Obiol *et al. Microbiome* (2023) 11:134
https://doi.org/10.1186/s40168-023-01571-5

Microbiome

RESEARCH

Open Access

Gene expression dynamics of natural assemblages of heterotrophic flagellates during bacterivory

Aleix Obiol[1*], David López-Escardó[1], Eric D. Salomaki[2], Monika M. Wiśniewska[2,3], Irene Forn[1], Elisabet Sà[1], Dolors Vaqué[1], Martin Kolísko[2,3] and Ramon Massana[1*]

Abstract

Background Marine heterotrophic flagellates (HF) are dominant bacterivores in the ocean, where they represent the trophic link between bacteria and higher trophic levels and participate in the recycling of inorganic nutrients for regenerated primary production. Studying their activity and function in the ecosystem is challenging since most of the HFs in the ocean are still uncultured. In the present work, we investigated gene expression of natural HF communities during bacterivory in four unamended seawater incubations.

Results The most abundant species growing in our incubations belonged to the taxonomic groups MAST-4, MAST-7, Chrysophyceae, and Telonemia. Gene expression dynamics were similar between incubations and could be divided into three states based on microbial counts, each state displaying distinct expression patterns. The analysis of samples where HF growth was highest revealed some highly expressed genes that could be related to bacterivory. Using available genomic and transcriptomic references, we identified 25 species growing in our incubations and used those to compare the expression levels of these specific genes.

Conclusions Our results indicate that several peptidases, together with some glycoside hydrolases and glycosyltransferases, are more expressed in phagotrophic than in phototrophic species, and thus could be used to infer the process of bacterivory in natural assemblages.

Keywords Bacterivory, Heterotrophic flagellates, Metatranscriptomics, Functional genes, Unamended incubations, Phagocytosis, Peptidases, Glycosidases

Background

Understanding the activity and functions of microbial communities in the ocean is fundamental to predict how marine ecosystems will be affected in the context of global change [1]. Marine microbes, both prokaryotes and microbial eukaryotes (protists), form the base of marine food webs and alterations in their composition and activities could directly impact biogeochemical cycles at a global scale [2]. Currently, increases in surface seawater temperature are amplifying the stratification of the water column, thus hampering mixing and the delivery of nutrients from the deep ocean to upper layers [3]. These changes are predicted to promote smaller

*Correspondence:
Aleix Obiol
obiol@icm.csic.es
Ramon Massana
ramonm@icm.csic.es
[1] Department of Marine Biology and Oceanography, Institut de Ciències del Mar (ICM-CSIC), Passeig Marítim de la Barceloneta 37-49, Barcelona, Catalonia 08003, Spain
[2] Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czech Republic
[3] Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

microorganisms in the ocean [3, 4] and uncouple bacterial production from grazing mortality [5]. Consequently, bacterial biomass could increase, thus producing an imbalance between carbon recycling and carbon export to the deep ocean. Even though bacterivory is central in marine food webs, both the players and the genes they use are still largely uncharacterized.

The use of multi-omics techniques has completely changed the field of microbial ecology, providing new approaches to study microbial diversity and functions. Given the complexity of eukaryotic genomes and the fact that the presence of an eukaryotic gene informs little about its in situ function, metatranscriptomics has been the preferred approach to study the activity of microbial eukaryotes [6, 7]. These studies can benefit substantially from the generation of reference genomes of dominant marine species, and new genomes from uncultured protists have been recently obtained by metagenomics [8–10] and single-cell genomics [11–14]. Recent studies using metatranscriptomics on protist communities have broadened our knowledge on different topics [15] such as trophic strategies [16], diel and seasonal cycles [17, 18], nutrient responses [19, 20], or functional biogeography [21]. The application of these tools to study bacterivory in the ocean is becoming very promising.

Heterotrophic flagellates (HF) are main bacterial grazers in the ocean [22], and at the same time, the most understudied component of the marine microbiome [23]. These microscopic unpigmented protists (2–5 μm), found in the photic ocean at concentrations up to 10^3 cells mL^{-1} [24], play a crucial role in the microbial loop by channeling carbon to higher trophic levels, remineralizing inorganic nutrients, and keeping bacterial abundances in balance [25, 26]. HFs predate on bacteria through phagocytosis, the uptake of the prey through membrane invagination and its digestion in an acidic environment [27]. Phagocytosis is an ancient evolutionarily conserved process with strong implications in the origin and evolution of eukaryotes [28, 29]. It has been deeply studied in metazoan immunity cells, given its role as a defense mechanism in human immunity [30], but only few studies have investigated the genes for phagocytosis as a nutritional process [12, 31–33]. In one of these, a machine-learning approach using reference genomes was performed to predict a phagotrophic trophic style [33].

Nevertheless, it is yet unclear which functional genes involved in phagocytosis are being expressed by marine HF species during bacterivory, mainly due to the lack of model organisms and the high phylogenetic diversity of HF assemblages. Some recent studies have explored this question by differential gene expression in cultured strains [34, 35]. However, as many of the dominant HF species in the ocean are not available in culture [36], we are still missing a large fraction of the molecular processes involved in marine bacterivory. Unamended incubations have proven to be a good approach to promote a pulse of bacterivory from natural HF assemblages [37], which facilitates a subsequent study of the gene expression of the stimulated uncultured species. We recently used this approach to follow the functional dynamics of a few uncultured MAST species by combining metatranscriptomics and single-cell amplified genomes [38]. Thus, the combination of unamended incubations and metatranscriptomics could allow identifying highly expressed genes by bacterivorous co-existing species.

In the present work, we performed four unamended dark incubations of surface seawater collected in the Blanes Bay Microbial Observatory (BBMO) at different seasons of the year to explore the functional dynamics of the HF community < 3 μm during bacterivorous growth. Our specific objectives were (i) to analyze whether HF communities showed similar expression patterns in different incubations, (ii) to identify the genes that are highly expressed during the process of bacterivory, and (iii) to validate if these could be used to infer bacterivory in natural assemblages. Altogether, our study advances the understanding on the biological processes shaping HF communities and shines a light on the characterization of bacterivory in the ocean.

Methods

Sampling at the BBMO and experimental setup

We performed four incubations: July 2017 ('Jul17'), March 2018 ('Mar18'), November 2018 ('Nov18'), and September 2020 ('Sep20'). For each incubation, surface seawater (see Fig. S1 for a schematic overview) was collected and pre-filtered through a 200-μm nylon mesh at the BBMO, a well-studied coastal sampling station located in the NW Mediterranean Sea [39]. In situ temperature was measured using a CTD probe. Carboys containing the seawater were covered with opaque plastic bags to avoid light penetration and transported to the laboratory in less than 2 h. There, 50 L of seawater was gravity-filtered through 3-μm pore-size polycarbonate filters into a polycarbonate carboy (Nalgene) and incubated in the dark for 5–10 days at in situ seawater temperature. The carboy was gently rolled on the floor once every 24 h to promote water mixing. Once a day, we sampled 2 L of seawater for RNA sequencing and filtered them through 0.6-μm pore-size (47/142 mm ø) polycarbonate filters using a peristaltic pump (~ 10-min filtration time). Samples were stored at − 80ºC after filtration. We also sampled 5 mL of seawater every 12–24 h and fixed them with glutaraldehyde (1% final concentration) for microbial counts. We stained the fixed samples

with 4',6-diamidino-2-phenylindole (DAPI) and filtered through 0.2-µm pore-size (25 mm ø) black polycarbonate filters. We manually counted cell abundances by epifluorescence microscopy of bacteria and heterotrophic flagellates (under UV light), phototrophic flagellates (UV and blue lights), and *Synechococcus* (blue and green lights).

RNA extraction, sequencing, and read analyses

We performed RNA extraction and library preparation for Illumina sequencing as detailed in [38]. Briefly, we cut and vortexed the frozen filters in tubes containing Power Soil beads (Mobio) and extracted RNA using RNeasy Mini Kit (Qiagen) followed by a DNAse treatment with Turbo DNA-free kit (Ambion). Extracts were kept at − 80ºC until processing. We selected 21 samples for sequencing according to their RNA extraction yields, determined using Qubit RNA HS Assay Kit (Thermo Fisher Scientific). Polyadenylated transcripts were reverse transcribed to cDNA and enriched by 15 polymerase chain reaction (PCR) cycles at CNAG (https://cnag.cat/). RNASeq libraries were prepared with KAPA-Stranded mRNA-Seq Illumina (Roche-KAPA Biosystems). Sequencing was carried out in Illumina platforms HiSeq2500 for "Jul17" incubation, HiSeq4000 for "Mar18" and "Nov18" incubations, and NovaSeq6000 for "Sep20" incubation. Paired-end reads (2 × 100 bp) were obtained with a sequencing depth of 15 Gbp, except for "Sep20" incubation, with 25 Gbp.

We trimmed metatranscriptomic Illumina reads for adapters and filtered them for phred scores of ≥ 20 and length ≥ 75 bp with Trimmomatic v0.38 [40] (Fig. S2). In order to characterize the taxonomic dynamics of the incubations, we followed the pipeline described in [41] to extract and classify 18S rRNA fragments from the obtained Illumina reads, using version 5 of eukaryotesV4 database (https://github.com/aleixop/eukaryotesV4).

