthesis. Moreover, the genomic data strongly suggested that most of the MAST species have the faculty to perform phagocytosis. MAST-3C-sp2 and MAST-1D-sp1 clustered with photosynthetic eukaryotes when the model was trained with the proteins representative of phagocytosis, but this was probably due to the poor genome completeness of both species. In addition, the model seems unable to differentiate between phagocytotic and osmotrophic strategies, as osmotrophic species in the original publication (i.e. oomycetes, see Fig. S1 in [46]) as well as *Hypochytrium* and Labyrinthulomycetes analyzed here (data not shown) were predicted to be phagocytotic. The grouping of osmotrophic genomes excluding MASTs in NMDS plots with complete gene data suggests that MAST species are phagotrophs and not osmotrophs. While the essential genes for photoautotrophy have been well documented either by comparative genomics or experimentally [64, 65], the identification of core proteins for phagocytosis is much less evident. Comparative proteomics have suggested a set of about 2000 proteins associated to the phagosomes [66]. However, the core genes associated to phagocytosis are still difficult to define [46] especially because these genes are used across multiple cellular functions. The assignment of a prototrophic lifestyle was also part of the model predictions, but we did not detect a high capacity to synthesize *de novo* low molecular-weight essential compounds in any MAST species, which might further support their dependency on phagocytosis.

Challenges in the quest for exclusive phagotrophic genetic tool-kits: Peptidases, as example.

As comparative genomics suggested that the MAST species investigated here were phagotrophs, we focused on genes putatively participating in the phagocytosis process. A previous study suggested distinctive functional capacities among heterotrophic picoeukaryotes, including some MASTs, related with glycoside hydrolases [17]; here we emphasized the role of peptidases. As anticipated, peptidases appeared in every Stramenopile genome tested. However, what was not expected is that both the number of peptidases per genome or the types of peptidases did not differ among trophic styles. The weak clustering of species by trophic strategy based on OGs could be due to the fact that species that share trophic role tend to be closer phylogenetically, which may cause that the same peptidase family formed different OGs (Fig S3). Correcting this effect by grouping OGs from the same peptidase family, we lose any pattern relating peptidases and trophic styles (Fig. 4). Thus, the amount and types of peptidases were similar in phagotrophic, phototrophic and osmotrophic species. This is in agreement with the fact that all eukaryotic species contain lysosome-related organelles used in autophagic process that promote the turnover and degradation of their own proteins. Therefore, it is unlikely to find distinct types of peptidases exclusively associated to phagotrophy.

High presence of V-PPases in MAST genomes

Extending our research towards the vacuole acidification, we focused on two widely known proton pumps: V-ATPases and V-PPases. V-ATPases are considered to be ubiquitous components of eukaryotic organisms [67, 68]. Accordingly, these genes were found in all Stramenopiles with complete genomes and in about half of the MASTs. Likely, their absence was due to genome incompleteness as these genes seem to be widespread and constrained (a single copy) along eukaryotes. V-PPases were initially described as a proton pump that acidifies the lumen of vacuoles in land-plants and microbial eukaryotes [69, 70]. Their role has been expanded to the acidification of the lumen of acidocalcisomes [21], an organelle that accumulates polyphosphate, calcium and other cationic metals in green and red algae [71, 21] as well as in trypanosomatid and apicomplexan parasites [72]. A recent analysis on the evolution of V-PPases showed that they are absent in Opisthokonts and Amoebozoans [21], the eukaryotic supergroups in which most of our understanding of phagotrophy comes from [73]. In contrast, they are highly represented in MASTs species. The presence and, in some cases, concrete expansions of V-PPases in MASTs, suggest an important role of this protein in modulating their cellular functions. In addition, clade 3 V-PPase seems to be the more enriched in MAST as compared to other Stramenopiles with different trophic modes. It has been recently found that a clade 3 V-PPase was overexpressed in *Cafeteria burkhardae* growing exponentially by bacterivory as compared to the stationary phase [74]. This again suggests that these V-PPases, particularly from clade 3, may exert a key role in the vacuole acidification towards digestion in MASTs.

Extensive presence of rhodopsin genes in MAST genomes

Microbial rhodopsins are a diverse group of photoactive proteins capable of solar energy usage independent of plastid photosystems. They act as light-driven ion pumps or light sensors [75]. Homologs of these seven-helix transmembrane proteins have been reported in many prokaryotic taxa as well as in various eukaryotes, including marine species of diatoms, dinoflagellates [76, 28], haptophytes, cryptophytes [77], and MAST-4 [27]. Phylogenetic clades with putatively distinct functions have been identified [78]. Thus, homologs of the proton-pumping proteorhodopsins, initially found in marine bacteria [79], such as bacteriorhodopsins, halorhodopsins, sensory rhodopsins, and xanthorhodopsins [80], have been identified in archaea, bacteria, protists, and viruses [81]. Other types of microbial rhodopsins include fungal rhodopsins [82] and, lately, the channelrhodopsins known for its use in optogenetics [83]. Here we extend the finding of diverse rhodopsins within uncultured MASTs belonging to distant Stramenopile clades.

By themselves, rhodopsins are not photoactive: it is only when coupled with the light-sensitive retinal chromophore that they can convert light into an electrical response. The chromophore binds covalently to the rhodopsin domain through a Schiff base linkage with a lysine in the middle of the seventh helix [84], and we observed this conserved position at the right place in the alignments of MAST rhodopsins. The pathway of retinal generation involves two critical steps: the biosynthesis of β-carotene from its precursor lycopene, and the cleavage β-carotene into retinal [85]. As expected, the early steps of carotenoid biosynthesis to lycopene were widely present in MAST species. However, the genes involved in the last two critical steps were poorly recovered: crtY was only found in MAST-3F-sp1 and bhl in 5 of the 15 MAST species. This suggests that MASTs rely on their diet as a constant supply of retinal as these compounds cannot be synthetized *de novo*. An alternative explanation would be that MASTs take advantage of the presence of the RPE65 gene, known to catalyze the formation of retinal in vertebrates by an alternative biosynthetic pathway [86, 87].

We identified rhodopsins in most MAST species. Their absence in MAST-1D-sp1 and MAST-C-sp1 could be explained by genome incompleteness, as these were the two genomes with lowest gene recovery (<20%), but they were also absent in MAST-4B-sp1 and MAST-11-sp1, which had BUSCO scores of 67% and 46%, respectively. Particularly intriguing was the absence of rhodopsin in MAST-4B-sp1, as this gene was present in the other three MAST-4 species; further work is needed to confirm the lack of rhodopsins in MAST-4B-sp1. Five MAST species contained xanthorhodopsins, a subtype of light-driven proton pumps derived from halophilic bacteria that contain an additional light-harvesting carotenoid antenna [80]. They formed a highly supported cluster together with xanthorhodopsins of marine haptophytes and dinoflagellates [76]. Two species (MAST-4E-sp1 and MAST-7B-sp1) appeared to contain MerMAIDS rhodopsins, a new type recently discovered by metagenomics [52]. The MerMAIDs are closely related to cation-channel rhodopsins but conduct anions, which make them unique. This is the first report of MerMAIDs rhodopsins in non-photosynthetic protists. Non-MerMaiD channelrhodopsins were found in other Stramenopiles like *Hyphochytrium catenoides* [88], *Cafeteria burkhardae*, and the labyrinthulomycetes *Schizochytrium aggregatum* and *Aurantiochytrium limacinum* (Fig. 6). Channelrhodopsins are involved in light-sensing functions like phototaxis in green algae [89], or even modulate the colony conformation of the choanoflagellate *Choanoeca flexa* [90]. Thus, these rhodopsins might present a different function than xanthorhodopsins and bacteriorhodopsins, whose activity as proton pumps might complement the role of V-ATPase and potentially V-PPase in their function to acidify digestive vacuoles [28]. The fact that we observed a high expression of rhodopsin genes in MAST-4A when growing by bacterivory strongly support this hypothesis [27], but this still needs to be validated experimentally. With the observed widespread presence of rhodopsin genes and the conserved transmembrane lysine for retinal binding, we tend to believe that light may play a much more important role for the phagotrophic MAST functions than we originally thought. At the individual genomic level, it is interesting to note that some species harbour more than one type of rhodopsin suggesting independent acquisitions. Thus, the physiological cell capabilities conferred by different rhodopsin types might contribute to the various functions of MASTs in marine ecosystems. Describing them is the first step to create hypothesis and better understand functional differences between MAST species and clades.

CONCLUSION

In part due to their inability to be cultured, the physiology and ecology of many MAST species is still little understood. By genome sequencing of single eukaryotic cells, we bypassed cultivation requirements and gained insights into these neglected microbial eukaryotes. Comparative genomic analyses indicated a phagocytotic capability of these uncultured lineages, consistent with what was expected. Genes clearly involved in phagocytosis, such as proton pumps for vacuole acidification and peptidases for prey digestion, were not exclusive of phagotrophic species, as they were equally represented in phototrophic and osmotrophic species. However, the remarkable presence of different types of V-PPases and rhodopsins suggests that these proton pumps might play a crucial role in different MAST species. Besides acidifying food vacuoles, a parallel scenario could be that MAST species couple rhodopsins proton pumping with the production of PPi thanks to V-PPases. This coupled pathway would confer them an alternative energy source, as occurs in glucose metabolism of the parasitic *Entamoeaba histolytica* that uses PPi instead of ATP [91]. A better clue of the involvement of proton pumps, digestive enzymes and rhodopsins in phagocytosis is needed and new evidences can be derived from gene expression studies with cultured species [74] or natural assemblages [27]. Finally, even though the physiological role of rhodopsins in MASTs still needs to be elucidated, their wide distribution and conserved functional structure suggest that light could play an unexpected role in phagotrophic MAST species, contributing to vacuole acidification, mediating phototaxis, or even providing alternative energy sources. This light usage by MAST species is consistent with the fact that they are restricted to the upper photic region of the oceans [92]. Overall, our data reveal that the MAST species analyzed contain a high metabolic plasticity that might facilitate to thrive in the oceans as very abundant bacterial grazers.

ACKNOWLEDGMENTS

This work was supported by the EU project SINGEK (H2020-MSCA-ITN-2015-675752). We thank Michiel Van Bel for his patience and time managing the Plaza instance. A special thanks to Oded Béjà for his expertise and eyes for the rhodopsins in eukaryotes.

SUPPLEMENTARY MATERIAL

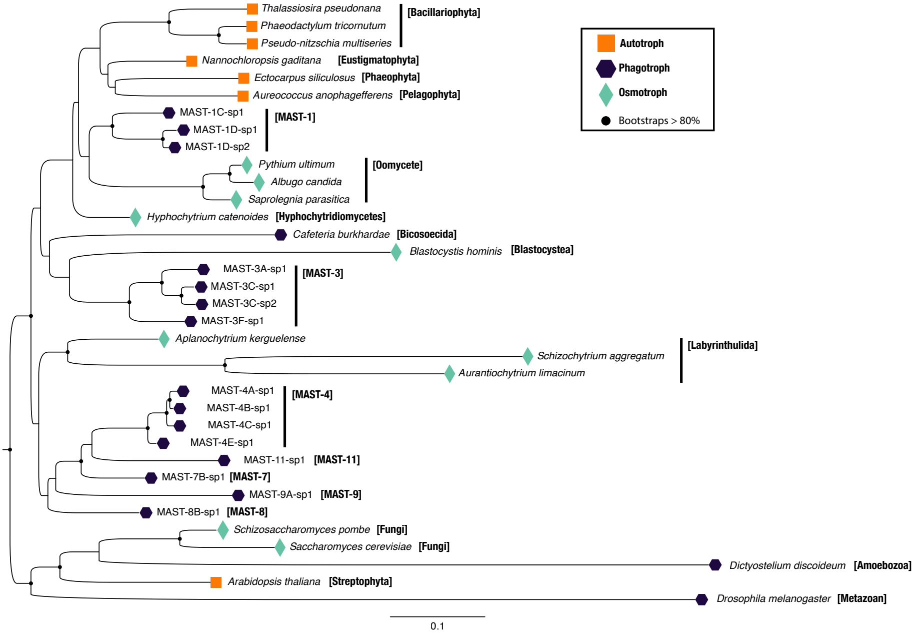

Thalassiosira pseudonana
Phaeodactylum tricornutum
Pseudo-nitzschia multiseries
[Bacillariophyta]
Nannochloropsis gaditana [Eustigmatophyta]
Ectocarpus siliculosus [Phaeophyta]
Aureococcus anophagefferens [Pelagophyta]
MAST-1C-sp1
MAST-1D-sp1
MAST-1D-sp2
[MAST-1]
Pythium ultimum
Albugo candida
Saprolegnia parasitica
[Oomycete]
Hyphochytrium catenoides [Hyphochytridiomycetes]
Cafeteria burkhardae [Bicosoecida]
Blastocystis hominis [Blastocystea]
MAST-3A-sp1
MAST-3C-sp1
MAST-3C-sp2
MAST-3F-sp1
[MAST-3]
Aplanochytrium kerguelense
Schizochytrium aggregatum
Aurantiochytrium limacinum
[Labyrinthulida]
MAST-4A-sp1
MAST-4B-sp1
MAST-4C-sp1
MAST-4E-sp1
[MAST-4]
MAST-11-sp1 [MAST-11]
MAST-7B-sp1 [MAST-7]
MAST-9A-sp1 [MAST-9]
MAST-8B-sp1 [MAST-8]
Schizosaccharomyces pombe [Fungi]
Saccharomyces cerevisiae [Fungi]
Dictyostelium discoideum [Amoebozoa]
Arabidopsis thaliana [Streptophyta]
Drosophila melanogaster [Metazoan]
0.1
Autotroph
Phagotroph
Osmotroph
Bootstraps > 80%

Figure S1. Prototrophy prediction of MAST species using a comparative genomics model [46]. PCA plot placing genomes based on their genes associated to GO categories defining a prototrophic lifestyle.

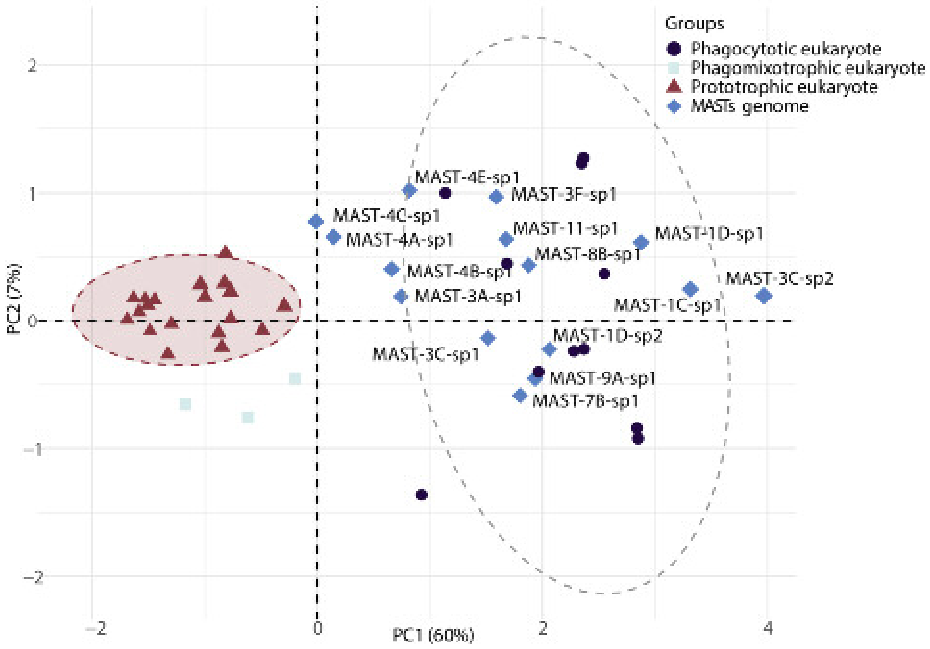


Figure S2. Phylogenetic tree of the taxa used for comparative genomics analysis, including the 15 uncultured MAST species, using the 18S rDNA gene. The tree was generated with IQTREE using 1000 trees for topology and 1000 trees for bootstrapping. Five non-stramenopile taxa were used as outgroup. Eukaryotic species were assigned to a trophic lifestyle.

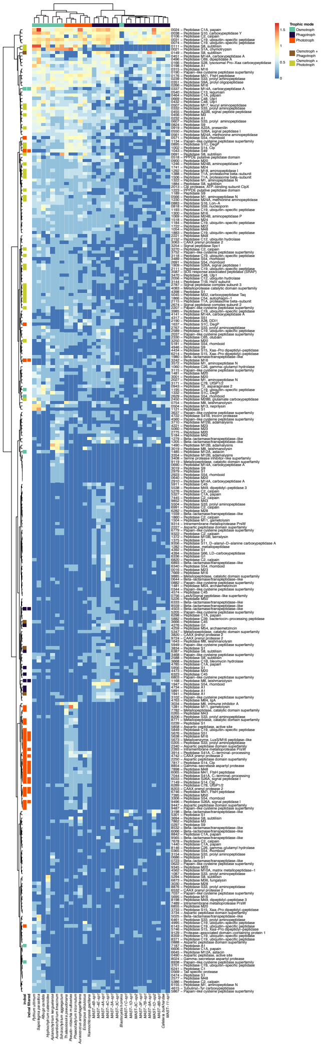


Figure S3. Distribution and abundance (log-transformed number of genes) of OGs annotated as peptidases in the 30 stramenopile genomes. Taxa are grouped according to their trophic strategy (upper part of the graph), while some of the OG clusters also indicate a given trophic lifestyle, as marked by the accumulation of OGs with IndVal scores. Filtered IndVal indicate those OGs which IPR code was not found within the other IndVal sets.

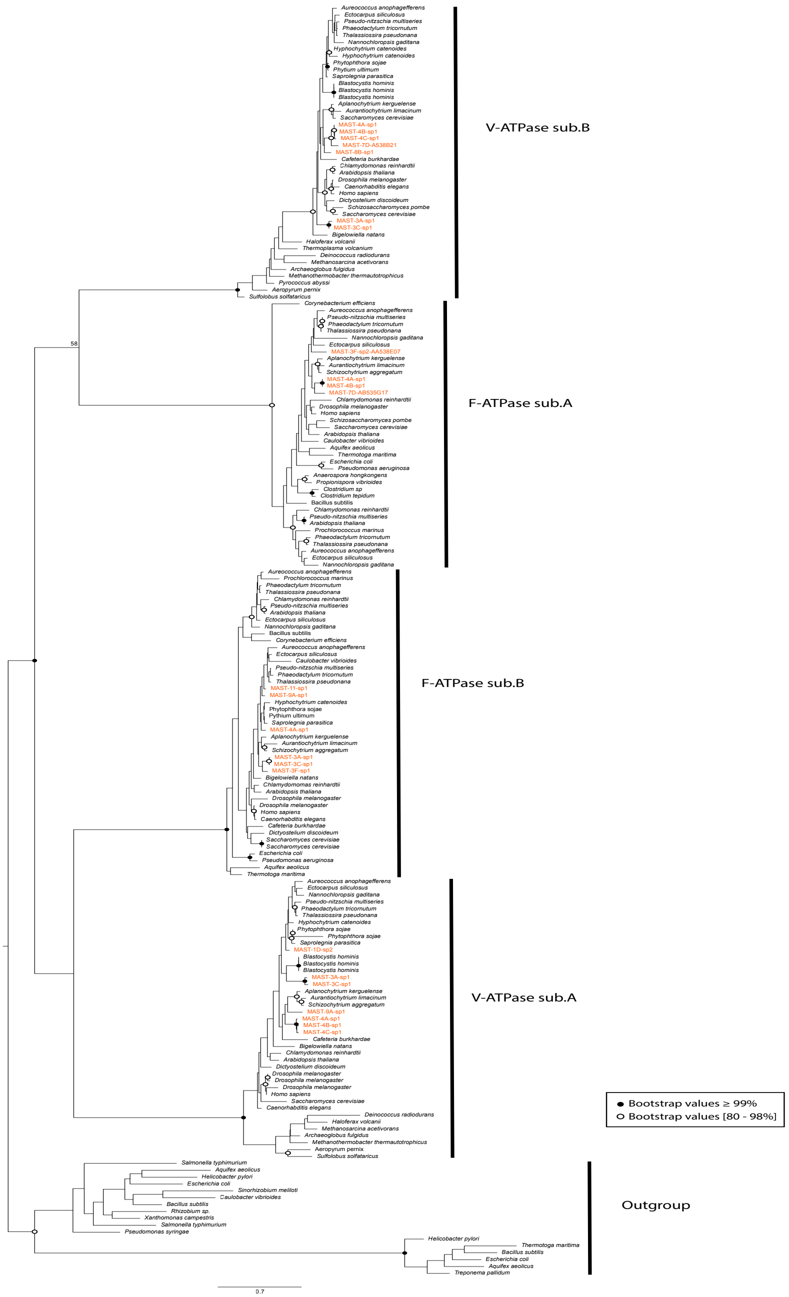


Figure S4. Phylogenetic tree of V-ATPases and the related F-ATPases genes constructed from recent bibliographical references (see Material and Methods). MASTs lineages are represented in orange. Values at nodes correspond to bootstraps > 80%.

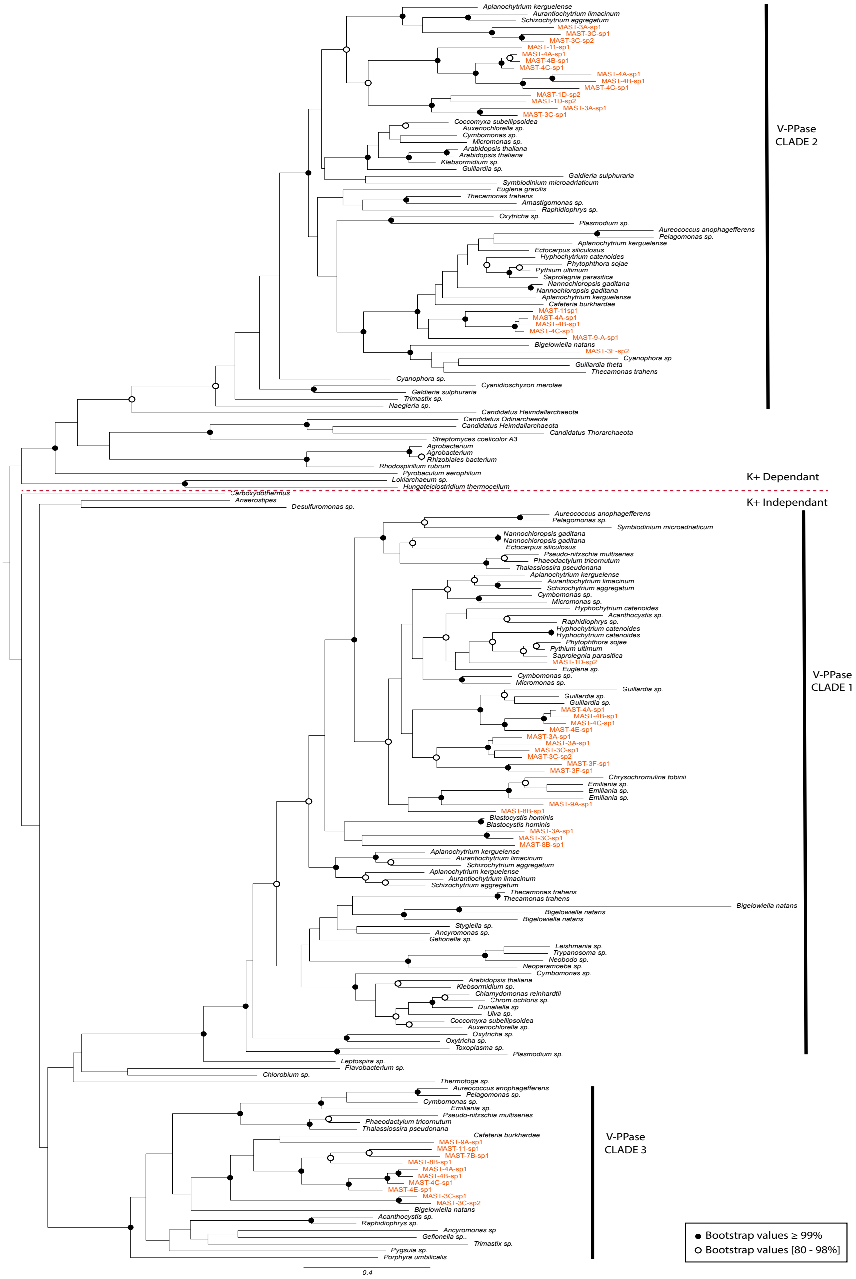


Figure S5. Phylogenetic tree of V-PPases genes constructed from recent bibliographical references (see Material and Methods). MASTs lineages are represented in orange. Values at nodes correspond to bootstraps > 80%.

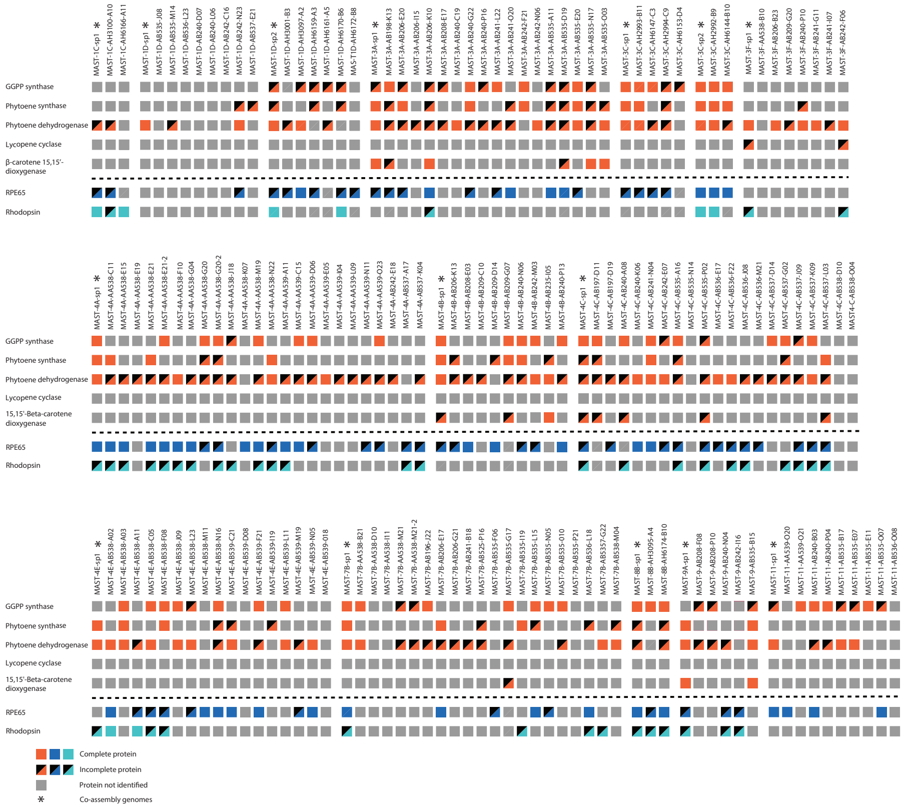


Figure S6. Presence of the genes needed for retinal biosynthesis in every individual MAST SAG. The presence of an enzyme for an alternative pathway (RPE65), as well as the presence of rhodopsin genes, is also indicated.

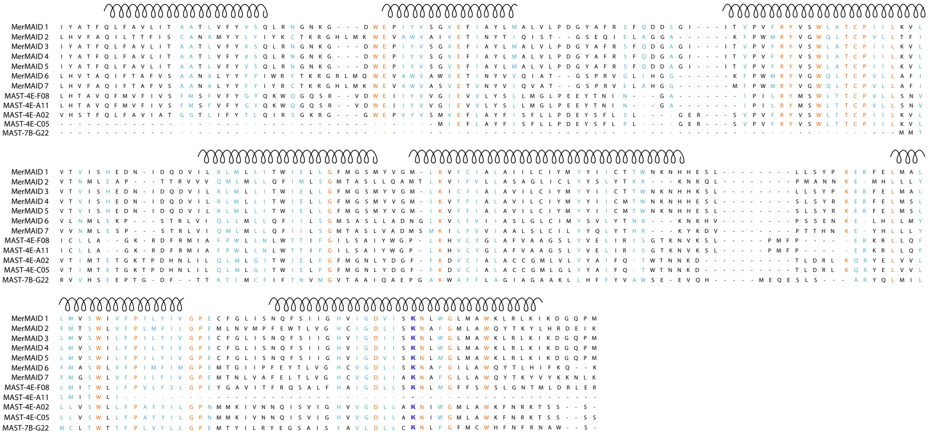


Figure S7. Sequences alignment of MerMAIDs channelrhodopsins. Highly conserved amino acids are shown in orange (identical) and light blue (in more than 60% of the sequences). The α-helices 1-7 were determined based on a previous publication [52]. The lysine Schiff base for retinal attachment found in the 7th helix is identified in dark blue.

REFERENCES

1. Bar-On YM, Milo R. The Biomass Composition of the Oceans: A Blueprint of Our Blue Planet. *Cell* 2019; 179: 1451–1454.

2. Field CB. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* 1998; 281: 237–240.

3. Zinger L, Gobet A, Pommier T. Two decades of describing the unseen majority of aquatic microbial diversity: Sequencing aquatic microbial diversity. *Mol Ecol* 2012; 21: 1878–1896.

4. Massana R, Castresana J, Balagué V, Guillou L, Romari K, Groisillier A, et al. Phylogenetic and Ecological Analysis of Novel Marine Stramenopiles. *Appl Environ Microbiol* 2004; 70: 3528–3534.

5. del Campo J, Balagué V, Forn I, Lekunberri I, Massana R. Culturing Bias in Marine Heterotrophic Flagellates Analyzed Through Seawater Enrichment Incubations. *Microb Ecol* 2013; 66: 489–499.

6. Massana R, del Campo J, Sieracki ME, Audic S, Logares R. Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J* 2014; 8: 854–866

7. Andersen KH, Aksnes DL, Berge T, Fiksen Ø, Visser A. Modelling emergent trophic strategies in plankton. *J Plankton Res* 2015; 37: 862–868.

8. Derelle R, López-García P, Timpano H, Moreira D. A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). *Mol Biol Evol* 2016; 33: 2890–2898.

9. Gómez F, Moreira D, Benzerara K, López-García P. Solenicola setigera is the first characterized member of the abundant and cosmopolitan uncultured marine stramenopile group MAST-3: Solenicola belongs to uncultured stramenopiles MAST-3. *Environ Microbiol* 2011; 13: 193–202.

10. Massana R, Unrein F, Rodríguez-Martínez R, Forn I, Lefort T, Pinhassi J, et al. Grazing rates and functional diversity of uncultured heterotrophic flagellates. *ISME J* 2009; 3: 588–596.

11. Stepanauskas R. Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* 2012; 15: 613–620.

12. Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. *Nat Rev Genet* 2016; 17: 175–188.

13. Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, et al. Single-Cell Genomics Reveals Organismal Interactions in Uncultivated Marine Protists. *Science* 2011; 332: 714–717.

14. Roy RS, Price DC, Schliep A, Cai G, Korobeynikov A, Yoon HS, et al. Single cell genome analysis of an uncultured heterotrophic stramenopile. *Sci Rep* 2015; 4: 4780.

15. Yilmaz S, Singh AK. Single cell genome sequencing. *Curr Opin Biotechnol* 2012; 23: 437–443.

16. Mangot J-F, Logares R, Sánchez P, Latorre F, Seeleuthner Y, Mondy S, et al. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci Rep* 2017; 7: 41498.

17. Seeleuthner Y, Mondy S, Lombard V, Carradec Q, Pelletier E, et al. Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nat Commun* 2018; 9: 310.

18. Rosales C, Uribe-Querol E. Phagocytosis: A Fundamental Process in Immunity. *BioMed Res Int* 2017; 2017: 1–18.

19. Underhill DM, Ozinsky A. Phagocytosis of Microbes: Complexity in Action. *Annu Rev Immunol* 2002; 20: 825–852.

20. Harikumar P, Reeves JP. The Lysosomal Proton Pump. In: Poste G, Crooke ST (eds). *New Insights into Cell and Membrane Transport Processes*. 1986. Springer US, Boston, MA, pp 61–74.

21. Goodenough U, Heiss AA, Roth R, Rusch J, Lee J-H. Acidocalcisomes: Ultrastructure, Biogenesis, and Distribution in Microbial Eukaryotes. *Protist* 2019; 170: 287–313.

22. Drobny M, Fischer-Schliebs E, Lüttge U, Ratajczak R. Coordination of V-ATPase and V-PPase at the Vacuolar Membrane of Plant Cells. In: Esser K, Lüttge U, Beyschlag W, Hellwig F (eds). *Progress in Botany*. 2003. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 171–216.

23. Kandori H. Ion-pumping microbial rhodopsins. *Front Mol Biosci* 2015; 2.

24. Govorunova EG, Sineshchekov OA, Li H, Spudich JL. Microbial Rhodopsins: Diversity, Mechanisms, and Optogenetic Applications. *Annu Rev Biochem* 2017; 86: 845–872.

25. Ruiz-González MX, Marín I. New Insights into the Evolutionary History of Type 1 Rhodopsins. *J Mol Evol* 2004; 58: 348–358.

26. Sharma AK, Spudich JL, Doolittle WF. Microbial rhodopsins: functional versatility and genetic mobility. *Trends Microbiol* 2006; 14: 463–469.

27. Labarre A, Obiol A, Wilken S, Forn I, Massana R. Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates. *Limnol Oceanogr* 2020; 65.

28. Slamovits CH, Okamoto N, Burri L, James ER, Keeling PJ. A bacterial proteorhodopsin proton pump in marine eukaryotes. *Nat Commun* 2011; 2: 183.

29. Sieracki ME, Poulton NJ, Jaillon O, Wincker P, de Vargas C, Rubinat-Ripoll L, et al. Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Sci Rep* 2019; 9: 6025.

30. Latorre F, Deutschmann IM, Labarre A, Obiol A, Krabberød A, Pelletier E, et al. Evolutionary diversification of tiny ocean predators. In preparation.

31. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; 30: 2114–2120

32. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J Comput Biol* 2012; 19: 455–477.

33. West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. Genome-reconstruction for eukaryotes from complex natural microbial communities. 2017. Genomics. *bioRxiv* 2017; 171355.

34. Laetsch DR, Blaxter ML. BlobTools: Interrogation of genome assemblies. *F1000Res* 2017; 6: 1287.

35. Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int J Syst Evol Microbiol* 2007; 57: 81–91.

36. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 2013; 29: 1072–1075.

37. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 2007; 23: 1061–1067.

38. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015; 31: 3210–3212.

39. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res* 2008; 18: 1979–1990.

40. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* 2007; 18: 188–196.

41. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 2011; 27: 757–763.

42. Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, et al. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucl Acids Res* 2018; 46: D1190–D1196.

43. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015; 12: 59–60.

44. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 2015; 16: 157.

45. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014; 30: 1236–1240.

46. Burns JA, Pittis AA, Kim E. Gene-based predictive models of trophic modes suggest Asgard archaea are not phagocytotic. *Nat Ecol Evol* 2018; 2: 697–704.
47. Oksanen AJ, Blanchet FG, Kindt R, et al Package "vegan." R package version 2.4-3, https://CRAN.R-project.org/package=vegan

48. Cáceres MD, Legendre P. Associations between species and groups of sites: indices and statistical inference. Ecology 2009; 90: 3566–3574.

49. Kolde , R. (2019) pheatmap: Pretty Heatmaps. R package version 1.0.12. https://CRAN.R-project.org/package=pheatmap

50. Mulkidjanian AY, Makarova KS, Galperin MY, Koonin EV. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat Rev Microbiol* 2007; 5: 892–899

51. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, et al. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proc Natl Acad Sci USA* 2019; 116: 20574–20583.

52. Oppermann J, Fischer P, Silapetere A, Liepe B, Rodriguez-Rozada S, Flores-Uribe J, et al. MerMAIDs: a family of metagenomically discovered marine anion-conducting and intensely desensitizing channelrhodopsins. *Nat Commun* 2019; 10: 3315.

53. Boeuf D, Audic S, Brillet-Guéguen L, Caron C, Jeanthon C. MicRhoDE: a curated database for the analysis of microbial rhodopsin diversity and evolution. *Database* 2015; 2015: bav080.

54. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 2013; 30: 772–780.

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

56. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* 2015; 32: 268–274.

57. Andersen RA. Biology and systematics of heterokont and haptophyte algae. *Am J Bot* 2004; 91: 1508–1522.

58. Massana R. Eukaryotic Picoplankton in Surface Oceans. *Annu Rev Microbiol* 2011; 65: 91–110.

59. de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, et al. Eukaryotic plankton diversity in the sunlit ocean. *Science* 2015; 348: 1261605–1261605.

60. Castillo YM, Mangot J, Benites LF, Logares R, Kuronishi M, Ogata H, et al. Assessing the viral content of uncultured picoeukaryotes in the global-ocean by single cell genomics. *Mol Ecol* 2019; 28: 4272–4289.

61. Martinez-Garcia M, Brazel D, Poulton NJ, Swan BK, Gomez ML, Masland D, et al. Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *ISME J* 2012; 6: 703–707

62. López-Escardó D, Grau-Bové X, Guillaumet-Adkins A, Gut M, Sieracki ME, Ruiz-Trillo I. Reconstruction of protein domain evolution using single-cell amplified genomes of uncultured choanoflagellates sheds light on the origin of animals. *Phil Trans R Soc B* 2019; 374: 20190088.

63. López-Escardó D, Grau-Bové X, Guillaumet-Adkins A, Gut M, Sieracki ME, Ruiz-Trillo I. Evaluation of single-cell genomics to address evolutionary questions using three SAGs of the choanoflagellate Monosiga brevicollis. *Sci Rep* 2017; 7: 11025.

64. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, et al. The Chlamydomonas Genome Reveals the Evolution of Key Animal and Plant Functions. *Science* 2007; 318: 245–250.

65. Rubin BE, Wetmore KM, Price MN, Diamond S, Shultzaberger RK, Lowe LC, et al. The essential gene set of a photosynthetic organism. *Proc Natl Acad Sci USA* 2015; 112: E6634–E6643.

66. Yutin N, Wolf MY, Wolf YI, Koonin EV. The origins of phagocytosis and eukaryogenesis. *Biol Direct* 2009; 4: 9.

67. Mulkidjanian AY, Makarova KS, Galperin MY, Koonin EV. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat Rev Microbiol* 2007; 5: 892–899.

68. Marshansky V, Rubinstein JL, Grüber G. Eukaryotic V-ATPase: Novel structural findings and functional insights. *Biochim Biophys Acta, Bioenergetics* 2014; 1837: 857–879.

69. Ikeda M, Rahman H, Moritani C, Umami K, Tanimura Y, Akagi R, et al. A vacuolar H+-pyrophosphatase in Acetabularia acetabulum: molecular cloning and comparison with higher plants and a bacterium. *J Exp Bot* 1999; 50: 139–140.

70. Gutiérrez-Luna FM, Hernández-Domínguez EE, Valencia-Turcotte LG, Rodríguez-Sotres R. Review: “Pyrophosphate and pyrophosphatases in plants, their involvement in stress responses and their possible relationship to secondary metabolism”. *Plant Science* 2018; 267: 11–19.

71. Yagisawa F, Nishida K, Yoshida M, Ohnuma M, Shimada T, Fujiwara T, et al. Identification of novel proteins in isolated polyphosphate vacuoles in the primitive red alga Cyanidioschyzon merolae: Novel proteins comprising polyphosphate vacuoles. *Plant J* 2009; 60: 882–893.

72. Docampo R, Huang G. Acidocalcisomes of eukaryotes. *Curr Opin Cell Biol* 2016; 41: 66–72.

73. Boulais J, Trost M, Landry CR, Dieckmann R, Levy ED, Soldati T, et al. Molecular characterization of the evolution of phagosomes. *Mol Syst Biol* 2010; 6: 423.

74. Massana R, Labarre A, López-Escardó D, Obiol A, Bucchini F, Hackl T, et al. Gene expression during bacterivorous growth of a widespread marine heterotrophic flagellate. ISME J.

75. Spudich JL, Jung K-H. Microbial Rhodopsins. In: Fersht AR (ed). *Protein Science Encyclopedia*. 2008. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, p mp16.

76. Lin S, Zhang H, Zhuang Y, Tran B, Gill J. Spliced leader-based metatranscriptomic analyses lead to recognition of hidden genomic features in dinoflagellates. *Proc Natl Acad Sci USA* 2010; **107**: 20033–20038.

77. Marchetti A, Schruth DM, Durkin CA, Parker MS, Kodner RB, Berthiaume CT, et al. Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proc Natl Acad Sci USA* 2012; 109: E317–E325.

78. Finkel OM, Béjà O, Belkin S. Global abundance of microbial rhodopsins. *ISME J* 2013; 7: 448–451.

79. Beja O. Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. *Science* 2000; 289: 1902–1906.

80. Balashov SP. Xanthorhodopsin: A Proton Pump with a Light-Harvesting Carotenoid Antenna. *Science* 2005; 309: 2061–2064.

81. Bratanov D, Kovalev K, Machtens J-P, Astashkin R, Chizhov I, Soloviov D, et al. Unique structure and function of viral rhodopsins. *Nat Commun* 2019; 10: 4939.

82. Bieszke JA, Spudich EN, Scott KL, Borkovich KA, Spudich JL. A Eukaryotic Protein, NOP-1, Binds Retinal To Form an Archaeal Rhodopsin-like Photochemically Reactive Pigment. *Biochemistry* 1999; 38: 14138–14145.

83. Kandori H. Biophysics of rhodopsins and optogenetics. *Biophys Rev* 2020.

84. Ernst OP, Lodowski DT, Elstner M, Hegemann P, Brown LS, Kandori H. Microbial and Animal Rhodopsins: Structures, Functions, and Molecular Mechanisms. *Chem Rev* 2014; 114: 126–163.

85. Lohr M. Carotenoid metabolism in phytoplankton. In: Roy S, Llewellyn C, Egeland ES, Johnsen G (eds). *Phytoplankton Pigments*. 2011. Cambridge University Press, Cambridge, pp 113–162.

86. Redmond. TM, Yu S, Lee E, Bok D, Hamasaki D, Chen N, et al. Rpe65 is necessary for production of 11-cis-vitamin A in the retinal visual cycle. *Nat Genet* 1998; 20: 344–351.

87. Redmond TM. Focus on Molecules: RPE65, the visual cycle retinol isomerase. *Exp Eye Res* 2009; 88: 846–847.

88. Leonard G, Labarre A, Milner DS, Monier A, Soanes D, Wideman JG, et al. Comparative genomic analysis of the ‘pseudofungus’ *Hyphochytrium catenoides*. *Open Biol* 2018; 8: 170184.

