Supplementary Information for "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[4], and Taylor Reiter[5]

[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]Population Health and Reproduction, University of California, Davis, Davis, CA, 95616
[*]Correspondence; halexander@whoi.edu

1 Development and performance of the trophic mode model

1.1 Description of the model and Heterotrophy Index

We used a variable selection algorithm and Random Forest machine learning model framework in order to predict the likely trophic mode of the eukaryotic TOPAZ MAGs described in this study. Transcriptomes from the MMETSP and EukProt were manually-annotated as phototroph, mixotroph, or heterotroph based on the literature (Supplementary Table 5). We tested our model with a randomly subset test set comprised of the 25% of MMETSP and EukProt transcriptomes (Keeling et al., 2014; Richter et al., 2020) that were excluded from the model building procedure. With this test subset we obtained an accuracy of 94.6% (Figure S23), meaning that nearly 95% of taxonomic annotations derived from the machine learning model aligned with their manually-assigned trophic mode annotation (Figure S23). When applied to the TOPAZ MAGs, all MAGs were either classified as phototrophs or heterotrophs, with none classified as mixotrophs. This likely reflects that the model was generally conservative when it came to assigning genomes or transcriptomes as mixotrophs (Figure S27). As a consequence, we developed a secondary metric for assessing the extent of heterotrophy in the test

genomes and transcriptomes using the KOs selected by the vita selection process ($n = 1787$), but instead of using the presence or absence of these KOs as a binary indicator to inform the classification of the MAGs, we used the presence or absence as part of an equation to more sensitively assess the number of KOs present that tend to be indicative or either heterotrophy or phototrophy. The result was the "Heterotrophy Index" (H-index), a metric for assessing trophy based on KEGG pathway presence or absence.

The H-index is a sliding scale that weights the presence of heterotrophy, phototrophy- and mixotrophy-indicative KOs to assess the overall likely trophic state of an organism, and will consequently better show when an organism is more likely mixotrophic or possessing traits from both heterotrophy and phototrophy. Because mixotrophs were less common in our test dataset, there are natural concerns about the skill of the model when it comes to identifying them (Vabalas et al., 2019).We did not attempt to address the imbalance of the categorical training data in the model as other papers have recently explored (Utkin, 2020; Collins et al., 2020), hence our model retains the bias of reduced sensitivity when the distribution of the training data categories cannot be compared directly to the "true" incidence of mixotrophy among eukaryotic organisms (Khalilia et al., 2011). Because the majority of test and training transcriptomes were phototrophs or heterotrophs, it is more conservative for a Random Forest model to assign these modes more frequently. As the TOPAZ MAGs covered lineages with known mixotrophic members (Jones, 2000), and with comparison and feedback from an alternative trophic model as described in Section 1.2, we applied the H-index to provide more sensitivity in the identification of likely mixotrophs. In particular, with the Random Forest design, if a MAG has characteristics of both heterotrophs and phototrophs that are present in the training data, but does not align with the limited sample of mixotroph transcriptomes (which is also problematic due to the opportunistic nature of the sampling, see Section 1.3), these MAGs would be assigned the best guess between phototrophy and heterotrophy, when in reality this combination of traits may indicate some form of mixotrophy. The H-index also serves as a confidence metric for the trophy estimate. For example, a MAG with a large positive H-index would more confidently be called a heterotroph, as this would indicate a strong frequency and degree of alignment with heterotroph references (and, specifically, alignment with those heterotrophic references for KOs identified by the vita algorithm as important for distinguishing trophic mode within the training set). By contrast, a small (or near zero) positive H-index may be mixotrophic or represent a less complete MAG.

1.2 Comparison to Burns and Lambert models

In order to assess the performance of our model, which relies solely on assessment of KOs (Kanehisa, 2019) that were determined computationally to be important and assessed for function after the fact (Figure S25), we applied the model from Burns et al. (2018) (heretofore referred to as the Burns model) to the same highly complete eukaryotic MAGs, as well as to the MMETSP transcriptomes. This model assigns a score from zero to one for individual characteristics related to trophy, including photosynthetic ability, phagocytosis, and prototrophy (Burns et al., 2018). Using Hidden Markov Models for selected genes which have known association with the aforementioned trophic strategies, the Burns model instead looks for a set of genes consistent with each trait, to assess the "completeness" of the genome or transcriptome with respect to the machinery known to be involved with each function. We found the Burns model results to be consistent with our H-index procedure in the following ways. Among the MMETSP transcriptomes, 93.1% (n=81) of the transcriptomes with

a positive H-index (indicative of net heterotrophy) also had a photosynthesis score of less than 0.5 as assigned by the Burns model, and 90.8% (n=79) had a photosynthesis score of less than 0.05 via the Burns model (Figure S26). Similarly, 87.3% (n=226) of MMETSP transcriptomes with a negative H-index (indicative of net phototrophy) had a photosynthesis score of greater than 0.5 as assigned by the Burns model (52.9% (n=137) had negative H-index and Burns photosynthesis score greater than 0.95). MMETSP transcriptomes with a zero or near-zero H-index, which corresponds to putative mixotrophy, had varying photosynthesis scores according to the Burns model, but were more likely to have high (0.6-1) photosynthesis scores, which is consistent with mixotrophy (Figure S26). However, several of the MAGs which were predicted to be mixotrophs by the Random Forest model and were annotated manually as mixotrophs from the available metadata had mid- to high- photosynthesis scores in the Burns model, yet negative-leaning H-index scores (Supplementary Table 10). These transcriptomes also tended to have high (>0.7) phagocytosis predictions per the Burns model, consistent with the presence of genetic resources for both heterotrophic and phototrophic strategies. Broadly, we found that, similarly to our H-index and Random Forest model annotations, the Burns photosynthesis prediction results tended to align with the expected lifestyle of each MAG based on EUKulele (Krinos et al., 2021) taxonomic annotations, with expected heterotrophs like Amoebozoa, Fungi, and Opisthokonta scoring low on photosynthetic ability, while expected phototrophs like Ochrophyta and Chlorophyta tended to score highly for photosynthetic ability (Figure S26). Most disagreement was found within the SAR clade and Crypophyta, wherein a range of photosynthesis scores were found by the Burns model, and sometimes these scores contradicted the annotation found by the H-index and Random Forest model (fig. S26 and Supplementary Tables 9 and 10). This would indicate potentially cryptic and variable trophic strategies and lifestyles within these annotated groups.

When split into classes of photosynthetic ability based on Burns model scores (to isolate "highly" or "not at all" photosynthetic: 0-0.1, 0.1-0.45, 0.45-0.55, 0.55-0.9, 0.9-1), MMETSP transcriptomes with "no" photosynthesis according to the Burns model (photosynthesis prediction < 0.1) had an average H-index of 38.85 ± 68.19, while transcriptomes with "high" photosynthesis (photosynthesis prediction > 0.9) had an average H-index of -272.34 ± 119.94 (Figure S24). As far as the TOPAZ MAGs, we similarly found that the MAGs predicted to be heterotrophic had low variance in photosynthesis prediction as reported by the Burns model, yet the variability in the photosynthetic prediction was high, in particular among those MAGs of higher (less negative, hence closer to "mixotrophic") H-index (Figure S27).

1.3 The future of trophic mode models

The model we developed relies solely upon references that were derived from expression-level data (transcriptomics). Additionally, we used the entire MMETSP (Keeling et al., 2014) and EukProt (Richter et al., 2020) databases with manually assigned trophic strategies based on the literature (Supplementary Table 5). Both of these choices carry issues that may be responsible for our under prediction of mixotrophy. First, as they were transcriptomic datasets, the experimental conditions that were used to generate the reference transcriptome are important, in that if a culture was maintained in phototrophy-favorable conditions (e.g. high light, sufficient inorganic nutrients) as opposed to heterotrophy-favorable conditions (e.g. low light, external carbon source), the transcripts reconstructed from these experiments may result in a reference that is skewed towards phototrophy or heterotrophy, respectively. Regardless of the conditions in which the organism was grown, it is likely

that genes present in the genome of the organism were not recovered by transcriptomics. While this means that important genes related to trophic strategy may be missed from the prediction workflow, this is also exciting, as there is much room for growth for these already accurate and skillful models. As genomic references become available and the number of transcriptomic experiments continue to grow, we can expect to further constrain the classes of genes that decide trophic mode. As databases grow, they may be pruned to include only experiments in which the trophy of the organism was known exactly, which would enable the patterns of expression which decide trophic mode to be pinpointed more precisely. Ideally, a complete core set of protein families necessary for heterotrophy and for phototrophs could be identified. A fundamental question that remains, however, is if there are any genes or protein families that are characteristic of mixotrophic organisms *only* or if these organisms are better characterized based on the co-occurrence of genes indicative of phototrophy and heterotrophy. If the latter, mixotrophy would be best characterized by the proportion of these sets which overlap in the genome or the expression profile of the candidate organism. While this is in principle the aim of the Burns model (Burns et al. (2018), Section 1.2), this approach might still be augmented by machine learning-based techniques that have the capacity to identify important genes that may not have yet been annotated.

2 Taxonomic group correlations with environmental parameters

The relative abundance of eukaryotic MAGs across samples was correlated with environmental parameters (Figure S12) and were found to broadly cluster by course taxonomic grouping, as such abundances were summed by broad taxonomic group to assess (Figure S13) general trends and correlations in patterns of abundance. Globally, we note that metazoan MAG abundance significantly (Bonferroni corrected $p < 0.001$) positively correlated with temperature, salinity, and retention time (defined as the average length of time that a particle has been trapped in an eddy), and significantly negatively correlated with a variety of nutrients (phosphate, nitrate+nitrite, silica), CDOM, and depth (meaning they are more abundant in surface samples). This suggests that the metazoan MAGs we are recovering are likely abundant in the subtropical and tropical waters of the large ocean gyres that are oligotrophic. Notably, Choanozoa and Amoebozoa clustered with Metazoans, forming a distinct cluster from other more typically photosynthetic and planktonic groups. Chlorophyta, Apusozoa, Cryptophyta, Dinoflagellata, Ochrophyta, SAR, and Haptophyta clustered separately from Metazoa and were typified by significant positive correlation with oxygen, *a*, and nitrite (Figure S13). Additionally, MAGs recovered that belonged to Ochrophyta were found to be significantly negatively correlated with temperature. This suggests that these MAGs were likely associated with more polar or subpolar regions.

Supplementary Tables

All Supplementary tables are available through the open science framework at `https://osf.io/twz2f/`.

- **Supplementary Table 1.** Assembly group description, sample inclusion, and basic assembly statistics.

- **Supplementary Table 2.** TOPAZ Eukaryotic MAG taxonomy, genomic characteristics (e.g., total length, GC content, N50, number of predicted proteins,), and estimated completeness and contamination.
- **Supplementary Table 3.** TOPAZ Prokaryotic MAG taxonomy as estimated by GTDB, dataset indication (non-redundant representatives (NR), total size, and estimated completeness and contamination.
- **Supplementary Table 4.** TOPAZ Prokaryotic MAG summary of recovered MAGs across phyla. Total counts are shown for all MAGs (All), the non-redundant subset of MAGs (AllNR), and the high-quality, non-redundant MAGs (HQNR).
- **Supplementary Table 5.** Reference transcriptomes used in the construction and testing of the trophic model. The manually annotated trophic status and details, relevant reference literature, and taxonomic information are noted. Additionally, a presence/absence matrix is provided for all eukaryotic KEGG ids considered here.
- **Supplementary Table 6.** Vita selected KOs and their associated heterotrophy, phototrophy, and mixotrophy ratios as described in equations 1-4.
- **Supplementary Table 7.** KEGG presence and absence for the TOPAZ eukaryotic MAGs, as was used in the trophic model.
- **Supplementary Table 8.** Pfam presence and absence across the TOPAZ eukaryotic MAGs.
- **Supplementary Table 9.** Eukaryotic TOPAZ MAG predicted trophic status and heterotrophy index (H-index).
- **Supplementary Table 10.** Burns model (Burns et al., 2018) predicted prototrophy, photosynthetic ability, and phagocytosis for the TOPAZ eukaryotic MAGs.
- **Supplementary Table 11.** Network analysis community composition.
- **Supplementary Table 12.** Eukaryotic cluster groups derived from average nucleotide identity clustering of eukaryotic TOPAZ MAGs with the Delmont eukaryotic MAGs (Delmont et al., 2020) based on an ANI cutoff of 99%.
- **Supplementary Table 13.** Environmental correlations with network derived communities.

Supplementary Figures

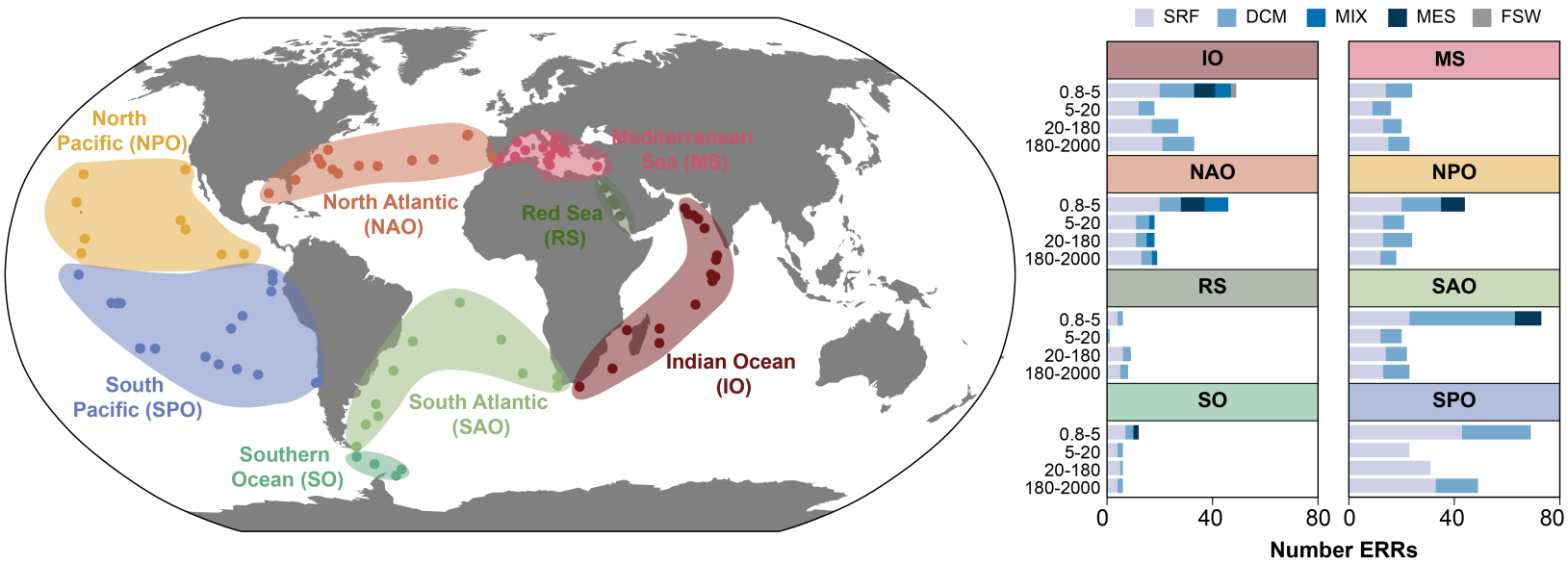

Figure S1: Sample map and distribution of sequence data sets across regions. Stations from *Tara* Oceans were categorized into general regions which are highlighted by color. The depth (surface (SRF), deep chlorophyll max (DCM), mixed surface sample (MIX), mesopelagic (MES), and filtered seawater (FSW)) and size fraction are shown as bar plots for each ocean region.

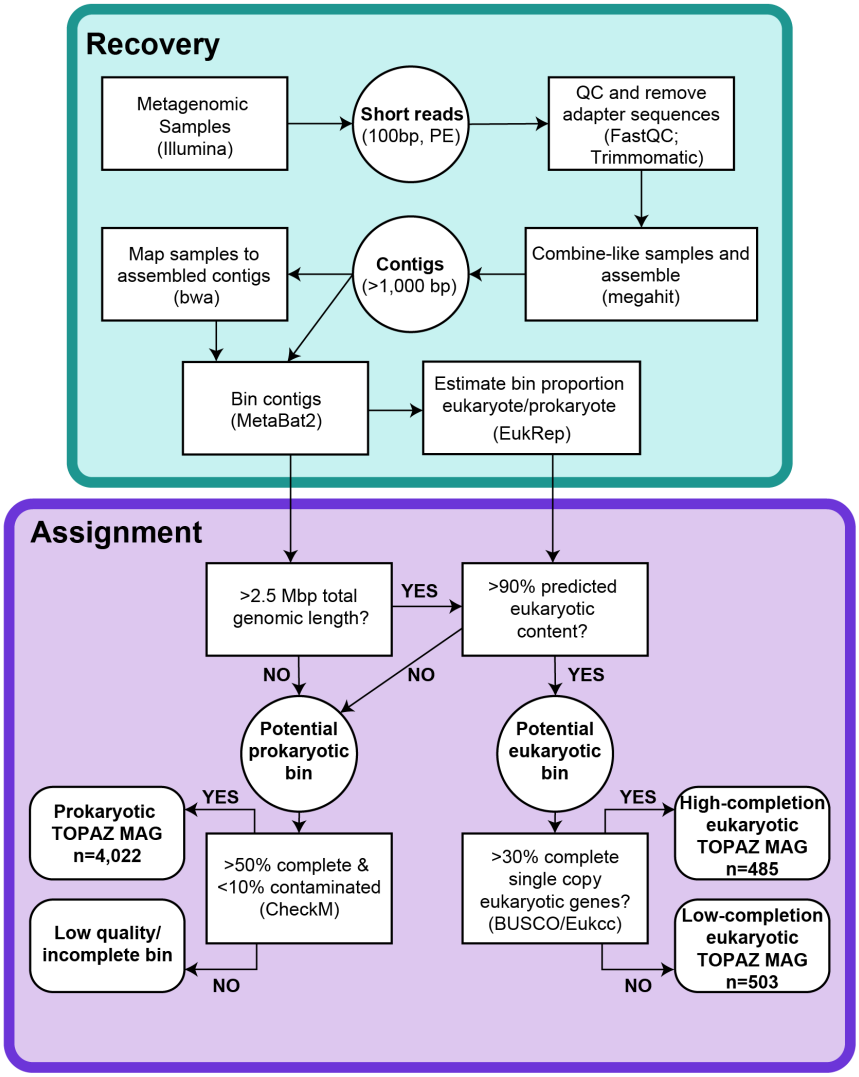

Figure S2: Flow chart of the EukHeist pipeline.

