Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics

Tom O. Delmont*[1,2], Morgan Gaia[1,2], Damien D. Hinsinger[1,2], Paul Fremont[1,2], Chiara Vanni[3], Antonio Fernandez Guerra[3,4], A. Murat Eren[5,6], Artem Kourlaiev[1,2], Leo d'Agata[1,2], Quentin Clayssen[1,2], Emilie Villar[1], Karine Labadie[1,2], Corinne Cruaud[1,2], Julie Poulain[1,2], Corinne Da Silva[1,2], Marc Wessner[1,2], Benjamin Noel[1,2], Jean-Marc Aury[1,2], *Tara* Oceans Coordinators, Colomban de Vargas[2,7], Chris Bowler[2,8], Eric Karsenti[2,7,9], Eric Pelletier[1,2], Patrick Wincker[1,2] and Olivier Jaillon[1,2]

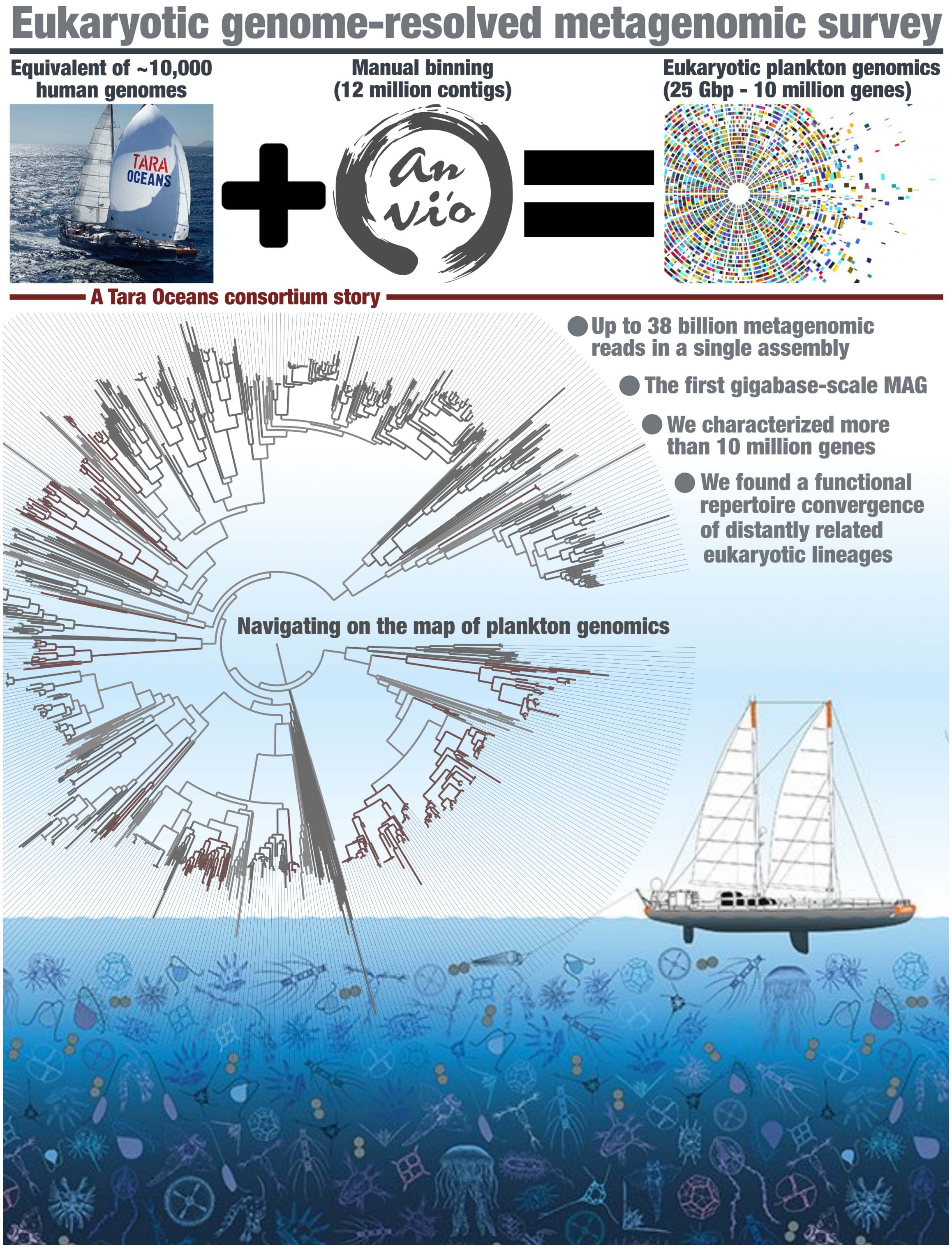
[1] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France.
[2] Research Federation for the study of Global Ocean systems ecology and evolution, FR2022/Tara GOsee, Paris, France.
[3] Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359, Bremen, Germany.
[4] Lundbeck GeoGenetics Centre, The Globe Institute, University of Copenhagen, 1350 Copenhagen, Denmark.
[5] Department of Medicine, University of Chicago, Chicago, Illinois 60637, USA
[6] Bay Paul Center, Marine Biological Laboratory, Woods Hole, Massachusetts 02543, USA
[7] Sorbonne Université and CNRS, UMR 7144 (AD2M), ECOMAP, station Biologique de Roscoff, Roscoff, France.
[8] Institut de Biologie de l'ENS, Département de Biologie, École Normale supérieure, CNRS, INSERM, Université PSL, Paris, France.
[9] Directors' research, European Molecular Biology Laboratory, Heidelberg, Germany.

*Corresponding author (Tom.Delmont@genoscope.fr)

Abstract: Marine planktonic eukaryotes play a critical role in global biogeochemical cycles and climate. However, their poor representation in culture collections limits our understanding of the evolutionary history and genomic underpinnings of planktonic ecosystems. Here, we used 280 billion *Tara* Oceans metagenomic reads from polar, temperate, and tropical sunlit oceans to reconstruct and manually curate more than 700 abundant and widespread eukaryotic environmental genomes ranging from 10 Mbp to 1.3 Gbp. This genomic resource covers a wide range of poorly characterized eukaryotic lineages that complement long-standing contributions from culture collections while better representing plankton in the upper layer of the oceans. We performed the first comprehensive genome-wide functional classification of abundant unicellular eukaryotic plankton, revealing four major groups connecting distantly related lineages. Neither trophic modes of plankton nor its vertical evolutionary history could explain the functional repertoire convergence of major eukaryotic lineages that coexisted within oceanic currents for millions of years.

Keywords: Marine eukaryotes, open ocean, plankton, genomics, metagenomics, *Tara* Oceans, anvi'o, single-cell genomics, evolution, phylogeny, functions, ecology

Cover: Navigating on the map of plankton genomics with *Tara* Oceans and anvi'o: a comprehensive genome-resolved metagenomic survey dedicated to eukaryotic plankton.



Eukaryotic genome-resolved metagenomic survey

Equivalent of ~10,000 human genomes

Manual binning (12 million contigs)

Eukaryotic plankton genomics (25 Gbp - 10 million genes)

A Tara Oceans consortium story

- Up to 38 billion metagenomic reads in a single assembly
- The first gigabase-scale MAG
- We characterized more than 10 million genes
- We found a functional repertoire convergence of distantly related eukaryotic lineages

Navigating on the map of plankton genomics

Introduction

Plankton in the sunlit ocean contributes about half of Earth's primary productivity, impacting global biogeochemical cycles and food webs[1,2]. Plankton biomass appears to be dominated by unicellular eukaryotes and small animals[3–6] including a phenomenal evolutionary and morphological biodiversity[5,7–9]. The composition of planktonic communities is highly dynamical and shaped by biotic and abiotic variables, some of which are changing abnormally fast in the Anthropocene[10–12]. Our understanding of marine eukaryotes has progressed substantially in recent years with the transcriptomic (e.g.,[13,14]) and genomic (e.g.,[15–17]) analyses of organisms isolated in culture, and the emergence of efficient culture-independent surveys (e.g.,[18,19]). However, most eukaryotic lineages' genomic content remains uncharacterized[20,21], limiting our understanding of their evolution, functioning, ecological interactions, and resilience to ongoing environmental changes.

