Supplementary Information for

Niche adaptation promoted the evolutionary diversification of tiny ocean predators

Francisco Latorre*, Ina M. Deutschmann, Aurelie Labarre, Aleix Obiol, Anders Krabberød, Eric Pelletier, Michael E. Sieracki, Corinne Cruaud, Olivier Jaillon, Ramon Massana, Ramiro Logares*

* Corresponding authors: Francisco Latorre and Ramiro Logares.
Email: latorre@icm.csic.es, ramiro.logares@icm.csic.es

This PDF file includes:

Supplementary text
Figures S1 to S3
Legends for Datasets S1 to S9
SI References

Other supplementary materials for this manuscript include the following:

Datasets S1 to S9

SI METHODS

S1. Geographic distribution of MAST-4 species and association patterns

The distribution of MAST-4 species as well as their association patterns were investigated using metabarcoding based on data from Logares et al. (1). This dataset includes surface water samples (3 m depth) from a total of 120 globally-distributed stations located in the tropical and sub-tropical ocean that were sampled as part of the *Malaspina 2010* expedition (2). Samples were obtained with a 20 L Niskin bottle deployed simultaneously to a CTD profiler that measured conductivity, temperature, oxygen, fluorescence, and turbidity. About 12 L of seawater were filtered to recover the smallest organismal size-fraction (0.2 - 3 µm; picoplankton). The concentration of inorganic nutrients (NO_3^-, NO_2^-, PO_4^{3-}, SiO_2) were included in our analyses (see Logares et al. (1) for details on their measurement).

Both the 18S (V4 region (3)) and 16S (V4-V5 region (4)) rRNA-genes were analyzed. Operational Taxonomic Units (OTUs) were delineated as Amplicon Sequence Variants (ASV) using DADA2 (5) and OTU tables were generated. Amplifications were performed with QIAGEN HotStar Taq master mix (Qigen Inc., Valencia, CA, USA). Amplicon libraries were paired-end sequenced using *Illumina* MiSeq (2 x 250 bp) at the Research and Testing Laboratory facility (see Logares et al. (1) for more details). We trimmed the 18S forward reads at 240 bp and the reverse reads at 180 bp, while for the 16S, forward reads were trimmed at 220 bp and reverse reads at 200 bp. Then, for the 18S, the maxEE was set to 7 and 8 for the forward and reverse reads respectively, while for the 16S, the maxEE was set to 2 for the forward reads and 4 for the reverse reads. OTUs were assigned taxonomy using the naïve Bayesian classifier method (6) together with the SILVA v132 database (7) as implemented in DADA2. Eukaryotic OTUs were also BLASTed against the Protist Ribosomal Reference database (PR^2, version 4.11.1 (8)). Streptophyta, Metazoa, nucleomorphs, chloroplasts, and mitochondria were removed from OTU tables.

To infer associations between OTUs we used eukaryotic and prokaryotic OTUs with total abundances >100 reads and occurrences >15% of the samples. All abundances were centered log-ratio (clr) transformed. Associations between OTUs were inferred using Maximal Information Coefficient (MIC) analyses as implemented in MICtools (9), which estimates the total information coefficient TIC_e and the maximal information coefficient MIC_e. TIC_e is used to estimate significant relationships, while their strength is calculated with MIC_e. TIC_e null distributions were estimated using 200,000 permutations and the significance level was set to 0.001 as suggested by Weiss et al. (10) $MIC_e = 0$ indicates no association between OTUs, while $MIC_e = 1$ indicates strong association. Environmentally-driven associations between OTUs were detected and removed using EnDED (11), with the methods Interaction Information and Data Processing Inequality. Furthermore, to account for data sparsity and the consequential correlations between zeros in the dataset, we removed associations between OTUs that were not present in $\geq$ 50% of the samples, i.e. less than half of the samples contained at least one of the two OTUs. We determined the Jaccard index for each association based on the presence of OTUs in the samples (intersection divided by union). We removed associations that featured a Jaccard index below 0.25. Moreover, only associations with $MIC_e > 0.4$ were considered. We used the Pearson and Spearman correlation coefficients to analyze the association type: positive Pearson or Spearman correlation coefficients point to co-occurrences, while negative values point to mutual exclusions. The distribution of OTUs across sea temperatures was explored using the *niche.val* function in the EcolUtils package (12). The abundance-weighted mean temperature was calculated for each OTU and used as an estimate of its temperature niche. We checked whether the obtained abundance-weighted mean temperature for each OTU was significantly different from chance ($p<0.05$) using a null model with 1,000 randomizations.

