bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Eukaryotic genomes from a global metagenomic dataset illuminate trophic modes and biogeography of ocean plankton

Harriet Alexander[1,*], Sarah K. Hu[2], Arianna I. Krinos[1,3], Maria Pachiadaki[1], Benjamin J. Tully[4], Christopher J. Neely[5], and Taylor Reiter[6]

[1]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA, USA, 02543

[2]Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA, USA, 02543

[3]MIT-WHOI Joint Program in Oceanography, Cambridge and Woods Hole, MA, 02540

[4]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089

[5]Department of Computational and Quantitative Biology, University of Southern California, Los Angeles, CA 90089

[6]Population Health and Reproduction, University of California, Davis, Davis, CA, 95616

[*]Correspondence; halexander@whoi.edu

Abstract

Molecular and genomic approaches that target mixed microbial communities (e.g., metagenomics or metatranscriptomics) provide insight into the ecological roles, evolutionary histories, and physiological capabilities of the microorganisms and the processes in the environment. Computational tools that harness large-scale sequence surveys have become a valuable resource for characterizing the genetic make-up of the bacterial and archaeal component of the marine microbiome. Yet, fewer studies have focused on the unicellular eukaryotic fraction of the community. Here, we developed the EukHeist automated computational pipeline, to retrieve eukaryotic and prokaryotic metagenome assembled genomes (MAGs). We applied EukHeist to the eukaryote-dominated large-size fraction data (0.8-2000μm) from the *Tara* Oceans survey to recover both eukaryotic and prokaryotic MAGs, which we refer to as TOPAZ (*Tara* Oceans Particle-Associated MAGs). The TOPAZ MAGs consisted of more than 900 eukaryotic MAGs representing environmentally-relevant microbial and multicellu-

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

lar eukaryotes in addition to over 4,000 bacterial and archaeal MAGs. The bacterial and archaeal TOPAZ MAGs retrieved with EukHeist complement previous efforts by expanding the existing phylogenetic diversity through the increase in coverage of many likely particle- and host-associated taxa. We also demonstrate how the novel eukaryotic genomic content recovered from this study might be used to infer functional traits, such as trophic mode. By coupling MAGs and metatranscriptomic data, we explored ecologically-significant protistan groups, such as the Stramenopiles. A global survey of both eukaryotic and prokaryotic MAGs enabled the identification of ecological cohorts, driven by specific environmental factors, and putative host-microbe associations. Accessible and scalable computational tools, such as EukHeist, are likely to accelerate the identification of meaningful genetic signatures from large datasets, ultimately expanding the eukaryotic tree of life.

Introduction

Unicellular microbial eukaryotes, or protists, play a critical part in all ecosystems found on the planet. In addition to their vast morphological and taxonomic diversity, protists exhibit a range of functional roles and trophic strategies (Caron et al., 2011). Protists are centrally important to global biogeochemical cycles, mediating the pathways for the synthesis and processing of carbon and nutrients in the environment (Mitra et al., 2014; Caron et al., 2017; Strom, 2008). Despite their importance across ecosystems and in the global carbon cycle, research on microbial eukaryotes typically lags behind that of bacteria and archaea (Caron and Countway, 2009; Keeling and Campo, 2017). Consequently, fundamental questions surrounding microbial eukaryotic ecological function remain unresolved. Novel approaches that enable genome retrieval from meta'omic data provide a means of bridging that knowledge gap.

Assembled genetic fragments (derived from metagenomic reads) can be grouped together based on their abundances, co-occurrences, and tetranucleotide frequency to reconstruct likely genomic collections, often called bins (Alneberg et al., 2014; Wu et al., 2014; Kang et al., 2019; Graham et al., 2017). These bins can then be refined through a series of steps to ultimately represent metagenome assembled genomes or MAGs (Parks et al., 2017; Delmont et al., 2018; Tully et al., 2018; Almeida et al., 2019). Binning metagenomic data into MAGs has revolutionized how researchers ask questions about microbial communities and has enabled the identification of novel bacterial and archaeal taxa and functional traits (Rinke et al., 2019; Tully, 2019), but the recovery of eukaryotic MAGs is less well established. The reason for this being arguably twofold: (1) eukaryotic genomic complexity (Zhang et al., 2011) complicates both metagenome assembly and MAG retrieval; and (2) there is a bias in currently available metagenomic computational tools towards the study of bacterial and archaeal members of the community. Much can be learned about the diversity and role of eukaryotes in our environment from eukaryotic MAG retrieval (Olm et al., 2019).

Here we developed and applied EukHeist, a scalable and reproducible pipeline to facilitate the reconstruction, taxonomic assignment, and annotation of prokaryotic and eukaryotic metagenome assembled genomes (MAGs) from mixed community metagenomes. The EukHeist pipeline was applied to a metagenomic dataset from the *Tara* Oceans expedition protist-size fractions samples (Carradec et al., 2018), which encompasses more than 20Tb of raw sequence data. From these large-size fraction metagenomic samples, we recovered over 4,000 prokaryotic MAGs and 900 eukaryotic MAGs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Results and Discussion

We developed the EukHeist metagenomic pipeline to automate the recovery and classification of eukaryotic and prokaryotic MAGs from large-scale environmental metagenomic datasets. EukHeist was applied to the metagenomic data from the large-size fraction metagenomic samples (0.8-2000μm) from *Tara* Oceans (Carradec et al., 2018), which is dominated by eukaryotic organisms. We generated 94 co-assembled metagenomes based on the ocean region, size fraction, and depth of the samples (Figure S1), which totaled 180 Gbp in length (Supplementary Table 1). A total of 988 eukaryotic MAGs and 4,022 prokaryotic MAGs were recovered; considering our efforts to target the larger size fractions, these MAGs have been made available under the name *Tara* Oceans Particle Associated MAGs, or TOPAZ (Supplementary Tables 2 and 3). The TOPAZ MAGs expand the current repertoire of publicly available eukaryotic genomic references for the marine environment and shed light on the biogeographical and functional potential of these eukaryotic-enriched marine communities.

Eukaryotic genome recovery from metagenomes covers major eukaryotic supergroups

The EukHeist classification pipeline identified 988 putative eukaryotic MAGs following the refinement of recovered metagenomic bins based on length ($>$ 2.5 Mbp) and proportion of base pairs predicted to be eukaryotic in origin by EukRep (West et al., 2018) (Figure S4). Protein coding regions in the eukaryotic MAGs were predicted using the EukMetaSanity pipeline (Neely et al., 2021), and the likely taxonomic assignment of each bin was made with MMSeqs (Steinegger and Söding, 2018) and EUKulele (Krinos et al., 2021) (Supplementary Table 2). Of the 988 eukaryotic MAGs recovered, 713 MAGs were estimated to be more than 10% complete based on the presence of core eukaryotic BUSCO orthologs (Simao et al., 2015). For the purposes of our subsequent analyses, we only consider the highly complete eukaryotic TOPAZ MAGs, or those that were greater than 30% complete based on BUSCO ortholog presence (n=485) (Figure 1).

Eukaryotic genomes are known to be both larger and have higher proportions of non-coding DNA than bacterial genomes (Zhang et al., 2011). On average across sequenced eukaryotic genomes, 33.1% of genomic content codes for genes (2.6% - 59.8% for the 1st and 3rd quartiles) (Hou and Lin, 2009), while bacterial genomes have a higher proportion of coding regions (86.9%; 83.9% - 89.3%) (Hou and Lin, 2009). The high-completion TOPAZ eukaryotic MAGs have an average of $73.7\% \pm 14.3\%$ gene coding regions (Figure S9). This trend of a higher proportion of coding regions was consistent across eukaryotic groups, where Haptophyta and Ochrophyta TOPAZ MAGs had an average coding region of $80.3 \pm 4.9\%$ and $78.1 \pm 6.3\%$, respectively. Genomes from cultured Haptophyta (*Emiliania huxleyi* CCMP1516 with 31 Mb or 21.9% (Read et al., 2013)), and Ochrophyta (*Phaeodactylum tricornutum* with 15.4 Mb or 57.3% (Bowler et al., 2008)) had significantly lower proportions of protein coding regions within their genomes compared to TOPAZ MAGs. The lowest percentages of gene coding were within Metazoan and Fungal TOPAZ MAGs, with $52.6 \pm 9.8\%$ and $58.8 \pm 6.7\%$, respectively. As a point of comparison, the human genome is estimated to have $\approx$ 34 Mb or $\approx$ 1.2% of the genome coding for proteins (Consortium, 2004). Globally, the higher gene coding percentages for the recovered eukaryotic TOPAZ MAGs likely reflect biases caused by the use of tetranucleotide frequencies in the initial binning (Kang et al., 2019) as well as challenges inherent in the assembly of non-coding and repeat-rich regions of eukaryotic genomes.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

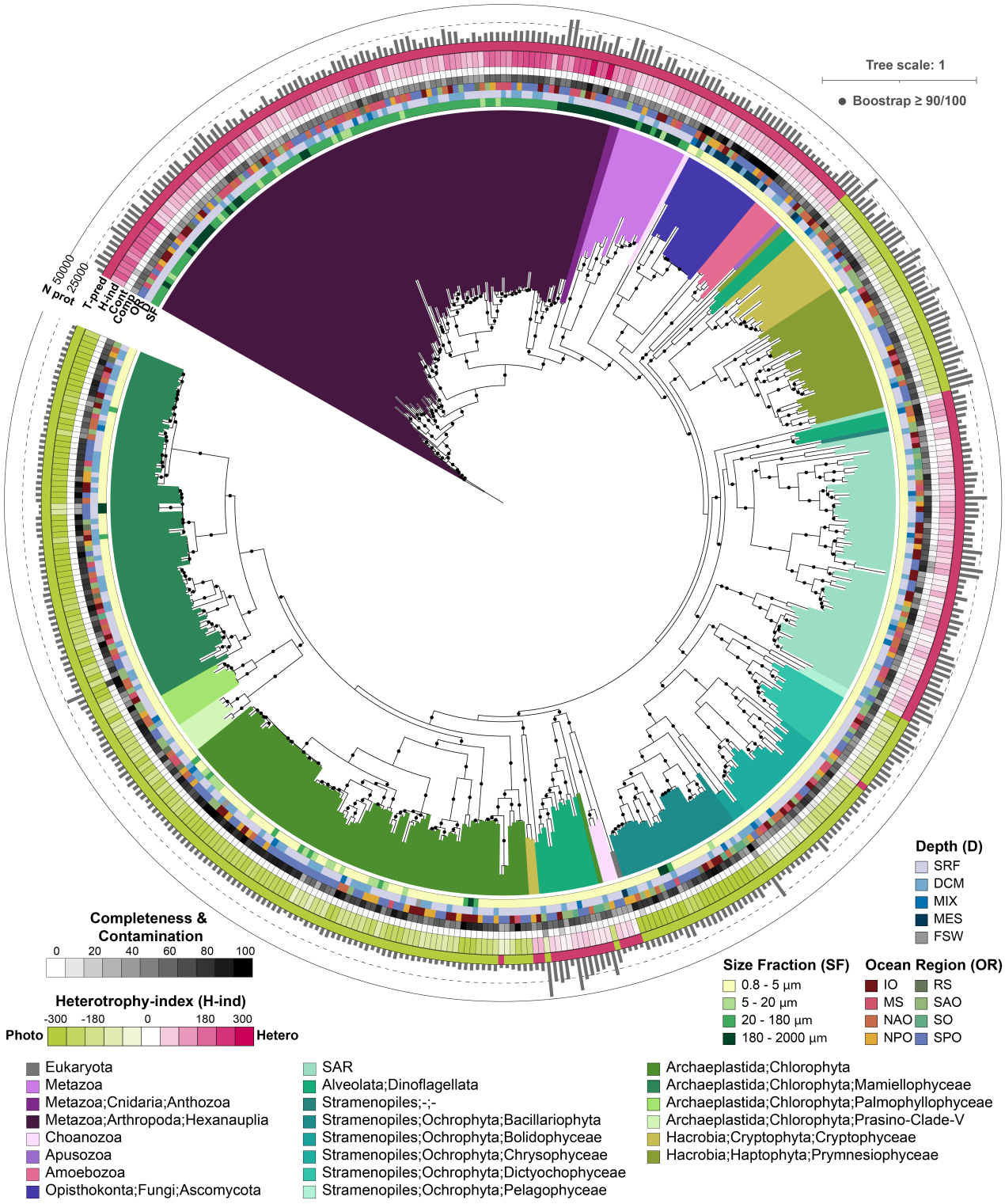

Figure 1: [Continued on next page.]

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 1: TOPAZ eukaryotic MAGs span the eukaryotic tree of life. The maximum likelihood tree was inferred from a concatenated protein alignment of 49 proteins from the eukaryotic BUSCO gene set that were found to be commonly present across at least 75% of the 485 TOPAZ eukaryotic MAGs that were estimated to be >30% complete based on BUSCO ortholog presence. The MAG names were omitted but the interactive version of the tree containing the MAG names can be accessed through iTOL (`https://itol.embl.de/shared/halexand`). Branches (nodes) are colored based on consensus protein annotation estimated by EUKulele and MMSeqs. The Ocean Region (OR), Depth (D), and Size Fraction (SF) of the co-assembly that a MAG was isolated from is color coded as colored bars. The completeness (comp) and contamination (cont) as estimated based on BUSCO presence are depicted as a heatmap. Predicted Heterotrophy Index (H-index), which ranges from phototroph-like (-300) to heterotroph-like (300) is shown as a heatmap. The predicted trophic mode (T-pred) based on the trophy random forest classifier with heterotroph (pink) and phototroph (green), is depicted. The number of proteins predicted with EukMetaSanity are shown as a bar graph along the outermost ring.

Phylogenetic placement of TOPAZ MAGs aligned with estimated taxonomy based on protein-consensus annotation (Figure 1). The recovered MAGs spanned 8 major eukaryotic supergroups: Archaeplastida (Chlorophyta), Opisthokonta (Metazoa, Choanoflagellata, and Fungi), Amoebozoa, Apusozoa, Haptista (Haptophyta), Cryptista (Cryptophyta), and the SAR supergroup (Stramenopiles, Alveolata, and Rhizaria) (Burki et al., 2020). Eukaryotic MAGs were retrieved from all ocean regions surveyed, with the largest number of high-completion TOPAZ MAGs recovered from the South Pacific Ocean Region (SPO) (n=143) and the fewest recovered from the Southern Ocean (SO) (n=11) and Red Sea (RS) (n=12) (Figure S8). These regional trends in MAG recovery and taxonomy aligned with the overall sequencing depth at each of these locations (Supplementary Table 1), with fewer, less diverse MAGs recovered from the SO and RS (Figures S5 and S8).

The largest number of MAGs was recovered from the smallest size fraction ($0.8-5\mu m$) (n=311) (Figures 1 and S5), and yielded the highest taxonomic diversity, including MAGs from all the major supergroups listed above (Figure S5). Chlorophyta (n=133), Ochrophyta (n=57), and taxa placed within the SAR group (Stramenopiles, Alveolata, and Rhizaria) (n=56) made up the the largest proportion of small size fraction MAGs. Chlorophyta MAGs were smaller and had fewer predicted proteins relative to other eukaryotic MAGs, despite demonstrating comparable completeness metrics; the average Chlorophyta MAG size was 13.9 Mbp with 7525 predicted proteins (Figure S9). By contrast, Cryptophyta and Haptophyta had the largest average MAG size with 50.8 Mbp and 44.4 Mbp with an average of 23500 and 24400 predicted proteins, respectively (Figure S9). Fewer eukaryotic MAGs were recovered from the other size fractions $5-20\mu m$ (n=20), $20-180\mu m$ (n=87), and $180-2000\mu m$ (n=39) (Figure S5), instead these larger size fractions recovered a higher total number of metazoan MAGs. Metazoan MAGs had the lowest average completeness ($50\pm13\%$) (Figures S7 and S9); where the average size of recovered metazoan MAGs was 43.2 Mb (6.5-177Mbp), encompassing an average of 14600 proteins (Figure S9). 76 of the 123 metazoan MAGs likely belong to the Hexanauplia (Copepoda) class; copepod genomes have been estimated to be up to 2.5 Gb with high variation (10-fold difference) across sequenced members (Jørgensen et al., 2019).

MAGs were also retrieved from all discrete sampling depths: surface, SRF (n=315), deep chloro-

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

phyll max, DCM (n=133), mesopelagic, MES (n=13), as well as samples with no discrete depth, MIX (n=21) and the filtered seawater controls, FSW (n=3). Notably, the FSW included 1 Chlorophyta MAG (TOPAZ_IOF1_E003) that was estimated to be 100% complete with no contamination (Supplementary Table 2).