Over the last decade, the *Tara* Oceans program has generated a homogeneous resource of marine plankton metagenomes and metatranscriptomes from the sunlit zone of all major oceans and two seas[22]. Critically, most of the sequenced plankton size fractions correspond to eukaryotic organismal sizes, providing a prime dataset to survey genomic traits and expression patterns from this domain of life. More than 100 million eukaryotic gene clusters have been characterized by the metatranscriptomes, half of which have no similarity to known proteins[5]. Most of them could not be linked to a genomic context[23], limiting their usefulness to gene-centric insights. The eukaryotic metagenomic dataset (the equivalent of ~10,000 human genomes) on the other hand has been partially used for plankton biogeographies[24,25], but remains unexploited for the characterization of genes and genomes due to a lack of robust methodologies to make sense of its diversity.

Genome-resolved metagenomics[26] has been extensively applied to the smallest *Tara* Oceans plankton size fractions, unveiling the ecology and evolution of thousands of viral, bacterial, and archaeal populations abundant in the sunlit ocean[27–32]. This approach may thus be appropriate also to characterize the genomes of the most abundant eukaryotic plankton. However, very few eukaryotic genomes have been resolved from metagenomes thus far[27,33–36], in part due to their complexity (e.g., high density of repeats[37]) and extended size[38] that might have convinced many of the unfeasibility of such a methodology. With the notable exception of some photosynthetic eukaryotes[27,33,36], metagenomics is lagging far behind cultivation for eukaryote genomics, contrasting with the two other domains of life. Here we fill this critical gap using hundreds of billions of metagenomic reads generated from the eukaryotic plankton size fractions of *Tara* Oceans and demonstrate that genome-resolved metagenomics is well suited for marine eukaryotic genomes of substantial complexity and length exceeding the emblematic gigabase. We used this new genomic resource to place major eukaryotic planktonic lineages in the tree of life and explore their evolutionary history based on both phylogenetic signals from conserved gene markers and present-day genomic functional landscape.

Results and discussion

A new resource of environmental genomes for eukaryotic plankton from the sunlit ocean

We performed the first comprehensive genome-resolved metagenomic survey of microbial eukaryotes from polar, temperate, and tropical sunlit oceans using 798 metagenomes (265 of which were released through the present study) derived from the *Tara* Oceans expeditions. They correspond to the surface and deep chlorophyll maximum layer of 143 stations from the Pacific, Atlantic, Indian, Arctic, and Southern Oceans, as well as the Mediterranean and Red Seas, encompassing eight eukaryote-enriched plankton size fractions ranging from 0.8 µm to 2 mm (Figure 1, Table S1). We used the 280 billion reads as inputs for 11 metagenomic co-assemblies (6-38 billion reads per co-assembly) using geographically bounded samples (Figure 1, Table S2), as previously done for the *Tara* Oceans 0.2–3 µm size fraction enriched in bacterial cells[27]. In addition, we used 158 eukaryotic single cells sorted by flow cytometry from seven *Tara* Oceans stations (Table S2) as input to perform complementary genomic assemblies (see Methods).

We thus created a culture-independent, non-redundant (average nucleotide identity <98%) genomic database for eukaryotic plankton in the sunlit ocean consisting of 683 metagenome-assembled genomes (MAGs) and 30 single-cell genomes (SAGs), all containing more than 10 million nucleotides (Table S3). These 713 MAGs and SAGs (hereafter dubbed SMAGs) were manually characterized and curated using a holistic framework within anvi'o[39] that relied heavily on differential coverage across metagenomes (see Methods and Supplemental Material). Nearly half the MAGs did not have vertical coverage >10x in any of the metagenomes, emphasizing the relevance of co-assemblies to gain sufficient coverage for relatively large eukaryotic genomes. Moreover, one-third of the SAGs remained undetected by *Tara* Oceans' metagenomic reads, emphasizing cell sorting's power to target less abundant lineages. Absent from the SMAGs are DNA molecules physically associated with the focal eukaryotic populations, but that did not correlate with their nuclear genomes across metagenomes. They include chloroplasts, mitochondria, and viruses generally present in multi-copy. Finally, some highly conserved multi-copy genes such as the 18S rRNA gene were also missing due to technical issues associated with assembly and binning, following the fate of 16S rRNA genes in bacterial MAGs[27].

This new genomic database for eukaryotic plankton has a total size of 25.2 Gbp and contains 10,207,450 genes according to a workflow combining metatranscriptomics, *ab-initio,* and protein-similarity approaches (see Methods). *Tara* Oceans SMAGs are, on average, ~40% complete (redundancy of 0.5%) and 35.4 Mbp long (up to 1.32 Gbp for the first Giga-scale eukaryotic MAG), with a GC-content ranging from 18.7% to 72.4% (Table S3). They are affiliated to Alveolata (n=44), Amoebozoa (n=4), Archaeplastida (n=64), Cryptista (n=31), Haptista (n=92), Opisthokonta (n=299), Rhizaria (n=2), and Stramenopiles (n=174). Only three closely related MAGs could not be affiliated to any known eukaryotic supergroup (see the phylogenetic section below). Among the 713 SMAGs, 271 contained multiple genes corresponding to chlorophyll *a-b* binding proteins and were considered phytoplankton (Table S3). Genome-wide comparisons with 484 reference transcriptomes from isolates of marine eukaryotes (the METdb database[40] which improved data from MMETSP[13] and added new transcriptomes from *Tara* Oceans, see Table S3) linked only 24 of the SMAGs (~3%) to a eukaryotic population already in culture (average nucleotide identity >98%). These include well-known Archaeplastida populations within the genera *Micromonas*, *Bathycoccus*, *Ostreococcus*, *Pycnococcus*, *Chloropicon* and *Prasinoderma* and a few taxa amongst Stramenopiles (e.g., the diatom *Minutocellus polymorphus*) and Haptista (e.g., *Phaeocystis cordata*). Thus, we found metagenomics, single-cell genomics, and culture highly complementary with very few overlaps for marine eukaryotic plankton's genomic characterization.