S2. Genome reconstruction using Single Amplified Genomes

Plankton samples were collected during the circumglobal *Tara Oceans* expedition and cryopreserved as described elsewhere (13). Individual picoplankton cells were isolated from water samples and stained with 1x SYBR Green I (Life Technologies Corporation) (14, 15) using a MoFlo (Dako Cytomation Carpinteria, CA, USA) flow cytometer equipped with the CyClone robotic arm for sorting into plates of 384 wells. Cells were lysed and their DNA denatured using cold KOH. The genome from each single cell was amplified using Multiple Displacement

Amplification (MDA) based on the Phi29 polymerase (RepliPHI™, Epicentre Biotechnologies, Madison, WI, USA) (16, 17). All single-cell work was performed at the Single Cell Genomics Center (https://scgc.bigelow.org). The obtained SAGs were taxonomically screened by PCR amplification and Sanger sequencing of the 18S rRNA gene using universal eukaryotic primers. A total of 69 SAGs affiliating to MAST-4 species A/B/C/E were selected for downstream analyses. Each selected MAST-4 SAG was sequenced in 1/8 of a lane using either *Illumina* HiSeq2000 or HiSeq4000 at either the Oregon Health & Science University (USA) or the French National Sequencing Center (Genoscope, France). A total of 424.1 Gb of sequencing data was produced, averaging 6.1 (± 0.22) Gb per SAG. For each SAG, sampling location, depth, and date are reported in **Supplementary Dataset S1**.

Each SAG was *de novo* assembled using SPAdes 3.10 (18) in single-cell mode "-sc" with default parameters. Contigs shorter than 1kbp were discarded. Quality control and general assembly statistics were computed with Quast v4.5 (19). Estimation of genome recovery was calculated with BUSCO v3 (Benchmarking Universal Single-Copy Orthologs) (20) using the Eukaryota_odb9 dataset (**Supplementary Dataset S2**). SAGs were also co-assembled to increase genome recovery. Only SAGs belonging to putatively the same species were co-assembled. Thus, SAGs had to fulfill three conditions to be co-assembled: *First*, their 18S rRNA-gene amplicon needed to be >99.5% similar. *Second*, their Average Nucleotide Identity (ANI) had to be >95%; ANI was computed using Enveomics (21) with the full-length contigs of all SAGs within each species. *Third*, SAGs had to display a homogeneous composition in Emergent Self-Organizing Maps (ESOM) (22) based on tetranucleotide frequencies. Tetranucleotide frequencies were computed using a 4 bp sliding window and 1 bp step length in fragmented contigs between 2.5 and 5 kbp in size considering both DNA strands and were subsequently clustered using ESOM. Raw data were normalized using robust estimates of mean and variance ("Robust ZT" option) and trained with the k-Batch algorithm and Euclidean grid distance. If fragments from a given SAG were mixed with those from another SAG in tetranucleotide ESOM representations, it indicated that their genomes were similar. SAGs fulfilling the previous three criteria were considered to belong to the same species and were subsequently co-assembled. Three MAST-4C SAGs (AB536_E17, AB536_F22, AB536_M21) showed more genomic divergence (ANI ~93%) compared to the others but were still included in the final co-assembly because the 18S and tetranucleotide frequencies passed the thresholds.