The composition of TOPAZ MAGs from basin-scale mesopelagic co-assemblies recovered a higher percentage of fungi relative to other depths. This is similar to other mesopelagic and bathypelagic molecular surveys, where the biomass of fungi is thought to outweigh other eukaryotes (Morales et al., 2019; Pernice et al., 2015; Edgcomb et al., 2010). Further, fungal MAGs had the highest overall average completeness ($87 \pm 15\%$) (Figures S7 and S9). A total of 16 highly complete fungal MAGs were also recovered, of those, 11 originated from the MES (Figures 1 and S8). Putative fungal TOPAZ MAGs were recovered from the phyla Ascomycota (n=10) and Basidiomycota (n=1) and ranged in size from 12.5-47.8 Mb (Figure S9), which are within range of known average genome sizes for these groups, 36.9 and 46.5 Mb, respectively (Mohanta and Bae, 2015).

The metagenome read recruitment to these TOPAZ MAGs paralleled MAG recovery, where metazoan MAGs dominated the larger size fractions ($20-180\mu m$ and $180-2000\mu m$) across both the surface and DCM for all stations, and Chlorophyta MAGs were dominant across most of the small size fraction stations ($0.8-5\mu m$) (Figure S11). A notable exception are the stations from the Southern Ocean, where Haptophyta and Ochrophyta were most abundant in all size fractions. Compared to the samples from the photic zone, SRF and DCM, the average recruitment of reads from the MES was far lower (24500 ± 34450 average CPM in the MES compared to 131000 ± 104000 and 136000 ± 85000 for the SRF and DCM, respectively (Figure S11). This suggests that the mesopelagic have high variability across communities (Pernice et al., 2015) and that we did not fully capture the eukaryotic MAGs that adequately describe all surveyed communities. Alternatively, this might suggest that the communities sampled were dominated by prokaryotic biomass (Pernice et al., 2014).

Trophic mode can be predicted from MAG gene content

Eukaryotic microbes can exhibit a diversity of functional traits and trophic strategies in the marine environment (Worden et al., 2015; Caron et al., 2011), including phototrophy, heterotrophy, and mixotrophy. Phototrophic protists are responsible for a significant fraction of the organic carbon synthesis via primary production; these phototrophs dominate the microbial biomass and diversity in the sunlit layer of the oceans (Worden et al., 2015; de Vargas et al., 2015). Phagotrophic protists (heterotrophs), which ingest bacteria, archaea, and smaller eukaryotes, and parasitic protists are known to account for a large percentage of mortality in food webs (Sherr and Sherr, 2002; Caron et al., 2011; Worden et al., 2015). Protists are also capable of mixed nutrition (mixotrophy), where a single-cell exhibits a combination of phototrophy and heterotrophy (Stoecker et al., 2017). Typically, the identification of trophic mode has relied upon direct observations of isolates within an lab setting, with more recent efforts including transcriptional profiling as a means of assessing trophic strategy (Keeling et al., 2014; Liu et al., 2016). Scaling up these culture-based observations to environmentally-relevant settings (Alexander et al., 2015; Hu et al., 2018; Gong et al., 2016) has been an important advance in the field for exploring complex communities without cultivation. An outcome from these studies has been the realization that trophic strategies are not governed by single genes (Labarre et al., 2020); in reality, trophic strategy will be shaped by an organisms' physiological potential and environmental setting. Therefore larger genomic and transcriptomic efforts to predict or characterize presumed

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

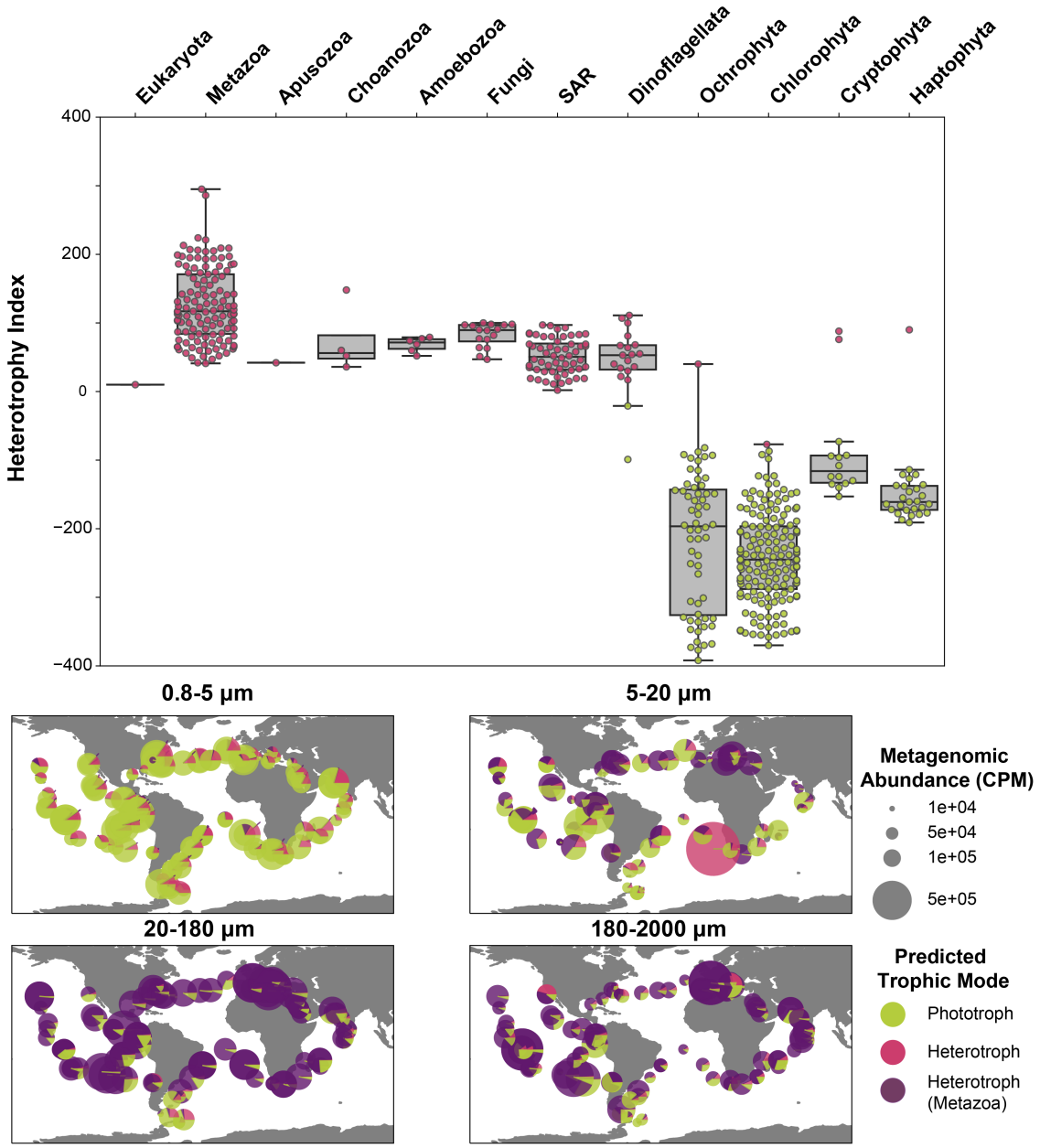

Figure 2: Estimated trophic status of TOPAZ eukaryotic MAGs. (Top) Trophic status was predicted for each high-completion TOPAZ eukaryotic MAG using a Random Forest model trained on the presence and absence of KEGG orthologs and is shown as a color (green, phototroph, pink, heterotroph). The Heterotrophy Index (H-index) (Equation (8)) for each MAG is plotted with a box plot showing the range of the H-index for each higher level group. (Bottom) The relative distribution and abundance of Phototroph (green), non-Metazoan Heterotroph (Pink), and Metazoan Heterotroph (Purple) is depicted across all surface samples. Plots are subdivided by size classes.

trophic strategies among mixed microbial communities will greatly contribute to our understanding of the role that microorganisms play in global biogeochemical cycles, by enabling the observation of functional traits and strategies *in situ*.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Large scale meta'omic results, such the TOPAZ MAGs recovered here, can be leveraged alongside presently available reference data to enable the prediction of biological traits (such as trophic mode) without *a priori* information. Machine learning (ML) applications can be implemented to access the potential of these large datasets. ML approaches have been recently shown to be capable of accurate functional prediction and cell type annotation using genetic input, in particular for cancer cell prediction (Shipp et al., 2002; Bashiri et al., 2017; Tabl et al., 2019), and functional gene and phenotype prediction in plants (Mahood et al., 2020). Recently, these approaches have been applied to culture and environmental transcriptomic data to predict trophic mode using currently available trophy annotations (Lambert et al., 2021; Burns et al., 2018; Jimenez et al., 2021). Here, we apply an independent machine learning model to the eukaryotic TOPAZ MAGs to predict each organisms' capacity for various metabolisms.

Using a reference set built from protistan transcriptomic data, we predicted the trophic mode of the highly complete TOPAZ MAGs using machine learning and direct estimation via presence of important KEGG pathways (eq. (8)). As the gradient of trophic mode among protists is not strictly categorical, we calculated a Heterotrophy Index (H-index) that places the TOPAZ MAGs on a scale of highly phototrophic (negative values) to highly heterotrophic (positive values) (Figures 1 and 2). Thus, for all sufficiently complete ($\geq 30\%$) TOPAZ MAGs we have predicted both a gross trophic category (heterotrophic ($n = 227$), mixotrophic ($n = 0$), or phototrophic ($n = 258$) as well as the quantitative extent of heterotrophy (H-index, eq. (8)). Broadly, the trophic predictions aligned well with the putative taxonomy of each MAG (Figures 1 and 2). For example, TOPAZ MAGs that had taxonomic annotation of well known heterotrophic lineages (Metazoa, Fungi), were predicted as heterotrophs based on our model. Further, our data-driven trophic mode predictions correlate well with an independent model designed to identify the presence of photosynthetic machinery and capacity for phagotrophy (Burns et al., 2018) (Figures S26 and S27).

Despite evidence that many lineages recovered include known mixotrophs, no TOPAZ MAGs were identified as mixotrophic using this approach. However, the utility of the H-index enables us to still consider mixotrophic-capable MAGs. We explore the likely reasons for this more deeply in the Section 1.3, but one potential explanation is that MAG recovery targets the genomic content of a eukaryotic lineage and the evolutionary history of phototrophy and heterotrophy is complicated and varies with respect to species (Flynn et al., 2019). Therefore, the genetic composition of MAGs may reflect encoded metabolisms that are not necessarily exhibited *in situ*. Additionally, mixotrophy is not a singular trait, but rather a spectrum of metabolic abilities that are largely driven by the microorganisms' nutritional needs and surrounding environment. Continued culturing combined with large-scale 'omic efforts will continue to improve such ML models focused on complex traits and ultimately our ability to predict trophic mode. We suggest that the integration of metagenomic and metatranscriptomic datasets might better reflect the active strategies being used.

Ecological niches of heterotrophic and phototrophic Stramenopiles using metatranscriptomic evidence

A total of 24 high-completion TOPAZ eukaryotic MAGs were subset as a case study to explore the utility of pairing metatranscriptome data with MAG results to characterize the ecological context of less resolved protistan lineages. Selected MAGs included 11 taxonomically-assigned as *Dic-*

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

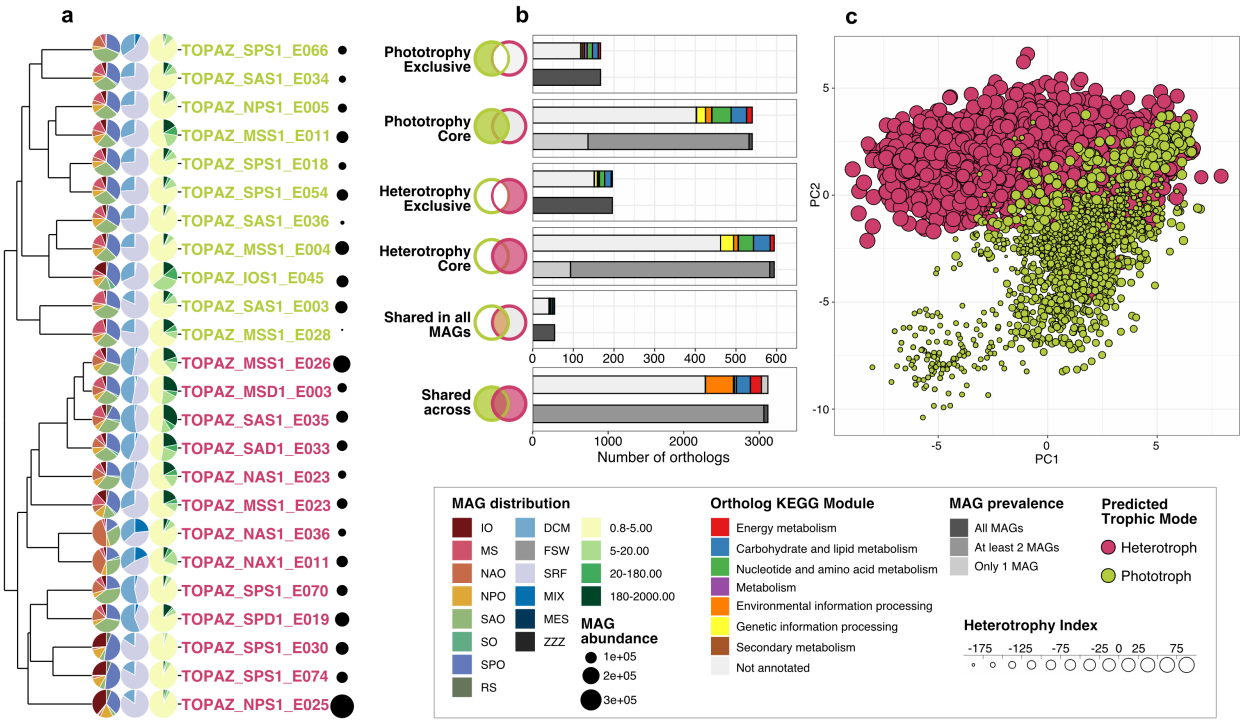

Figure 3: Dictyochophyceae and stramenopile MAGs A subset of 24 highly-complete TOPAZ MAGs taxonomically classified as stramenopiles and *Dictyochophyceae*. (a) Cluster dendrogram derived from the presence or absence of orthologs grouped MAGs by predicted trophic mode (green indicates phototrophy and pink indicates heterotrophy). Pie charts to the left of each TOPAZ MAG name indicate the relative CPM abundance of each MAG for (left to right pies) ocean region, depth sampled, and size fraction. Bubble plots to the right of each TOPAZ MAG name indicate the total MAG CPM abundance. (b) Summary of shared and unique orthologs based on occurrence in phototrophy- and heterotrophy-predicted TOPAZ MAGs. Venn diagrams indicate category of orthologs shown in each panel, while panels report the KEGG module and prevalence among MAGs in each category (bar plot). (c) Principle component analysis derived from metatranscriptome reads, from the surface and smallest size fraction, mapped to shared orthologs (Shared in all MAGs in (b)) among all 24 MAGs. Symbol size designates Heterotrophy Index, while symbol color denotes predicted trophic mode.

tyochophyceae (silicoflagellates), which were putatively classified as phototrophs with our trophic model (Figure 2), and 13 MAGs within a phylogenetically-related stramenopile clade and classified as heterotrophs (Figures 1 and 3 a).

All 24 MAGs had a cosmopolitan distribution primarily originating from surface samples from the smallest size fraction, but some individual MAGs had relatively higher CPM abundances suggesting environmental selection based on oceanographic region (Figure 3a, Figure S18, Figure S17). The biogeography of the *Dictyochophyceae* aligned with existing literature, where *Dictyochophyceae* are globally distributed and typically found in the euphotic layer of the world ocean (Vaulot et al., 2008; Obiol et al., 2020; Massana, 2011). Global sampling efforts have also recovered genetic signatures of *Dictyochophyceae* as a prominent, but not abundant, member of the Stramenopile group that does not demonstrate remarkable seasonality (Giner et al., 2019; Obiol et al., 2020). While grazing on bac-

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

teria and picocyanobacteria has been observed within mixotrophic *Dictyochophyceae*, previous work to quantify grazing rates were unsuccessful to due to the low cell abundance (Unrein et al., 2014). Therefore, metagenomic and metatranscriptomic datasets, such as the recovered TOPAZ MAGs, are well suited to illuminate the biogeography and functional potential of the *Dictyochophyceae*. *Dictyochophyceae* and heterotrophic Stramenopiles have previously been reported in analyses of the *Tara* Oceans data (Carradec et al., 2018; Pierella Karlusich et al., 2020; Vorobev et al.; Sieracki et al., 2019).