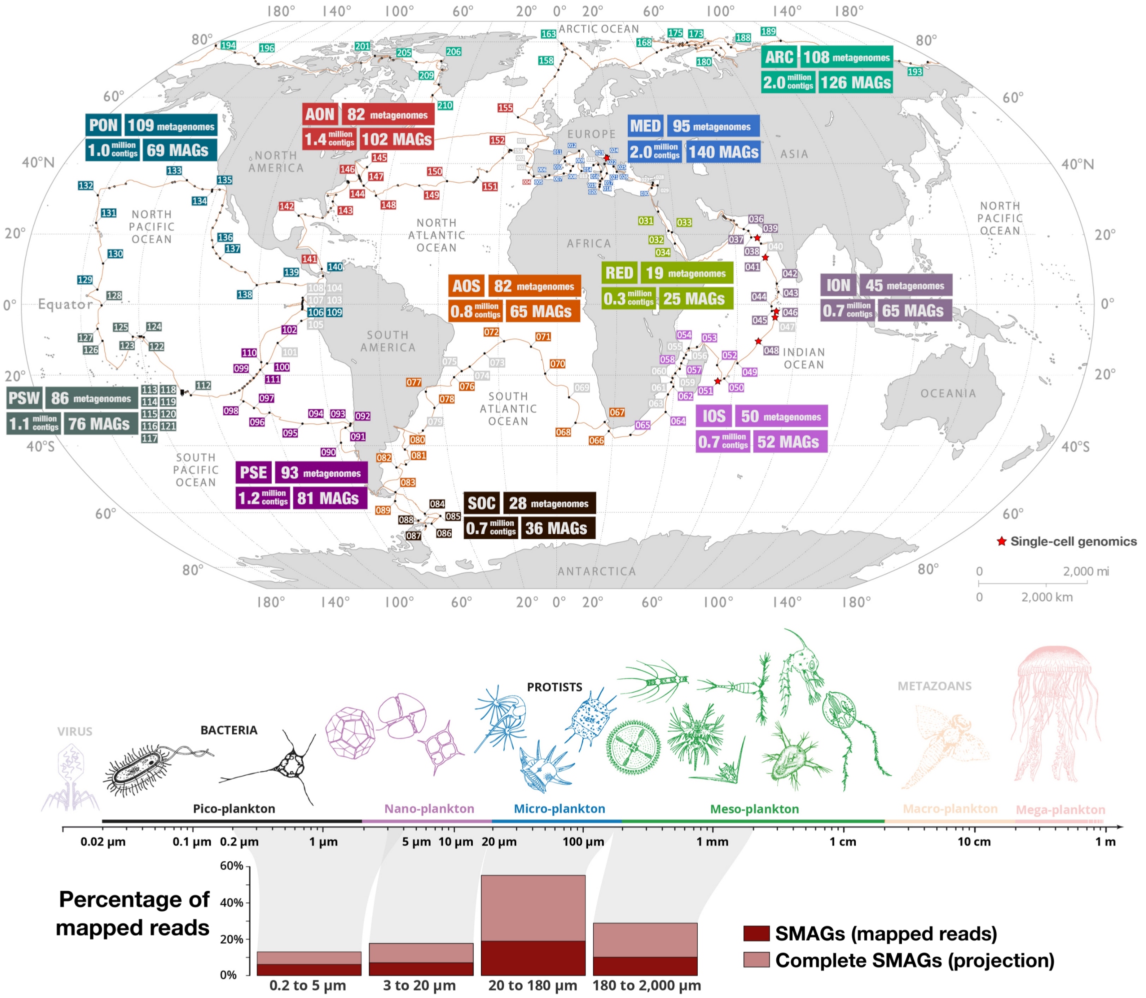

ARC 108 metagenomes 2.0 million contigs 126 MAGs
MED 95 metagenomes 2.0 million contigs 140 MAGs
RED 19 metagenomes 0.3 million contigs 25 MAGs
ION 45 metagenomes 0.7 million contigs 65 MAGs
IOS 50 metagenomes 0.7 million contigs 52 MAGs
SOC 28 metagenomes 0.7 million contigs 36 MAGs
AON 82 metagenomes 1.4 million contigs 102 MAGs
AOS 82 metagenomes 0.8 million contigs 65 MAGs
PON 109 metagenomes 1.0 million contigs 69 MAGs
PSW 86 metagenomes 1.1 million contigs 76 MAGs
PSE 93 metagenomes 1.2 million contigs 81 MAGs
ARCTIC OCEAN
NORTH AMERICA
EUROPE
ASIA
AFRICA
SOUTH AMERICA
OCEANIA
ANTARCTICA
NORTH PACIFIC OCEAN
NORTH ATLANTIC OCEAN
SOUTH ATLANTIC OCEAN
SOUTH PACIFIC OCEAN
INDIAN OCEAN
Equator
180° 140° 100° 60° 20° 0° 20° 60° 100° 140° 180°
80° 60° 40°N 20° 0° 20° 40°S 60° 80°
★ Single-cell genomics
0 2,000 mi
0 2,000 km
VIRUS
BACTERIA
PROTISTS
METAZOANS
Pico-plankton
Nano-plankton
Micro-plankton
Meso-plankton
Macro-plankton
Mega-plankton
0.02 µm 0.1 µm 0.2 µm 1 µm 5 µm 10 µm 20 µm 100 µm 1 mm 1 cm 10 cm 1 m
Percentage of mapped reads
60% 40% 20% 0%
0.2 to 5 µm
3 to 20 µm
20 to 180 µm
180 to 2,000 µm
SMAGs (mapped reads)
Complete SMAGs (projection)

Figure 1. A genome-resolved metagenomic survey dedicated to eukaryotes in the sunlit ocean. The map displays *Tara* Oceans stations used to perform genome-resolved metagenomics, summarizes the number of metagenomes, contigs longer than 2,500 nucleotides, and eukaryotic MAGs characterized from each co-assembly and outlines the stations used for single-cell genomics. ARC: Arctic Ocean; MED: Mediterranean Sea; RED: Red Sea, ION: Indian Ocean North; IOS: Indian Ocean South; SOC: Southern Ocean; AON: Atlantic Ocean North; AOS: Atlantic Ocean South; PON:

Pacific Ocean North; PSE: Pacific South East; PSW: Pacific South West. The bottom panel summarizes mapping results from the SMAGs across 939 metagenomes organized into four size fractions. The mapping projection of complete SMAGs is described in the Methods and Supplemental Material.

The SMAGs recruited 39.1 billion reads with >90% identity (average identity of 97.4%) from 939 metagenomes, representing 11.8% of the *Tara* Oceans metagenomic dataset dedicated to unicellular and multicellular organisms ranging from 0.2 µm to 2 mm (Table S4). In contrast, METdb with a total size of ~23 Gbp recruited less than 7 billion reads (average identity of 97%), indicating that the collection of *Tara* Oceans SMAGs reported herein better represents the diversity of open ocean eukaryotes as compared to genomic data from decades of culture efforts worldwide. The majority of *Tara* Oceans metagenomic reads were still not recruited, which could be explained by eukaryotic genomes that our methods failed to reconstruct, the occurrence of abundant bacterial, archaeal, and viral populations in the large size fractions we considered (e.g., *Trichodesmium*[41]), and the incompleteness of the SMAGs. Indeed, with the assumption of correct completion estimates, complete SMAGs would have recruited ~26% of all metagenomic reads, including >50% of reads for the 20-180 µm size fraction alone due in part to an important contribution of hundreds of large copepod MAGs abundant within this cellular range (see Figure 1 and Table S4).

Expanding the genomic representation of the eukaryotic tree of life

We then determined the phylogenetic distribution of the new ocean SMAGs in the tree of eukaryotic life. METdb was chosen as a taxonomically curated reference transcriptomic database from culture collections, and the two largest subunits of the three DNA-dependent RNA polymerases (six multi-kilobase genes found in all modern eukaryotes and hence already present in the Last Eukaryotic Common Ancestor) were used as evolutionary marker genes given their relevance to our understanding of eukaryogenesis[42]. Protein sequences for these genes were manually extracted and curated for the SMAGs (n=2,150) and METdb (n=2,032) (see Methods and Supplemental Material). BLAST results provided a novelty score for each of them (see Methods and Table S3), expanding our analysis scope to eukaryotic genomes stored in NCBI as of August 2020. Our final phylogenetic analysis included 416 reference transcriptomes and 576 environmental SMAGs that contained at least one of the six genes (Figure 2). The concatenated DNA-dependent RNA polymerase protein sequences effectively reconstructed a coherent tree of eukaryotic life, comparable to previous large-scale phylogenetic analyses based on other gene markers[43], and to a complementary BUSCO-centric phylogenomic analysis using protein sequences corresponding to hundreds of smaller gene markers (Figure S1). As a noticeable difference, the Haptista were most closely related to Archaeplastida, while Cryptista was most closely related to the TSAR supergroup (Telonemia not represented here, Stramenopiles, Alveolata and Rhizaria), albeit with weaker supports. This view of the eukaryotic tree of life using a previously underexploited universal marker is by no means conclusive by itself but contributes to ongoing efforts to understand deep evolutionary relationships amongst eukaryotes while providing an effective framework to assess the phylogenetic positions of a large number of the *Tara* Oceans SMAGs.