89. Sineshchekov OA, Jung K-H, Spudich JL. Two rhodopsins mediate phototaxis to low- and high-intensity light in Chlamydomonas reinhardtii. *Proc Natl Acad Sci USA* 2002; 99: 8689–8694.

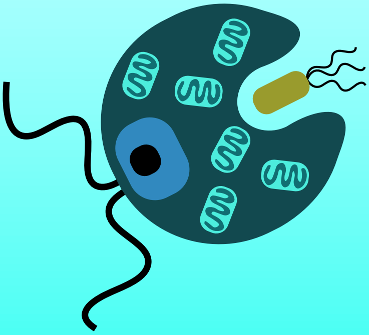
90. Brunet T, Larson BT, Linden TA, Vermeij MJA, McDonald K, King N. Light-regulated collective contractility in a multicellular choanoflagellate. *Science* 2019; 366: 326–334.

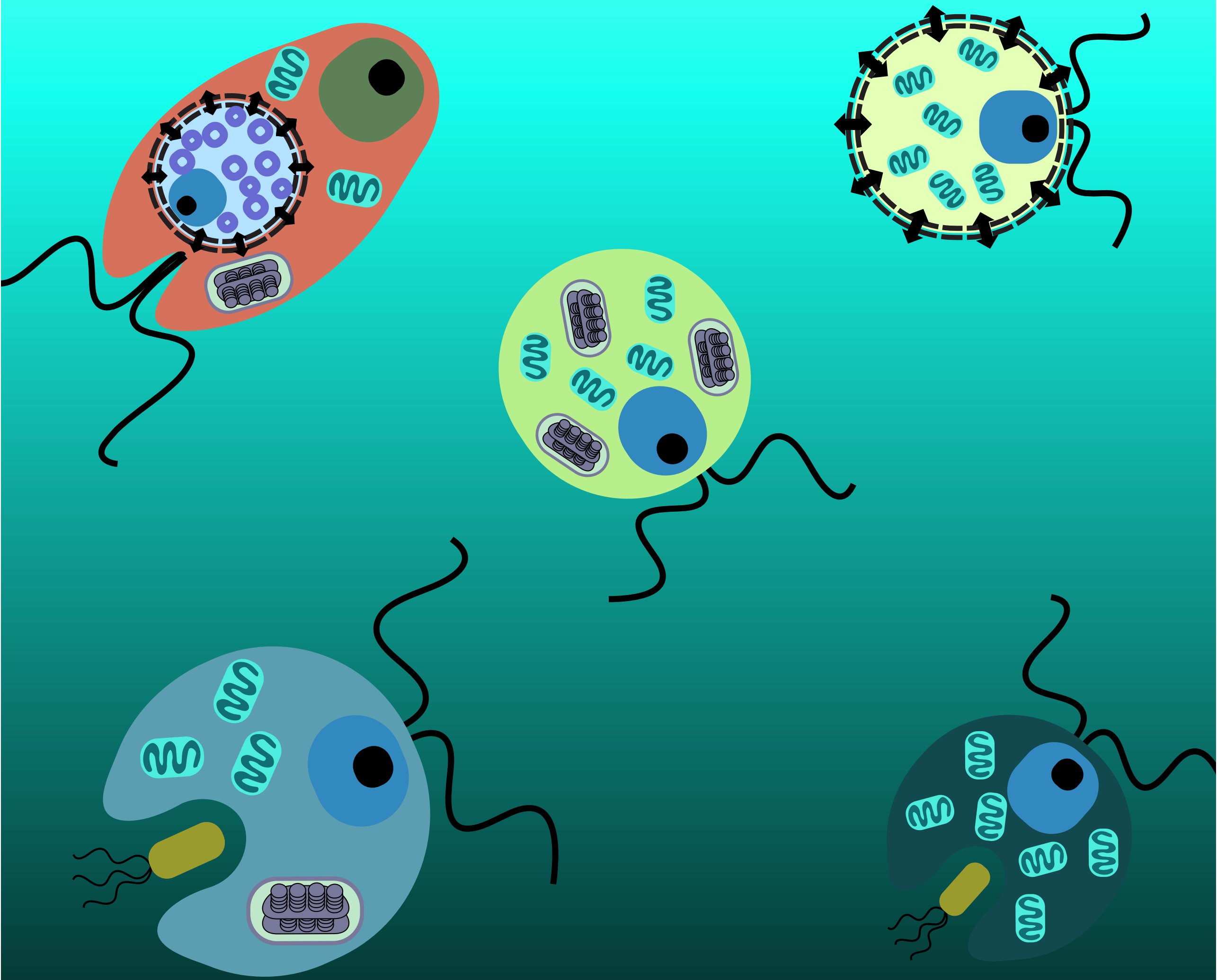
91. Saavedra E, Encalada R, Vázquez C, Olivos-García A, Michels PAM, Moreno-Sánchez R. Control and regulation of the pyrophosphate-dependent glucose metabolism in Entamoeba histolytica. *Mol Biochem Parasitol* 2019; 229: 75–87.

92. Obiol A, Giner CR, Sánchez P, Duarte CM, Acinas SG, Massana R. A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Mol Ecol Resour* 2020; 20: 718–731.

Chapter 3

Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates





Published in Limnology and Oceanography
Special Issue: Linking Metagenomics to Aquatic Microbial Ecology and Biogeochemical Cycles

Aurelie Labarre, Aleix Obiol, Susanne Wilken, Irene Forn & Ramon Massana

ABSTRACT

Environmental molecular sequencing has revealed an abundance of microorganisms that were previously unknown, mainly because most had not been cultured in the laboratory. Within this novel diversity, there are the uncultured MAST clades (MArine STramenopiles), which are major components of marine heterotrophic flagellates (HFs) thought to be active bacterial grazers. In this study, we investigated the gene expression of natural HFs in a mixed community where bacterivory was promoted. Using fluorescence in situ hybridization and 18S rDNA derived from metatranscriptomics, we followed the taxonomic dynamics during the incubation, and confirmed the increase in relative abundance of different MAST lineages. We then used single cell genomes of several MAST species to gain an insight into their most expressed genes, with a particular focus on genes related to phagocytosis. The genomes of MAST-4A and MAST-4B were the most represented in the metatranscriptomes, and we identified highly expressed genes of these two species involved in motility and cytoskeleton remodeling, as well as many lysosomal enzymes. Particularly relevant were the cathepsins, which are characteristic digestive enzymes of the phagolysosome and the rhodopsins, perhaps used for vacuole acidification. The combination of single cell genomics and metatranscriptomics gives insights on the phagocytic capacity of uncultured and ecologically relevant HF species.

INTRODUCTION

Heterotrophic flagellates (HFs) are widespread throughout the eukaryotic tree of life and may represent the most ancient eukaryote lifestyle (Cavalier-Smith 2006; Jürgens and Massana 2008; Adl et al. 2019). These colorless flagellated protists are important consumers of primary and secondary production in marine ecosystems (Arndt et al. 2000; Worden et al. 2015), and play a pivotal role in microbial food webs by ensuring the recycling of nutrients. Although they make a significant contribution, it is difficult to assess their diversity because they cannot be easily differentiated by microscopy and most remain uncultured. Moreover, HFs are often ignored in quantitative studies because they are less abundant than photosynthetic protists, and are not well represented in sequence databases, especially of sequenced genomes (del Campo et al. 2014). To study the gene expression and elucidate functional characteristics of these diverse and complex assemblages, metatranscriptomics provides a promising but also challenging opportunity.

Most HFs are considered to be bacterivorous, that is, they consume bacteria by phagocytosis. The engulfment and digestion of a foreign organism as prey, is an important nutritional process in many species of unicellular eukaryotes, and consists of three main steps: (1) motility and prey recognition, (2) internalization, formation, and maturation of the phagosome, and (3) digestion of prey within the phagolysosome (Levin et al. 2016; Uribe-Querol and Rosales 2017). In phagocytosis, prey is internalized by invagination of the plasma membrane to form an intracellular vacuole known as the phagosome (Niedergang and Grinstein 2018). Engulfment is controlled by the actin cytoskeleton and coordinated by phagocytic receptors that activate the GTPases Rac, Rho, and Cdc42 genes (Vieira et al. 2002; Niedergang and Grinstein 2018). The phagosomes then undergo a maturation process, acquire different proteins (like the Rab GTPase) (Rink et al. 2005; Fairn and Grinstein 2012), and become acidified. Mature phagosomes become functional phagolysosomes by fusing with lysosomes, which provide digestive enzymes and further acidify the environment to optimize the performance of these enzymes. Thus, a mature phagolysosome is characterized by the presence of a range of lysosomal acid hydrolases such as proteases, lysozymes, and lipases (Vieira et al. 2002; Fairn and Grinstein 2012). Most of our understanding of phagocytosis at the genomic level is limited to studies on a few species of eukaryotes: several mammals (Boulais et al. 2010), choanoflagellates (Dayel and King 2014), green algae (Burns et al. 2015), and amoebae (Okada et al. 2005). The molecular machinery of phagocytosis in environmentally relevant HFs has not been described, which prevents us from quantitatively evaluating particular traits representative for this lifestyle.

A substantial part of marine HF assemblages are MArine STramenopiles (MASTs) (Massana et al. 2014), a set of largely uncultured clades within Stramenopiles, a taxa-rich supergroup including autotrophic (Ochrophyta) and heterotrophic (Pseudofungi, Sagenista, and Opalozoa) high-rank lineages. To date, 18 MAST clades have been identified within these three heterotrophic lineages, and each one may have a distinct ecological niche (Massana et al. 2014). Some MAST lineages are widespread and highly abundant in the surface ocean (de Vargas et al. 2015; Mangot et al. 2018). Most MASTs are assumed to be bacterial feeders, as this has been demonstrated in a few of them (e.g., MAST-1 and MAST-4) by microscopic inspection of bacterial preys inside the cells (Massana et al. 2009; Lin et al. 2012). However, it is not clear that all MASTs can perform phagocytosis, nor has it been confirmed which sets of genes and proteins are used for this process. Because of the difficulty of cultivating the most relevant species of marine HFs, it is not possible to perform direct physiological and gene expression studies on single species, so there has been little progress in understanding the genetic basis of phagocytosis in these organisms.

In this article, we circumvent the need for culture-based approaches using a combination of molecular tools to study a set of MAST species growing in near-natural conditions. First, we established an unamended incubation of a coastal surface sample in which active HF cells were growing by feeding on bacteria (Massana et al. 2006), and obtained metatranscriptomic data at several time points during the incubation. Second, by using the cell counts and the 18S rDNA signatures of the metatranscriptomes, we analyzed the temporal dynamics of the taxonomic groups succeeding in the incubation, tracing their positive or negative response.

Third, using genomic data obtained by single cell genomics (SCG) (Mangot et al. 2017), we recruited transcripts from several MAST species, allowing the analysis of expressed genes likely contributing to the phagocytosis pathway. Thus, the combination of metatranscriptomics with single cell genomes, available for a few dominant species, allowed us for the first time to study the expression profile of uncultured HFs in natural assemblages.

MATERIALS AND METHODS

Growth of marine HFs in an unamended incubation

Approximately 100 L of surface seawater were sampled from Blanes Bay (41°40′N, 2°48′E) on 4th July 2017, prefiltered by gravity through a nylon mesh of 200 μm, and transported to the institute within 2 h. In the lab, 50 L of seawater were gravity-filtered through 3 μm pore size polycarbonate filters (47 mm diameter) into a polycarbonate bottle (Nalgene). The bottle was incubated for 5 d in the dark at 24°C, the in situ temperature of the sampling site, and sampled twice a day for cell counts and once a day for molecular data. For total cell counts, aliquots were fixed with glutaraldehyde and stained with 4′,6-diamidino-2-phenylindole (DAPI). Cell counts of heterotrophic bacteria, *Synechococcus*, and phototrophic and heterotrophic flagellates (2–3 μm in size) were obtained using epifluorescence microscopy, with excitation by UV radiation (DAPI stained DNA signal) and blue light (to confirm the presence of chlorophyll) (Giner et al. 2016). For cell counts by fluorescence in situ hybridization (FISH) (Amann et al. 1995), aliquots were fixed with formaldehyde and hybridized with oligonucleotide probes specific to MAST-4, MAST-7, *Minorisa minuta*, and Prymnesiophyceae, as described previously (Cabello et al. 2016; Giner et al. 2016). Samples were then examined by epifluorescence microscopy with blue light excitation. For molecular analyses, 2 L of the incubation were filtered through 0.6 μm pore size polycarbonate filters of 47 mm diameter, which were then flash frozen in liquid nitrogen and stored at −80°C until RNA extraction.

RNA extraction and Illumina sequencing

For RNA extraction, the filters were shattered and vortexed in a tube containing Power Soil beads (Mobio) as described previously (Alonso-Sáez et al. 2018). RNA extraction was performed using the RNeasy Mini Kit (Qiagen). About 60 μL of the primary RNA extract were processed using Turbo DNAse (Ambion, Turbo DNA-free kit) to completely remove residual DNA. The RNA extract was purified by ethanol precipitation, and the pellet resuspended in 40 μL 10 mmol L^{-1} TRIS. Metatranscriptomic sequencing was performed using 200–400 ng of total RNA extract. Illumina RNASeq libraries were prepared at CNAG (**https://www.cnag.crg.eu/**) using KAPA Stranded mRNA-Seq Illumina (Roche-KAPA Biosystems). The polyadenylated eukaryotic transcripts were first isolated using poly-T oligonucleotides attached to beads. Then, the mRNA was fragmented using heat and magnesium, converted to complementary DNA (cDNA) by reverse transcription, and sequencing adaptors were ligated to the ends of cDNA molecules. Ligation products were enriched by 15 polymerase chain reaction (PCR) cycles and final libraries were validated with an Agilent 2100 Bioanalyzer. Sequencing was performed at CNAG using an Illumina HiSeq 2500 (TruSeq SBS Kit v4) system, which generated 73 million paired-end reads (2 × 100 bp) for a final sequencing depth of 15 Gbp per sample. The sequence data are available as raw reads at the NCBI BioSample database (SAMN11783926).

Taxonomic characterization of the microbial community

Focusing on the entire eukaryotic domain, we built a reference database of the hypervariable V4 region of the 18S rRNA gene using sequences from SILVA (Quast et al. 2013), and from data sets of environmental marine protists based on 454 (Massana et al. 2015), and Illumina sequencing (Giner et al. 2019). The database is available at https://github.com/aleixop/eukaryotesV4. These references were assigned to several "class level" taxonomic ranks using manual curation. We used local alignment by USEARCH (Edgar 2010) to retrieve reads from the metatranscriptomes related to this reference database (> 97% similarity and > 90%

maximal score and filtered them to obtain the correct taxonomic classification: we assigned a read to a taxonomic class when all top hits belonged to that class; otherwise, the read remained unclassified. The relative abundance of a taxonomic class was obtained by dividing its number of V4 reads by the total number of V4 reads in the sample. To obtain a finer classification of reads within MAST lineages, we prepared a second reference data set using sequences classified within separate MAST lineages (Massana et al. 2014) and applied the same criteria as for read classification.

Assembling the metatranscriptome and read mapping

We performed quality/adapter trimming of the Illumina HiSeq raw reads using Trimmomatic v0.33 (Bolger et al. 2014) with default settings. We then used SortMeRNA v2.1 (Kopylova et al. 2012) to identify and remove ribosomal RNA reads by comparing all reads against the SILVA SSU and LSU rDNA databases (Quast et al. 2013) and the PR^2 database (Guillou et al. 2013), using a match score e-value of < 0.01. The resulting set of rRNA free reads allowed us to perform a de novo metatranscriptomic coassembly of the six time points sampled using Trinity (Grabherr et al. 2011) with default parameters. We then used the Burrows-Wheeler aligner BWA software (Li and Durbin 2009) to map the individual cleaned paired-end reads from each sample back to the metatranscriptome assembly, allowing up to two mismatches per read, and then we estimated the expression level of each transcript in TPM units (transcripts per million) using the Salmon software (Patro et al. 2017). We computed TPM values for each isoform defined by Trinity (Grabherr et al. 2011), and for subsequent analyses we summed the signal from all isoforms from the same gene and kept the longest isoform as the representative sequence of each gene.

Extracting MAST transcripts using functionally annotated genes from single amplified genomes (SAGs)

To perform taxonomic binning of the final metatranscriptome, we used the Bowtie2 algorithm with default parameters (Langmead and Salzberg 2012), to map the assembled transcripts to predicted open reading frames of reference genomes from 10 representative MAST species obtained by SCG (Supporting Information Table S1). The single cell amplified genomes of MAST-4A and -4E are published (Mangot et al. 2017), while others have been sequenced, assembled and annotated as part of a separate study (Labarre et al. unpubl.). Gene prediction and functional annotation was performed using Augustus (Stanke et al. 2004) and InterProScan (Jones et al. 2014). Based on the gene annotation obtained using the single cell genome, we associated each MAST-specific transcript to a gene family that has a given function. We also assigned the proteins encoded by the genomes to the eggNOG database (Huerta-Cepas et al. 2016) that provides general functional overview classified into main categories of biological metabolism.

Genome assemblies are available at **doi.org/10.6084/m9.figshare.c.4534562**.

RESULTS

Growth of marine HFs in an unamended incubation

We examined a mixed community of picoplanktonic microbes (≤3 μm) growing in a closed system, where higher trophic levels like larger predatory flagellates or ciliates had been filtered out. Using direct epifluorescence microscopy counts, we evaluated the temporal dynamics of several components of this mixed assemblage (Fig. 1a). Bacteria developed progressively during the incubation, obtaining their highest abundance at day 3 (~2 × 10^6 cells mL^{-1}), and appearing to decrease at day 4. During the 5 d of incubation, HFs increased continuously, multiplying by nearly 10-fold (from 10^3 to 10^4 cells mL^{-1}). Photosynthesis was inhibited by incubating the samples in the dark, such that the abundance of photosynthetic flagellates and *Synechococcus* decreased markedly from their initial abundance. As intended, the dark unamended incubation promoted the growth of HFs, which became the most important eukaryotic component of the system.

We constructed metatranscriptomic data for samples taken at six time points during the incubation. Despite the enrichment of mRNA in these transcripts (reverse transcription was based on the mRNA poly-A tail), we obtained many 18S rRNA reads (5–12% per sample), which we used to assess the taxonomic composition and dynamics of the assemblage by classifying individual reads to broad taxonomic classes (Fig. 1b). Initially, the samples were dominated by groups composed mainly of photosynthetic species, principally Prymnesiophyceae, Dictyochophycea, and Dinoflagellates, and these decreased markedly over time. In contrast, groups with heterotrophic taxa became more abundant, namely Choanomonada and several MAST lineages (MAST-1, -7, and -3). The increase in abundance of Chlorarachniophyta, which is a generally photosynthetic class, was due to the presence of its only known heterotrophic member, *M. minuta*.

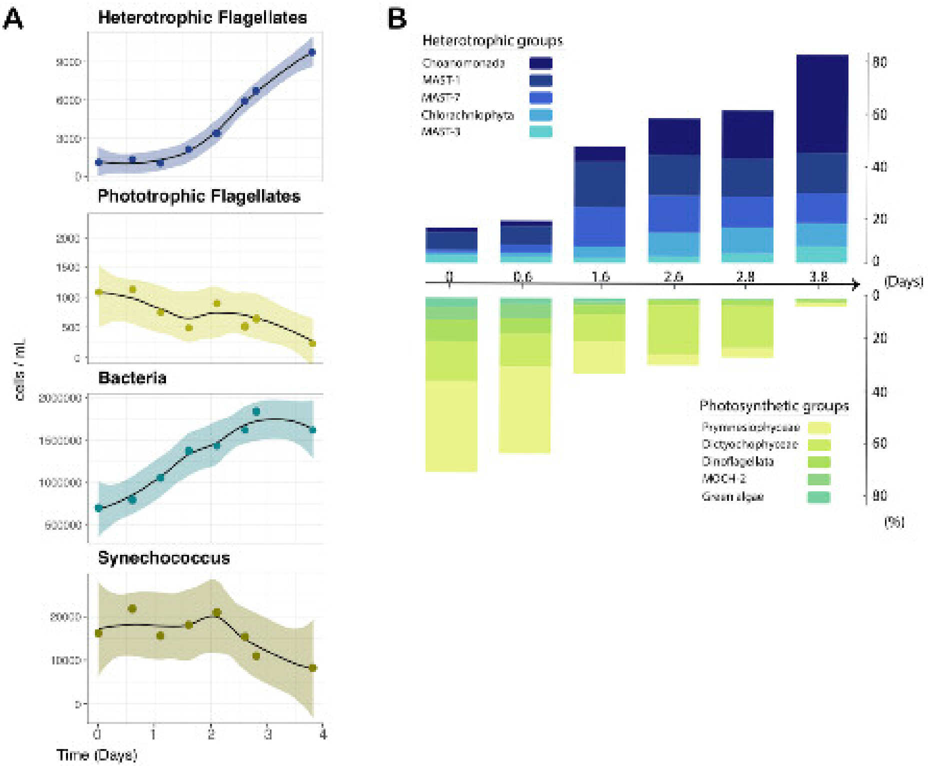


Figure 1. Temporal dynamics of the mixed microbial assemblage. (**a**) Changes in cell abundance of HFs (2–3 μm), photosynthetic flagellates (2–3 μm), heterotrophic bacteria, and *Synechococcus* obtained by epifluorescence microscopy after DAPI staining. Dots represent actual cell counts while shades show the overall trend using the estimate of the conditional mean function with the R package ggplot using a linear model (lm). (**b**) Relative abundance of the most important taxonomic groups during the enrichment as measured by their contribution to V4 18S rDNA reads in the metatranscriptomes. The heterotrophic groups are shown in the upper graph while the photosynthetic groups are in the lower graph.

We used taxonomy profiles based on 18S rRNA data to characterize the dynamics of the 40 most abundant classes, which collectively accounted for 99% of the community (Fig. 2a). The metatranscriptomic data set showed a large reduction in relative abundance of virtually all autotrophic groups, including Archaeplastida, Cryptomonadales, Prymnesiophyceae, Diatomea, Pelagophyceae, and the three MOCH lineages. In contrast, heterotrophic groups showed a more varied response to the incubation, with some decreasing in abundance (Picozoa, MAST-11, MALV-II), others remaining stable (Telonema, Katablepharids, Cercozoa), et al increasing (Choanomonada, most MAST lineages, Bicosoecida and Labyrinthulomycetes). As expected, groups containing both autotrophic and heterotrophic species did not show a clear trend. To support these observations, we targeted a few groups using FISH (Fig. 2b), and observed a marked decrease in prymnesiophytes (from 400 to 3 cells mL^{-1}), and an increase in *M. minuta*, MAST-7, and MAST-4 (increasing from 65 to 1300 cells mL^{-1}, 11 to 300 cells mL^{-1}, and 47 to 200 cells mL^{-1}, respectively). Our metatranscriptomics data revealed a very complex assemblage containing a large diversity of taxonomic groups and confirmed the growth of HFs and the decrease of photosynthetic groups.

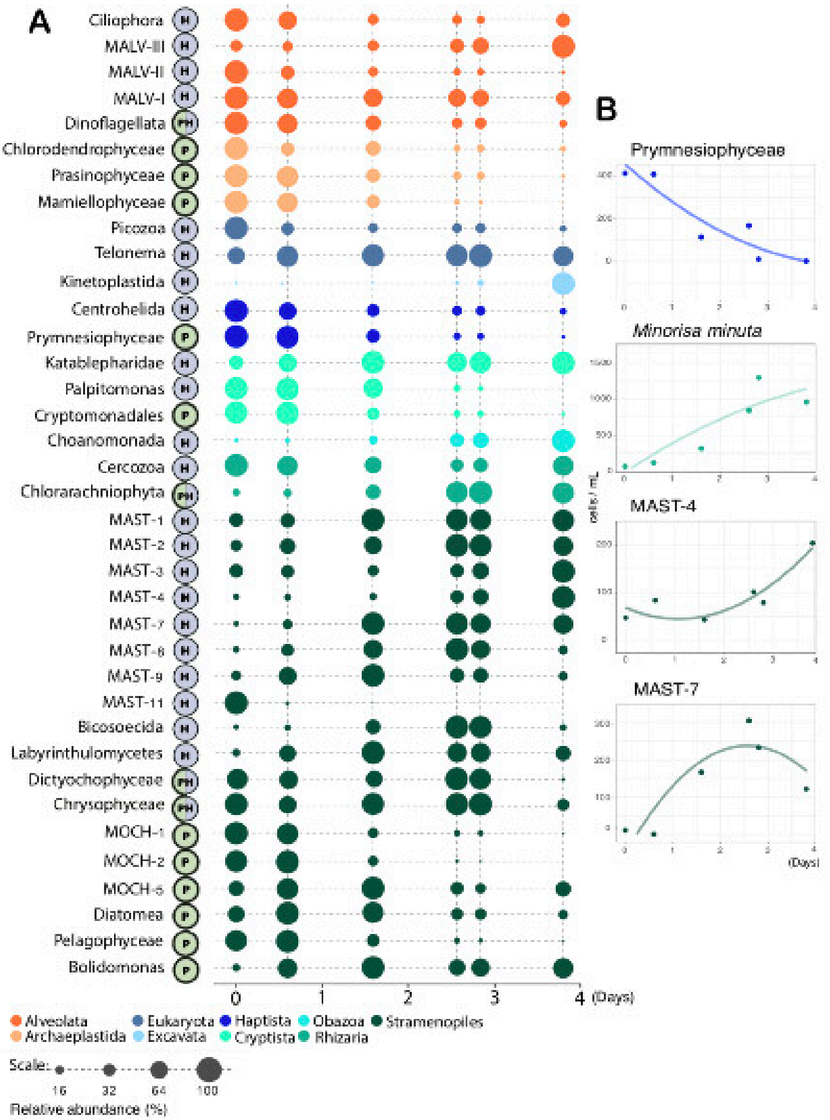


Figure 2. Temporal changes in the relative abundance of all taxonomic groups. (**a**) The relative abundance of a given group is calculated by dividing the number of 18S rDNA reads from the group by the total number of 18S rDNA reads in the sample; these values are normalized within each group (being the highest abundance scaled to 100). The main trophic mode of the groups is shown as cartoons next to their names: P (phototrophs), H (heterotrophs), PH (groups containing phototrophs and heterotrophs). (**b**) Actual cell abundances estimated by FISH for MAST-4, MAST-7, *M. minuta*, and Prymnesiophyceae.

Characterization of HFs using metatranscriptomics

After excluding rRNA reads, we built a de novo metatranscriptome assembly by merging the paired-end RNA-Seq data for the six time points sampled during incubation. We obtained 3,812,907 transcripts with an average length of 414 bp. We computed the TPM value for each transcript in every sample, and summed the values for all isoforms from the same gene, resulting in 3,338,309 transcript values. The resulting gene expression profile of the mixed community is very heterogeneous, as it represents hundreds of species from distant phylogenetic groups, each one expressing a different set of specific genes.

To assess gene expression and their putative function in individual species, we mapped the assembled transcripts to the predicted open reading frames of the SAGs of 10 MAST species. The two species that retrieved most transcripts were MAST-4A and MAST-4B (10,419 and 3789, respectively), and these had the highest expression level (Fig. 3a). Noticeably, the two closely related MAST-4C and -E retrieved very few transcripts and very little expression signal, as well as the remaining MAST genomes tested, with perhaps the exception of MAST-11 for which the expression signal was still considerable. Overall, we obtained a low recovery of transcripts mapping the SAGs, with about 0.22% of the entire metatranscriptome assigned to MAST-4A and 0.17% to MAST-4B. This low retrieval was generally consistent with the temporal changes in MAST relative abundances identified by the finest taxonomic classification of 18S rDNA reads (Supporting Information Fig. S1), which suggested little dominance of the 10 species represented by our SAGs. Thus, while MAST-7 increased in relative abundance (Figs. 1b, 2), only 144 transcripts mapped to the MAST-7A SAG (Fig. 3a), as the subclades growing in the incubation were MAST-7B, -D, and -E (Supporting Information Fig. S1).

Similarly, few transcripts were fetched using the MAST-3A and -3F SAGs (the subclades growing during incubation were -3E, -I, and -J) and MAST-9A SAG (the subclade growing was -9D). The signal represented by MAST-4 was moderate and increased during the incubation both by FISH cell counts and by 18S rDNA mapping mostly to clades-A and -B. Finally, results were not consistent in two cases: MAST-11 was barely detectable by the 18S rRNA mapping but retrieved 823 transcripts with a moderate expression profile, while MAST-1D retrieved very few transcripts but abundant 18S rDNA reads.

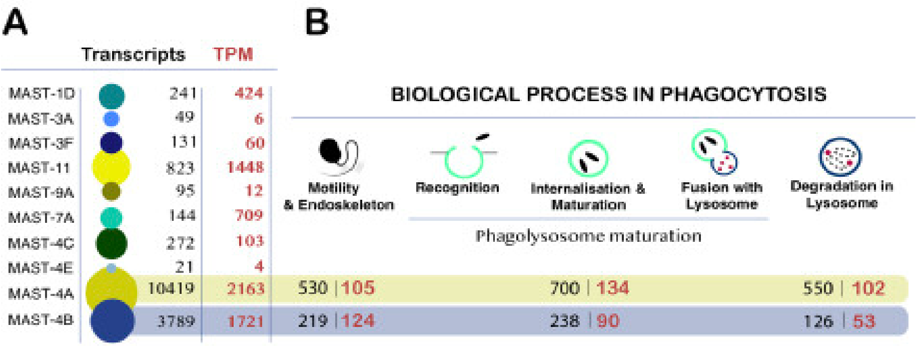
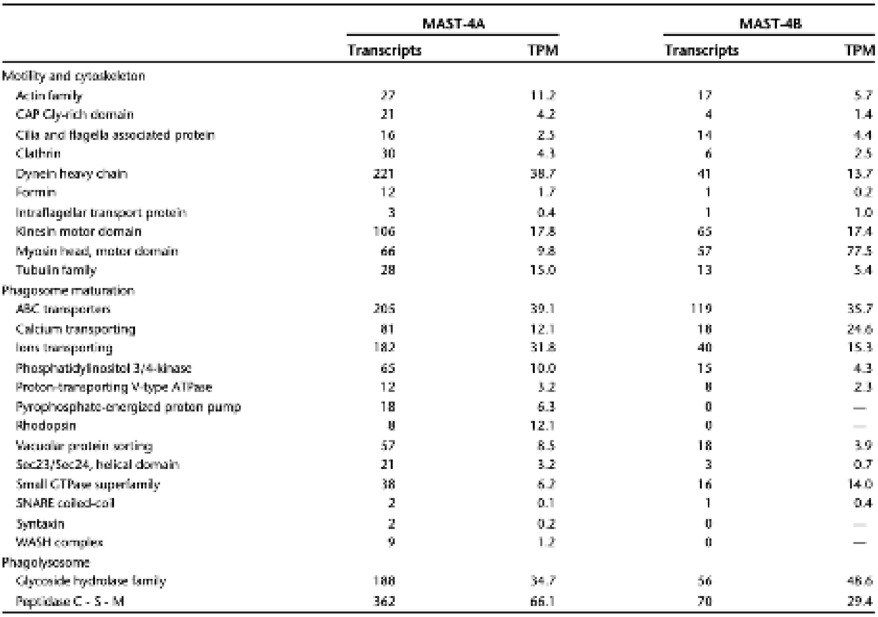


Figure 3. Mapping of the metatranscriptome toward MAST genomes. (**a**) Number of transcripts recruited using SCG of 10 MAST species and their expression level (averaged TPM values over the six time-points) shown in red. Circles illustrate the relative count of the transcripts in every species. (**b**) A schematic representation of the main steps of phagocytosis, together with the number of transcripts (and their averaged TPM values) detected in each of them within the MAST-4A and -B species.

Gene expression of two uncultured HFs

The transcripts associated to MAST-4A (10,419 transcripts) and to MAST-4B (3789) were annotated using the eggNOG database, which categorizes their functions into roles in metabolism, cellular and signaling processes, and information storage (Supporting Information Table S2). The large majority of transcripts were identified through the eggNOG database (only 1652 transcripts in MAST-4A and 707 in MAST-4B did not have a match) but many of them affiliated to uncharacterized proteins (3679 and 1544 transcripts). Most general functions were well represented by the expressed genes, being posttranslational modification and cytoskeleton the most represented categories.

The next step was to target expressed genes potentially involved in the phagocytosis pathway, focusing on the three main steps: prey recognition and motility, phagosome maturation, and degradation in lysosome (Table 1). Many of these genes were associated with motility and/or cytoskeleton functions (475 transcripts in MAST-4A and 219 in MAST-4B). Genes coding for actin and tubulin, which are structural components of the flagella and the cytoskeleton, were abundant and highly expressed in both species. We also detected genes associated with microtubule formation, such as myosin, dynein, and kinesin, as well as several flagella-associated proteins and intraflagellar transporters that are essential for flagellar growth. In both species, we found highly expressed genes likely involved in the phagolysosome maturation step (614 transcripts in MAST-4A and 238 in MAST-4B), although some have more diverse cellular functions. These include phosphatidylinositol 3/4 kinase and several proton pumps that are potentially responsible for phagosome acidification (Table 1). For example, a rhodopsin was the third most highly expressed gene in MAST-4A, along with the vacuolar pyrophosphatase and GTPase Arf type in MAST-4B, which in turn are well known as regulators of vesicular traffic and actin remodeling. Although less expressed, the presence of the SNARE complex (Soluble N-ethylmaleimide-sensitive factor activating protein receptor) is very informative, as it is responsible for intracellular membrane fusion and trafficking steps interacting with vacuolar protein sorting. Finally, we also observed the expression of a set of genes encoding digestive enzymes (437 transcripts in MAST-4A and 126 in MAST-4B), such as the glycoside hydrolase family and peptidases, especially the lysosomal proteases cysteine cathepsins. These genes were among the most expressed in MAST-4B and were also important in MAST-4A (Table 1).



	MAST-4A		MAST-4B	
	Transcripts	TPM	Transcripts	TPM
Motility and cytoskeleton				
Actin family	27	11.2	17	5.7
CAP Gly-rich domain	21	4.2	4	1.4
Cilia and flagella associated protein	16	2.5	14	4.4
Clathrin	30	4.3	6	2.5
Dynein heavy chain	221	38.7	41	13.7
Formin	12	1.7	1	0.2
Intraflagellar transport protein	3	0.4	1	1.0
Kinesin motor domain	106	17.8	65	17.4
Myosin head, motor domain	66	9.8	57	77.5
Tubulin family	28	15.0	13	5.4
Phagosome maturation				
ABC transporters	205	39.1	119	35.7
Calcium transporting	81	12.1	18	24.6
Ions transporting	182	31.8	40	15.3
Phosphatidylinositol 3/4-kinase	65	10.0	15	4.3
Proton-transporting V-type ATPase	12	3.2	8	2.3
Pyrophosphate-energized proton pump	18	6.3	0	—
Rhodopsin	8	12.1	0	—
Vacuolar protein sorting	57	8.5	18	3.9
Sec23/Sec24, helical domain	21	3.2	3	0.7
Small GTPase superfamily	38	6.7	16	14.0
SNARE coiled-coil	2	0.1	1	0.4
Syntaxin	2	0.2	0	—
WASH complex	9	1.2	0	—
Phagolysosome				
Glycoside hydrolase family	188	34.7	56	46.6
Peptidase C - S - M	362	66.1	70	29.4

Table 1. Genes involved in the phagocytosis pathway identified in MAST-4A and -4B. The number of transcripts and their averaged TPM values within each category defined within the phagocytosis pathway are displayed within the three identified steps.

Finally, as we followed the gene expression on several time points along the incubation, we explored the possibility of changes in the expression profiles of the two species. The abundance of transcripts was relatively constant over time for MAST-4A, and increased markedly for MAST-4B (Fig. 4a). This was consistent with the 18S rDNA signal (Supporting Information Fig. S1), and suggested that the proportion of MAST-4A cells remained stable in the community, while that of MAST-4B cells increased. Then we focused on the subset of the 100 most highly expressed genes, with the expression signal normalized within each species to account for the changes of species abundance along the incubation (Fig. 4b). No clear changes seemed to occur during time, with the few temporal clusters appearing unrelated to any gene function in particular. Therefore, it seemed that the two species were transcriptionally at a similar stage throughout the incubation. Interestingly, in this subset of most highly expressed genes, we detected several that were related to the three main steps of phagocytosis described, and more specifically the presence of digestive enzymes as the cathepsins, confirming the relevance of this process for the flagellate ecophysiology.

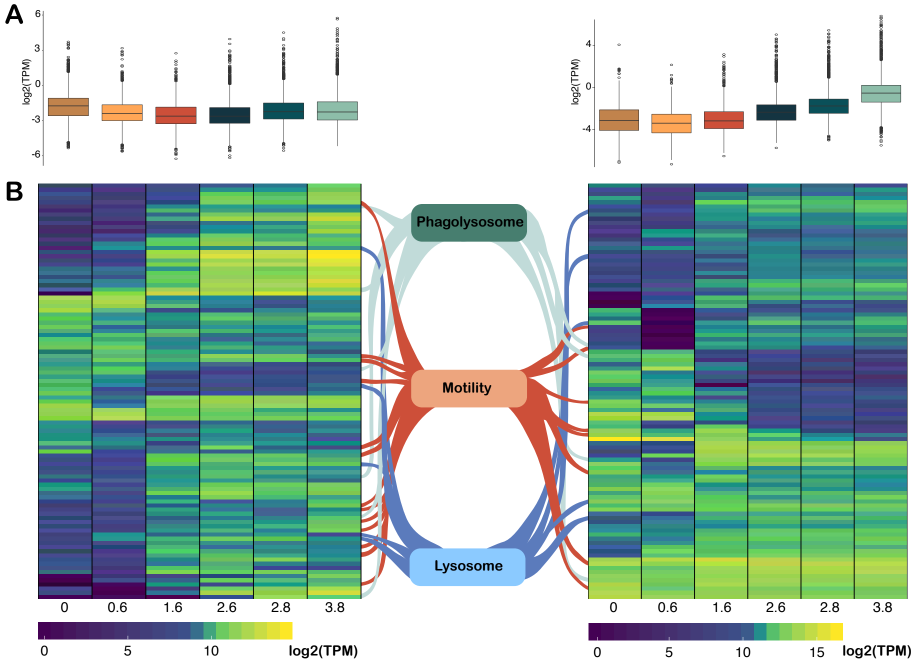


Figure 4. Gene expression of MAST-4A (left) and MAST-4B (right) during the incubation. (**a**) Box plots showing the expression values of all transcripts as TPM data. (**b**) Expression dynamics of the 100 most expressed transcripts of MAST-4A and MAST-4B. Within each species, the TPMs for every genes are added up and this sum is scaled to 1 million, in order to get its expression profile as if it was the only member of the community. Genes involved in phagocytosis are marked and classified into three broad categories. Heatmaps were done with the R package superheat and the hierarchical clustering method.

DISCUSSION

Using a mixed natural community sampled from surface water in Blanes Bay, we successfully promoted the growth of bacterivorous HFs in order to study their gene expression profiles. Based on both cell counts and 18S rDNA analyses, we observed a pronounced change in community composition during the incubation, most notably an increase in the abundance of HFs and a decrease in photosynthetic taxa. The switch from autotrophy to heterotrophy was expected as the experiment was conducted in the dark. This suppressed the growth of photosynthetic taxa, which in turn could be grazed by the heterotrophs. The development of the HFs was further supported by the growth of bacteria that proliferated early in the incubation as well as the absence of higher trophic grazers, following similar dynamics to those previously observed in other incubations with different initial communities (Massana et al. 2006; Weber et al. 2012). Larger taxa (i.e., dinoflagellates and ciliates) were nevertheless detected at the beginning in our incubation, and this signal most likely derived from broken cells, explaining the modest 18S rDNA signal detected and their decrease along time (Fig. 2). By integrating taxonomic classification of transcripts with microscopical cell counts, we detected a large diversity of efficient grazers in a predator dynamics scenario.

While 18S rDNA sequences can be classified phylogenetically and are useful for ecological studies on species distribution, many lineages such as the marine stramenopiles are understudied because they remain uncultured. In particular, they generally lack genomic information because most genomic research is biased toward a few cultured model species (Pawlowski et al. 2012; del Campo et al. 2014). This gap can now be filled by SCG, which allows retrieving the genomes of individually sorted cells without the need for culturing (Mangot et al. 2017). When combined with metatranscriptomics, which is a reliable approach for investigating metabolically active cells, SCG allowed the distinction between unicellular individual lineages present in the incubation that were so far barely recognized as most genomics focuses on a few model eukaryotes (del Campo et al. 2014). While SCG is essential to assign gene functions to uncultured species, it is also known to have some limitations, such as lack of coverage in some genomic regions (Rinke et al. 2014), and the presence of contaminant sequences that can compromise the quality of the final assembly. Thus, the absence of a particular gene in the metatranscriptomics analysis could be because it was not amplified by the multiple displacement amplification, was not assembled, or was not annotated in the final SAG used as reference. At any rate, in this study, SCG has allowed unprecedented insights into MASTs gene expression profiles, which could have not been revealed with the metatranscriptomics on its own.

Phagocytosis is a specific form of endocytosis (uptake of extracellular material) that involves engulfing large particles (Niedergang and Grinstein 2018) and that originated billions of years ago (Yutin et al. 2009). It is a complex process found in diverse eukaryotes, and that involves a variety of functional genes that are often not unique to the phagocytosis pathway but are shared with other processes (e.g., actin filament, ABC transporters). This mode of feeding has been studied in depth in macrophages in the mammalian immune system, and also in some unicellular microbial eukaryotes (Dayel and King 2014). Gotthardt et al. (2002), performed a targeted study of the proteins involved in phagocytosis (and their corresponding genes) using direct proteomic analyses of extracted phagosomes. Recently, a complementary effort using comparative genomics has attempted to identify the set of genes that are unique and representative of different trophic modes, including phagotrophy (Burns et al. 2015, 2018). These insights showed a complex process controlled by regulatory mechanisms involving numerous genes, revealing the potential for molecular detection of specific markers of phagocytosis. Linking gene expression and ecosystem function is feasible in marine bacteria, where marker genes for given biogeochemical functions in the oceans have been identified (Ferrera et al. 2015), allowing targeted studies of, for example, ammonia oxidation or phosphorous uptake (Imhoff 2016). Similar efforts toward the identification of genes indicative of a trophic strategy in microbial eukaryotes have recently been published (Alexander et al. 2015; Liu et al. 2016), including the search of specific traits driving heterotrophy (Beisser et al. 2017; Hu et al. 2018). Phagocytosis offers a unique case where genes can be related to microbial food webs. Therefore, to analyze genes that participate in phagocytosis in uncultured HFs, we used a metatranscriptomics approach on a community enrichment combined with a selection of specific transcripts provided by single cell genomes. Our method allowed to identify genes that control multiple aspects of the phagocytosis. We have detected many known characterized genes, and even though we sampled at different times during the enrichment, we did not see noticeable transcriptional changes. Therefore, in our targeted species, the phagocytosis machinery showed similar functional signal along the sampled time.