This analysis enabled both targeted and untargeted approaches to investigate the physiology of SAR TOPAZ MAGs. First, we assessed ortholog co-occurrence across the MAGs to identify genes that were present only among the phototrophy-predicted *Dictyochophyceae* MAGs (Phototrophy Exclusive; Figure 3b). Genes deemed Phototrophy Exclusive (Figure 3b) were related to chlorophyll biosynthesis (Por; protochlorophyllide reductase), enzymes integral to the pentose phosphate pathways, and acyltransferases and carboxylases, which are involved in *de novo* fatty acid biosynthesis functions (e.g., ACACA; acetyl-CoA carboxylase biotin, ACSS; acetyl-coA synthetase). Inversely, genes detected only within putative stramenopile MAGs classified as heterotrophs (*i.e.*, Heterotrophy Exclusive; Figure 3 b) included enzymes integral for the breakdown of large sugar molecules such as glycosaminoglycans (*e.g.*, IDUA; L-iduronidase, UDP-glucose:O-linked fucose beta-1,3-glucosyltransferase, NAGLU; alpha N acetylglucosaminidase, IDS; iduronate-2sulfatase, and GALC; galactosylceramidases). Enzymes associated with glycosaminoglycan metabolism may be associated with cell adhesion or the intracellular re-processing of glycosaminoglycan; the latter of which may be a genetic attribute for more heterotrophic lifestyles. By isolating the presence and absence of specific genes across MAGs with varied predicted trophic modes, we can identify sets of genes that may be indicative of a species' ecological role.

Ordination results based on the expression of transcripts common across all 24 MAGs (Figure 3 c; based on 'Shared in all MAGs' in Figure 3b) clustered by H-index and TOPAZ MAG identity (Figure S19). While there was some overlap among MAGs predicted to be heterotrophic versus phototrophic, trends dictating the PCA results appeared to be driven primarily by the trophic mode of individual MAGs, rather than region sampled (Figure 3 c, Figure S19). Further, ordination results resembled previously observed trends from transcript-based efforts to separate phototrophic, mixotrophic, and heterotrophic protistan species from cultivation (Koid et al., 2014; Beisser et al., 2017) and the environment (Hu et al., 2018). Results from the mapped metatranscriptome reads revealed populations to exhibit more heterotrophic or phototrophic traits depending on the environment (Figure S21, Figure S22). For instance, among the *Dictyochophyceae*-predicted MAGs variable phototrophic versus heterotrophic relative abundances may reflect a mixotrophic-capable population responding to the environment (Figure S21). The 13 MAGs classified as heterotrophic (ML model in this study) were phylogenetically similar to other Stramenopiles (Figure 3 b & Figure 1) and H-indices reported for each MAG aligned with what was seen in the metatranscriptome signal, where MAGs with the highest heterotrophy scores (e.g., TOPAZ_SAS1_E035 and TOPAZ_NAX_E011; Figure S19) had higher CPM associated with heterotrophic traits in all samples (Figure S22). While the identity of the *Dictyochophyceae* MAGs was further supported by phylogenetic similarity with other *Dictyochophyceae* MAGs and single-cell genome-informed MAGs (Figure S20), the taxonomic identity of the presumed heterotrophic Stramenopile TOPAZ MAGs was less resolved. While we cannot confidently annotate beyond the taxonomic classification of 'SAR', these MAGs were distinct from those in culture and likely include a mixotrophic-capable group of protists distinct from the MA-

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

rine STramenopiles (MAST; Figure S20). These findings demonstrate the value of large untargeted genetic approaches to gain insight into the *in situ* metabolisms of less explored branches of the eukaryotic tree of life. Paired metagenomic and metatranscriptome results, alongside the environmental context provided by a large-scale global sampling effort, and the predicted nutritional strategies we can gain a more comprehensive understanding of protists in our oceans.

TOPAZ prokaryotic MAGs distinct from previous marine MAG recovery efforts

The vast majority of the retrieved prokaryotic MAGs belonged to Bacteria. High-quality non-redundant TOPAZ (HQ-NR-TOPAZ) MAGs were comprised of 711 bacterial and 5 archaeal MAGs belonging to 30 different phyla (Figure 4 and Supplementary Table 4); an additional 15 phyla were recovered in the medium quality (MQ) MAGs. Of the 716 HQ-NR-TOPAZ MAGs, 507 were unique based on a 99% ANI comparison threshold with MAGs generated from previous binning efforts from *Tara* Oceans metagenomic data, including Delmont et al. (2018) (TARA), Tully et al. (2018) (TOBG), and Parks et al. (2017) (UBA) (Figure 4). The phylogenetic diversity captured by the TOPAZ MAGs was quantified by a comparison to a "neutral" reference set of genomes; these neutral references approximate the state of marine microbial genomes, dominated by isolate genomes, previous to the incorporation of the *Tara* Oceans-derived MAGs (Table 1). Relative to the neutral genomic references, the entire TOPAZ NR (includes both HQ and MQ) set represented a 42.8% phylogenetic gain (as measured by additional branch length contributed by a set of data) and 59.9% phylogenetic diversity (as measured by the total branch length spanned by a set of taxa), as compared to efforts focused solely on the smaller size fractions such as TARA and UBA, which had a smaller degree of gain (31.0% and 25.8%, respectively) and diversity (44.4% and 40.5%, respectively) (Table 1). An inclusive tree containing the neutral reference and all *Tara* Oceans MAGs (TOBG + UBA + TARA + TOPAZ), the TOPAZ NR MAGs represented 14.4% of the phylogenetic gain and 44.7% phylogenetic diversity, suggesting that the TOPAZ MAGs offer the largest increase in phylogenetic novelty when compared to MAGs reconstructed from the metagenomes of the smaller size fractions ($< 5.00\mu m$). The TOPAZ MAGs primarily originated from the larger *Tara* Ocean size fraction samples, and thus include a higher proportion of more complex host- and particle-associated bacterial communities. The novelty of the HQ- and MQ-NR-TOPAZ MAGs here, suggests that these particle-associated MAGs are overlooked and current genome databases are largely skewed towards free-living bacteria.

To confirm the hypothesis that the prokaryotic TOPAZ MAGs included particle-associated members, we examined the genomic features of several selected groups that were well-recovered here and in single-cell amplified genomic datasets (i.e., GORG) (Pachiadaki et al., 2019). To avoid potential biases related to completeness and contamination of the genomes, only the HQ-NR MAGs were compared to the GORG SAGs, and analyses were limited to groups with sufficient representation within both datasets (Bacteriodota, Cyanobacteria, and Proteobacteria). For these well represented groups, the average GC% and estimated genome size of the TOPAZ MAGs were significantly higher than the ones typically reported in free-living marine bacteria (Dufresne et al., 2005; Swan et al., 2013; Luo et al., 2015) and those observed within the GORG dataset (Pachiadaki et al., 2019). TOPAZ MAGs were found to encode more tRNAs on average per genome than GORG (39.5 vs 30). Additionally, Carbohydrate-Active Enzymes (CAZy) and peptidases were enriched within the TOPAZ MAGs relative to GORG (Figure S29). Larger genomes have been considered diagnostic for a copiotrophic

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

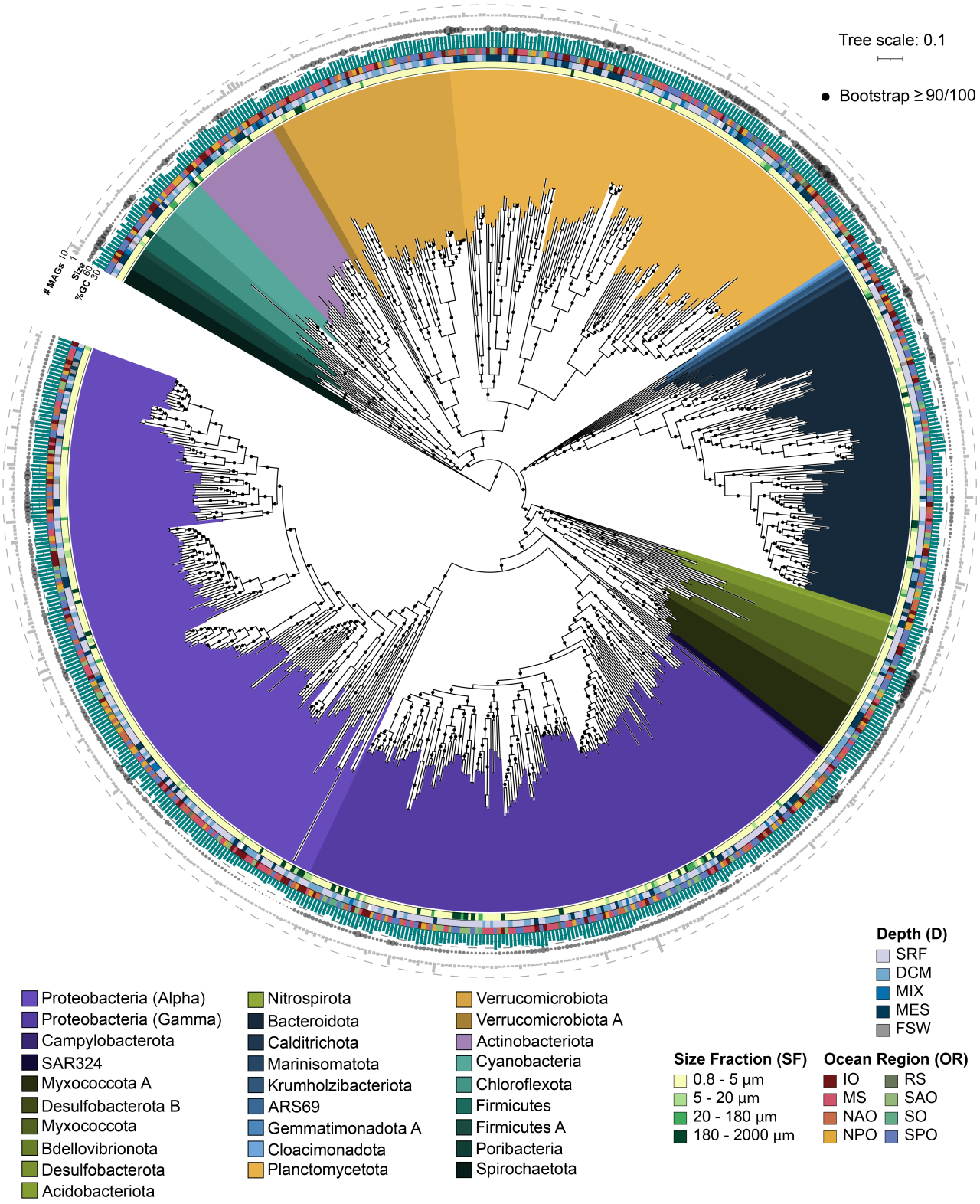

Figure 4: [Continued on next page.]

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figure 4: Diversity of the high-quality non-redundant bacterial TOPAZ MAGs. The approximately-maximum-likelihood phylogenetic tree was inferred form a concatenated protein alignment of 75 proteins using FastTree and GToTree workflow. The MAG names were omitted but the interactive version of the tree containing the MAG names can be accessed through iTOL (`https://itol.embl.de/shared/halexand`). Branches (nodes) are colored based on taxonomic annotations estimated by GTDBtk. The Ocean Region (OR), Size Fraction (SF), and Depth (D) of the co-assembly that a MAG was isolated from is color coded as colored bars. The GC (%) content is shown as a bar graph (in green), the genome size as bubble plot (the estimated size of the smallest genome included in this tree is 1.00Mbp and the largest is 13.24Mbp) and the number of MAGs in each genomic cluster (of 99 or higher %ANI) as a bar plot (in grey)

lifestyle in bacteria (Okie et al., 2020), since the more extended and flexible gene repertoire can facilitate substrate catabolism in organic rich niches such as particles. Genomes of copiotrophs are also commonly found to have higher copy numbers of genes associated with replication and protein biosynthesis such as tRNAs and rRNAs (Rocha, 2004) which facilitate higher growth rates. In contrast, the streamlined genomes of SAR11 and other groups that have free-living oligotrophic lifestyles require fewer resources to maintain and replicate their genomes and have higher carbon-use efficiency (Giovannoni et al., 2014). Similarly, G and C have higher energy cost of production and more limited intracellular availability compared to A and T (Moore et al., 2013; Luo et al., 2015). The genomic trends observed support our findings that TOPAZ MAGs represent both particle associated and free-living microbes, and are relatively enriched for copiotrophic microbes.

Environmental factors structure TOPAZ MAG co-occurrence

The co-retrieval of eukaryotic and prokaryotic MAGs from across the global ocean allows the unique opportunity to assess the biogeographical and ecological associations and potential co-occurrence of these organisms while also being able to infer likely function. To identify communities of associated organisms that co-occur across the surface ocean metagenomes, we performed a correlation clustering based on the abundances of the eukaryotic TOPAZ MAGS and the HQ-NR-TOPAZ MAGs (Figure 5 a). We employed a modularity optimization algorithm to the correlation analysis Blondel et al. (2008) to identify distinct communities of co-occurring organisms. This approach identified seven distinct communities, which included 379 of the non-redundant TOPAZ MAGs (Figure 5 b). The communities were variably connected to each other, as defined by Equation (11), with high connectedness among Communities 1, 2, and 3, and patchy connectedness among Communities 4-7 (Figure 5 b). Community 6 (the smallest of the communities, which consisted of a single Bacillariophyta TOPAZ MAG and five distinct *Synechococcus* TOPAZ MAGs) had the highest inter-connectedness (0.733 connectedness), suggesting that members of this community co-occurred across samples with high fidelity. Moreover, this community was largely distinct from other communities and only shared significant connections to Community 7. The other six communities showed lesser degrees of inter-connectedness (range: $0.181 - 0.550$; mean: 0.365 ± 0.204), suggesting that they co-occur less consistently across samples.

The seven communities that we identified based on metagenomic abundance correlations also significantly correlated with environmental factors, which consequently define the environmental niches

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table 1: Phylogenetic diversity and gain of various MAGs originating from *Tara* Oceans. Phylogenetic diversity and gain of prokaryotic MAGs was assessed for this study (TOPAZ), TOBG (Tully et al., 2018), UBA (Parks et al., 2017), and TARA (Delmont et al., 2018) relative to each other as well as a "Neutral" tree comprised of relevant marine bacteria.

Base tree	MAGs of interest	No. of MAGs	Phylogenetic diversity*	Phylogenetic gain°
Neutral	TOPAZ (MQ, NR)	1,571	59.9%	42.8%
Neutral	TOPAZ (HQ, NR)	634	41.6%	25.8%
Neutral	TOBG	1,974	61.3%	46.7%
Neutral	UBA	1,052	40.5%	25.8%
Neutral	TARA	722	44.4%	31.0%
Neutral	TOBG + UBA + TARA	3,750	66.6%	51.8%
Neutral + *Tara* Oceans MAGs^{HQ}	TOPAZ (HQ, NR)	634	26.1%	6.2%
Neutral + *Tara* Oceans MAGs^{MQ}	TOPAZ (MQ, NR)	1,572	44.7%	14.4%
Neutral + *Tara* Oceans MAGs^{MQ}	TOBG	1,977	48.5%	11.1%
Neutral + *Tara* Oceans MAGs^{MQ}	UBA	1,055	23.8%	1.6%
Neutral + *Tara* Oceans MAGs^{MQ}	TARA	722	28.0%	3.4%

* total branch length spanned by a set of taxa
° additional branch length contributed by a set of taxa
^{HQ} includes Neutral, TOBG, UBA, TARA, and TOPAZ HQ, NR MAGs
^{MQ} includes Neutral, TOBG, UBA, TARA, and TOPAZ MQ, NR MAGs

where the communities were most abundant (Figure 5 c, Supplementary Table 13). Temperature was a primary factor defining the community correlations, significantly correlating with five of the seven communities. Community 4 correlated with colder temperatures and Communities, 1, 3, 5, and 6 correlated with warmer temperatures (Figure 5 c). In Community 4, we found significant positive correlations with chlorophyll and net primary productivity (Chla: $\rho = 0.401$, $p = 2.34e-35$, NPP: $\rho = 0.166$, $p = 1.28e-3$), while we found negative correlations with "residence time" ($\rho = -0.347$, $p = 1.98e-14$), indicating a likely occurrence in newly formed eddies. Thus, Community 4 was largely found within colder, productive regions, and had enhanced metagenomic abundance in the Southern Ocean and the North Atlantic (Figure S32). Community 4 was comprised of MAGs from Chlorophyta, Cryptophyta, Haptophyta, and Ochrophyta, the major groups containing primarily phototrophic eukaryotic microbes. 18 prokaryotic MAGs were also contained in this community, including both photosynthetic (Synechococcales) and non-photosynthetic lineages (e.g. Myxococcota and Planctomycetota). All told, this guild of MAGs comprises likely photosynthesizers often found in cold, but not necessarily nutrient-rich, environments.