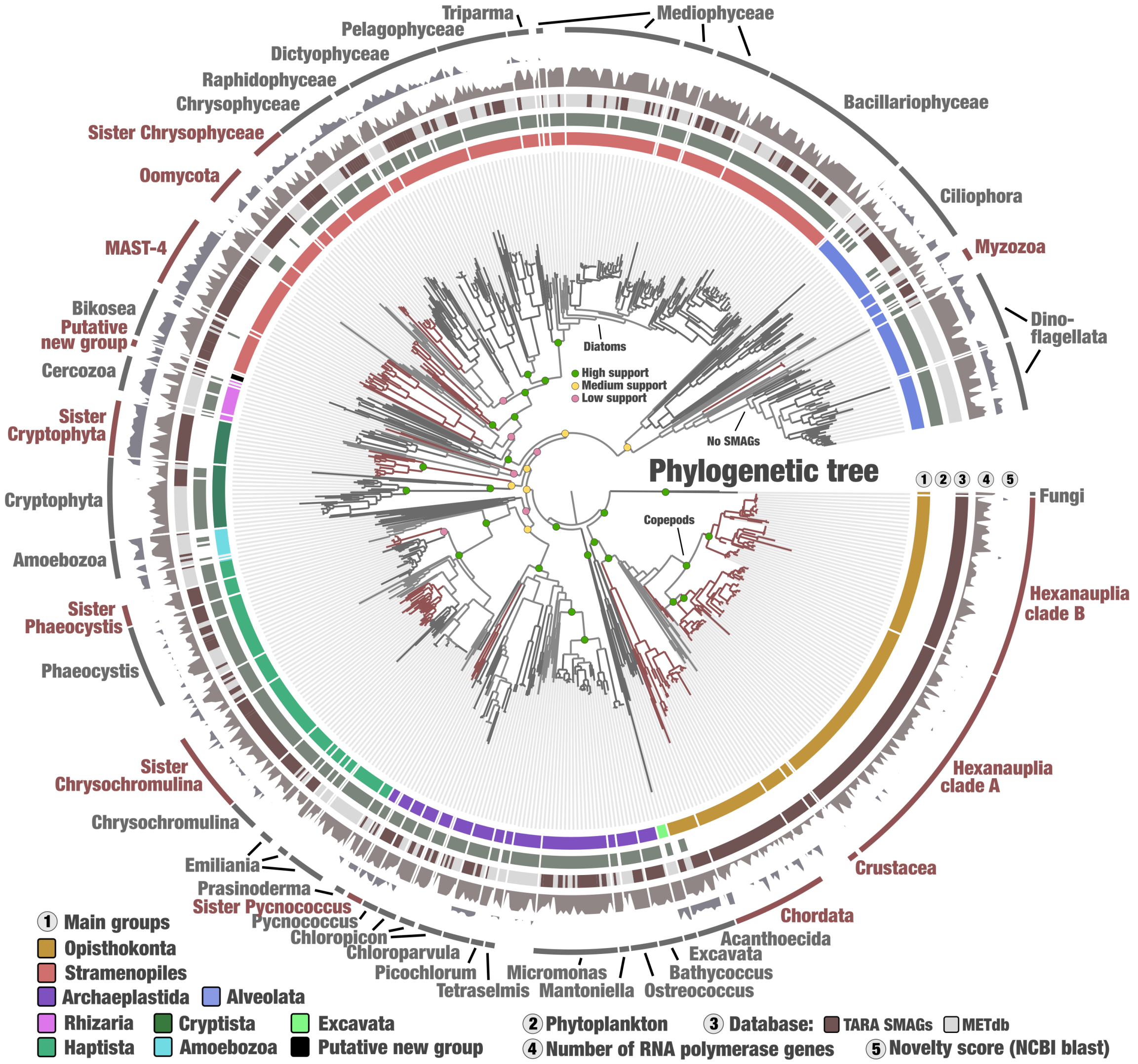


Figure 2: Phylogenetic analysis of concatenated DNA-dependent RNA polymerase protein sequences from eukaryotic plankton. The maximum-likelihood phylogenetic tree of the concatenated two largest subunits from the three DNA-dependent RNA polymerases (six genes in total) included *Tara* Oceans SMAGs and METdb transcriptomes and was generated using a total of 7,243 sites in the alignment and LG+F+R10 model; Opisthokonta was used as the outgroup. Supports for selected clades are displayed. Phylogenetic supports were considered high (aLRT>=80 and UFBoot>=95), medium (aLRT>=80 or UFBoot>=95) or low (aLRT<80 and UFBoot<95) (see Methods). The tree was decorated with additional layers using the anvi'o interface. The novelty score layer (see Methods) was set with a minimum of 30 (i.e., 70% similarity) and a maximum of 60 (i.e., 40% similarity). Branches and names in red correspond to main lineages lacking representatives in METdb.

Amongst small planktonic animals, the *Tara* Oceans SMAGs recovered one lineage of Chordata related to the Oikopleuridae family, and Crustacea including a wide range of copepods (Figure 2, Table S3). Copepods dominate large size fractions of plankton[8] and represent some of the most abundant animals on the planet[44,45]. They actively feed on unicellular plankton and are a significant food source for larger animals such as fish, thus representing a key trophic link within the global carbon cycle[46]. For now, less than ten copepod genomes have been characterized by isolates[47,48]. The additional 8.4 Gbp of genomic material unveiled herein is split into 217 MAGs, and themselves organized into two main phylogenetic clusters that we dubbed marine Hexanauplia clades A and B. The two clades were equally abundant and detected in all oceanic regions. Copepod MAGs typically had broad geographic distributions, being detected on average in 25% of the globally distributed *Tara* Oceans stations. In comparison, Opisthokonta MAGs affiliated to Chordata and Choanoflagellatea (Acanthoecida) were, on average detected in less than 10% of sampling sites.

Generally occurring in smaller size fractions, SMAGs corresponding to unicellular eukaryotes considerably expanded our genomic knowledge of known genera within Alveolata, Archaeplastida, Haptista and Stramenopiles (Table S3). Just within the diatoms for instance (Stramenopiles), MAGs were reconstructed for *Fragilariopsis* (n=5), *Pseudo-nitzschia* (n=7), *Chaetoceros* (n=11), *Thalassiosira* (n=5) and seven other genera, all of which are known to contribute significantly to photosynthesis in the sunlit ocean[49]. Beyond this genomic expansion of known planktonic genera, the SMAGs covered various lineages lacking representatives in METdb. These included (1) sister clades to the Cryptophyta division (putative Katablepharidophyta division[50] according to their relatively abundant 18S amplicons in small size fractions of *Tara* Oceans[8]), to the class Chrysophyceae, and the genera *Phaeocystis* and *Pycnococcus*, (2) basal lineages of Oomycota within Stramenopiles and Myzozoa within Alveolata, (3) multiple branches within the MAST-4 lineage, (4) and a small cluster possibly at the root of Rhizaria we dubbed “putative new group” (Figure 2). The BUSCO-centric phylogenomic analysis placed it at the root of Haptista (Figure S1), supporting its high novelty while stressing the difficulty placing it accurately in the eukaryotic tree of life. The novelty score of individual DNA-dependent RNA polymerase genes was supportive of the topology of the tree. Significantly, the sister clade to the Cryptophyta division, diverse MAST-4 lineage and putative new group all displayed a deep branching distance from cultures and a high novelty score.

The most conspicuous lineage lacking any SMAGs was the Dinoflagellata, a prominent and extremely diverse phylum in small and large eukaryotic size fractions of *Tara* Oceans[8]. These organisms harbor very large and complex genomes[51] that likely require much deeper sequencing efforts to be recovered by genome-resolved metagenomics.

A complex interplay between the evolution and functioning of marine eukaryotes

SMAGs provided a broad genomic assessment of the eukaryotic tree of life within the sunlit ocean by covering a wide range of marine plankton eukaryotes distantly related to cultures but abundant in the open ocean. Thus, the resource provided an opportunity to explore the interplay between the phylogenetic signal and functional repertoire of eukaryotic plankton with genomics. With EggNOG[52–54], we identified orthologous groups corresponding to known (n=15,870) and unknown functions (n=12,567, orthologous groups with no assigned function at http://eggnog5.embl.de/) for 4.7 million genes (nearly 50% of the genes, see Methods) and used their genomic distributions to classify the SMAGs based on their functional profiles (Table S5). Our hierarchical clustering analysis using Euclidean distance and Ward linkage (an approach to organize genomes based on pangenomic traits[55]) first split the SMAGs into small animals (Chordata, Crustacea, copepods) and putative unicellular eukaryotes (Figure 3). Fine-grained functional clusters exhibited a highly coherent taxonomy within the unicellular eukaryotes. For instance, SMAGs affiliated to the coccolithophore *Emiliana* and the sister clade to *Phaeocystis* formed distinct clusters. The sister clade to Cryptophyta was also confined to a single cluster that could be explained partly by a considerable radiation of genes related to dioxygenase activity (up to 644 genes). Most strikingly, the Archaeplastida SMAGs not only clustered with respect to their genus-level taxonomy, but the organization of these clusters was highly coherent with their evolutionary relationships (see Figure 2), confirming not only the novelty of the sister clade to *Pycnococcus*, but also the sensitivity of our framework to draw the functional landscape of unicellular marine eukaryotes.