For two uncultured HF species, MAST-4A and MAST-4B, we detected genes involved in several of the main steps of the phagocytosis pathway as have been described mostly for mammalian macrophages. Phagocytosis is an actin-dependent process that is required to initiate prey capture. We identified genes that control the nucleation of new actin filaments, namely the assembly factors of the Arp2/3 complex (May and Machesky 2001; Lai et al. 2008). Actin polymerization is also promoted by the small GTPases of the Rho family, of which we detected Cdc42 here; this is generally followed by Rac1 and Rac2 activation, but these were not observed. These GTPases interact with proteins of the WASP and Scar/WAVE family that stimulate the Arp2/3 complex (Castellano et al. 2001), and both were actively expressed in the community assemblage. Involved in phagosome maturation, the Rab-family GTPases (Vieira et al. 2003; Fairn and Grinstein 2012) participate in the formation of the phagolysosome, and here we failed to detect important genes involved in this step, such as Rab5 and Rab7. These gene absences could be due to genome incompleteness of the SAGs. In addition to GTPases, we found a putative phosphatidylinositol 3-kinase (PI3K), which is necessary for successful phagolysosome formation (Vieira et al. 2003). The food vacuole seems to be acidified by highly expressed proton and cation pumps that are organized by vacuolar ATPases (Kissing et al. 2015). In addition, we identified high expression of a rhodopsin gene in MAST-4A, and we hypothesize that the corresponding protein could act as a light-driven proton pump contributing to phagosome acidification (Slamovits et al. 2011; Kandori 2015). Microbial rhodopsins were initially found in

Archaebacteria and are now known to be widely dispersed light-driven ion transporters across all domains of life (Beja et al. 2000; Finkel et al. 2013). Finally, the strongly acidified phagolysosome becomes rich in hydrolytic enzymes that promote the degradation of the ingested microbial prey, and here we found several highly expressed digestive enzymes such as cathepsins and glycoside hydrolases. Our results indicate a set of genes related to phagocytosis that were highly expressed in environmentally relevant bacterivorous uncultured HFs. The genes defined here are often not exclusive to phagocytosis, but represent a continuum of proteins involved in different types of fusion, vesicle transport, and digestive processes. In addition to markers of phagosome acidification (V-ATPase and rhodopsins), the digestive enzymes cathepsins would be ideal candidates for detecting phagocytosis in assemblages of marine HFs. Future work will need to assess their suitability to target this trophic mode in natural communities.

Historically, the study of microorganisms, including ecogenomics and gene expression profiling, has focused on single species in pure culture. Metatranscriptomics allows us to circumvent the culture-dependent analysis of many microbial species and to perform functional studies in complex communities. In our analysis, SCG has facilitated the targeting of the transcriptional profile of uncultured HFs. This approach is a new opportunity to examine the heterogeneity of microbial communities, recover their true diversity, and better understand specific biological processes performed by particular species.

ACKNOWLEDGMENTS

This project was supported by the EU project SINGEK (H2020-MSCA-ITN-2015-675752) and the Spanish project ALLFLAGS (CTM2016-75083-R, MINECO). S.W. was supported by a Postdoctoral Fellowship from the European Union's Seventh Framework Program under Marie Sklodowska-Currie grant agreement 6261882. We thank Ramiro Logares and Fran Latorre for useful discussions and sharing data.

SUPPLEMENTARY MATERIAL

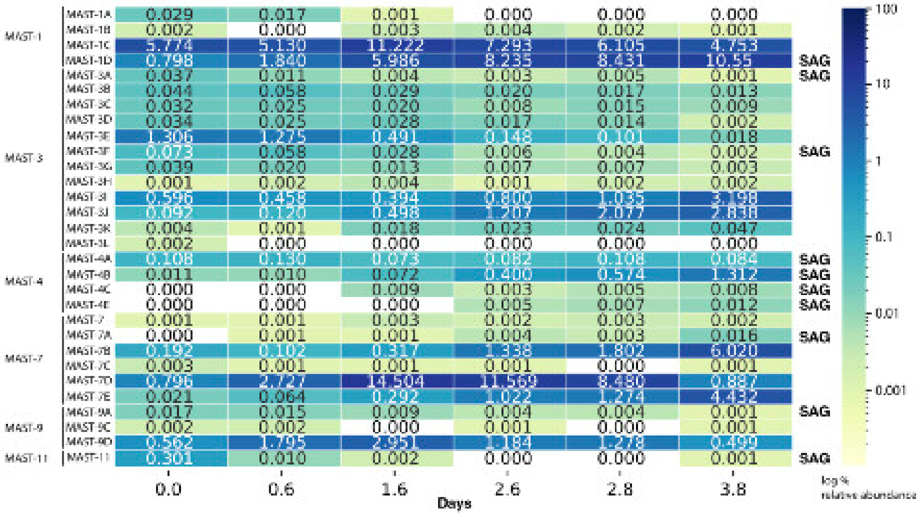
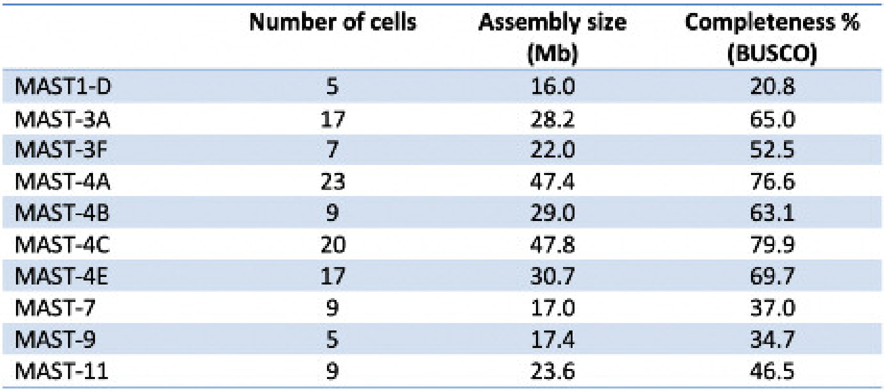
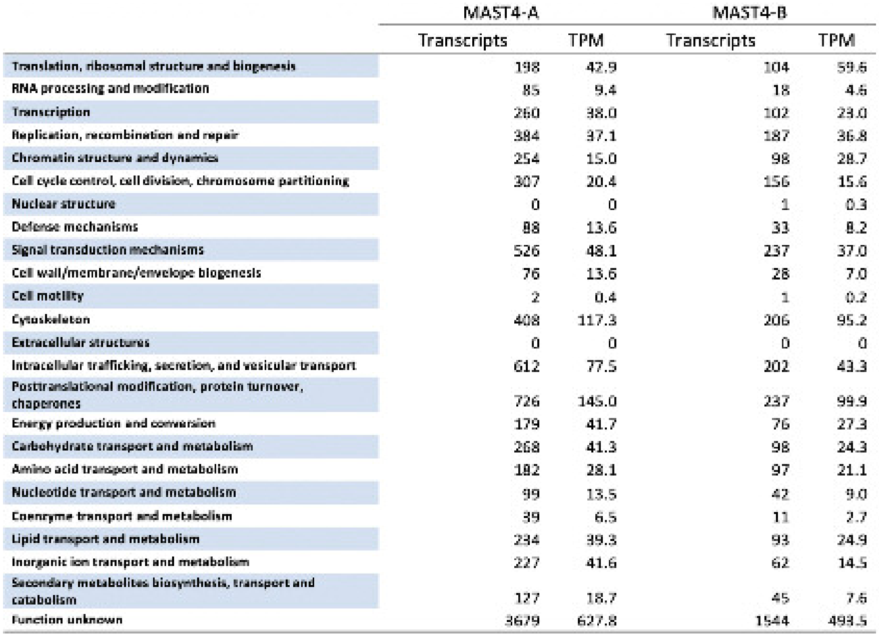


Figure S1. Relative abundance of MAST groups and subgroups in the incubation by using V4 18S rDNA metatranscriptomics reads. The last column indicates whether or not the subgroup contains a SAG representative. The color gradient helps to visualize the progression in time of the relative abundance of different taxa.



	Number of cells	Assembly size (Mb)	Completeness % (BUSCO)
MAST1-D	5	16.0	20.8
MAST-3A	17	28.2	65.0
MAST-3F	7	22.0	52.5
MAST-4A	23	47.4	76.6
MAST-4B	9	29.0	63.1
MAST-4C	20	47.8	79.9
MAST-4E	17	30.7	69.7
MAST-7	9	17.0	37.0
MAST-9	5	17.4	34.7
MAST-11	9	23.6	46.5

Table S1. Single cell genome co-assemblies of uncultured marine stramenopiles. The table shows the number of cells sequenced, the final co-assembly size, and the genome completeness score from BUSCO using the Eukaryote/Protist database. Note that, as the number of co-assembled cells increases, we obtain better completion from BUSCO. This is expected as the MDA protocol for single-celled genomics randomly amplifies portions of the genome, and as we add more cells we are increasing the genome coverage.



	MAST4-A		MAST4-B	
	Transcripts	TPM	Transcripts	TPM
Translation, ribosomal structure and biogenesis	198	42.9	104	59.6
RNA processing and modification	85	9.4	18	4.6
Transcription	260	38.0	102	23.0
Replication, recombination and repair	384	37.1	187	36.8
Chromatin structure and dynamics	254	15.0	98	28.7
Cell cycle control, cell division, chromosome partitioning	307	20.4	156	15.6
Nuclear structure	0	0	1	0.3
Defense mechanisms	88	13.6	33	8.2
Signal transduction mechanisms	526	48.1	237	37.0
Cell wall/membrane/envelope biogenesis	76	13.6	28	7.0
Cell motility	2	0.4	1	0.2
Cytoskeleton	408	117.3	206	95.2
Extracellular structures	0	0	0	0
Intracellular trafficking, secretion, and vesicular transport	612	77.5	202	43.3
Posttranslational modification, protein turnover, chaperones	726	145.0	237	99.9
Energy production and conversion	179	41.7	76	27.3
Carbohydrate transport and metabolism	268	41.3	98	24.3
Amino acid transport and metabolism	182	28.1	97	21.1
Nucleotide transport and metabolism	99	13.5	42	9.0
Coenzyme transport and metabolism	39	6.5	11	2.7
Lipid transport and metabolism	234	39.3	93	24.9
Inorganic ion transport and metabolism	227	41.6	62	14.5
Secondary metabolites biosynthesis, transport and catabolism	127	18.7	45	7.6
Function unknown	3679	627.8	1544	493.5

Table S2. Functional annotation of all identified MAST-4A and MAST-4B transcripts. Transcripts affiliate to the 24 COG categories using the EggNOG database. The TPM values provided for each species give the expression assessment cumulated considering the transcripts identified. The function was selected based on the best e-value (with a minimum threshold of 0.001). When the transcript was attributed to different functions with the same best e-value, we considered this transcript in every category and therefore was counted more than once.

REFERENCES

Adl, S. M., and others. 2019. Revisions to the classification, nomenclature, and diversity of eukaryotes. J. Eukaryot. Microbiol. 66: 4–119. doi:10.1111/jeu.12691

Alexander, H., M. Rouco, S. T. Haley, S. T. Wilson, D. M. Karl, and S. T. Dyhrman. 2015. Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. Proc. Natl. Acad. Sci. USA 112: E5972–E5979. doi:10.1073/pnas.1518165112

Alonso-Sáez, L., X. A. G. Morán, and M. R. Clokie. 2018. Low activity of lytic pelagiphages in coastal marine waters. ISME J. 12: 2100–2102. doi:10.1038/s41396-018-0185-y

Amann, R. I., W. Ludwig, and K.-H. Schleifer. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59: 143–169.

Arndt, H., D. Dietrich, B. Auer, E.-J. Cleven, T. Gräfenhan, M. Weitere, and A. P. Mylnikov. 2000. Functional diversity of heterotrophic flagellates in aquatic ecosystems, p. 240–268. In The flagellates.

Beisser, D., and others. 2017. Comprehensive transcriptome analysis provides new insights into nutritional strategies and phylogenetic relationships of chrysophytes. PeerJ 5: e2832. doi:10.7717/peerj.2832

Beja, O., and others. 2000. Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. Science 289: 1902–1906. doi:10.1126/science.289.5486.1902

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

Boulais, J., and others. 2010. Molecular characterization of the evolution of phagosomes. Mol. Syst. Biol. 6: 423. doi:10.1038/msb.2010.80

Burns, J. A., A. Paasch, A. Narechania, and E. Kim. 2015. Comparative genomics of a bacterivorous green alga reveals evolutionary causalities and consequences of phago-mixotrophic mode of nutrition. Genome Biol. Evol. 7: 3047–3061. doi:10.1093/gbe/evv144

Burns, J. A., A. A. Pittis, and E. Kim. 2018. Gene-based predictive models of trophic modes suggest Asgard archaea are not phagocytotic. Nat. Ecol. Evol. 2: 697–704. doi:10.1038/s41559-018-0477-7

Cabello, A. M., M. Latasa, I. Forn, X. A. G. Morán, and R. Massana. 2016. Vertical distribution of major photosynthetic picoeukaryotic groups in stratified marine waters: Photo-picoeukaryotes community structure in the DCM. Environ. Microbiol. 18: 1578–1590. doi:10.1111/1462-2920.13285

Castellano, F., C. Le Clainche, D. Patin, M.-F. Carlier, and P. Chavrier. 2001. A WASp-VASP complex regulates actin polymerization at the plasma membrane. EMBO J. 20: 5603–5614. doi:10.1093/emboj/20.20.5603

Cavalier-Smith, T. 2006. Cell evolution and Earth history: Stasis and revolution. Philos. Trans. R. Soc. Lond. B Biol. Sci. 361: 969–910. doi:10.1098/rstb.2006.1842

Dayel, M. J., and N. King. 2014. Prey capture and phagocytosis in the choanoflagellate Salpingoeca rosetta. PLoS One 9: e95577. doi:10.1371/journal.pone.0095577

de Vargas, C., and others. 2015. Eukaryotic plankton diversity in the sunlit ocean. Science 348: 1261605. doi:10.1126/science.1261605 del Campo, J., M. E. Sieracki, R. Molestina, P. Keeling, R. Massana, and I. Ruiz-Trillo. 2014. The others: Our biased perspective of eukaryotic genomes. Trends Ecol. Evol. 29: 252–259. doi:10.1016/j.tree.2014.03.006

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

Fairn, G. D., and S. Grinstein. 2012. How nascent phagosomes mature to become phagolysosomes. Trends Immunol. 33: 397–405. doi:10.1016/j.it.2012.03.003

Ferrera, I., M. Sebastian, S. G. Acinas, and J. M. Gasol. 2015. Prokaryotic functional gene diversity in the sunlit ocean: Stumbling in the dark. Curr. Opin. Microbiol. 25: 33–39. doi:10.1016/j.mib.2015.03.007

Finkel, O. M., O. Béjà, and S. Belkin. 2013. Global abundance of microbial rhodopsins. ISME J. 7: 448–451. doi:10.1038/ismej.2012.112

Giner, C. R., I. Forn, S. Romac, R. Logares, C. de Vargas, and R. Massana. 2016. Environmental sequencing provides reasonable estimates of the relative abundance of specific picoeukaryotes. Appl. Environ. Microbiol. 82: 4757–4766. doi:10.1128/AEM.00560-16

Giner, C. R., and others. 2019. Quantifying long-term recurrence in planktonic microbial eukaryotes. Mol. Ecol. 28: 923–935. doi:10.1111/mec.14929

Gotthardt, D., H. J. Warnatz, O. Henschel, F. Brückert, M. Schleicher, and T. Soldati. 2002. High-resolution dissection of phagosome maturation reveals distinct membrane trafficking phases. Mol. Biol. Cell 13: 3508–3520. doi:10.1091/mbc.e02-04-0206

Grabherr, M. G., and others. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29: 644–652. doi:10.1038/nbt.1883

Guillou, L., and others. 2013. The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res. 41: D597–D604. doi:10.1093/nar/gks1160

Hu, S. K., Z. Liu, H. Alexander, V. Campbell, P. E. Connell, S. T. Dyhrman, K. B. Heidelberg, and D. A. Caron. 2018. Shifting metabolic priorities among key protistan taxa within and below the euphotic zone: Depth-related protistan metatranscriptomes. Environ. Microbiol. 20: 2865–2879. doi:10.1111/1462-2920.14259

Huerta-Cepas, J., and others. 2016. eggNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res. 44: D286–D293. doi:10.1093/nar/gkv1248

Imhoff, J. 2016. New dimensions in microbial ecology-functional genes in studies to unravel the biodiversity and role of functional microbial groups in the environment. Microorganisms 4: 19. doi:10.3390/microorganisms4020019

Jones, P., and others. 2014. InterProScan 5: Genome-scale protein function classification. Bioinformatics 30: 1236–1240. doi:10.1093/bioinformatics/btu031

Jürgens, K., and R. Massana. 2008. Protistan grazing on marine bacterioplankton, p. 383–441. In D. L. Kirchman [ed.], Microbial ecology of the oceans. John Wiley & Sons.

Kandori, H. 2015. Ion-pumping microbial rhodopsins. Front. Mol. Biosci. 2: 52. doi:10.3389/fmolb.2015.00052

Kissing, S., and others. 2015. Vacuolar ATPase in phagosome-lysosome fusion. J. Biol. Chem. 290: 14166–14180. doi:10.1074/jbc.M114.628891

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

Lai, F. P., and others. 2008. Arp2/3 complex interactions and actin network turnover in lamellipodia. EMBO J. 27: 982–992. doi:10.1038/emboj.2008.34

Langmead, B., and S. L. Salzberg. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9: 357–359. doi:10.1038/nmeth.1923

Levin, R., S. Grinstein, and J. Canton. 2016. The life cycle of phagosomes: Formation, maturation, and resolution. Immunol. Rev. 273: 156–179. doi:https://doi.org/10.1111/imr.12439

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. doi:10.1093/bioinformatics/btp324

Lin, Y.-C., T. Campbell, C.-C. Chung, G.-C. Gong, K.-P. Chiang, and A. Z. Worden. 2012. Distribution patterns and phylogeny of marine stramenopiles in the North Pacific Ocean. Appl. Environ. Microbiol. 78: 3387–3399. doi:10.1128/AEM.06952-11

Liu, Z., V. Campbell, K. B. Heidelberg, and D. A. Caron. 2016. Gene expression characterizes different nutritional strategies among three mixotrophic protists. FEMS Microbiol. Ecol. 92: fiw106. doi:10.1093/femsec/fiw106

Mangot, J.-F., and others. 2017. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. Sci. Rep. 7: 41498. doi:10.1038/srep41498

Mangot, J.-F., I. Forn, A. Obiol, and R. Massana. 2018. Constant abundances of ubiquitous uncultured protists in the open sea assessed by automated microscopy: Automatic microscopic counts of small protists. Environ. Microbiol. 20: 3876–3889. doi:10.1111/1462-2920.14408

Massana, R., L. Guillou, R. Terrado, I. Forn, and C. Pedrós-Alió. 2006. Growth of uncultured heterotrophic flagellates in unamended seawater incubations. Aquat. Microb. Ecol. 45: 171–180. doi:10.3354/ame045171

Massana, R., F. Unrein, R. Rodríguez-Martínez, I. Forn, T. Lefort, J. Pinhassi, and F. Not. 2009. Grazing rates and functional diversity of uncultured heterotrophic flagellates. ISME J. 3: 588–596. doi:10.1038/ismej.2008.130

Massana, R., J. del Campo, M. E. Sieracki, S. Audic, and R. Logares. 2014. Exploring the uncultured microeukaryote majority in the oceans: Reevaluation of ribogroups within stramenopiles. ISME J. 8: 854–866. doi:10.1038/ismej.2013.204

Massana, R., and others. 2015. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing: Protist diversity in European coastal areas. Environ. Microbiol. 17: 4035–4049. doi:10.1111/1462-2920.12955

May, R. C., and L. M. Machesky. 2001. Phagocytosis and the actin cytoskeleton. J. Cell Sci. 114: 1061–1077.

Niedergang, F., and S. Grinstein. 2018. How to build a phagosome: New concepts for an old process. Curr. Opin. Cell Biol. 50: 57–63. doi:10.1016/j.ceb.2018.01.009

Okada, M., and others. 2005. Proteomic analysis of phagocytosis in the enteric protozoan parasite Entamoeba histolytica. Eukaryot. Cell 4: 827–831. doi:10.1128/EC.4.4.827-831.2005

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

Pawlowski, J., and others. 2012. CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. PLoS Biol. 10: e1001419. doi:10.1371/journal.pbio.1001419

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner. 2013. The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res. 41: D590–D596. doi:10.1093/nar/gks1219

Rink, J., E. Ghigo, Y. Kalaidzidis, and M. Zerial. 2005. Rab conversion as a mechanism of progression from early to late endosomes. Cell 122: 735–749. doi:10.1016/j.cell.2005.06.043

Rinke, C., and others. 2014. Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. Nat. Protoc. 9: 1038–1048. doi:10.1038/nprot.2014.067

Slamovits, C. H., N. Okamoto, L. Burri, E. R. James, and P. J. Keeling. 2011. A bacterial proteorhodopsin proton pump in marine eukaryotes. Nat. Commun. 2: 183. doi:10.1038/ncomms1188

Stanke, M., R. Steinkamp, S. Waack, and B. Morgenstern. 2004. AUGUSTUS: A web server for gene finding in eukaryotes. Nucleic Acids Res. 32: W309–W312. doi:10.1093/nar/gkh379

Uribe-Querol, E., and C. Rosales. 2017. Control of phagocytosis by microbial pathogens. Front. Immunol. 8: 1368. doi:10.3389/fimmu.2017.01368

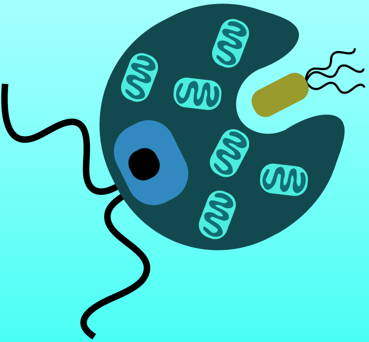
Vieira, O. V., R. J. Botelho, and S. Grinstein. 2002. Phagosome maturation: Aging gracefully. Biochem. J. 366: 689–704. doi:10.1042/bj20020691

Vieira, O. V., and others. 2003. Modulation of Rab5 and Rab7 recruitment to phagosomes by phosphatidylinositol 3-kinase. Mol. Cell. Biol. 23: 2501–2514. doi:10.1128/MCB.23.7.2501-2514.2003

Weber, F., J. del Campo, C. Wylezich, R. Massana, and K. Jürgens. 2012. Unveiling trophic functions of uncultured protist taxa by incubation experiments in the brackish Baltic Sea. PLoS One 7: e41970. doi:10.1371/journal.pone.0041970

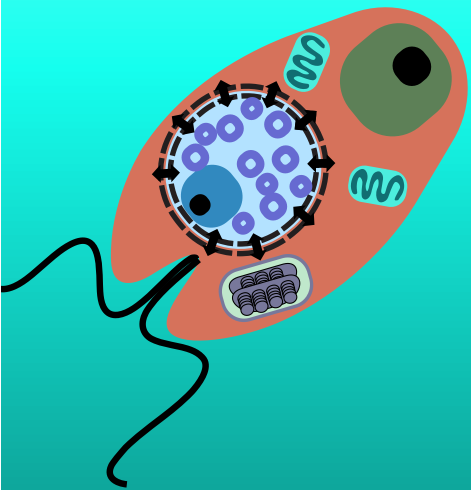
Worden, A. Z., M. J. Follows, S. J. Giovannoni, S. Wilken, A. E. Zimmerman, and P. J. Keeling. 2015. Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. Science 347: 1257594–1257594. doi:10.1126/science.1257594

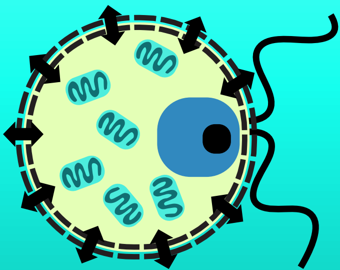
Yutin, N., M. Y. Wolf, Y. I. Wolf, and E. V. Koonin. 2009. The origins of phagocytosis and eukaryogenesis. Biol. Direct 4: 9. doi:10.1186/1745-6150-4-9

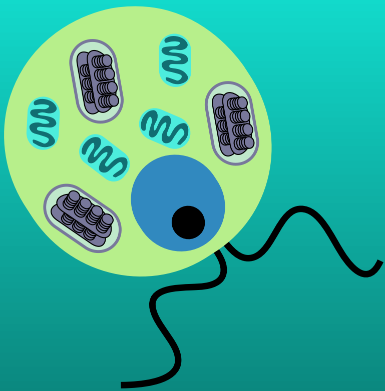


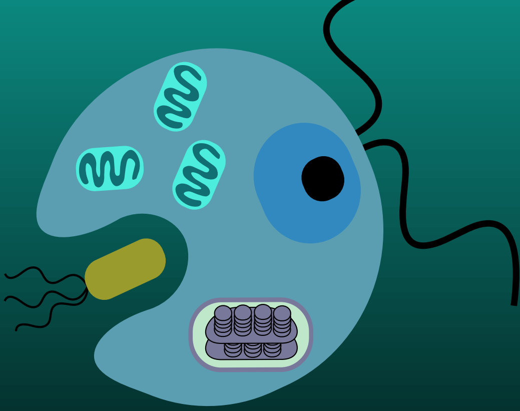
Chapter 4

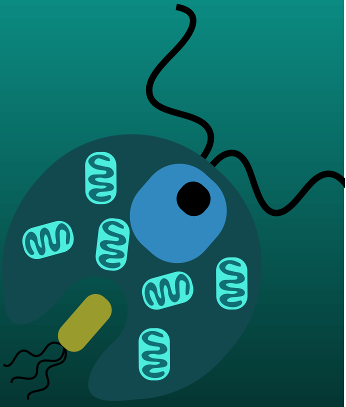
Gene expression during bacterivorous growth of a widespread marine heterotrophic flagellate











Published in in The ISME Journal: Multidisciplinary Journal of Microbial Ecology

Ramon Massana, **Aurelie Labarre**, David López-Escardó, Aleix Obiol, François Bucchini,Thomas Hack, Matthias G. Fischer, Klaas Vandepoele, Denis V. Tikhonenkov, Filip Husnik & Patrick J. Keeling

ABSTRACT

Phagocytosis is a fundamental process in marine ecosystems by which prey organisms are consumed and their biomass incorporated in food webs or remineralized. However, studies searching for the genes underlying this key ecological process in free-living phagocytizing protists are still scarce, in part due to the lack of appropriate ecological models. Our reanalysis of recent molecular datasets revealed that the cultured heterotrophic flagellate *Cafeteria burkhardae* is widespread in the global oceans, which prompted us to design a transcriptomics study with this species, grown with the cultured flavobacterium *Dokdonia* sp. We compared the gene expression between Exponential and Stationary phases, which were complemented with three starvation by dilution phases that appeared as intermediate states. We found distinct expression profiles in each condition and identified 2056 differentially expressed genes between Exponential and Stationary samples. Upregulated genes at the Exponential phase were related to DNA duplication, transcription and translational machinery, protein remodeling, respiration and phagocytosis, whereas upregulated genes in the Stationary phase were involved in signal transduction, cell adhesion and lipid metabolism. We identified a few highly expressed phagocytosis genes, like peptidases and proton pumps, which could be used to target this ecologically relevant process in marine ecosystems.

.

INTRODUCTION

Eukaryotic microbes (protists) include a diverse collection of unicellular organisms that are involved in crucial food web processes such as primary production, predation, and parasitism [1, 2]. A particular functional group, referred as heterotrophic flagellates, are known to be primary agents of bacterivory. As such, they keep bacterial abundances in check, direct bacterial production to higher trophic levels, and release inorganic nutrients that sustain regenerated primary production [3, 4]. For years, the abundance, distribution, and activity of heterotrophic flagellates was studied as a group property and their diversity addressed by morphological and culturing approaches [5, 6]. The advent of molecular tools revealed many uncultured and undescribed species [7, 8], highlighted a prevalent culturing bias, and suggested many of the isolated species were rare in nature and perhaps poor models for more dominant ones [9]. Little work has been done linking physiological studies of cultured heterotrophic flagellates with the genes responsible for ecologically relevant processes, despite the great promise of transcriptomics to provide new insights into the ecology of eukaryotic species [10].

Heterotrophic flagellates feed on bacteria through phagocytosis, the engulfment and digestion of a prey cell in a food vacuole. Phagocytosis is an ancient trait that marked the origin of eukaryotic cells [11] and allowed critical evolutionary innovations [12, 13]. It is a complex process involving hundreds of proteins operating in consecutive steps: sensing and motility, prey recognition, cytoskeleton remodeling for food vacuole formation, vacuole maturation, and acidic enzymatic digestion. Given its importance in immunity [14], phagocytosis has been mostly investigated at the cellular and molecular level in metazoan immune cells [15, 16], where identified genes have been placed in functional maps [17]. The few studies done with free-living protists, like ciliates and amoebozoans [18, 19], indicate that the basic machinery for phagocytosis and many of the genes involved are evolutionarily conserved [20]. However, these studies do not provide a detailed model of how gene expression changes during phagocytic growth, and this could be readily studied by differential expression analyses of cells actively preying versus starved ones. This experiment has rarely been performed [21, 22], due to the lack of cultured ecological models.

We studied the bicosoecid *Cafeteria burkhardae*, an efficient suspension feeder that preys on bacteria by creating a current with its anterior flagellum. Although the used strain E4-10 was named *Cafeteria roenbergensis*, a recent paper that sequenced the 18S rDNA of the type species *C. roenbergensis* [23] showed that both strains had different 18S rDNA, which led to the description of *C. burkhardae* [24]. *Cafeteria burkhardae* strain E4-10 was used in the MMETSP transcriptome initiative [25] and its high-quality draft genome has been recently released [26]. Moreover, the strains easily cultured from seawater [5] and often used in growth and grazing experiments [27, 28] also correspond to *C. burkhardae* [24]. Previous studies suggested this species was a minor member of marine heterotrophic flagellates [29], but we describe here more extensive molecular surveys that reveal a widespread distribution. We grew *C. burkhardae* in batch cultures with a known bacterium and collected transcriptomic samples at the Exponential and Stationary phases, together with additional states where the cells were starved by dilution. Differential expression analysis identified genes correlated with Exponential growth, when cells were feeding, converting bacterial food to biomass and dividing. Some of these genes, particularly those that were highly expressed, are promising targets for future exploration of phagocytosis in marine ecosystems.

MATERIAL AND METHODS

C. burkhardae in the Malaspina dataset

Marine microbes (0.2-3 µm size fraction) were collected during the Malaspina expedition in 120 stations at surface and in 13 profiles of 7 depths from surface to the bathypelagic zone. Eukaryotic diversity was assessed by sequencing the V4 18S rDNA region. Details of sample collection, nucleic acid extraction, V4 amplification, and Illumina sequencing are presented elsewhere for surface data [30] and vertical profiles [31]. Here, we processed the reads using DADA2 [32] with parameters *truncLen* 240,210 and *maxEE* 6,8 and identified the ASV (Amplicon Sequence Variant) corresponding to *C. burkhardae*. Its relative abundance was calculated against the number of reads per sample after removal of metazoan and plant reads. Metagenomes of the same size fraction in vertical profiles were generated from the same cruise [33] and used in BLAST [34] fragment recruitment analysis against the *C. burkhardae* genome [24]. Direct cell counts were performed in 13 surface samples by FISH as explained before [29, 35].

Growth of Cafeteria burkhardae on Dokdonia sp.

The flavobacterium *Dokdonia* sp. MED134 was isolated on Zobell agar plates from the Blanes Bay Microbial Observatory [36]. To prepare cell concentrates, a colony was inoculated in 50 mL of Zobell medium and incubated at 22°C for 3 days. Cells were collected by centrifugation (4500 rpm for 15 min), resuspended in sterile seawater (filtered by 0.2 µm and autoclaved), centrifuged again, resuspended in 100 mL of sterile seawater, and kept at 4°C for one week. To calculate the cell abundance of the concentrate, one aliquot was fixed with ice-cold glutaraldehyde (1% final concentration), stained with DAPI, and filtered on a 0.2 µm pore-size polycarbonate filter. Filters were mounted on a slide and counts were performed by epifluorescence microscopy by exciting with UV radiation [37].

Cafeteria burkhardae strain E4-10 was isolated in 1989 [38] and maintained on a rice grain with artificial seawater. The culture was acclimated to grow on *Dokdonia* MED134 as prey in two steps. First 0.1 mL of the culture was inoculated in a flask with 20 mL of sterile seawater and 10^8 bacteria mL^{-1} for 5 days. Second, 1 mL of this culture was inoculated to 400 mL of sterile seawater and 2.4 x 10^7 bacteria mL^{-1} for one week. Flagellate growth was inspected by light microscopy through the culture flasks. Incubations were done at 22°C on the lab bench.

Batch cultures, dilution event, and RNA extraction and sequencing

Three batch cultures were prepared with 400 mL of sterile seawater, *Dokdonia* MED134 at 2.5 x 10^7 cells mL^{-1}, and 1 mL of *C. burkhardae* from the last acclimation bottle. Three mL aliquots were fixed with glutaraldehyde to count, just after sampling, the abundance of flagellates and bacteria by epifluorescence microscopy. Flagellate growth rates were calculated as the slope of the linear part of logarithmic cell numbers versus time. Grazing rates were calculated using growth rates, the slope of the logarithmic decrease of bacteria, and the geometric mean of flagellates and bacteria abundances using the formulas of Frost [39] and Heinbokel [40]. Growth efficiency was calculated from growth and grazing rates and the estimated carbon per cell of both species obtained from cell sizes measured at the microscope [41].

Samples for transcriptomics were taken in triplicates from the last acclimation bottle (Inoculum), and in duplicates in the three bottles at the Exponential (day 2.3) and Stationary (day 3.7) phases. Cells were collected in microfiltration units of 0.8 µm pore size (Vivaclear MINI 0.8µm PES, Sartorius, Göttingen, Germany). For each sample, four units were filled with 0.5 mL of culture, spun down for 30 sec at 1000 rpm, and the step repeated until processing 10 mL. Next, 100 µL of lysis buffer from the RNAqueous-Micro kit (Thermo Fisher Scientific, Waltham, Massachusetts, US) were added to each unit, vortexed, left for 1 min, and the lysate was spun down at 13,000 rpm for 30 sec. The four cell lysates from the same sample were combined and the RNA was extracted following the kit's protocol. Genomic DNA was removed with DNase I. RNA quantity and purity was assessed with a NanoDrop 1000 Spectrophotometer (Thermo Fisher Scientific) and the RNA extracts were kept at -80°C.

During the exponential phase, three dilutions (10 mL of culture in 190 mL sterile seawater) were prepared from each batch culture, and they were processed after 0.4, 1.4 and 3.3 days for cell counts (5 mL) and RNA extraction (195 mL). As these large volumes prevented the use of microfiltration units, cell collection was done on 47 mm polycarbonate filters of 0.8 µm pore size. Filters were cut in 4 pieces, submerged in 1 mL of lysis buffer, vortexed, and left for 30 sec. The lysate was recovered and the RNA was extracted as before.

Polyadenylated RNA transcripts were converted into cDNA following the Smart-seq2 protocol [42] designed for very low RNA amounts. In brief, Oligo-$dT_{30}VN$ primers annealed to all mRNAs containing a poly(A) tail, then reverse transcription and template-switching was done, followed by 9-cycles of PCR amplification using IS PCR oligos linked at the two ends of the cDNA molecules [42]. Amplified cDNA was purified and quantified with a Qubit fluorometer (Thermo Fisher Scientific). The complete set of 24 cDNA samples (15 µl at 2-4 ng l^{-1}) was sent to the Sequencing + Bioinformatics Consortium at UBC and, based on the BioAnalyzer results (Agilent, Santa Clara, California, US), 21 samples were chosen for sequencing (Table S1). Illumina Nextera XT libraries with a dual index were prepared and pooled on a single lane of a NextSeq Illumina sequencer yielding, on average, 14.1 million 150 bp pair-ended reads per sample (Table S1). Raw reads have been deposited in ENA under the accession number PRJEB36247.

Transcriptome assembly, functional annotation, and differential expression analysis

Quality trimming of Illumina reads was done using Trimmomatic 0.33 [43] with parameters set to crop:149 slidingwindow:6:25 minlen:50. This removed about one third of the reads per sample (Table S1). High quality reads were mapped with Bowtie2 [44] towards the genome of *Dokdonia* MED134 (3.3 Mb; CP009301) and the *C. burkhardae* rDNA operon (5800 bp; extracted from a genome contig with the 18S rDNA [KY886365] and the 28S rDNA [FJ032656]). We used Bowtie2 in the sensitive mode, which restricts to zero the mismatches in seed alignment, and removed the mapped reads from the sequencing files. Reads mapping the bacterial genome were highest in Exponential, intermediate in Dilution, and lowest in Stationary stages (Fig. S1a), while reads mapping to eukaryotic rDNA operon were similar in all cases (Fig. S1b). Cleaned reads from all samples (4.9 million on average, Table S1) were co-assembled using Trinity-v2.4.0 [45]. The initial transcriptome consisted of 70 652 isoforms, for which the longest one of each gene was retained, resulting in 48 502 transcripts. These were compared using BLAST against the genome [26] and the transcriptome [25] of *C. burkhardae*, and annotated by Trinotate using UniProt [46], Pfam [47] and eggNOG [48] databases. We retained transcripts having a match to the genome or the transcriptome, or annotated as Eukaryota (19 215 left). Cleaned reads were mapped to this set with RSEM [49] and we kept 15 887 transcripts that appeared in at least 3 samples (0.3% of the signal removed). An additional BLASTn search removed obvious bacterial and viral genes (15 123 left). Transcripts with several ORFs identified by TransDecoder [45] were split when a different function was predicted for each ORF: 866 were split in two, 92 in three and 12 in four parts. The expression level of split regions was often very different (Fig. S2). Gene space completeness of the final curated transcriptome of 16 209 genes was estimated with BUSCO V3 [50].

The curated transcriptome was further processed using TRAPID [51] to annotate sequences with InterPro domains [52]. The processing strategy outlined in the original publication was slightly modified: sequence similarity search was performed using DIAMOND [53] in 'more-sensitive' mode (e-value cutoff of 10^{-5}) against a stramenopile-oriented PLAZA database [54] comprising genomic data of 35 organisms including *C. burkhardae* (Table S2). Functional annotation was transferred from the top protein hit and its assigned gene family.

Cleaned reads were mapped to the curated transcriptome using RSEM. The TPM (Transcripts Per Million) table was used for sample comparison by NMDS and for differential expression (DE) analyses with EdgeR [55]. The latter tool detects DE genes (logFC >2 and FDR corrected p-values $<10^{-3}$) in pair-wise sample comparisons. InterPro domain enrichment analysis of gene sets showing a specific expression profile (e.g. genes upregulated in the Exponential versus the Stationary phase) was performed with TRAPID using the hypergeometric distribution, with a maximum Benjamini–Hochberg corrected p-value cutoff of 0.05 and the entire curated transcriptome used as background. Enriched protein domains were manually assigned to given general processes and cellular functions.

RESULTS

Distribution of Cafeteria burkhardae in the global ocean

We took advantage of recently published protist diversity surveys to study the distribution of *C. burkhardae* in the global ocean (Fig. 1a). The ASV of this species was detected in most epipelagic samples (154 out of 172) with a wide variation in its relative abundance (Table 1), often below 0.1% and sometimes above 1% (median of 0.03%). The presence and relative abundance of this ASV was intermediate at the mesopelagic (found in 58 out of 61 samples; median of 0.09%) and maximal at the bathypelagic (in 58 of 60 samples; median of 0.49%). The patchy distribution of this ASV was evident in the three layers, as revealed by the huge differences between average and median values (Table 1). For instance, 22% of bathypelagic samples showed an abundance above 10%, while in 20% of samples it was below 0.1%. Performing FISH counts on 13 surface samples along the cruise track, we found cells in only 5 samples (Table 1), with abundances from 0.7 to 10.7 cells mL^{-1}.

We then used the *C. burkhardae* genome to perform a fragment recruitment analysis against 66 metagenomes of the same expedition. This PCR-free survey detected *C. burkhardae* in all samples and confirmed the increase in relative abundance along the water column (Table 1). In three bathypelagic samples, the *C. burkhardae* genome recruited ~0.6% of reads, suggesting a high dominance of this species in their microbial assemblage that also included prokaryotes. Metagenomic reads mapped along the complete genome and were mostly placed at the 99-100% similarity interval (Fig. 1b). This occurred in the three water layers (Fig. S3), albeit at surface some genomic regions recovered reads at lower similarity, probably from highly conserved genes of other species. This metagenomic analysis indicates that the cultured strain is widespread in the global ocean.