The four communities (1, 3, 5, and 6) that were positively correlated with temperature were distin-

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

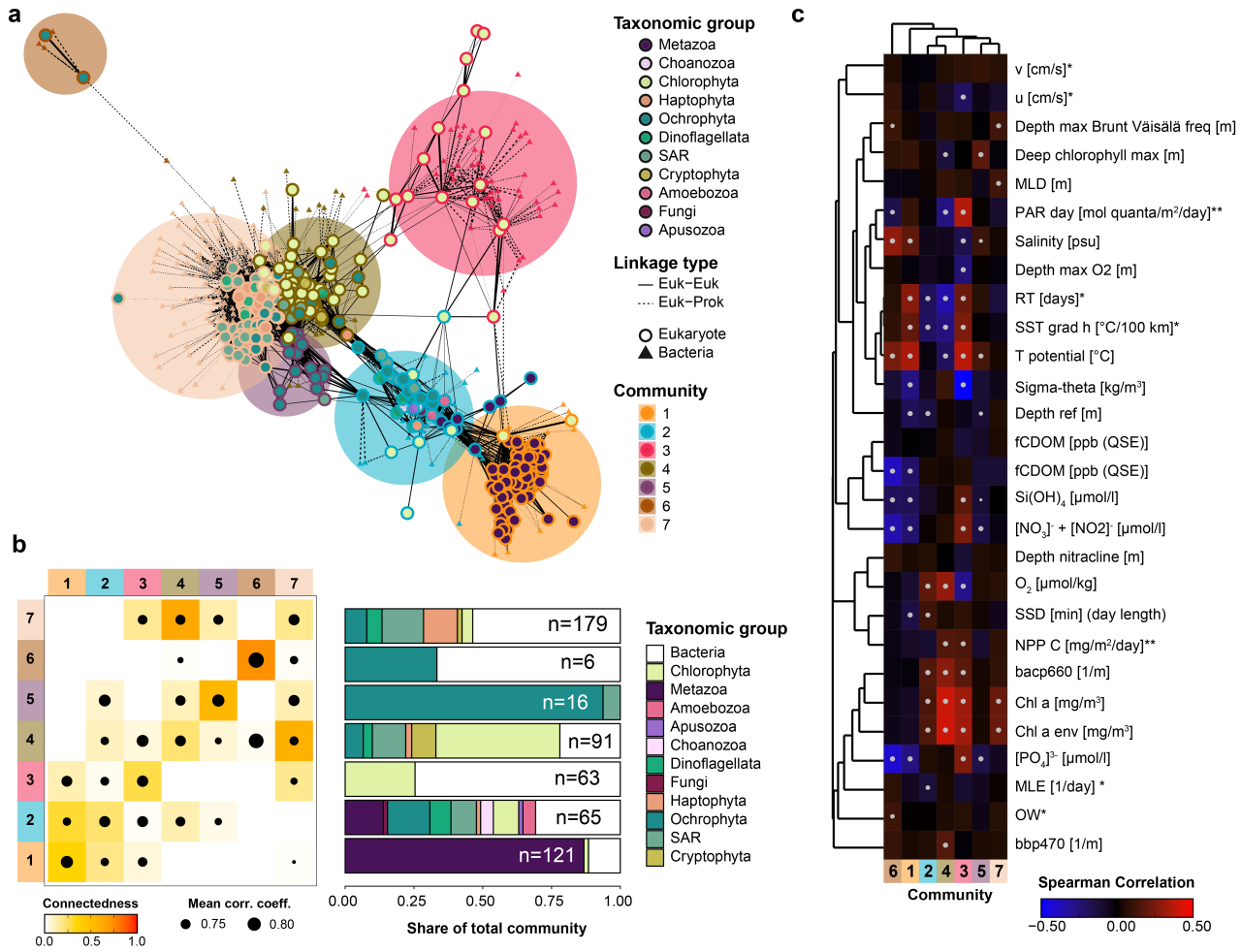

Figure 5: Distinct communities recovered from the TOPAZ MAGs. a) A network analysis performed on the metagenomic abundance of all recovered eukaryotic and prokaryotic TOPAZ MAGs based on Spearman Correlation analysis, identifying 7 distinct communities (see materials and methods). A force-directed layout of the seven communities is shown with eukaryotes (circles) and bacteria (triangles). Only linkages between eukaryotes are visualized. b) The connectedness and taxonomic composition of each community is depicted. Connectedness was calculated based on Equation (11). c) A Spearman correlation between the summed metagenomic abundance of each community and environmental parameters from the sampling (Tara Oceans Consortium and Tara Oceans Expedition, 2016), modeled mesoscale physical features based on d'Ovidio et al. (2010) (indicated with *), and averaged remote sensing products (indicated with **). Significant Spearman correlations, those with a Bonferroni adjusted $p < 0.01$, are indicated with a dot on the heatmap.

guished by different correlations with nutrients and physical features. Communities 1 and 3 tended to be found in longer-lived eddies (according to the calculation by d'Ovidio et al. (2010) as reported in the *Tara* Oceans metadata Tara Oceans Consortium and Tara Oceans Expedition (2016) as "residence time") (Figure 5 c; Community 1: $\rho = 0.274$, $p = 1.74e-8$, Community 3: $\rho = 0.216$, $p = 7.15e-5$). However, these two communities differed both in their association with nutrients and their taxonomic compositions. Community 1 was dominated by Metazoa and bacteria and correlated with oligotrophic

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

conditions (nitrate and nitrite: $\rho = -0.234$, $p = 1.21e-10$, phosphate: $\rho = -0.269$, $p = 1.76e-14$, silica: $\rho = -0.190$, $p = 1.02e-6$), and was most abundant in the larger size fraction samples (20-2000 μm) (Figure S32). By contrast, Community 3 was largely comprised of phototrophic chlorophytes and bacteria and was more likely to be found in high-nutrient environments (nitrate and nitrite: $\rho = 0.230$, $p = 3.04e-10$, phosphate: $\rho = -0.266$, $p = 3.82e-14$, silica: $\rho = 0.182$, $p = 4.39e-6$), and was most abundant in smaller size fraction samples (0.8-20 μm), particularly around the tropics (Figure S32). The other two warm-associated communities were comprised by SAR and bacteria (Community 5), and Ochrophyta and SAR (Community 6) (Figure 5 b). These communities were negatively correlated with nutrient concentrations (Community 5: nitrate and nitrite: $\rho = -0.140$, $p = 3.17e-3$, phosphate: $\rho = -0.147$, $p = 1.28e-3$, silica not significant; Community 6: nitrate and nitrite: $\rho = -0.355$, $p = 1.68e-26$, phosphate: $\rho = -0.418$, $p = 6.01e-38$, silica: $\rho = -0.187$, $p = 1.92e-6$;), suggesting communities that thrive in oligotrophic regions.

While many of the communities recovered appeared to be driven largely by environmental forces, the genomic signatures of Community 1 members suggest that this community was comprised by MAGs from hosts and likely microbiome-associated microbes. Community 1 was comprised primarily of Metazoa, specifically Hexanauplia, and bacterial MAGs (Figure 5 b). Many of the bacterial MAGs in Community 1 had genes that suggest adaptations to microaerophic niches such as those which might be experienced when living in close host association (e.g. high affinity oxygen cytochromes, and reductases) (Figure S33). The bacterial MAGs in Community 1 could be broadly broken into two apparent functional types: those with larger genomes typical of copiotrophic bacteria and those with small genomes indicative potentially of reductive evolution. The first group was comprised of MAGs from family Saprospiraceae in phylum Bacteriodota (n=2, 3.0 Mbp average genome size), the family UBA2386 in phylum Plactomycetota, which lacks cultured representatives (n=2, 3.3 Mbp), the order Opitutales in phylum Verrucomicrobiota (n=2, 3.4Mbp), and the family Vibrionaceae (n=2, 4.5 Mbp) and order Pseudomonadales (n=2; 3.2 Mbp), both in phylum Gammaproteobacteria (Figure S33). In addition to their relatively large size, the Saprospiraceae, Plactomycetota and Vibrionaceae MAGs were found to encode for genes involved in the hydrolysis and utilization of various complex carbon sources including chitin and other carbohydrates (Figure S33), such as those that might be shed or excreted by zooplankton such as copepods (Corte et al., 2017). By contrast, the second group of bacterial MAGs within Community 1 with smaller genomes, included MAGs from the Proteobacteria order Rickettsiales (n=3, 0.6-1.2 Mbp), Gammaproteobacteria family Francisellaceae (n=1, 1.2 Mbp), and the Bacteriodota family Amoebophilaceae (n=1; 0.8Mbp) Figure S33). The smaller genome sizes exhibited by these groups may be indicative of a genome streamlining which occurred with reductive evolution due to obligate or facultative symbiosis (Giovannoni et al., 2014). Rickettsiales, Francisellaceae, and Amoebophilaceae all contain well-described obligate intracellular symbionts (Santos-Garcia et al., 2014; Darby et al., 2007; Li et al., 2021) and zoonotic pathogens (Celli and Zahrt, 2013; Darby et al., 2007).

Conclusion

Sequence datasets are revolutionizing how we form new hypotheses and explore environments on the planet. Here, we demonstrated a critical advance in the recovery of MAGs from environmentally-relevant eukaryotic organisms with EukHeist. The retrieval and study of MAGs to study the role

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

of microorganisms in environmentally significant biogeochemical cycling is promising; however the current lack of eukaryotic reference genomes and transcriptomes complicates our ability to interpret the eukaryotic component of the microbial community. We recovered 988 total eukaryotic MAGs, 485 of which were deemed highly complete. Our findings demonstrate that specific branches of the eukaryotic tree were more likely to be resolved at the MAG-level due to their smaller genome size, distribution in the water column, and biological complexity. A substantial portion of the recovered eukaryotic MAGs were distinct from existing sequenced representatives, demonstrating that these large-scale surveys are a critical step towards characterizing less-resolved branches of the eukaryotic tree of life.

The continuing expansion of global-scale meta'omic surveys, such as BioGeoTraces Biller et al. (2018) and Bio-GO-SHIP (Ustick et al., 2021), highlights the importance of developing scalable and automated methods to enable more complete analysis of these data. Metagenomic pipelines that specifically integrate steps for handling eukaryotic biology, such as the EukHeist pipeline, are vital as eukaryotes are important members of microbial communities, ranging from the ocean, to soil (Bailly et al., 2007) and human- (Lukeš et al., 2015) and animal-associated (Campo et al., 2019) environments. The application of eukaryotic-sensitive methods such as EukHeist to other systems stands to greatly increase our understanding of the diversity and function of the "eukaryome".

Materials and Methods

Data acquisition

The metagenomic and metatranscriptomic data corresponding to the size fractions dominated by eukaryotic organisms ranging from microbial eukaryotes and zooplankton ($0.8-2000\mu m$) as originally published by Carradec et al. (2018) were retrieved from European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) under the accession numbers PRJEB4352 (large size fraction metagenomic data) and PRJEB6603 (large size fraction metatranscriptomic data) on November 20, 2018. Only samples with paired end reads (forward and reverse) were used in the subsequent analyses (Supplementary Table 1). After an initial sample-to-sample comparison with sourmash (`sourmash compare -k 31 -scaled 10000`) (Brown and Irber, 2016) (Figure S3), it was determined that samples largely clustered by depth and size fraction. Samples were grouped for co-assembly by size fraction ($0.8-5\mu m$, $5-20\mu m$, $20-180\mu m$, and $180-2000\mu m$) as per Carradec et al. (2018), depth or sample type (surface (SRF), deep chlorophyll maximum (DCM), mesopelagic (MES), mixed surface sample (MIX), and filtered seawater (FSW)), and geographic location (Supplementary Table 1). In cases where a sample did not fall directly within one of the size classes, it was assigned to an existing size class based on the upper μm limit of the sample. This grouping resulted in the combination of 824 cleaned, paired FASTQ files samples into 94 distinct co-assembly groups, which were used downstream for co-assembly (Supplementary Table 1).

EukHeist pipeline for metagenome assembly and binning

The metagenomic analysis, assembly, binning, and all associated quality control steps were carried out with a bioinformatic pipeline, EukHeist, that enables user-guided analysis of stand-alone metagenomic or paired metagenomic and metatranscriptomic sequence data. EukHeist is a streamlined and

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

scalable pipeline currently based on the Snakemake workflow engine (Koster and Rahmann, 2012) that is configured to facilitate deployment on local HPC systems. Figure S2 outlines the structure and outputs of the existing EukHeist pipeline. EukHeist is designed to retrieve and identify both eukaryotic and prokaryotic MAGs from large, metagenomic and metatranscriptomic datasets (Figure S2). EukHeist takes input of sequence meta-data, user-specified assembly pairings (co-assembly groups), and raw sequence files, and returns MAGs that are characterized as either likely eukaryotic or prokaryotic.

Here, all raw sequences accessed from the EMBL-EBI were quality assessed with FastQC and MultiQC (Andrews, 2010). Sequences were trimmed using Trimmomatic (v. 0.36; parameters: `ILLUMINACLIP:2:30:7, LEADING:2, TRAILING:2, SLIDINGWINDOW:4:2, MINLEN:50`) (Bolger et al., 2014). Passing mate paired reads were maintained for assembly and downstream analyses. Quality trimmed reads co-assembled based on assembly groups (Supplementary Table 1) with MEGAHIT (v1.1.3, parameters: `k= 29,39,59,79,99,119`) (Li et al., 2015). Basic statistics were assessed for all assemblies with Quast (v. 5.0.2) (Gurevich et al., 2013) (Supplementary Table 1). Cleaned reads from assembly-group-associated metagenomic and metatranscriptomic samples were mapped back against the assemblies with bwa mem (v.0.7.17) (Li and Durbin, 2010). The bwa-derived abundances were summarized with MetaBat2 (v. 2.12.1) script `jgi_summarize_bam_contig_depths` (with default parameters). The output contig abundance tables were used along with tetranucleotide frequencies to associate contigs into putative genomic bins using MetaBat2 (v. 2.12.1) (Kang et al., 2019). The Snakemake profile used to conduct this analysis is available at `https://www.github.com/alexanderlabwhoi/tara-euk-metag`. A generalized version of the Snakemake pipeline (called EukHeist) that might be readily applied to other datasets is available at `https://www.github.com/alexanderlabwhoi/EukHeist`. MAGs here are subsequently named and referred to as **T**ara **O**ceans **P**article **A**ssociated MAGs (TOPAZ) and are individually named based on their assembly group (Supplementary Tables 2 and 3).

Identification of putative Eukaryotic MAGs

The binning process described above recovered a total of 16,385 putative bins. These bins were screened to identify high completion eukaryotic and prokaryotic bins. All bins were first screened for length, assuming that eukaryotic bins would likely be greater than 2.5Mbp in size (modeled off of the size of the smallest known eukaryotic genome, $\sim$ 2.3Mbp *Microsporidian Encephalitozoon intestinalis* (Corradi et al., 2010)). Bins larger than 2.5Mbp were screened for relative eukaryotic content using EukRep (West et al., 2018), a k-mer based strategy that estimates the likely domain-origin of metagenomic contigs. EukRep was used to classify the relative proportion of eukaryotic and prokaryotic content in each bin in a contig-by-contig manner. This approach identified 907 candidate eukaryotic bins that were greater than 2.5Mb in length and estimated to have more than 90% eukaryotic content by length. Protein coding domains were predicted in all 907 putative eukaryotic bins using EukMetaSanity (Neely et al., 2021).

Protein prediction in Eukaryotic MAGs with EukMetaSanity

Taxonomy. The MMseqs2 v12.113e3 (Steinegger and Söding, 2017, 2018; Mirdita et al., 2019) `taxonomy` module (parameters: `-s 7 -min-seq-id 0.40 -c 0.3 -cov-mode 0`) was used to pro-

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

vide a first-pass taxonomic assignment of the input MAG for use in a downstream element of EukMetaSanity pipeline that requires an input NCBI taxon id or a taxonomic level (i.e. Order, Family, etc.). We created a custom database comprising both OrthoDB (Kriventseva et al., 2018) and MMETSP (Keeling et al., 2014) protein databases (OrthoDB-MMETSP) that integrates NCBI taxon ids. MMseqs2 was used to query each MAG against the OrthoDB-MMETSP database to identify a first-pass taxonomic assignment. The lowest common ancestor of top scoring hits was identified to provide taxonomic assignment to each candidate eukaryotic bin. The `taxonomyreport` module generates a taxon tree that includes the percent of MMseqs mappings that correspond to each taxonomic level. A taxonomic identifier and scientific name are selected to the strain level or when total mapping exceeds 8%, whichever comes first. The assigned NCBI taxon id is retained for downstream analyses.

Repeats identification. RepeatModeler (Flynn et al., 2020a; Smit and Hubley, 2008-2015) was used to provide *ab initio* prediction of transposable elements, including short and long interspersed nuclear repeats, as well as other DNA transposons, small RNA, and satellite repeats. RepeatMasker (Smit et al., 2013-2015) was then used to hard-mask these identified regions, as well as any Family-level (as identified above) repeats from the DFam 3.2 database (Flynn et al., 2020b). RepeatMasker commands `ProcessRepeats` (parameter: `-nolow`) and `rmOutToGff3` (parameter: `-nolow`) were used to output masked sequences (excluding low-complexity repeat DNA from the mask) as FASTA and gene-finding format (GFF3) files, respectively.