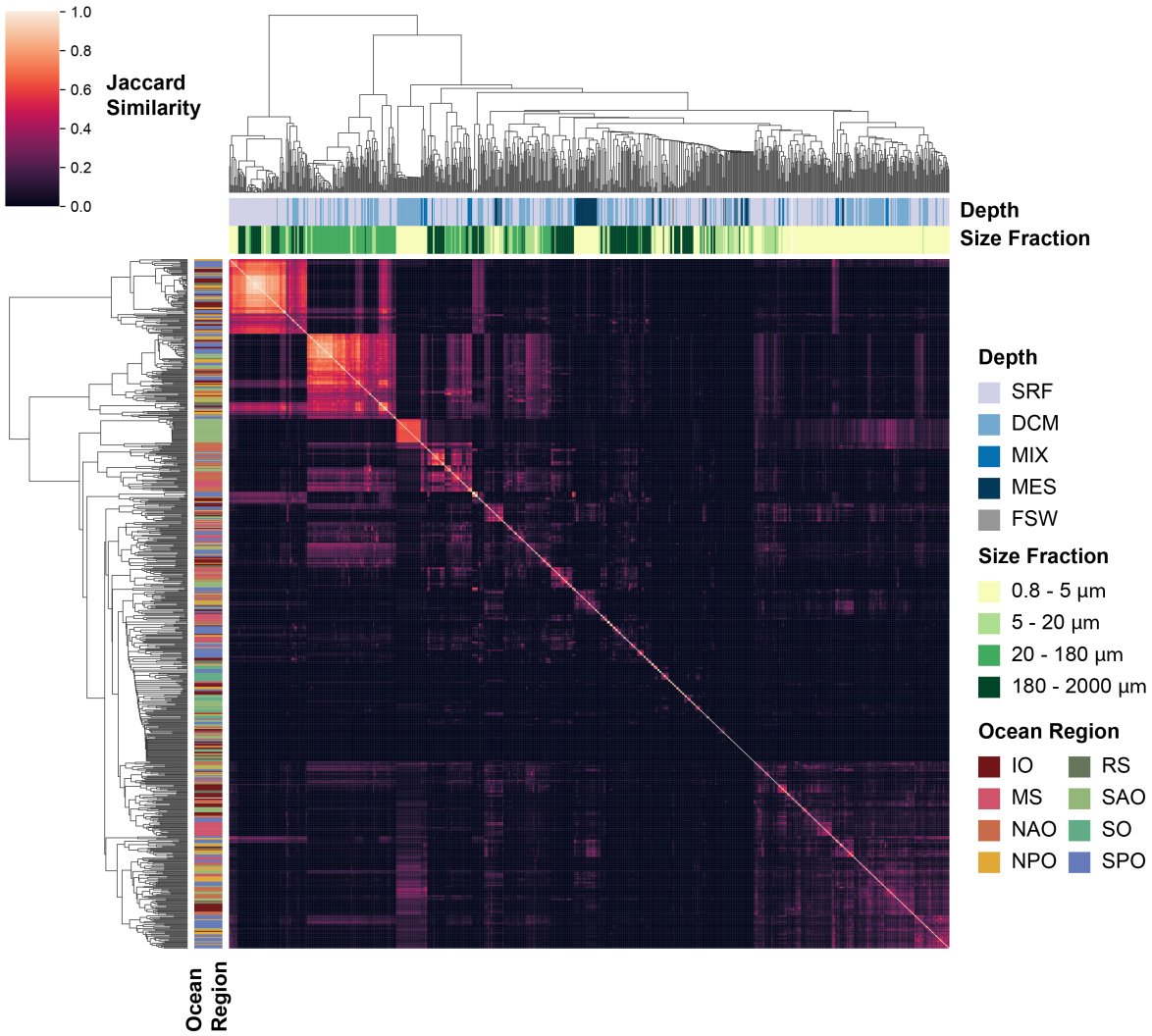

Figure S3: Sourmash comparison of all metagenomic samples from *Tara* Oceans large size fraction dataset. A minhash comparison was calculated using sourmash (k=31, scale=10,000) of the 824 metagenomic samples corresponding to the large size fraction metagenomic data from *Tara* Oceans (PRJEB4352) (Brown and Irber, 2016). The relative sequence content similarity is shown as Jaccard similarity. Hierarchical clustering of samples based on sequence content is shown and sample identity (sample depth, size fraction, and ocean region) is highlighted by colored blocks.

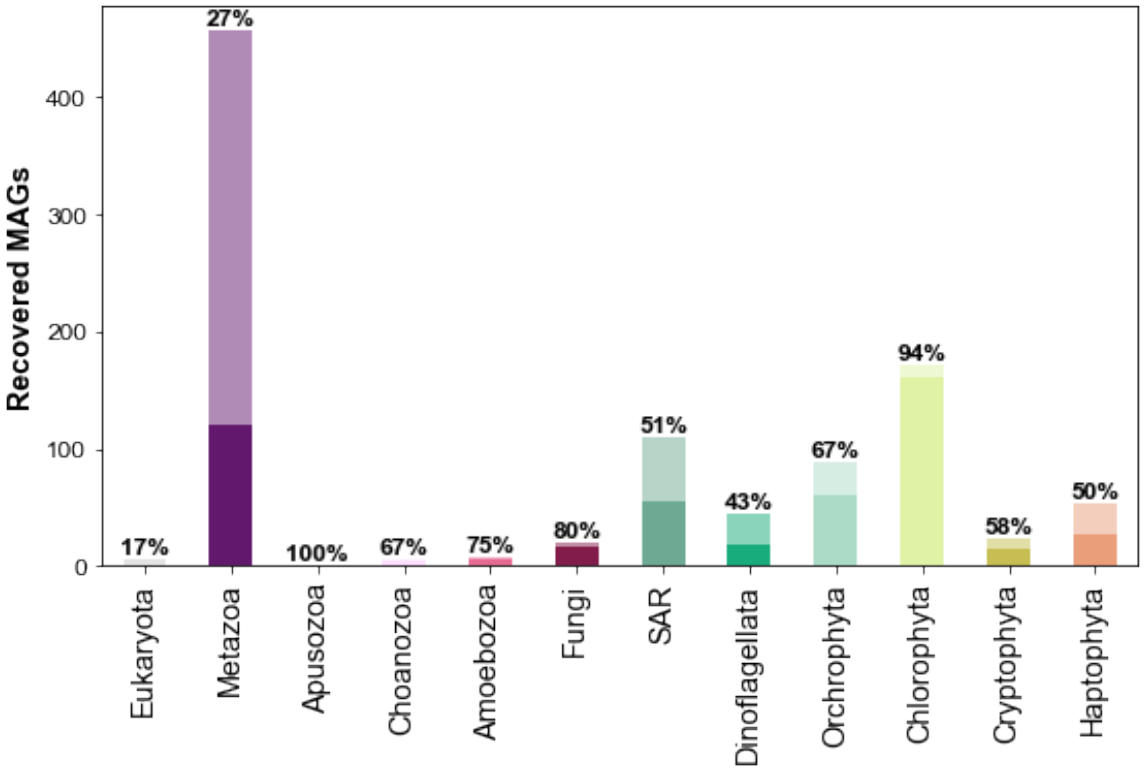

Figure S4: TOPAZ Recovered Eukaryotic MAGs. Course level taxonomic categorization of recovered eukaryotic TOPAZ MAGs (n=988). For each taxonomic group, the total number of MAGs is depicted. MAGs within a taxonomic group that were highly complete ($>$ 30% BUSCO completeness) are shaded and the percentage of highly complete MAGs for each taxonomic group is reported.

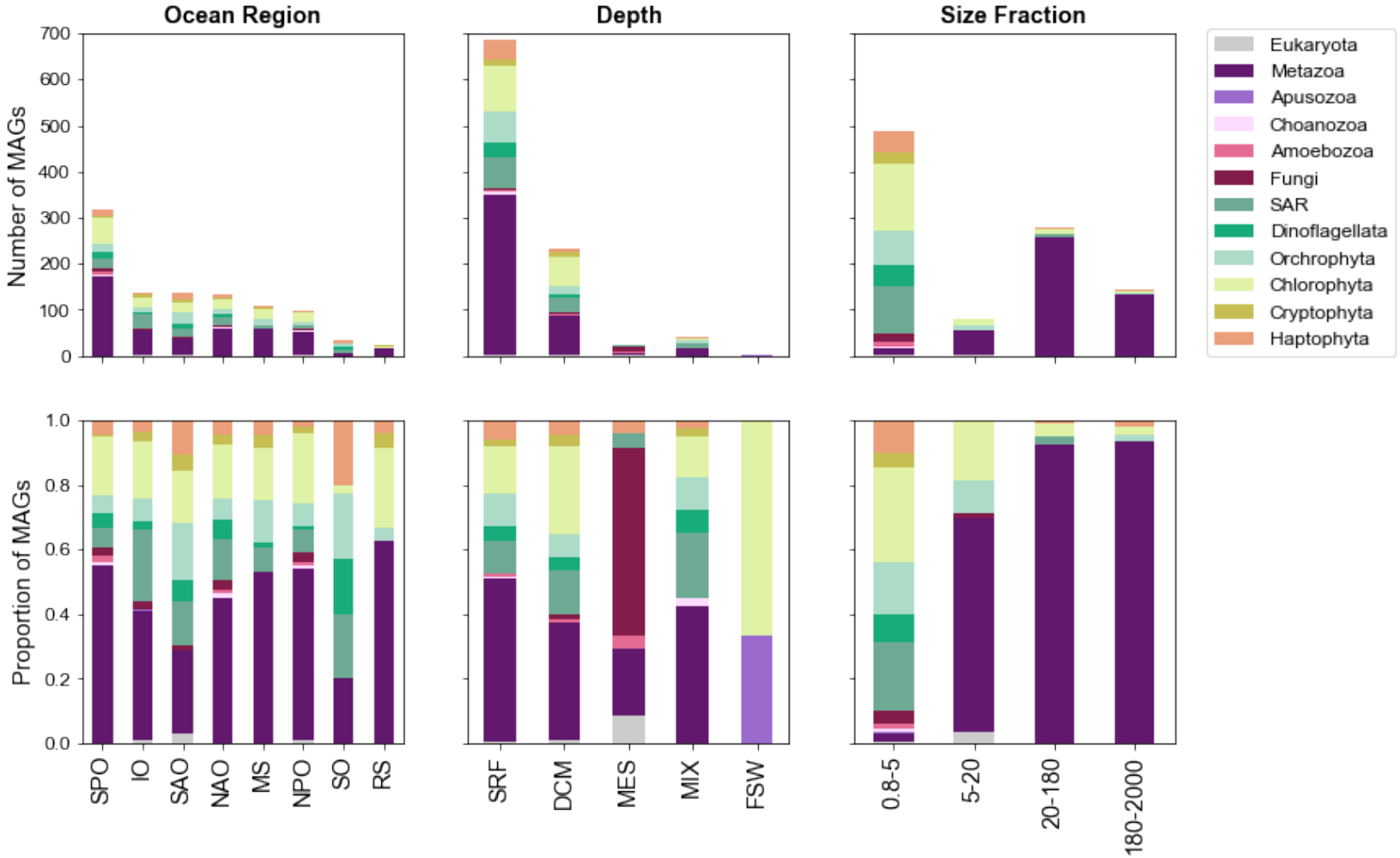

Figure S5: TOPAZ Eukaryotic MAG as recovered by assembly group. The taxonomic breakdown of eukaryotic MAGs recovered within each general type of assembly group (based on Ocean Region, Depth, and Size Fraction) for all eukaryotic MAGs recovered in this study (n=988). Taxonomy is shown both as a total number recovered (top) and as a proportion of MAGs recovered for a given category (bottom).

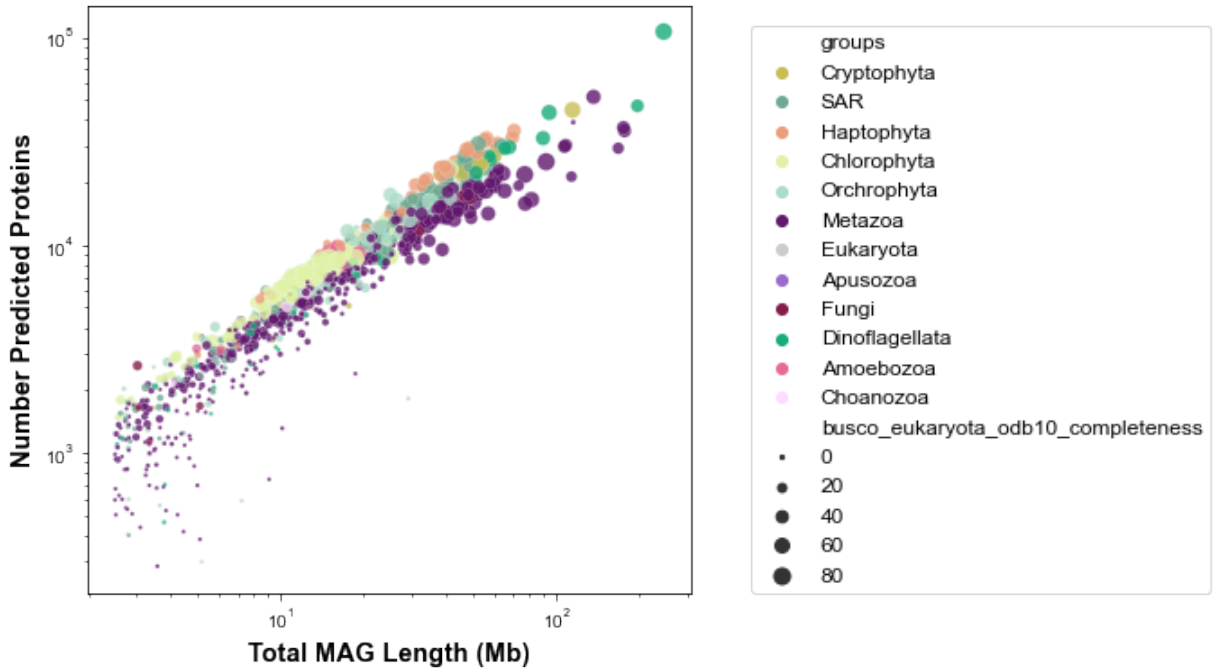

Figure S6: The number of predicted proteins as a function of total MAG length.The number of predicted proteins for each eukaryotic TOPAZ mag (n=988) is plotted against the total MAG length (Mb). Each MAG is colored by its taxonomic group and the size of the circle is scaled by the estimated BUSCO completeness.