A total of 69 SAGs belonging to MAST-4 were co-assembled: MAST-4A (23 SAGs), MAST-4B (9 SAGs), MAST-4C (20 SAGs), and MAST-4E (17 SAGs). Prior to co-assembly, reads were digitally-normalized using BBNorm (23), considering a minimum coverage depth of 5x and a maximum target coverage depth of 100x. Normalized reads were co-assembled with SPAdes 3.10 using the single-cell mode ("--sc") running only the assembly module ("--only-assembler"). To extend contigs, they were re-scaffolded with SSPACE v3 (24). Repetitive regions were masked, along with tRNA sequences, using RepeatMasker (25) and tRNAscan-SE-1.3 (26). Quality and assembly statistics were computed with Quast (19) and are shown in **Supplementary Dataset S2**. Parameters not mentioned were set to default. Co-assembled SAGs were carefully checked for foreign DNA. Based on the premise that sequences from the same species have virtually the same tetra-nucleotide frequencies, a second tetra-nucleotide ESOM map was built for the four MAST-4 co-assemblies with the same parameters as previously described. Contigs that did not cluster together with the majority of contigs from a given SAG co-assembly were removed. Subsequently, co-assembled contigs that were classified as prokaryotic were removed based on the 5-mer profiles using EukRep (27) with mild stringency. Lastly, eukaryotic contigs with extreme GC content values, *i.e.,* values outside the range of GC content mean ± 10 % (Standard deviation) in each SAG co-assembly, were removed as well (**Supplementary Dataset S2**). Co-assembled genome completeness was estimated with BUSCO v3 (28). For each co-assembly, protein-coding genes were predicted *de novo* with AUGUSTUS 3.2.3 (29, 30) using the identified BUSCO v3 proteins as the training set. Predicted genes were functionally annotated using 1) CAZy database from dbCAN v6 (31) and HMMER 3.1b2 (32) (e-value ≤ 10^{-5}) , 2) KEGG (Release 2015-10-12; (33, 34)) and 3) eggNOG v4.5 (35) , both using BLAST 2.2.28+ and considering hits with >25% identity, >60% query coverage, <10^{-5} e-value and amino acid alignment lengths >200. Gene sequences (nucleotides) were also mapped against the Marine Atlas of Tara Oceans Unigenes (MATOU) Version 1 (20171115)(36) using BLAST 2.2.28+ with the same thresholds as the ones above used for the amino acid sequences, except for the identity threshold, which was increased to 75%, to consider nucleotide sequence variation instead of

amino acid. MAST-4 genomes were clustered in terms of their GH composition with the *hclust* function in R based on "manhattan" distances.

S3. Phylogenomics and genome differentiation

We used two approaches to analyze the phylogenetic vs. whole-genome differentiation among MAST-4 species. In the *first* approach, we randomly selected 30 conserved proteins (included in eukaryota_odb9, BUSCO v3) that were identified in all MAST-4 species (**Supplementary Dataset S3**) as well as in other publicly available Stramenopile genomes: *Phytophtora sojae* (NCBI:txid67593), *Phytophtora infestans* (NCBI:txid403677), *Schizochytrium aggregatum* (JGI:Schag1)*, Aurantiochytrium limacinum* (JGI:Aurli1) and *Cafeteria roenbergensis* (37). Genes were aligned individually with Mafft (38) using the '—auto' mode and concatenated with catfasta2phyml (39). Poorly aligned sequences and regions were removed using trimAl v1.4.rev22 (40) with "-automated1" mode and default parameters. The phylogenetic tree was built with RAxML version 8.0.0 (41) using the General Time Reversible model with a gamma-distributed rate variation among sites (GTR+G). Initial seed was "-p 666". In addition, we used the automatic bootstrap criterion (-autoMRE) and rapid Bootstrap mode (-f a). The *second* approach consisted of computing the Average Amino-acid Identity (AAI) for each pair of MAST-4 using Enveomics based on the predicted genes (amino acids). Genomes were clustered by similarity using the *pvclust* (42) package in R with "maximum" as the distance method.