Assembly, annotation, and quantification of transcripts

We first identified and removed ribosomal RNA fragments from quality-filtered reads using SortMeRNA v3.0.3 [42] with default parameters. We then co-assembled the remaining reads using rnaSPAdes v3.14.1 [43] with default parameters and obtained a single assembled metatranscriptome per incubation (Fig. S2). For each one, we removed sequences shorter than 200 bp with VSEARCH v2.17.0 [44] and kept the longest isoform of each gene. We taxonomically classified the transcripts using Kaiju v1.8.2 [45] in MEM mode with the nr + euk database and parameters '-x -m 11' as in [46] and removed transcripts associated to prokaryotic taxa and viruses. We translated the transcripts to proteins using GeneMarkS-T v5.1 [47] with minimum length of 200 bp and default parameters and removed transcripts that could not be translated to a protein. We quantified the expression of the preliminary obtained transcripts per incubation in each sample using Salmon v1.8.0 [48] in mapping-based mode. We considered a gene as expressed if it had ≥ 2 transcripts per million (TPM) in at least one sample [49] and removed the transcripts below this threshold. With these final metatranscriptomes, we did a second quantification with Salmon and obtained the expression profiles for each sample. We functionally annotated the predicted proteins using eggNOG-mapper v2.1.2 [50, 51]. We performed most functional analyses using KEGG ortholog (KO) annotations [52] and its higher-order associated classifications (BRITE and Pathway). To do so, we generated an expression profile KO table by pooling the TPM values from transcripts associated to each individual KO, each related to a distinct functional gene with all its intraspecific and interspecific variants.

Detection in the incubations of species with known genomic data

We built a reference protein database of 1038 eukaryotic genomes and transcriptomes by combining EukProt database version 3 [53] and 50 single amplified genomes prepared from BBMO samples. In order to detect the presence of these species in our incubations, we aligned the metatranscriptomes to this protein database using DIAMOND blastp v2.0.14 [54] in "sensitive" mode. We kept the top scoring hits for each transcript after selecting alignments with > 90% identity and a minimum of 50 amino acids. We removed transcripts having more than one top hit (same *e* value) with different reference genomes (1.6% of the cases). With this, we identified 51 species that had a reasonable similarity in the incubations and at least 100 transcripts in one incubation. Then, we verified if the signal detected derived from the same exact species (or a closely related one) by mapping the retrieved transcripts to the coding sequences (CDS) of the reference species at the nucleotide level using blastn v2.7.1 [55]. We considered a species present in the incubations when the transcripts and the reference CDS had a median identity > 99%. We selected 25 species, annotated their reference genomes/transcriptomes using eggNOG-mapper and quantified their gene expression by mapping the unassembled metatranscriptomic reads to their CDS sequences using Salmon. We normalized the obtained read counts using the effective length of each mapped CDS sequence (i.e., we divided mapped read counts by the effective length reported by Salmon) and converted them to integer counts ranging from 0 to 10^6 using a pseudo-count as in Salazar et. al [56]. Then, we corrected the obtained expression profiles using TMM transformation and

converted them to pseudocounts per million using edgeR [57]. We assigned trophic mode to selected species according to common assumptions in the literature. Thus, we considered Diatomeae and all Archaeplastida species to be phototrophs, except for Picozoa (heterotrophs) [14]. In groups of pigmented microorganisms known to harbor multiple mixotrophic species (Prymnesiophyceae, Bolidophyceae, and Dictyochophyceae), we considered all species to be mixotrophs [58–60]. We labeled as heterotrophs the rest of groups, as they encompass unpigmented species. For taxa with no detailed species assignment, we used the trophic mode of their closest cultured match in GenBank. ChrysophyceaeNA-sp1 was closest to *Paraphysomonas imperforata* at 99.1% identity (heterotroph), while Micromonas-sp1 was closest to *Micromonas bravo* at 99.8% identity (phototroph).

Statistical analyses

We performed all general analyses using R v4.1.1 [61] with packages tidyverse v1.3.1 [62] and vegan v2.5.7 [63]. We divided the samples of each incubation into "lag", "growth", and "decline" states according to their position in the HF growth curve assessed by microscopy. We validated this classification using plotPCA function in package DESeq2 v1.32.0 [64] using transcript counts obtained by Salmon followed by a variance stabilizing transformation (Fig. S3). From the list of 359 highly expressed genes (i.e., KO entries) at the growth state in the incubations (Table S1), we identified 104 housekeeping genes, later used for normalizing gene expression functions, by choosing ribosomal proteins and other genes generally used in the literature [65, 66].

Results

Growth and taxonomic dynamics

We conducted a total of four unamended incubations between July 2017 and September 2020 following similar experimental procedures (Table 1; Fig. S1). For each one, we followed the growth dynamics of bacteria and picoeukaryotes by epifluorescence microscopy. In all cases, there was an initial peak of bacteria followed by a peak of heterotrophic flagellates (HF) and a continuous decrease of photosynthetic populations (Fig. 1A). Initial bacterial abundances were around $7–9\times10^5$ cells ml^{-1} and peaked to ~ 2×10^6 cells ml^{-1} at 4–5 days in the first three incubations, while in "Sep20" the bacterial peak was earlier and lower. The initial peak of bacteria was followed by a peak of HF, which increased from $5–10\times10^2$ cells ml^{-1} to $7–13\times10^3$ cells ml^{-1}. In the absence of light, the phototrophic flagellates (PF) presented a steady decrease from the initial counts of $1–3\times10^3$ cells ml^{-1}, with the exception of "Mar18" incubation, where initial cell abundances were one order magnitude higher due to a bloom of very small cells (1–2 µm). *Synechococcus* cell abundances steadily decreased from initial abundances about 2×10^4 cells ml^{-1} in the first three incubations and twice these initial values (and a more marked decrease) in "Sep20". Following the HF dynamics in Fig. 1A, we classified the incubations into three states ("lag", "growth", and "decline").

Taxonomic groups with chloroplast-harboring species—Prymnesiophyceae, Dictyochophyceae, MOCH-2, Pelagophyceae, Chlorophyta, and Dinoflagellata—showed a clear decreasing trend along the incubations (bottom panels in Fig. 1B). These accounted for 43–70% of relative read abundance at the beginning of the incubations and were nearly absent towards the end. Dictyochophyceae in "Nov18" was the only exception to this trend, as its relative abundance increased from 1% at around day 4 to 22% at the final time. Heterotrophic protists (upper panels in Fig. 1B) initially represented 17–42% of the relative read abundance and explained most of the read signal during the "growth" state (83–91% total relative read abundance). The most abundant groups considering the 4 incubations were MAST-1 and MAST-7. Chrysophyceae and Telonemia were also very abundant in "Mar18" and "Nov18" incubations (the latter also in "Sep20"). In terms of the overall development of main taxonomic groups,

Table 1 Overview of the incubations

Incubation	Sampling date	Water temp (°C)	metaT samples	Num. transcripts (10^3)	Total size (Mbp)	N50 (bp)	KEGG annotated (%)
Mar18	06/03/2018	14	5	294	187	1112	31.4
Jul17	04/07/2017	24	6	225	162	1438	32.9
Sep20	15/09/2020	24	4	410	201	799	22.4
Nov18	05/11/2018	19	6	350	228	1216	30.6

General information of the incubations and statistics of the final metatranscriptomes