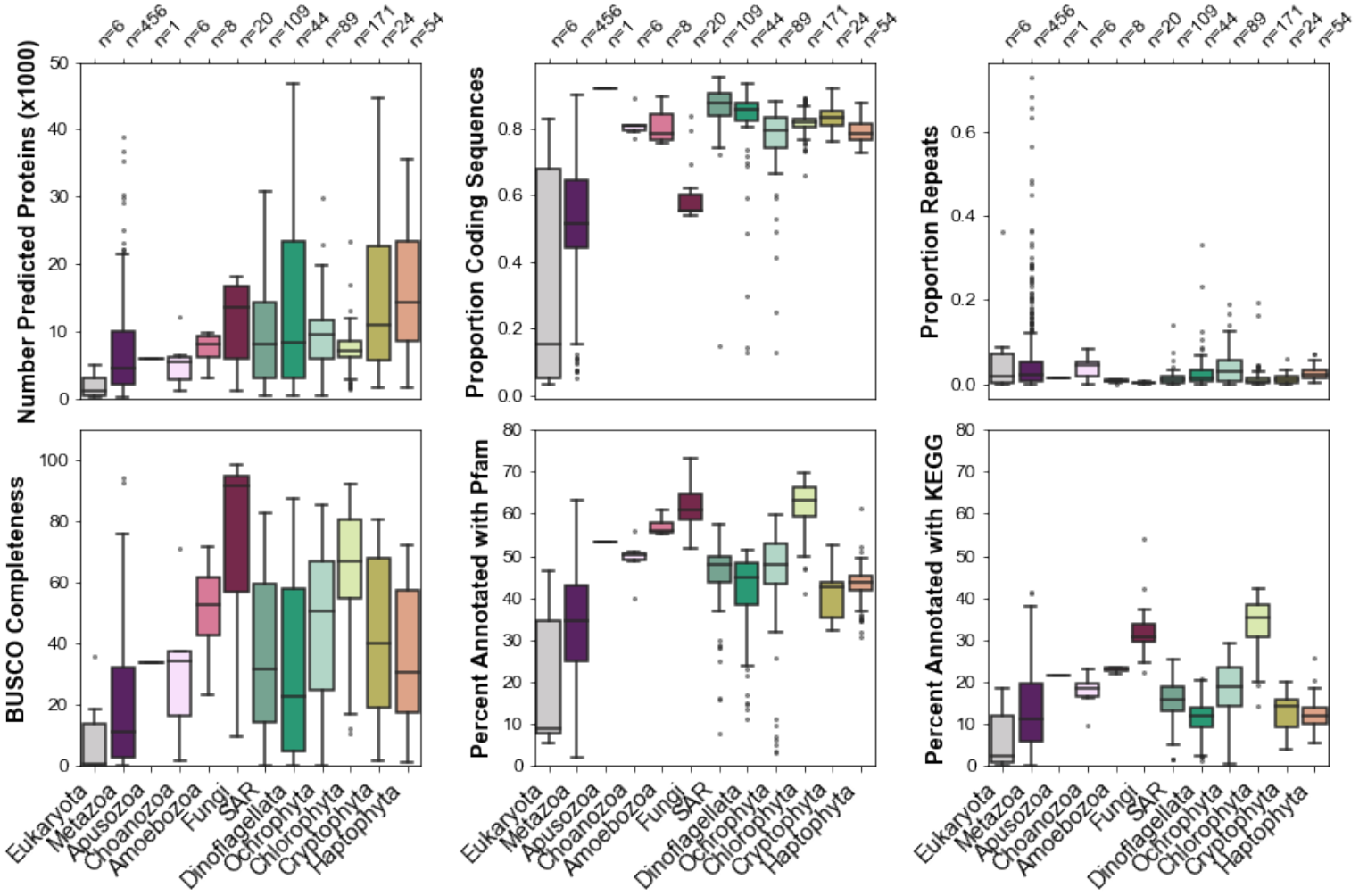

Figure S7: Genomic traits of recovered eukaryotic completeness, protein predictions, and annotation of all eukaryotic TOPAZ MAGs (n=988). The number of predicted proteins, proportion coding sequences, proportion repeat content, BUSCO completeness, percent annotation with Pfam and KEGG ontology are shown as box and whisker plots for the major higher-level groups that we define for this paper.

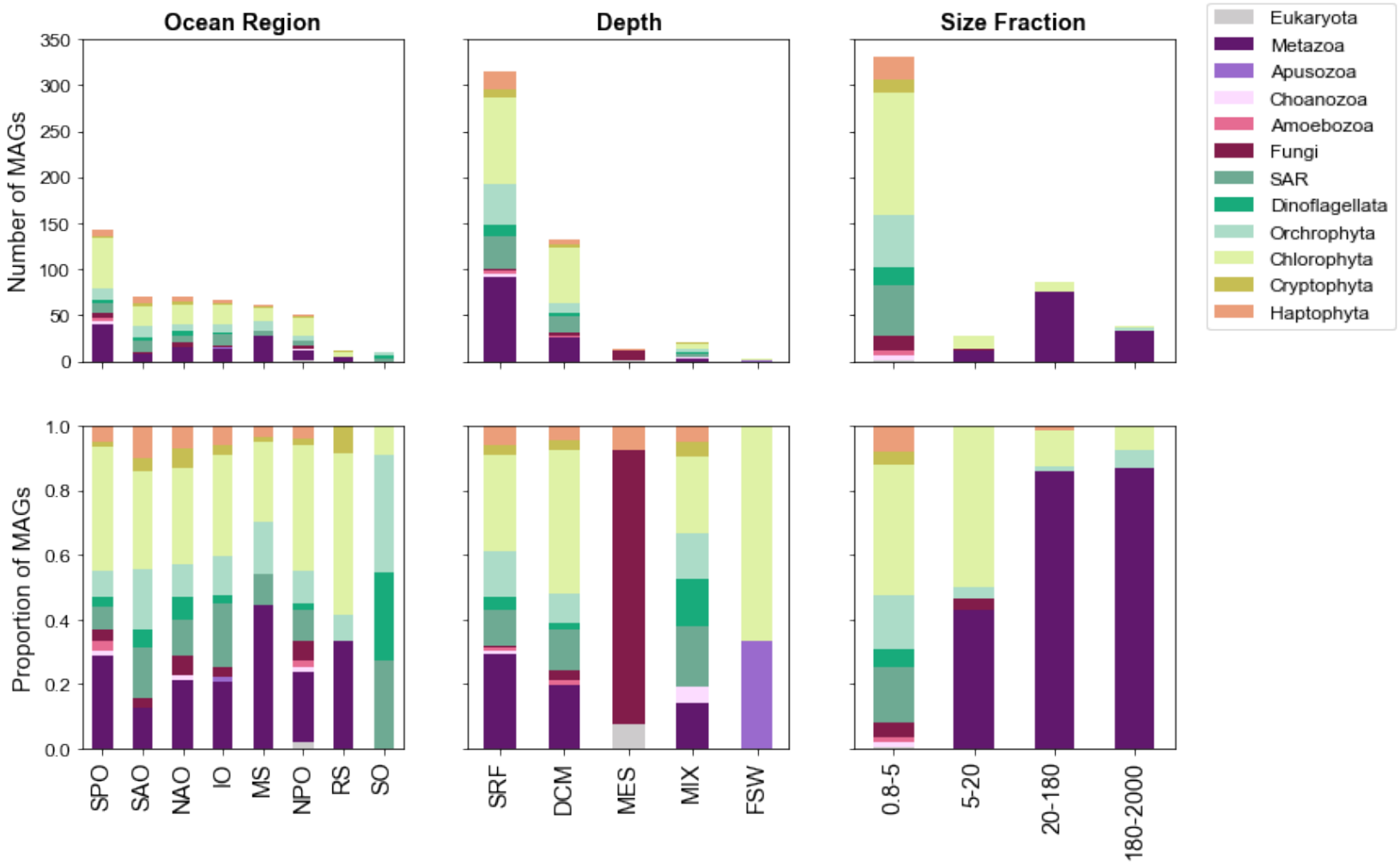

Figure S8: TOPAZ Highly Complete Eukaryotic MAGs as recovered by assembly group. The taxonomic breakdown of eukaryotic MAGs recovered within each general type of assembly group (based on Ocean Region, Depth, and Size Fraction) for highly complete eukaryotic MAGs (> 30% BUSCO completeness)recovered in this study (n=485)). Taxonomy is shown both as a total number recovered (top) and as a proportion of MAGs recovered for a given category (bottom).

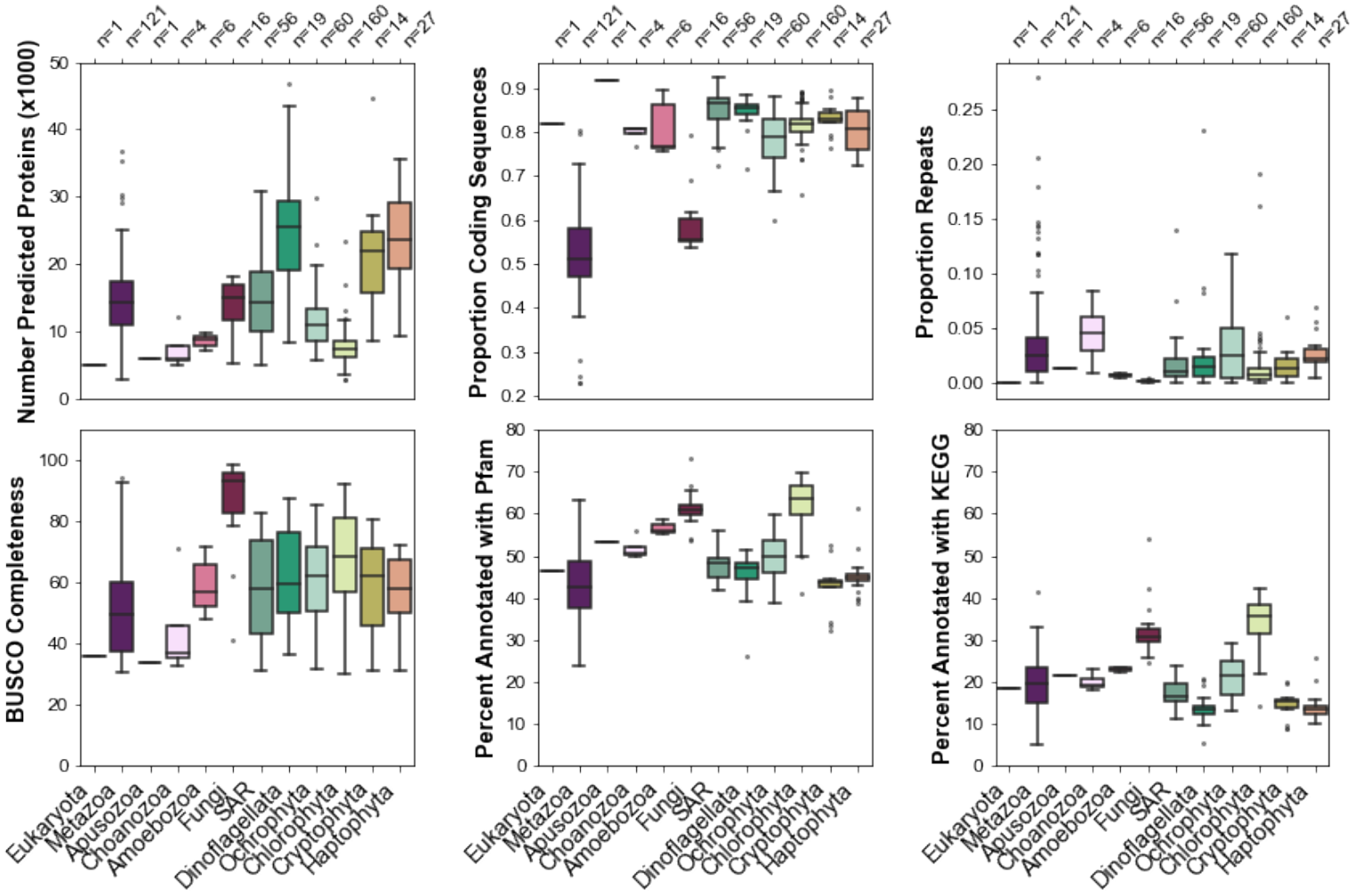

Figure S9: Genomic traits of recovered eukaryotic completeness, protein predictions, and annotation of the highly complete eukaryotic TOPAZ MAGs (n=485). The number of predicted proteins, proportion coding sequences, proportion repeat content, BUSCO completeness, percent annotation with Pfam and KEGG ontology are shown as box and whisker plots for the major higher-level groups that we define for this paper.

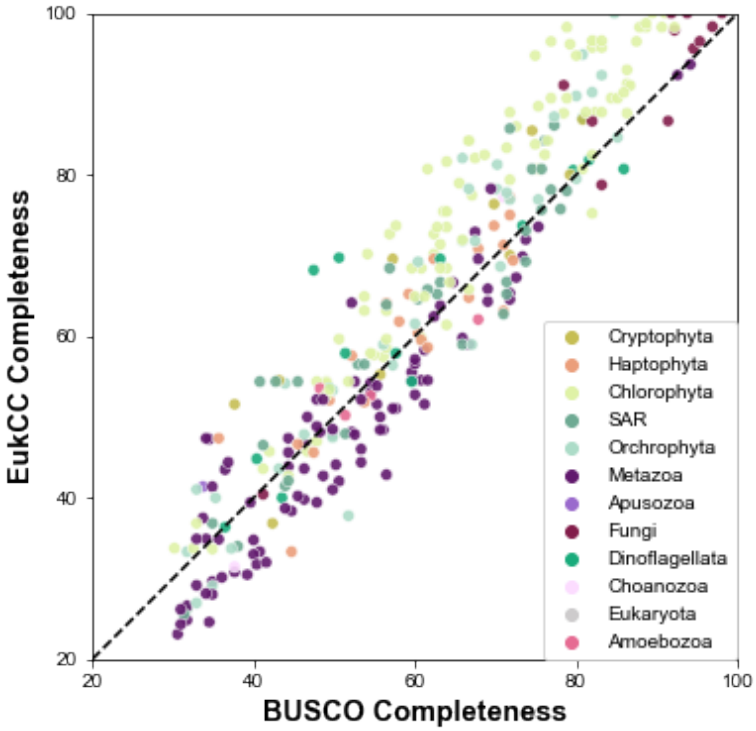

Figure S10: A comparison of two metrics of eukaryotic completeness. BUSCO completeness as defined based on the presence and absence of genes within the eukaryota_odb10 dataset was compared against the estimated EukCC completeness based on estimated lineages of given MAGs. Generally, it was observed that EukCC performed particularly well for certain groups (e.g. chlorophytes and fungi) and less well for others (e.g. metazoa).

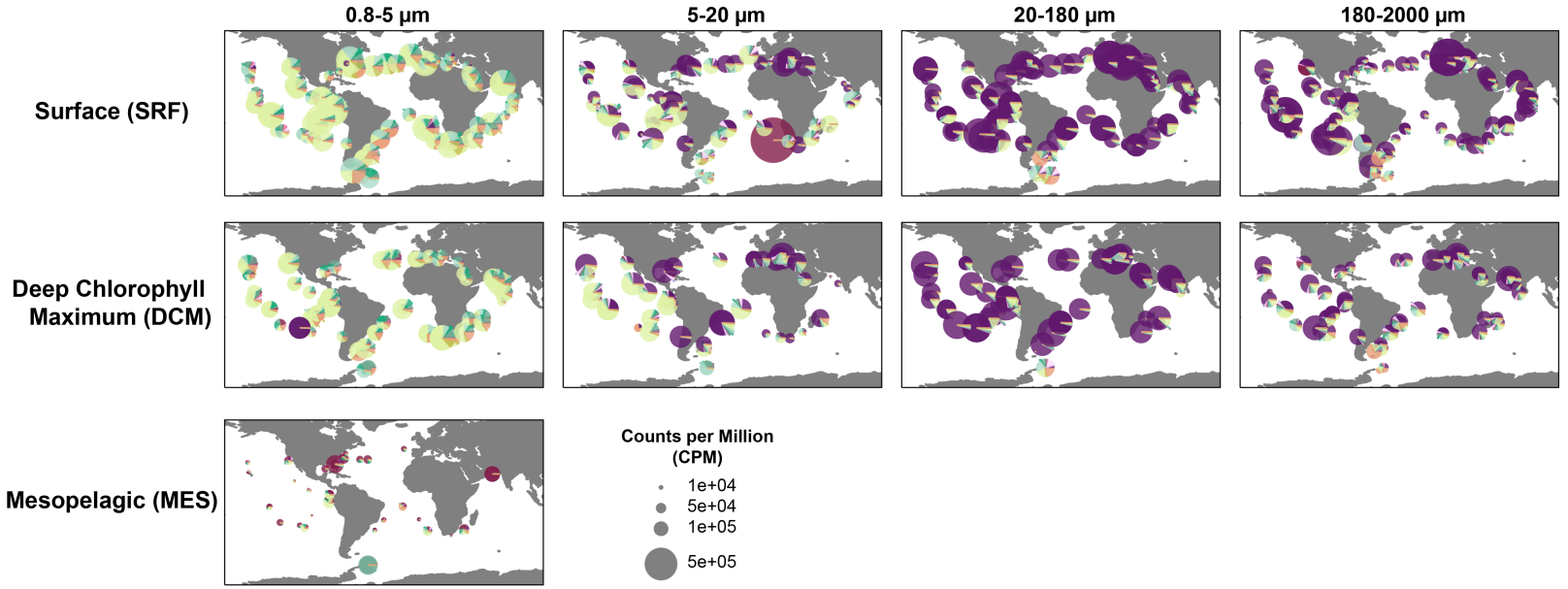

Figure S11: The distribution of the major lineages of eukaryotic TOPAZ MAGs recovered across the *Tara* Oceans metagenomic datasets. The counts per million (CPM) is depicted for all stations depicted by depth and size fraction.