S4. Abundance and expression of selected MAST-4 ERGs in the ocean

We investigated the distribution, abundance, and expression in the global ocean of selected Ecological Relevant Genes (ERGs), in this case, lysosomal enzymes (glycoside hydrolases). For that, we mapped metagenomic and metatranscriptomic reads from *Tara Oceans* (a total of 52 surface water stations encompassing the 0.8 – 5 μm size fraction (total 104 samples), the organismal size range where MAST-4 is found) against predicted genes from each MAST-4 species (**Supplementary Dataset S4**). Metatranscriptomic reads derived from sequencing polyA-enriched RNA (14, 36). The mapping was done with BWA (43) and only hits with identity > 95% and an alignment length > 80 bp were considered. Reads that mapped to more than one target were discarded. Gene abundance and expression estimates were normalized by dividing the Reads Per Kilobase (RPK) of each gene [number of mapped reads (counts) / gene length (kbp)] by the Scaling Factor (SF) [Sum of all considered RPKs in a sample / 10^6]. Hereafter, the abundance of genes and transcripts is expressed as Counts Per Million (CPM) or Transcripts Per Million (TPM) respectively. The comparison between the mean TPM values of the 20 selected MAST-4 GHs vs. the 152 single-copy housekeeping genes (from BUSCO v3's eukaryota_odb9 database) for each *TARA Oceans* station was performed using a two sample Wilcoxon test from the *matrixTests* R package (44) (**Supplementary Dataset S8**).

S5. Calculation of dN/dS ratios in homologous genes

Homologous MAST-4 genes were identified using reciprocal protein BLAST (v. 2.2.28+) with the following thresholds: >25% identity, >60% of query coverage, $<10^{-5}$ e-value, and an alignment length >200 amino acids. Gene sequences (amino acid) were aligned using Mafft 7.402 with default parameters and then converted into a codon-based nucleotide alignment with Pal2nal (45). Alignments with one or more unknown nucleotides (Ns) were discarded. For each homolog, a nucleotide-based phylogenetic tree was built using RAxML 8.2.12 (41), with the model GTR+CAT, including bootstrap analyses, and a starting seed "-p 12345" as well as the optimization "-d" parameter. Positive selection was tested on each homolog with HyPhy 2.3.14 (46) using aBSREL (branch) (47) and MEME (site) (48) models considering the codon-based nucleotide alignment and the previous phylogenetic tree. Parameters included options for universal code and testing in all branches. A *p-value* of 0.1 (default) was used for the analysis with the MEME model.

SI FIGURES

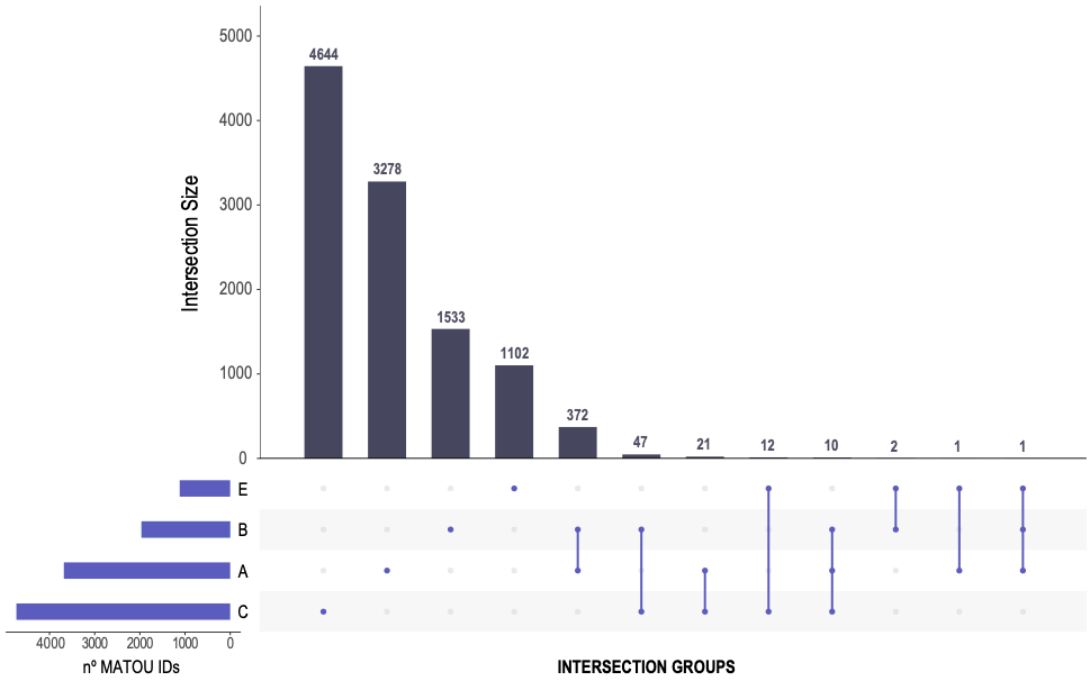

Fig. S1. Number of Unigenes (i.e. representative genes after clustering genes at 95% identity) from the MATOU database found in MAST-4 and the number of genes shared by the four species. Note that the different groups are ordered by group size and that the biggest groups are those including only one MAST-4 species, followed by the groups constituted by the combination of two or more species.