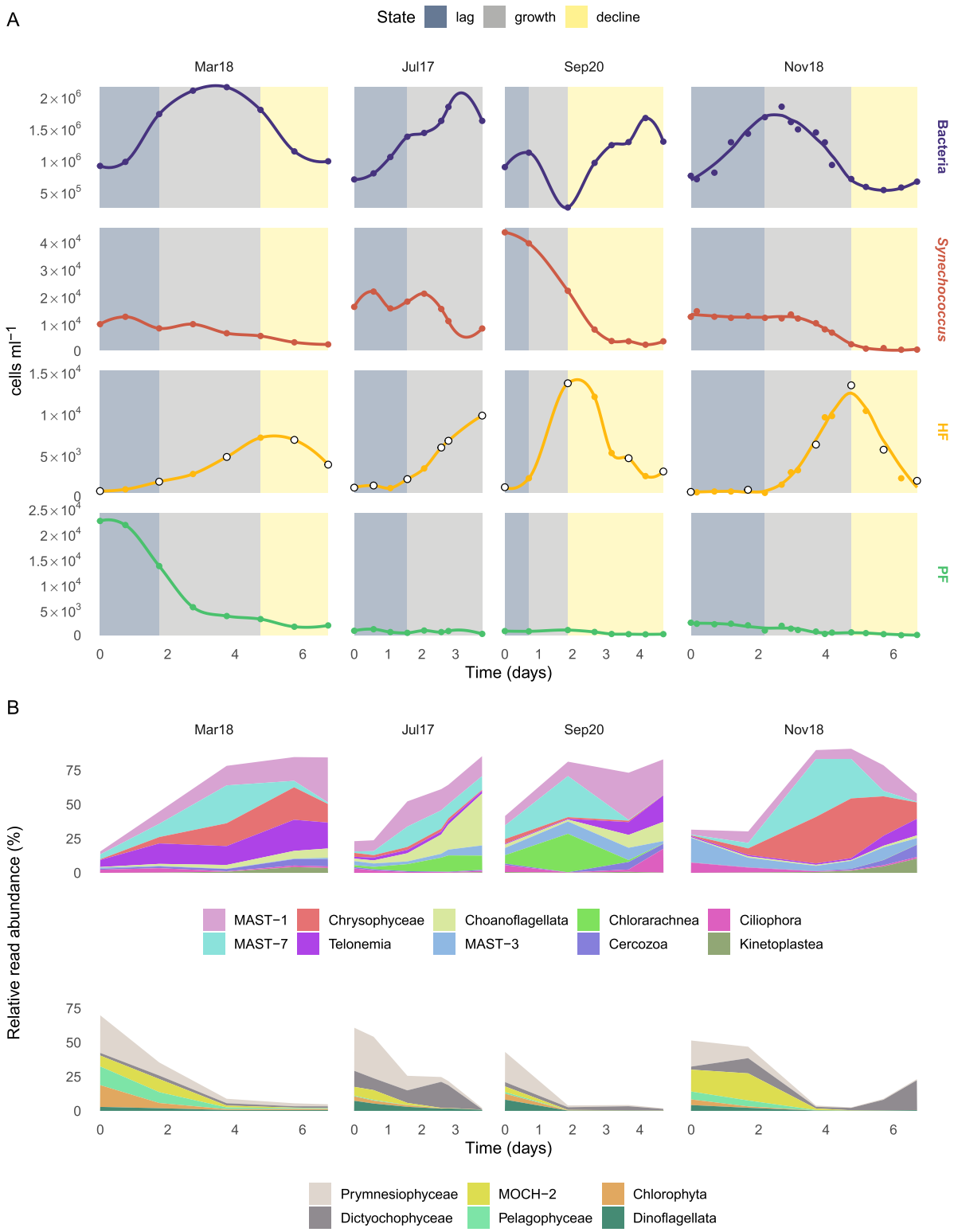


Fig. 1 Cell counts and taxonomic dynamics during the four incubations. **A** Cell counts of heterotrophic bacteria, *Synechococcus*, heterotrophic flagellates (HF), and phototrophic flagellates (PF) conducted by epifluorescence microscopy. The background of the plots is colored by the different incubation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

incubations grouped by pairs: "Jul17" and "Sep20" were closer, as well as "Mar18" and "Nov18".

General functional dynamics

We generated four de novo metatranscriptomes by coassembling reads and curating the transcripts sets (see Fig. S2 for a schematic overview of the process). The final datasets contained $2–4 \times 10^5$ transcripts with a N50 ranging from 799 to 1438 bp (Table 1). Using KEGG database, we could functionally annotate approximately a third of the transcripts (Table 1), which represented around half of total TPM in each incubation (Fig. S4). A PCA plot with the normalized counts for these transcripts roughly validated the three states considered here (Fig. S3).

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

Higher-level KO annotation, as represented by KEGG BRITE categories, showed similar structuring between states with "growth" and "decline" samples clustering together and "lag" samples forming a separate cluster (Fig. 2C). The decrease along incubation time in the expression of photosynthesis-related proteins was very apparent (Fig. 2C), with virtually no expression in "decline" samples. Apart from this clear pattern, the dynamics of the remaining categories were less marked and could be roughly divided into 3 different trends: (i) functions with genes exhibiting high expression through all incubation states, (ii) those showing higher expression in both "growth" and "decline" states, and (iii) those more expressed in the "growth" state. The first group included constitutive processes of the cell, such as spliceosome, translation factors, or chaperones, as well as membrane trafficking processes and transporters. In the second category, ribosome was the most expressed category, followed by ubiquitin, GTP-binding and transcription- and replication-related proteins. The third group contained peptidases, GPI-anchored proteins, and glycosyltransferases.

Most expressed genes during "growth" state

Based on the dynamics of bacteria and HF in all incubations, it could be deduced that the "growth" state was the period when the bacterivory process had the highest relative importance. So, we focused on the most expressed genes (KOs) during that state. We obtained a list of 359 highly expressed KOs (those with > 500 TPM on average in one of the incubations), which we then grouped into a custom system of 24 categories partially based on KEGG BRITE and Pathway hierarchies, assigning each KO to a single category (Table S1) to avoid duplicated signal. These general functions presented comparable gene expression levels when putting together all "growth" samples from the four incubations (Fig. 3A). Some of the most expressed categories were related to processes generally considered as constitutive or housekeeping, as already seen in general dynamics (Fig. 2): ribosomal and cytoskeleton proteins, protein processing, translation, replication, or transcription. However, other categories related to metabolism also emerged, such as peptidases, CAZy enzymes, or other hydrolases, as well as translocases (mainly including proton pumps) and membrane trafficking proteins (Fig. 3A).

Following the data revealed in Fig. 3A and past evidence of their role in phagocytosis (see the "Discussion" section for further details), we analyzed in detail the KOs of three promising functional categories, namely peptidases, translocases (proton pumps), and CAZy enzymes (Fig. 3B). We first assessed the overlap in the functional annotation of transcripts to avoid having different KOs with the same transcript associated. This was particularly critical within peptidases, in which the 20 KOs were grouped in 6 unique categories (Fig. S5). We then computed the average fold change (FC) of the 104 housekeeping genes (HK; Table S1) between "lag" and "growth" states per incubation and used this value as baseline to compare the expression FC of the rest of the genes. All peptidase genes had a higher expression in the "growth" state than HK genes in at least 3 of the 4 incubations (Fig. 3B), suggesting an upregulation during growth by bacterivory, with an average FC > 3 (i.e., their expression increased more than 3 times from "lag" to "growth" states). Among these, cysteine peptidases (including 7 cathepsin types, see Fig. S5) were by far the

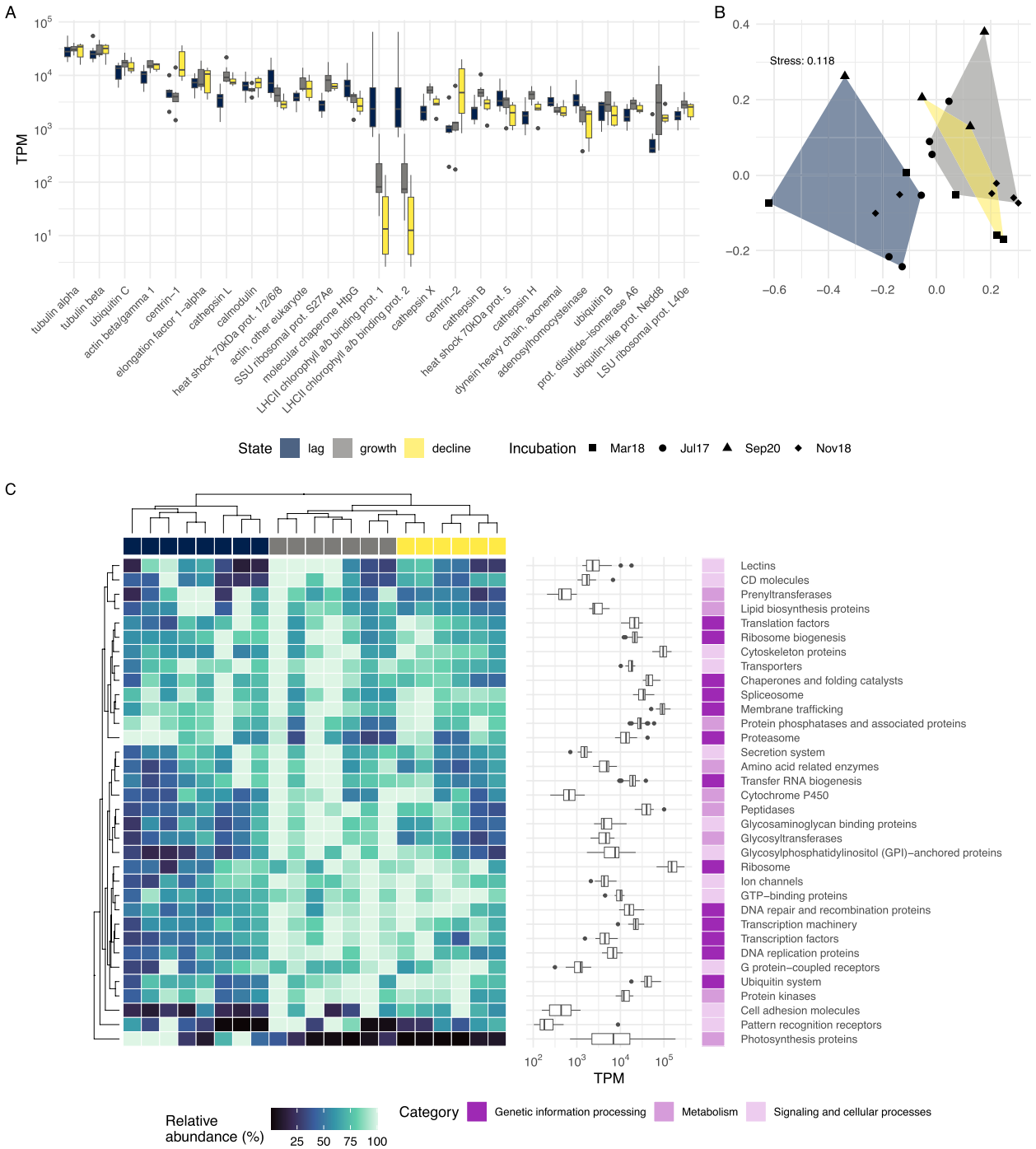


Fig. 2 General functional dynamics of the four incubations. **A** Expression values per sample of the 25 most expressed KOs (KEGG orthologs) in the incubations represented by boxplots colored by incubation state. **B** Non-metric multidimensional scaling (NMDS) plot using Bray–Curtis dissimilarities between the expression of KOs in the different samples. Samples are grouped by their incubation state. **C** Heatmap displaying the expression of main categories as represented by KEGG BRITE classifications in all samples (i.e., each column represents a sample). Values shown are computed by scaling TPM values to a 0–100 scale per category and incubation (i.e., TPM values belonging to a category and incubation are divided by their maximum value). Boxplots display the actual TPM values per sample of each category