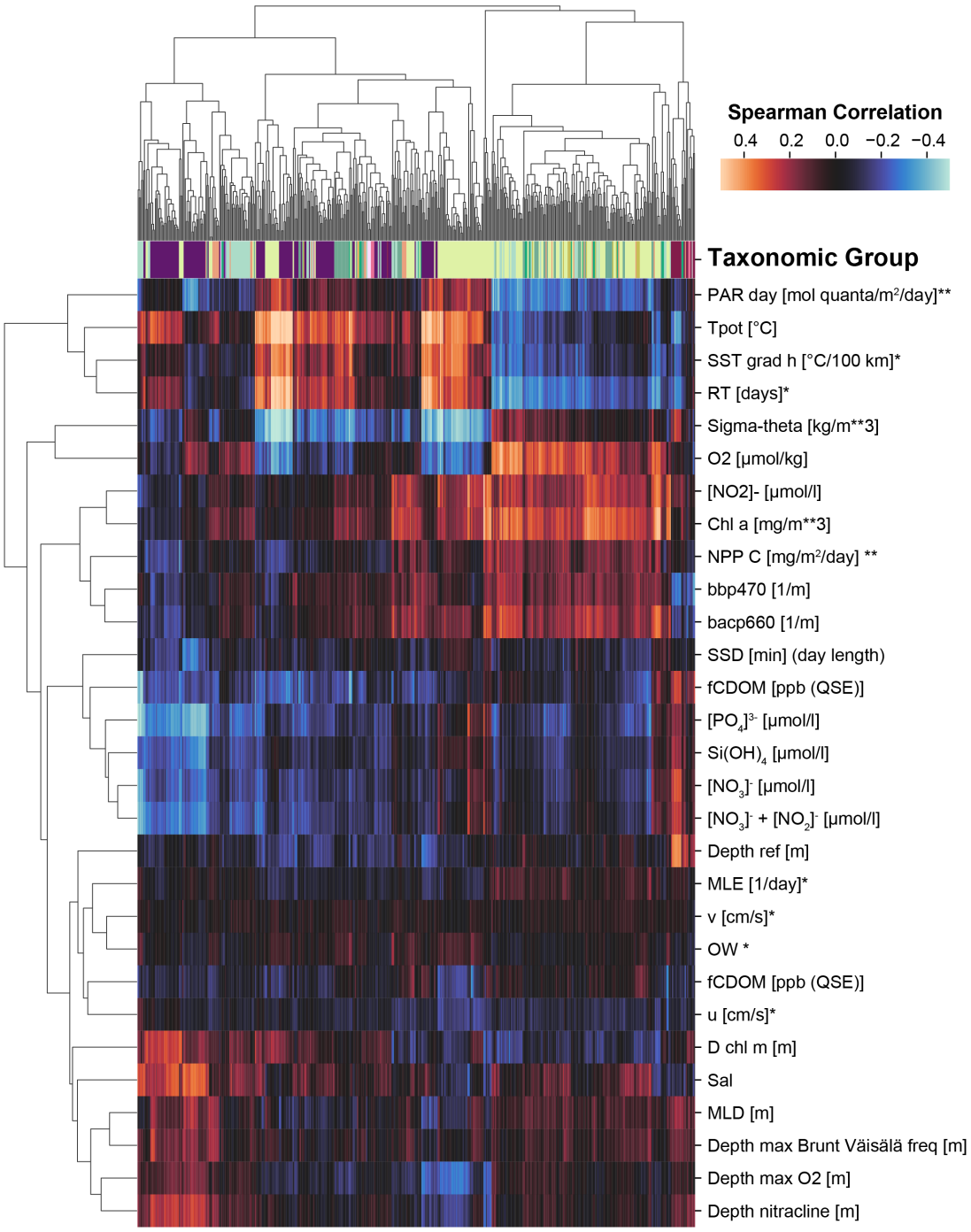

Figure S12: A Spearman correlation between the metagenomic abundance of each of the 485 high-completion eukayrotic TOPAZ MAGs and environmental parameters from the sampling (Tara Oceans Consortium and Tara Oceans Expedition, 2016), modeled mesoscale physical features based on d'Ovidio et al. (2010) (indicated with *) , and averaged remote sensing products (indicated with **).

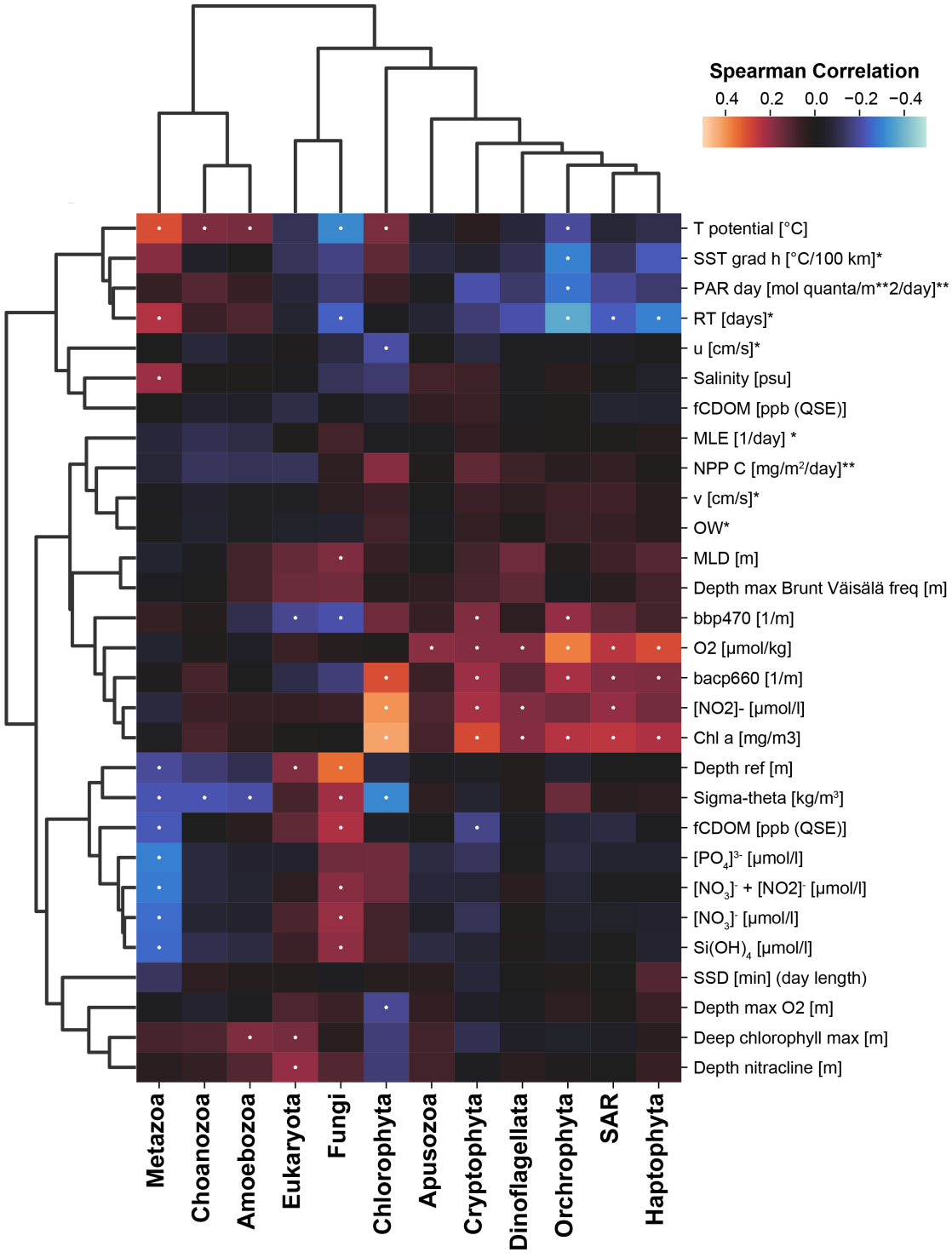

Figure S13: A Spearman correlation between the summed metagenomic abundance of each taxonomic group and environmental parameters from the sampling (Tara Oceans Consortium and Tara Oceans Expedition, 2016), modeled mesoscale physical features based on d'Ovidio et al. (2010) (indicated with *) , and averaged remote sensing products (indicated with **). Significant Spearman correlations, those with a Bonferroni adjusted $p < 0.01$, are indicated with a dot on the heatmap.

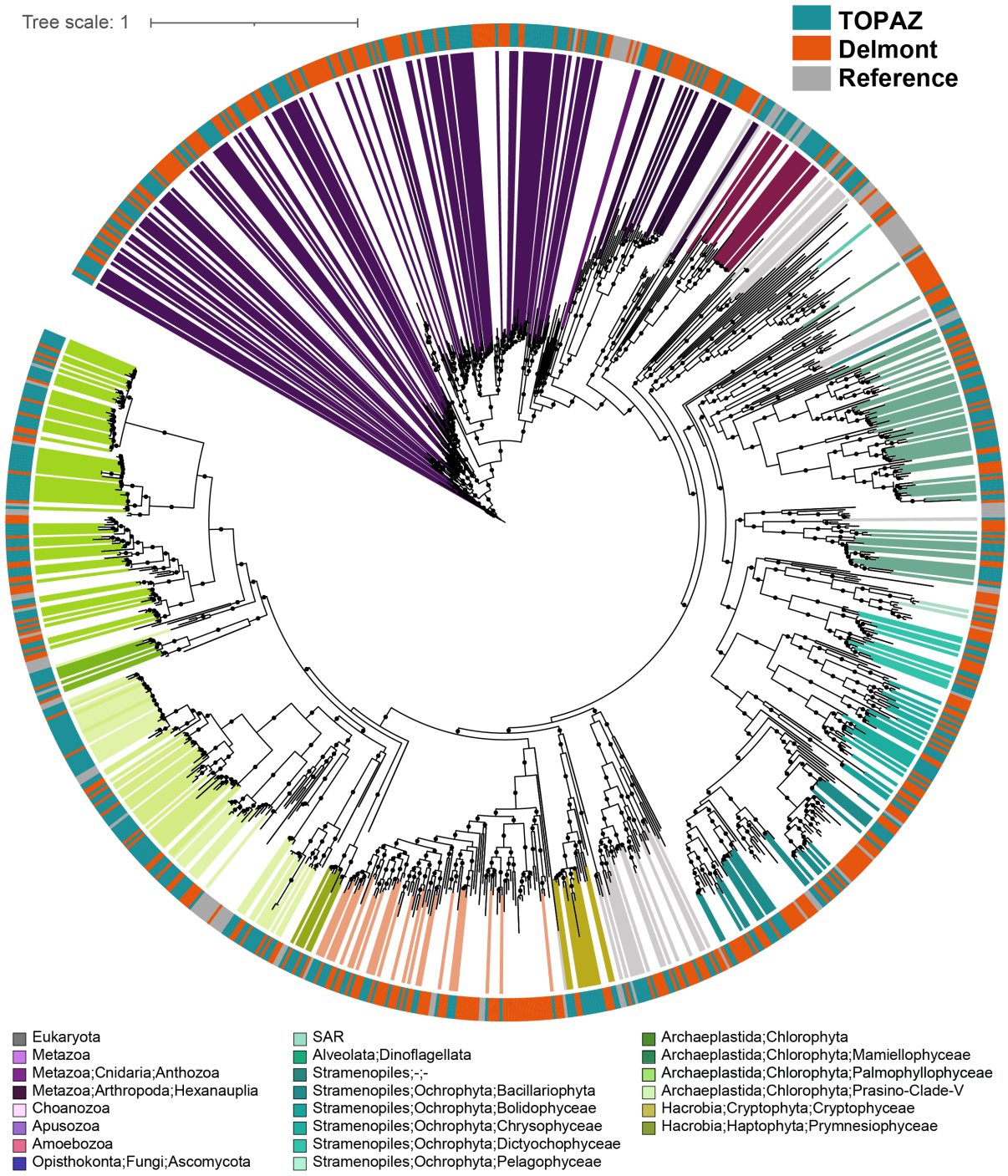

Figure S14: A concatenated protein phylogeny containing TOPAZ and Delmont (Delmont et al., 2020) eukaryotic MAGs that were estimated to be greater than 30% complete as well as reference genomes from cultured isolates. The maximum likelihood tree was inferred from a concatenated protein alignment of 49 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 same proteins that were used in Figure 1). Branches (nodes) corresponding to TOPAZ MAGs are colored based on consensus protein annotation estimated by EUKulele and MMSeqs.

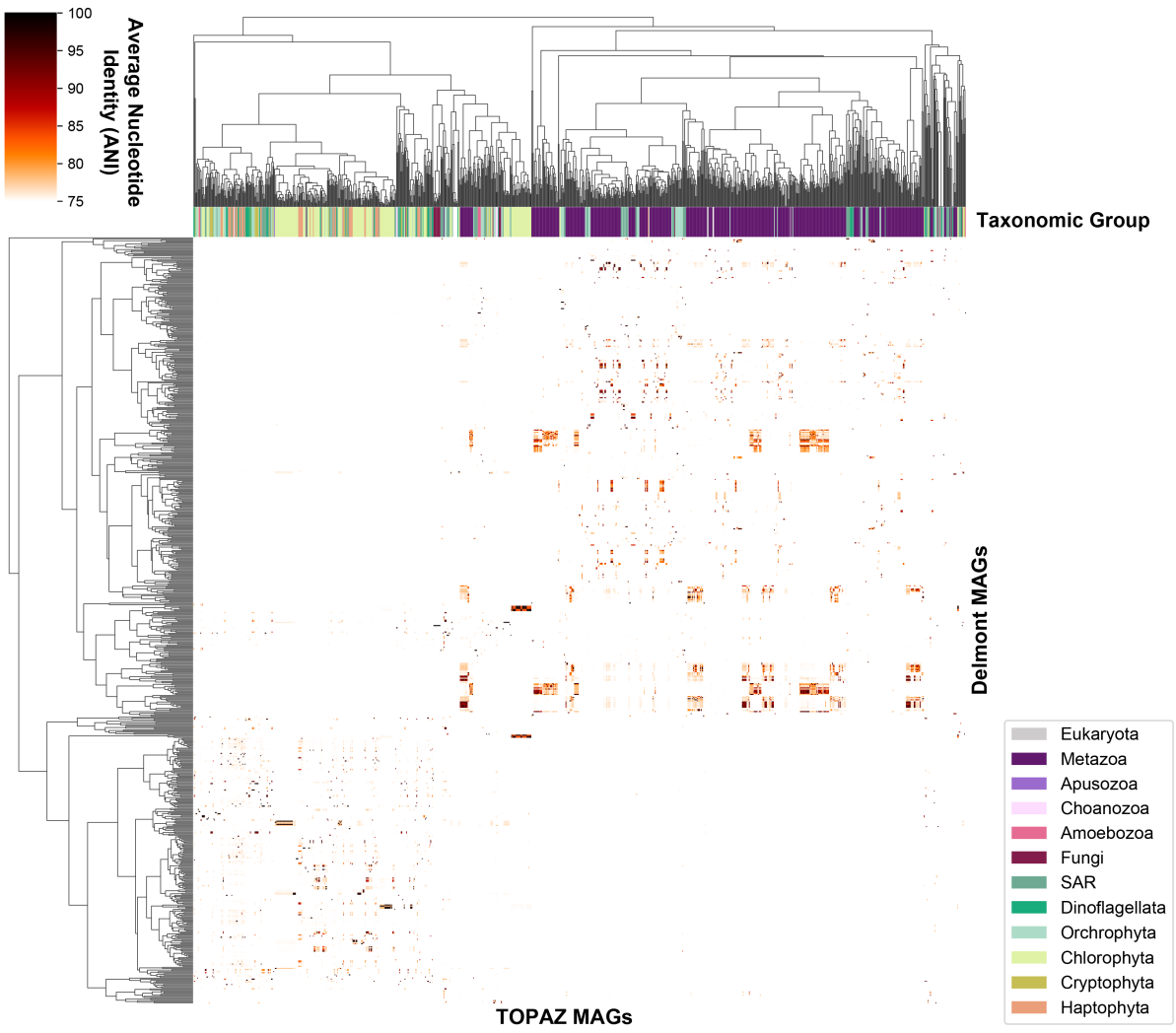

Figure S15: Average nucleotide identity (ANI) between the High Completion TOPAZ and Delmont MAGs. The ANI as estimated by FastANI between all eukaryotic TOPAZ MAGs and Delmont Eukaryotic MAGs is depicted. The taxonomic affiliation of the TOPAZ MAGs is depicted by color along the x-axis. A clustering was done with a cutoff of 99% ANI generating 83 unique clusters of MAGs, of which 46 contained a representative from both TOPAZ and Delmont (Supplementary Table 12).