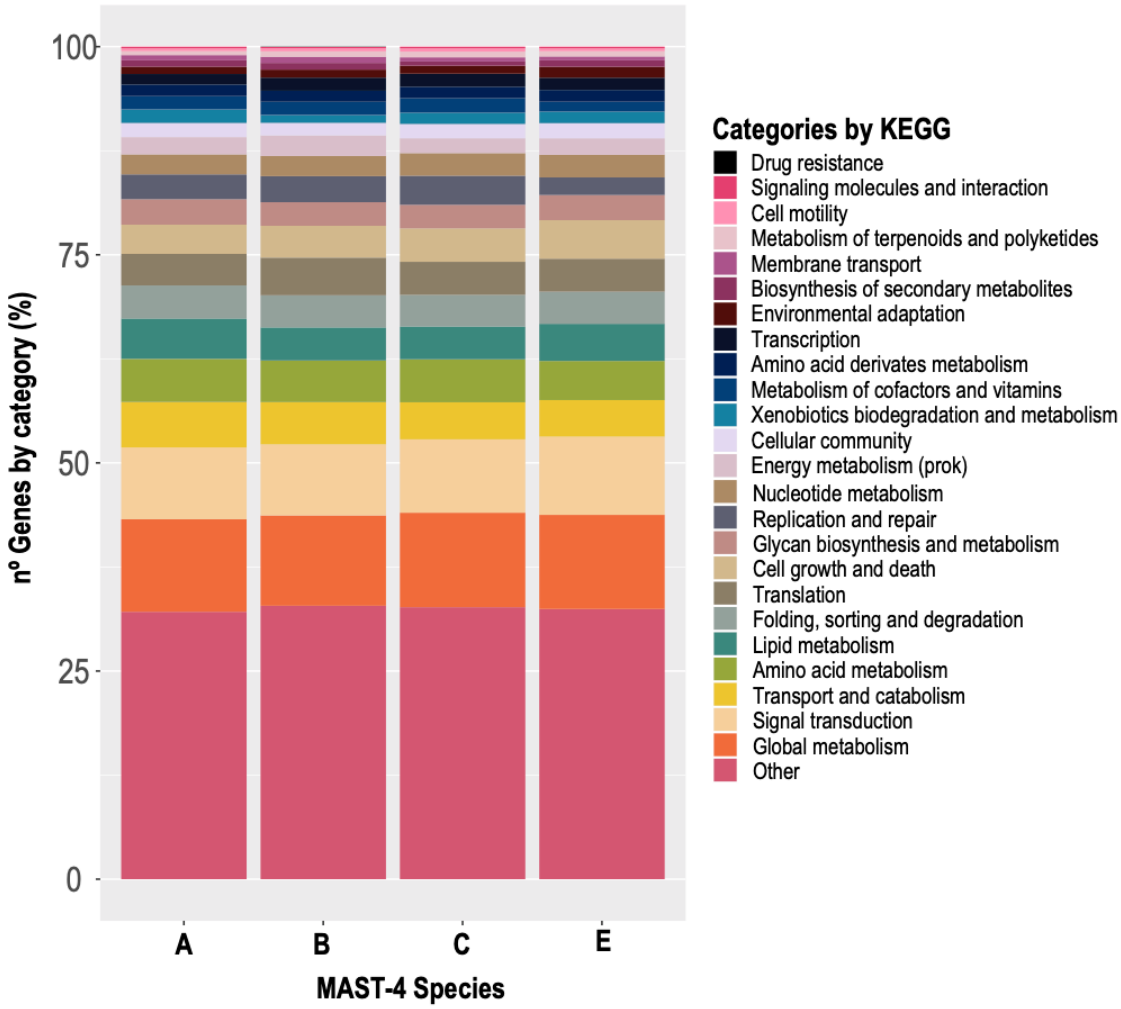

Fig. S2. Functional profile of MAST-4 genes according to KEGG. KEGG annotations are indicated as percentage of genes falling into functional categories. The category "Other" is an artificial grouping including all the annotations belonging to human related pathways such as 'Alzheimer' or 'Influenza A'.