Four major functional groups of unicellular eukaryotes emerged from the hierarchical clustering (Figure 3). Importantly, the taxonomic coherence observed in fine-grained clusters vanished when moving towards the root of these functional groups. Group A was an exception since it only covered the Haptista (including the highly cosmopolitan sister clade to *Phaeocystis*). Group B, on the other hand, encompassed a highly diverse and polyphyletic group of distantly related heterotrophic (e.g., MAST-4 and MALV) and mixotrophic (e.g., Myzozoa and Cryptophyta) lineages of various genomic size, suggesting that broad genomic functional trends may not only be explained by the trophic mode of plankton. Group C was mostly photosynthetic and covered the diatoms (Stramenopiles of various genomic size) and Archaeplastida (small genomes) as sister clusters. This finding likely reflects that diatoms are the only group with an obligatory photoautotrophic lifestyle within the Stramenopiles, like the Archaeplastida. Finally, Group D encompassed three distantly related lineages of heterotrophs (those systematically lacked gene markers for photosynthesis) exhibiting rather large genomes: Oomycota, Acanthoecida choanoflagellates, and the Cryptophyta's sister clade. Those four functional groups have similar amounts of detected functions and contained both cosmopolite and rarely detected SMAGs across the *Tara* Oceans stations. While attempts to classify marine eukaryotes based on genomic functional traits have been made in the past (e.g., using a few SAGs[56]), our resource therefore provided a broad enough spectrum of genomic material for a first genome-wide functional classification of abundant lineages of unicellular eukaryotic plankton in the upper layer of the ocean.

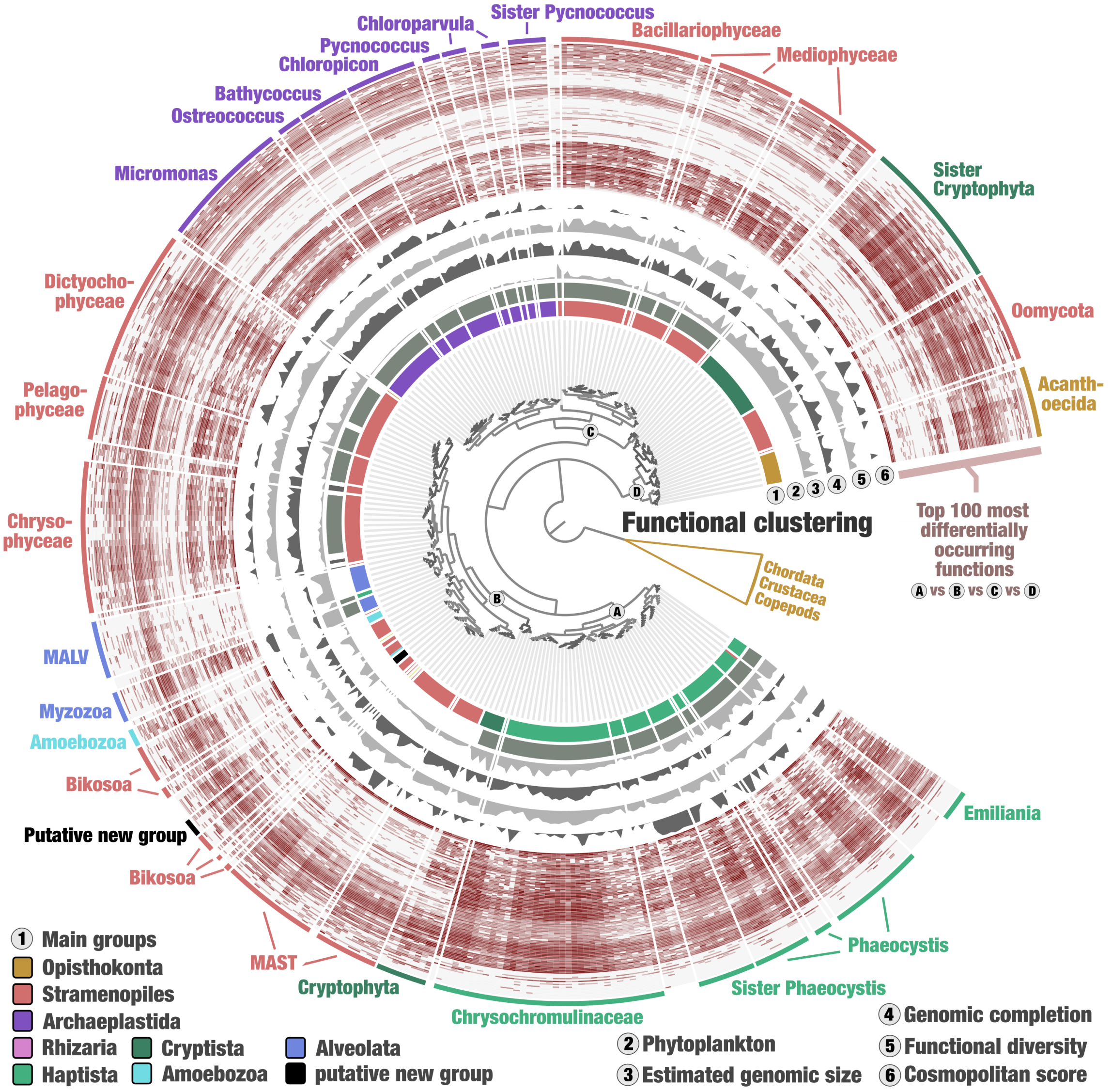


Figure 3. The genomic functional landscape of unicellular eukaryotes in the sunlit ocean. The figure displays a hierarchical clustering (Euclidean distance with Ward's linkage) of 681 SMAGs based on the occurrence of ~28,000 functions identified with EggNOG[52–54], rooted with small animals (Chordata, Crustacea and copepods) and decorated with layers of information using the anvi'o interactive interface. Layers include the occurrence in log 10 of 100 functions with lowest p-value when performing Welch's ANOVA between the functional groups A, B, C and D (see nodes in the tree). Removed from the analysis were Ciliophora MAGs (gene calling is problematic for this lineage), two less complete MAGs affiliated to Opisthokonta, and functions occurring more than 500 times in the gigabase-scale MAG and linked to retrotransposons connecting otherwise unrelated SMAGs.

A total of 2,588 known and 680 unknown functions covering 1.94 million genes (~40% of the annotated genes) were significantly differentially occurring between the four functional groups (Welch's ANOVA tests, p-value <1.e^{-05}, Table S5). We displayed the occurrence of the 100 functions with lowest p-values in the hierarchical clustering presented in Figure 3 to illustrate and help convey the strong signal between groups. However, more than 3,000 functions contributed to the basic partitioning of SMAGs. They cover all high-level functional categories identified in the 4.7 million genes with similar proportions (Figure S2), indicating that a wide range of functions related to information storage and processing, cellular processes and signaling, and metabolism contribute to the partitioning of the groups. As a notable difference, functions related to transcription (-50%) and RNA processing and modification (-47%) were less represented, while those related to carbohydrate transport and metabolism were enriched (+43%) in the differentially occurring functions. Interestingly, we noticed within Group C a scarcity of various functions otherwise occurring in high abundance among unicellular eukaryotes. These included functions related to ion channels (e.g., extracellular ligand-gated ion channel activity, intracellular chloride channel activity, magnesium ion transmembrane transporter activity, calcium ion transmembrane transport, calcium sodium antiporter activity) that may be linked to flagellar motility and the response to external stimuli[57], reflecting the lifestyle of true autotrophs. Group D, on the other hand, had significant enrichment of various functions associated with carbohydrate transport and metabolism (e.g., alpha and beta-galactosidase activities, glycosyl hydrolase families, glycogen debranching enzyme, alpha-L-fucosidase), denoting a distinct carbon acquisition strategy. Overall, the properties of thousands of differentially occurring functions suggest that eukaryotic plankton's complex functional diversity is vastly intertwined within the tree of life, as inferred from phylogenies. This reflects the complex nature of the genomic structure and phenotypic evolution of organisms, which rarely fit their evolutionary relationships.