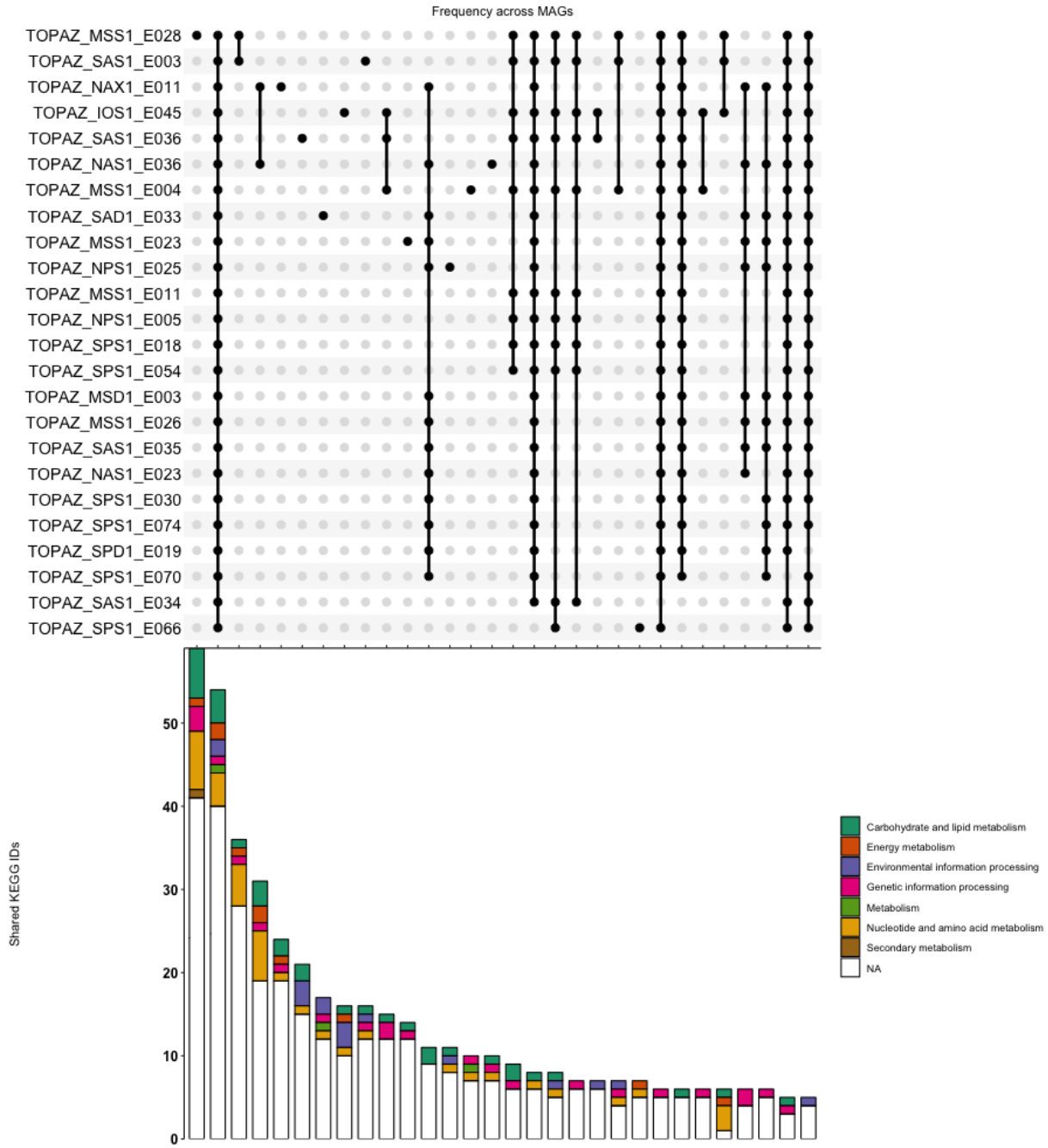

Figure S16: Intersection of shared or unique KEGG IDs across 24 SAR and *Dictyochophyceae* TOPAZ MAGs. Subset of MAGs (named on the left) indicated by dot grid (top; ggupset) and the total number of shared KEGG IDs (bottom) as barplots. The largest 30 sets of shared or unique KEGG ID intersections is shown. Combination of MAG intersections show by colors denote the KEGG module, where NA (white) indicates that the KEGG ID did not belong to a KEGG module.

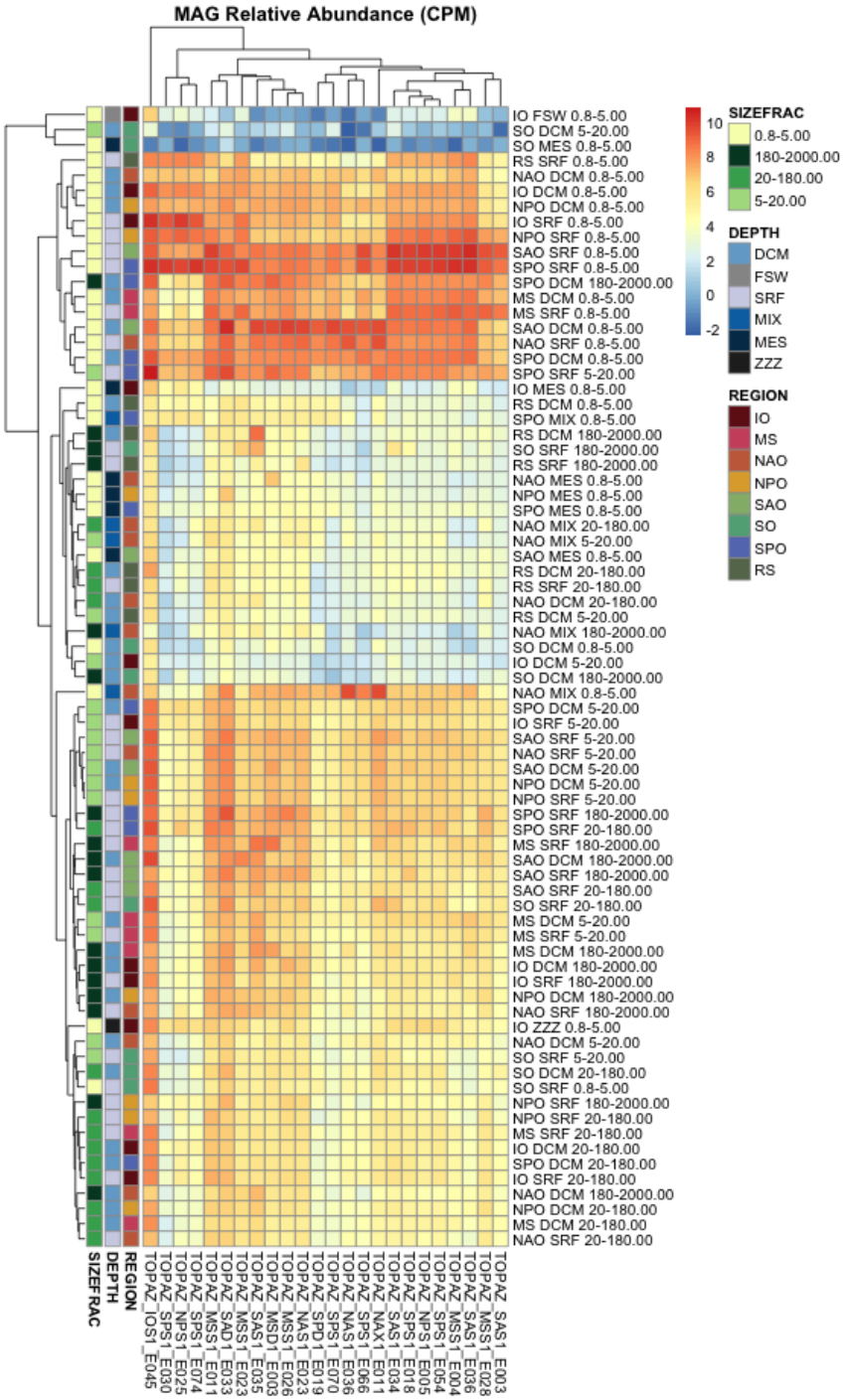

Figure S17: Relative counts per million (CPM) of 25 selected SAR and *Dictyochophyceae* TOPAZ MAGs (x-axis) by sample (y-axis and color bar). CPM of MAGs were summed by sample type, size fraction, depth, and ocean region, and Z-score transformed (heatmap signature) for visualization with pheatmap (R; v1.0.12). MAGs and samples are clustered based on average distance.

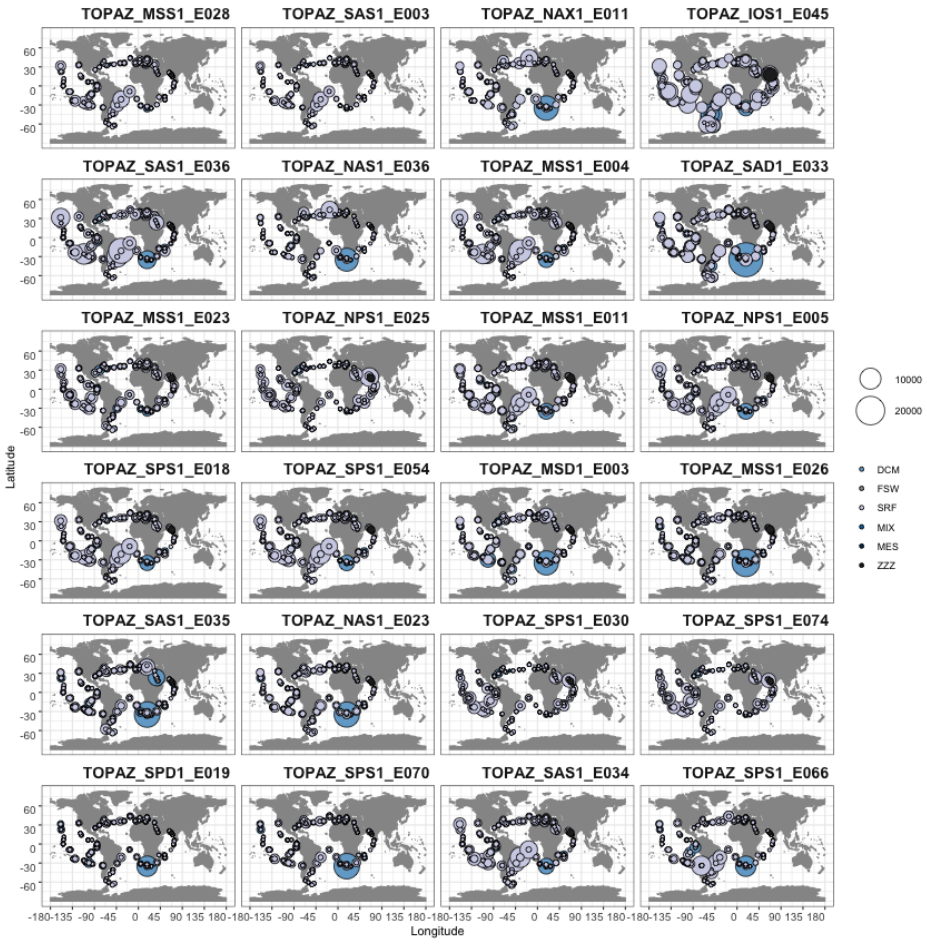

Figure S18: Metagenomic abundance plots for the 24 SAR and *Dictyochophyceae* TOPAZ MAGs considered. The relative abundance in CPM is shown as a bubble size. Color indicates the depth of the sample.

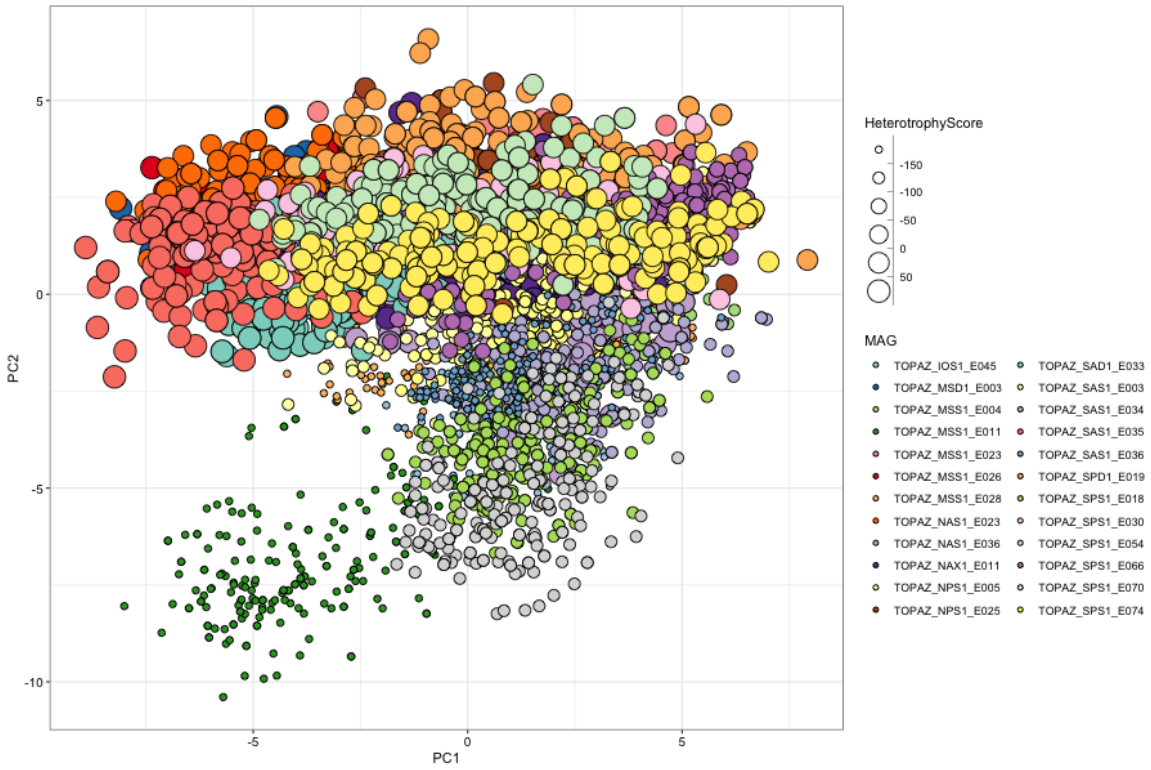

Figure S19: Principle component analysis derived from metatranscriptome reads, from the surface and smallest size fraction, mapped to shared orthologs (Shared in all MAGs in (Figure 5 b)) among selected 24 TOPAZ MAGs. Symbol size designates Heterotrophy Index, while symbol color denotes each TOPAZ MAG.

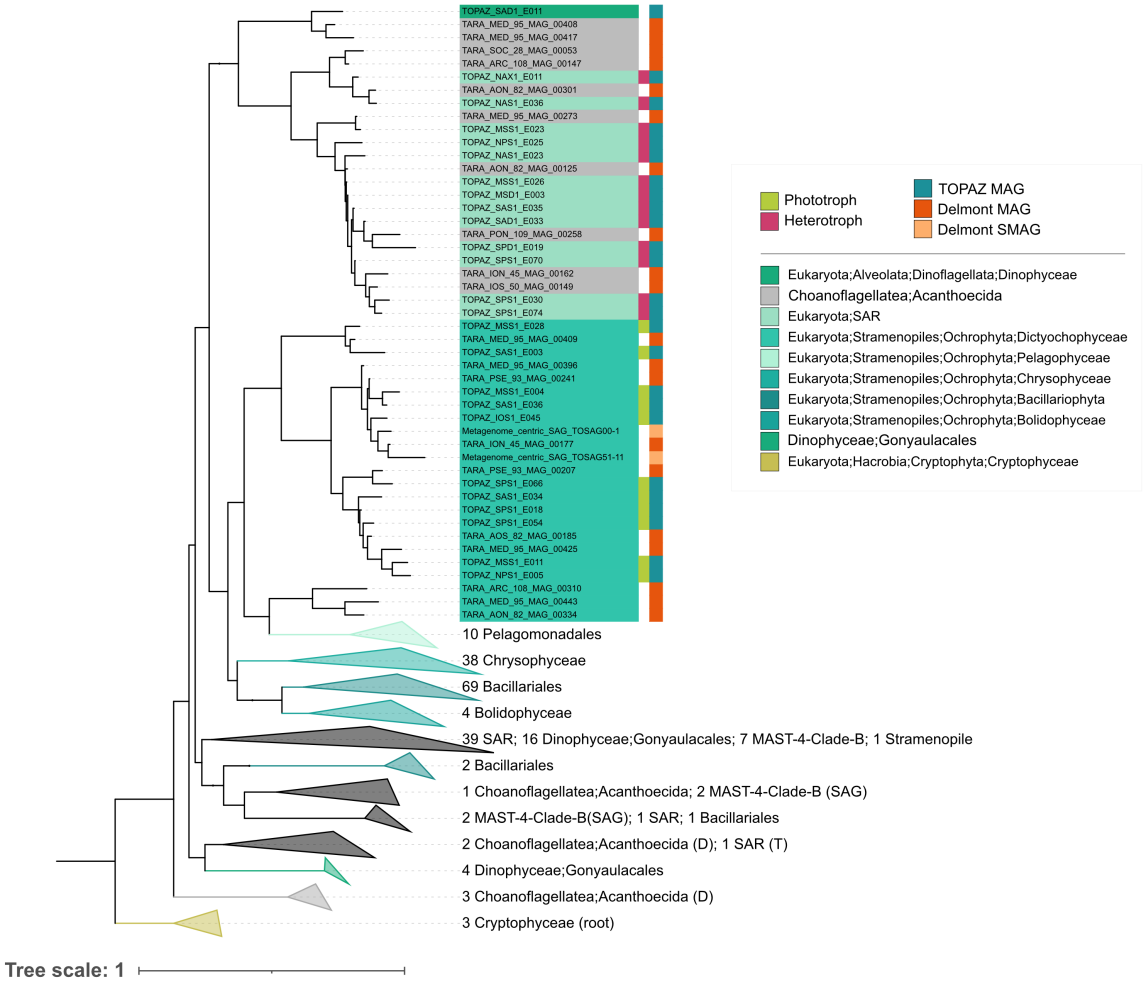

Figure S20: Concatenated BUSCO phylogeny of the 24 selected TOPAZ MAGs with a selection of other stramenopile MAGs and SMAGs from Delmont et al. (2020). The phylogeny is pictured with iTOL. Collapsed clades include a mixture of TOPAZ and Delmont-generated MAGs that are listed.