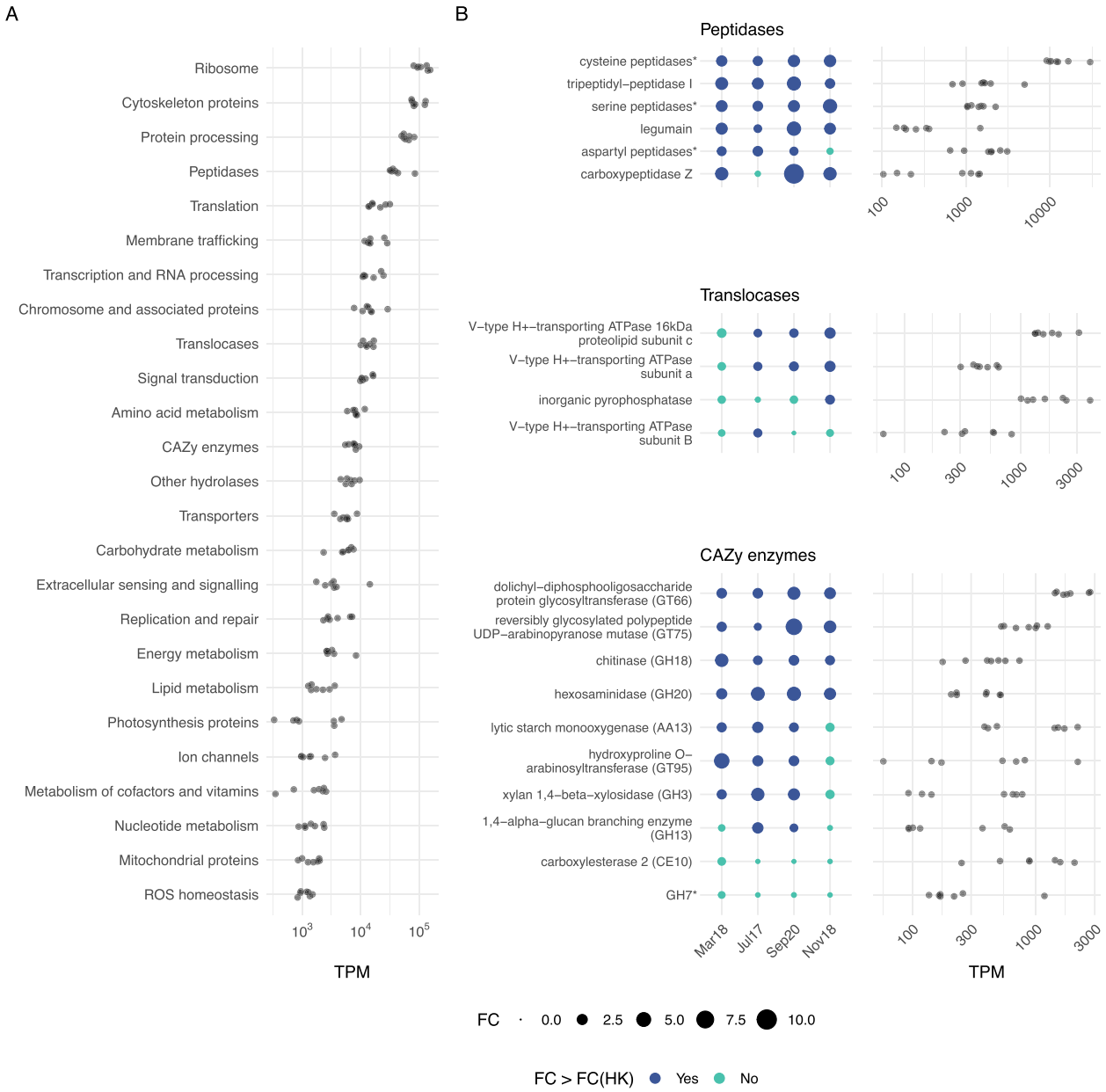


Fig. 3 Most expressed functions and genes in the growth state of the incubations. **A** Expression values (TPM) of the 359 most expressed genes in "growth" samples pooled into custom categories, delineated to report each KO to a single category (see Table S1 for further details). **B** Fold change (FC) between "lag" and "growth" incubation states and TPM expression values of genes (KOs) annotated as peptidases, translocases (proton pumps), and CAZy enzymes. Dots representing FC larger than the average FC of housekeeping genes are colored in dark blue. KOs displaying overlap in functional annotations (i.e., different KOs associated to the same transcript) needed to be grouped into broader sets (asterisk, see Fig. S5 for further details). "Cysteine peptidases" include cathepsins B, F, H, K, L, O, and X, as well as KDEL-tailed endopeptidase and xylem cysteine peptidase; "aspartyl peptidase" includes cathepsins D and E, phytepsin and saccharopepsin; "serine peptidase" includes cathepsin A, serine carboxypeptidase-like clades I and II and vitellogenic carboxypeptidase-like protein. For CAZy enzymes, "GH7" groups cellulose 1,4-beta-cellobiosidase and cellulase

most expressed in "growth" state (Fig. 3B). For translocases, genes related to V-type ATPase displayed different patterns, with two subunits (a and c) being more expressed than HK genes in 3 incubations and subunit B in only one. We also detected this latter trend with inorganic pyrophosphatase, which was one of the most expressed genes related to proton pumps in the incubations (Fig. 3B). For CAZy enzymes, two glycoside hydrolases (GH18 and GH20) and two glycosyltransferases (GT66 and GT75) were more expressed than

HK genes in all the incubations. Carboxylesterase 2 and GH7 genes showed the opposite trend, with a lower FC than HK in all cases. GT66, AA13, and CE10 were the most expressed CAZy genes in "growth" state (Fig. 3B).

Gene expression at the species level

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

Taking advantage of the different trophic modes represented by the 25 species (12 heterotrophs, 8 mixotrophs, and 5 phototrophs), we analyzed whether the expression of the genes identified as putatively relevant for

Table 2 Species with genomic data well represented in the metatranscriptomes

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

The 25 species are displayed with general taxonomy, genome completeness (BUSCO), and quantification information in the incubations. Species marked with an asterisk were retrieved using our own BBMO SAG collection, while the rest were retrieved using EukProt database

bacterivory varied between nutritional strategies. The majority of these genes displayed higher levels of expression in phagotrophs compared to phototrophs (Fig. 5; Table S2). Within peptidases, cysteine peptidases were the ones with the highest relative expression values, with approximately one order of magnitude higher than the

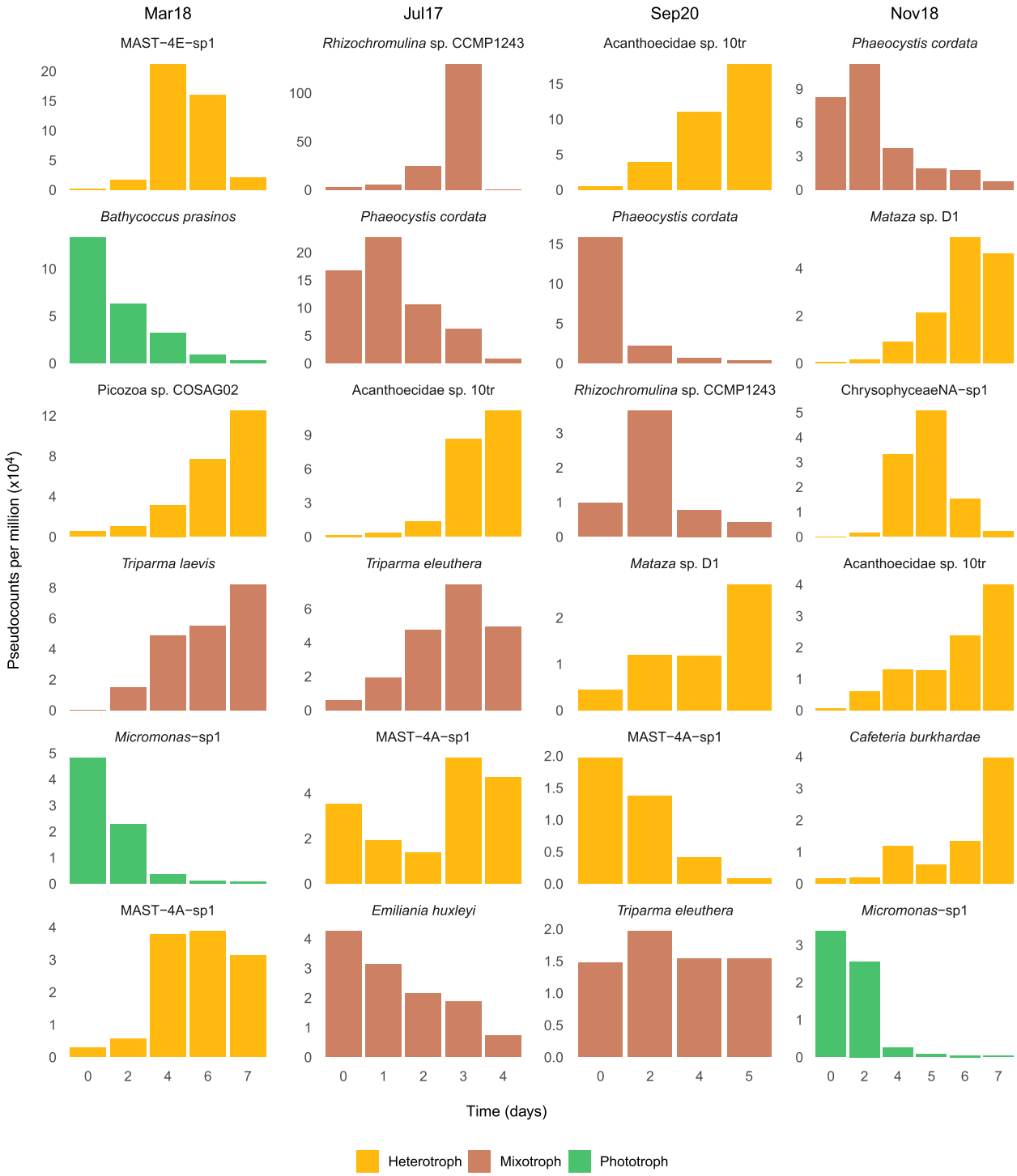


Fig. 4 Expression dynamics of a selection of species with genomic data found in the metatranscriptomes. See the full list of the 25 detected species in Table 2 and the display of their expression dynamics in Fig. S7. Values represent pseudocounts per million, obtained after correcting the abundance profiles by gene length and sequencing depth (see the "Methods" section for details)