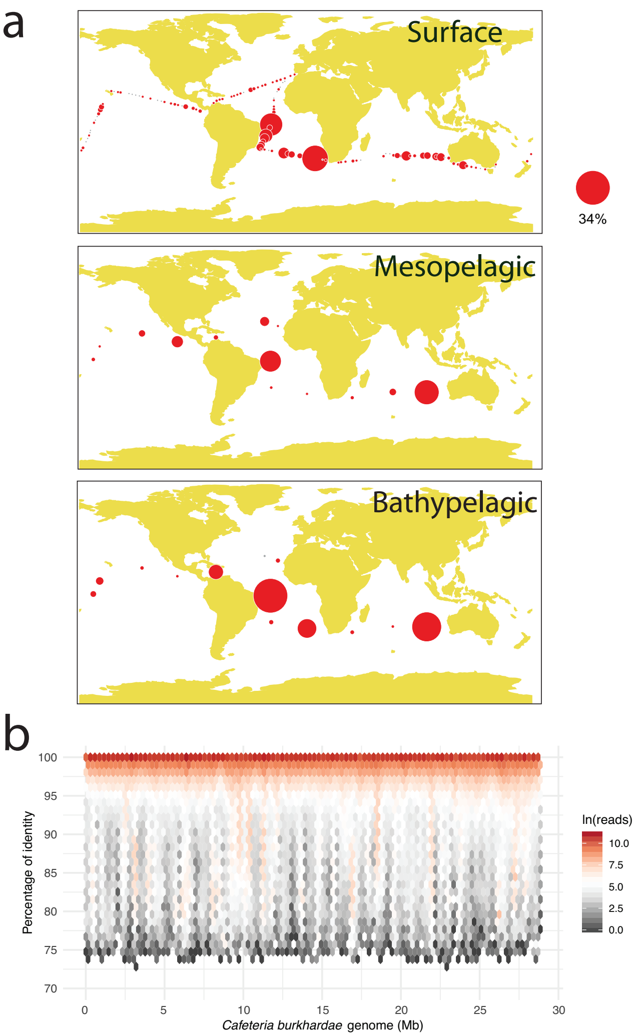
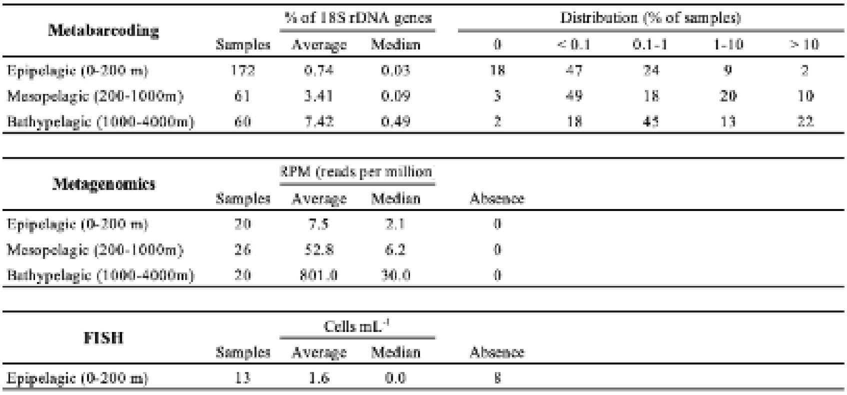


Figure 1. Widespread distribution of *Cafeteria burkhardae* in the global ocean. a) Relative abundance in three vertical regions of the ASV identical to *C. burkhardae* from a study of picoplankton diversity using V4 18S rDNA amplicons. Grey circles indicate absence of the ASV, while the area of red circles is proportional to the relative abundance (the scale applies to the three panels). b) Fragment recruitment analysis done with 66 metagenomes from the same expedition and the *C. burkhardae* genome as reference. All genome regions are mapped, with most metagenomic reads being >99% similar.



Metabarcoding		% of 18S rDNA genes		Distribution (% of samples)				
	Samples	Average	Median	0	< 0.1	0.1-1	1-10	> 10
Epipelagic (0-200 m)	172	0.74	0.03	18	47	24	9	2
Mesopelagic (200-1000m)	61	3.41	0.09	3	49	18	20	10
Bathypelagic (1000-4000m)	60	7.42	0.49	2	18	45	13	22

Metagenomics		RPM (reads per million		
	Samples	Average	Median	Absence
Epipelagic (0-200 m)	20	7.5	2.1	0
Mesopelagic (200-1000m)	26	52.8	6.2	0
Bathypelagic (1000-4000m)	20	801.0	30.0	0

FISH		Cells mL^{-1}		
	Samples	Average	Median	Absence
Epipelagic (0-200 m)	13	1.6	0.0	8

Table 1. Distribution of *C. burkhardae* in the global Malaspina survey by metabarcoding, metagenomics and FISH counts.

Dynamics of Cafeteria burkhardae in batch cultures

The cell dynamics of *C. burkhardae* and *Dokdonia* MED134 in the three batch cultures were highly reproducible (Fig. 2). After a short latency phase, there was a very fast growth of the flagellate population, so that over a 34 hour period densities increased from a few hundreds to 8 x 10^4 cells mL^{-1} in a perfect exponential growth curve (R^2 ≥0.99), yielding doubling times of 4.2-4.6 hours (Table S3). Parallel to the flagellate growth there was an exponential decay of bacteria, whose abundance fell from 25 to 3.5 x 10^6 cells mL^{-1}. The grazing rates in the three cultures were 40-49 bacteria flagellate^{-1} h^{-1}, and the estimated growth efficiencies were ~40%. Cultures remained relatively stable after the exponential phase, with similar bacterial numbers for weeks and a slow decrease of flagellate numbers, with half-life exponential decay of 121-140 hours. Flagellate cell size changed during the batch culture (Fig. 3), with larger cells at the exponential phase than at the stationary phase.

The three batch cultures were diluted 20-fold in the middle of the exponential phase to reduce bacterial abundances below the level supporting flagellate growth. Cell counts at different times after the dilution showed one or two divisions of the flagellate population, likely at the expense of what they had ingested before dilution, until they stopped growing (Fig. 2). Bacterial counts doubled only once, indicating no bacterial growth in sterile seawater. Flagellate cell sizes at the different dilution times were in between the exponential and stationary states (Fig. 3b). We regarded these dilutions as a different way of entering starvation, more gradual than the abrupt stationary state.

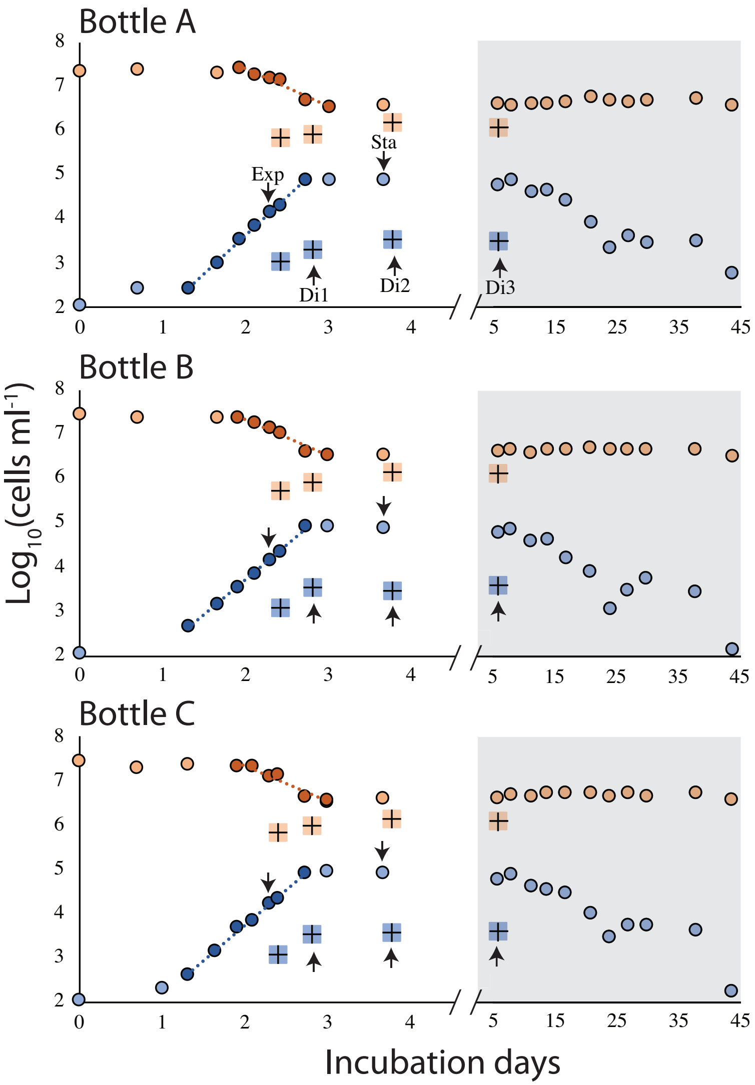


Figure 2. Abundance of bacteria (orange circles) and *C. burkhardae* (blue circles) in three parallel batch cultures. Points used to calculate the flagellate growth rate and the bacteria exponential decay are darker and display the derived linear regression. The abundances of both components during the dilution treatments are also shown (as colored crosses). Note the change of scale in the x-axis at the shaded area. Samples for transcriptomics are marked with an arrow.

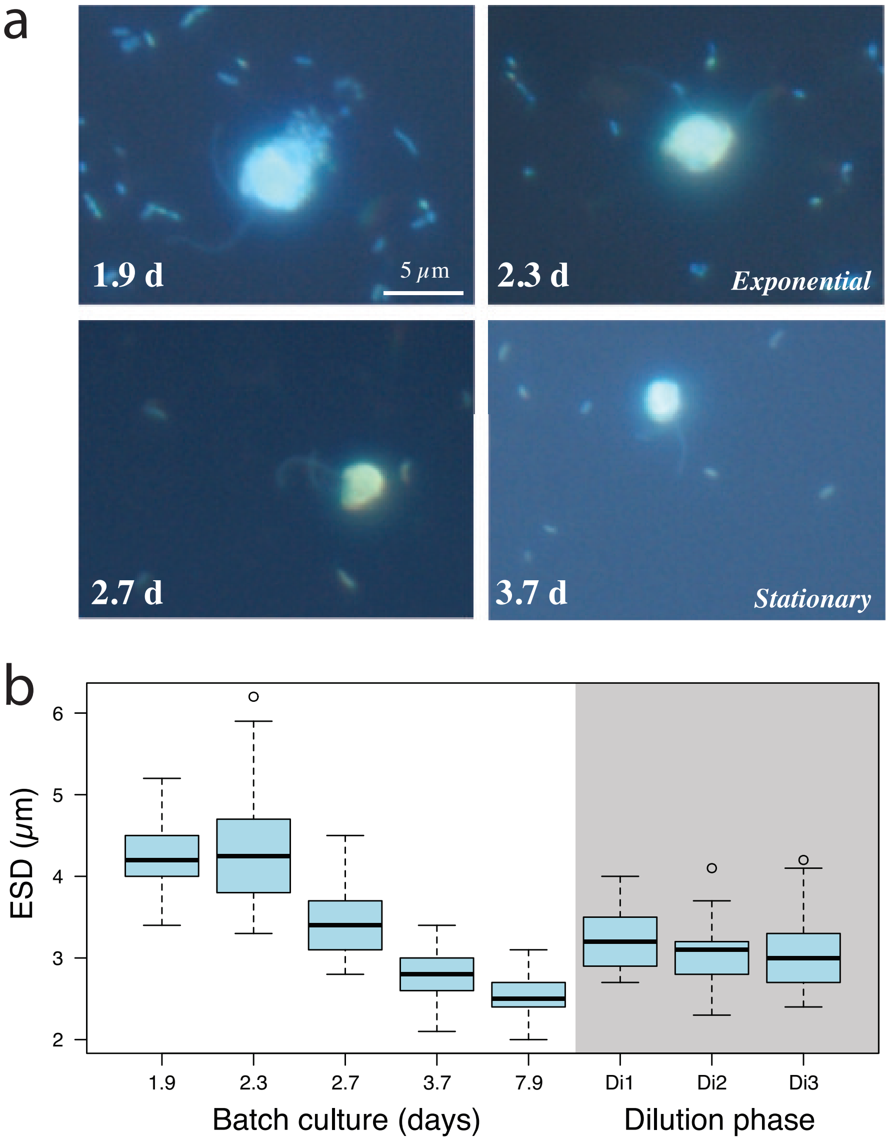


Figure 3. Cell size changes of *C. burkhardae* at different growth states. a) Epifluorescence microscope images of flagellates and bacteria in different days of the batch culture. The scale bar applies to all images. b) Box plots of the ESD (Equivalent Spherical Diameter) of about 50 cells during the batch culture and in the three dilution events.

De novo transcriptome of C. burkhardae and overall expression profiles

Gene expression analysis was performed in 21 samples from six phases: the Exponential phase, the Stationary phase, three states after starving by dilution for different times, and the Inoculum (Table S1). Each phase included a mix of biological replicates (different bottles) and technical replicates (same bottle). Poor quality raw Illumina reads and those mapping the *Dokdonia* sp. genome or the *C. burkhardae* rDNA operon were removed, leaving only about one third of the reads. These were assembled to generate a *de novo* transcriptome, which was then curated to keep transcripts with a high likelihood to belong to *C. burkhardae* based on genomic data, transcriptomic data, and functional annotations. The *de novo* transcriptome had 16 209 genes and an estimated BUSCO completeness of 82.2% (for comparison, the annotated genome has a BUSCO score of 83.8%, [26]).

Cleaned reads were mapped to the *de novo* transcriptome to get the TPM values of each transcript per sample (74.3% mapped reads on average, Table S1). We focused on the expression profiles of the 5 phases derived from well-controlled conditions. Samples from the same phase grouped together, while each phase occupied a different position in the NMDS plot (Fig. 4a). The three dilution events were placed orderly between Exponential and Stationary phases, following an apparent temporal trend of transcriptional activity. We then computed the differentially expressed (DE) genes between all phases (Table S4). Grouping of samples based on DE genes was consistent with their NMDS placement and showed that biological and technical replicates were indistinguishable, with Pearson correlation coefficients close to 1 (Fig. 4b), so they could all be treated as replicates of the experimental condition. Further analyses including the Inoculum and the MMETSP transcriptome (for which the culture state was undetermined) showed these two states were far from Exponential samples (Fig. S4). In particular, the Inoculum was placed between Dilution-3 and Stationary, while the MMETSP had a more distant position.

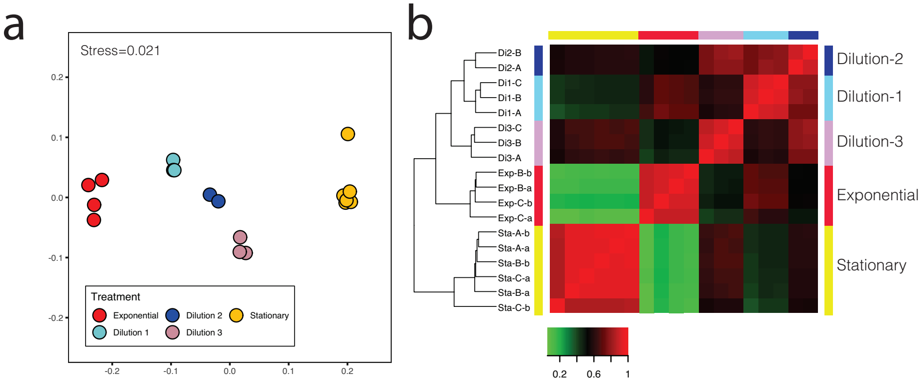
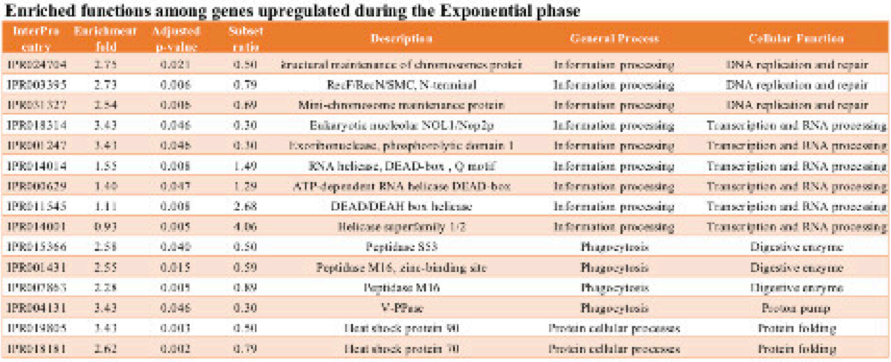


Figure 4. Comparison of the expression profiles of all samples in the five main states. **a** NMDS (non-metric multidimensional scaling) plot placing samples in a two dimensional space based on TPM values of all genes. **b** Heatmap showing Pearson correlation coefficients in sample pairwise comparisons based on differently expressed genes.

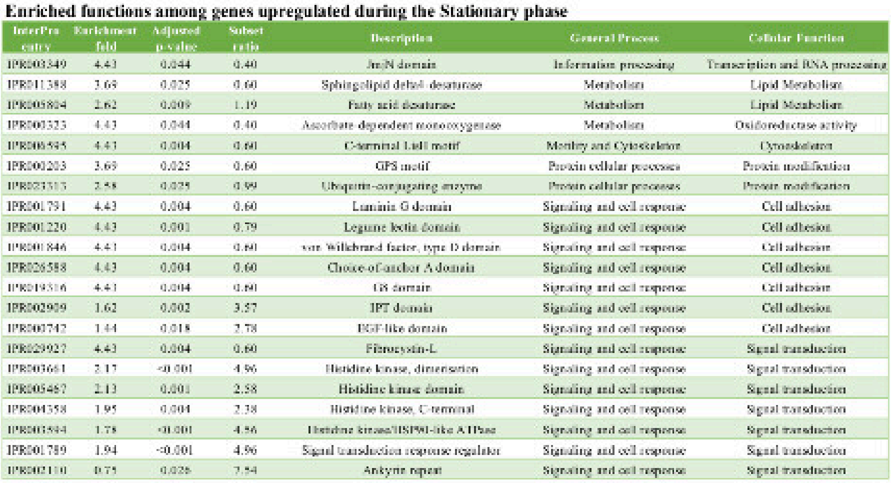
Differentially expressed genes and highly expressed genes

As the Exponential to Stationary pair presented the highest number of DE genes, with 1231 and 825 upregulated genes respectively, an enrichment analysis was performed to identify the biological functions associated to these DE gene sets (Table 2). Enriched functions among genes upregulated during the Exponential phase invoked a population of actively dividing cells, with proteins involved in DNA replication (structural maintenance of chromosome), transcription and RNA processing (RNA helicases, exoribonucleases) and protein remodeling (heat shock proteins). Phagocytosis was the other general process enriched in the Exponential phase, represented by digestive enzymes (Peptidases M16 and S53), and proton pumps (V-PPase). Among genes upregulated during the Stationary phase there was a striking enrichment of functions related to signaling and cell response, in particular signal transduction (histidine kinases) and cell adhesion (VWF and extracellular protein domains like EGF, laminin or lectin). Other intriguing functions enriched in the Stationary phase were those related to lipid metabolism (fatty acid desaturases).



Enriched functions among genes upregulated during the Exponential phase

InterPro entry	Enrichment fold	Adjusted p-value	Subset ratio	Description	General Process	Cellular Function
IPR024704	2.75	0.021	0.50	tructural maintenance of chromosomes protei	Information processing	DNA replication and repair
IPR003395	2.73	0.006	0.79	RecF/RecN/SMC, N-terminal	Information processing	DNA replication and repair
IPR031327	2.54	0.006	0.69	Mini-chromosome maintenance protein	Information processing	DNA replication and repair
IPR018314	3.43	0.046	0.30	Eukaryotic nucleolar NOL1/Nop2p	Information processing	Transcription and RNA processing
IPR001247	3.43	0.046	0.30	Exoribonuclease, phosphorolytic domain 1	Information processing	Transcription and RNA processing
IPR014014	1.55	0.008	1.49	RNA helicase, DEAD-box , Q motif	Information processing	Transcription and RNA processing
IPR000629	1.40	0.047	1.29	ATP-dependent RNA helicase DEAD-box	Information processing	Transcription and RNA processing
IPR011545	1.11	0.008	2.68	DEAD/DEAH box helicase	Information processing	Transcription and RNA processing
IPR014001	0.93	0.005	4.06	Helicase superfamily 1/2	Information processing	Transcription and RNA processing
IPR015366	2.58	0.040	0.50	Peptidase S53	Phagocytosis	Digestive enzyme
IPR001431	2.55	0.015	0.59	Peptidase M16, zinc-binding site	Phagocytosis	Digestive enzyme
IPR007863	2.28	0.005	0.89	Peptidase M16	Phagocytosis	Digestive enzyme
IPR004131	3.43	0.046	0.30	V-PPase	Phagocytosis	Proton pump
IPR019805	3.43	0.003	0.50	Heat shock protein 90	Protein cellular processes	Protein folding
IPR018181	2.62	0.002	0.79	Heat shock protein 70	Protein cellular processes	Protein folding



Enriched functions among genes upregulated during the Stationary phase

InterPro entry	Enrichment fold	Adjusted p-value	Subset ratio	Description	General Process	Cellular Function
IPR003349	4.43	0.044	0.40	JmjN domain	Information processing	Transcription and RNA processing
IPR011388	3.69	0.025	0.60	Sphingolipid delta4-desaturase	Metabolism	Lipid Metabolism
IPR005804	2.62	0.009	1.19	Fatty acid desaturase	Metabolism	Lipid Metabolism
IPR000323	4.43	0.044	0.40	Ascorbate dependent monooxygenase	Metabolism	Oxidoreductase activity
IPR006595	4.43	0.004	0.60	C-terminal LisH motif	Motility and Cytoskeleton	Cytoskeleton
IPR000203	3.69	0.025	0.60	GPS motif	Protein cellular processes	Protein modification
IPR023313	2.58	0.025	0.99	Ubiquitin-conjugating enzyme	Protein cellular processes	Protein modification
IPR001791	4.43	0.004	0.60	Laminin G domain	Signaling and cell response	Cell adhesion
IPR001220	4.43	0.001	0.79	Legume lectin domain	Signaling and cell response	Cell adhesion
IPR001846	4.43	0.004	0.60	von Willebrand factor, type D domain	Signaling and cell response	Cell adhesion
IPR026588	4.43	0.004	0.60	Choice-of-anchor A domain	Signaling and cell response	Cell adhesion
IPR019316	4.43	0.004	0.60	G8 domain	Signaling and cell response	Cell adhesion
IPR002909	1.62	0.002	3.57	IPT domain	Signaling and cell response	Cell adhesion
IPR000742	1.44	0.018	2.78	EGF-like domain	Signaling and cell response	Cell adhesion
IPR029927	4.43	0.004	0.60	Fibrocystin-L	Signaling and cell response	Signal transduction
IPR003661	2.17	<0.001	4.96	Histidine kinase, dimerisation	Signaling and cell response	Signal transduction
IPR005467	2.13	0.001	2.58	Histidine kinase domain	Signaling and cell response	Signal transduction
IPR004358	1.95	0.004	2.38	Histidine kinase, C-terminal	Signaling and cell response	Signal transduction
IPR003594	1.78	<0.001	4.56	Histidine kinase/HSP90-like ATPase	Signaling and cell response	Signal transduction
IPR001789	1.94	<0.001	4.96	Signal transduction response regulator	Signaling and cell response	Signal transduction
IPR002110	0.75	0.026	7.54	Ankyrin repeat	Signaling and cell response	Signal transduction

Table 2. Enriched functions based on InterPro domains in the subset of upregulated genes at the exponential phase (1231) or the stationary phase (825) as compared with the complete transcriptome. Enrichment fold values are reported in log2 scale. The subset ratio indicates the percentage of DE genes within each function.

We finally focused on the most highly expressed genes, those with an average TPM value >500 in any of the five phases. The selected 432 genes accounted for a considerable share of the expression signal in all samples (from 52 to 66%; 62% on average) and were manually assigned to a cellular function included in a general process. Comparing the Exponential and Stationary phases, we found that 79 of these highly expressed genes were upregulated in the Exponential phase, 94 in the

Stationary phase, and 259 were similarly expressed. These genes generally followed a regular expression pattern from Exponential to Stationary, with the dilution phases in between (Fig. S5). From this list, we selected a few relevant genes that may be optimal cornerstones to study specific process (Fig. 5). The function of many of them corresponded to the enriched functions found before (Table 2), and we also point to additional cases of genes upregulated in the Exponential phase (myosin, ubiquitin, elongation factor, peroxidase), or in the Stationary phase (chitin synthase, thiolase, cadherin, dehydrogenase).

The classification of highly expressed genes in functional categories allowed us to analyze functional expression changes in the different states (by adding up the TPM values of genes within each category). On a broad level (Fig. 6a), there were several general processes that decreased their expression from Exponential, through dilutions to the Stationary phase: protein cellular processes (which displayed the highest expression), phagocytosis, motility and cytoskeleton. The remaining general processes exhibited the opposite trend. On a more specific level (Fig. 6b), cellular functions that reduced their expression from Exponential to Stationary formed two groups, those with a sudden decrease (cytoskeleton, protein folding and proton pump) and those with a gradual decrease (transcription and translation machinery, TCA cycle, digestive enzymes, motility). Genes stimulated during starvation also displayed two distinct groups: those with a highly increased expression (lipid metabolism, cell adhesion, bactericidal proteins) and those with a moderate increase (transporters, amino acid and carbohydrate metabolism, signal transduction).

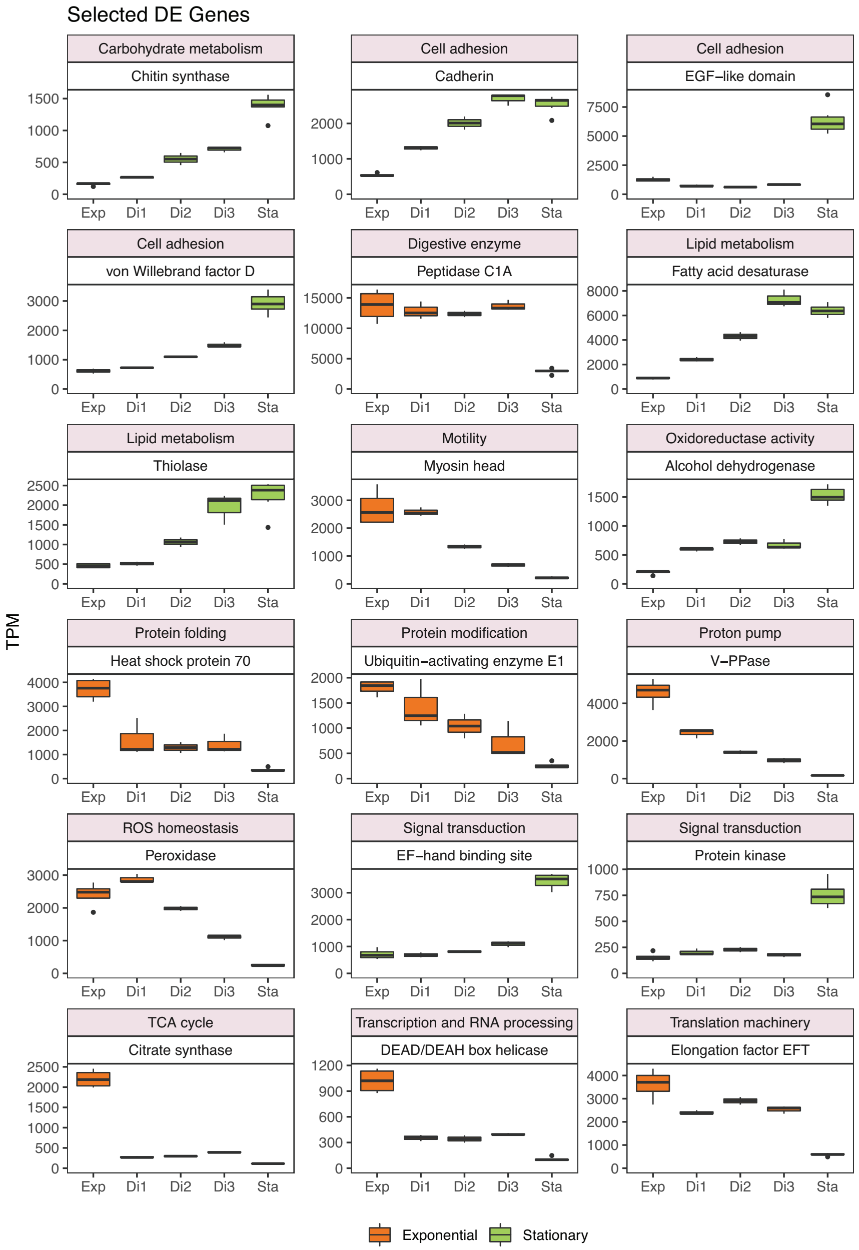


Figure 5. Box plots displaying the transcriptional changes along the five states of a few highly expressed genes. Genes are selected because they are differentially expressed in Exponential versus Stationary phases and appeal for important cellular functions.

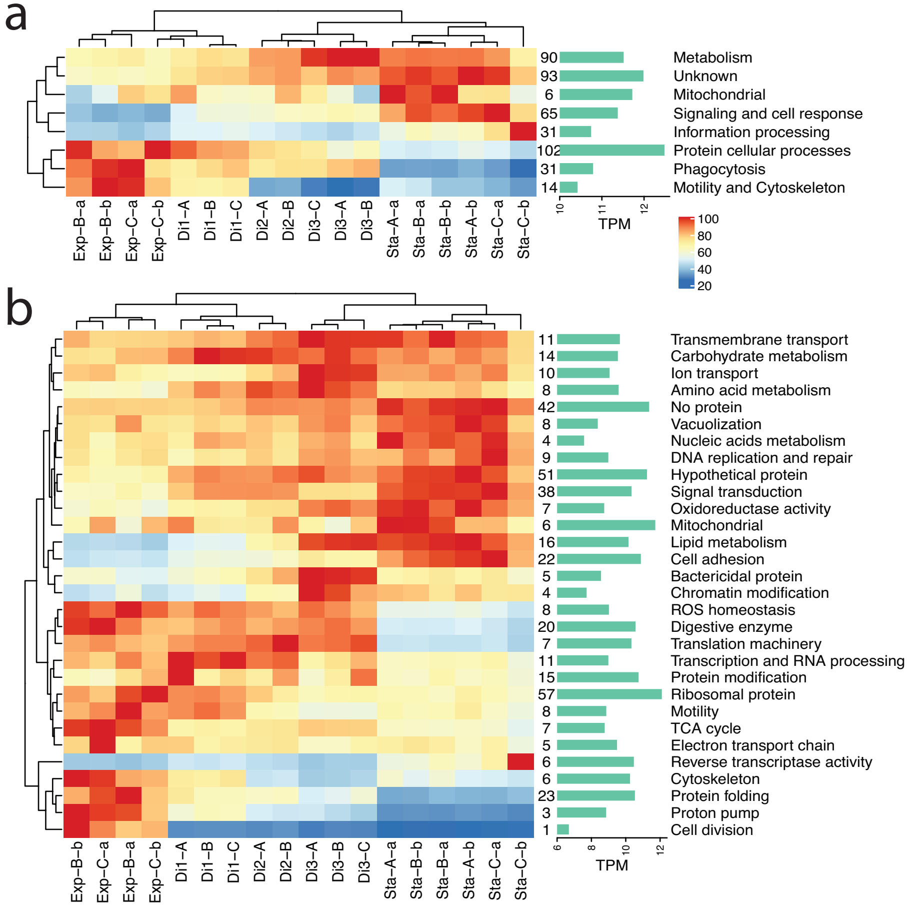


Figure 6. Gene expression changes of general processes (**a**) and associated cellular functions (**b**) computed by adding up the TPM values of highly expressed genes within these categories (numbers of genes per category shown after the heatmap). The data displayed in each cell represents the percentage with respect to the highest value in the process/function (considered 100%). Bar plots on the right display the actual TPM value of this highest cell.

DISCUSSION

An opportunistic and widely distributed heterotrophic flagellate

Marine microbial ecology has accepted the "uncultured majority" problem [56], where many ecologically relevant species are uncultured, and as a result we lack optimal ecophysiological models to interpret ecosystem processes. The genus *Cafeteria* was described decades ago [23], is easily cultured from marine samples [5], but was considered to be of little ecological relevance [29]. The analysis of sequencing data from the global Malaspina expedition, however, showed that *C. burkhardae* was a widespread species, often at very low abundance but with a few cases of high abundance. This patchiness contrasted with the log-normal distribution of other uncultured heterotrophic flagellates [35]. Its relative abundance increased through the water column, which does not need to imply an increase in cell counts, because of the drastic decrease of heterotrophic flagellates numbers with depth [57]. In addition, the metagenomic signal in the open sea matched perfectly with the genome of the cultured strain, indicating that this strain is a good representative of a widespread marine species.

Batch cultures allow a simple and quick evaluation of the growth and grazing kinetics of heterotrophic flagellates. In our cultures, *C. burkhardae* was a fast growing and ferocious predator, with grazing (50 bacteria h^{-1}) and growth rates (0.16 h^{-1}) comparable to the rates of cultured heterotrophic flagellates [27, 58]. Grazing rates of cultured species are higher than typical community rates, 2-20 bacteria h^{-1} [3]. *C. burkhardae* had a long survival at the stationary phase, with thousands of cells mL^{-1} still present after 40 days. Another interesting aspect was that the growth ceased at bacterial abundances of 3 x 10^{6} cells mL^{-1}, a density higher than typical bacterioplankton abundances of 10^{5}-10^{6} cells mL^{-1} in surface and 10^{4}-10^{5} in deep waters. This suggests that *C. burkhardae* may grow in patches of high food abundance, such as those found in permanent or ephemeral particles [59, 60]. The increase in cell volume during fast growth can be an adaptation to exploit temporary enriched environments. After explosive growth, this species can survive for weeks until a new particle is colonized. This feast and famine existence [61] is consistent with its patchy distribution and its increase with depth, as the relative importance of particles in microbial processes seems to increase with depth [62].

Transcriptional profiles in different physiological states

Transcriptomics is a promising and accessible way to gather new evolutionary and ecological insights into microbial eukaryotes [10], but few studies have been done with bacterivorous flagellates [21, 63, 64]. In some cases, the transcriptome is designed to retrieve genes for multigene phylogenies and, as seen here, many genes are expressed in all growth states. To fulfill our aim of identifying genes involved in phagocytosis, it was essential to link gene expression with the growth status. Accordingly, we put a considerable effort into sampling the exponential phase, which was challenging because only few hours separated the start of apparent growth and the stationary phase. Without a dedicated microscopic inspection, it would have been easy to miss this short window of time and sample dense and stationary cultures. That was likely the case for the MMETSP sample (and most bulk transcriptomes focused on gene discovery) that had a transcriptional profile closer to stationary samples. We also artificially "synchronized" cells to a gradual transition to starvation by dilution (by reducing bacterial encounter). The dilution samples had distinct expression profiles and were placed in an ordered manner between Exponential and Stationary phases (Figs. 4-6).

We identified a large number of genes (12.7% of total) that were differentially expressed between the Exponential and Stationary phases. Many of the DE genes upregulated in the Exponential phase were related to the functions expected in the scenario of a population of cells feeding, converting food to biomass and dividing: DNA replication, transcription, translation, protein modification, respiration, cytoskeleton reorganization, and phagocytosis. In the Stationary state, when cells had miniaturized to adapt to starvation, many upregulated genes related to signaling and cell response, with signal transduction across membranes and cell adhesion being the most significant, suggesting a crucial role in sensing the environment for hotspots to restart grazing and growth. The gene coding for fatty acid desaturase, which forms double bonds in fatty acids to increase membrane fluidity [65], was upregulated in the Stationary phase, perhaps to accommodate extracellular protein domains like cadherin, lectin and laminin in the membrane, also upregulated at this phase. Also intriguing was the high expression of chitin synthase, a gene that has been found in other stramenopiles that were not thought to contain chitin [66]. It could be speculated that chitin might provide cell rigidity to this species, contributing to its survival during starvation. Finally, many unknown genes were highly expressed (Fig. S5), some with homologous in other eukaryotes (hypothetical protein; 51 genes) and others with no match at all (no protein; 42 genes). More than half were differentially expressed, some upregulated at the Exponential (11 genes) but the majority at the Stationary (57 genes). These unknown DE genes represent interesting grounds for future functional genomics explorations.

Upregulated genes in Exponential state targeting phagocytosis

Phagocytosis is a very complex process involving the coordinated action of many proteins [16]. It is of great evolutionary and ecological significance, so one major aim of our study was to identify highly expressed genes functionally related to phagocytosis. The upregulated gene in the Exponential phase with the highest expression level coded for a digestive enzyme of the Peptidase C1A family, a group of cysteine peptidases that typically include lysosomal or secreted proteins [67, 68]. The majority of cathepsins, known to be activated in the acidic lysosomes, belong to this family. Other peptidases were also highly expressed in the Exponential phase: Peptidase S53, a serine peptidase with optimal pH of 3, and Peptidase M16, a metal dependent peptidase. Other upregulated digestive enzymes were Adenosylhomocysteine hydrolase, which hydrolyses the biosynthetic precursor of homocysteine, and the alpha/beta hydrolase fold that is common to hydrolytic enzymes of varied catalytic function.

Digestive enzymes used in phagocytosis operate in the acidic environment of mature phagosomes, which are acidified by the action of the transmembrane proton pumps V-ATPases and V-PPases [69]. Although both types were found in *C. burkhardae*, the V-PPase (vacuolar pyrophosphatase) exhibited a higher expression, being the fifth most highly expressed gene in the Exponential state. So, this proton pump seems to be responsible for phagosome acidification in this species. In a recent experiment we identified a high expression of rhodopsin in the uncultured MAST-4 heterotrophic flagellate [70], and hypothesized that the coding protein acted as a light-driven proton pump that contributed to phagosome acidification. Even though rhodopsin genes were found in the *C. burkhardae* transcriptome, they were never highly expressed. This may explain why this species is not restricted to photic waters.

Finally, two of the highly expressed genes in the Exponential phase were peroxidases. The canonical function of these enzymes is to detoxify deleterious reactive oxygen species (ROS). In the reverse action, peroxidases can produce ROS radicals, which in phagocytes of the animal immune system participate in killing pathogens [71]. In free-living protists that use phagocytosis for nutrition, such as the amoebozoan *Dictyostelium*, the involvement of ROS radicals in prey processing has not been demonstrated [18], but our data suggest they may possibly play a role in prey digestion, although this is currently speculative.

Concluding remarks

Functional and genomic analyses with marine bacterivorous heterotrophic flagellates have been limited by the lack of appropriate model species. Using molecular diversity surveys, we show that the well-known cultured species *Cafeteria burkhardae* is widespread in the ocean and seems to be an opportunistic species that grows fast in patches of high bacterial density and becomes a good survivor in the diluted surrounding seawater. In batch cultures, *C. burkhardae* presents marked changes in gene expression when actively growing and when starving, and we identified promising gene sets specific for each state. Whether or not this match with the genetic machinery at play in natural communities, where this species faces complex biotic and abiotic interactions, remains an open question. Among the most interesting genes during active grazing are those related to phagocytosis, such as digestive enzymes, proton pumps, and perhaps peroxidases. Future studies with other cultured heterotrophic flagellates, or even more interestingly with natural or manipulated assemblages [70], will be necessary to evaluate if these genes are functionally relevant in other species as well, in which case they will represent promising markers to study bacterivory in the oceans.

ACKNOWLEDGMENTS

This research was supported by the Spanish Ministry of Economy and Competitiveness project ALLFLAGS (CTM2016-75083-R) to RM, by the Russian Foundation for Basic Research (grant No 18-504-51028) within the framework of Ministry of Education and Science of the Russian Federation project No AAAA-A18-118012690098-5 to DVT, and by the European Union Marie Skłodowska-Curie Grant agreement H2020-MSCA-ITN-2015-675752 to AL and FB. We are grateful to Christoph Deeg and Curtis Suttle (University of British Columbia, Canada) who kindly provided the *Cafeteria burkhardae* culture.

SUPPLEMENTARY INFORMATION

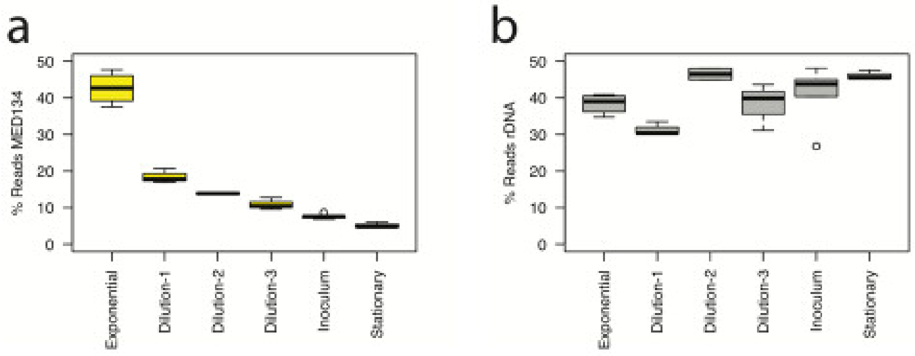
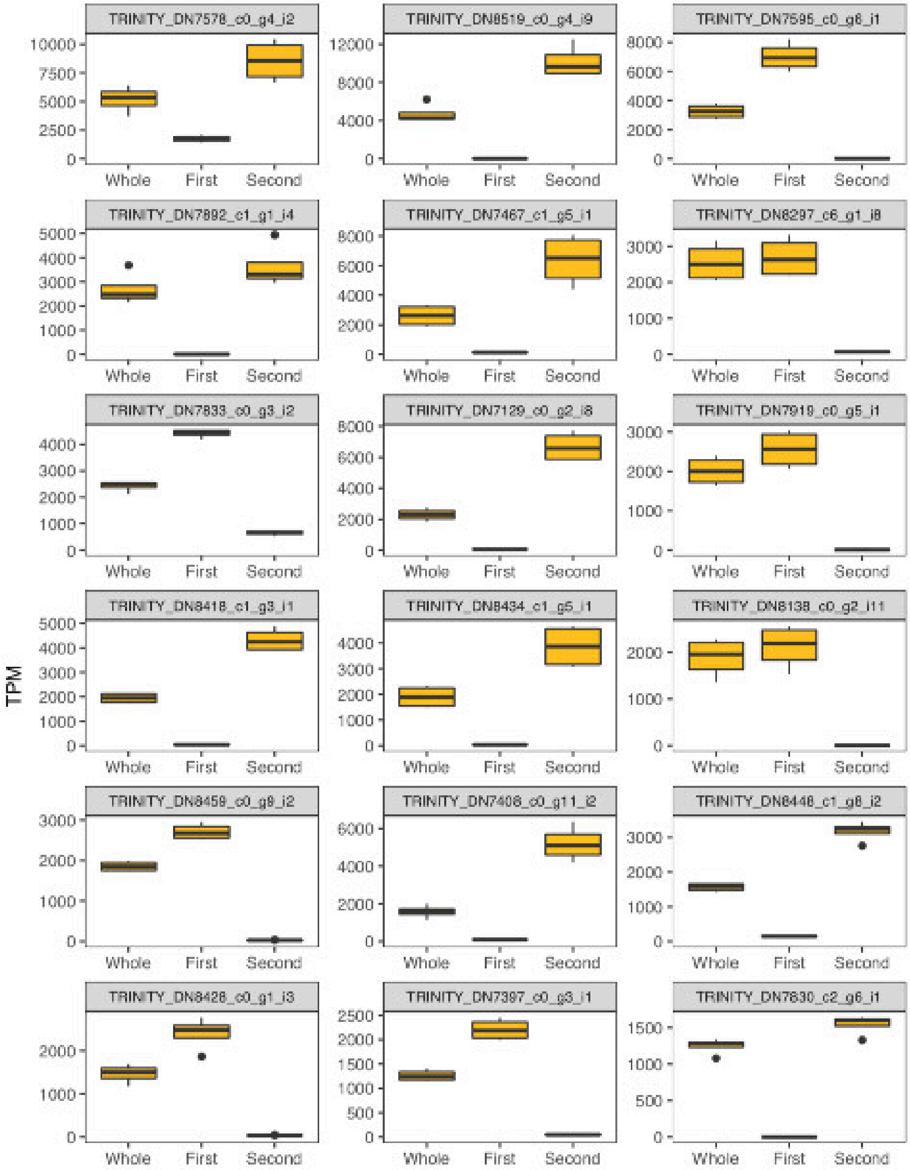
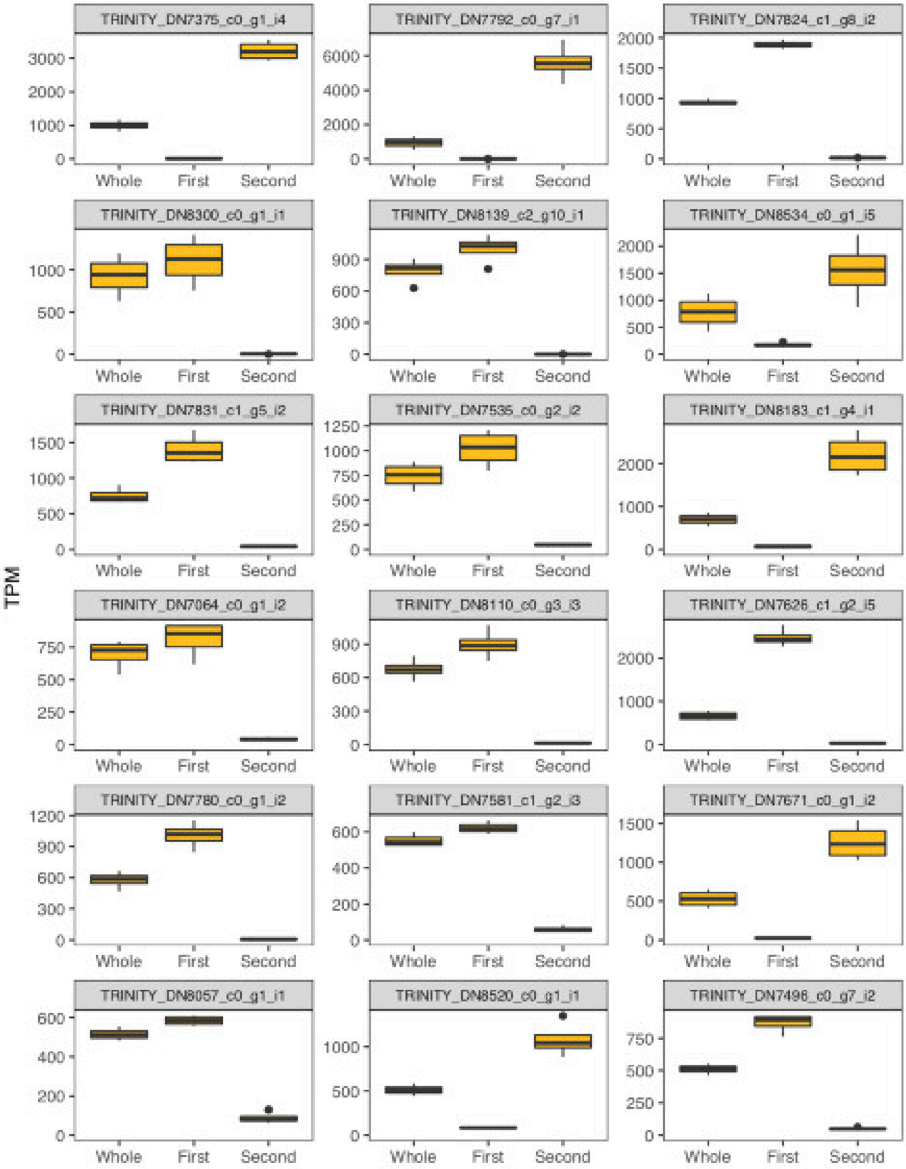


Figure S1. Cleaning of sequencing reads before *de novo* transcriptome assembly. **a** Percentage of reads mapping the *Dokdonia* MED134 genome. **b** Percentage of reads mapping the operon of *C. burkhardae* (after excluding *Dokdonia* reads).