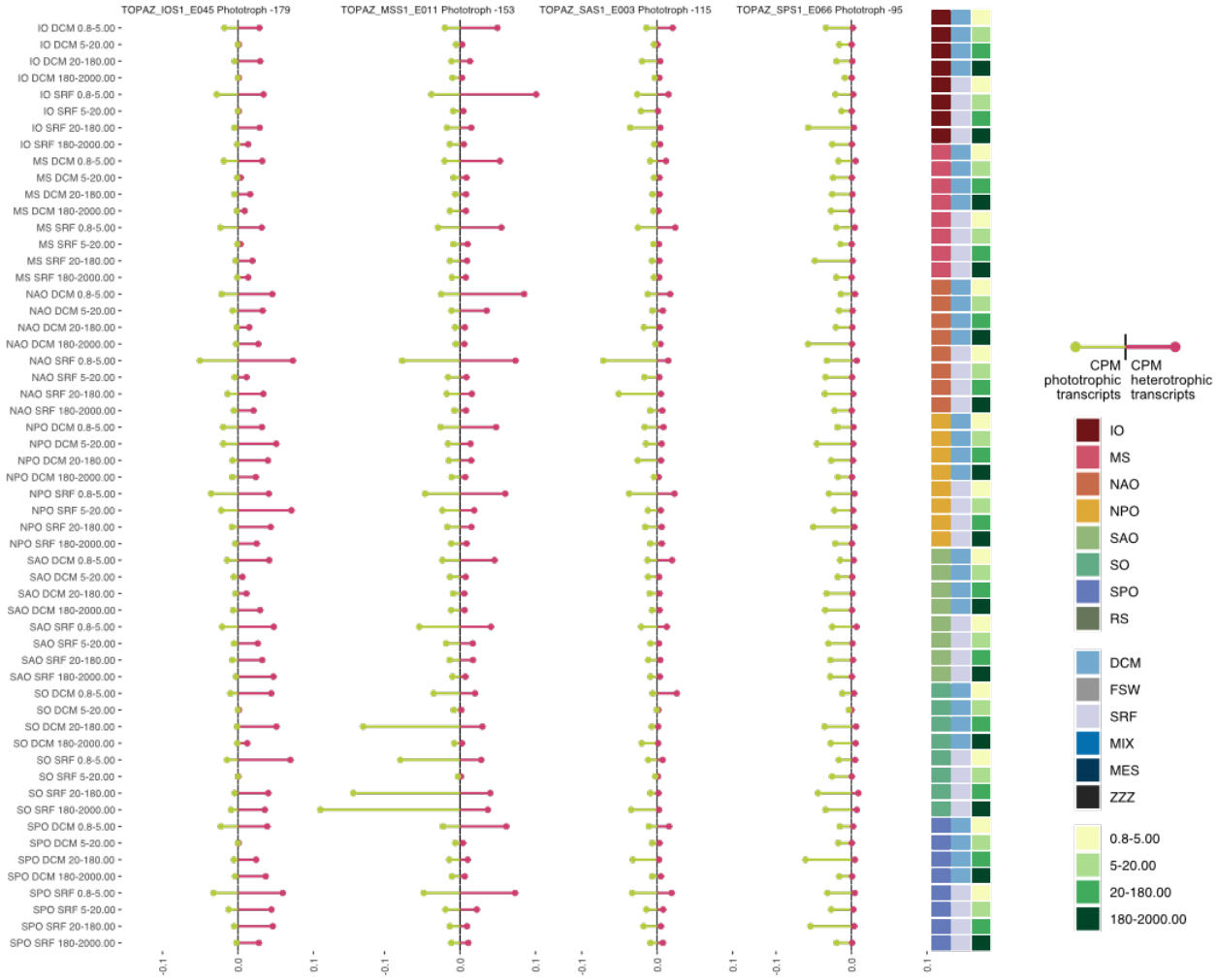

Figure S21: Lollipop plot showing the relative CPM of mapped transcripts associated with more phototrophic (left) or heterotrophic (right) traits for 4 *Dictyochophyceae* MAGs (left to right). Relative abundance is shown by MAG for each sample (y-axis and color bar). Shown TOPAZ MAGs were classified as primarily phototrophic (low H-index, reported next to each TOPAZ MAG name), but relative abundance of mapped transcripts varies by sample type.

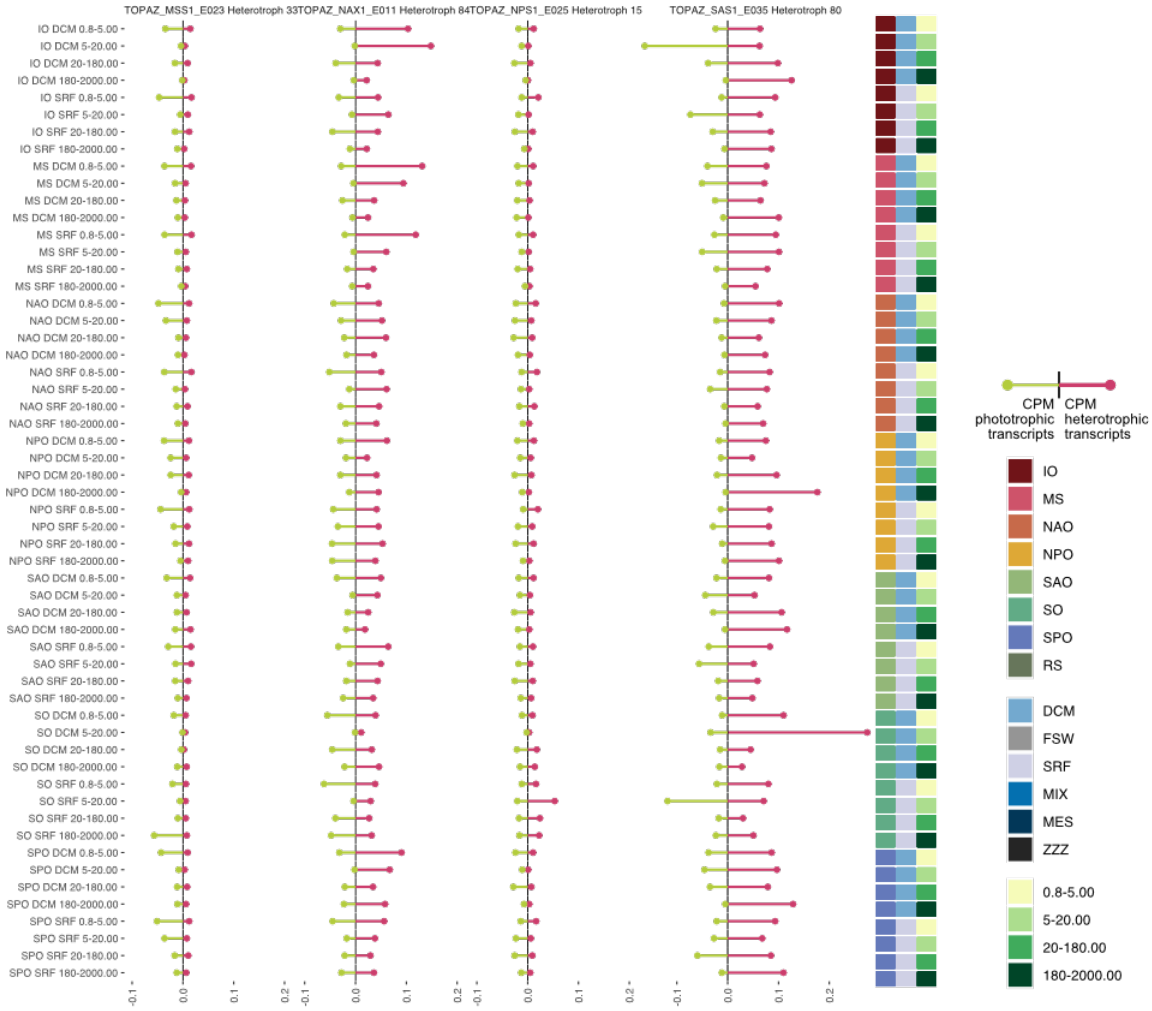

Figure S22: Lollipop plot showing the relative CPM of mapped transcripts associated with more phototrophic (left) or heterotrophic (right) traits for 4 SAR TOPAZ MAGs (left to right). Relative abundance is shown by MAG for each sample (y-axis and color bar). TOPAZ MAGs selected are a subset from the SAR clade closely related to stramenopiles and were predicted to be heterotrophic (high H-index; reported next to each TOPAZ MAG name).

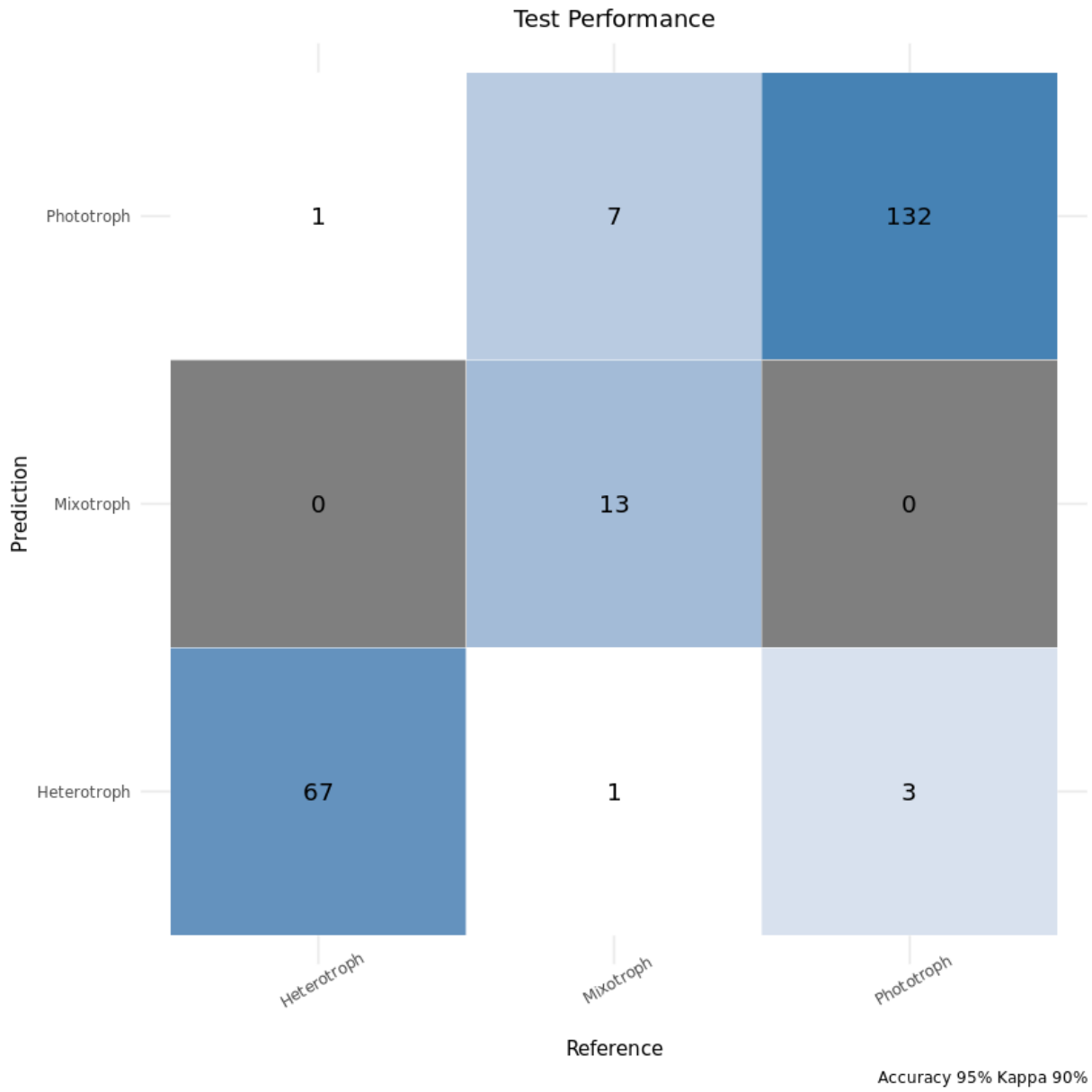

Figure S23: Confusion matrix generated from Random Forest model construction on the 25% of the reference transcriptomes used for testing. The *x* categorical axis is the manually-annotated trophic mode, while the *y* categorical axis is the predicted trophic mode via the Random Forest model. The greatest number of erroneous predictions involved mixotrophy (8 of 12).

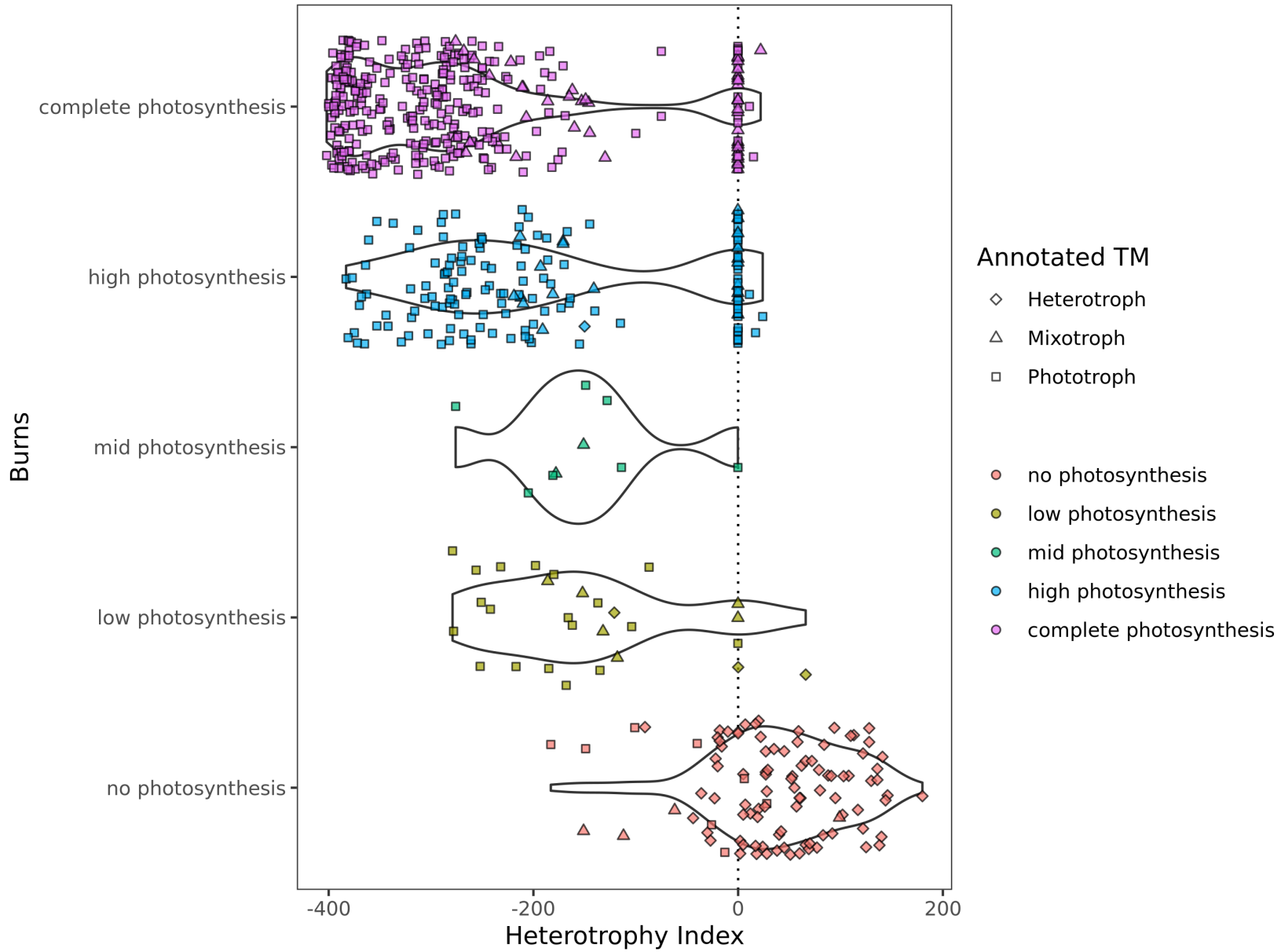

Figure S24: Comparison of Heterotrophy Index scores with the Burns model prediction for photosynthetic ability. On the *y* axis, MAGs are split into categories based on their photosynthesis score from the Burns model (which ranges from zero to one): "complete photosynthesis" (0.9-1), "high photosynthesis" (0.55-0.9), "mid photosynthesis" (0.45-0.55), "low photosynthesis" (0.1-0.45) and "no photosynthesis" (0-0.1).