rest (Fig. 5). Relative gene expression differed depending on trophic mode, with heterotrophic species always displaying slightly higher values (around 2% of gene expression) than mixotrophic species and markedly higher than phototrophic ones (< 0.5%). The remaining peptidase genes displayed lower expression levels but generally followed a similar trend. In translocases, inorganic pyrophosphatase had the highest levels of relative expression and, together with the V-ATPase subunit a, was more expressed in heterotrophs than in phototrophs (Fig. 5). CAZy enzymes showed different trends. GH13 seemed to be more expressed in phototrophs, CE10 in mixotrophs, and GT66, GH18, GH20, AA13, GT95, and GH3 in heterotrophs. As revealed by functional annotations, genes encoding enzymes CE10 and GH7 were not present in genomes/transcriptomes from the selected phototrophic species, and AA13 had a single sequence annotated in *Pycnococcus provasolii*. Therefore, these did not display any relative expression value for phototrophs (Fig. 5). We selected the best candidates as marker genes for bacterivory by combining the results of a randomization test with 10,000 permutations (p value < 0.05), and the results of fold change reported in the previous section (Fig. 5; Table S2).

Discussion

Marine heterotrophic flagellates (HF) remain largely undersampled, and very few ecologically relevant species are available in culture [23]. As a consequence, key processes in global biogeochemical cycles, such as the

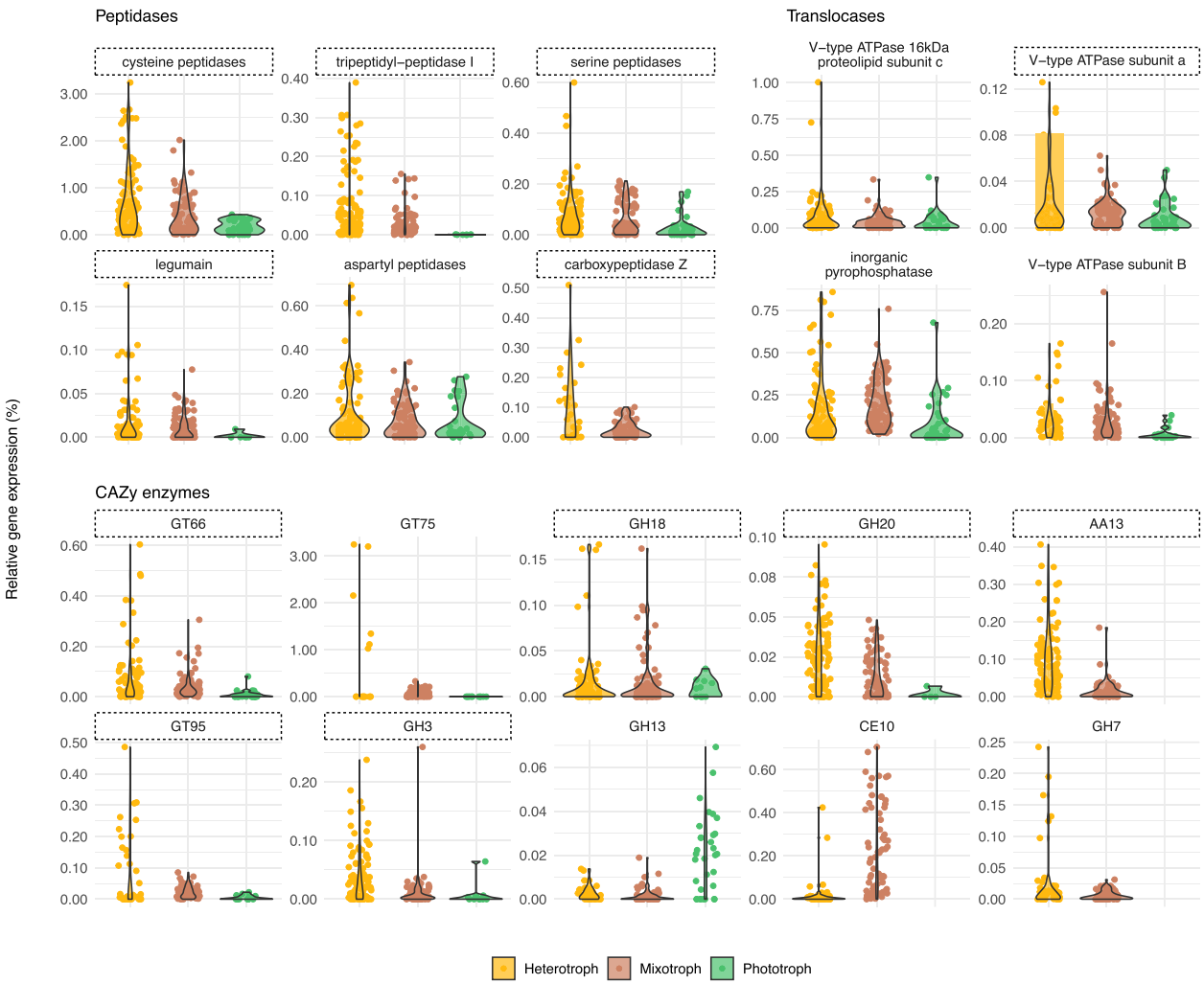


Fig. 5 Expression of selected genes in species with different trophic modes. Points represent the relative expression of the gene in a single species and sample. Values were computed by dividing the expression of the selected gene by the total expression for each species and sample. Values are separated by the trophic mode of the species they come from (Table 2). Genes within dashed rectangles have a higher relative expression in phagotrophs than in phototrophs and could be good candidates to infer bacterivory in natural assemblages (see Table S2 for more details)

grazing on marine bacteria by small predators, still need to be well characterized. In this study, we performed four unamended incubations of coastal seawater in the dark to promote the growth of natural HF assemblages and assess their gene expression during bacterivory.

Circumventing the lack of cultured representative HF species

The use of unamended incubations allowed a more than tenfold increase in cell densities of natural HF assemblages. Even after performing a polyA selection towards eukaryotic messenger RNA, the obtained metatranscriptomes contained a remarkable ribosomal RNA signal (around 3% of reads on average matched the V4 region of the 18S rDNA) due to its huge abundance in the cell [67]. We took advantage of this to assess the general taxonomic dynamics in the incubations, an approach supported by a previous study where we reported that the relative abundance of this rRNA signal was well correlated with the FISH counts of the target cells [38]. The most represented taxa belonged to MAST clades, Chrysophyceae and Telonemia, groups that have been identified to be highly abundant and widespread in the surface ocean [36]. Despite the general dominance of these taxonomic groups, differences emerged between the individual species growing in the incubations, thus highlighting the seasonality of protist communities in BBMO [68] and the large diversity of natural assemblages.

Most of the groups present in the incubations are known to be poorly represented in public databases, so we opted for a dual strategy to analyze our dataset. First, we performed de novo assemblies to analyze the expression patterns of the whole community. We could functionally annotate half of the transcripts in the assemblies, while the other half remained completely unknown. This issue was also reported in a global eukaryotic metatranscriptomic survey [21] and agrees with the estimation that currently 40–60% of microbial predicted genes cannot be assigned to a known function [69]. Second, we used taxonomic binning, as carried out in other studies [19, 70], to answer specific questions on some bacterivory-related genes. Although less than 15% of metatranscriptomic reads mapped to our reference genome dataset with > 90% identity (Fig. S8), this approach revealed interesting expression insights in relevant species-specific functional genes.

The general functional dynamics of the 4 incubations showed seemingly similar patterns of gene expression that could reflect some kind of functional redundancy, by which different sets of co-existing taxa are performing the same function in the ecosystem [71]. The most visible trend in functional dynamics was the expression decrease of photosynthesis-related genes to virtually zero, in agreement with the observed decrease of phototrophic flagellates and the fact that incubations were performed in the dark. Despite representing an obvious outcome of the experimental setup, this transition from phototrophy to heterotrophy in all the incubations highlights the reliability of our data to analyze the expression of bacterivory-related genes.

Highly expressed genes during bacterivory

Some of the most expressed categories in "growth" samples (where bacterivory should be highest) belonged to constitutive processes, with genes involved in several cellular functions (like actin or tubulin), and we did not analyze them any further. Instead, we focused on other sets of highly expressed genes such as peptidases, translocases, and CAZy enzymes, as these could be promising targets in the study of bacterivory.