***Ab initio* prediction.** GeneMark (Lomsadze et al., 2005) was used to generate *ab initio* gene predictions with the repeat-masked eukaryotic candidate bin sequences output from the prior step. The GeneMark subprogram ProtHint attempts to use Order-level proteins from OrthoDB-MMETSP database to generate intron splice-site predictions for *ab initio* modeling using GeneMark EP (Bruna et al., 2020). If ProtHint fails to generate predictions, then GeneMark will default to ES mode. Due to the fragmented nature of metagenomic assemblies, the prediction parameter stringency was drastically reduced relative to what is recommended for draft genome projects (parameters: `-min_contig 500 -min_contig_in_predict 500 -min_gene_in_predict 100`). These parameters can be easily modified within the EukMetaSanity config file. GeneMark outputs predictions of protein coding sequences (CDS) and exon/intron structure as GFF3 files.

Integrating protein evidence. MetaEuk (Levy Karin et al., 2020) was used to directly map the repeat-masked eukaryotic candidate bins sequences against proteins from the MMETSP (Keeling et al., 2014; Johnson et al., 2018) and eukaryotes included in the OrthoDB v10 dataset (Kriventseva et al., 2018), hereafter referred to as the OrthoDB-MMETSP database. MetaEuk `easy-predict` (parameters: `-min-length 30 -metaeuk-eval 0.0001 -s 7 -cov-mode 0 -c 0.3 -e 100 -max-overlap 0`) used Order-level proteins to identify putative CDS and exon/intron structure. MetaEuk encodes this output as headers in FASTA sequences that are then parsed into GFF3 files.

Merging final results. GFF3 output from the previous *ab initio* and MetaEuk protein evidence steps were input into Gffread (Pertea and Pertea, 2020) (parameters: `-G -merge`) to localize predictions from both lines of evidence into a single GFF3 output file. Each locus was then merged together using a Python (Foundation) script and the BioPython API (Cock et al., 2009) within EukMetaSanity.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

The set of *ab initio* generated exons in each locus is used as a prediction of the underlying exon/intron structure of the gene locus to which it is assigned. If there are any protein-evidence-generated exons present at the same locus, and if the total numbers of exons predicted by each line of evidence have $\geq$ 70% agreement, *ab initio* generated exons lacking a corresponding protein-evidence-generated exon are removed (the first and last exon(s) of a locus are not removed). Conversely, any protein-evidence-generated exon present that lacks a corresponding *ab initio* generated exon is added to the predicted exon/intron structure. The final gene structure for each locus is then processed into GFF3 and FASTA format.

Functional and taxonomic annotation of eukaryotic MAGs

Predicted proteins from EukMetaSanity were annotated for function against protein families in Pfam with PfamScan (Finn et al., 2014) and KEGG using kofamscan (Kanehisa, 2019; Aramaki et al., 2019) (Supplementary Tables 7 and 8). The relative completeness and contamination of each putative Eukaryotic MAG was assessed based on protein content using BUSCO v 4.0.5 against the eukaryota_odb10 gene set using default parameters (Simao et al., 2015) and EukCC v 0.2 using the EukCC database (created 22 October 2019 (Saary et al., 2020)). Annotation and completeness assessment were carried out using a EukHeist-Annotate (`https://www.github.com/halexand/EukHeist-annotate`). EukCC (Saary et al., 2020) was also used to calculate MAG completeness and contamination. The average completeness across groups increased in all cases with EukCC except for metazoans, which on average had a lower estimated completeness (Figure S10).

The taxonomic affiliation of the high- and low-completion bins was estimated using MMSeqs taxonomy through EukMetaSanity and EUKulele (Krinos et al., 2021), an annotation tool that takes a protein-consensus approach, leveraging a Last Common Ancestor (LCA) estimation of protein taxonomy, as well as MMSeqs2 taxonomy module (Steinegger and Söding, 2017, 2018; Mirdita et al., 2019). Taxonomic level estimation in EUKulele was assessed based on e-value derived best-hits, where percent id was used as a means of assessing taxonomic level, with the following cutoffs: species, >95%; genus, 95-80%; family, 80-65%; order, 65-50%; class, 50-30% modeled off of Carradec et al. (2018). All MAGs were searched against the MarMetZoan combining the MarRef, MMETSP, and metazoan orthoDB databases (Johnson et al., 2018; Keeling et al., 2014; Kriventseva et al., 2018; Klemetsen et al., 2017). This database is available for download through EUKulele.

Phylogeny of eukaryotic MAGs

A total of 49 BUSCO proteins were found to be present across 80% or more of the highly complete eukaryotic TOPAZ MAGs and were selected for the construction of the tree. Amino acid sequences from all genomes and transcriptomes of interest were collected and aligned individually using mafft (v7.471) (parameters: `-thread -8 -auto`) (Katoh and Standley, 2013). Individual protein alignments were trimmed to remove sections of the alignment that were poorly aligned with trimAl (v1.4.rev15) (parameters: `-automated1`) (Capella-Gutierrez et al., 2009a). Protein sequences were then concatenated and trimmed again with trimAl (parameters: `-automated1`). A final tree was then constructed using RAxML (v 8.2.12; parameters: `raxmlHPC-PTHREADS-SSE3 -T 16 -f a -m PROTGAMMAJTT -N 100 -p 42 -x 42`) (Stamatakis, 2014). The amino acid alignment and construction was controlled with a Snakemake workflow: `https://github.com/halexand/`

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

`BUSCO-MAG-Phylogeny/`. Trees were visualized and finalized with iTOL (Letunic and Bork, 2016).

Prokaryotic MAG assessment and analysis

The 15,478 bins that were not identified as putative eukaryotic bins based on length and EukRep metrics were screened to identify quality prokaryotic bins. The quality and phylogenetic-association of these bins was assessed with a modified version of MAGpy (Stewart et al., 2019), which was altered to include taxonomic annotation with GTDB-TK v.0.3.2 (Chaumeil et al., 2019). Bins were assessed based on single copy ortholog content with CheckM v (Parks et al., 2015) to identify 2 different bin quality sets: 1) high-quality (HQ) prokaryotic bins (>90% completeness, <5% contamination), and 2) medium-quality (MQ) prokaryotic bins (90-75% completeness, <10% contamination). A total of 4022 prokaryotic MAGs met the above criteria. A final set of 2,407 non-redundant (NR) HQ-MQ MAGs were identified using dRep v2.6.2 (Olm et al., 2017), which performs pairwise genome comparisons in two steps. First, a rapid primary algorithm, Mash v1.1.1 (Ondov et al., 2016) is applied. Genomes with Mash values equivalent to 90% Average Nucleotide Identity (ANI) or higher were then compared with MUMmer v3.23 (Marçais et al., 2018). Genomes with ANI $\geq$ 99% were considered to belong to the same cluster. The best representative MAGs were selected based on the dRep default scoring equation (Olm et al., 2017). Out of the final set of 2,407 NR MAGs, 716 were HQ. The same pipeline was used to determine the HQ and MQ NR MAGs reconstructed from the *Tara* Oceans metagenomes in previous studies (Tully et al., 2018; Parks et al., 2017; Delmont et al., 2018).

Phylogeny of bacterial non-redundant high-quality MAGs

Only 5 out of the 716 HQ NR MAGs were found to belong to Archaea, thus only bacterial MAGs were used for the construction of the phylogenetic tree with GToTree v.1.4.10 (Lee, 2019) and the gene set (HMM file) for Bacteria (74 targets). GToTree pipeline uses Prodigal v2.6.3 (Hyatt et al., 2010) to retrieve the coding sequences in the genomes, and HMMER3 v3.2.1 (Eddy, 2011) to identify the target genes based on the provided HMM file. MUSCLE v3.8 (Edgar, 2004) was then used for the gene alignments, and Trimal v1.4 (Capella-Gutierrez et al., 2009b) for trimming. The concatenated aligned is used for the tree constructions using FastTree v2.1 (Price et al., 2010). Three genomes were excluded from the analysis due to having too few of the target genes. The tree was visualized using the Interactive Tree of Life (iToL) (Letunic and Bork, 2016).

Prokaryote MAG phylogeny comparison

A set of 8,644 microbial genomes were collected from the MarDB database (Klemetsen et al., 2017)(accessed 31 May 2018) encompassing the publicly available marine microbial genomes. Genomes were assessed using CheckM v1.1.1 (Parks et al., 2015)(parameters: `lineage_wf`) and genomes estimated to be <70% complete or >10% contamination were discarded. The remaining genomes (n = 5,878) were assessed using CompareM v0.0.23 (parameters: `aai_wf`; `https://github.com/dparks1134/CompareM`) and near identical genomes were identified using a cutoff of $\geq$ 95% average amino acid identity (AAI) with $\geq$ 85% orthologous fraction (determined as one standard deviation from the average orthologous fraction for genomes with $97-100\%$ AAI). Based on CheckM quality, the genome with the highest completion and/or lowest contamination were retained. From the remaining genomes (n = 3,843), all MAGs derived from the *Tara* Oceans dataset, specifically from Tully et al. Tully et al.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(2018) and Parks et al. Parks et al. (2017), were removed. The remaining genomes (n = 2,275) would be used to form the base of a phylogenetic tree representing the available genome diversity prior to the release of previous *Tara* Oceans related MAG datasets Tully et al. (2018); Parks et al. (2017); Delmont et al. (2018), termed the "neutral" component of subsequent phylogenetic trees.

For the comparisons, phylogenetic trees were constructed using GToTree v1.4.7 (Lee, 2019) (default parameters; 25 Bacteria_and_Archaea markers). Any genome added to a tree that did not meet the default 50% marker presence requirement was excluded from that tree. Five iterations of phylogenetic trees were constructed using the neutral genomes paired with each *Tara* Oceans MAG dataset, the high-quality TOPAZ prokaryote MAGs, and the medium-quality TOPAZ prokaryote MAGs, individually, and two larger trees were constructed containing all neutral genomes and *Tara* Oceans MAGs, with additions of either high- or medium-quality TOPAZ MAGs. Phylogenetic trees were assessed using genometreetk (parameter: `pd`; `https://github.com/dparks1134/GenomeTreeTk`) to determine the phylogenetic diversity (i.e., the total branch length traversed by a set of leaves) and phylogenetic gain (i.e., the additional branch length added by a set of leaves) (Parks et al., 2017) for each set of MAGs compared against the neutral genomes and for the TOPAZ prokaryote MAGs compared against the neutral genomes and the other *Tara* Oceans MAGs.

MAG abundance profiling

Raw reads from all metagenomic and metatranscriptomic samples were mapped against the eukaryotic and prokaryotic TOPAZ MAGs to estimate relative abundances with CoverM (v. 0.5.0; parameters: `-min-read-percent-identity 0.95 -min-read-aligned-percent 0.75 -min-covered-fraction 0 -contig-end-exclusion 75 -trim-min 0.05 -trim-max 0.95 -proper-pairs-only`; `https://github.com/wwood/CoverM`). The total number of reads mapped to each MAG was then used to calculate Reads Per Kilobase Million (*RPKM*), where for some MAG, i : $RPKM_i = X_i/l_i N 10^9$, with X = total number of reads recruiting to a MAG, l = length of MAG in Kb, N = total number of trimmed reads mapping to a sample in millions. We also calculated counts per million (CPM), a normalization of the RPKM to the sum of all RPKMs in a sample. CPM, a modification of transcripts per million (TPM) was first proposed by Wagner et al. (2012) as an alternative to RPKM that reduces statistical bias. The metric has since been applied to metagenomics data, sometimes called GPM (genes per million) (Gradoville et al., 2017).

Nutritional modelling

To predict the trophic mode of the high quality TOPAZ eukaryotic MAGs (n=485), a Random Forest model (Breiman, 2001) was constructed and calibrated using the ranger (Wright and Ziegler, 2017) and tuneRanger packages in R (Probst et al., 2018), respectively. The model was trained using KEGG Orthology (KO) annotations (Kanehisa, 2019) from a manually-curated reference trophic mode transcriptomic dataset consisting of the MMETSP (Keeling et al., 2014) and EukProt (Richter et al., 2020) (Supplementary Table 5). 644 of the transcriptomes in this reference dataset came from the MMETSP (Keeling et al., 2014), after 22 transcriptomes were removed due to low coverage of KEGG and Pfam annotations (Finn et al., 2014). The remaining 266 came from the EukProt database (Richter et al., 2020), after 162 were removed due having fewer than 500 present KOs. Nutritional strategy (phototrophy, heterotrophy, or mixotrophy) was assessed for each reference transcriptome individually

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

based on the literature, 25% of the combined reference transcriptomes were excluded from model training as testing data.

A subset of KEGG Orthologs (KOs) that were predictive for trophic mode classification was determined computationally with the vita variable selection package in R (Janitza et al., 2016) (Supplementary Table 6), which was tested and justified by Degenhardt et al. Degenhardt et al. (2017). This process was carried out by the algorithm without regard to the predicted function of the KOs, but we found that many of these KOs were implicated in carbohydrate and energy metabolism, with preference for those KOs that differ strongly between heterotrophs and phototrophs (particularly for energy metabolism; Figure S25). The model was built using the selected KOs ($n = 1787$ of a total 21585 KOs) with the 75% of the combined database assigned as training data.

Additionally, we developed a secondary metric for assessing the extent of heterotrophy of a transcriptome or MAG. As opposed to the trinary classification scheme of the Random Forest model, this approach quantifies the extent that the MAG aligns with heterotrophic, phototrophic, or mixotrophic references by assigning a composite score. We calculated the likelihood of vita selected KOs used in the Random Forest model above to be present within heterotrophic, phototrophic, or mixotrophic reference transcriptomes. Three scores (h, p, m), one corresponding to each trophic mode, were hence calculated for each vita-selected KO (k) ($n = 1787$) (Supplementary Table 6). In Equation (1), K is the number of references the KO was present in for each trophic mode category, while n is the total number of references available for each trophic mode category.

$$h_k = \mathbf{g}\left(\frac{K_{\text{het}}}{n_{\text{het}}}\right) \tag{1}$$

$$p_k = \mathbf{g}\left(\frac{K_{\text{photo}}}{n_{\text{photo}}}\right) \tag{2}$$

$$m_k = \mathbf{g}\left(\frac{K_{\text{mixo}}}{n_{\text{mixo}}}\right) \tag{3}$$

$$\text{where, } \mathbf{g}(a) = \begin{cases} a & \text{if } a > 0.5 \\ -(0.5 - a) & \text{otherwise} \end{cases} \tag{4}$$

If a given KO occurred in fewer than 50% of the reference transcriptomes for a trophic mode, it was considered not to be characteristic of that trophic mode and as such the score, which we represent as the variable a, the ratio of the present KOs to the total for the subset of transcriptomes annotated some trophic mode (Equation (11)), was transformed ($-(0.5-a)$, if $a < 0.5$), to reflect the absence without valuing absence over presence. In the test transcriptome dataset, the ratio-transformed scores were negated when a given KO was absent from the transcriptome. For instance, if a KO was absent from 90% of reference transcriptomes assigned to heterotrophy ($a = 0.1$), and absent in the MAG or transcriptome being evaluated, it would receive a score of $h_k = -1 * (-(0.5-0.1)) = 0.4$ (Equation (1)) for that KO. This reflects that the absence of the KO in the evaluated MAG or transcriptome aligned well with the high probability that the KO was absent among the reference transcriptomes.

The scores for all KOs selected by vita were then used to scale the presence/absence patterns observed across transcriptomes and MAGs. Thus, for each transcriptome or MAG a single score was calculated

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

for each trophic mode heterotrophy (H), phototrophy (P), and mixotrophy (M) for all KOs present within the transcriptome or MAG (K):

$$H = \sum_{k \in K} h_k \tag{5}$$

$$P = \sum_{k \in K} p_k \tag{6}$$

$$M = \sum_{k \in K} m_k \tag{7}$$

These calculated values can then be aggregated to a composite heterotrophy score (H_{ind}) (Supplementary Table 9). The score was computed as follows:

$$H_{ind} = \begin{cases} -1^{(H-P)}\sqrt{(H-P)^2}, & \text{if } M - \max(H,P) < 50, \\ \frac{-1^{(H-P)}\sqrt{(H-P)^2}}{M}, & \text{if } M - \max(H,P) \geq 50 \end{cases} \tag{8}$$

Ecological analysis of SAR and Dictyochophyceae MAGs

24 highly-complete TOPAZ eukaryotic MAGs were subset to characterize the biogeography and putative trophic modes of *Dictyochophyceae* and closely-related stramenopiles. MAGs included 11 *Dictyochophyceae* MAGs and 13 MAGs belonging to a phylogenetically similar branch (Eukaryotic SAR; derived from the `MMSeqs` and `EUKulele` assignment). Metagenomic CPM abundance of MAGs was used to compare the biogeography and distribution of *Dictyochophyceae* and closely-related stramenopile MAGs. MAGs were further classified based on trophic prediction and heterotrophy score.