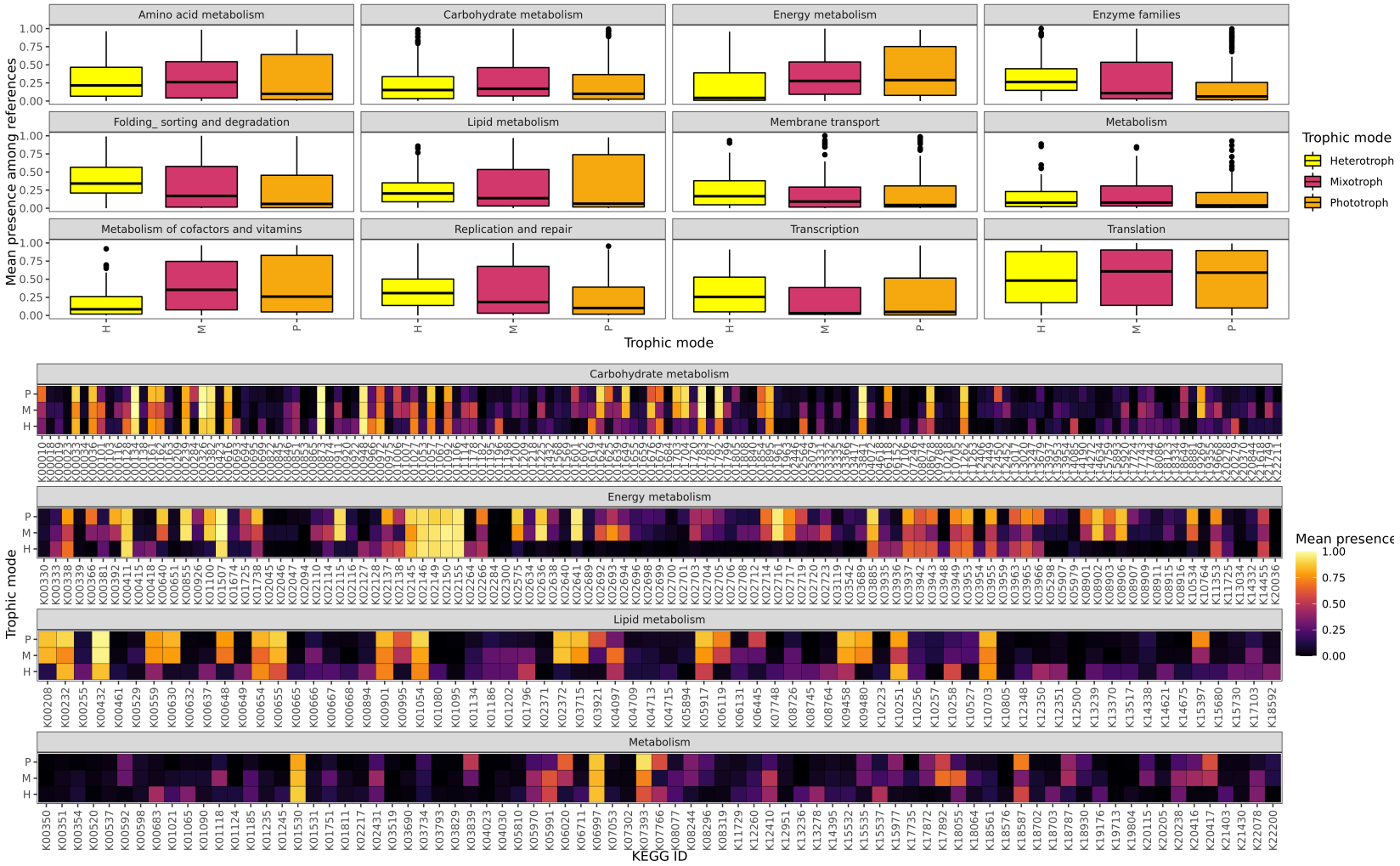

Figure S25: Presence-absence of KEGG IDs as aggregated across all reference transcriptomes (MMETSP and EukProt) used in the trophic mode modeling process. The top boxplot panel shows the mean presence of the KEGG IDs implicated in the listed pathways for each of the three identified trophic modes, while the bottom heatmap panel shows the average presence of each individual KO along the horizontal axis for reference transcriptomes annotated as each major trophic mode (vertical axis).

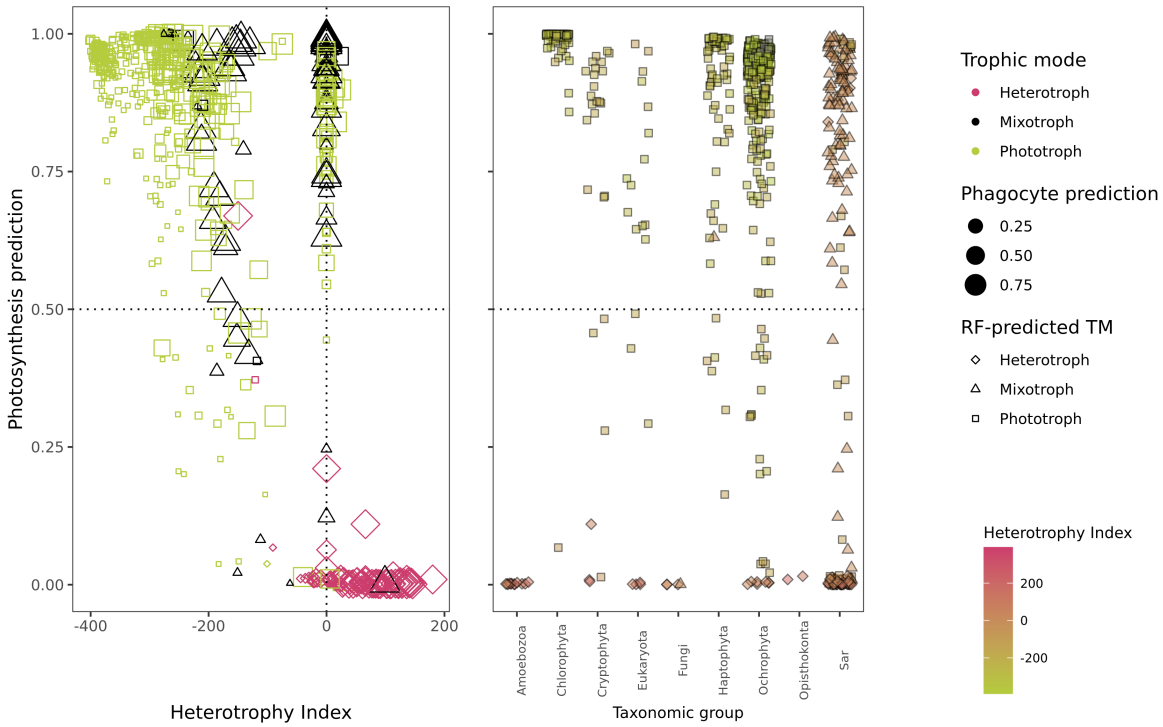

Figure S26: Comparison of the Burns (Burns et al., 2018) model to our Random Forest predictions and Heterotrophy Index calculations for the reference MMETSP transcriptomes. Left: Burns Burns et al. (2018) photosynthesis predictions vs. composite Heterotrophy Index, scaled by the phagocytosis prediction and colored by the manually-annotated trophic mode. Right: predicted photosynthetic ability by taxonomic grouping, colored by the calculated Heterotrophy Index.

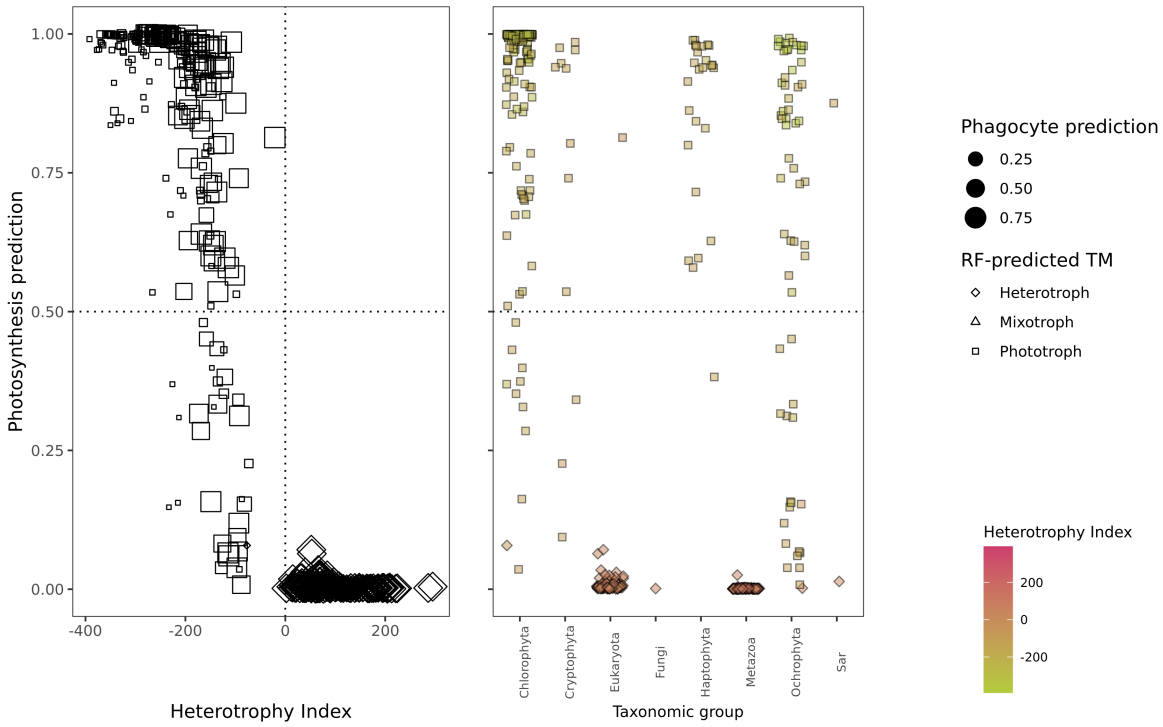

Figure S27: Comparison of the Burns (Burns et al., 2018) model to our Random Forest predictions and Heterotrophy Index calculations for the TOPAZ MAGs. Left: Burns Burns et al. (2018) photosynthesis predictions vs. composite Heterotrophy Index, scaled by the phagocytosis prediction and shape indicating the Random Forest-derived predicted trophic mode (note there is no color because trophic mode could not be manually annotated). Right: predicted photosynthetic ability by taxonomic grouping, colored by the calculated Heterotrophy Index.

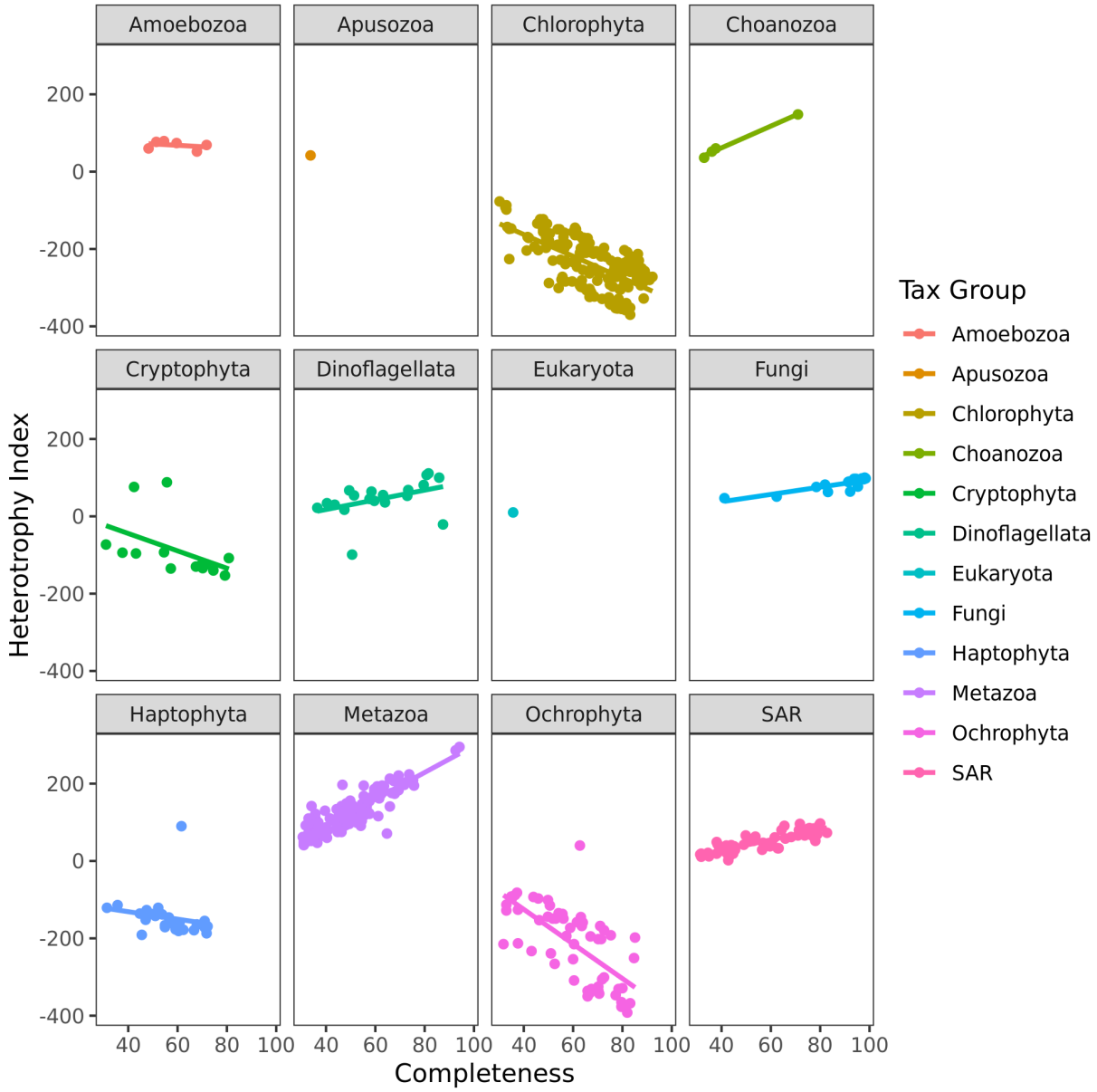

Figure S28: Heterotrophy Index scores as a function of BUSCO completeness, faceted by taxonomy phylum. The strongest association of Heterotrophy Index with completeness was found within Metazoa, for which it appears that the magnitude of the Heterotrophy Index is tightly coupled to the level of completeness of the MAG. By contrast, Cryptophyta (which are known to be mixotrophic (Jones, 2000)) showed a much weaker coupling of completeness with the magnitude of the Heterotrophy Index, and all values of the Index were closer to zero.

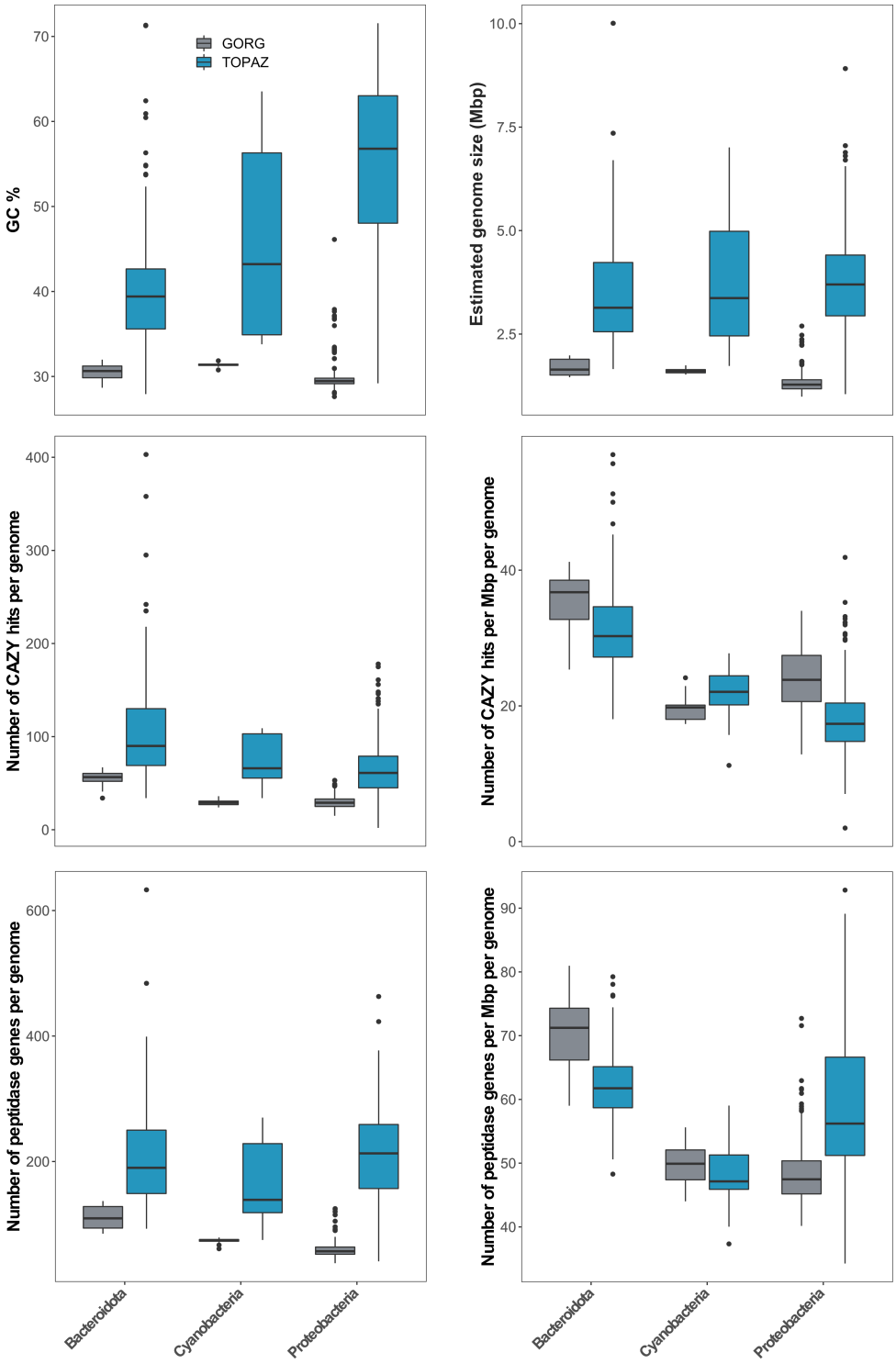

Figure S29: Comparison of the genomic characteristics between the TOPAZ MAGs and GORG SAGs belonging to the phyla, Bacteroidota, Cyanobacteria and Proteobacteria. The distributions of GC % content, estimated genome size (Mbp), number of CAZY hits per genome, number of CAZY hits per Mbp per genome, number of peptidase genes per genome and number of peptidase genes per Mbp per genome are presented as box plots.