To this point, our analysis focused on the 4.4 million genes that were functionally annotated to EggNOG, which discarded more than half of the genes we identified in the SMAGs. Our current lack of understanding of many eukaryotic functional genes even within the scope of model organisms[58] can explain the limits of reference-based approaches to study the gene content of eukaryotic plankton. Thus, to gain further insights and overcome these limitations, we partitioned and categorized the eukaryotic gene content with AGNOSTOS[59]. AGNOSTOS grouped 5.4 million genes in 424,837 groups of genes sharing remote homologies, adding 2.3 million genes left uncharacterized by the EggNOG annotation. AGNOSTOS applies a strict set of parameters for the grouping of genes discarding 575,053 genes by its quality controls and 4,264,489 genes in singletons. The integration of the EggNOG annotations into AGNOSTOS resulted in a combined dataset of 25,703 EggNOG orthologous groups (singletons and gene clusters) and 271,464 AGNOSTOS groups of genes, encompassing 6.4 million genes, 45% more genes that the original dataset (see Methods). The genome-wide functional classification of SMAGs based on this extended set of genes supported most trends previously observed with EggNOG annotation alone (Figure S3; Table S6), straightening our observations. But most interestingly, classification based solely on 23,674 newly identified groups of genes of unknown function (Table S7, a total of 1.3 million genes discarded by EggNOG) were also supportive of the overall trends, including notable links between diatoms and green algae and between sister clade to Cryptophyta and Acanthoecida (Figure S4). Thus, we identified a functional repertoire convergence of distantly related eukaryotic plankton lineages in both the known and unknown coding sequence space, the latter representing a substantial amount of biologically relevant gene diversity.

Niche and biogeography of individual eukaryotic populations

Besides insights into organismal evolution and genomic functions, the SMAGs provided an opportunity to evaluate the present and future geographical distribution of eukaryotic planktonic populations (close to species-level resolution) using the genome-wide metagenomic read recruitments. Here, we determined the niche characteristics (e.g., temperature range) of 374 SMAGs (~50% of the resource) detected in at least five stations (Table S8) and used climate models to project world map distributions (https://gigaplankton.shinyapps.io/TOENDB/) based on climatologies for the periods of 2006-2015 and 2090-2099[25] (see Methods and Supplemental Material).

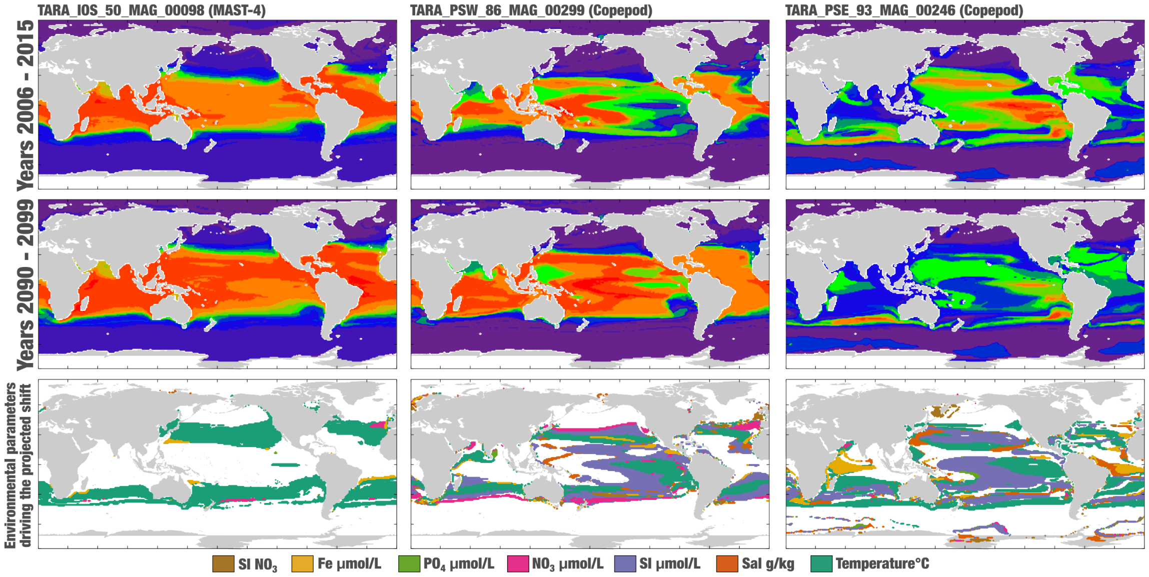

TARA_IOS_50_MAG_00098 (MAST-4)
TARA_PSW_86_MAG_00299 (Copepod)
TARA_PSE_93_MAG_00246 (Copepod)
Years 2006 - 2015
Years 2090 - 2099
Environmental parameters driving the projected shift
SI NO_3
Fe µmol/L
PO_4 µmol/L
NO_3 µmol/L
SI µmol/L
Sal g/kg
Temperature°C

Figure 4. World map distribution projections for three eukaryotic MAGs during the periods of 2006-2015 and 2090-2099. The probability of presence ranges from 0 (purple) to 1 (red), with green corresponding to a probability of 0.5. The bottom row displays first-rank region-dependent environmental parameters driving the projected shifts of distribution (in regions where |ΔP|>0.1). Noticeably, projected decreases of silicate in equatorial regions drive 34% of the expansion of TARA_PSW_MAG_00299 while driving 34% of the reduction of TARA_PSE_93_MAG_00246, possibly reflecting different life strategies of these copepods (e.g., grazing). In contrast, the expansion of TARA_IOS_50_MAG_00098 is mostly driven by temperature (74%).

Each of these SMAGs was estimated to occur in a surface averaging 42 and 39 million km^2 for the first and second period, respectively, corresponding to ~12% of the surface of the ocean. Our data suggest that most eukaryotic populations in the database will remain widespread for decades to come. However, many changes in biogeography are projected to occur. For instance, the most widespread population in the first period (a MAST-4 MAG) would still be ranked first at the end of the century but with a surface area increasing from 37% to 46% (Figure 4), a gain of 28 million km^2 corresponding to the surface of North America. Its expansion from the tropics towards more temperate oceanic regions regardless of longitude is mostly explained by temperature and reflects the expansion of tropical niches due to global warming, echoing recent predictions made with amplicon surveys and imaging data[60]. As an extreme case, the SMAG benefiting the most between the two periods (a copepod) could experience a gain of 55 million km^2 (Figure 4), more than the surface of Asia and Europe combined. On the other hand, the SMAG losing most ground (also a copepod) could undergo a decrease of 47 million km^2. Projected changes in these two examples correlated with various variables (including a notable contribution of silicate), an important reminder that temperature alone cannot explain plankton's biogeography in the ocean. Our integration of genomics, metagenomics, and climate models provided the resolution needed to project individual eukaryotic population niche trajectories in the sunlit ocean.

Conclusion

Following methodological advances for viral, bacterial and archaeal lineages, we are experiencing a shift from cultivation to metagenomics for the genomic characterization of marine eukaryotes *en masse*. Our culture-independent and manually curated genomic characterization of unicellular eukaryotic populations and small animals abundant in the sunlit ocean covered a wide range of poorly characterized lineages and provided the first gigabase-scale metagenome-assembled genome, a landmark for both genome-resolved metagenomics and plankton genomics. These lineages cover multiple trophic levels (e.g., copepods and their prey, mixotrophs, autotrophs, and parasites) and appear to be abundant and widespread in the sunlit ocean. In summary, most eukaryotic genomes we characterized with different degrees of completion are not only different from past genomic surveys of isolated marine organisms but also better represent eukaryotic plankton in the open photic ocean. As a result, our survey represents an innovative step towards using genomics to explore in concert the ecological and evolutionary underpinnings of environmentally relevant eukaryotic organisms, using metagenomics to fill critical gaps in our remarkable culture porfolio[22].

Phylogenetic gene markers such as the DNA-dependent RNA polymerases (the basis of our phylogenetic analysis) provide a critical understanding of the origin of eukaryotic lineages and allowed us to place most environmental genomes in a comprehensible evolutionary framework. However, this framework is based on sequence variations within core genes that in theory are inherited from the last eukaryotic common ancestor representing the vertical evolution of eukaryotes, disconnected from the structure of genomes. As such, it does not recapitulate the functional evolutionary journey of plankton, as demonstrated in our genome-wide functional classification of unicellular eukaryotes in both the known and unknown coding sequence space. The dichotomy between phylogeny and function was already well described with morphological and other phenotypic traits and could be explained in part by secondary endosymbiosis events that have spread plastids and genes for their photosynthetic capabilities across the eukaryotic tree of life[61–64]. Here we moved beyond morphological inferences and disentangled the phylogeny of gene markers and broad genomic functional repertoire of a comprehensive collection of marine eukaryotic lineages. We identified four major genomic functional groups of unicellular eukaryotes made of distantly related lineages. The Stramenopiles proved particularly effective in terms of genomic functional diversification, possibly explaining part of their remarkable success in this biome[8,65].