The majority of the highly expressed peptidase genes belonged to cathepsins. According to the overlap displayed in functional annotations, these clustered into separate groups by their catalytic type (cysteine, serine, and aspartyl peptidases). An exhaustive analysis should be performed to identify the phylogenetic relationships within these groups, as most of them were described in humans or model organisms [72, 73] and little is known for uncultured protists species. Cathepsins are mainly found in the lysosome, where they act as digestive enzymes degrading proteins in acidic conditions [74]. Their high gene expression during bacterivory is consistent with the fact that more than 60% of bacterial dry weight is composed by proteins [75]. Some peptidases have been previously found in phagosomes, such as cathepsin L in metazoan macrophages [76] or cathepsin D in the amoeba *Dictyostelium discoideum* [77]. A recent paper working with the bicosoecid flagellate *Cafeteria burkhardae* detected an abundant cysteine peptidase gene being differentially expressed during bacterivory [34], and other studies have reported their presence in mixotrophic algae [78], mixotrophic dinoflagellates [79], and MAST groups [38].

Within translocases, V-type ATPase and inorganic pyrophosphatase genes were among those highly-expressed in "growth" samples. V-type ATPase proton pumps are involved in the acidification of the lysosome, among other organelles [80], and it has been hypothesized that pyrophosphatase could play a similar role in protists not belonging to Opisthokonta and Amoebozoa supergroups [12, 81], where this protein is absent. The gene for this proton pump was more expressed than the V-type ATPase in the abovementioned *C. burkhardae* study, thus supporting this view [34]. We did not find a clear pattern of upregulation of the pyrophosphatase

proton pump between "growth" and "lag" states, whereas V-ATPases-related genes showed contrasting trends depending on the targeted subunits. These belonged to different domains of the proton pump [80] with distinct roles (ATP hidrolysis in subunit B and proton translocation in subunits a and c) that might explain the observed patterns.

Carbohydrate-active (CAZy) enzymes are related to carbohydrate metabolism [82], including glycoside hydrolases (GH) and glycosyltransferases (GT). GTs create glycosidic bonds with a variety of organic substrates [83], although some of the reactions they catalyze can be reversible [84]. Actually, some GTs have been reported to have distinct enzymatic activities (hydrolysis among them) with differing pH optima [85]. GT66 gene, involved in the *N*-glycosylation of proteins [86]—a highly conserved metabolic process obligatory for viability in eukaryotes [87]—as well as GT75 gene, which can be associated to cell wall metabolism [88], were upregulated in "growth" state. The upregulation of GT66 entails an increase in protein glycosylation by *N*-glycans, which, among various functions, are involved in the functionality of some lysosomal proteins [89]. The high expression of GT75 could be due to either a biosynthetic pathway or an undescribed degradative process. In the case of GHs, they are responsible for carbohydrate hydrolysis [83], and they could be involved in bacterivory as digestive enzymes. It was recently reported that GHs account on average for 3% of predicted genes in four MAST-4 species [13], which are among the most abundant HFs in the ocean [36]. A gene encoding a chitinase belonging to GH18 was the most expressed GH in "growth" samples. Chitin is mostly found in metazoans, fungi, and diatoms [90], and the presence of chitinase in pico-sized HFs indicates that these enzymes may have other physiological functions that are still unknown [91]. Other highly expressed GH genes in our incubations were hexosaminidase (GH20) and xylan 1,4 – beta – xylosidase (GH3). Hexosaminidases are abundant components found in phagosomes of the parasite *Entamoeba histolytica* [32] and several GH20 genes were found in the genomes of MAST-4 and MAST-3 species [92]. Xylan is an important component in plant cell walls [93], and the enzyme 1,4 – beta – xylosidase is relevant to the production of biofuels [94]. Xylan also occurs in some red and green algae [95] and the high expression of GH3 genes in the incubations could be due to microscopic algae being ingested in our incubations. A CAZy enzyme not belonging to GH or GT, the lytic starch monooxygenase (AA13), also displayed high levels of expression. This enzyme was recently discovered [96] and is related to the degradative activity of recalcitrant polysaccharides [97].

Only three of the highly expressed genes reported here—cellulase (GH7), 1,4 – beta – xylosidase (GH3) and tripeptidyl-peptidase I—were present in the set of 474 genes predicting phagotrophic trophic mode in a previous comparative genomics model [33]. Considering that this model was built using reference genomes, differences between the approaches were expected, and this highlights the importance of adding expression data in the search of marker genes for phagocytosis.

Expression of bacterivory genes at the species level

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

With the expression data retrieved from the 25 species we could test whether selected peptidases, translocases, and CAZy genes represented a larger share of the species expression in bacterivorous species compared to phototrophic ones, and thus could be related to the process of bacterivory. Given that phototrophs may have been in suboptimal conditions under dark conditions, one could argue that this analysis could be biased towards heterotrophic modes. However, we did not detect strong differences in relative expression for phototrophic species between initial samples, where PF abundance was highest and the absence of light was just starting, and the rest of the incubations (Fig. S10). Transcriptomic studies in phototrophic species growing in optimal light conditions need to be done to further validate their low expression of bacterivory genes.

Our analysis revealed that several peptidase genes were generally more expressed in species having a phagotrophic mode of nutrition (heterotrophs and mixotrophs), with cysteine peptidases representing up to 3% of the total gene expression of some species. This suggests that, although these could participate in other cellular processes, they have a key role in bacterivory as digestive enzymes. Proton pumps (translocases) did not display these marked differences between trophic modes, except for one V-ATPase subunit. For CAZy enzymes, GT66, GH18, GH20, AA13, GT95, and GH3 were the genes displaying the most marked differences of relative expression between phagotrophic and phototrophic species. Thus, these could be key players in bacterivory, with a role in the digestion of ingested bacteria in the phagolysosome. Although the transcripts associated to the 25 identified species only represent 2.5% of the total signal in the 4 metatranscriptomes (data not shown), this analysis represents a proof-of-concept of what can be achieved with more reference genomes from species that are currently uncultured.

Conclusions

Using a combination of incubations to promote the growth of natural HF assemblages and metatranscriptomic data, we obtained high-quality gene expression data related to bacterivory, a critical process in global biogeochemical cycles that still needs to be well characterized. Our results indicate that different HF communities follow similar functional profiles and that genes encoding peptidases and CAZy enzymes are highly expressed during the process of bacterivory. A more detailed analysis at the species level revealed that among others, cysteine peptidases, together with some glycoside hydrolases and glycosyltransferases, are key players in this process and could be used to infer bacterivory in natural assemblages. A further study of these genes and their expression in multiple and diverse taxa could reveal the intricate mechanisms of marine bacterivory, which in turn could improve predictions on the alteration of marine ecosystems due to global change.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40168-023-01571-5.

Additional file 1: Fig. S1. Overview of the experimental setup. **Fig. S2.** Overview of the bioinformatic processing of the metatranscriptomic reads. **Fig. S3.** PCA plots used for validation of the different incubation states based on normalized read counts per transcript. Each point represents a sample, and the value next to it represents the time of incubation expressed as days. **Fig. S4.** (A) Percentage of transcripts annotated with different databases. (B) Percentage of total TPM explained by the transcripts annotated with KEGG. **Fig. S5.** Overlap in functional KEGG annotations in three groups of peptidases (cysteine, aspartyl and serine peptidases) using the whole metatranscriptomic dataset. Values represent the percentage of overlap for each KO pair computed as shared transcripts (i.e., transcripts annotated with both KOs) divided by the total transcripts of the KO displayed in the y-axis. Abbreviations: c (cysteine); s (serine); cp (carboxypeptidase). **Fig. S6.** Nucleotide identity of transcripts associated to 51 represented species in the metatranscriptomes. We selected species having a median identity higher than 99% (green bars), while we discarded the rest (red bars). **Fig. S7.** Expression dynamics of the 25 species with genomic data found in the metatranscriptomes. Abundance values represent pseudocounts per million, obtained after correcting the abundance profiles by gene lengths and sequencing depth (see Methods for details). Note that some species appear in several incubations. **Fig. S8.** Per-sample summary of the mapping of the unassembled metatranscriptomic reads to the database EukProt+SAGs using DIAMOND blastx. Only alignments with >90% query coverage are considered. **Fig. S9.** Histogram of the CDS identity between the genes of MAST-4A and MAST-4B species available in EukProt v3. Nucleotide percentage identities were calculated using blastn. **Fig. S10.** Expression of selected genes in species with different trophic modes separated by the state of the incubations. Points represent the relative expression of the gene in a single species and sample. Values were computed by dividing the expression of the selected gene by the total expression for each species and sample. Values are separated by the trophic mode of the species they come from (Table 2).

Additional file 2: Table S1. List of the 359 most expressed genes (KO, KEGG Orthologs) in "growth" samples and general gene expression statistics associated.

Additional file 3: Table S2. Statistical test supporting the selection of candidate genes related to bacterivory. The two first columns show the number of gene expression observations for the 12 heterotrophic and 5 phototrophic species in all possible samples (dots in Fig. 5). For genes with at least 15 observations in the two groups, we performed a randomization test with 10000 permutations to test whether each gene was more expressed in heterotrophs than in phototrophs. In addition, when there was a much larger number of observations in heterotrophs than in phototrophs, the gene was considered to be more expressed in heterotrophs. In this selection, we also included the patterns of overexpression reported in Fig. 3 displayed in the 'FC > FC(HK)' column ('Yes' represents overexpression in at least 3 out of 4 incubations). The data leading to the negative selection form this list are marked in red.

Acknowledgements

Bioinformatic analyses were performed at the Marbits (ICM-CSIC; https://marbits.icm.csic.es), FinisTerrae II (CESGA; https://www.cesga.es/), and Mole platforms (https://www.paru.cas.cz/).

Authors' contributions

AO, IF and RM designed the study. AO, DL, IF, ES, DV and RM performed the incubations and collected the samples. AO and IF performed the microscopic counts. AO and DL extracted the RNA and organized its sequencing. AO, DL, EDS, MMW, MK and RM analyzed the data and interpreted the results. AO and RM wrote the manuscript. The authors read and approved the final manuscript.