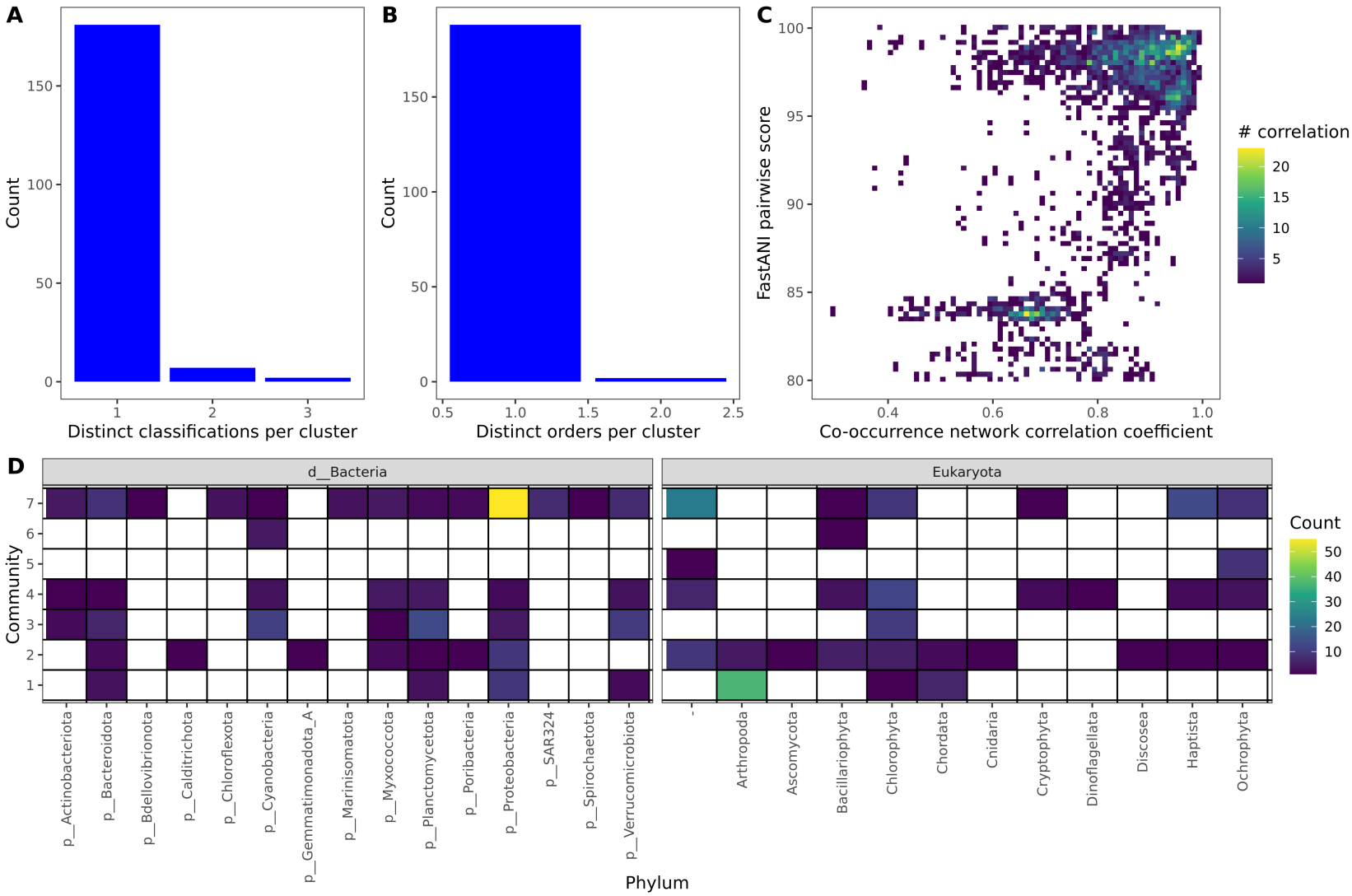

Figure S30: Supporting figures for the network analysis section. A and B: FastANI clustering does not typically group together eukaryotic MAGs of different overall taxonomic classification (A) or taxonomic order (B). C: Justification for ANI cutoff of 95% (and correlation coefficient 0.504) for considering eukaryotic MAGs to be sufficiently similar to be clustered. The majority of MAGs with pairwise ANI scores of $\geq$95 have correlation coefficients of $\geq$0.8. D: Taxonomic composition of the 7 identified communities from the main text.

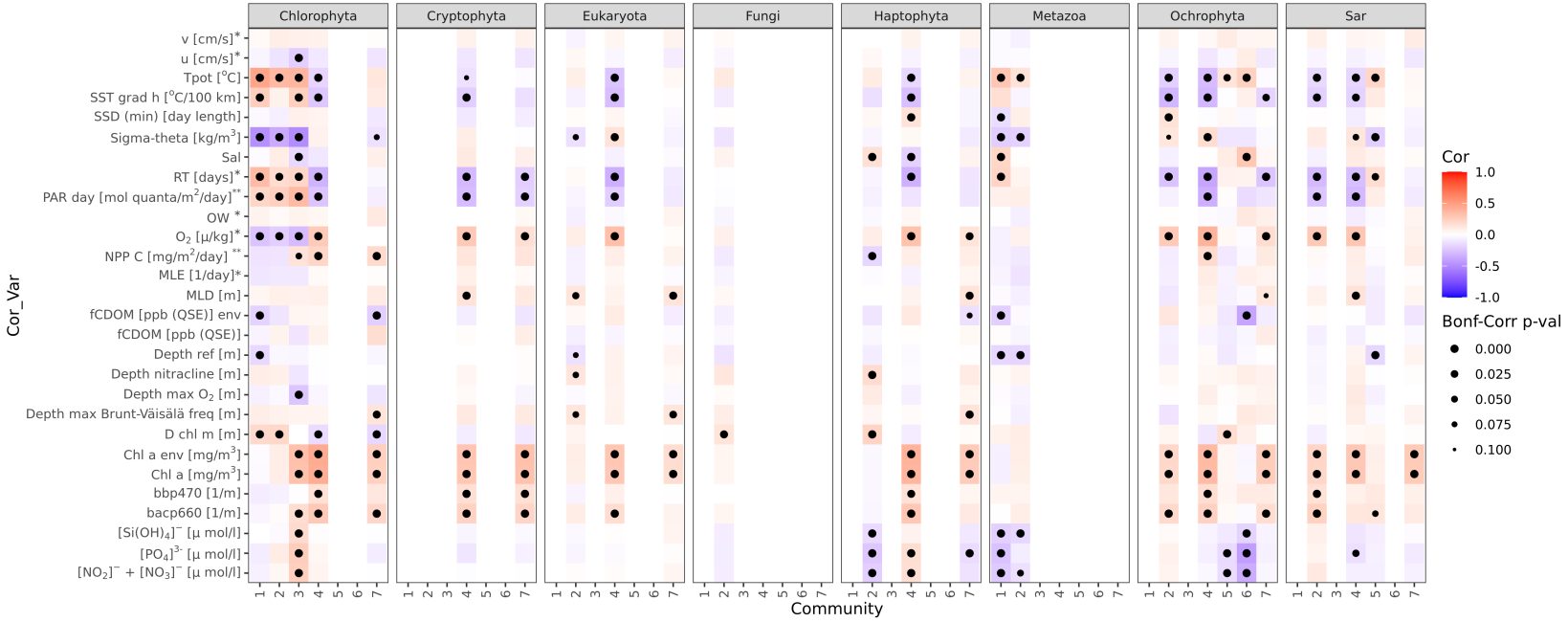

Figure S31: Extended environmental correlations figure displaying the strength and directions of environmental correlations as tracked by both taxonomic group and network-based community.

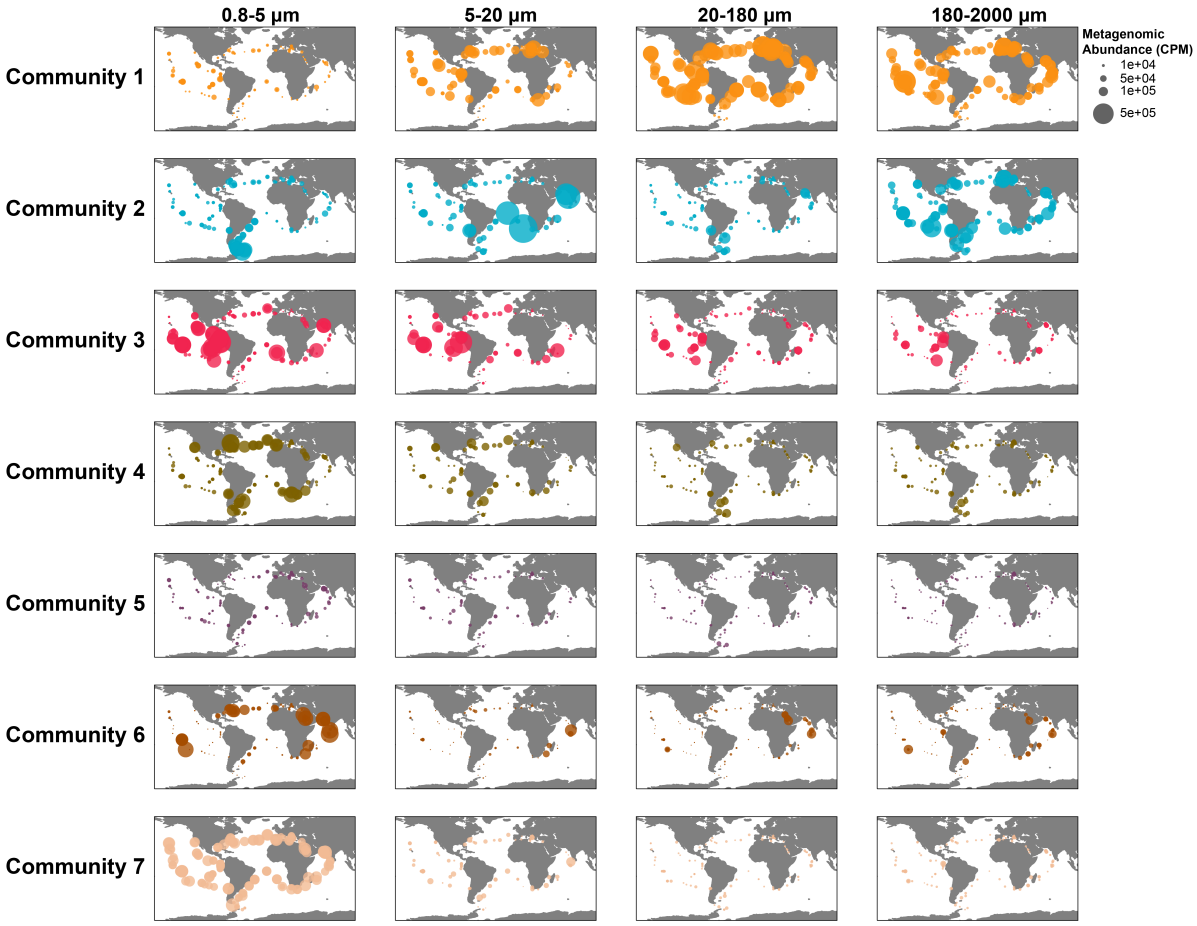

Figure S32: Distribution of identified communities across *Tara* Oceans Samples. The summed metagenomic abundance (CPM) of each of the seven identified communities from the network analysis (Figure 6) is shown for each size fraction. The size of the bubble indicates the relative abundance of the community in a given sample.

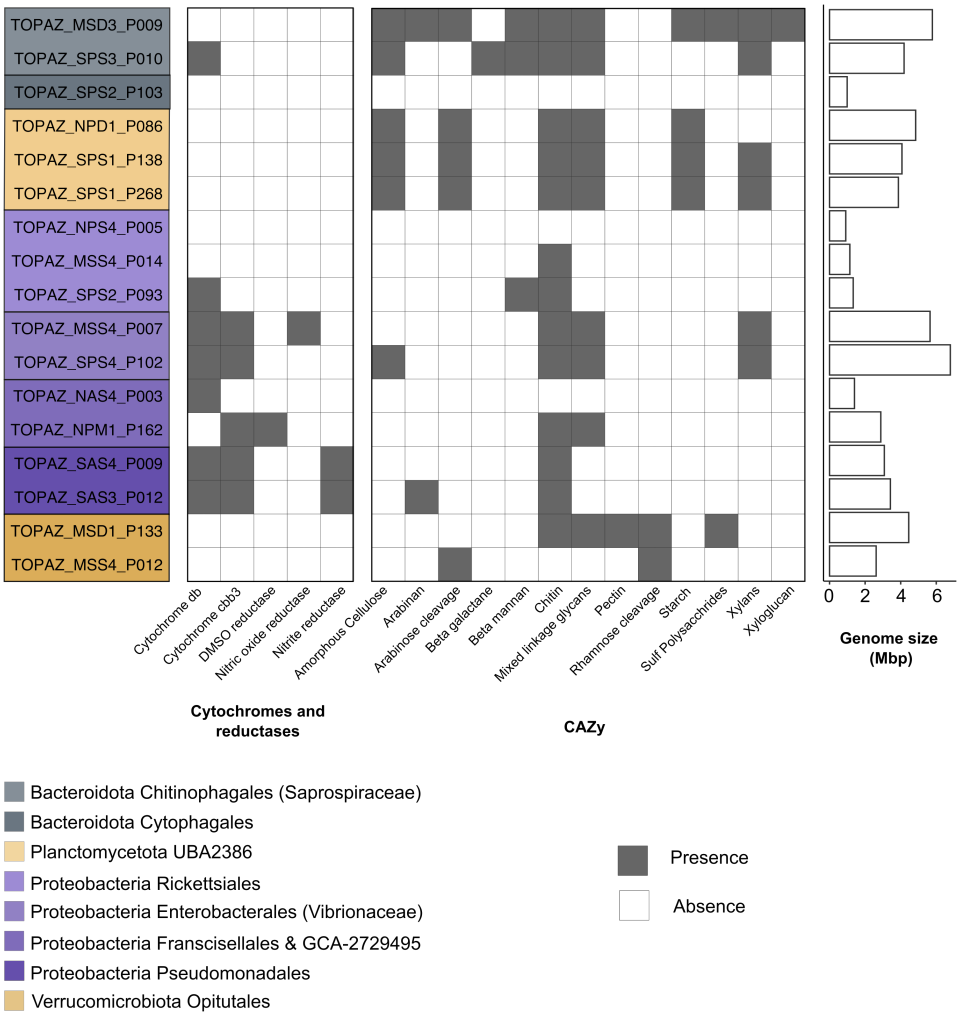

Figure S33: Genomic characteristics of the Community 1 MAGs. MAG names are color-coded based on the taxonomic affiliations estimated by GTDB-tk. The presence of the high oxygen affinity cytochromes db ubiquinol oxidase and cbb3, as well as the reductases (involved in anaerobic processes) DMSO family type II, nitric oxide and nitrite are noted with grey in the left panel of the heat map. The presence of genes involved in CAZy pathways for the hydrolysis/degradation of amorphous cellulose, arabinan, arabinose cleavage, beta-galactan, beta-mannan, chitin, mixed-linkage glucans, pectin, rhamnose cleavage, starch, sulf-polysachharides, xylans, xyloglucan are noted with grey in the right panel of the heat map. The estimated size of the genomes is shown as a bar plot.

References

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.

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.

L. Collins, G. McCarthy, A. Mellor, G. Newell, and L. Smith. Training data requirements for fire severity mapping using Landsat imagery and random forest. *Remote Sensing of Environment*, 245: 111839, 2020.

T. O. Delmont, M. Gaia, D. D. Hinsinger, P. Fremont, C. Vanni, A. F. Guerra, A. M. Eren, A. Kourlaiev, L. d'Agata, Q. Clayssen, E. Villar, K. Labadie, C. Cruaud, J. Poulain, C. D. Silva, M. Wessner, B. Noel, J.-M. Aury, C. de Vargas, C. Bowler, E. Karsenti, E. Pelletier, P. Wincker, and O. J. and. Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. oct 2020. doi: 10.1101/2020.10.15.341214. URL https://doi.org/10.1101%2F2020.10.15.341214.

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.

R. I. Jones. Mixotrophy in planktonic protists: an overview. *Freshwater Biology*, 45(2):219–226, 2000.

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.

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, 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.

M. Khalilia, S. Chakraborty, and M. Popescu. Predicting disease risks from highly imbalanced data using random forest. *BMC medical informatics and decision making*, 11(1):1–13, 2011.

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.

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. 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.

L. V. Utkin. An imprecise deep forest for classification. *Expert Systems with Applications*, 141: 112978, 2020.

A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson. Machine learning algorithm validation with a limited sample size. *PloS ONE*, 14(11):e0224365, 2019.