The topology of phylogenetic trees compared to the functional clustering of a wide range of eukaryotic lineages has revealed contrasting evolutionary journeys for widely scrutinized gene markers of evolution and less studied genomic functions of plankton. The apparent functional convergence of distantly related lineages that coexisted in the same biome for millions of years could not be explained by neither a vertical evolutionary history of unicellular eukaryotes nor their trophic modes (phytoplankton versus heterotrophs), shedding new lights into the complex functional dynamics of plankton over evolutionary time scales. Convergent evolution is a well-known phenomenon of independent origin of biological traits such as molecules and behaviors[66,67] that has been observed in the morphology of microbial eukaryotes[68] and is often driven by common selective pressures within similar environmental conditions. However, an independent origin of similar functional profiles is not the only possible explanation for organisms sharing the same habitat. Indeed, one could wonder if lateral gene transfers between eukaryotes[69,70] have played a central role in these processes, as previously observed between eukaryotic plant pathogens[71] or grasses[72]. As a case in point, secondary endosymbiosis events are known to have resulted in massive gene transfers between endosymbionts and their hosts in the oceans[61,62]. In particular, these events involved transfers of genes from green algae to diatoms[73], two lineages clustering together in our genomic functional classification of eukaryotic plankton. However, lineages sharing the same secondary endosymbiotic history did not always fall in the same functional group. This was the case for diatoms, Haptista and Cryptista that have different functional trends yet originate from a common ancestor that likely acquired its plastid from red and green algae[61,62,74]. Surveying phylogenetic trends for functions derived from the ~10 million genes identified here will likely contribute to new insights regarding the extent of lateral gene transfers between eukaryotes[75,76], the independent emergence of functional traits (convergent evolution), as well as functional losses between lineages[77], that altogether might have driven the functional convergences of distantly related eukaryotic lineages abundant in the sunlit ocean.

Regardless of the mechanisms involved, the functional repertoire convergences we observed likely highlight primary organismal functioning, which have fundamental impacts on plankton ecology, and their functions within marine ecosystems and biogeochemical cycles. Thus, the apparent dichotomy between phylogenies (a vertical evolutionary framework) and genome-wide functional repertoires (genome structure evolution) depicted here should be viewed as a fundamental attribute of marine unicellular eukaryotes that we suggest warrants a new rationale for studying the structure and state of plankton, a rationale also based on present-day genomic functions rather than phylogenetic and morphological surveys alone.

STAR Methods

***Tara* Oceans metagenomes.** We analyzed a total of 943 *Tara Oceans* metagenomes available at the EBI under project PRJEB402 (https://www.ebi.ac.uk/ena/browser/view/PRJEB402). 265 of these metagenomes have been released through this study. Table S1 reports accession numbers and additional information (including the number of reads and environmental metadata) for each metagenome.

Genome-resolved metagenomics. We organized the 798 metagenomes corresponding to size fractions ranging from 0.8 µm to 2 mm into 11 'metagenomic sets' based upon their geographic coordinates. We used those 0.28 trillion reads as inputs for 11 metagenomic co-assemblies using MEGAHIT[78] v1.1.1, and simplified the scaffold header names in the resulting assembly outputs using anvi'o[39] v.6.1 (available from http://merenlab.org/software/anvio). Co-assemblies yielded 78 million scaffolds longer than 1,000 nucleotides for a total volume of 150.7 Gbp. We performed a combination of automatic and manual binning on each co-assembly output, focusing only on the 11.9 million scaffolds longer than 2,500 nucleotides, which resulted in 837 manually curated eukaryotic metagenome-assembled genomes (MAGs) longer than 10 million nucleotides. Briefly, (1) anvi'o profiled the scaffolds using Prodigal[79] v2.6.3 with default parameters to identify an initial set of genes, and HMMER[80] v3.1b2 to detect genes matching to 83 single-copy core gene markers from BUSCO[81] (benchmarking is described in a dedicated blog post[82]), (2) we used a customized database including both NCBI's NT database and METdb to infer the taxonomy of genes with a Last Common Ancestor strategy[5] (results were imported as described in http://merenlab.org/2016/06/18/importing-taxonomy), (3) we mapped short reads from the metagenomic set to the scaffolds using BWA v0.7.15[83] (minimum identity of 95%) and stored the recruited reads as BAM files using samtools[84], (4) anvi'o profiled each BAM file to estimate the coverage and detection statistics of each scaffold, and combined mapping profiles into a merged profile database for each metagenomic set. We then clustered scaffolds with the automatic binning algorithm CONCOCT[85] by constraining the number of clusters per metagenomic set to a number ranging from 50 to 400 depending on the set. Each CONCOCT clusters (n=2,550, ~12 million scaffolds) was manually binned using the anvi'o interactive interface. The interface considers the sequence composition, differential coverage, GC-content, and taxonomic signal of each scaffold. Finally, we individually refined each eukaryotic MAG >10 Mbp as outlined in Delmont and Eren[86], and renamed scaffolds they contained according to their MAG ID. Table S2 reports the genomic features (including completion and redundancy values) of the eukaryotic MAGs. The supplemental material provides more information regarding this workflow and describes examples for CONCOCT clusters' binning and curation.

A first Giga scale eukaryotic MAG. We performed targeted genome-resolved metagenomics to confirm the biological relevance and improve statistics of the single MAG longer than 1 Gbp with an additional co-assembly (five Southern Ocean metagenomes for which this MAG had vertical coverage >1x) and by considering contigs longer than 1,000 nucleotides, leading to a gain of 181,8 million nucleotides. To our knowledge, we describe here the first successful characterization of a Gigabase-scale MAG (1.32 Gbp with 419,520 scaffolds), which we could identify using two distinct metagenomic co-assemblies.

MAGs from the 0.2–3 µm size fraction. We incorporated into our database 20 eukaryotic MAGs longer than 10 million nucleotides previously characterized from the 0.2–3 µm size fraction[27], providing a set of MAGs corresponding to eukaryotic cells ranging from 0.2 µm (picoeukaryotes) to 2 mm (small animals).

Single-cell genomics: We used 158 eukaryotic single cells sorted by flow cytometry from seven *Tara* Oceans stations as input to perform genomic assemblies (up to 18 cells with identical 18S rRNA genes per assembly to optimize completion statistics, see Supplementary Table 2), providing 34 single-cell genomes (SAGs) longer than 10 million nucleotides. Cell sorting, DNA amplification, sequencing and assembly were performed as described elsewhere[19]. In addition, manual curation was performed using sequence composition and differential coverage across 100 metagenomes in which the SAGs were most detected, following the methodology described in the genome-resolved metagenomics section. For SAGs with no detection in *Tara* Oceans metagenomes, only sequence composition and taxonomical signal could be used, limiting this curation effort's scope. Notably, manual curation of SAGs using the genome-resolved metagenomic workflow turned out to be highly valuable, leading to the removal of more than one hundred thousand scaffolds for a total volume of 193.1 million nucleotides. This metagenomic-guided decontamination effort contributes to previous efforts characterizing eukaryotic SAGs from the same cell sorting material[19,56,87–89] and provides new marine eukaryotic guidelines SAGs. The supplemental material provides more information regarding this workflow and describes an example for the curation of SAGs using metagenomics.

Characterization of a non-redundant database of SMAGs. We determined the average nucleotide identity (ANI) of each pair of SMAGs using the dnadiff tool from the MUMmer package[90] v.4.0b2. SMAGs were considered redundant when their ANI was >98% (minimum alignment of >25% of the smaller SMAG in each comparison). We then selected the longest SMAG to represent a group of redundant SMAGs. This analysis provided a non-redundant genomic database of 713 SMAGs.