To investigate physiological potential, quality trimmed metatranscriptome reads were mapped using `Salmon` Patro et al. (2017) to the 24 MAGs. Comparison of of the putative metabolic capabilities of the predicted phototrophic versus heterotrophic MAGs was conducted in R3.6.1, where MAGs were clustered using the average distance between MAGs based on the presence and absence of known orthologs (KO annotations (Kanehisa, 2019)). Principle coordinate analysis of center log-ratio transformed TPM abundances of mapped reads from the smallest size fraction ($0.8 - 5.00\mu m$) of all surface samples revealed the degree of overlap between *Dictyochophyceae* and closely-related Stramenopile MAGs. Specific genes shared among all MAGs, shared across predicted trophic modes, and those vita selected KOs used in the trophic model (described above) were further targeted. All analyses described above are available at `https://alexanderlabwhoi.github.io/2021-TOPAZ-MAG-Figures/`.

Network Analysis

To identify co-occurring MAGs across the stations surveyed by *Tara* Oceans, the CPM abundance of each highly-complete eukaryotic MAG (> 30% BUSCO completeness) and each non-redundant, highly complete bacterial MAG was assessed at each station at all available depths and size fractions as described above. CPM was used because of the power of this metric for comparing samples directly: the sum of all CPM values per sample will be the same, as sequencing depth is accounted for after gene length. This makes it easier to compare the abundances of MAGs originally recovered

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

from different sites (Gradoville et al., 2017). A Spearman correlation matrix was generated to identify monotonic relationships between MAGs. Correlations were filtered based first on p-value, using the Šidák correction (Šidák, 1967), a slightly less stringent metric than the Bonferroni correction. The Šidák correlation adjusts for multiple comparisons and is given by $p < 1-(1-\alpha)^{1/n}$, where n is the total number of comparisons, and α is the significance value, in this case 0.05. We considered only those correlations within the 90th percentile of CPM correlations, thus correlations with absolute value less than 0.504 were removed from the analysis. Subsequently, we further filtered interactions to those with coefficient of correlation > 0.70 for the construction of the network diagram. Because it was expected for several of the eukaryotic MAGs to be closely related (based on ANI), the relationships in the network were further filtered to exclude interactions between MAGs of exceedingly high similarity (having both 99% ANI similarity and > 0.70 coefficient of correlation in the network analysis) (Supplementary Table 12). ANI-based group members tended to have identical taxonomic classifications: only 2 of 94 clusters had different classifications at the order level per EUKulele (Figure S30).

We generated a network from this reduced set of labeled interactions (cut off at > 0.70 coefficient of correlation, focusing on interactions between eukaryotes and prokaryotes or eukaryotes and eukaryotes, and using ANI-based clusters instead of MAG names when applicable) using `igraph` (Csardi and Nepusz, 2006; Team) (Supplementary Table 11). Communities of highly associated MAGs were identified using a modularity optimization algorithm introduced in Blondel et al. (2008) and implemented in `igraph` (Csardi and Nepusz, 2006).

We assessed the connectedness within and between communities by calculating a connectedness metric as follows. For the connectedness within a community (one community to itself), we identified the number of "dense" connections by counting up the total number of links found between community members, regardless of how many times the particular MAG had been connected to its own community, and divided that number by the total possible "dense", meaning the number of connections which would exist if all community members were connected to all other community members. Between different communities, we defined connectedness by qualifying that a "connection" is made the first time each MAG from a given community is linked to another community, and calculated this quantity by dividing the number of realized links between community members by the maximum total size of the two involved communities (Figure 5 b; Equation (9) - Equation (11)).

$$C_{x,x} = \frac{\Sigma_{x=1}^{n_x}\Sigma_{y=1}^{n_y} f(x,x)}{\frac{n_x(n_{y-1})}{2}} \tag{9}$$

$$C_{x,y} = \frac{\Sigma_{x=1}^{n_x}\Sigma_{y=x+1}^{n_y} f(x,y)}{\max(n_x,n_y)} \tag{10}$$

$$f(a,b) = \begin{cases} 1 & \text{if } a \text{ and } b \text{ are connected} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

We calculated Spearman correlation coefficients for the relationship between the abundance of communities between stations and several environmental parameters of interest from the *Tara* Oceans

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

metadata (Pesant et al., 2015; Tara Oceans Consortium and Tara Oceans Expedition, 2016) (Figure 5). We considered the measured physical and chemical parameters, the modeled mesoscale physical oceanographic parameters, and averaged remote sensing products (Tara Oceans Consortium and Tara Oceans Expedition, 2016; d'Ovidio et al., 2010; Pesant et al., 2015). We adjusted the p-value of these comparisons using a Bonferroni adjustment within the statistics package in R (Team).

Data and code availability

The eukaryotic and prokaryotic TOPAZ MAGs and Supplementary Tables 1-13 are available through the Open Science Framework (OSF) at `https://osf.io/gm564/` with the DOI: 10.17605/OSF.IO/GM564. EukHeist, which was used to recover the reported TOPAZ MAGs can be found at `https://github.com/AlexanderLabWHOI/EukHeist` and EukMetaSanity which was used for protein prediction in eukaryotic MAGs can be found at `https://github.com/cjneely10/EukMetaSanity`. Code used to generate the figures in this paper can be found at `https://github.com/AlexanderLabWHOI/2021-TOPAZ-MAG-Figures`. An interactive visualizer for the TOPAZ eukaryotic MAGs is available at `https://share.streamlit.io/cjneely10/tara-analysis/main/TARAVisualize/main.py` with source code at `https://github.com/cjneely10/TARA-Analysis`.

Acknowledgements

This research would not have been possible without the community driven efforts to provide open and freely available data by the *Tara* Oceans Consortium. This research was supported by a National Science Foundation grant (NSF-OCE-1948025) to HA and a WHOI Independent Research and Development award to HA. SKH was supported through a Postdoctoral Fellowship (OCE-0939564) provided through the NSF Center for Dark Energy Biosphere Investigations and through an NSF grant (OCE-1947776). AIK was supported by the Computational Science Graduate Fellowship (DOE; DE-SC0020347). BJT was supported by the Center for Dark Energy Biosphere Investigations (C-DEBI) through NSF-OCE-0939654.

Author contributions statement

HA and SKH conceived of and designed the study. HA carried out the assembly and binning. HA, AIK, SKH, MP, TR, and CJN, and BJT analyzed the data. HA and SKH wrote the manuscript with input from all authors. All authors edited and commented on the manuscript.

Ethics Declaration

The authors declare no conflicts of interest.

References

H. Alexander, M. Rouco, S. T. Haley, S. T. Wilson, D. M. Karl, and S. T. Dyhrman. Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. *Proceedings of the National Academy of Sciences*, 112(44):E5972–E5979, nov 2015. ISSN

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

0027-8424. doi: 10.1073/pnas.1518165112. URL http://www.pnas.org/content/early/2015/10/09/1518165112.abstract{%}5Cnhttp://www.pnas.org/lookup/doi/10.1073/pnas.1518165112http://www.pnas.org/lookup/doi/10.1073/pnas.1518165112http://www.pnas.org/content/early/2015/10/09/1518165112.abstracthttp://www.p.

A. Almeida, A. L. Mitchell, M. Boland, S. C. Forster, G. B. Gloor, A. Tarkowska, T. D. Lawley, and R. D. Finn. A new genomic blueprint of the human gut microbiota. *Nature*, 568(7753):499–504, 4 2019. ISSN 14764687. doi: 10.1038/s41586-019-0965-1.

J. Alneberg, B. S. Bjarnason, I. De Bruijn, M. Schirmer, J. Quick, U. Z. Ijaz, L. Lahti, N. J. Loman, A. F. Andersson, and C. Quince. Binning metagenomic contigs by coverage and composition. *Nature Methods*, 11(11):1144–1146, 10 2014. ISSN 15487105. doi: 10.1038/nmeth.3103.

S. Andrews. Fastqc: A quality control tool for high throughput sequence data., 2010. URL http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:FastQC+a+quality+control+tool+for+high+throughput+sequence+data.#0. [Online; accessed 2014-03-31].

T. Aramaki, R. Blanc-Mathieu, H. Endo, K. Ohkubo, M. Kanehisa, S. Goto, and H. Ogata. KofamKOALA: KEGG ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics*, 36(7):2251–2252, nov 2019. doi: 10.1093/bioinformatics/btz859. URL https://doi.org/10.1093%2Fbioinformatics%2Fbtz859.

J. Bailly, L. Fraissinet-Tachet, M.-C. Verner, J.-C. Debaud, M. Lemaire, M. Wésolowski-Louvel, and R. Marmeisse. Soil eukaryotic functional diversity, a metatranscriptomic approach. *The ISME journal*, 1(7):632–642, 2007.

A. Bashiri, M. Ghazisaeedi, R. Safdari, L. Shahmoradi, and H. Ehtesham. Improving the prediction of survival in cancer patients by using machine learning techniques: experience of gene expression data: a narrative review. *Iranian Journal of Public Health*, 46(2):165, 2017.

D. Beisser, N. Graupner, C. Bock, S. Wodniok, L. Grossmann, M. Vos, B. Sures, S. Rahmann, and J. Boenigk. Comprehensive transcriptome analysis provides new insights into nutritional strategies and phylogenetic relationships of chrysophytes. 5:e2832, 2017. ISSN 2167-8359. doi: 10.7717/peerj.2832. URL https://peerj.com/articles/2832.

S. J. Biller, P. M. Berube, K. Dooley, M. Williams, B. M. Satinsky, T. Hackl, S. L. Hogle, A. Coe, K. Bergauer, H. A. Bouman, T. J. Browning, D. D. Corte, C. Hassler, D. Hulston, J. E. Jacquot, E. W. Maas, T. Reinthaler, E. Sintes, T. Yokokawa, and S. W. Chisholm. Marine microbial metagenomes sampled across space and time. *Sci Data*, 5(1), sep 2018. doi: 10.1038/sdata.2018.176. URL https://doi.org/10.1038%2Fsdata.2018.176.

V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiments*, 2008(10):P10008, 2008.

A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 8 2014. ISSN 14602059. doi: 10.1093/bioinformatics/btu170.

C. Bowler, A. E. Allen, J. H. Badger, J. Grimwood, K. Jabbari, A. Kuo, U. Maheswari, C. Martens, F. Maumus, R. P. Otillar, E. Rayko, A. Salamov, K. Vandepoele, B. Beszteri, A. Gruber, M. Heijde,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

M. Katinka, T. Mock, K. Valentin, F. Verret, J. A. Berges, C. Brownlee, J.-P. Cadoret, A. Chiovitti, C. J. Choi, S. Coesel, A. D. Martino, J. C. Detter, C. Durkin, A. Falciatore, J. Fournet, M. Haruta, M. J. J. Huysman, B. D. Jenkins, K. Jiroutova, R. E. Jorgensen, Y. Joubert, A. Kaplan, N. Kröger, P. G. Kroth, J. L. Roche, E. Lindquist, M. Lommer, V. Martin–Jézéquel, P. J. Lopez, S. Lucas, M. Mangogna, K. McGinnis, L. K. Medlin, A. Montsant, M.-P. O. Secq, C. Napoli, M. Obornik, M. S. Parker, J.-L. Petit, B. M. Porcel, N. Poulsen, M. Robison, L. Rychlewski, T. A. Rynearson, J. Schmutz, H. Shapiro, M. Siaut, M. Stanley, M. R. Sussman, A. R. Taylor, A. Vardi, P. von Dassow, W. Vyverman, A. Willis, L. S. Wyrwicz, D. S. Rokhsar, J. Weissenbach, E. V. Armbrust, B. R. Green, Y. V. de Peer, and I. V. Grigoriev. The phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, 456(7219):239–244, oct 2008. doi: 10.1038/nature07410. URL `https://doi.org/10.1038%2Fnature07410`.

L. Breiman. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/a:1010933404324. URL `https://doi.org/10.1023%2Fa%3A1010933404324`.

C. T. Brown and L. Irber. sourmash: a library for MinHash sketching of DNA. *JOSS*, 1(5):27, sep 2016. doi: 10.21105/joss.00027. URL `https://doi.org/10.21105%2Fjoss.00027`.

T. Bruna, A. Lomsadze, and M. Borodovsky. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics and Bioinformatics*, 2(2), jun 2020. ISSN 2631-9268. doi: 10.1093/nargab/lqaa026. URL `https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqaa026/5836691`.

F. Burki, A. J. Roger, M. W. Brown, and A. G. Simpson. The new tree of eukaryotes. *Trends in Ecology & Evolution*, 35(1):43–55, jan 2020. doi: 10.1016/j.tree.2019.08.008. URL `https://doi.org/10.1016%2Fj.tree.2019.08.008`.

J. A. Burns, A. A. Pittis, and E. Kim. Gene-based predictive models of trophic modes suggest Asgard archaea are not phagocytotic. *Nature Ecology & Evolution*, 2(4):697–704, 2018.

J. Campo, D. Bass, and P. J. Keeling. The eukaryome: Diversity and role of microeukaryotic organisms associated with animal hosts. *Funct Ecol*, 34(10):2045–2054, dec 2019. doi: 10.1111/1365-2435.13490. URL `https://doi.org/10.1111%2F1365-2435.13490`.

S. Capella-Gutierrez, J. M. Silla-Martinez, and T. Gabaldon. trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, 8 2009a. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp348.

S. Capella-Gutierrez, J. M. Silla-Martinez, and T. Gabaldon. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, jun 2009b. doi: 10.1093/bioinformatics/btp348. URL `https://doi.org/10.1093%2Fbioinformatics%2Fbtp348`.

D. A. Caron and P. D. Countway. Hypotheses on the role of the protistan rare biosphere in a changing world. *Aquatic Microbial Ecology*, 57(3):227–238, 2009. ISSN 09483055. doi: 10.3354/ame01352.

D. A. Caron, P. D. Countway, A. C. Jones, D. Y. Kim, and A. Schnetzer. Marine protistan diversity. *Annual Review of Marine Science*, 4(1):467–493, 2011. ISSN 1941-1405. doi: 10.1146/annurev-marine-120709-142802.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

D. A. Caron, H. Alexander, A. E. Allen, J. M. Archibald, E. V. Armbrust, C. Bachy, C. J. Bell, A. Bharti, S. T. Dyhrman, S. M. Guida, et al. Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology*, 15(1):6–20, 2017.

Q. Carradec, E. Pelletier, C. Da Silva, A. Alberti, Y. Seeleuthner, R. Blanc-Mathieu, G. Lima-Mendez, F. Rocha, L. Tirichine, K. Labadie, A. Kirilovsky, A. Bertrand, S. Engelen, M.-A. Madoui, R. Méheust, J. Poulain, S. Romac, D. J. Richter, G. Yoshikawa, C. Dimier, S. Kandels-Lewis, M. Picheral, S. Searson, S. G. Acinas, E. Boss, M. Follows, G. Gorsky, N. Grimsley, L. Karp-Boss, U. Krzic, S. Pesant, E. G. Reynaud, C. Sardet, M. Sieracki, S. Speich, L. Stemmann, D. Velayoudon, J. Weissenbach, O. Jaillon, J.-M. Aury, E. Karsenti, M. B. Sullivan, S. Sunagawa, P. Bork, F. Not, P. Hingamp, J. Raes, L. Guidi, H. Ogata, C. de Vargas, D. Iudicone, C. Bowler, and P. Wincker. A global ocean atlas of eukaryotic genes. *Nature Communications*, 9(1):373, 2018. ISSN 2041-1723. doi: 10.1038/s41467-017-02342-1.

J. Celli and T. C. Zahrt. Mechanisms of francisella tularensis intracellular pathogenesis. *Cold Spring Harbor Perspectives in Medicine*, 3(4):a010314–a010314, apr 2013. doi: 10.1101/cshperspect.a010314. URL `https://doi.org/10.1101%2Fcshperspect.a010314`.

P.-A. Chaumeil, A. J. Mussig, P. Hugenholtz, and D. H. Parks. GTDB-tk: a toolkit to classify genomes with the genome taxonomy database. *Bioinformatics*, nov 2019. doi: 10.1093/bioinformatics/btz848. URL `https://doi.org/10.1093%2Fbioinformatics%2Fbtz848`.

P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, jun 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp163. URL `https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp163`.

I. H. G. S. Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, oct 2004. doi: 10.1038/nature03001. URL `https://doi.org/10.1038%2Fnature03001`.

N. Corradi, J. F. Pombert, L. Farinelli, E. S. Didier, and P. J. Keeling. The complete sequence of the smallest known nuclear genome from the microsporidian encephalitozoon intestinalis. *Nature Communications*, 1(6), 2010. ISSN 20411723. doi: 10.1038/ncomms1082.

D. D. Corte, A. Srivastava, M. Koski, J. A. L. Garcia, Y. Takaki, T. Yokokawa, T. Nunoura, N. H. Elisabeth, E. Sintes, and G. J. Herndl. Metagenomic insights into zooplankton-associated bacterial communities. *Environ Microbiol*, 20(2):492–505, oct 2017. doi: 10.1111/1462-2920.13944. URL `https://doi.org/10.1111%2F1462-2920.13944`.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006. URL `https://igraph.org`.