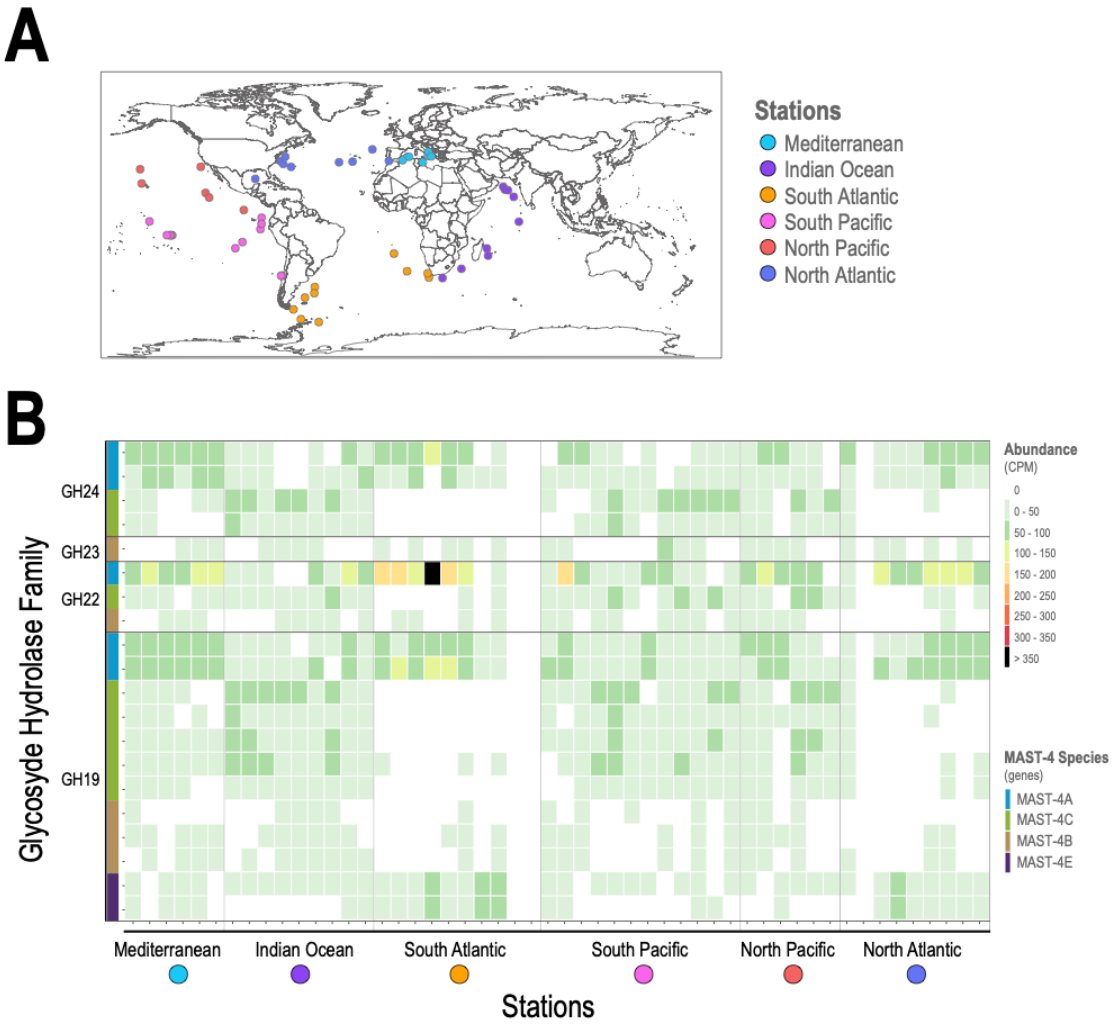

Fig. S3. Abundance of GH genes in MAST-4A/B/C/E. **Panel A)** Geographic location of metagenomic samples of *Tara Oceans*. **Panel B)** Heatmap of the Glycoside Hydrolase family abundances in MAST-4 (see their expression in Figure 5C). Samples are in the x-axis grouped by the ocean region and ordered following the expedition's trajectory. Genes in the y-axis are organized by family and each species is indicated with a color. GH22, GH23 and GH24 are families of lysozymes and GH19 is a family of chitinases that can also act as lysozymes in some organisms.