TPM
TRINITY_DN7578_c0_g4_i2
10000
7500
5000
2500
0
TRINITY_DN8519_c0_g4_i9
12000
8000
4000
0
TRINITY_DN7595_c0_g6_i1
8000
6000
4000
2000
0
TRINITY_DN7892_c1_g1_i4
5000
4000
3000
2000
1000
0
TRINITY_DN7467_c1_g5_i1
8000
6000
4000
2000
0
TRINITY_DN8297_c6_g1_i8
3000
2000
1000
0
TRINITY_DN7833_c0_g3_i2
4000
3000
2000
1000
0
TRINITY_DN7129_c0_g2_i8
8000
6000
4000
2000
0
TRINITY_DN7919_c0_g5_i1
3000
2000
1000
0
TRINITY_DN8418_c1_g3_i1
5000
4000
3000
2000
1000
0
TRINITY_DN8434_c1_g5_i1
4000
3000
2000
1000
0
TRINITY_DN8138_c0_g2_i11
2000
1000
0
TRINITY_DN8459_c0_g8_i2
3000
2000
1000
0
TRINITY_DN7408_c0_g11_i2
6000
4000
2000
0
TRINITY_DN8448_c1_g8_i2
3000
2000
1000
0
TRINITY_DN8428_c0_g1_i3
2000
1000
0
TRINITY_DN7397_c0_g3_i1
2500
2000
1500
1000
500
0
TRINITY_DN7830_c2_g6_i1
1500
1000
500
0
Whole
First
Second


TPM
TRINITY_DN7375_c0_g1_i4
0 1000 2000 3000
TRINITY_DN7792_c0_g7_i1
0 2000 4000 6000
TRINITY_DN7824_c1_g8_i2
0 500 1000 1500 2000
TRINITY_DN8300_c0_g1_i1
0 500 1000
TRINITY_DN8139_c2_g10_i1
0 300 600 900
TRINITY_DN8534_c0_g1_i5
0 500 1000 1500 2000
TRINITY_DN7831_c1_g5_i2
0 500 1000 1500
TRINITY_DN7535_c0_g2_i2
0 250 500 750 1000 1250
TRINITY_DN8183_c1_g4_i1
0 1000 2000
TRINITY_DN7064_c0_g1_i2
0 250 500 750
TRINITY_DN8110_c0_g3_i3
0 300 600 900
TRINITY_DN7626_c1_g2_i5
0 1000 2000
TRINITY_DN7780_c0_g1_i2
0 300 600 900 1200
TRINITY_DN7581_c1_g2_i3
0 200 400 600
TRINITY_DN7671_c0_g1_i2
0 500 1000 1500
TRINITY_DN8057_c0_g1_i1
0 200 400 600
TRINITY_DN8520_c0_g1_i1
0 500 1000
TRINITY_DN7496_c0_g7_i2
0 250 500 750
Whole First Second

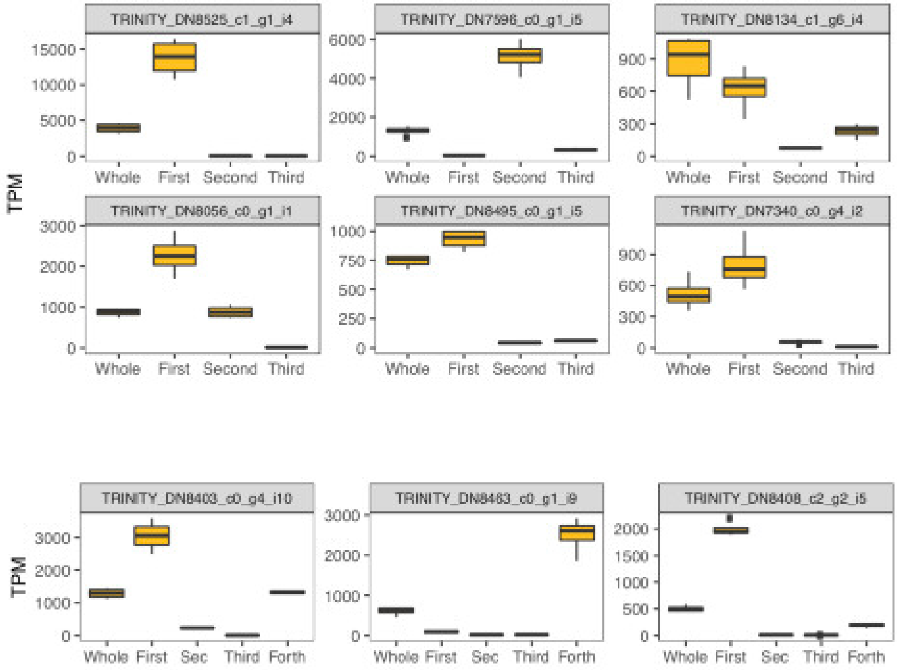


Figure S2. Expression level of selected transcripts before and after being split in 2-4 fragments based on the presence of ORFs with different functional predictions. The list shows highly expressed transcripts in the Exponential phase (average TPM >500 before splitting).

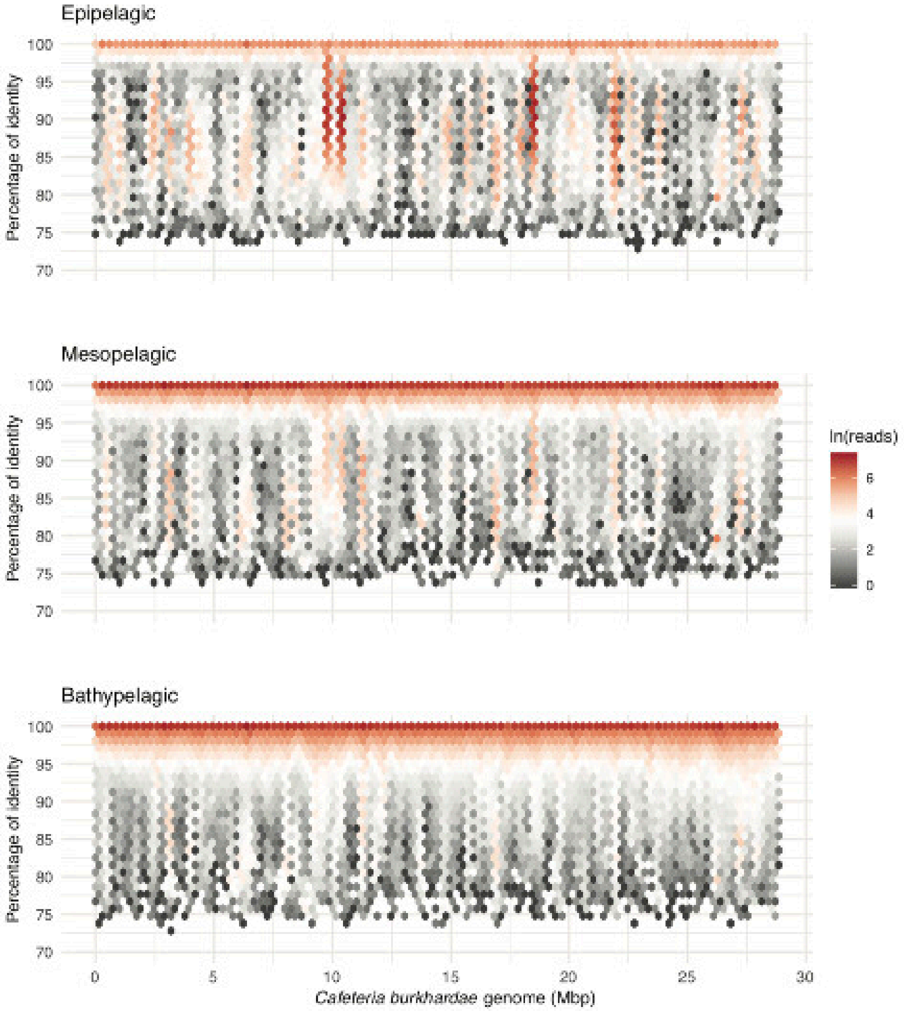


Figure S3. Fragment recruitment analysis done with metagenomes from the Malaspina expedition and the *C. burkhardae* genome as reference. Data is separated in the three water column regions, epipelagic (20 metagenomes), mesopelagic (26), and bathypelagic (20).

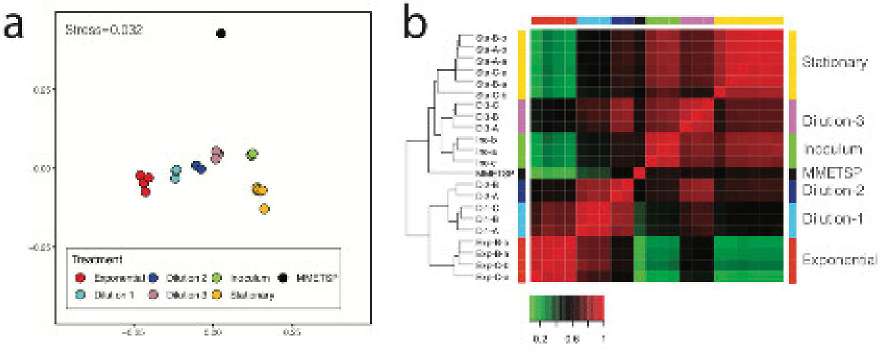
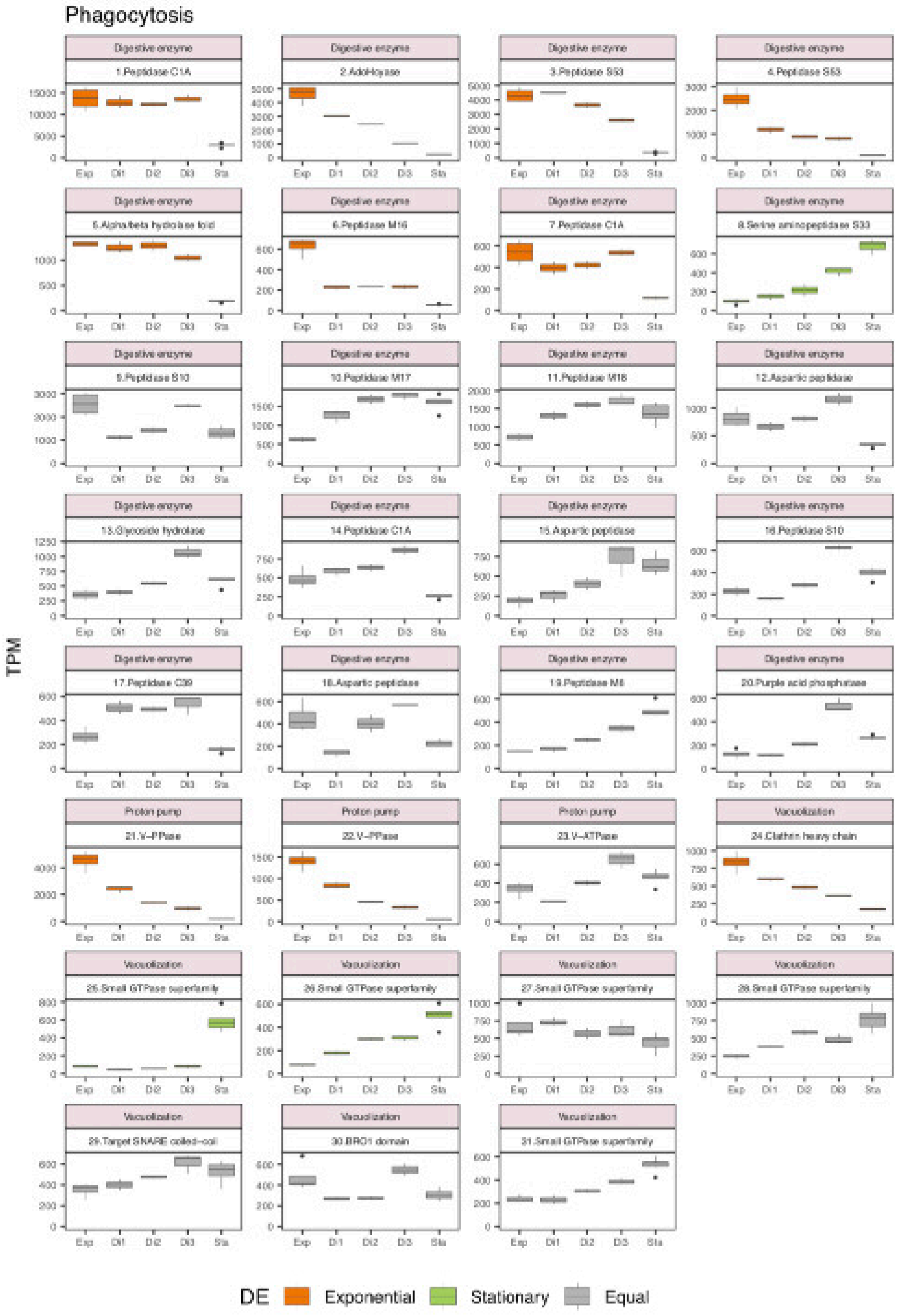
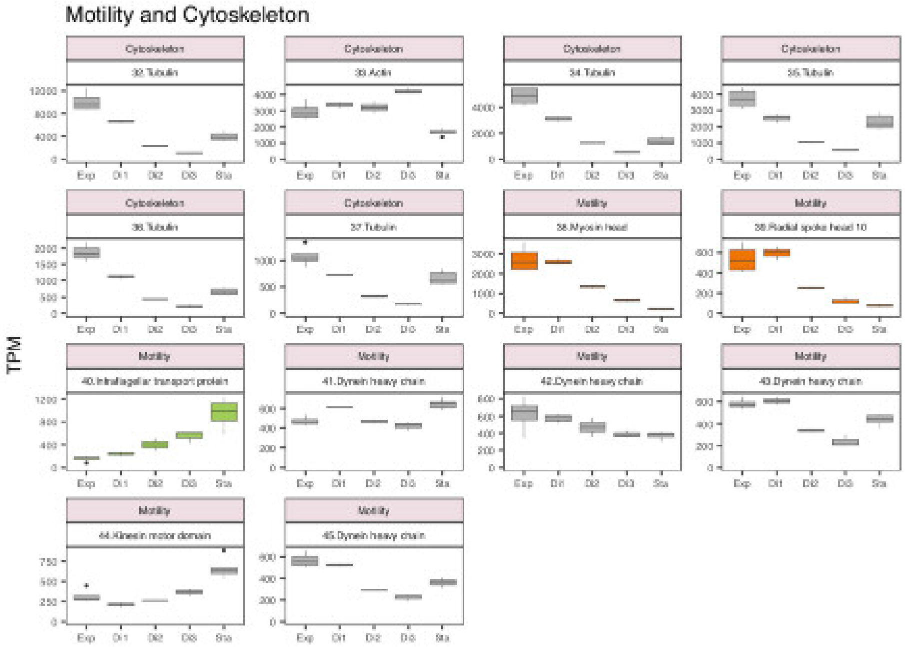
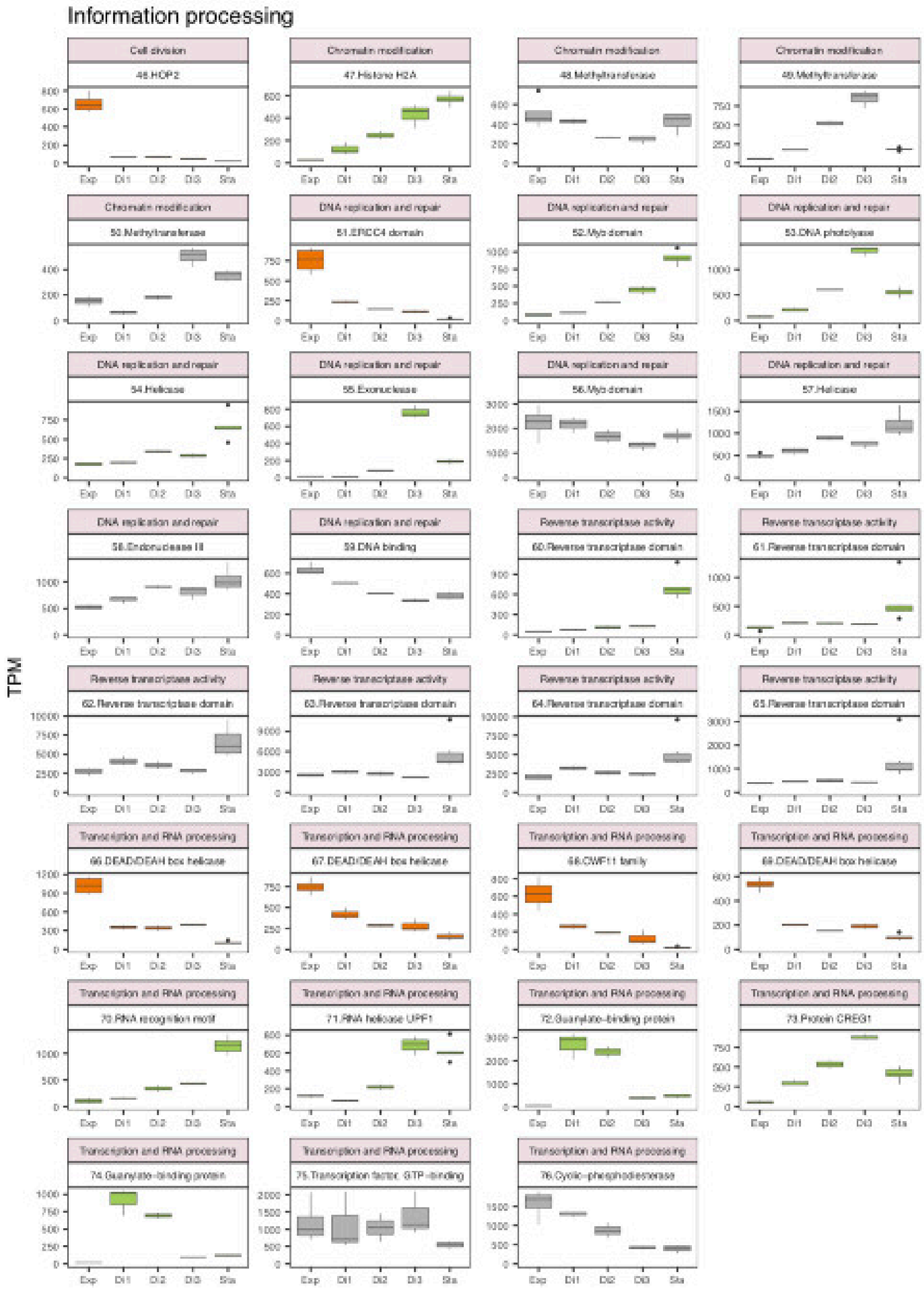
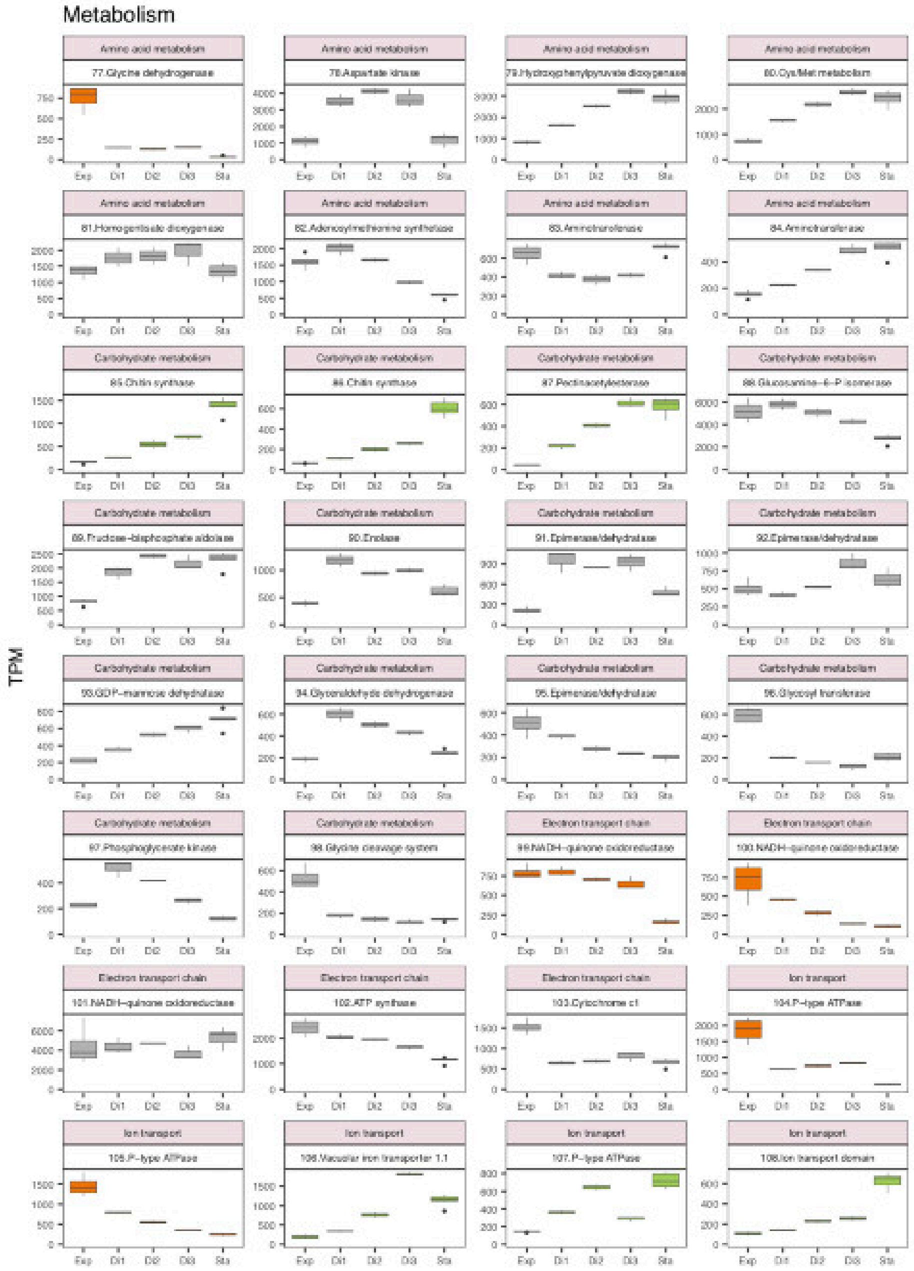


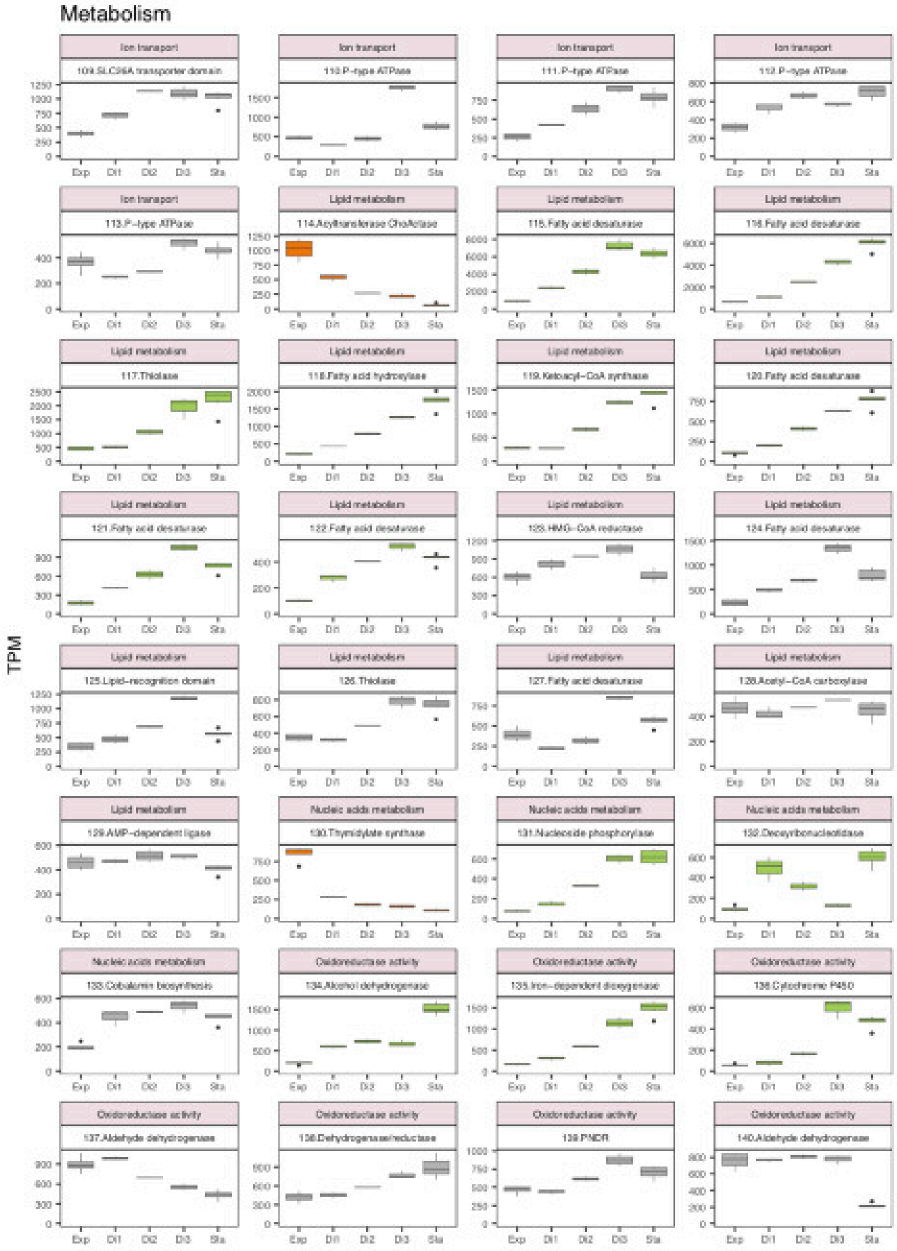
Figure S4. Comparison of the expression profiles of all samples in the five main states plus the Inoculum, and the MMETSP transcriptome. **a** NMDS plot placing samples in a two dimensional space based on TPM values of all genes. **b** Heatmap showing Pearson correlation coefficients in sample pairwise comparisons based on differently expressed genes.


Phagocytosis
TPM
Digestive enzyme
1.Peptidase C1A
Digestive enzyme
2.AdoHcyase
Digestive enzyme
3.Peptidase S53
Digestive enzyme
4.Peptidase S53
Digestive enzyme
5.Alpha/beta hydrolase fold
Digestive enzyme
6.Peptidase M16
Digestive enzyme
7.Peptidase C1A
Digestive enzyme
8.Serine aminopeptidase S33
Digestive enzyme
9.Peptidase S10
Digestive enzyme
10.Peptidase M17
Digestive enzyme
11.Peptidase M18
Digestive enzyme
12.Aspartic peptidase
Digestive enzyme
13.Glycoside hydrolase
Digestive enzyme
14.Peptidase C1A
Digestive enzyme
15.Aspartic peptidase
Digestive enzyme
16.Peptidase S10
Digestive enzyme
17.Peptidase C39
Digestive enzyme
18.Aspartic peptidase
Digestive enzyme
19.Peptidase M8
Digestive enzyme
20.Purple acid phosphatase
Proton pump
21.V–PPase
Proton pump
22.V–PPase
Proton pump
23.V–ATPase
Vacuolization
24.Clathrin heavy chain
Vacuolization
25.Small GTPase superfamily
Vacuolization
26.Small GTPase superfamily
Vacuolization
27.Small GTPase superfamily
Vacuolization
28.Small GTPase superfamily
Vacuolization
29.Target SNARE coiled–coil
Vacuolization
30.BRO1 domain
Vacuolization
31.Small GTPase superfamily
Exp
Di1
Di2
Di3
Sta
DE
Exponential
Stationary
Equal

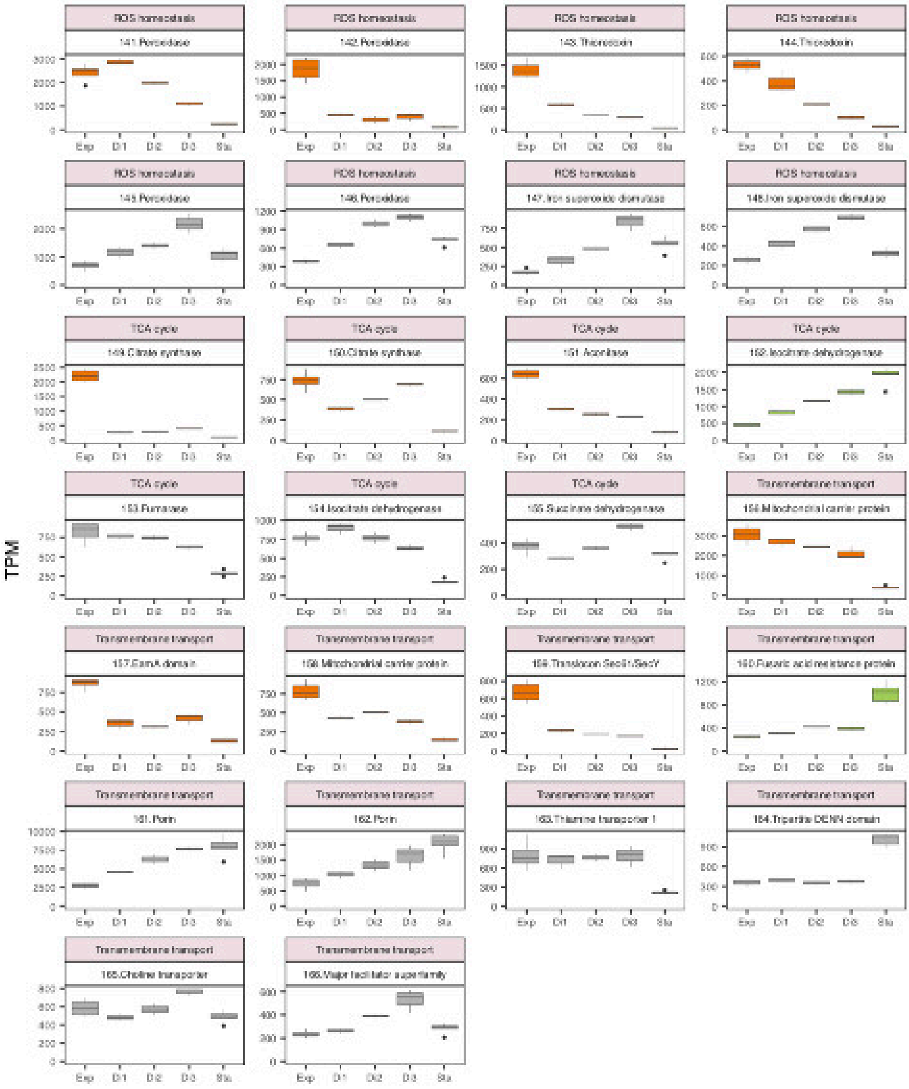

Motility and Cytoskeleton
TPM
Cytoskeleton
32.Tubulin
Cytoskeleton
33.Actin
Cytoskeleton
34.Tubulin
Cytoskeleton
35.Tubulin
Cytoskeleton
36.Tubulin
Cytoskeleton
37.Tubulin
Motility
38.Myosin head
Motility
39.Radial spoke head 10
Motility
40.Intraflagellar transport protein
Motility
41.Dynein heavy chain
Motility
42.Dynein heavy chain
Motility
43.Dynein heavy chain
Motility
44.Kinesin motor domain
Motility
45.Dynein heavy chain
Exp
Di1
Di2
Di3
Sta

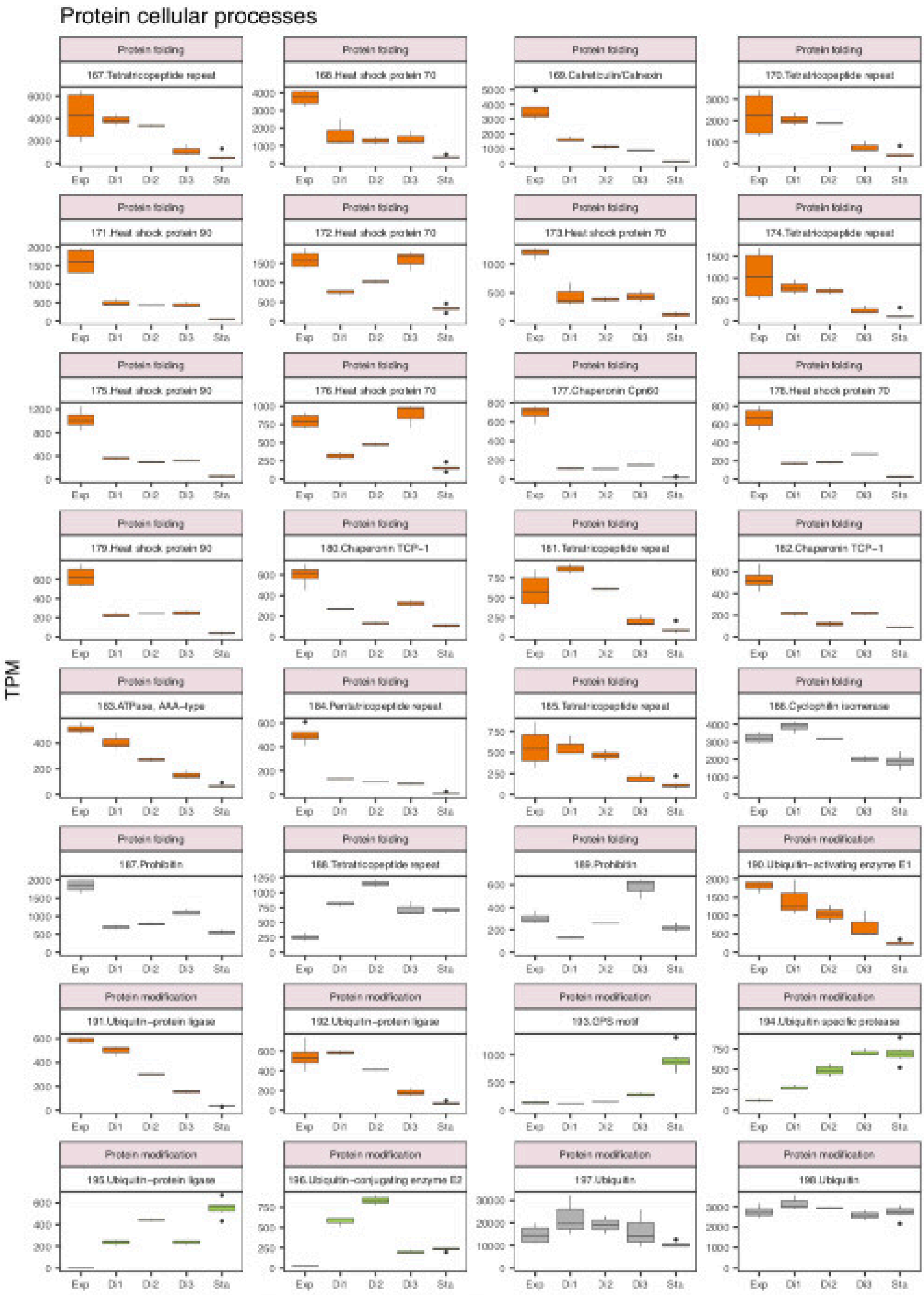

Information processing
TPM
Cell division
46.HOP2
Chromatin modification
47.Histone H2A
Chromatin modification
48.Methyltransferase
Chromatin modification
49.Methyltransferase
Chromatin modification
50.Methyltransferase
DNA replication and repair
51.ERCC4 domain
DNA replication and repair
52.Myb domain
DNA replication and repair
53.DNA photolyase
DNA replication and repair
54.Helicase
DNA replication and repair
55.Exonuclease
DNA replication and repair
56.Myb domain
DNA replication and repair
57.Helicase
DNA replication and repair
58.Endonuclease III
DNA replication and repair
59.DNA binding
Reverse transcriptase activity
60.Reverse transcriptase domain
Reverse transcriptase activity
61.Reverse transcriptase domain
Reverse transcriptase activity
62.Reverse transcriptase domain
Reverse transcriptase activity
63.Reverse transcriptase domain
Reverse transcriptase activity
64.Reverse transcriptase domain
Reverse transcriptase activity
65.Reverse transcriptase domain
Transcription and RNA processing
66.DEAD/DEAH box helicase
Transcription and RNA processing
67.DEAD/DEAH box helicase
Transcription and RNA processing
68.CWF11 family
Transcription and RNA processing
69.DEAD/DEAH box helicase
Transcription and RNA processing
70.RNA recognition motif
Transcription and RNA processing
71.RNA helicase UPF1
Transcription and RNA processing
72.Guanylate-binding protein
Transcription and RNA processing
73.Protein CREG1
Transcription and RNA processing
74.Guanylate-binding protein
Transcription and RNA processing
75.Transcription factor, GTP-binding
Transcription and RNA processing
76.Cyclic-phosphodiesterase
Exp
Di1
Di2
Di3
Sta