A. C. Darby, N.-H. Cho, H.-H. Fuxelius, J. Westberg, and S. G. Andersson. Intracellular pathogens go extreme: genome evolution in the rickettsiales. *Trends in Genetics*, 23(10):511–520, oct 2007. doi: 10.1016/j.tig.2007.08.002. URL `https://doi.org/10.1016%2Fj.tig.2007.08.002`.

C. de Vargas, S. Audic, N. Henry, J. Decelle, F. Mahe, R. Logares, E. Lara, C. Berney, N. L. Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. Chaf-

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

fron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horak, O. Jaillon, G. Lima-Mendez, J. Luke, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, J. Raes, M. E. Sieracki, S. Speich, L. Stemmann, S. Sunagawa, J. Weissenbach, P. Wincker, E. Karsenti, E. Boss, M. Follows, L. Karp-Boss, U. Krzic, E. G. Reynaud, C. Sardet, M. B. Sullivan, and D. V. and. Eukaryotic plankton diversity in the sunlit ocean. *Science*, 348(6237):1261605–1261605, may 2015. doi: 10.1126/science.1261605. URL `https://doi.org/10.1126%2Fscience.1261605`.

F. Degenhardt, S. Seifert, and S. Szymczak. Evaluation of variable selection methods for random forests and omics data sets. *Briefings in Bioinformatics*, 20(2):492–503, oct 2017. doi: 10.1093/bib/bbx124. URL `https://doi.org/10.1093%2Fbib%2Fbbx124`.

T. O. Delmont, C. Quince, A. Shaiber, O. C. Esen, S. T. Lee, M. S. Rappé, S. L. McLellan, S. Lücker, and A. M. Eren. Nitrogen-fixing populations of planctomycetes and proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, 3(7):804–813, 7 2018. ISSN 2058-5276. doi: 10.1038/s41564-018-0176-9.

F. d'Ovidio, S. D. Monte, S. Alvain, Y. Dandonneau, and M. Levy. Fluid dynamical niches of phytoplankton types. *Proceedings of the National Academy of Sciences*, 107(43):18366–18370, oct 2010. doi: 10.1073/pnas.1004620107. URL `https://doi.org/10.1073%2Fpnas.1004620107`.

A. Dufresne, L. Garczarek, and F. Partensky. *Genome Biol*, 6(2):R14, 2005. doi: 10.1186/gb-2005-6-2-r14. URL `https://doi.org/10.1186%2Fgb-2005-6-2-r14`.

F. d'Ovidio, S. De Monte, S. Alvain, Y. Dandonneau, and M. Lévy. Fluid dynamical niches of phytoplankton types. *Proceedings of the National Academy of Sciences*, 107(43):18366–18370, 2010.

S. R. Eddy. Accelerated profile HMM searches. *PLoS Computational Biology*, 7(10):e1002195, oct 2011. doi: 10.1371/journal.pcbi.1002195. URL `https://doi.org/10.1371%2Fjournal.pcbi.1002195`.

R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, mar 2004. doi: 10.1093/nar/gkh340. URL `https://doi.org/10.1093%2Fnar%2Fgkh340`.

V. P. Edgcomb, D. Beaudoin, R. Gast, J. F. Biddle, and A. Teske. Marine subsurface eukaryotes: the fungal majority. *Environmental Microbiology*, 13(1):172–183, aug 2010. doi: 10.1111/j.1462-2920.2010.02318.x. URL `https://doi.org/10.1111%2Fj.1462-2920.2010.02318.x`.

R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, and M. Punta. Pfam: The protein families database. *Nucleic Acids Research*, 42(D1):D222–D230, 1 2014. ISSN 03051048. doi: 10.1093/nar/gkt1223.

J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, 117(17):9451–9457, apr 2020a.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

ISSN 10916490. doi: 10.1073/pnas.1921046117. URL https://www.pnas.org/content/117/17/9451https://www.pnas.org/content/117/17/9451.abstract.

J. M. Flynn, R. Hubley, C. Goubert, J. Rosen, A. G. Clark, C. Feschotte, and A. F. Smit. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, 117(17):9451–9457, apr 2020b. ISSN 10916490. doi: 10.1073/pnas.1921046117. URL https://www.pnas.org/content/117/17/9451https://www.pnas.org/content/117/17/9451.abstract.

K. J. Flynn, A. Mitra, K. Anestis, A. A. Anschütz, A. Calbet, G. D. Ferreira, N. Gypens, P. J. Hansen, U. John, J. L. Martin, J. S. Mansour, M. Maselli, N. Medić, A. Norlin, F. Not, P. Pitta, F. Romano, E. Saiz, L. K. Schneider, W. Stolte, and C. Traboni. Mixotrophic protists and a new paradigm for marine ecology: where does plankton research go now? *Journal of Plankton Research*, 41(4): 375–391, jul 2019. doi: 10.1093/plankt/fbz026. URL https://doi.org/10.1093%2Fplankt%2Ffbz026.

P. S. Foundation. Python language reference, version 3.6. http://www.python.org.

C. R. Giner, V. Balagué, A. K. Krabberø d, I. Ferrera, A. Reñé, E. Garcés, J. M. Gasol, R. Logares, and R. Massana. Quantifying long-term recurrence in planktonic microbial eukaryotes. 28(5):923–935, 2019. ISSN 1365-294X. doi: 10.1111/mec.14929. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.14929.

S. J. Giovannoni, J. C. Thrash, and B. Temperton. Implications of streamlining theory for microbial ecology. *ISME Journal*, 8(8):1553–1565, apr 2014. doi: 10.1038/ismej.2014.60. URL https://doi.org/10.1038%2Fismej.2014.60.

W. Gong, J. Browne, N. Hall, D. Schruth, H. Paerl, and A. Marchetti. Molecular insights into a dinoflagellate bloom. *ISME Journal*, 11(2):439–452, dec 2016. doi: 10.1038/ismej.2016.129. URL https://doi.org/10.1038%2Fismej.2016.129.

M. R. Gradoville, B. C. Crump, R. M. Letelier, M. J. Church, and A. E. White. Microbiome of trichodesmium colonies from the north pacific subtropical gyre. *Front. Microbiol.*, 8, jul 2017. doi: 10.3389/fmicb.2017.01122. URL https://doi.org/10.3389%2Ffmicb.2017.01122.

E. D. Graham, J. F. Heidelberg, and B. J. Tully. Binsanity: unsupervised clustering of environmental microbial assemblies using coverage and affinity propagation. *PeerJ*, 5:e3035, 3 2017. ISSN 2167-8359. doi: 10.7717/peerj.3035. NULL.

A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–1075, feb 2013. doi: 10.1093/bioinformatics/btt086. URL https://doi.org/10.1093%2Fbioinformatics%2Fbtt086.

Y. Hou and S. Lin. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: Gene content estimation for dinoflagellate genomes. *PLoS ONE*, 4(9):e6978, sep 2009. doi: 10.1371/journal.pone.0006978. URL https://doi.org/10.1371%2Fjournal.pone.0006978.

S. K. Hu, Z. Liu, H. Alexander, V. Campbell, P. E. Connell, S. T. Dyhrman, K. B. Heidelberg, and D. A. Caron. Shifting metabolic priorities among key protistan taxa within and below the euphotic

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

zone. *Environmental Microbiology*, 20(8):2865–2879, aug 2018. ISSN 14622912. doi: 10.1111/1462-2920.14259. URL http://doi.wiley.com/10.1111/1462-2920.14259.

D. Hyatt, G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer, and L. J. Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1), mar 2010. doi: 10.1186/1471-2105-11-119. URL https://doi.org/10.1186%2F1471-2105-11-119.

S. Janitza, E. Celik, and A.-L. Boulesteix. A computationally fast variable importance test for random forests for high-dimensional data. *Adv Data Anal Classif*, 12(4):885–915, nov 2016. doi: 10.1007/s11634-016-0276-4. URL https://doi.org/10.1007%2Fs11634-016-0276-4.

V. Jimenez, J. A. Burns, F. L. Gall, F. Not, and D. Vaulot. No evidence of phago-mixotropy in micromonas polaris (mamiellophyceae), the dominant picophytoplankton species in the arctic. *J. Phycol.*, 57(2):435–446, mar 2021. doi: 10.1111/jpy.13125. URL https://doi.org/10.1111%2Fjpy.13125.

L. K. Johnson, H. Alexander, and C. T. Brown. Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *GigaScience*, 12 2018. ISSN 2047-217X. doi: 10.1093/gigascience/giy158.

T. S. Jørgensen, B. L. H. Nielsen, B. Petersen, P. D. Browne, B. W. Hansen, and L. H. Hansen. The whole genome sequence and mRNA transcriptome of the tropical cyclopoid copepod apocyclops royi. *G3 Genes|Genomes|Genetics*, 9(5):1295–1302, mar 2019. doi: 10.1534/g3.119.400085. URL https://doi.org/10.1534%2Fg3.119.400085.

M. Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11):1947–1951, sep 2019. doi: 10.1002/pro.3715. URL https://doi.org/10.1002%2Fpro.3715.

D. D. Kang, F. Li, E. Kirton, A. Thomas, R. Egan, H. An, and Z. Wang. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, 7:e7359, jul 2019. doi: 10.7717/peerj.7359. URL https://doi.org/10.7717%2Fpeerj.7359.

K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, 4 2013. ISSN 07374038. doi: 10.1093/molbev/mst010.

P. J. Keeling and d. J. Campo. Marine protists are not just big bacteria. *Current Biology*, 27(11):R541–R549, 6 2017. ISSN 09609822. doi: 10.1016/j.cub.2017.03.075.

P. J. Keeling, F. Burki, H. M. Wilcox, B. Allam, E. E. Allen, L. A. Amaral-Zettler, E. V. Armbrust, J. M. Archibald, A. K. Bharti, C. J. Bell, B. Beszteri, K. D. Bidle, C. T. Cameron, L. Campbell, D. A. Caron, R. A. Cattolico, J. L. Collier, K. Coyne, S. K. Davy, P. Deschamps, S. T. Dyhrman, B. Edvardsen, R. D. Gates, C. J. Gobler, S. J. Greenwood, S. M. Guida, J. L. Jacobi, K. S. Jakobsen, E. R. James, B. Jenkins, U. John, M. D. Johnson, A. R. Juhl, A. Kamp, L. A. Katz, R. Kiene, A. Kudryavtsev, B. S. Leander, S. Lin, C. Lovejoy, D. Lynn, A. Marchetti, G. McManus, A. M. Nedelcu, S. Menden-Deuer, C. Miceli, T. Mock, M. Montresor, M. A. Moran, S. Murray, G. Nadathur, S. Nagai, P. B. Ngam, B. Palenik, J. Pawlowski, G. Petroni, G. Piganeau, M. C. Posewitz,

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

K. Rengefors, G. Romano, M. E. Rumpho, T. Rynearson, K. B. Schilling, D. C. Schroeder, A. G. B. Simpson, C. H. Slamovits, D. R. Smith, G. J. Smith, S. R. Smith, H. M. Sosik, P. Stief, E. Theriot, S. N. Twary, P. E. Umale, D. Vaulot, B. Wawrik, G. L. Wheeler, W. H. Wilson, Y. Xu, A. Zingone, and A. Z. Worden. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biology*, 12(6):1–6, 06 2014. doi: 10.1371/journal.pbio.1001889. URL `https://doi.org/10.1371/journal.pbio.1001889`.

T. Klemetsen, I. A. Raknes, J. Fu, A. Agafonov, S. V. Balasundaram, G. Tartari, E. Robertsen, and N. P. Willassen. The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Research*, 46(D1):D692–D699, Nov. 2017.

A. E. Koid, Z. Liu, R. Terrado, A. C. Jones, D. A. Caron, and K. B. Heidelberg. Comparative Transcriptome Analysis of Four Prymnesiophyte Algae. 9(6):e97801, 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0097801. URL `https://dx.plos.org/10.1371/journal.pone.0097801`.

J. Koster and S. Rahmann. Snakemake–a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 10 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts480.

A. I. Krinos, S. K. Hu, N. R. Cohen, and H. Alexander. Eukulele: Taxonomic annotation of the unsung eukaryotic microbes. *Journal of Open Source Software*, 2021. doi: 10.21105/joss.02817.

E. V. Kriventseva, D. Kuznetsov, F. Tegenfeldt, M. Manni, R. Dias, F. A. SimÃ£o, and E. M. Zdobnov. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research*, 47(D1):D807–D811, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1053. URL `https://doi.org/10.1093/nar/gky1053`.

A. Labarre, A. Obiol, S. Wilken, I. Forn, and R. Massana. Expression of genes involved in phagocytosis in uncultured heterotrophic flagellates. *Limnol Oceanogr*, 65(S1), jan 2020. doi: 10.1002/lno.11379. URL `https://doi.org/10.1002%2Flno.11379`.

B. Lambert, R. Groussman, M. Schatz, S. Coesel, B. Durham, A. J. Alverson, A. E. White, and E. V. Armbrust. The dynamic trophic architecture of open-ocean protist communities revealed through machine-guided metatranscriptomics. *bioRxiv*, 2021.

M. D. Lee. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics*, 35(20):4162–4164, mar 2019. doi: 10.1093/bioinformatics/btz188. URL `https://doi.org/10.1093%2Fbioinformatics%2Fbtz188`.

I. Letunic and P. Bork. Interactive tree of life (itol) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic acids research*, 44(W1):W242–W245, 7 2016. ISSN 13624962. doi: 10.1093/nar/gkw290.

E. Levy Karin, M. Mirdita, and J. Söding. MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome*, 8(1):48, apr 2020. ISSN 20492618. doi: 10.1186/s40168-020-00808-x. URL `https://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-020-00808-x`.

D. Li, C.-M. Liu, R. Luo, K. Sadakane, and T.-W. Lam. Megahit: an ultra-fast single-node solution

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31 (10):1674–1676, 5 2015. ISSN 1460-2059. doi: 10.1093/bioinformatics/btv033.

D. Li, J. Fang, B. Wen, and X. Wu. Molecular identification of a novel intracellular proteobacteria from scallop chlamys farreri. *Aquaculture*, 539:736565, jun 2021. doi: 10.1016/j.aquaculture.2021.736565. URL https://doi.org/10.1016%2Fj.aquaculture.2021.736565.

H. Li and R. Durbin. Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics*, 26(5):589–595, 3 2010. ISSN 13674803. doi: 10.1093/bioinformatics/btp698.

Z. Liu, V. Campbell, K. B. Heidelberg, and D. A. Caron. Gene expression characterizes different nutritional strategies among three mixotrophic protists. *FEMS Microbiology Ecology*, 92(7): fiw106, may 2016. doi: 10.1093/femsec/fiw106. URL https://doi.org/10.1093%2Ffemsec%2Ffiw106.

A. Lomsadze, V. Ter-Hovhannisyan, Y. O. Chernoff, and M. Borodovsky. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Research*, 33(20):6494–6506, 2005. ISSN 03051048. doi: 10.1093/nar/gki937. URL /pmc/articles/PMC1298918/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC1298918/.

J. Lukeš, C. R. Stensvold, K. Jirků-Pomajbíková, and L. W. Parfrey. Are human intestinal eukaryotes beneficial or commensals? *PLoS Pathogens*, 11(8):e1005039, aug 2015. doi: 10.1371/journal.ppat.1005039. URL https://doi.org/10.1371%2Fjournal.ppat.1005039.

H. Luo, L. R. Thompson, U. Stingl, and A. L. Hughes. Selection maintains low genomic GC content in marine SAR11 lineages. *Mol Biol Evol*, 32(10):2738–2748, jun 2015. doi: 10.1093/molbev/msv149. URL https://doi.org/10.1093%2Fmolbev%2Fmsv149.

E. H. Mahood, L. H. Kruse, and G. D. Moghe. Machine learning: A powerful tool for gene function prediction in plants. *Applications in Plant Sciences*, 8(7):e11376, 2020.

G. Marçais, A. L. Delcher, A. M. Phillippy, R. Coston, S. L. Salzberg, and A. Zimin. MUMmer4: A fast and versatile genome alignment system. *PLoS Computational Biology*, 14(1):e1005944, jan 2018. doi: 10.1371/journal.pcbi.1005944. URL https://doi.org/10.1371%2Fjournal.pcbi.1005944.

R. Massana. Eukaryotic Picoplankton in Surface Oceans. 65(1):91–110, 2011. ISSN 0066-4227, 1545-3251. doi: 10.1146/annurev-micro-090110-102903. URL http://www.annualreviews.org/doi/10.1146/annurev-micro-090110-102903.

M. Mirdita, M. Steinegger, and J. Söding. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, 35(16):2856–2858, aug 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty1057. URL https://academic.oup.com/bioinformatics/article/35/16/2856/5280135.

A. Mitra, K. J. Flynn, J. M. Burkholder, T. Berge, A. Calbet, J. A. Raven, E. Granéli, P. M. Glibert, P. J. Hansen, D. K. Stoecker, et al. The role of mixotrophic protists in the biological carbon pump. *Biogeosciences*, 11(4):995–1005, 2014.