SI DATASETS

Dataset S1 (separate file). Summary of each SAG's environmental data from the *TARA Oceans* expedition. Legend: Sample Depths: D - DCM, S - SUR; Platform: GS - National Sequencing Center of Genoscope; OR - Oregon Health & Science University.

Dataset S2 (separate file). Basic assembly statistics from QUAST, BUSCO and AUGUSTUS for all the co-assemblies before and after the cleaning pipeline. Legend: Norm and Non-norm indicate whether or not the raw reads were normalized using BBNORM prior to the co-assembly.

Dataset S3 (separate file). BUSCO v3 proteins used to generate the multi-gene phylogeny of MAST-4 and from which contig they were retrieved in the co-assemblies.

Dataset S4 (separate file). Samples of metagenomes and metatranscriptomes from the *TARA Oceans* expeditions mapped against MAST-4 genes.

Dataset S5 (separate file). OTUs used in the Network Association with MICTools. Low and Upp CI_values determine the Confidence Interval, while the Sign implies the position of the observed value based on the CI: HIGHER - Observed value is > Upp CI, LOWER - Observed value is < Low CI, NON-SIGNIFICANT - Observed value is within both boundaries.

Dataset S6 (separate file). Summary of all hits of MAST-4 genes to different databases.

Dataset S7 (separate file). Summary of all GHs gene families found in MAST-4 genomes. A value of 0 indicates that no gene was annotated as part of such GH family, while a value of 1 indicates that at least one gene was found.

Dataset S8 (separate file). Expression (TPM) means for the 20 most expressed GH genes (Figure 5) and the 152 single-copy housekeeping genes (from BUSCO eukaryota_odb9) found in MAST-4 for each *Tara Oceans* station. p-values corresponding to the difference of the means for each station are indicated (Wilcoxon test).

Dataset S9 (separate file). List of all homologous genes between all four MAST-4 species. For each homolog alignment, the number of branches and sites with positive selection is given along with the function from dbCAN (CAZymes).

SI References

1. R. Logares, *et al.*, Disentangling the mechanisms shaping the surface ocean microbiota. *Microbiome* **8**, 55 (2020).
2. C. M. Duarte, Seafaring in the 21st century: The Malaspina 2010 circumnavigation expedition. *Limnol. Oceanogr. Bull.* (2015) https:/doi.org/10.1002/lob.10008.
3. T. Stoeck, *et al.*, Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* (2010) https:/doi.org/10.1111/j.1365-294X.2009.04480.x.
4. A. E. Parada, D. M. Needham, J. A. Fuhrman, Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Microbiol.* (2016) https:/doi.org/10.1111/1462-2920.13023.
5. B. J. Callahan, *et al.*, DADA2: High-resolution sample inference from Illumina amplicon data. *Nat. Methods* (2016) https:/doi.org/10.1038/nmeth.3869.
6. Q. Wang, G. M. Garrity, J. M. Tiedje, J. R. Cole, Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* (2007) https:/doi.org/10.1128/AEM.00062-07.
7. C. Quast, *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
8. L. Guillou, *et al.*, The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res.* (2013) https:/doi.org/10.1093/nar/gks1160.
9. D. Albanese, S. Riccadonna, C. Donati, P. Franceschi, A practical tool for maximal information coefficient analysis. *Gigascience* (2018)