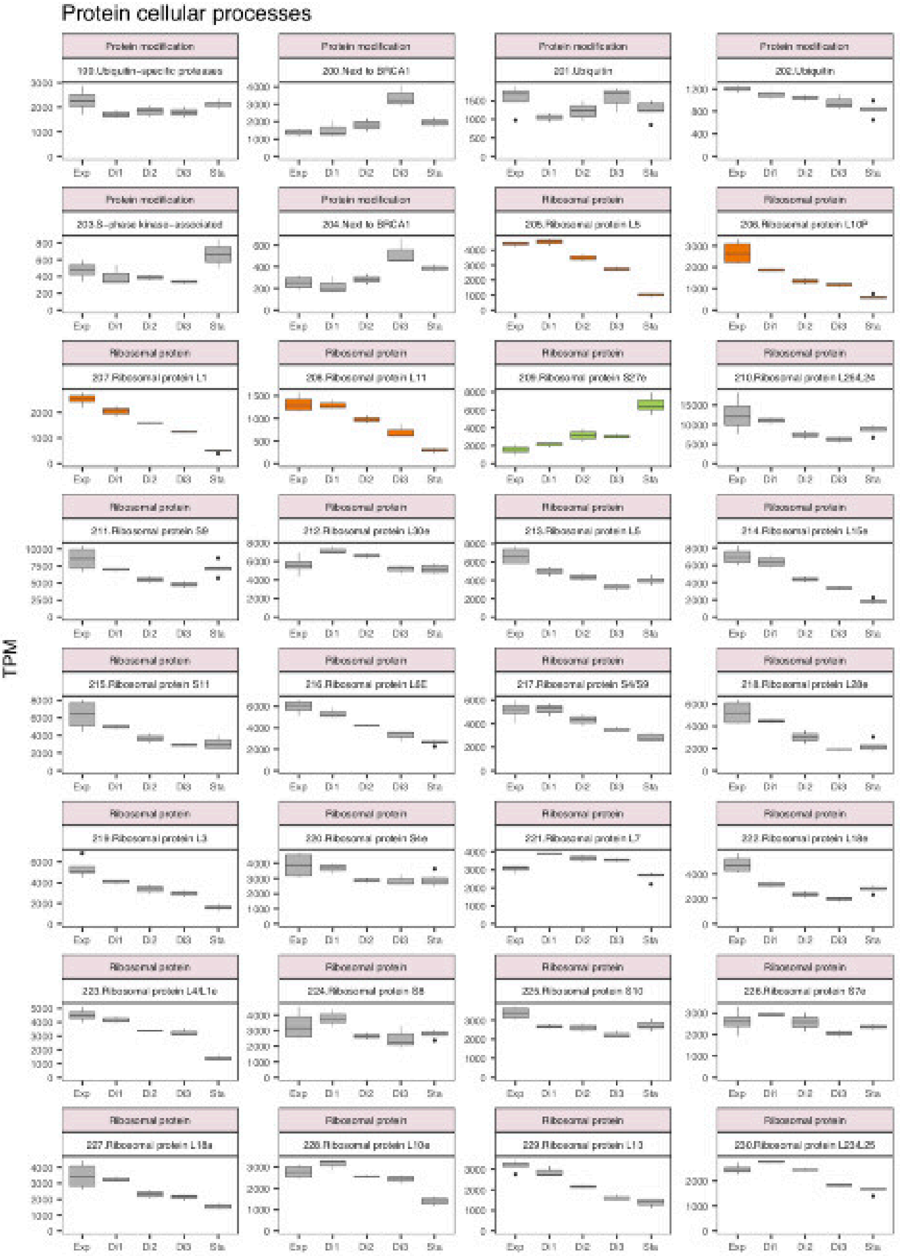

Metabolism
TPM
Amino acid metabolism
77.Glycine dehydrogenase
0
250
500
750
Amino acid metabolism
78.Aspartate kinase
0
1000
2000
3000
4000
Amino acid metabolism
79.Hydroxyphenylpyruvate dioxygenase
0
1000
2000
3000
Amino acid metabolism
80.Cys/Met metabolism
0
1000
2000
Amino acid metabolism
81.Homogentisate dioxygenase
0
500
1000
1500
2000
Amino acid metabolism
82.Adenosylmethionine synthetase
0
500
1000
1500
2000
Amino acid metabolism
83.Aminotransferase
0
200
400
600
Amino acid metabolism
84.Aminotransferase
0
200
400
Carbohydrate metabolism
85.Chitin synthase
0
500
1000
1500
Carbohydrate metabolism
86.Chitin synthase
0
200
400
600
Carbohydrate metabolism
87.Pectinacetylesterase
0
200
400
600
Carbohydrate metabolism
88.Glucosamine−6−P isomerase
0
2000
4000
6000
Carbohydrate metabolism
89.Fructose−bisphosphate aldolase
0
500
1000
1500
2000
2500
Carbohydrate metabolism
90.Enolase
0
500
1000
Carbohydrate metabolism
91.Epimerase/dehydratase
0
300
600
900
Carbohydrate metabolism
92.Epimerase/dehydratase
0
250
500
750
1000
Carbohydrate metabolism
93.GDP−mannose dehydratase
0
200
400
600
800
Carbohydrate metabolism
94.Glyceraldehyde dehydrogenase
0
200
400
600
Carbohydrate metabolism
95.Epimerase/dehydratase
0
200
400
600
Carbohydrate metabolism
96.Glycosyl transferase
0
200
400
600
Carbohydrate metabolism
97.Phosphoglycerate kinase
0
200
400
Carbohydrate metabolism
98.Glycine cleavage system
0
200
400
600
Electron transport chain
99.NADH−quinone oxidoreductase
0
250
500
750
Electron transport chain
100.NADH−quinone oxidoreductase
0
250
500
750
Electron transport chain
101.NADH−quinone oxidoreductase
0
2000
4000
6000
Electron transport chain
102.ATP synthase
0
1000
2000
Electron transport chain
103.Cytochrome c1
0
500
1000
1500
Ion transport
104.P−type ATPase
0
500
1000
1500
2000
Ion transport
105.P−type ATPase
0
500
1000
1500
Ion transport
106.Vacuolar iron transporter 1.1
0
500
1000
1500
Ion transport
107.P−type ATPase
0
200
400
600
800
Ion transport
108.Ion transport domain
0
200
400
600
Exp
Di1
Di2
Di3
Sta

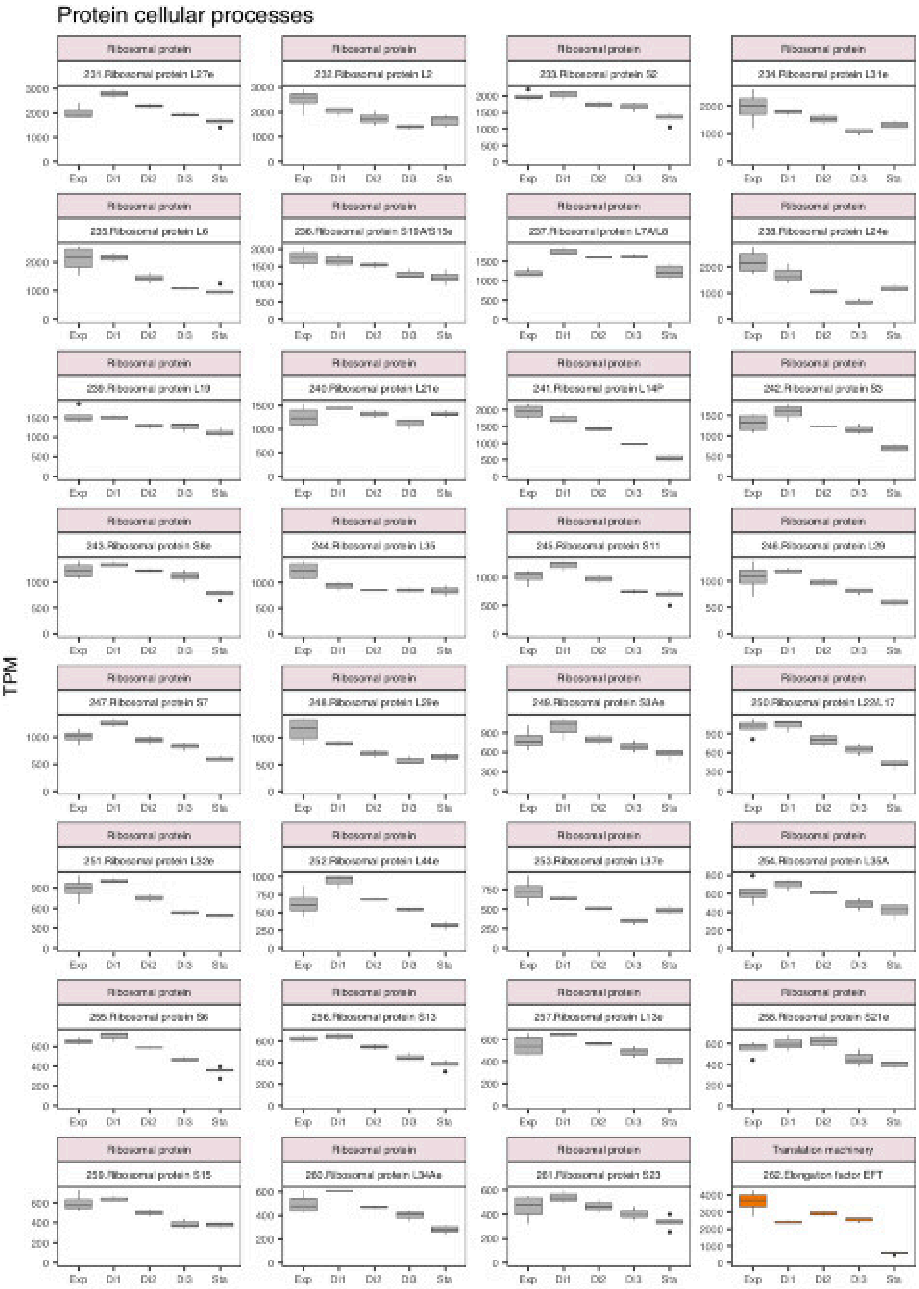

Metabolism
TPM
Ion transport
109.SLC38A transporter domain
Ion transport
110.P-type ATPase
Ion transport
111.P-type ATPase
Ion transport
112.P-type ATPase
Ion transport
113.P-type ATPase
Lipid metabolism
114.Acyltransferase ChoActase
Lipid metabolism
115.Fatty acid desaturase
Lipid metabolism
116.Fatty acid desaturase
Lipid metabolism
117.Thiolase
Lipid metabolism
118.Fatty acid hydroxylase
Lipid metabolism
119.Ketoacyl-CoA synthase
Lipid metabolism
120.Fatty acid desaturase
Lipid metabolism
121.Fatty acid desaturase
Lipid metabolism
122.Fatty acid desaturase
Lipid metabolism
123.HMG-CoA reductase
Lipid metabolism
124.Fatty acid desaturase
Lipid metabolism
125.Lipid-recognition domain
Lipid metabolism
126.Thiolase
Lipid metabolism
127.Fatty acid desaturase
Lipid metabolism
128.Acetyl-CoA carboxylase
Lipid metabolism
129.AMP-dependent ligase
Nucleic acids metabolism
130.Thymidylate synthase
Nucleic acids metabolism
131.Nucleoside phosphorylase
Nucleic acids metabolism
132.Deoxyribonucleotidase
Nucleic acids metabolism
133.Cobalamin biosynthesis
Oxidoreductase activity
134.Alcohol dehydrogenase
Oxidoreductase activity
135.Iron-dependent dioxygenase
Oxidoreductase activity
136.Cytochrome P450
Oxidoreductase activity
137.Aldehyde dehydrogenase
Oxidoreductase activity
138.Dehydrogenase/reductase
Oxidoreductase activity
139.PNDR
Oxidoreductase activity
140.Aldehyde dehydrogenase
Exp
Di1
Di2
Di3
Sta

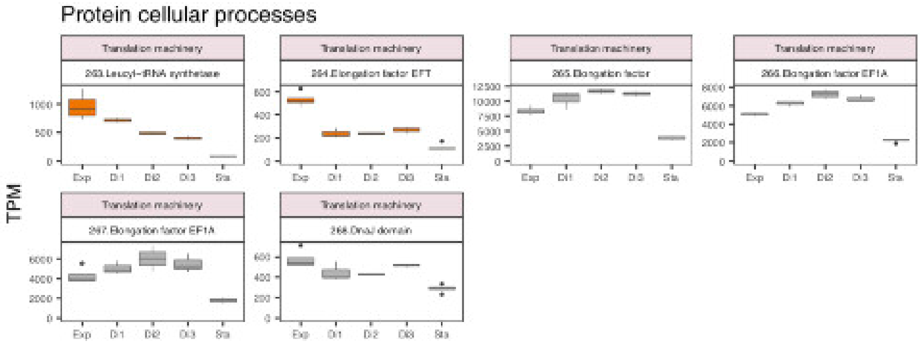
Metabolism

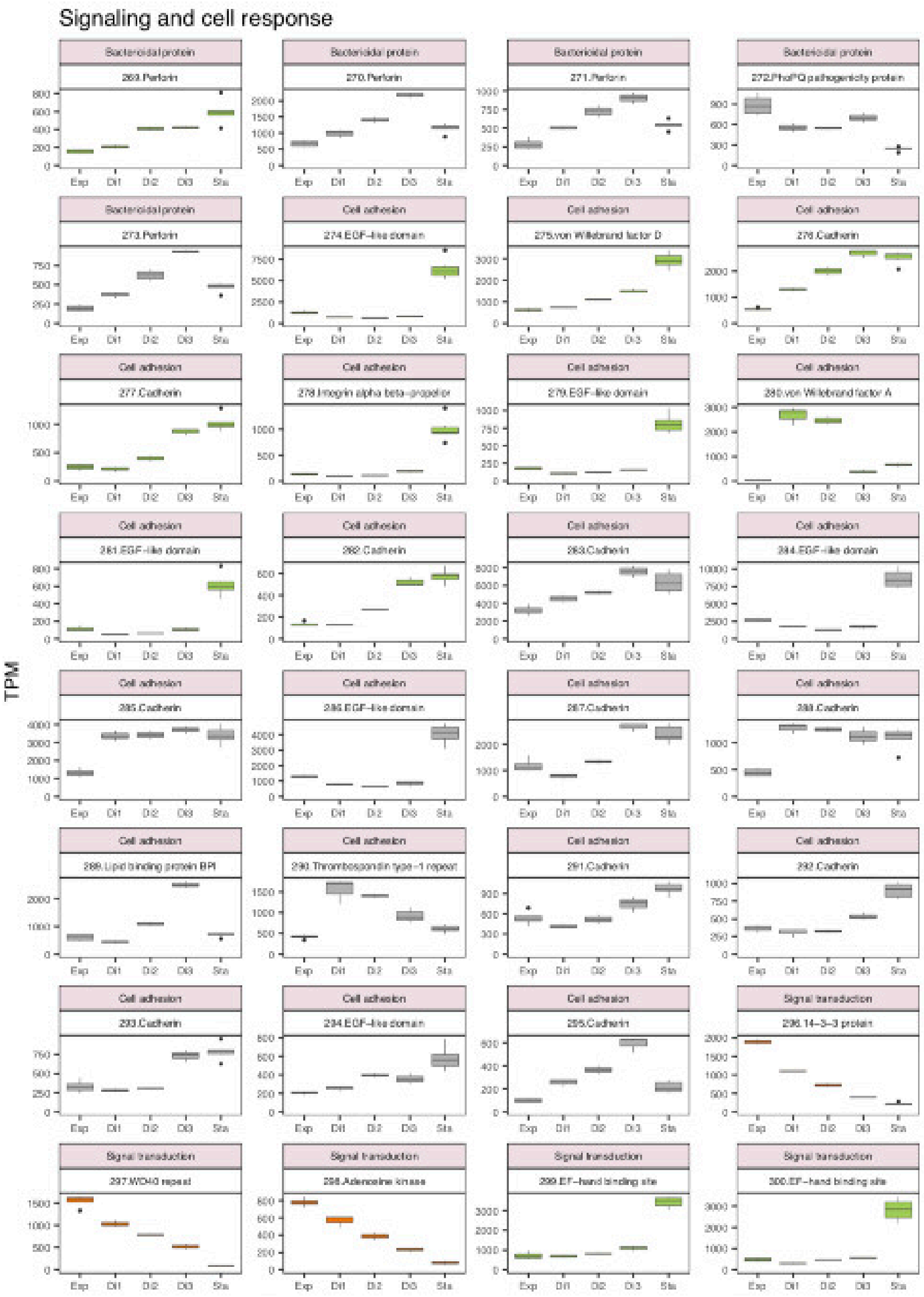

TPM
ROS homeostasis
141.Peroxidase
ROS homeostasis
142.Peroxidase
ROS homeostasis
143.Thioredoxin
ROS homeostasis
144.Thioredoxin
ROS homeostasis
145.Peroxidase
ROS homeostasis
146.Peroxidase
ROS homeostasis
147.Iron superoxide dismutase
ROS homeostasis
148.Iron superoxide dismutase
TCA cycle
149.Citrate synthase
TCA cycle
150.Citrate synthase
TCA cycle
151.Aconitase
TCA cycle
152.Isocitrate dehydrogenase
TCA cycle
153.Fumarase
TCA cycle
154.Isocitrate dehydrogenase
TCA cycle
155.Succinate dehydrogenase
Transmembrane transport
156.Mitochondrial carrier protein
Transmembrane transport
157.SamA domain
Transmembrane transport
158.Mitochondrial carrier protein
Transmembrane transport
159.Translocon Sec61/SecY
Transmembrane transport
160.Fusaric acid resistance protein
Transmembrane transport
161.Porin
Transmembrane transport
162.Porin
Transmembrane transport
163.Thiamine transporter 1
Transmembrane transport
164.Tripartite DENN domain
Transmembrane transport
165.Choline transporter
Transmembrane transport
166.Major facilitator superfamily
Exp
D1
D2
D3
Sta


Protein cellular processes
TPM
Exp
Di1
Di2
Di3
Sta
Protein folding
167.Tetratricopeptide repeat
Protein folding
168.Heat shock protein 70
Protein folding
169.Calreticulin/Calnexin
Protein folding
170.Tetratricopeptide repeat
Protein folding
171.Heat shock protein 90
Protein folding
172.Heat shock protein 70
Protein folding
173.Heat shock protein 70
Protein folding
174.Tetratricopeptide repeat
Protein folding
175.Heat shock protein 90
Protein folding
176.Heat shock protein 70
Protein folding
177.Chaperonin Cpn60
Protein folding
178.Heat shock protein 70
Protein folding
179.Heat shock protein 90
Protein folding
180.Chaperonin TCP-1
Protein folding
181.Tetratricopeptide repeat
Protein folding
182.Chaperonin TCP-1
Protein folding
183.ATPase, AAA-type
Protein folding
184.Pentatricopeptide repeat
Protein folding
185.Tetratricopeptide repeat
Protein folding
186.Cyclophilin isomerase
Protein folding
187.Prohibitin
Protein folding
188.Tetratricopeptide repeat
Protein folding
189.Prohibitin
Protein modification
190.Ubiquitin-activating enzyme E1
Protein modification
191.Ubiquitin-protein ligase
Protein modification
192.Ubiquitin-protein ligase
Protein modification
193.OPS motif
Protein modification
194.Ubiquitin specific protease
Protein modification
195.Ubiquitin-protein ligase
Protein modification
196.Ubiquitin-conjugating enzyme E2
Protein modification
197.Ubiquitin
Protein modification
198.Ubiquitin

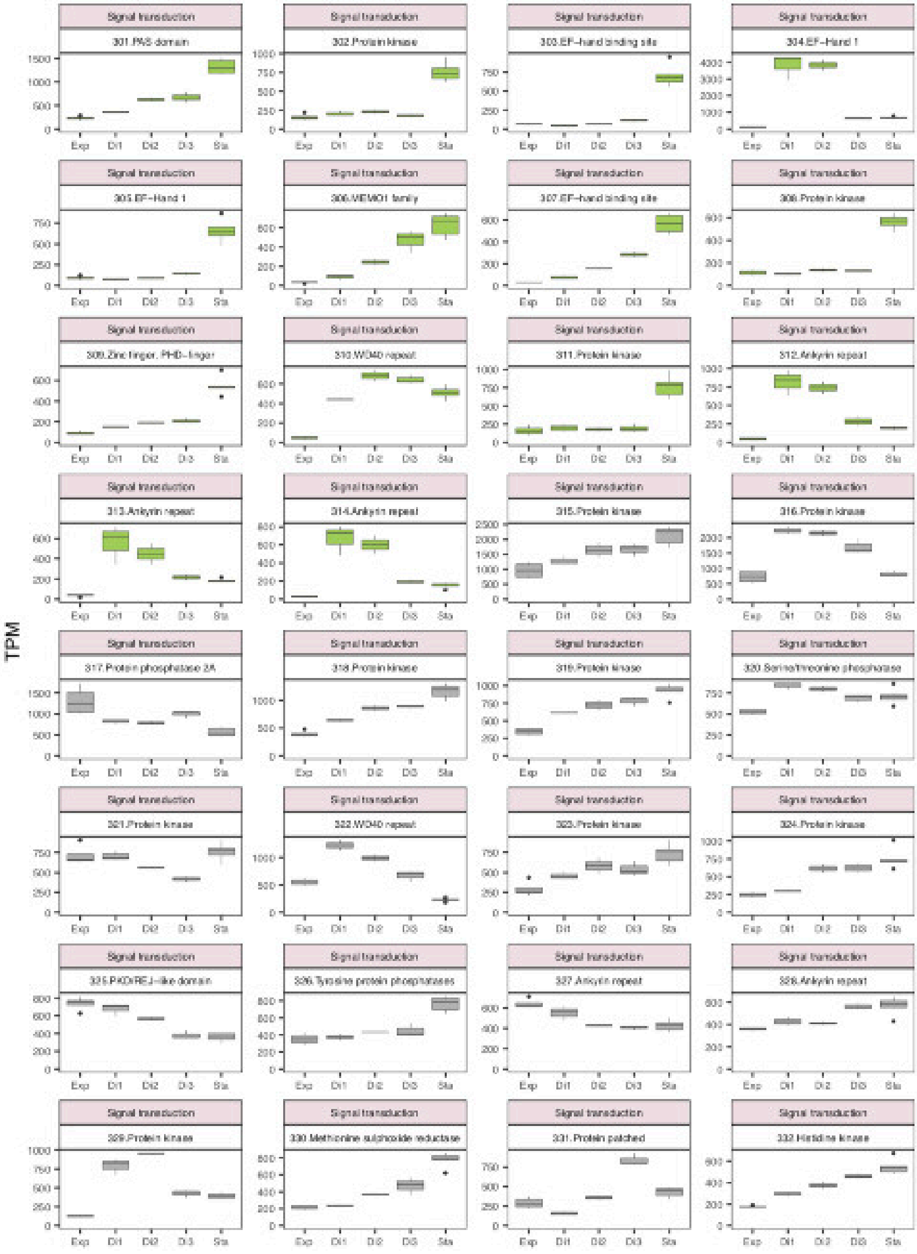

Protein cellular processes
TPM
Protein modification
199.Ubiquitin-specific proteases
Protein modification
200.Next to BRCA1
Protein modification
201.Ubiquitin
Protein modification
202.Ubiquitin
Protein modification
203.S-phase kinase-associated
Protein modification
204.Next to BRCA1
Ribosomal protein
205.Ribosomal protein L5
Ribosomal protein
206.Ribosomal protein L10P
Ribosomal protein
207.Ribosomal protein L1
Ribosomal protein
208.Ribosomal protein L11
Ribosomal protein
209.Ribosomal protein S27e
Ribosomal protein
210.Ribosomal protein L26/L24
Ribosomal protein
211.Ribosomal protein S9
Ribosomal protein
212.Ribosomal protein L30e
Ribosomal protein
213.Ribosomal protein L6
Ribosomal protein
214.Ribosomal protein L15e
Ribosomal protein
215.Ribosomal protein S11
Ribosomal protein
216.Ribosomal protein L6E
Ribosomal protein
217.Ribosomal protein S4/S9
Ribosomal protein
218.Ribosomal protein L28e
Ribosomal protein
219.Ribosomal protein L3
Ribosomal protein
220.Ribosomal protein S4e
Ribosomal protein
221.Ribosomal protein L7
Ribosomal protein
222.Ribosomal protein L18e
Ribosomal protein
223.Ribosomal protein L4/L1e
Ribosomal protein
224.Ribosomal protein S8
Ribosomal protein
225.Ribosomal protein S10
Ribosomal protein
226.Ribosomal protein S7e
Ribosomal protein
227.Ribosomal protein L18a
Ribosomal protein
228.Ribosomal protein L10e
Ribosomal protein
229.Ribosomal protein L13
Ribosomal protein
230.Ribosomal protein L23/L25
Exp
Di1
Di2
Di3
Sta

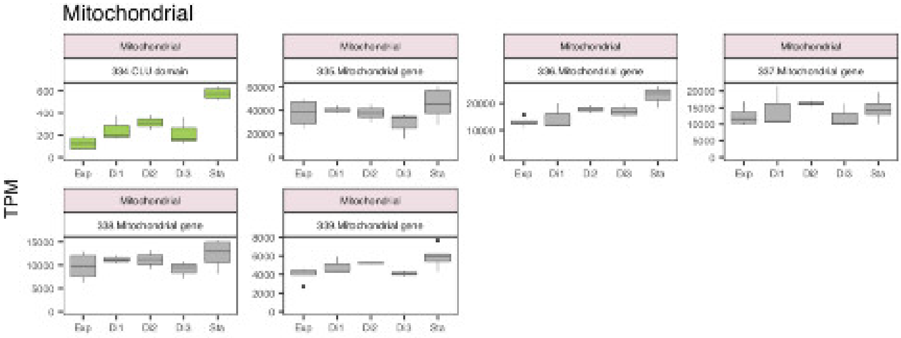

Protein cellular processes
TPM
Ribosomal protein
231.Ribosomal protein L27e
Ribosomal protein
232.Ribosomal protein L2
Ribosomal protein
233.Ribosomal protein S2
Ribosomal protein
234.Ribosomal protein L31e
Ribosomal protein
235.Ribosomal protein L6
Ribosomal protein
236.Ribosomal protein S10A/S15e
Ribosomal protein
237.Ribosomal protein L7A/L8
Ribosomal protein
238.Ribosomal protein L24e
Ribosomal protein
239.Ribosomal protein L19
Ribosomal protein
240.Ribosomal protein L21e
Ribosomal protein
241.Ribosomal protein L14P
Ribosomal protein
242.Ribosomal protein S3
Ribosomal protein
243.Ribosomal protein S6e
Ribosomal protein
244.Ribosomal protein L35
Ribosomal protein
245.Ribosomal protein S11
Ribosomal protein
246.Ribosomal protein L29
Ribosomal protein
247.Ribosomal protein S7
Ribosomal protein
248.Ribosomal protein L28e
Ribosomal protein
249.Ribosomal protein S3Ae
Ribosomal protein
250.Ribosomal protein L22/L17
Ribosomal protein
251.Ribosomal protein L32e
Ribosomal protein
252.Ribosomal protein L44e
Ribosomal protein
253.Ribosomal protein L37e
Ribosomal protein
254.Ribosomal protein L35A
Ribosomal protein
255.Ribosomal protein S8
Ribosomal protein
256.Ribosomal protein S13
Ribosomal protein
257.Ribosomal protein L13e
Ribosomal protein
258.Ribosomal protein S21e
Ribosomal protein
259.Ribosomal protein S15
Ribosomal protein
260.Ribosomal protein L34Ae
Ribosomal protein
261.Ribosomal protein S23
Translation machinery
262.Elongation factor EFT
Exp
Di1
Di2
Di3
Sta

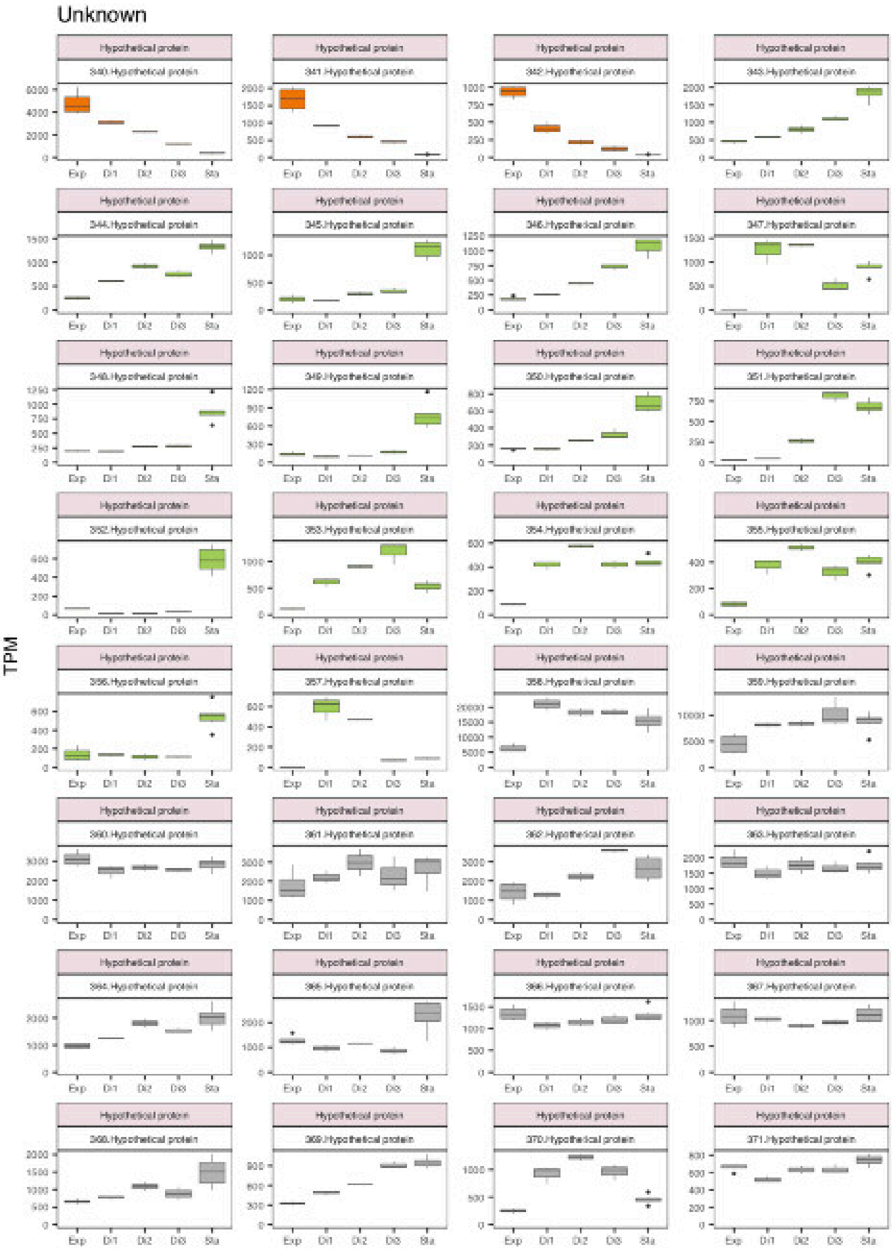

Protein cellular processes
TPM
Translation machinery
263.Leucyl−tRNA synthetase
0
500
1000
Translation machinery
264.Elongation factor EFT
0
200
400
600
Translation machinery
265.Elongation factor
0
2500
5000
7500
10000
12500
Translation machinery
266.Elongation factor EF1A
0
2000
4000
6000
8000
Translation machinery
267.Elongation factor EF1A
0
2000
4000
6000
Translation machinery
268.DnaJ domain
0
200
400
600
Exp
D1
D2
D3
Sta

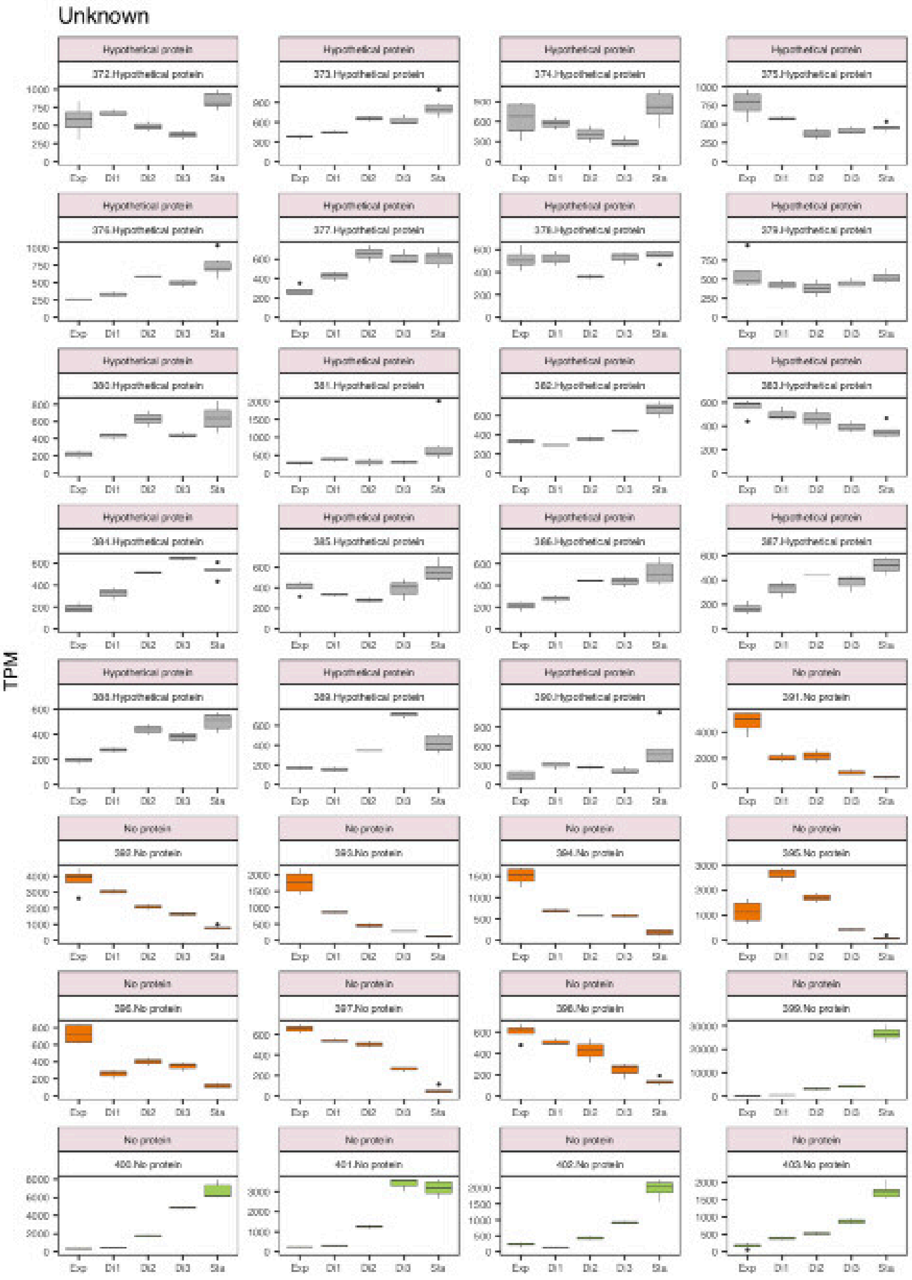

Signaling and cell response
TPM
Exp
Di1
Di2
Di3
Sta
Bactericidal protein
269.Perforin
Bactericidal protein
270.Perforin
Bactericidal protein
271.Perforin
Bactericidal protein
272.PhoPQ pathogenicity protein
Bactericidal protein
273.Perforin
Cell adhesion
274.EGF-like domain
Cell adhesion
275.von Willebrand factor D
Cell adhesion
276.Cadherin
Cell adhesion
277.Cadherin
Cell adhesion
278.Integrin alpha beta-propellor
Cell adhesion
279.EGF-like domain
Cell adhesion
280.von Willebrand factor A
Cell adhesion
281.EGF-like domain
Cell adhesion
282.Cadherin
Cell adhesion
283.Cadherin
Cell adhesion
284.EGF-like domain
Cell adhesion
285.Cadherin
Cell adhesion
286.EGF-like domain
Cell adhesion
287.Cadherin
Cell adhesion
288.Cadherin
Cell adhesion
289.Lipid binding protein BPI
Cell adhesion
290.Thrombospondin type-1 repeat
Cell adhesion
291.Cadherin
Cell adhesion
292.Cadherin
Cell adhesion
293.Cadherin
Cell adhesion
294.EGF-like domain
Cell adhesion
295.Cadherin
Signal transduction
296.14-3-3 protein
Signal transduction
297.WD40 repeat
Signal transduction
298.Adenosine kinase
Signal transduction
299.EF-hand binding site
Signal transduction
300.EF-hand binding site

Signaling and cell response




Mitochondrial
TPM
Mitochondrial
334 CLU domain
600
400
200
0
Mitochondrial
335 Mitochondrial gene
60000
40000
20000
0
Mitochondrial
336 Mitochondrial gene
20000
10000
0
Mitochondrial
337 Mitochondrial gene
20000
15000
10000
5000
0
Exp
Di1
Di2
Di3
Sta
Mitochondrial
338 Mitochondrial gene
15000
10000
5000
0
Mitochondrial
339 Mitochondrial gene
8000
6000
4000
2000
0
Exp
Di1
Di2
Di3
Sta


Unknown
TPM
Hypothetical protein
340.Hypothetical protein
341.Hypothetical protein
342.Hypothetical protein
343.Hypothetical protein
344.Hypothetical protein
345.Hypothetical protein
346.Hypothetical protein
347.Hypothetical protein
348.Hypothetical protein
349.Hypothetical protein
350.Hypothetical protein
351.Hypothetical protein
352.Hypothetical protein
353.Hypothetical protein
354.Hypothetical protein
355.Hypothetical protein
356.Hypothetical protein
357.Hypothetical protein
358.Hypothetical protein
359.Hypothetical protein
360.Hypothetical protein
361.Hypothetical protein
362.Hypothetical protein
363.Hypothetical protein
364.Hypothetical protein
365.Hypothetical protein
366.Hypothetical protein
367.Hypothetical protein
368.Hypothetical protein
369.Hypothetical protein
370.Hypothetical protein
371.Hypothetical protein
Exp
Di1
Di2
Di3
Sta


Unknown
TPM
Hypothetical protein
372.Hypothetical protein
373.Hypothetical protein
374.Hypothetical protein
375.Hypothetical protein
376.Hypothetical protein
377.Hypothetical protein
378.Hypothetical protein
379.Hypothetical protein
380.Hypothetical protein
381.Hypothetical protein
382.Hypothetical protein
383.Hypothetical protein
384.Hypothetical protein
385.Hypothetical protein
386.Hypothetical protein
387.Hypothetical protein
388.Hypothetical protein
389.Hypothetical protein
390.Hypothetical protein
No protein
391.No protein
392.No protein
393.No protein
394.No protein
395.No protein
396.No protein
397.No protein
398.No protein
399.No protein
400.No protein
401.No protein
402.No protein
403.No protein
Exp
Di1
Di2
Di3
Sta

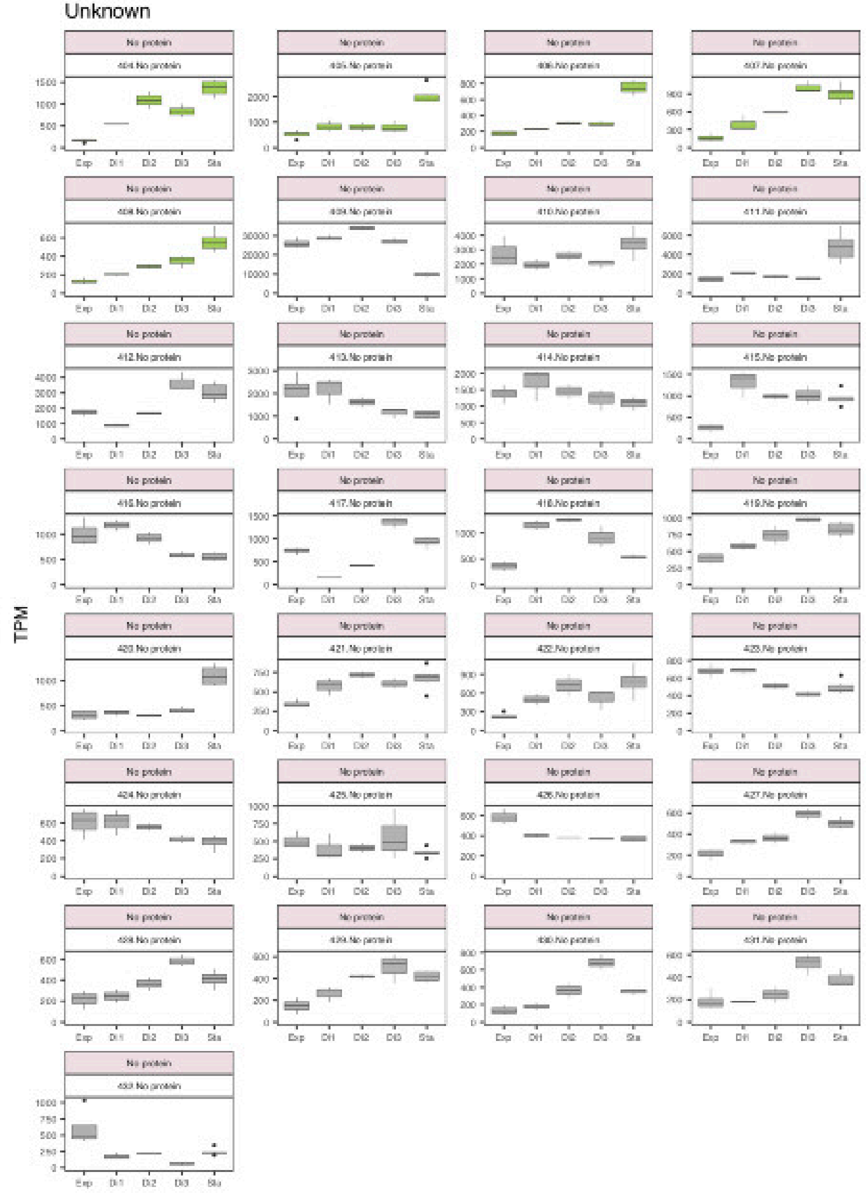
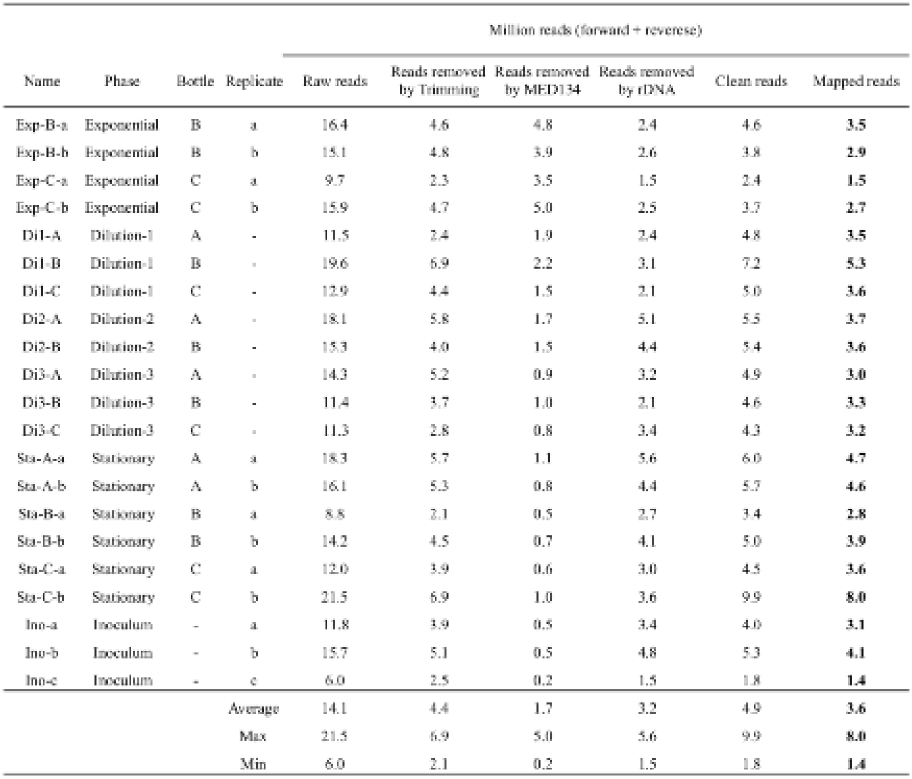
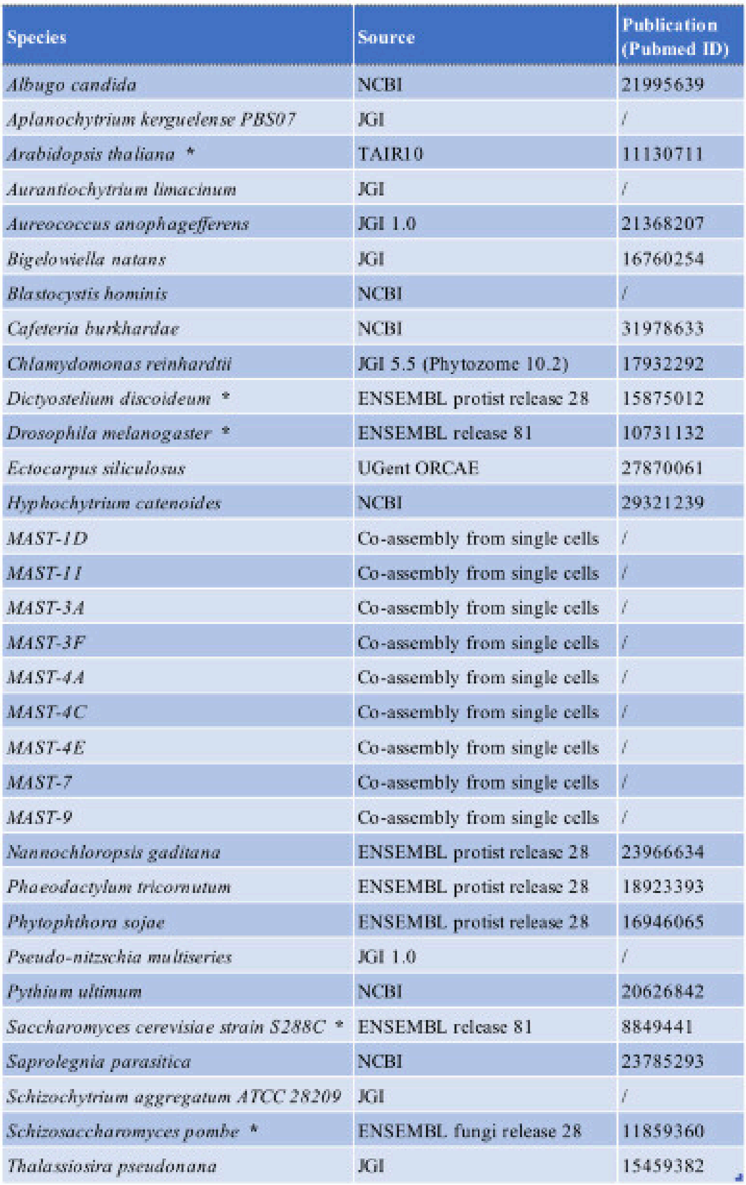


Figure S5. Box plots displaying expression changes in the five states of the 432 highly expressed genes, ordered based on their general process and cellular function and then by differential expression between Exponential and Stationary. Orange: genes upregulated in Exponential; Green; genes upregulated in Stationary; Grey: genes equally expressed.