Funding

Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research was supported by the Spanish Ministry of Science and Innovation projects ALLFLAGS (CTM2016-75083-R), DIVAS (PID2019-108457RB-I00) and the "Severo Ochoa Centre of Excellence" accreditation (CEX2019-000928-S), and the European Union projects SINGEK (H2020-MSCA-ITN-2015–675752) and GENEXLAB (EUR2022-134047, MICINN). AO was supported by a Spanish FPI grant. E.D.S. was supported by the MSCA-IF SMART (CZ.02.2.69/0.0/0.0/20_079/0017809).

Availability of data and materials

Raw data for "Jul17" incubation were already published at the NCBI BioSample database with accession number SAMN11783926 [38]. For the rest of the incubations, raw data are deposited at NCBI with accession number PRJNA973582. Assemblies, quantification, and functional annotation tables are available at FigShare (https://doi.org/10.6084/m9.figshare.22801697), and all code used for data processing and analyses is available at GitHub (https://github.com/aleixop/metaT_bacterivory).

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 15 November 2022 Accepted: 12 May 2023
Published online: 15 June 2023

References

1. Cavicchioli R, Ripple WJ, Timmis KN, Azam F, Bakken LR, Baylis M, et al. Scientists' warning to humanity: microorganisms and climate change. Nat Rev Microbiol. 2019;17:569–86. https://doi.org/10.1038/s41579-019-0222-5. Springer US.
2. Hutchins DA, Fu F. Microorganisms and ocean global change. Nat Microbiol. 2017;2:17058 Macmillan Publishers Limited.
3. Li WKW, McLaughlin FA, Lovejoy C, Carmack EC. Smallest algae thrive as the arctic ocean freshens. Science. 2009;326:539–539 American Association for the Advancement of Science.
4. Daufresne M, Lengfellner K, Sommer U. Global warming benefits the small in aquatic ecosystems. Proc Natl Acad Sci. 2009;106:12788–93.
5. Sarmento H, Montoya JM, Vázquez-Domínguez E, Vaqué D, Gasol JM. Warming effects on marine microbial food web processes: how far can we go when it comes to predictions? Philos Trans R Soc Lond B Biol Sci. 2010;365:2137–49.
6. 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. 2016;15:6–20 Nature Publishing Group.
7. Keeling PJ, del Campo J. Marine protists are not just big bacteria. Curr Biol. 2017;27:R541–9 Elsevier Ltd.
8. Alexander H, Hu SK, Krinos AI, Pachiadaki M, Tully BJ, Neely CJ, et al. Eukaryotic genomes from a global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton. bioRxiv. 2022:2021.07.25.453713. Available from: https://www.biorxiv.org/content/10.1101/2021.07.25.453713v2 [cited 2 Jul 2022].
9. Delmont TO, Gaia M, Hinsinger DD, Frémont P, Vanni C, Fernandez-Guerra A, et al. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. Cell Genomics. 2022;2:100123.
10. Duncan A, Barry K, Daum C, Eloe-Fadrosh E, Roux S, Schmidt K, et al. Metagenome-assembled genomes of phytoplankton microbiomes from the Arctic and Atlantic Oceans. Microbiome. 2022;10:67.
11. Gawryluk RMR, del Campo J, Okamoto N, Strassert JFH, Lukeš J, Richards TA, et al. Morphological identification and single-cell genomics of marine diplonemids. Curr Biol. 2016;26:3053–9.
12. Labarre A, López-Escardó D, Latorre F, Leonard G, Bucchini F, Obiol A, et al. Comparative genomics reveals new functional insights in uncultured MAST species. ISME J. 2021;15:1767–81 Nature Publishing Group.
13. Latorre F, Deutschmann IM, Labarre A, Obiol A, Krabberød AK, Pelletier E, et al. Niche adaptation promoted the evolutionary diversification of tiny ocean predators. PNAS. National Academy of Sciences; 2021;118. Available from: https://www.pnas.org/content/118/25/e2020955118 [cited 22 Jun 2021].
14. Schön ME, Zlatogursky VV, Singh RP, Poirier C, Wilken S, Mathur V, et al. Single cell genomics reveals plastid-lacking Picozoa are close relatives of red algae. Nat Commun. 2021;12:6651.
15. Cohen NR, Alexander H, Krinos AI, Hu SK, Lampe RH. Marine microeukaryote metatranscriptomics: sample processing and bioinformatic workflow recommendations for ecological applications. Front Marine Sci. 2022;9:858 Available from: https://www.frontiersin.org/article/10.3389/fmars.2022.867007 [cited 2022 Jun 30].
16. Lambert BS, Groussman RD, Schatz MJ, Coesel SN, Durham BP, Alverson AJ, et al. The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. PNAS. National Academy of Sciences; 2022;119. Available from: https://www.pnas.org/content/119/7/e2100916119 [cited 15 Feb 2022].
17. Kolody BC, McCrow JP, Allen LZ, Aylward FO, Fontanez KM, Moustafa A, et al. Diel transcriptional response of a California current plankton microbiome to light, low iron, and enduring viral infection. ISME J. 2019;13:2817–33. https://doi.org/10.1038/s41396-019-0472-2. Springer US.
18. Louyakis AS, Gourlé H, Casaburi G, Bonjawo RME, Duscher AA, Foster JS. A year in the life of a thrombolite: comparative metatranscriptomics reveals dynamic metabolic changes over diel and seasonal cycles. Environ Microbiol. 2018;20:842–61.
19. Alexander H, Jenkins BD, Rynearson TA, Dyhrman ST. Metatranscriptome analyses indicate resource partitioning between diatoms in the field. Proc Natl Acad Sci U S A. 2015;112:E2182–90.
20. Muratore D, Boysen AK, Harke MJ, Becker KW, Casey JR, Coesel SN, et al. Complex marine microbial communities partition metabolism of scarce resources over the diel cycle. Nat Ecol Evol. 2022;6:218–29 Nature Publishing Group.
21. Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, et al. A global ocean atlas of eukaryotic genes. Nat Commun. 2018;9:373.
22. Fenchel T. The ecology of heterotrophic microflagellates. In: Marshall KC, editor. Advances in Microbial Ecology. Boston: Springer US; 1986. p. 57–97 Available from: http://link.springer.com/10.1007/978-1-4757-0611-6_2.
23. del Campo J, Sieracki ME, Molestina R, Keeling PJ, Massana R, Ruiz-Trillo I. The others: our biased perspective of eukaryotic genomes. Trends Ecol Evol. 2014;29:252–9.
24. Jürgens K, Massana R. Protistan grazing on marine bacterioplankton. In: Kirchman DL, editor. Microbial ecology of the oceans. 2nd ed. Hoboken: Wiley; 2008. p. 383–441 Available from: http://doi.wiley.com/10.1002/9780470281840.ch11.
25. Pernthaler J. Predation on prokaryotes in the water column and its ecological implications. Nat Rev Microbiol. 2005;3:537–46.
26. Sherr EB, Sherr BF. Significance of predation by protists in aquatic microbial food webs. Antonie Van Leeuwenhoek. 2002;81:293–308.
27. Botelho RJ, Grinstein S. Phagocytosis. Curr Biol. 2011;21:R533–8.
28. Mills DB. The origin of phagocytosis in Earth history. Interface Focus. 2020;10:20200019.
29. 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 John Wiley & Sons, Ltd.
30. Flannagan RS, Cosío G, Grinstein S. Antimicrobial mechanisms of phagocytes and bacterial evasion strategies. Nat Rev Microbiol. 2009;7:355–66 Nature Publishing Group.
31. 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 Available from: https://www.sciencedirect.com/science/article/pii/S1937644808012069 [cited 2021 Apr 14]. Academic Press.
32. Okada M, Huston CD, Mann BJ, Petri WA, Kita K, Nozaki T. Proteomic analysis of phagocytosis in the Enteric protozoan parasite Entamoeba histolytica. Eukaryotic Cell. 2005;4:827–31 American Society for Microbiology Journals.
33. 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 Springer US.
34. 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. 2021;15:154–67 Nature Publishing Group.
35. Prokopchuk G, Korytář T, Juricová V, Majstorović J, Horák A, Šimek K, et al. Trophic flexibility of marine diplonemids - switching from osmotrophy to bacterivory. ISME J. 2022;16:1409–19.
36. Obiol A, Muhovic I, Massana R. Oceanic heterotrophic flagellates are dominated by a few widespread taxa. Limnol Oceanogr. 2021;66:4240–53.
37. Massana R, Guillou L, Terrado R, Forn I, Pedrós-Alió C. Growth of uncultured heterotrophic flagellates in unamended seawater incubations. Aquat Microb Ecol. 2006;45:171–80.