T. K. Mohanta and H. Bae. The diversity of fungal genome. *Biol Proced Online*, 17(1), apr 2015. doi: 10.1186/s12575-015-0020-z. URL https://doi.org/10.1186%2Fs12575-015-0020-z.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

C. M. Moore, M. M. Mills, K. R. Arrigo, I. Berman-Frank, L. Bopp, P. W. Boyd, E. D. Galbraith, R. J. Geider, C. Guieu, S. L. Jaccard, T. D. Jickells, J. L. Roche, T. M. Lenton, N. M. Mahowald, E. Marañón, I. Marinov, J. K. Moore, T. Nakatsuka, A. Oschlies, M. A. Saito, T. F. Thingstad, A. Tsuda, and O. Ulloa. Processes and patterns of oceanic nutrient limitation. *Nature Geosci*, 6(9): 701–710, mar 2013. doi: 10.1038/ngeo1765. URL `https://doi.org/10.1038%2Fngeo1765`.

S. E. Morales, A. Biswas, G. J. Herndl, and F. Baltar. Global structuring of phylogenetic and functional diversity of pelagic fungi by depth and temperature. *Front. Mar. Sci.*, 6, mar 2019. doi: 10.3389/fmars.2019.00131. URL `https://doi.org/10.3389%2Ffmars.2019.00131`.

C. J. Neely, S. K. Hu, H. Alexander, and B. J. Tully. The high-throughput gene prediction of more than 1,700 eukaryote genomes using the software package eukmetasanity. *bioRxiv*, 2021. doi: 10.1101/.

A. Obiol, C. R. Giner, P. Sánchez, C. M. Duarte, S. G. Acinas, and R. Massana. A metagenomic assessment of microbial eukaryotic diversity in the global ocean. 20(3):1755–0998.13147, 2020. ISSN 1755-098X, 1755-0998. doi: 10.1111/1755-0998.13147. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13147`.

J. G. Okie, A. T. Poret-Peterson, Z. M. Lee, A. Richter, L. D. Alcaraz, L. E. Eguiarte, J. L. Siefert, V. Souza, C. L. Dupont, and J. J. Elser. Genomic adaptations in information processing underpin trophic strategy in a whole-ecosystem nutrient enrichment experiment. *eLife*, 9, jan 2020. doi: 10.7554/elife.49816. URL `https://doi.org/10.7554%2Felife.49816`.

M. R. Olm, C. T. Brown, B. Brooks, and J. F. Banfield. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME Journal*, 11(12):2864–2868, jul 2017. doi: 10.1038/ismej.2017.126. URL `https://doi.org/10.1038%2Fismej.2017.126`.

M. R. Olm, P. T. West, B. Brooks, B. A. Firek, R. Baker, M. J. Morowitz, and J. F. Banfield. Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome*, 7(1):26, 12 2019. ISSN 2049-2618. doi: 10.1186/s40168-019-0638-1.

B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, and A. M. Phillippy. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol*, 17(1), jun 2016. doi: 10.1186/s13059-016-0997-x. URL `https://doi.org/10.1186%2Fs13059-016-0997-x`.

M. G. Pachiadaki, J. M. Brown, J. Brown, O. Bezuidt, P. M. Berube, S. J. Biller, N. J. Poulton, M. D. Burkart, J. J. L. Clair, S. W. Chisholm, and R. Stepanauskas. Charting the complexity of the marine microbiome through single-cell genomics. *Cell*, 179(7):1623–1635.e11, dec 2019. doi: 10.1016/j.cell.2019.11.017. URL `https://doi.org/10.1016%2Fj.cell.2019.11.017`.

D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, and G. W. Tyson. Checkm: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055, 7 2015. ISSN 15495469. doi: 10.1101/gr.186072.114.

D. H. Parks, C. Rinke, M. Chuvochina, P.-A. Chaumeil, B. J. Woodcroft, P. N. Evans, P. Hugenholtz, and G. W. Tyson. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the tree of life. *Nature Microbiology*, 2(11):1533–1542, 11 2017. ISSN 2058-5276. doi: 10.1038/s41564-017-0012-7.

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

M. C. Pernice, I. Forn, A. Gomes, E. Lara, L. Alonso-Sáez, J. M. Arrieta, F. del Carmen Garcia, V. Hernando-Morales, R. MacKenzie, M. Mestre, E. Sintes, E. Teira, J. Valencia, M. M. Varela, D. Vaqué, C. M. Duarte, J. M. Gasol, and R. Massana. Global abundance of planktonic heterotrophic protists in the deep ocean. *ISME Journal*, 9(3):782–792, oct 2014. doi: 10.1038/ismej.2014.168. URL `https://doi.org/10.1038%2Fismej.2014.168`.

M. C. Pernice, C. R. Giner, R. Logares, J. Perera-Bel, S. G. Acinas, C. M. Duarte, J. M. Gasol, and R. Massana. Large variability of bathypelagic microbial eukaryotic communities across the world's oceans. *ISME Journal*, 10(4):945–958, oct 2015. doi: 10.1038/ismej.2015.170. URL `https://doi.org/10.1038%2Fismej.2015.170`.

M. Pertea and G. Pertea. GFF Utilities: GffRead and GffCompare. *F1000Research*, 9:304, sep 2020. ISSN 1759796X. doi: 10.12688/f1000research.23297.1. URL `https://doi.org/10.12688/f1000research.23297.1`.

S. Pesant, , F. Not, M. Picheral, S. Kandels-Lewis, N. L. Bescot, G. Gorsky, D. Iudicone, E. Karsenti, S. Speich, R. Troublé, C. Dimier, and S. Searson. Open science resources for the discovery and analysis of tara oceans data. *Sci Data*, 2(1), may 2015. doi: 10.1038/sdata.2015.23. URL `https://doi.org/10.1038%2Fsdata.2015.23`.

J. J. Pierella Karlusich, F. M. Ibarbalz, and C. Bowler. Phytoplankton in the *Tara* Ocean. 12(1):233–265, 2020. ISSN 1941-1405, 1941-0611. doi: 10.1146/annurev-marine-010419-010706. URL `https://www.annualreviews.org/doi/10.1146/annurev-marine-010419-010706`.

M. N. Price, P. S. Dehal, and A. P. Arkin. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE*, 5(3):e9490, mar 2010. doi: 10.1371/journal.pone.0009490. URL `https://doi.org/10.1371%2Fjournal.pone.0009490`.

P. Probst, M. Wright, and A.-L. Boulesteix. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2018. doi: 10.1002/widm.1301.

B. A. Read, J. Kegel, M. J. Klute, A. Kuo, S. C. Lefebvre, F. Maumus, C. Mayer, J. Miller, A. Monier, A. Salamov, J. Young, M. Aguilar, J. M. Claverie, S. Frickenhaus, K. Gonzalez, E. K. Herman, Y. C. Lin, J. Napier, H. Ogata, A. F. Sarno, J. Shmutz, D. Schroeder, C. De Vargas, F. Verret, P. Von Dassow, K. Valentin, Y. Van De Peer, G. Wheeler, J. B. Dacks, C. F. Delwiche, S. T. Dyhrman, G. Glöckner, U. John, T. Richards, A. Z. Worden, X. Zhang, I. V. Grigoriev, A. E. Allen, K. Bidle, M. Borodovsky, C. Bowler, C. Brownlee, J. Mark Cock, M. Elias, V. N. Gladyshev, M. Groth, C. Guda, A. Hadaegh, M. D. Iglesias-Rodriguez, J. Jenkins, B. M. Jones, T. Lawson, F. Leese, E. Lindquist, A. Lobanov, A. Lomsadze, S. B. Malik, M. E. Marsh, L. MacKinder, T. Mock, B. Mueller-Roeber, A. Pagarete, M. Parker, I. Probert, H. Quesneville, C. Raines, S. A. Rensing, D. M. Riaño-Pachón, S. Richier, S. Rokitta, Y. Shiraiwa, D. M. Soanes, M. Van Der Giezen, T. M. Wahlund, B. Williams, W. Wilson, G. Wolfe, and L. L. Wurch. Pan genome of the phytoplankton

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Emiliania underpins its global distribution. *Nature*, 499(7457):209–213, jun 2013. ISSN 00280836. doi: 10.1038/nature12221. URL http://www.ncbi.nlm.nih.gov/pubmed/23760476.

D. J. Richter, C. Berney, J. F. H. Strassert, F. Burki, and d. C. Vargas. Eukprot: a database of genome-scale predicted proteins across the diversity of eukaryotic life. *bioRxiv*, page 2020.06.30.180687, 7 2020. doi: 10.1101/2020.06.30.180687.

C. Rinke, F. Rubino, L. F. Messer, N. Youssef, D. H. Parks, M. Chuvochina, M. Brown, T. Jeffries, G. W. Tyson, J. R. Seymour, and P. Hugenholtz. A phylogenomic and ecological analysis of the globally abundant marine group ii archaea (ca. poseidoniales ord. nov.). *ISME Journal*, 13(3): 663–675, 3 2019. ISSN 17517370. doi: 10.1038/s41396-018-0282-y.

E. P. Rocha. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genome Research*, 14(11):2279–2286, nov 2004. doi: 10.1101/gr.2896904. URL https://doi.org/10.1101%2Fgr.2896904.

P. Saary, A. L. Mitchell, and R. D. Finn. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with eukcc. *Genome Biology*, 21(1), 9 2020. ISSN 1474760X. doi: 10.1186/s13059-020-02155-4.

D. Santos-Garcia, P.-A. Rollat-Farnier, F. Beitia, E. Zchori-Fein, F. Vavre, L. Mouton, A. Moya, A. Latorre, and F. J. Silva. The genome of cardinium cBtQ1 provides insights into genome reduction, symbiont motility, and its settlement in bemisia tabaci. *Genome Biology and Evolution*, 6 (4):1013–1030, apr 2014. doi: 10.1093/gbe/evu077. URL https://doi.org/10.1093%2Fgbe%2Fevu077.

E. B. Sherr and B. F. Sherr. *Antonie van Leeuwenhoek*, 81(1):293–308, 2002. doi: 10.1023/a:1020591307260. URL https://doi.org/10.1023%2Fa%3A1020591307260.

M. A. Shipp, K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine*, 8(1):68–74, 2002.

Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, jun 1967. doi: 10.1080/01621459.1967.10482935. URL https://doi.org/10.1080%2F01621459.1967.10482935.

M. E. Sieracki, N. J. Poulton, O. Jaillon, P. Wincker, C. de Vargas, L. Rubinat-Ripoll, R. Stepanauskas, R. Logares, and R. Massana. Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. 9(1):6025, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-42487-1. URL http://www.nature.com/articles/s41598-019-42487-1.

F. A. Simao, R. M. Waterhouse, P. Ioannidis, V. E. Kriventseva, and E. M. Zdobnov. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31 (19):3210–3212, 10 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv351.

A. Smit and R. Hubley. Repearmodeler open-1.0. http://www.repeatmasker.org, 2008-2015.

A. Smit, R. Hubley, and P. Green. Repeatmasker open-4.0. http://www.repeatmasker.org, 2013-2015.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

A. Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 5 2014. ISSN 1460-2059. doi: 10.1093/bioinformatics/btu033.

M. Steinegger and J. Söding. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets, nov 2017. ISSN 15461696. URL `https://www.nature.com/articles/nbt.3988`.

M. Steinegger and J. Söding. Clustering huge protein sequence sets in linear time. *Nature Communications*, 9(1):1–8, dec 2018. ISSN 20411723. doi: 10.1038/s41467-018-04964-5. URL `www.nature.com/naturecommunications`.

R. D. Stewart, M. D. Auffret, T. J. Snelling, R. Roehe, and M. Watson. Magpy: a reproducible pipeline for the downstream analysis of metagenome-assembled genomes (mags). *Bioinformatics*, 35(12):2150–2152, 6 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty905.

D. K. Stoecker, P. J. Hansen, D. A. Caron, and A. Mitra. Mixotrophy in the marine plankton. *Annu. Rev. Mar. Sci.*, 9(1):311–335, jan 2017. doi: 10.1146/annurev-marine-010816-060617. URL `https://doi.org/10.1146%2Fannurev-marine-010816-060617`.

S. L. Strom. Microbial ecology of ocean biogeochemistry: a community perspective. *Science*, 320 (5879):1043–1045, 2008.

B. K. Swan, B. Tupper, A. Sczyrba, F. M. Lauro, M. Martinez-Garcia, J. M. Gonzalez, H. Luo, J. J. Wright, Z. C. Landry, N. W. Hanson, B. P. Thompson, N. J. Poulton, P. Schwientek, S. G. Acinas, S. J. Giovannoni, M. A. Moran, S. J. Hallam, R. Cavicchioli, T. Woyke, and R. Stepanauskas. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences*, 110(28):11463–11468, jun 2013. doi: 10.1073/pnas.1304246110. URL `https://doi.org/10.1073%2Fpnas.1304246110`.

A. A. Tabl, A. Alkhateeb, W. ElMaraghy, L. Rueda, and A. Ngom. A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. *Frontiers in Genetics*, 10:256, 2019.

C. Tara Oceans Consortium and P. Tara Oceans Expedition. Environmental context of all samples from the Tara Oceans Expedition (2009-2013), about water column features. PANGAEA, 2016. doi: 10.1594/PANGAEA.858207. URL `https://doi.org/10.1594/PANGAEA.858207`. In: Tara Oceans Consortium, C; Tara Oceans Expedition, P (2016): Registry of all samples from the Tara Oceans Expedition (2009-2013). PANGAEA, https://doi.org/10.1594/PANGAEA.859953.

R. C. Team. R version 3.6.2: A language and environmental for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria: USBN*, pages 3–900051.

B. J. Tully. Metabolic diversity within the globally abundant marine group ii euryarchaea offers insight into ecological patterns. *Nature Communications*, 10(1):1–12, 12 2019. ISSN 20411723. doi: 10.1038/s41467-018-07840-4.

B. J. Tully, E. D. Graham, and J. F. Heidelberg. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, 5:170203, 1 2018. ISSN 2052-4463. doi: 10.1038/sdata.2017.203.

bioRxiv preprint doi: https://doi.org/10.1101/2021.07.25.453713; this version posted July 25, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

F. Unrein, J. M. Gasol, F. Not, I. Forn, and R. Massana. Mixotrophic haptophytes are key bacterial grazers in oligotrophic coastal waters. 8(1):164–176, 2014. ISSN 1751-7362, 1751-7370. doi: 10.1038/ismej.2013.132. URL http://www.nature.com/articles/ismej2013132.

L. J. Ustick, A. A. Larkin, C. A. Garcia, N. S. Garcia, M. L. Brock, J. A. Lee, N. A. Wiseman, J. K. Moore, and A. C. Martiny. Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. *Science*, 372(6539):287–291, apr 2021. doi: 10.1126/science.abe6301. URL https://doi.org/10.1126%2Fscience.abe6301.

D. Vaulot, W. Eikrem, M. Viprey, and H. Moreau. The diversity of small eukaryotic phytoplankton ($\leq$3 *M*m) in marine ecosystems. 32(5):795–820, 2008. ISSN 1574-6976. doi: 10.1111/j.1574-6976.2008.00121.x. URL https://academic.oup.com/femsre/article-lookup/doi/10.1111/j.1574-6976.2008.00121.x.

A. Vorobev, M. Dupouy, Q. Carradec, T. O. Delmont, A. Annamalé, P. Wincker, and E. Pelletier. Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. 30(4):647–659. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.253070.119. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.253070.119.

G. P. Wagner, K. Kin, and V. J. Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, 131(4):281–285, aug 2012. doi: 10.1007/s12064-012-0162-3. URL https://doi.org/10.1007%2Fs12064-012-0162-3.

P. T. West, A. J. Probst, V. I. Grigoriev, B. C. Thomas, and J. F. Banfield. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome research*, 28(4):569–580, 4 2018. ISSN 1549-5469. doi: 10.1101/gr.228429.117.

A. Z. Worden, M. J. Follows, S. J. Giovannoni, S. Wilken, A. E. Zimmerman, and P. J. Keeling. Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science*, 347(6223):1257594, 2015. ISSN 0036-8075. doi: 10.1126/science.1257594. URL http://www.sciencemag.org/content/347/6223/1257594.short.

M. N. Wright and A. Ziegler. ranger: A fast implementation of random forests for high dimensional data in c++ and r. *J. Stat. Soft.*, 77(1), 2017. doi: 10.18637/jss.v077.i01. URL https://doi.org/10.18637%2Fjss.v077.i01.

Y.-W. Wu, Y.-H. Tang, S. G. Tringe, B. A. Simmons, and S. W. Singer. Maxbin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, 2(1):26, 12 2014. ISSN 2049-2618. doi: 10.1186/2049-2618-2-26.

W. Zhang, J. Chen, Y. Yang, Y. Tang, J. Shang, and B. Shen. A practical comparison of de novo genome assembly software tools for next-generation sequencing technologies. *PLoS ONE*, 6(3), 2011. ISSN 19326203. doi: 10.1371/journal.pone.0017915.