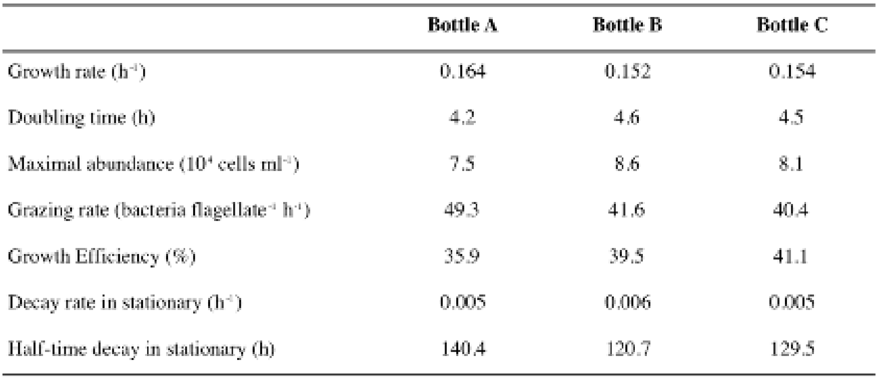
				Million reads (forward + reverese)					
Name	Phase	Bottle	Replicate	Raw reads	Reads removed by Trimming	Reads removed by MED134	Reads removed by rDNA	Clean reads	Mapped reads
Exp-B-a	Exponential	B	a	16.4	4.6	4.8	2.4	4.6	**3.5**
Exp-B-b	Exponential	B	b	15.1	4.8	3.9	2.6	3.8	**2.9**
Exp-C-a	Exponential	C	a	9.7	2.3	3.5	1.5	2.4	**1.5**
Exp-C-b	Exponential	C	b	15.9	4.7	5.0	2.5	3.7	**2.7**
Di1-A	Dilution-1	A	-	11.5	2.4	1.9	2.4	4.8	**3.5**
Di1-B	Dilution-1	B	-	19.6	6.9	2.2	3.1	7.2	**5.3**
Di1-C	Dilution-1	C	-	12.9	4.4	1.5	2.1	5.0	**3.6**
Di2-A	Dilution-2	A	-	18.1	5.8	1.7	5.1	5.5	**3.7**
Di2-B	Dilution-2	B	-	15.3	4.0	1.5	4.4	5.4	**3.6**
Di3-A	Dilution-3	A	-	14.3	5.2	0.9	3.2	4.9	**3.0**
Di3-B	Dilution-3	B	-	11.4	3.7	1.0	2.1	4.6	**3.3**
Di3-C	Dilution-3	C	-	11.3	2.8	0.8	3.4	4.3	**3.2**
Sta-A-a	Stationary	A	a	18.3	5.7	1.1	5.6	6.0	**4.7**
Sta-A-b	Stationary	A	b	16.1	5.3	0.8	4.4	5.7	**4.6**
Sta-B-a	Stationary	B	a	8.8	2.1	0.5	2.7	3.4	**2.8**
Sta-B-b	Stationary	B	b	14.2	4.5	0.7	4.1	5.0	**3.9**
Sta-C-a	Stationary	C	a	12.0	3.9	0.6	3.0	4.5	**3.6**
Sta-C-b	Stationary	C	b	21.5	6.9	1.0	3.6	9.9	**8.0**
Ino-a	Inoculum	-	a	11.8	3.9	0.5	3.4	4.0	**3.1**
Ino-b	Inoculum	-	b	15.7	5.1	0.5	4.8	5.3	**4.1**
Ino-c	Inoculum	-	c	6.0	2.5	0.2	1.5	1.8	**1.4**
			Average	14.1	4.4	1.7	3.2	4.9	**3.6**
			Max	21.5	6.9	5.0	5.6	9.9	**8.0**
			Min	6.0	2.1	0.2	1.5	1.8	**1.4**

Table S1. Naming of samples, indicating the phase, the biological and technical replicate, and number of reads at different steps: raw reads, reads removed (after quality control or because they affiliate with the bacterial genome or the rDNA operon), clean reads, and reads finally mapping to the *C. burkhardae de novo* transcriptome



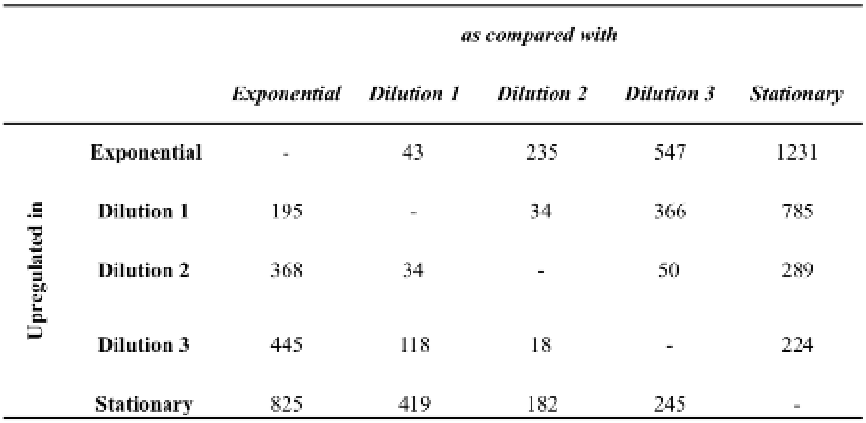
Species	Source	Publication (Pubmed ID)
Albugo candida	NCBI	21995639
Aplanochytrium kerguelense PBS07	JGI	/
Arabidopsis thaliana *	TAIR10	11130711
Aurantiochytrium limacinum	JGI	/
Aureococcus anophagefferens	JGI 1.0	21368207
Bigelowiella natans	JGI	16760254
Blastocystis hominis	NCBI	/
Cafeteria burkhardae	NCBI	31978633
Chlamydomonas reinhardtii	JGI 5.5 (Phytozome 10.2)	17932292
Dictyostelium discoideum *	ENSEMBL protist release 28	15875012
Drosophila melanogaster *	ENSEMBL release 81	10731132
Ectocarpus siliculosus	UGent ORCAE	27870061
Hyphochytrium catenoides	NCBI	29321239
MAST-1D	Co-assembly from single cells	/
MAST-11	Co-assembly from single cells	/
MAST-3A	Co-assembly from single cells	/
MAST-3F	Co-assembly from single cells	/
MAST-4A	Co-assembly from single cells	/
MAST-4C	Co-assembly from single cells	/
MAST-4E	Co-assembly from single cells	/
MAST-7	Co-assembly from single cells	/
MAST-9	Co-assembly from single cells	/
Nannochloropsis gaditana	ENSEMBL protist release 28	23966634
Phaeodactylum tricornutum	ENSEMBL protist release 28	18923393
Phytophthora sojae	ENSEMBL protist release 28	16946065
Pseudo-nitzschia multiseries	JGI 1.0	/
Pythium ultimum	NCBI	20626842
Saccharomyces cerevisiae strain S288C *	ENSEMBL release 81	8849441
Saprolegnia parasitica	NCBI	23785293
Schizochytrium aggregatum ATCC 28209	JGI	/
Schizosaccharomyces pombe *	ENSEMBL fungi release 28	11859360
Thalassiosira pseudonana	JGI	15459382

Table S2 Species used to build the stramenopile-oriented PLAZA genome database. For species marked by an asterisk, reference GO annotation was retrieved from the GO website.



	Bottle A	Bottle B	Bottle C
Growth rate (h^{-1})	0.164	0.152	0.154
Doubling time (h)	4.2	4.6	4.5
Maximal abundance (10^4 cells ml^{-1})	7.5	8.6	8.1
Grazing rate (bacteria $flagellate^{-1}$ h^{-1})	49.3	41.6	40.4
Growth Efficiency (%)	35.9	39.5	41.1
Decay rate in stationary (h^{-1})	0.005	0.006	0.005
Half-time decay in stationary (h)	140.4	120.7	129.5

Table S3. Growth properties of *C. burkhardae* growing on *Dokdonia* MED134 in the three batch cultures established.



		as compared with				
		Exponential	*Dilution 1*	*Dilution 2*	*Dilution 3*	*Stationary*
Upregulated in	**Exponential**	-	43	235	547	1231
	Dilution 1	195	-	34	366	785
	Dilution 2	368	34	-	50	289
	Dilution 3	445	118	18	-	224
	Stationary	825	419	182	245	-

Table S4. Number of differentially expressed genes in pairwise comparisons among the five phases. Each line indicates the number of upregulated genes of the phase labeled per line (against the phase labeled per columns).

REFERENCES

1. Sherr BF, Sherr EB, Caron D, Vaulot D, Worden A. Oceanic protists. Oceanography. 2007;20:130-34.

2. Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ. Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. Science 2015;347:1257594.

3. Jürgens, K., and R. Massana. Protistan Grazing on Marine Bacterioplankton, p. 383–441. In D.L. Kirchman [ed.], Microbial Ecology of the Oceans 2008. John Wiley & Sons, Inc.

4. Pernthaler J. Predation on prokaryotes in the water column and its ecological implications. Nat Rev Microbiol. 2005;3:537-46.

5. Boenigk J, Arndt H. Bacterivory by heterotrophic flagellates: community structure and feeding strategies. Ant van Leeuw. 2002;81:465-80.

6. Vørs N, Buck KR, Chavez FP, Eikrem W, Hansen LE, Østergaard JB, et al. Nanoplankton of the equatorial Pacific with emphasis on the heterotrophic protists. Deep-Sea Res. II 1995;42:585-602.

7. Massana R, Guillou L, Díez B, Pedrós-Alió C. Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean. Appl Environ Microbiol. 2002;68:4554-58.

8. Rodríguez-Martínez R, Rocap G, Logares R, Romac S, Massana R. Low evolutionary diversification in a widespread and abundant uncultured protist (MAST-4). Mol Biol Evol. 2012;29:1393-406.

9. del Campo J, Balagué V, Forn I, Lekunberri I, Massana R. Culturing bias in marine heterotrophic flagellates analyzed through seawater enrichment incubations. Microb Ecol. 2013;66:489-99.

10. Caron DA, Alexander H, Allen AE, Archibald JM, Armbrust EV, Bachy C, et al. Probing the evolution, ecology and physiology of marine protists using transcriptomics. Nature Rev Microb. 2017;15:6-20.

11. Yutin N, Wolf MY, Wolf YI, Koonin EV. The origins of phagocytosis and eukaryogenesis. Biol Direct. 2009;4:9.

12. Keeling PJ. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. Annu Rev Plant Biol. 2013;64:583-607.

13. Martin WF, Tielens AGM, Mentel M, Garg SG, Gould SB. The physiology of phagocytosis in the context of mitochondrial origin. Microb Mol Biol Rev. 2017;81:e00008-17.

14. Rosales C, Uribe-Querol E. Phagocytosis: A fundamental process in immunity. BioMed Res Int. 2017:9042851.

15. Gotthardt D, Warnatz HJ, Henschel O, Brückert F, Schleicher M, Soldati T. (2002). High-resolution dissection of phagosome maturation reveals distinct membrane trafficking phases. Mol Biol Cell. 2002;13:3508–20.

16. Niedergang F, Grinstein S. How to build a phagosome: new concepts for an old process. Curr Opin Cell Biol. 2018;50:57–63.

17. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 2019;47:D590–D595.

18. Bozzaro S, Bucci C, Steinert M. Phagocytosis and host-pathogen interactions in *Dictyostelium* with a look at macrophages. Int Rev Cell Mol Biol. 2008;271:253-300.

19. Jacobs ME, DeSouza LV, Samaranayake H, Pearlman RE, Siu KWM, Klobutcher LA. The *Tetrahymena thermophila* phagosome proteome. Eukaryot Cell. 2006;5:1990-2000.

20. Boulais J, Trost M, Landry CR, Dieckmann R, Levy ED, Soldati T, et al. Molecular characterization of the evolution of phagosomes. Mol Syst Biol. 2010;6:423.

21. Lie AAY, Liu Z, Terrado R, Tatters AO, Heidelberg KB, Caron DA. Effect of light and prey availability on gene expression of the mixotrophic chrysophyte *Ochromonas* sp. BMC Genomics. 2017;18:163.

22. Rubin ET, Cheng S, Montalbano AL, Menden-Deuen S, Rynearson TA. Transcriptomic response to feeding and starvation in a herbivorous dinoflagellate. Front Mar Sci. 2019;6:246.

23. Fenchel T, Patterson DJ. *Cafeteria roenbergensis* nov. gen., nov. sp., a heterotrophic microflagellate from marine plankton. Mar Microb Food Webs. 1988;3:9-19.

24. Schoenle A, Hohlfeld M, Rosse M, Filz P, Wylezich C, Nitsche F, et al. Global comparison of bicosoecid *Cafeteria*-like flagellates from the deep ocean and surface waters, with reorganization of the family Cafeteriaceae. Europ J Protistol. 2020;73:125665.

25. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral- Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. 2014;12:e1001889.

26. Hackl T, Martin R, Barenhoff K, Duponchel S, Heider D, Fischer MG. Four high-quality draft genome assemblies of the marine heterotrophic nanoflagellate *Cafeteria roenbergensis*. Sci Data. 2020;7:29.

27. Anderson R, Kjelleberg S, Mcdougald D, Jürgens K. Species-specific patterns in the vulnerability of carbon-starved bacteria to protist grazing. Aquat Microb Ecol. 2011;64:105-16.

28. de Corte D, Paredes G, Yokokawa T, Sintes E, Herndl GJ. Differential response of *Cafeteria roenbergensis* to different bacterial and archaeal characteristics. Microb Ecol. 2019:78:1-5.

29. Massana R, del Campo J, Dinter C, Sommaruga R. Crash of a population of the marine heterotrophic flagellate *Cafeteria roenbergensis* by viral infection. Environ Microbiol. 2007;9:2660-69.

30. Logares R, Deutschmann IM, Junger PC, Giner CR, Krabberød AK, Schmidt TSB, et al. Disentangling the mechanisms shaping the surface ocean microbiota. Microbiome. 2020;8:55.

31. Giner CR, Pernice MC, Balagué V, Duarte CM, Gasol JM, Logares R, et al. Marked changes in diversity and relative activity of picoeukaryotes with depth in the world ocean. ISME J. 2020;14:437-449.

32. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Meth. 2016;13:581-83.

33. Obiol A, Giner CR, Sánchez P, Duarte CM, Acinas SG, Massana R. A metagenomic assessment of microbial eukaryotic diversity in the global ocean. Mol Ecol Res. 2020;20:718–731.

34. Altschul SF, W Gish, W Miller, EW Myers, DJ Lipman. Basic local alignment search tool. J Mol Biol. 1990;215:403-10.

35. Mangot J-F, Forn I, Obiol A, Massana R. Constant abundances of ubiquitous uncultured protists in the open sea assessed by automated microscopy. Environ Microbiol. 2018;20:3876-89.

36. Lekunberri I, Gasol JM, Acinas SG, Gómez-Consarnau L, Crespo BG, Casamayor EO, et al. The phylogenetic and ecological context of cultured and whole genome-sequenced planktonic bacteria from the coastal NW Mediterranean Sea. Syst Appl Microbiol. 2014;37:216-28.

37. Porter KG, Feig YS. The use of DAPI for identifying aquatic microfloral. Limnol Oceanogr. 1980;25:943-48.

38. González JM, Suttle CA. Grazing by marine nanoflagellates on viruses and virus-sized particles: Ingestion and digestion. Mar Ecol Prog Ser. 1993;94:1-10.

39. Frost BW. Effects of size and concentration of food particles on the feeding behavior of the marine planktonic copepod *Calanus pacificus*. Limnol Oceanogr. 1972;17:805-15.

40. Heinbokel JF. Studies on the functional role of tintinnids in the Southern California Bight. I. Grazing and growth rates in laboratory cultures. Mar Biol. 1978;47:177–189.

41. Menden-Deuer S, Lessard EJ. 2000. Carbon to volume relationships for dinoflagellates, diatoms, and other protist plankton. Limnol Oceanogr. 2000;45:569–79.

42. Picelli S, Faridani OR, Björklund Å, Winberg G, Sagasser S, Sandberg R. Full-length RNA-seq from single cells using Smart-seq2. Nature Protocols. 2014;9:171-81.

43. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-20.

44. Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Meth. 2012;9:357-59.

45. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;8:1494-512.

46. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge . Nucleic Acids Res. 2019;47:D506–D515.

47. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012;40:D290-D301.

48. Powell S, Szklarczyk D, Trachana K, Roth A, Kuhn M, Muller J, et al. eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. Nucleic Acids Res. 2012;40:D284-D289.

49. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.

50. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. Mol Biol Evol. 2018;35:543–48.

51. van Bel M, Proost S, van Neste C, Deforce D, van de Peer Y, Vandepoele K. TRAPID, an efficient online tool for the functional and comparative analysis of de novo RNA-Seq transcriptomes. Genome Biology. 2013;14:R134.

52. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res. 2017;45:D190–D199.

53. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nature Meth. 2015;12:59–60.

54. Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, et al. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Res. 2018;46:D1190–D1196.

55. McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. 2012;40:4288-97.

56. Zinger L, Gobet A, Pommier T. Two decades of describing the unseen majority of aquatic microbial diversity. Mol Ecol. 2012;21:1878-96.

57. Pernice MC, Forn I, Gomes A, Lara E, Alonso-Sáez L, Arrieta JM, et al. Global abundance of planktonic heterotrophic protists in the deep ocean. ISME J. 2015;9:782-92.

58. Eccleston-Parry JD, Leadbeater BSC. A comparison of the growth-kinetics of 6 marine heterotrophic nanoflagellates fed with one bacterial species. Mar Ecol Prog Ser. 1994;105:167-77.

59. Arndt H, Hausmann K, Wolf M. Deep-sea heterotrophic nanoflagellates of the Eastern Mediterranean Sea: qualitative and quantitative aspects of their pelagic and benthic occurrence. Mar Ecol Prog Ser. 2003;256:45-56.

60. Azam F, Long RA. Sea snow microcosms. Nature. 2001;414:495-98.

61. Fenchel T. Ecology of protozoa: The biology of free-living phagotrophic protists. 1987. Science Tech Publishers, Madison and Springer-Verlag.

62. Mestre M, Ruiz-González C, Logares R, Duarte CM, Gasol JM, Sala MM. Sinking particles promote vertical connectivity in the ocean microbiome. Proc Natl Acad Sci USA. 2018;115:E6799-E6807.

63. Beisser D, Graupner N, Bock C, Wodniok S, Grosmann L, Vos M, et al. Comprehensive transcriptome analysis provides new insights into nutritional strategies and phylogenetic relationships of chrysophytes. PeerJ 2017;5:e2832.

64. Liu Z, Campbell V, Heidelberg KB, Caron DA. Gene expression characterizes different nutritional strategies among three mixotrophic protists. FEMS Microb Ecol. 2016;92:fiw106.

65. Garba L, Ali MSM, Oslan SN, Abdul Rahman RNZRB. Review on fatty acid desaturases and their roles in temperature acclimatisation. J Appl Sci. 2017;17:282-95.

66. Cheng W, Lin M, Qiu M, Kong L, Xu Y, Li Y, et al. Chitin synthase is involved in vegetative growth, asexual reproduction and pathogenesis of *Phytophthora capsici* and *Phytophthora.* Environ Microbiol. 2019;21:4537-47.

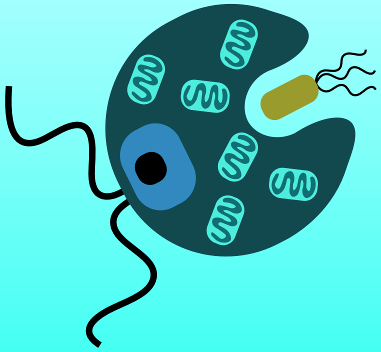
67. Rawlings ND, Barrett AJ. Families of cysteine peptidases. Methods Enzymol. 1994;244:461-86.

68. Rawlings ND, Barrett AJ, Finn R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 2015;44:D343-D350.

69. Baltscheffsky M, Schultz A, Baltscheffsky H. H+-proton-pumping inorganic pyrophosphatase: a tightly membrane-bound family. FEBS Lett. 1999;457:527–33.

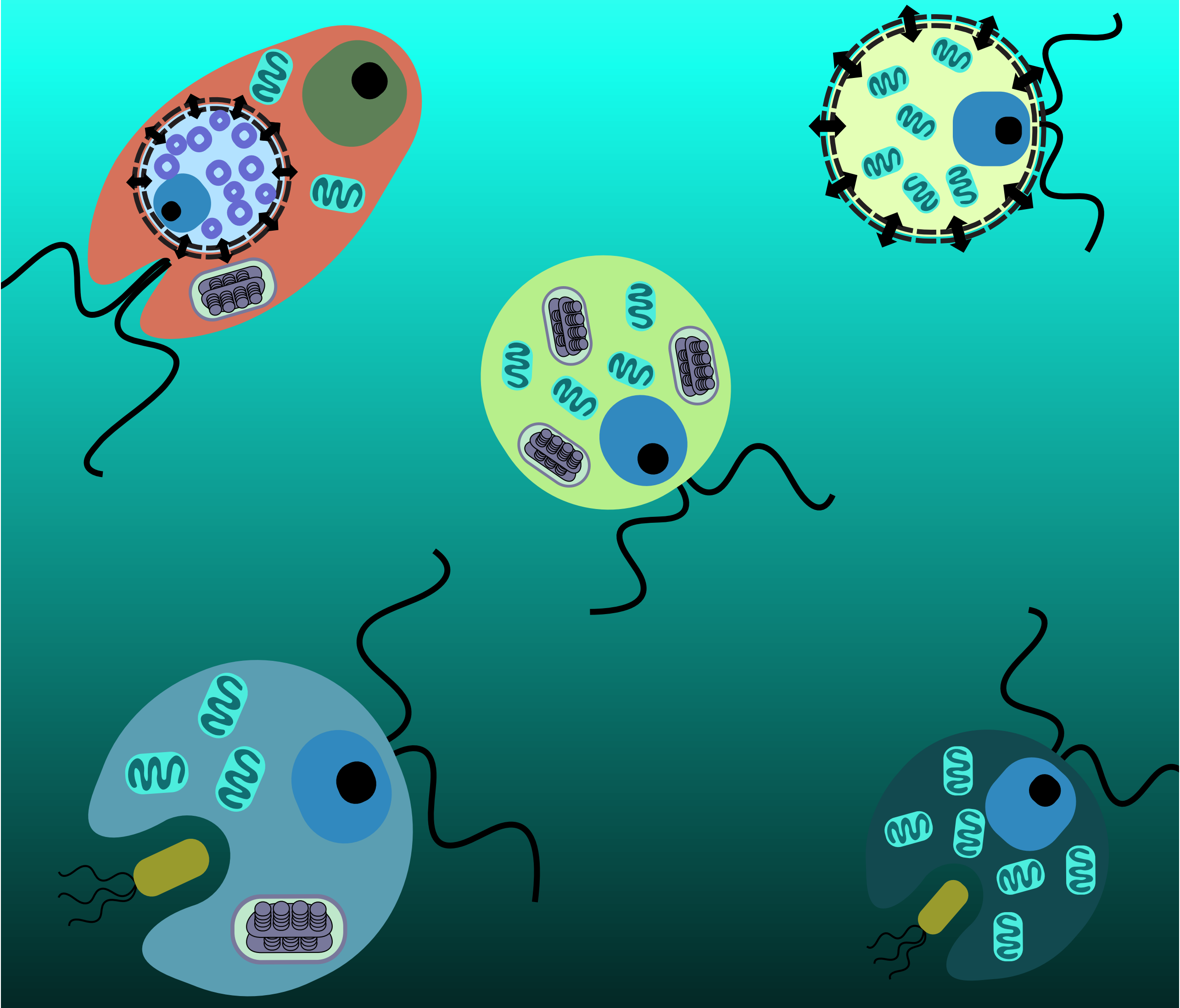
70. Labarre A, Obiol A, Wilken S, Forn I, Massana R. Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates. Limnol Oceanogr. 2020;65:S149-S160.

71. Minakami R, Sumimotoa H. Phagocytosis-coupled activation of the superoxide-producing phagocyte oxidase, a member of the NADPH oxidase (nox) family. Int J Hematol. 2006;84,193-98.



Chapter 5

Synthesis of results and General Discussion



SYNTHESIS OF RESULTS AND GENERAL DISCUSSION

This thesis was initiated within the context of the broad revolution of culture-independent genomic techniques that make accessible the study of neglected uncultured lineages. Thus, new insights in marine eukaryotic unicellular organisms became possible towards a better understanding of microbes roles and interactions, enlightening the functioning of the ocean. With the objective to increase our knowledge on unappreciated microbes and their ecological implications, this work focused on the study of heterotrophic flagellates, the most important grazers of bacteria in aquatic ecosystems.

Collectively, the present thesis explores two setups covering different DNA and RNA-based methodologies that proved to be successful in the investigation of unculturable MArine STramenopile (MASTs) lineages. First, MAST cells were isolated with a single cell sorting technique from samples taken during the field campaign TARA Oceans, and also from the Blanes Bay Microbial Observatory in the Mediterranean Sea (Chapter 2). We aimed to provide reference genomes of MArine Stramenopiles, for which previous genomic information is limited and/or uncharacterized. These genomes were used in downstream comparative genomic analyses, whereby assembled genomes of MASTs allowed us to get access to their gene repertoire and to provide evidence of their heterotrophic behavior in the marine environment. In the next chapter (Chapter 3) an experimental approach was developed to assess the activity of uncultured MASTs directly in their natural environment by tracing the expression of specific genes involved in phagocytosis. We followed the growth of heterotrophic versus phototrophic organisms in a controlled incubation and collected samples for metatranscriptomics during the active grazing phase by bacterivory. We emphasized the use of reference genomes (obtained in Chapter 2) in tandem with environmental metatranscriptomics information to discern the gene expression profile of MASTs within their microbial community. Hence, this work produced crucial evidence of the genes used during phagocytosis by several abundant members of the microbial assemblage. As supportive data, we validated previous assumptions employing a cultured heterotrophic flagellate model species, *Cafeteria burkhardae* (Chapter 4). Towards a direct perception of bacterivory in a marine eukaryote, we designed a controlled experimental work where *Cafeteria burkhardae* fed on a specific diet. The genes that were upregulated whilst grazing on bacteria were revealed by transcriptomics. Altogether, the results obtained in the three chapters give an overview of the ecology of heterotrophic flagellates that are part of the community composition in oceans. In addition, this information provides new evidence for the molecular mechanism of phagocytosis in protists.

Each chapter of this thesis has been submitted as an article for publication, therefore each one has its own discussion section. Nonetheless, in the following section the main results of each chapter will be covered in a general discussion, focusing on some common aspects.

UNRAVELING THE UNCULTURED MARINE STRAMENOPILES

By the end of the twentieth century, it became clear that all organisms live in close metabolic and functional relation with other organisms. This led to the growing need to expand our knowledge on unicellular protists, in order to shed some light on their spectacular capacities. A challenge with microbes has been that over 99% of microbial species on Earth cannot be cultured and expanded in the lab (Lasken and McLean, 2014). New approaches were then necessary to resolve microbial ecology. The last decade has been as rich in the discovery of new protist organisms by DNA-based analyses, however, most of them are still generally poorly described. Although DNA-based taxonomies of protists have been used to describe the protistan diversity across many ecosystems (de Vargas et al. 2015; Massana et al. 2015; Mahé et al. 2017), only a minority of them have been studied deeply enough to reveal their specific biological functions in the environment, whether they are heterotrophs (Jürgens and Massana, 2008) or phototrophs (Stoecker et al. 2009), or simply discovering new ecological potentials (Montagnes, 2012; Suzuki and Not, 2015; de Vargas et al. 2015). Indeed, the cellular and functional characterization has been arduous work as this novel diversity is mostly represented by organisms of very small size (pico- and nano-), coupled with the difficulties of growing them *in-vitro*. As a consequence, many novel protist lineages have been neglected (Caron et al. 2009). Therefore, in this work we have attempted to remedy this gap, similarly to other initiatives (Stern et al. 2018), and favoring the study of picoeukaryotes and in particular the heterotrophic flagellates.

Single cell isolation and sequencing techniques have shown tremendous growth over the last years and have impacted many diverse areas of biological research (Wang and Navin, 2015). They proved to be of great promise in the access of morphologically indistinct miocrobes, and have been widely used to study microbial metabolic potential (Yoon et al. 2011; Benites et al. 2019; Ku and Sebé-Pedrós, 2019). Here we provide 15 draft genomes from novel and previously uncharacterized marine stramenopiles (MASTs); each one co-assembled from several single amplified genomes (SAGs) of cells from the same population (based on identical 18S rDNA and other genomic features). Co-assembly increases the completeness of genomes (by allowing the random coverage within each cell to complement each other), and shows the recovery of enough conserved genes to characterize the function of the MASTs. It also allowed the access to the genomes of some species, such as MAST-1D, MAST-7B or MAST-11, for which individual assemblies were poorly resolved. We observed that a certain number of SAGs was necessary for obtaining a near-complete genome with a quality equivalent to that obtained from a cultured species (the genomes of MAST-4A and MAST-4C, using 23 and 20 SAGs respectively, were very complete). The new inferred reference MAST genomes provide the backbone to target directly the activity of Marine Stramenopiles.

A single microbial cell contains only femtograms of DNA (and RNA), too low to be processed by current DNA sequencers. Therefore, whole genome amplification (WGA) is a prerequisite for massively parallel DNA sequencing of single cells. In eukaryotic cells, this amplification step (using multiple displacement amplification, MDA) has very often resulted in a biased and partial coverage of the genome sequence (López-Escardó et al. 2017, Mangot et al. 2017). This bias, resulting in low genome completeness even after co-assembly, needs to be taken into consideration when interpreting our results. The throughput and quality of high-throughput sequencing platforms are continuously increasing. In the interest of recovering the function of a target organism, an alternative to single-cell DNA sequencing would be the sequencing of the transcriptomes of single cells (Kolisko et al. 2014; Liu et al. 2017; Ku and Sebé-Pedrós, 2019).

With the advent of '-omic' methods, access to large-scale datasets (metagenomics, metatranscriptomics, metaproteomics, metabolomics) has revolutionized microbial ecology and has accelerated our understanding of biological processes by the analysis of environmental DNA and RNA. Omics data integration has allowed to reveal the functions and physiology of protist species in their natural environment (Marchetti et al. 2012; Alexander et al. 2015; Caron et al. 2017).

In this work, we applied a metatranscriptomic analysis with the idea that the genomic function of a microbe reflects its fundamental ecological niche. Therefore, to address key aspects of functional ecology of MAST protists, we developed a controlled microcosm in which we followed the dynamics of protistan groups of different ecological strategies (phototrophs versus heterotrophs). Subsequent 18S rRNA analysis revealed that we succeeded in enriching heterotrophic flagellates while minimizing the culturing bias. By integrating genomic reconstructions obtained from single cells with metatranscriptomics, the unamended seawater incubations in the dark represented a valuable alternative to identify gene expression over time of uncultured MASTs in their natural assemblage. We also had the opportunity to use a cultured heterotrophic flagellate for a differential gene expression study. The two phases that describe the growth of *Cafeteria burkhardae*, the exponential and starvation phases, were characterized by the expression of different genes.

THE CHALLENGE OF BIG DATA

Following sequencing, the workflow for genomics includes read cleaning and filtering, assembly, alignment (*de novo* or reference-based), gene annotation and functional prediction. Every step aims to reach the gene repertoire and functionality of the species of interest. In general, the assembly and annotation steps are the most critical. An accurate reconstruction is crucial, as the base accuracy of an assembly can affect all downstream analyses (Liao et al. 2018). Assembly involves the merging of reads from the same genome into a single contiguous sequence ’contig’ and contigs into ‘scaffolds’. One of the biggest challenges in assembly is the handling of sequencing errors like substitutions, insertions and deletions. Error rates of Illumina DNA sequencing are 0.02–0.05% (Kelley et al. 2010), which can result in uneven sequencing depth and thus hampering a proper assembly. Single cell sequencing is still under development and has not been yet optimised to avoid non-uniform coverage, which is typical from amplification methods, including ‘blackout’ regions, which are contiguous regions of the genome for which no reads are available. Computational methods to overcome these errors have been designed, such as digital normalisation (KHMer red, BBTools ref) to allow for a better average coverage of unique k-mers across sequencing libraries.

Once assembled, genes can be predicted and functionally annotated. Genome annotation consists of attaching biological meaningful information to genome sequences. The first step, called “gene prediction”, consists of properly determining the location and structure of the protein coding regions in a genome. Typically, genes can be predicted in one of three ways: 1) intrinsic (or ab-initio), 2) extrinsic and/or 3) the combination of the two. The intrinsic approach only focuses on information that can be extracted from the genomic sequence itself such as the coding potential (e.g. start and stop codons producing Open Reading Frames, ORFs), while the extrinsic method uses similarity to other sequence types (e.g. transcripts and/or polypeptides) as input information (Dominguez Del Angel et al. 2018). Lacking usable model protist gene data in the databases, we performed intrinsic gene predictions, further used for functional annotations. Functional annotation consists of associating biological information (function) to genomic elements. Here, we inferred functional annotations and homologous gene families from predicted genes using a state-of-the-art pipeline that combine references from model organism.

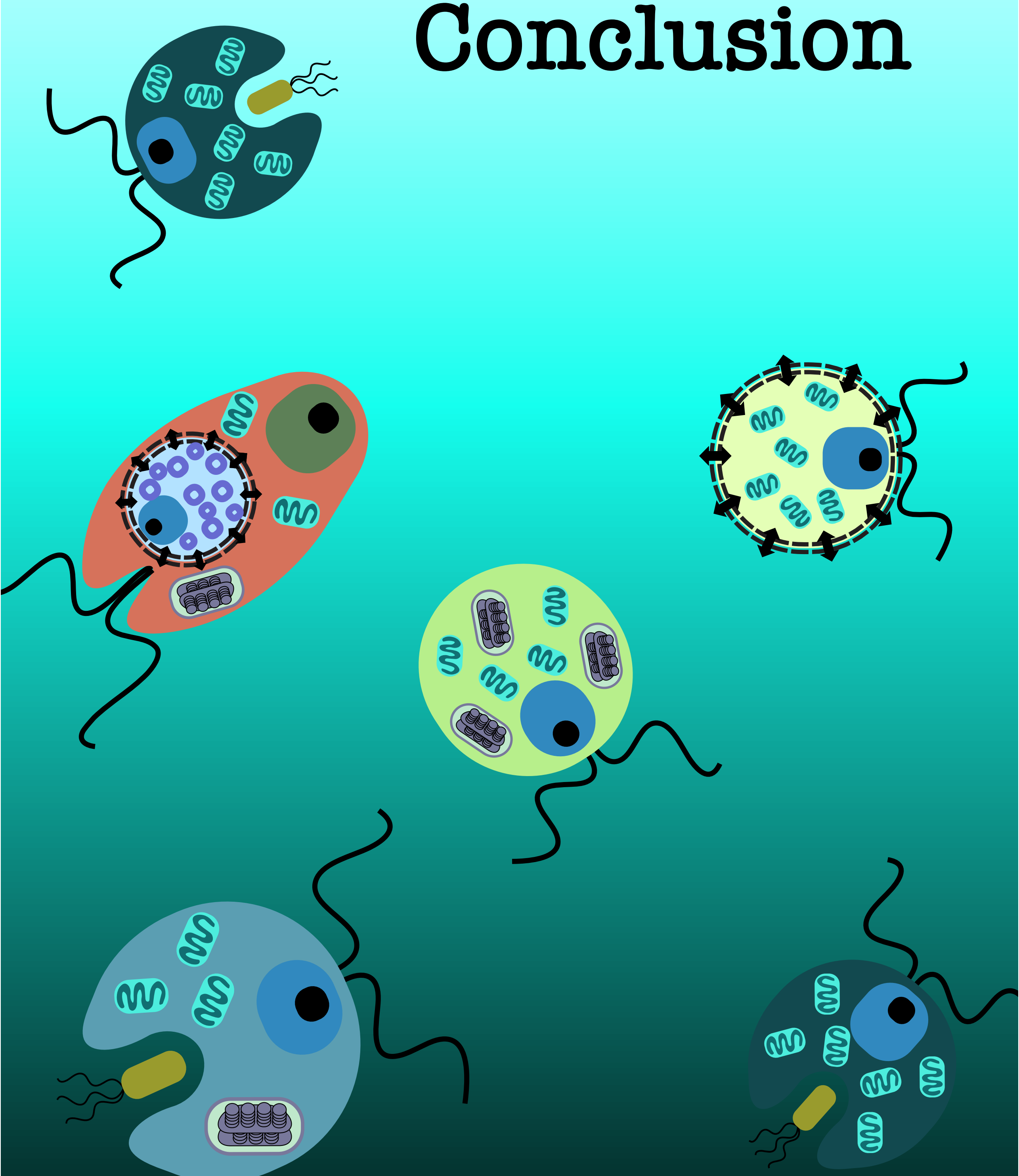
Both steps created immense challenges and, in a few cases, we recovered too few predicted genes, and so the data was insufficient to successfully reach an indication of their functional capacity – for example, the trophic strategies of MAST-1C, MAST-1D or MAST-3C remained elusive.

PHAGOCYTOSIS IN HETEROTROPHIC FLAGELLATES

Heterotrophic flagellates are the smallest and least studied groups of protists both at the morphological and molecular levels. Characterized by a variety of metabolisms and ecological potentials, their functioning in aquatic ecosystems is very important, but not fully understood. Collectively, they are very important bacterial grazers, but they may play other roles. It has been hypothesized that the ecological roles of marine protists remain rooted in trophic behaviors or ecological preferences (Worden et al. 2015). In addition to their ecological role, HFs are also central to important evolutionary questions. Hence, HFs have been essential in the study of the origins of photosynthesis and parasitism (Gawryluk et al. 2019; Janouškovec et al. 2015), origin of multicellular animals (Tikhonenkov et al. 2020) or in helping rooting the tree of eukaryotes and clarification of their relationships (Strassert et al. 2019).

Phagotrophy is part of the trophic strategies of microbes and is central in food webs, however our understanding of this process is mainly based on animal performance (Boulais et al. 2010). A few transcriptomic studies have identified the genes involved in phagocytosis in protists, for example in the slime mold *Dictyostelium discoideum* (Sillo et al. 2008). Here we present a novel investigation of phagocytosis-promoting genes based on the functional genetics of heterotrophic flagellates. By using *Cafeteria burkhardae* we were able to manipulate the culture medium and demonstrate the changes in gene expression of this heterotrophic flagellate during bacterivory. Thus, analysing the genes that were highly expressed during the digestion phase revealed a high implication of the proton pumps vacuolar-type H+ translocating pyrophosphatase (V-PPase) and of several cathepsins. Characteristic genes were therefore accessible. Moreover, we showed the unexpected potential role of light in Marine Stramenopiles digestion. Rhodopsin genes were found in the genome of several MAST species, responding to different functions, and potentially explaining clade diversification. MAST-4E, being separated phylogenetically from the remaining MAST-4 species, contained the MerMAIDS type of rhodopsin. Thus, a future focus towards the rhodopsin machinery could be promising to explain the ecological niches in Marine Stramenopiles.

Conclusion



CONCLUSION

1. Single cell genomics is a promising approach to retrieve the genomes of uncultured picoeukaryotes such as Marine Stramenopiles. It allowed quality assemblies and the recovery of a large number of genes, necessary to address further ecological questions.