https:/doi.org/10.1093/gigascience/giy032.
10. S. Weiss, *et al.*, Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* (2016) https:/doi.org/10.1038/ismej.2015.235.
11. I. M. Deutschmann, EnDED - - Environmentally-Driven Edge Detection Program (2019) https:/doi.org/doi.org/10.5281/zenodo.3271730.
12. G. Salazar, EcolUtils: Utilities for community ecology analysis (2019).
13. J. L. Heywood, M. E. Sieracki, W. Bellows, N. J. Poulton, R. Stepanauskas, Capturing diversity of marine heterotrophic protists: one cell at a time. *ISME J.* **5**, 674–84 (2011).
14. A. Alberti, *et al.*, Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Sci. Data* (2017) https:/doi.org/10.1038/sdata.2017.93.
15. J. F. Mangot, *et al.*, Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci. Rep.* **7** (2017).
16. M. Martinez-Garcia, *et al.*, Unveiling in situ interactions between marine protists and bacteria through single cell sequencing. *ISME J.* **6**, 703–7 (2012).
17. R. Stepanauskas, M. E. Sieracki, Matching phylogeny and metabolism in the uncultured marine bacteria, one cell at a time. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 9052–7 (2007).
18. A. Bankevich, *et al.*, SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–77 (2012).
19. A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–5 (2013).
20. R. M. Waterhouse, *et al.*, BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.* (2018) https:/doi.org/10.1093/molbev/msx319.
21. L. M. Rodriguez-R, K. T. Konstantinidis, The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes (2016) https:/doi.org/10.7287/peerj.preprints.1900v1 (January 26, 2020).
22. A. Ultsch, F. Morchen, ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM (2009) (May 15, 2016).
23. B. Bushnell, J. Rood, E. Singer, BBMerge – Accurate paired shotgun read merging via overlap. *PLoS One* (2017) https:/doi.org/10.1371/journal.pone.0185056.
24. M. Boetzer, C. V Henkel, H. J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–9 (2011).
25. A. Smit, R. Hubley, P. Green, RepeatMasker Open-4.0. 2013-2015 . *http://www.repeatmasker.org* (2013).
26. T. M. Lowe, P. P. Chan, tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* (2016) https:/doi.org/10.1093/nar/gkw413.
27. P. T. West, A. J. Probst, I. V. Grigoriev, B. C. Thomas, J. F. Banfield, Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* (2018) https:/doi.org/10.1101/gr.228429.117.
28. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
29. M. Stanke, B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, W465–W467 (2005).
30. M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–44 (2008).
31. Y. Yin, *et al.*, DbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* (2012) https:/doi.org/10.1093/nar/gks479.
32. S. R. Eddy, BIOINFORMATICS REVIEW Profile hidden Markov models. *Bioinforma. Rev.* (1998) https:/doi.org/btb114 [pii].
33. M. Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
34. M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, M. Tanabe, KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457-62 (2015).
35. J. Huerta-Cepas, *et al.*, EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* (2016) https:/doi.org/10.1093/nar/gkv1248.
36. Q. Carradec, *et al.*, A global ocean atlas of eukaryotic genes. *Nat. Commun.* (2018)

https:/doi.org/10.1038/s41467-017-02342-1.
37. T. Hackl, *et al.*, Four high-quality draft genome assemblies of the marine heterotrophic nanoflagellate Cafeteria roenbergensis. *Sci. Data* **7**, 29 (2020).
38. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* (2013) https:/doi.org/10.1093/molbev/mst010.
39. J. A. A. Nylander, catfasta2phyml.
40. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* (2009) https:/doi.org/10.1093/bioinformatics/btp348.
41. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* (2014) https:/doi.org/10.1093/bioinformatics/btu033.
42. R. Suzuki, H. Shimodaira, Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* (2006) https:/doi.org/10.1093/bioinformatics/btl117.
43. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–60 (2009).
44. K. Koncevičius, matrixTests: Fast Statistical Hypothesis Tests on Rows and Columns of Matrices (2020).
45. M. Suyama, D. Torrents, P. Bork, PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* (2006) https:/doi.org/10.1093/nar/gkl315.
46. S. L. Kosakovsky Pond, S. D. W. Frost, S. V. Muse, HyPhy: Hypothesis testing using phylogenies. *Bioinformatics* (2005) https:/doi.org/10.1093/bioinformatics/bti079.
47. M. D. Smith, *et al.*, Less is more: An adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32**, 1342–1353 (2015).
48. B. Murrell, *et al.*, Detecting individual sites subject to episodic diversifying selection. *PLoS Genet.* **8**, 1002764 (2012).