38. Labarre A, Obiol A, Wilken S, Forn I, Massana R. Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates. Limnol Oceanogr. 2020;65:lno.11379.
39. Gasol JM, Cardelús C, Morán XAG, Balagué V, Forn I, Marrasé C, et al. Seasonal patterns in phytoplankton photosynthetic parameters and primary production at a coastal NW Mediterranean site. Sci Mar. 2016;80:63–77.
40. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
41. 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;20:718–31.
42. Kopylova E, Noé L, Touzet H. SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28:3211–7.
43. Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience. 2019;8:100. https://doi.org/10.1093/gigascience/giz100. [cited 2021 Jan 22].
44. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584 PeerJ Inc.
45. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016;7:11257 Nature Publishing Group.
46. Bucchini F, Del Cortona A, Kreft Ł, Botzki A, Van Bel M, Vandepoele K. TRAPID 2.0: a web application for taxonomic and functional analysis of de novo transcriptomes. Nucleic Acids Res. 2021;49:101. https://doi.org/10.1093/nar/gkab565. [cited 2021 Aug 4].
47. Tang S, Lomsadze A, Borodovsky M. Identification of protein coding regions in RNA transcripts. Nucleic Acids Res. 2015;43:e78.
48. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9 Nature Publishing Group.
49. Mika K, Marinić M, Singh M, Muter J, Brosens JJ, Lynch VJ. Evolutionary transcriptomics implicates new genes and pathways in human pregnancy and adverse pregnancy outcomes. eLife. 2021;10:e69584 Rokas A, Perry GH, Stevens A, Wildman DE, Mesiano S, editors eLife Sciences Publications, Ltd.
50. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol. 2021;38:5825–9.
51. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 50: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47:D309-14.
52. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021;49:D545–51.
53. Richter DJ, Berney C, Strassert JFH, Poh Y-P, Herman EK, Muñoz-Gómez SA, et al. EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotes. Peer Community Journal. 2. https://doi.org/10.24072/pcjournal.173.
54. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18:366–8.
55. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
56. Salazar G, Paoli L, Alberti A, Huerta-Cepas J, Ruscheweyh HJ, Cuenca M, et al. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. Cell. 2019;179:1068-1083.e21.
57. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009;26:139–40.
58. Wilken S, Yung CCM, Hamilton M, Hoadley K, Nzongo J, Eckmann C, et al. The need to account for cell biology in characterizing predatory mixotrophs in aquatic environments. Philos Trans R Soc Lond Biol Sci. 2019;374:20190090.
59. Li Q, Edwards KF, Schvarcz CR, Selph KE, Steward GF. Plasticity in the grazing ecophysiology of Florenciella (Dichtyochophyceae), a mixotrophic nanoflagellate that consumes Prochlorococcus and other bacteria. Limnol Oceanogr. 2021;66:47–60.
60. Li Q, Edwards KF, Schvarcz CR, Steward GF. Broad phylogenetic and functional diversity among mixotrophic consumers of Prochlorococcus. ISME J. 2022;16:1557–69 Nature Publishing Group.
61. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/.
62. Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4:1686.
63. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: community ecology package. 2020. Available from: https://CRAN.R-project.org/package=vegan.
64. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
65. Alexander H, Jenkins B, Rynearson T, Saito M, Mercier M, Dyhrman S. Identifying reference genes with stable expression from high throughput sequence data. Front Microbiol. 2012;3:385 Available from: https://www.frontiersin.org/article/10.3389/fmicb.2012.00385 [cited 2022 May 19].
66. Li Z, Zhang Y, Li W, Irwin AJ, Finkel ZV. Conservation and architecture of housekeeping genes in the model marine diatom Thalassiosira pseudonana. New Phytol. 2022;234:1363–76.
67. Cui P, Lin Q, Ding F, Xin C, Gong W, Zhang L, et al. A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing. Genomics. 2010;96:259–65.
68. Giner CR, Balagué V, Krabberød AK, Ferrera I, Reñé A, Garcés E, et al. Quantifying long-term recurrence in planktonic microbial eukaryotes. Mol Ecol. 2019;28:923–35.
69. Vanni C, Schechter MS, Acinas SG, Barberán A, Buttigieg PL, Casamayor EO, et al. Unifying the known and unknown microbial coding sequence space. eLife. 2022;11:e67667 Brown CT, Storz G, Brown CT, Smith B, editors eLife Sciences Publications, Ltd.
70. Metegnier G, Paulino S, Ramond P, Siano R, Sourisseau M, Destombe C, et al. Species specific gene expression dynamics during harmful algal blooms. Sci Rep. 2020;10:6182 Nature Publishing Group.
71. Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, O'Connor MI, et al. Function and functional redundancy in microbial systems. Nat Ecol Evol. 2018;2:936–43 Nature Publishing Group.
72. Ritonja A, Popović T, Kotnik M, Machleidt W, Turk V. Amino acid sequences of the human kidney cathepsins H and L. FEBS Lett. 1988;228:341–5.
73. Berti PJ, Storer AC. Alignment/phylogeny of the papain superfamily of cysteine proteases. J Mol Biol. 1995;246:273–83.
74. Turk V, Stoka V, Vasiljeva O, Renko M, Sun T, Turk B, et al. Cysteine cathepsins: From structure, function and regulation to new frontiers. Biochim Biophys Acta Proteins Proteom. 2012;1824:68–88.
75. Simon M, Azam F. Protein content and protein synthesis rates of planktonic marine bacteria. Mar Ecol Prog Ser. 1989;51:201–13.
76. Szulc-Dąbrowska L, Bossowska-Nowicka M, Struzik J, Toka FN. Cathepsins in bacteria-macrophage interaction: defenders or victims of circumstance? Front Cell Infect Microbiol. 2020;10:601072.
77. Gotthardt D, Warnatz HJ, Henschel O, Brückert F, Schleicher M, Soldati T. High-resolution dissection of phagosome maturation reveals distinct membrane trafficking phases. Mol Biol Cell. 2002;13:3508–20.
78. McKie-Krisberg ZM, Sanders RW, Gast RJ. Evaluation of mixotrophy-associated gene expression in two species of polar marine algae. Front Mar Sci. 2018;5:273 Available from: https://www.frontiersin.org/articles/10.3389/fmars.2018.00273/full [cited 2021 Mar 25]Frontiers.
79. Cohen NR, McIlvin MR, Moran DM, Held NA, Saunders JK, Hawco NJ, et al. Dinoflagellates alter their carbon and nutrient metabolic strategies across environmental gradients in the central Pacific Ocean. Nat Microbiol. 2021;6:173–86 Nature Publishing Group.
80. Nelson N, Perzov N, Cohen A, Hagai K, Padler V, Nelson H. The cellular biology of proton-motive force generation by V-ATPases. J Exp Biol. 2000;203:89–95.
81. Baltscheffsky M, Schultz A, Baltscheffsky H. H+-PPases: a tightly membrane-bound family. FEBS Lett. 1999;457:527–33.
82. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42:D490–5.
83. Bourne Y, Henrissat B. Glycoside hydrolases and glycosyltransferases: families and functional modules. Curr Opin Struct Biol. 2001;11:593–600.
84. Zhang C, Griffith BR, Fu Q, Albermann C, Fu X, Lee I-K, et al. Exploiting the reversibility of natural product glycosyltransferase-catalyzed reactions.

Science. 2006;313:1291–4 American Association for the Advancement of Science.
85. Lairson LL, Henrissat B, Davies GJ, Withers SG. Glycosyltransferases: structures, functions, and mechanisms. Annu Rev Biochem. 2008;77:521–55.
86. Henrissat B, Sulzenbacher G, Bourne Y. Glycosyltransferases, glycoside hydrolases: surprise, surprise! Curr Opin Struct Biol. 2008;18:527–33.
87. Kukuruzinska MA, Lennon K. Protein N-glycosylation: molecular genetics and functional significance. Crit Rev Oral Biol Med. 1998;9:415–48 SAGE Publications Inc.
88. Popper ZA, Michel G, Hervé C, Domozych DS, Willats WGT, Tuohy MG, et al. Evolution and diversity of plant cell walls: from algae to flowering plants. Annu Rev Plant Biol. 2011;62:567–90.
89. Reggiori F, Gabius H-J, Aureli M, Römer W, Sonnino S, Eskelinen E-L. Glycans in autophagy, endocytosis and lysosomal functions. Glycoconj J. 2021;38:625–47.
90. Cheng H, Shao Z, Lu C, Duan D. Genome-wide identification of chitinase genes in Thalassiosira pseudonana and analysis of their expression under abiotic stresses. BMC Plant Biol. 2021;21:87.
91. Taira T, Gushiken C, Sugata K, Ohnuma T, Fukamizo T. Unique GH18 chitinase from Euglena gracilis: full-length cDNA cloning and characterization of its catalytic domain. Biosci Biotechnol Biochem. 2018;82:1090–100.
92. Seeleuthner Y, Mondy S, Lombard V, Carradec Q, Pelletier E, Wessner M, et al. Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. Nat Commun. 2018;9:310.
93. Faik A. Xylan Biosynthesis: news from the Grass1. Plant Physiol. 2010;153:396–402.
94. Romero AM, Mateo JJ, Maicas S. Characterization of an ethanol-tolerant 1,4-β-xylosidase produced by Pichia membranifaciens. Lett Appl Microbiol. 2012;55:354–61.
95. Hsieh YSY, Harris PJ. Xylans of red and green algae: what is known about their structures and how they are synthesised? Polymers. 2019;11:354 Multidisciplinary Digital Publishing Institute.
96. Vaaje-Kolstad G, Westereng B, Horn SJ, Liu Z, Zhai H, Sørlie M, et al. An oxidative enzyme boosting the enzymatic conversion of recalcitrant polysaccharides. Science. 2010;330:219–22 American Association for the Advancement of Science.
97. Lo Leggio L, Simmons TJ, Poulsen J-CN, Frandsen KEH, Hemsworth GR, Stringer MA, et al. Structure and boosting activity of a starch-degrading lytic polysaccharide monooxygenase. Nat Commun. 2015;6:5961 Nature Publishing Group.
98. Nanjappa D, Sanges R, Ferrante MI, Zingone A. Diatom flagellar genes and their expression during sexual reproduction in Leptocylindrus danicus. BMC Genomics. 2017;18:813.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