2. Unamended incubation in the dark is a reliable tool to study heterotrophic flagellates in their natural environment. Taxonomic groups were discriminated after few days of incubation, and this allowed us to get access to the expressed genes of the enriched phagotrophic assemblage and therefore their bacterivory activity.

3. The combination of metatranscriptomics and genomics (using reference genomes obtained from single cells) helped to target the uncultured MASTs in their direct habitat while providing novel information of their diverse functions.

4. Transcripts of MAST-4A and MAST-4B were the most abundant in the unamended incubation and revealed highly expressed genes related to motility and cytoskeleton remodeling, the first steps involved in vacuole formation, and to lysosomal enzymes such as cathepsins, characteristic of the phagolysosome.

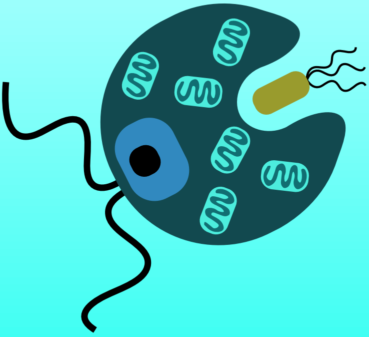
5. The growth of *Cafeteria burkhardae* showed that the Exponential phase was marked by upregulated digestive enzymes (Peptidases), and proton pumps (V-PPase), which were poorly expressed in the Stationary phase. This strong change provides a better understanding of the molecular mechanisms regulating phagocytosis.

6. The focus on digestive enzymes as marker genes for phagocytosis in a comparative genomics analysis within Stramenopiles did not show specificity. These genes were equally present in heterotrophic and phototrophic species. We did not identify unique genes characterizing phagocytosis in heterotrophic flagellates.

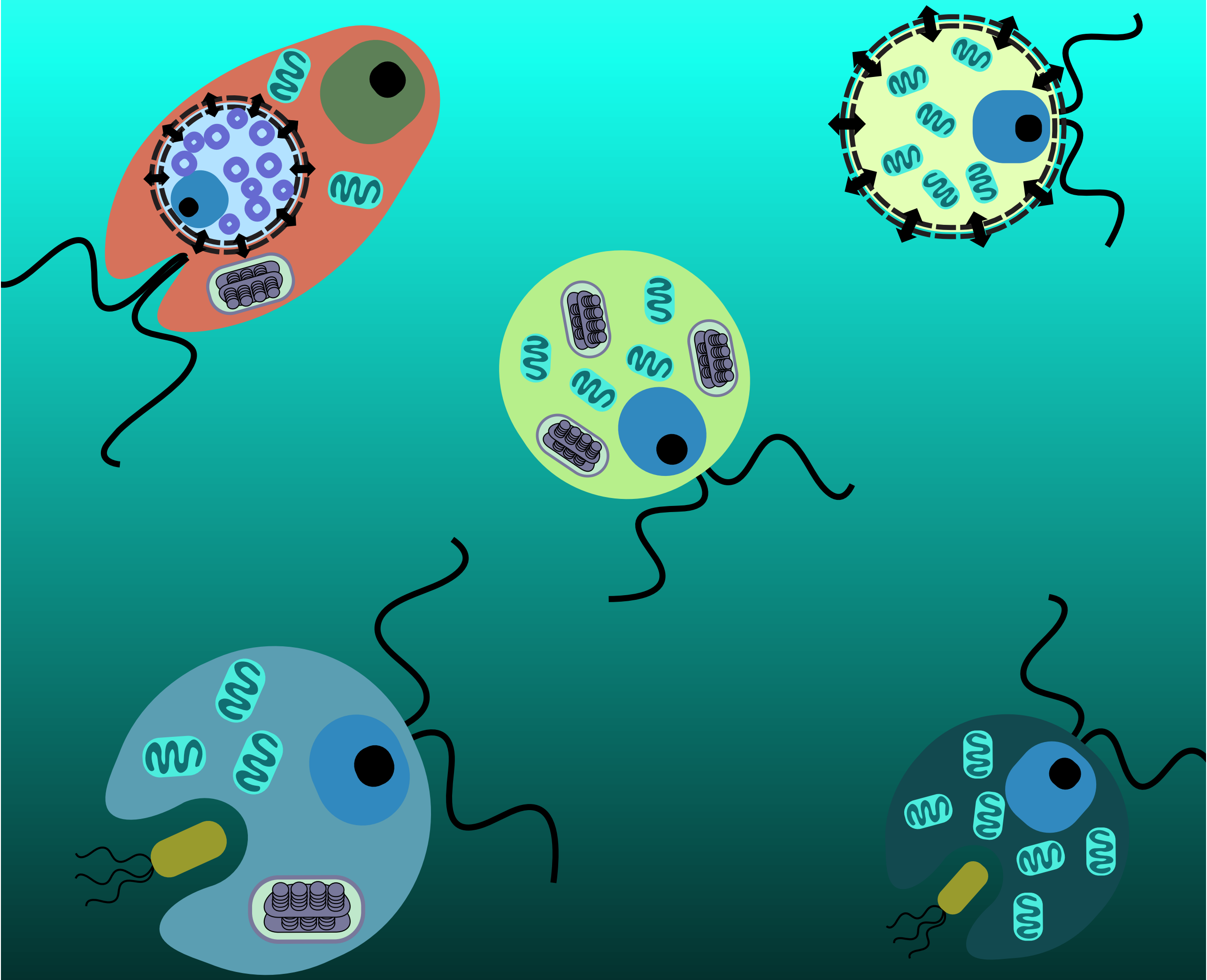
7. Our experiments with *Cafeteria burkhardae* show a higher gene expression of V-PPase than V-ATPase, the proton pump assumed to be involved in vacuole

acidification. Also, the presence of rhodopsins in Marine Stramenopiles suggests a potential role of light during phagocytosis and particularly during the digestion phase - changing our point of view in the use of light by phagotrophs.

8. *Cafeteria burkhardae* cells are very abundant in the sea of microbes and have proved to be convenient to study the functional diversity of small sized protists, allowing to address ecological concepts by molecular biology. *Cafeteria burkhardae* is a great example of a copiotrophic heterotrophic flagellate and could be used as a future model organism.



General References



REFERENCES (Includes Introduction and General discussion)

Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, et al. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. *J Eukaryot Microbiol* 2018; jeu.12691.

Adl SM, Leander BS, Simpson AGB, Archibald JM, Anderson ORoger, Bass D, et al. Diversity, Nomenclature, and Taxonomy of Protists. *Systematic Biology* 2007; **56**: 684–689.

Adl SM, Simpson AGB, Farmer MA, Andersen RA, Anderson OR, Barta JR, et al. The New Higher Level Classification of Eukaryotes with Emphasis on the Taxonomy of Protists. *J Eukaryotic Microbiology* 2005; **52**: 399–451.

Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, et al. The Revised Classification of Eukaryotes. *J Eukaryot Microbiol* 2012; **59**: 429–514.

Alberti, Adriana, Pesant, Stephane, Tara Oceans Consortium, Coordinators, Tara Oceans Expedition, Participants. Methodology used in the lab for molecular analyses and links to the Sequence Read Archive of selected samples from the Tara Oceans Expedition (2009-2013). 2017. PANGAEA - Data Publisher for Earth & Environmental Science. , 631.7 kBytes

Alexander H, Rouco M, Haley ST, Wilson ST, Karl DM, Dyhrman ST. Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. *Proc Natl Acad Sci USA* 2015; **112**: E5972–E5979.

Altermatt F, Fronhofer EA, Garnier A, Giometto A, Hammes F, Klecka J, et al. Big answers from small worlds: a user's guide for protist microcosms as a model system in ecology and evolution. *Methods Ecol Evol* 2015; **6**: 218–231.

Andersen RA. Biology and systematics of heterokont and haptophyte algae. *Am J Bot* 2004; **91**: 1508–1522.

Archibald JM. Endosymbiosis and Eukaryotic Cell Evolution. *Current Biology* 2015; **25**: R911–R921.

Armbrust EV. The Genome of the Diatom Thalassiosira Pseudonana: Ecology, Evolution, and Metabolism. *Science* 2004; **306**: 79–86.

Arndt, H., Dietrich, D., Auer, B., Cleven, E.-J., Gräfenhan, T., Weitere, M., Mylnikov, A.P.,. Functional diversity of heterotrophic flagellates in aquatic ecosystems, in: Leadbeater, B.S., Green, J.C. (Eds.), The Flagellates - Unity, Diversity and Evolution. Taylor & Francis Ltd, London, 2000, pp. 240–268.

Bar-On YM, Milo R. The Biomass Composition of the Oceans: A Blueprint of Our Blue Planet. *Cell* 2018; **179**: 1451–1454.

Batz W, Wunderlich F. Structural transformation of the phagosomal membrane in Tetrahymena cells endocytosing latex beads. *Arch Microbiol* 1976; **109**: 215–220.

Benites LF, Poulton N, Labadie K, Sieracki ME, Grimsley N, Piganeau G. Single cell ecogenomics reveals mating types of individual cells and ssDNA viral infections in the smallest photosynthetic eukaryotes. *Phil Trans R Soc B* 2019; **374**: 20190089.

Boulais J, Trost M, Landry CR, Dieckmann R, Levy ED, Soldati T, et al. Molecular characterization of the evolution of phagosomes. *Mol Syst Biol* 2010; **6**: 423.

Burki F. The Eukaryotic Tree of Life from a Global Phylogenomic Perspective. *Cold Spring Harbor Perspectives in Biology* 2014; **6**: a016147–a016147.

Burki F, Kaplan M, Tikhonenkov DV, Zlatogursky V, Minh BQ, Radaykina LV, et al. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc R Soc B* 2016; **283**: 20152802.

Burki F, Roger AJ, Brown MW, Simpson AGB. The New Tree of Eukaryotes. *Trends in Ecology & Evolution* 2019; **35**: 43–55.

Burns JA, Paasch A, Narechania A, Kim E. Comparative Genomics of a Bacterivorous Green Alga Reveals Evolutionary Causalities and Consequences of Phago-Mixotrophic Mode of Nutrition. *Genome Biol Evol* 2015; **7**: 3047–3061.

Burns JA, Pittis AA, Kim E. Gene-based predictive models of trophic modes suggest Asgard archaea are not phagocytotic. *Nat Ecol Evol* 2018; **2**: 697–704.

Calbet A. The trophic roles of microzooplankton in marine systems. *ICES Journal of Marine Science* 2008; **65**: 325–331.

Calbet A, Landry MR. Phytoplankton growth, microzooplankton grazing, and carbon cycling in marine systems. *Limnol Oceanogr* 2004; **49**: 51–57.

Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J* 2017; **11**: 2639–2643.

Caron DA, Worden AZ, Countway PD, Demir E, Heidelberg KB. Protists are microbes too: a perspective. *ISME J* 2009; **3**: 4–12.

Caron DA, Alexander H, Allen AE, Archibald JM, Armbrust EV, Bachy C, et al. Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nat Rev Microbiol* 2017; **15**: 6–20.

Caron DA, Countway PD, Jones AC, Kim DY, Schnetzer A. Marine Protistan Diversity. *Annu Rev Mar Sci* 2012; **4**: 467–493.

Cavalier-Smith T. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *International Journal of Systematic and Evolutionary Microbiology* 2002; **52**: 297–354.

Cavan EL, Belcher A, Atkinson A, Hill SL, Kawaguchi S, McCormack S, et al. The importance of Antarctic krill in biogeochemical cycles. *Nat Commun* 2019; **10**: 4742.

Chan TSY, Nasser F, St-Denis CH, Mandal HS, Ghafari P, Hadjout-Rabi N, et al. Carbon nanotube compared with carbon black: effects on bacterial survival against grazing by ciliates and antimicrobial treatments. *Nanotoxicology* 2012; **7**: 251–258.

Christaki U, Courties C, Massana R, Catala P, Lebaron P, Gasol JM, et al. Optimized routine flow cytometric enumeration of heterotrophic flagellates using SYBR Green I: FC analysis of HF. *Limnol Oceanogr Methods* 2011; **9**: 329–339.

Collier JL, Rest JS. Swimming, gliding, and rolling toward the mainstream: cell biology of marine protists. *MBoC* 2019; **30**: 1245–1248.

Cosson P, Soldati T. Eat, kill or die: when amoeba meets bacteria. *Current Opinion in Microbiology* 2008; **11**: 271–276.

Dawson SC, Fritz-Laylin LK. Sequencing free-living protists: the case for metagenomics. *Environmental Microbiology* 2009; **11**: 1627–1631.

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

del Campo J, Kolisko M, Boscaro V, Santoferrara LF, Nenarokov S, Massana R, et al. EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLoS Biol* 2018; **16**: e2005849.

del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I. The others: our biased perspective of eukaryotic genomes. *Trends in Ecology & Evolution* 2014; **29**: 252–259.

Dent RM, Haglund CM, Chin BL, Kobayashi MC, Niyogi KK. Functional Genomics of Eukaryotic Photosynthesis Using Insertional Mutagenesis of *Chlamydomonas reinhardtii*. *Plant Physiol* 2005; **137**: 545–556.

Derelle R, López-García P, Timpano H, Moreira D. A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). *Molecular Biology and Evolution* 2016; **33**: 2890–2898.

Dolan JR, Šimek K. Ingestion and digestion of an autotrophic picoplankter, Synechococcus, by a heterotrophic nanoflagellate, Bode saLtans. *Limnol Oceanogr* 1998; **43**: 1740–1746.

Dominguez Del Angel V, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Vinnere Pettersson O, et al. Ten steps to get started in Genome Assembly and Annotation. *F1000Res* 2018; **7**: 148.

Falkowski PG. The Evolution of Modern Eukaryotic Phytoplankton. *Science* 2004; **305**: 354–360.

Finlay BJ. Global Dispersal of Free-Living Microbial Eukaryote Species. *Science* 2002; **296**: 1061–1063.

Flynn KJ, Mitra A, Anestis K, Anschütz AA, Calbet A, Ferreira GD, et al. Mixotrophic protists and a new paradigm for marine ecology: where does plankton research go now? *Journal of Plankton Research* 2019; **41**: 375–391.

Fok AK, Lee Y, Allen RD. The Correlation of Digestive Vacuole pH and Size with the Digestive Cycle in *Paramecium caudatum* [1]. *The Journal of Protozoology* 1982; **29**: 409–414.

Freeman SA, Grinstein S. Phagocytosis: receptors, signal integration, and the cytoskeleton. *Immunol Rev* 2014; **262**: 193–215.

Gawryluk RMR, Tikhonenkov DV, Hehenberger E, Husnik F, Mylnikov AP, Keeling PJ. Non-photosynthetic predators are sister to red algae. *Nature* 2019; **572**: 240–243.

Gawryluk RMR, del Campo J, Okamoto N, Strassert JFH, Lukeš J, Richards TA, et al. Morphological Identification and Single-Cell Genomics of Marine Diplonemids. *Current Biology* 2016; **26**: 3053–3059.

Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and Structural Biotechnology Journal* 2020; **18**: 9–19.

Giner CR, Pernice MC, Balagué V, Duarte CM, Gasol JM, Logares R, et al. Marked changes in diversity and relative activity of picoeukaryotes with depth in the world ocean. *ISME J* 2020; **14**: 437–449.

Gómez F, Moreira D, Benzerara K, López-García P. Solenicola setigera is the first characterized member of the abundant and cosmopolitan uncultured marine stramenopile group MAST-3: Solenicola belongs to uncultured stramenopiles MAST-3. *Environmental Microbiology* 2011; **13**: 193–202.

Gonda K, Komatsu M, Numata O. Calmodulin and Ca2+/Calmodulin-Binding Proteins are Involved in Tetrahymena thermophila Phagocytosis. *Cell Structure and Function* 2000; **25**: 243–251.

Gooday AJ, Schoenle A, Dolan JR, Arndt H. Protist diversity and function in the dark ocean - challenging the paradigms of deep-sea ecology with special emphasis on foraminiferans and naked protists. *European Journal of Protistology* 2020; 125721.

Grattepanche J-D, Walker LM, Ott BM, Paim Pinto DL, Delwiche CF, Lane CE, et al. Microbial Diversity in the Eukaryotic SAR Clade: Illuminating the Darkness Between Morphology and Molecular Data. *BioEssays* 2018; **40**: 1700198.

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, et al. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research* 2012; **41**: D597–D604.

Hackett JD, Anderson DM, Erdner DL, Bhattacharya D. Dinoflagellates: a remarkable evolutionary experiment. *Am J Bot* 2004; **91**: 1523–1534.

Hennemuth W, Rhoads LS, Eichelberger H, Watanabe M, Van Bell KM, Ke L, et al. Ingestion and Inactivation of Bacteriophages by *Tetrahymena*. *Journal of Eukaryotic Microbiology* 2008; **55**: 44–50.

Hofstatter PG, Ribeiro GM, Porfírio-Sousa AL, Lahr DJG. The Sexual Ancestor of all Eukaryotes: A Defense of the “Meiosis Toolkit”: A Rigorous Survey Supports the Obligate Link between Meiosis Machinery and Sexual Recombination. *BioEssays* 2020; 2000037.

Jacobs ME, DeSouza LV, Samaranayake H, Pearlman RE, Siu KWM, Klobutcher LA. The Tetrahymena thermophila Phagosome Proteome. *Eukaryotic Cell* 2006; **5**: 1990–2000.

Janouškovec J, Tikhonenkov DV, Burki F, Howe AT, Kolísko M, Mylnikov AP, et al. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proc Natl Acad Sci USA* 2015; **112**: 10200–10207.

Jurgens K, Massana R. Protistan Grazing on Marine Bacterioplankton. In: Kirchman DL (ed). *Microbial Ecology of the Oceans*. 2008. John Wiley & Sons, Inc., Hoboken, NJ, USA, pp 383–441.

Kamies R, Martinez-Jimenez CP. Advances of single-cell genomics and epigenomics in human disease: where are we now? *Mamm Genome* 2020; **31**: 170–180.

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, et al. The tree of eukaryotes. *Trends in Ecology & Evolution* 2005; **20**: 670–676.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol* 2014; **12**: e1001889.

Keeling PJ, Campo J del. Marine Protists Are Not Just Big Bacteria. *Current Biology* 2017; **27**: R541–R549.

Kelley DR, Schatz MC, Salzberg SL. Quake: quality-aware detection and correction of sequencing errors. *Genome Biology* 2010; **11**: R116.

Kilias ES, Peeken I, Metfies K. Insight into protist diversity in Arctic sea ice and melt-pond aggregate obtained by pyrosequencing. *Polar Research* 2014; **33**: 23466.

Kinchen JM, Ravichandran KS. Phagosome maturation: going through the acid test. *Nat Rev Mol Cell Biol* 2008; **9**: 781–795.

King N. The Unicellular Ancestry of Animal Development. *Developmental Cell* 2004; **7**: 313–325.

Kolisko M, Boscaro V, Burki F, Lynn DH, Keeling PJ. Single-cell transcriptomics for microbial eukaryotes. *Current Biology* 2014; **24**: R1081–R1082.

Ku C, Sebé-Pedrós A. Using single-cell transcriptomics to understand functional states and interactions in microbial eukaryotes. *Phil Trans R Soc B* 2019; **374**: 20190098.

Kuspa A, Sucgang R, Shaulsky G. The promise of a protist: the Dictyostelium genome project. *Functional & Integrative Genomics* 2001; **1**: 279–293.

Lasken RS, McLean JS. Recent advances in genomic DNA sequencing of microbial species from single cells. *Nat Rev Genet* 2014; **15**: 577–584.

Laursen L. Spain's ship comes in. *Nature* 2011; **475**: 16–17.

Lax G, Eglit Y, Eme L, Bertrand EM, Roger AJ, Simpson AGB. Hemimastigophora is a novel supra-kingdom-level lineage of eukaryotes. *Nature* 2018; **564**: 410–414.

Lee KH, Jeong HJ, Yoon EY, Jang SH, Kim HS, Yih W. Feeding by common heterotrophic dinoflagellates and a ciliate on the red-tide ciliate Mesodinium rubrum. *ALGAE* 2014; **29**: 153–163.

Leonard G, Labarre A, Milner DS, Monier A, Soanes D, Wideman JG, et al. Comparative genomic analysis of the 'pseudofungus' *Hyphochytrium catenoides*. *Open Biol* 2018; **8**: 170184.

Levin R, Grinstein S, Canton J. The life cycle of phagosomes: formation, maturation, and resolution. *Immunol Rev* 2016; **273**: 156–179.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, et al. The Diploid Genome Sequence of an Individual Human. *PLoS Biol* 2007; **5**: e254.

Li J, Montagnes DJS. Restructuring Fundamental Predator-Prey Models by Recognising Prey-Dependent Conversion Efficiency and Mortality Rates. *Protist* 2015; **166**: 211–223.

Liao X, Li M, Zou Y, Wu F-X, Yi-Pan, Wang J. Current challenges and solutions of de novo assembly. *Quant Biol* 2018; **7**: 90–109.

Liu Z, Hu SK, Campbell V, Tatters AO, Heidelberg KB, Caron DA. Single-cell transcriptomics of small microbial eukaryotes: limitations and potential. *ISME J* 2017; **11**: 1282–1285.

Logares R, Deutschmann IM, Junger PC, Giner CR, Krabberød AK, Schmidt TSB, et al. Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome* 2020; **8**: 55.

López-Escardó D, Grau-Bové X, Guillaumet-Adkins A, Gut M, Sieracki ME, Ruiz-Trillo I. Evaluation of single-cell genomics to address evolutionary questions using three SAGs of the choanoflagellate Monosiga brevicollis. *Sci Rep* 2017; **7**: 11025.

López-Escardó D, Grau-Bové X, Guillaumet-Adkins A, Gut M, Sieracki ME, Ruiz-Trillo I. Evaluation of single-cell genomics to address evolutionary questions using three SAGs of the choanoflagellate Monosiga brevicollis. *Sci Rep* 2017; **7**: 11025.

López-García P, Rodríguez-Valera F, Pedrós-Alió C, Moreira D. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* 2001; **409**: 603–607.

Mahé F, de Vargas C, Bass D, Czech L, Stamatakis A, Lara E, et al. Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nat Ecol Evol* 2017; **1**: 0091.

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

Maicher MT, Tiedtke A. Biochemical analysis of membrane proteins from an early maturation stage of phagosomes. Electrophoresis. 1999;20(4-5):1011-1016.

Malik S-B, Pightling AW, Stefaniak LM, Schurko AM, Logsdon JM. An Expanded Inventory of Conserved Meiotic Genes Provides Evidence for Sex in Trichomonas vaginalis. *PLoS ONE* 2008; **3**: e2879.

Mangot J-F, Logares R, Sánchez P, Latorre F, Seeleuthner Y, Mondy S, et al. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci Rep* 2017; **7**: 41498.

Mangot J-F, Logares R, Sánchez P, Latorre F, Seeleuthner Y, Mondy S, et al. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci Rep* 2017; **7**: 41498.

Marchetti A, Schruth DM, Durkin CA, Parker MS, Kodner RB, Berthiaume CT, et al. Comparative metatranscriptomics identifies molecular bases for the physiological responses of phytoplankton to varying iron availability. *Proceedings of the National Academy of Sciences* 2012; **109**: E317–E325.

Marshansky V, Futai M. The V-type H+-ATPase in vesicular trafficking: targeting, regulation and function. *Current Opinion in Cell Biology* 2008; **20**: 415–426.

Massana R, BalaguÃ© V, Guillou L, PedrÃ³s-AliÃ³ C. Picoeukaryotic diversity in an oligotrophic coastal site studied by molecular and culturing approaches. *FEMS Microbiology Ecology* 2004; **50**: 231–243.

Massana R, del Campo J, Sieracki ME, Audic S, Logares R. Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J* 2014; **8**: 854–866.

Massana, R. “Protistan diversity in environmental molecular surveys” in Marine Protists, eds S. Ohtsuka, T. Suzaki, T. Horiguchi, N. Suzuki, and F. Not (Tokyo: Springer), 2015; 3–21.

Massana R, Gobet A, Audic S, Bass D, Bittner L, Boutte C, et al. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing: Protist diversity in European coastal areas. *Environ Microbiol* 2015; **17**: 4035–4049.

Massana R, Terrado R, Forn I, Lovejoy C, Pedros-Alio C. Distribution and abundance of uncultured heterotrophic flagellates in the world oceans. *Environ Microbiol* 2006; **8**: 1515–1522.

Massana R, Unrein F, Rodríguez-Martínez R, Forn I, Lefort T, Pinhassi J, et al. Grazing rates and functional diversity of uncultured heterotrophic flagellates. *The ISME Journal* 2009; **3**: 588–596.

Massana R, Unrein F, Rodríguez-Martínez R, Forn I, Lefort T, Pinhassi J, et al. Grazing rates and functional diversity of uncultured heterotrophic flagellates. *ISME J* 2009; **3**: 588–596.

Matz C, Boenigk J, Arndt H, Jürgens K. Role of bacterial phenotypic traits in selective feeding of the heterotrophic nanoflagellate Spumella sp. *Aquat Microb Ecol* 2002; **27**: 137–148.

Metzker ML. Sequencing technologies — the next generation. *Nat Rev Genet* 2010; **11**: 31–46.

Montagnes DJS. Ecophysiology and Behavior of Tintinnids. In: Dolan JR, Montagnes DJS, Agatha S, Coats DW, Stoecker DK (eds). *The Biology and Ecology of Tintinnid Ciliates*. 2012. John Wiley & Sons, Ltd, Chichester, UK, pp 85–121.

Montagnes D, Roberts E, Lukeš J, Lowe C. The rise of model protozoa. *Trends in Microbiology* 2012; **20**: 184–191.

Montagnes D, Barbosa A, Boenigk J, Davidson K, Jürgens K, Macek M, et al. Selective feeding behaviour of key free-living protists: avenues for continued study. *Aquat Microb Ecol* 2008; **53**: 83–98.

More K, Simpson AGB, Hess S. Two New Marine Species of *Placopus* (Vampyrellida, Rhizaria) That Perforate the Theca of *Tetraselmis* (Chlorodendrales, Viridiplantae). *J Eukaryot Microbiol* 2019; **66**: 560–573.

Needham DM, Sachdeva R, Fuhrman JA. Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *ISME J* 2017; **11**: 1614–1629.

Niedergang F, Grinstein S. How to build a phagosome: new concepts for an old process. *Current Opinion in Cell Biology* 2018; **50**: 57–63.

Niedergang F, Grinstein S. Phagocytosis: receptors, signal integration, and the cytoskeleton. *Immunol Rev* 2018; **262**: 193–215.

Nilsson JR. On Food Vacuoles in *Tetrahymena pyriformis* GL. *The Journal of Protozoology* 1977; **24**: 502–507.

Not F, del Campo J, Balagué V, de Vargas C, Massana R. New Insights into the Diversity of Marine Picoeukaryotes. *PLoS ONE* 2009; **4**: e7143.

O'Malley MA, Leger MM, Wideman JG, Ruiz-Trillo I. Concepts of the last eukaryotic common ancestor. *Nat Ecol Evol* 2019; **3**: 338–344.

Parfrey LW, Lahr DJG, Katz LA. The Dynamic Nature of Eukaryotic Genomes. *Molecular Biology and Evolution* 2008; **25**: 787–794.

Patil S, Moeys S, von Dassow P, Huysman MJJ, Mapleson D, De Veylder L, et al. Identification of the meiotic toolkit in diatoms and exploration of meiosis-specific SPO11 and RAD51 homologs in the sexual species Pseudo-nitzschia multistriata and Seminavis robusta. *BMC Genomics* 2015; **16**: 930.

Pauwels A-M, Trost M, Beyaert R, Hoffmann E. Patterns, Receptors, and Signals: Regulation of Phagosome Maturation. *Trends in Immunology* 2017; **38**: 407–422.

Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, et al. CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. *PLoS Biol* 2012; **10**: e1001419.

Pernthaler J, Posch T. Microbial Food Webs. *Encyclopedia of Inland Waters*. 2009. Elsevier, pp 244–251.

Pernthaler J. Predation on prokaryotes in the water column and its ecological implications. *Nat Rev Microbiol* 2005; **3**: 537–546.

Pesant S, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D, et al. Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* 2015; **2**: 150023.

Pinard R, de Winter A, Sarkis GJ, Gerstein MB, Tartaro KR, Plant RN, et al. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 2006; **7**: 216.

Podar M, Keller M, Hugenholtz P. Single Cell Whole Genome Amplification of Uncultivated Organisms. In: Epstein SS (ed). *Uncultivated Microorganisms*. 2009. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 241–256.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* 2012; **41**: D590–D596.

Ramesh MA, Malik S-B, Logsdon JM. A Phylogenomic Inventory of Meiotic Genes. *Current Biology* 2005; **15**: 185–191.

Richards TA, Jones MDM, Leonard G, Bass D. Marine Fungi: Their Ecology and Molecular Diversity. *Annu Rev Mar Sci* 2012; **4**: 495–522.

Richter DJ, Fozouni P, Eisen MB, King N. Gene family innovation, conservation and loss on the animal stem lineage. *eLife* 2018; **7**: e34226.

Richter DJ, Berney C, Strassert JFH, Burki F, de Vargas C. EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotic life. 2020. Evolutionary Biology.

Roberts EC, Zubkov MV, Martin-Cereceda M, Novarino G, Wootton EC. Cell surface lectin-binding glycoconjugates on marine planktonic protists. *FEMS Microbiology Letters* 2006; **265**: 202–207.

Roberts RL, Barbieri MA, Ullrich J, Stahl PD. Dynamics of rab5 activation in endocytosis and phagocytosis. *Journal of Leukocyte Biology* 2000; **68**: 627–632.

Rohatgi R, Ma L, Miki H, Lopez M, Kirchhausen T, Takenawa T, et al. The Interaction between N-WASP and the Arp2/3 Complex Links Cdc42-Dependent Signals to Actin Assembly. *Cell* 1999; **97**: 221–231.

Roy RS, Price DC, Schliep A, Cai G, Korobeynikov A, Yoon HS, et al. Single cell genome analysis of an uncultured heterotrophic stramenopile. *Sci Rep* 2014; **4**: 4780.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* 2007; **5**: e77.

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences* 1977; **74**: 5463–5467.

Schoenle A, Hohlfeld M, Rosse M, Filz P, Wylezich C, Nitsche F, et al. Global comparison of bicosoecid Cafeteria-like flagellates from the deep ocean and surface waters, with reorganization of the family Cafeteriaceae. *European Journal of Protistology* 2020; **73**: 125665.

Sherr BF, Sherr BF. Encyclopedia of microbiology, 3. ed. 2009. Elsevier, Acad. Press, Amsterdam.

Sherr BF, Sherr EB, Rassoulzadegan F. Rates of digestion of bacteria by marine phagotrophic protozoa: temperature dependence. *Appl Environ Microbiol* 1988; **54**: 1091–1095.

Sherr EB, Sherr BF. [No title found]. *Antonie van Leeuwenhoek* 2002; **81**: 293–308.

Shiratori T, Suzuki S, Kakizawa Y, Ishida K. Phagocytosis-like cell engulfment by a planctomycete bacterium. *Nat Commun* 2019; **10**: 5529.

Sibbald SJ, Archibald JM. More protist genomes needed. *Nat Ecol Evol* 2017; **1**: 0145.

Sillo A, Bloomfield G, Balest A, Balbo A, Pergolizzi B, Peracino B, et al. Genome-wide transcriptional changes induced by phagocytosis or growth on bacteria in Dictyostelium. *BMC Genomics* 2008; **9**: 291.

Simek K, Jezbera J, Hornák K, Vrba J, Sed’a J. Role of diatom-attached choanoflagellates of the genus Salpingoeca as pelagic bacterivores. *Aquat Microb Ecol* 2004; **36**: 257–269.

Simpson AGB, Slamovits CH, Archibald JM. Protist Diversity and Eukaryote Phylogeny. In: Archibald JM, Simpson AGB, Slamovits CH, Margulis L, Melkonian M, Chapman DJ, et al. (eds). *Handbook of the Protists*. 2017. Springer International Publishing, Cham, pp 1–21.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proceedings of the National Academy of Sciences* 2006; **103**: 12115–12120.

Speijer D, Lukeš J, Eliáš M. Sex is a ubiquitous, ancient, and inherent attribute of eukaryotic life. *Proc Natl Acad Sci USA* 2015; **112**: 8827–8834.

Stern R, Kraberg A, Bresnan E, Kooistra WHCF, Lovejoy C, Montresor M, et al. Molecular analyses of protists in long-term observation programmes—current status and future perspectives. *Journal of Plankton Research* 2018; **40**: 519–536.

Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner H-W, et al. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology* 2010; **19**: 21–31.

Stoecker D, Johnson M, deVargas C, Not F. Acquired phototrophy in aquatic protists. *Aquat Microb Ecol* 2009; **57**: 279–310.

Strassert JFH, Jamy M, Mylnikov AP, Tikhonenkov DV, Burki F. New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life. *Molecular Biology and Evolution* 2019; **36**: 757–765.

Strassert JFH, Jamy M, Mylnikov AP, Tikhonenkov DV, Burki F. New Phylogenomic Analysis of the Enigmatic Phylum Telonemia Further Resolves the Eukaryote Tree of Life. *Molecular Biology and Evolution* 2019; **36**: 757–765.

Strassert JFH, Karnkowska A, Hehenberger E, del Campo J, Kolisko M, Okamoto N, et al. Single cell genomics of uncultured marine alveolates shows paraphyly of basal dinoflagellates. *ISME J* 2018; **12**: 304–308.

Suzuki N, Not F. Biology and Ecology of Radiolaria. In: Ohtsuka S, Suzaki T, Horiguchi T, Suzuki N, Not F (eds). *Marine Protists*. 2015. Springer Japan, Tokyo, pp 179–222.

Tashyreva D, Prokopchuk G, Yabuki A, Kaur B, Faktorová D, Votýpka J, et al. Phylogeny and Morphology of New Diplonemids from Japan. *Protist* 2018; **169**: 158–179.

Thurman J, Drinkall J, Parry J. Digestion of bacteria by the freshwater ciliate Tetrahymena pyriformis. *Aquat Microb Ecol* 2010; **60**: 163–174.

Tikhonenkov DV, Hehenberger E, Esaulov AS, Belyakova OI, Mazei YA, Mylnikov AP, et al. Insights into the origin of metazoan multicellularity from predatory unicellular relatives of animals. *BMC Biol* 2020; **18**: 39.

Torruella G, de Mendoza A, Grau-Bové X, Antó M, Chaplin MA, del Campo J, et al. Phylogenomics Reveals Convergent Evolution of Lifestyles in Close Relatives of Animals and Fungi. *Current Biology* 2015; **25**: 2404–2410.

van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: Tone down the bias. *Experimental Cell Research* 2014; **322**: 12–20.

Verity PG. Feeding In Planktonic Protozoans: Evidence For Non-Random Acquisition of Prey. *The Journal of Protozoology* 1991; **38**: 69–76.

Vieira OV, Bucci C, Harrison RE, Trimble WS, Lanzetti L, Gruenberg J, et al. Modulation of Rab5 and Rab7 Recruitment to Phagosomes by Phosphatidylinositol 3-Kinase. *MCB* 2003; **23**: 2501–2514.

Waller RF, Cleves PA, Rubio-Brotons M, Woods A, Bender SJ, Edgcomb V, et al. Strength in numbers: Collaborative science for new experimental model systems. *PLoS Biol* 2018; **16**: e2006333.

Wang Y, Navin NE. Advances and Applications of Single-Cell Sequencing Technologies. *Molecular Cell* 2015; **58**: 598–609.

Ward RM, Schmieder R, Highnam G, Mittelman D. Big data challenges and opportunities in high-throughput sequencing. *Systems Biomedicine* 2013; **1**: 29–34.

Weber F, del Campo J, Wylezich C, Massana R, Jürgens K. Unveiling Trophic Functions of Uncultured Protist Taxa by Incubation Experiments in the Brackish Baltic Sea. *PLoS ONE* 2012; **7**: e41970.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* 2008; **452**: 872–876.

Wideman JG, Monier A, Rodríguez-Martínez R, Leonard G, Cook E, Poirier C, et al. Unexpected mitochondrial genome diversity revealed by targeted single-cell genomics of heterotrophic flagellated protists. *Nat Microbiol* 2020; **5**: 154–165.

Wootton EC, Zubkov MV, Jones DH, Jones RH, Martel CM, Thornton CA, et al. Biochemical prey recognition by planktonic protozoa. *Environ Microbiol* 2007; **9**: 216–222.

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

Worden AZ, Nolan JK, Palenik B. Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnol Oceanogr* 2004; **49**: 168–179.

Woyke T, Doud DFR, Schulz F. The trajectory of microbial single-cell sequencing. *Nat Methods* 2017; **14**: 1045–1054.

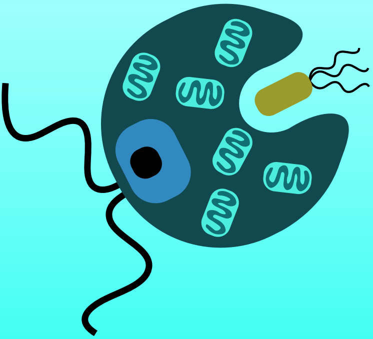
Yates RM, Hermetter A, Russell DG. The Kinetics of Phagosome Maturation as a Function of Phagosome/Lysosome Fusion and Acquisition of Hydrolytic Activity: Assays of Phagosome Maturation. *Traffic* 2005; **6**: 413–420.

Yilmaz S, Singh AK. Single cell genome sequencing. *Current Opinion in Biotechnology* 2012; **23**: 437–443.

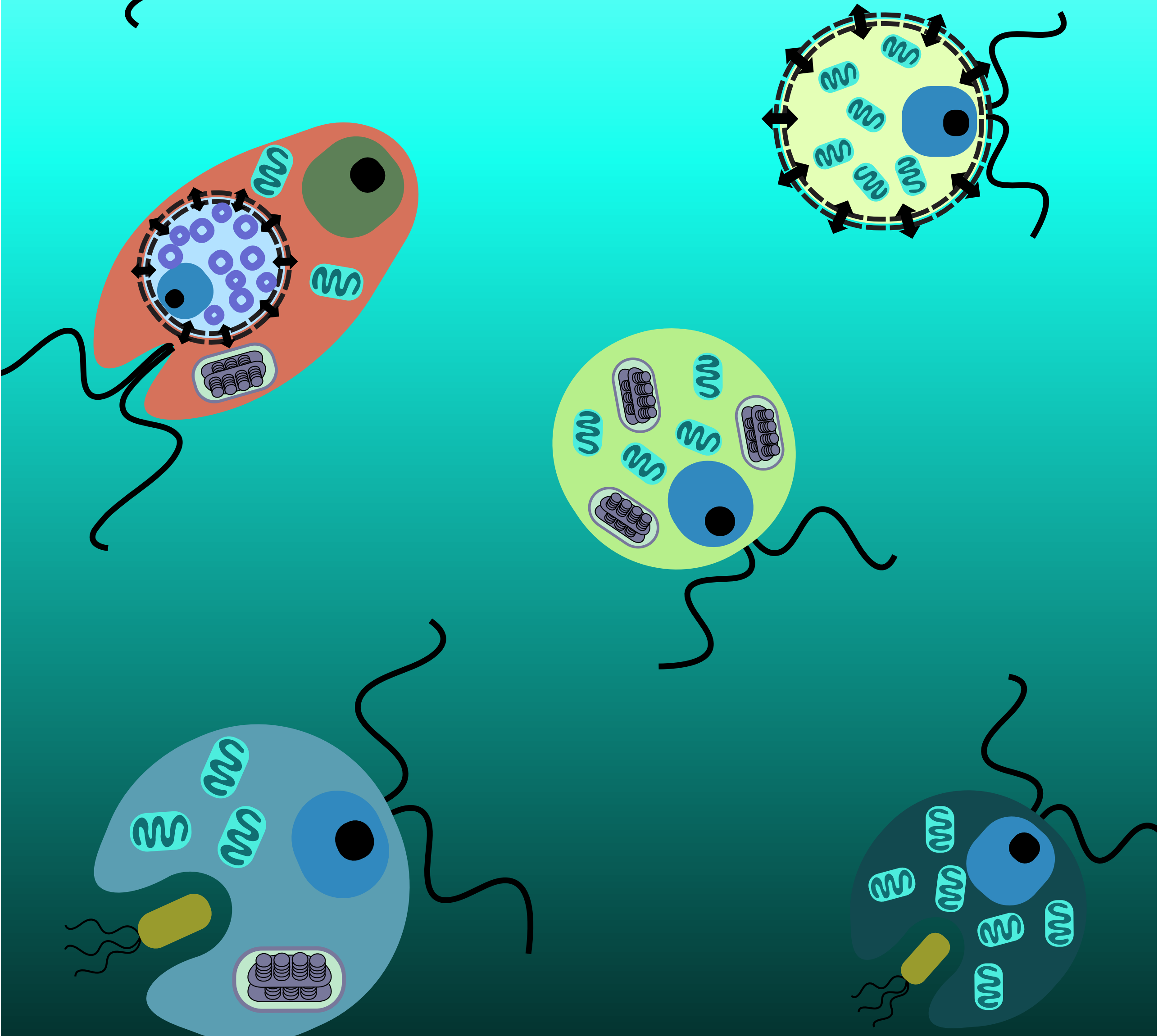
Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, et al. Single-Cell Genomics Reveals Organismal Interactions in Uncultivated Marine Protists. *Science* 2011; **332**: 714–717.

Yubuki N, Leander BS, Silberman JD. Ultrastructure and Molecular Phylogenetic Position of a Novel Phagotrophic Stramenopile from Low Oxygen Environments: Rictus lutensis gen. et sp. nov. (Bicosoecida, incertae sedis). *Protist* 2010; **161**: 264–278.

Yutin N, Wolf MY, Wolf YI, Koonin EV. The origins of phagocytosis and eukaryogenesis. *Biol Direct* 2009; **4**: 9.



Acknowledgments



ACKNOWLEDGMENTS

This research was made possible by a grant from the Marie Curie European Union; I have been very happy to be part of this ITN project “SINGEK” and I would like to thank my supervisor Ramon Massana for trusting me in being part of this PhD experience. I would also like to express my gratitude for the continuous support in my PhD study and research. A special thanks to Elena, our project coordinator for always being here whenever I needed.

Thank you to my friends from SINGEK and my fellow mates in Barcelona, iin France, everyone across the oceans.

I am grateful to my committee members Dr Iñaki Ruiz-Trillo, Dr Gwenaelle Piganeau, Dr Javier del Campo and Dr Jerome Conte for their time and advice all along my PhD.

I would like to thank Dr Colomban de Vargas for the opportunity to be part of the TARA crew a few times - It has been an amazing experience to be on field work in the middle of the Pacific ocean.

Finally, many thanks to my family for their help and support.

UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

Institut
de Ciències
del Mar