Taxonomical inference of SMAGs. We manually determined the taxonomy of SMAGs using a combination of approaches: (1) taxonomical signal from the initial gene calling (Prodigal), (2) phylogenetic approaches using the RNA polymerase and METdb, (3) ANI within the SMAGs and between SMAGs and METdb, (4) local blasts using BUSCO gene markers, (5) and lastly the functional clustering of SMAGs to gain knowledge into very few SMAGs lacking gene markers and ANI signal.

Protein coding genes. Protein coding genes for the SMAGs were characterized using three complementary approaches: protein alignments using reference databases, metatranscriptomic mapping from *Tara* Oceans and *ab-initio* gene predictions. While the overall framework was highly similar for MAGs and SAGs, the methodology slightly differed to take the best advantage of those two databases when they were processed (see the two following sections).

Protein-coding genes for the MAGs. Protein alignments: Since the alignment of a large protein database on all the MAG assemblies is time greedy, we first detected the potential proteins of Uniref90 + METdb that could be aligned to the assembly by using MetaEuk[91] with default parameters. This subset of proteins was aligned using BLAT with default parameters, which localized each protein on the MAG assembly. The exon/intron structure was refined using genewise[92] with default parameters to detect splice sites accurately. Each MAG's GeneWise alignments were converted into a standard GFF file and given as input to gmove. **Metatranscriptomic mapping from *Tara* Oceans:** A total of 905 individual *Tara* Oceans metatranscriptomic assemblies (mostly from large planktonic size fractions) were aligned on each MAG assembly using Minimap2[93] (version 2.15-r905) with the "-ax splice" flag. BAM files were filtered as follows: low complexity alignments were removed and only alignments covering at least 80% of a given metatranscriptomic contig with at least 95% of identity were retained. The BAM files were converted into a standard GFF file and given as input to gmove. ***Ab-initio* gene predictions:** A first gene prediction for each MAG was performed using gmove and the GFF file generated from metatranscriptomic alignments. From these preliminary gene models, 300 gene models with a start and a stop codon were randomly selected and used to train AUGUSTUS[94] (version 3.3.3). A second time, AUGUSTUS was launched on each MAG assembly using the dedicated calibration file, and output files were converted into standard GFF files and given as input to gmove. Each individual line of evidence was used as input for gmove (http://www.genoscope.cns.fr/externe/gmove/) with default parameters to generate the final protein-coding genes annotations.

Protein coding genes for the SAGs. Protein alignments: The Uniref90 + METdb database of proteins was aligned using BLAT[95] with default parameters, which localized protein on each SAG assembly. The exon/intron structure was refined using GeneWise[92] and default parameters to detect splice sites accurately. The GeneWise alignments of each SAG were converted into a standard GFF file and given as input to gmove. **Metatranscriptomic mapping from *Tara* Oceans:** The 905 *Tara* Oceans metatranscriptomic individual fastq files were filtered with kfir (http://www.genoscope.cns.fr/kfir) using a k-mer approach to select only reads that shared 25-mer with the input SAG assembly. This subset of reads was aligned on the corresponding SAG assembly using STAR[96] (version 2.5.2.b) with default parameters. BAM files were filtered as follows: low complexity alignments were removed and only alignments covering at least 80% of the metatranscriptomic reads with at least 90% of identity were retained. Candidate introns and exons were extracted from the BAM files and given as input to gmorse[97]. ***Ab-initio* gene predictions:** *Ab-initio* models were predicted using SNAP[98] (v2013-02-16) trained on complete protein matches and gmorse models, and output files were converted into standard GFF files and given as input to gmove. Each line of evidence was used as input for gmove (http://www.genoscope.cns.fr/externe/gmove/) with default parameters to generate the final protein-coding genes annotations.

BUSCO completion scores for protein-coding genes in SMAGs. BUSCO[81] v.3.0.4 with the set of eukaryotic single-copy core gene markers (n=255). Completion and redundancy (number of duplicated gene markers) of SMAGs were computed from this analysis.

Biogeography of SMAGs. We performed a final mapping of all metagenomes to calculate the mean coverage and detection of the SMAGs (Table S4). Briefly, we used BWA v0.7.15 (minimum identity of 90%) and a FASTA file containing the 713 non-redundant SMAGs to recruit short reads from all 943 metagenomes. We considered SMAGs were detected in a given filter when >25% of their length was covered by reads to minimize non-specific read recruitments[27]. The number of recruited reads below this cut-off was set to 0 before determining vertical coverage and percent of recruited reads. Regarding the projection of mapped reads, if SMAGs were to be complete, we used BUSCO completion scores to project the number of mapped reads. Note that we preserved the actual number of mapped reads for the SMAGs with completion <10% to avoid substantial errors to be made in the projections.

Identifying the environmental niche of SMAGs. Seven physicochemical parameters were used to define environmental niches: sea surface temperature (SST), salinity (Sal), dissolved silica (Si), nitrate (NO_3), phosphate (PO_4), iron (Fe), and a seasonality index of nitrate (SI NO_3). Except for Fe and SI NO3, these parameters were extracted from the gridded World Ocean Atlas 2013 (WOA13)[99]. Climatological Fe fields were provided by the biogeochemical model PISCES-v2[100]. The seasonality index of nitrate was defined as the range of nitrate concentration in one grid cell divided by the maximum range encountered in WOA13 at the Tara sampling stations. All parameters were co-located with the corresponding stations and extracted at the month corresponding to the Tara sampling. To compensate for missing physicochemical samples in the Tara *in situ* data set, climatological data (WOA) were favored. More details are available in the supplemental material.

Cosmopolitan score. Using metagenomes from the Station subset 1 (n=757), SMAGs were assigned a “cosmopolitan score” based on their detection across 119 stations (see the supplemental material for more details).

A database of manually curated DNA-dependent RNA polymerase genes. A eukaryotic dataset[101] was used to build HMM profiles for the two largest subunits of the DNA-dependent RNA polymerase (RNAP-a and RNAP-b). These two HMM profiles were incorporated within the anvi’o framework to identify RNAP-a and

RNAP-b genes (Prodigal[79] annotation) in the SMAGs and METdb transcriptomes. Alignments, phylogenetic trees and blast results were used to organize and manually curate those genes. Finally, we removed sequences shorter than 200 amino-acids, providing a final collection of DNA-dependent RNA polymerase genes for the SMAGs (n=2,150) and METdb (n=2,032) with no duplicates (see the supplemental material for more details).

Novelty score for the DNA-dependent RNA polymerase genes. We compared both the RNA-Pol A and RNA-Pol B peptides sequences identified in SMAGs and MetDB to the nr database (retrieved on October 25, 2019) using blastp, as implemented in blast+[102] v.2.10.0 (e-value of $1e^{-10}$). We kept the best hit and considered it as the closest sequence present in the public database. For each SMAG, we computed the average percent identity across RNA polymerase genes (up to six genes) and defined the novelty score by subtracting this number from 100. For example, with an average percent identity is 64%, the novelty score would be 36%.

Phylogenetic analyses of SMAGs. The protein sequences included for the phylogenetic analyses (either the **DNA-dependent RNA polymerase genes** we recovered manually or the **BUSCO set of 255 eukaryotic single-copy core gene markers** we recovered automatically from the ~10 million protein coding genes) were aligned with MAFFT[103] v.764 and the FFT-NS-i algorithm with default parameters. Sites with more than 50% of gaps were trimmed using Goalign v0.3.0-alpha5 (http://www.github.com/evolbioinfo/goalign). The phylogenetic trees were reconstructed with IQ-TREE[104] v1.6.12, and the model of evolution was estimated with the ModelFinder[105] Plus option: for the concatenated tree, the LG+F+R10 model was selected. Supports were computed from 1,000 replicates for the Shimodaira-Hasegawa (SH)-like approximation likelihood ratio (aLRT)[106] and ultrafast bootstrap approximation (UFBoot)[107]. As per IQ-TREE manual, we deemed the supports good when SH-aLRT >= 80% and UFBoot >= 95%. Anvi'o v.6.1 was used to visualize and root the phylogenetic trees.

EggNOG functional inference of SMAGs. We performed the functional annotation of protein-coding genes using the EggNog-mapper[53,54] v2.0.0 and the EggNog5 database[52]. We used Diamond[108] v0.9.25 to align proteins to the database. We refined the functional annotations by selecting the orthologous group within the lowest taxonomic level predicted by EggNog-mapper.

Eukaryotic SMAGs integration in the AGNOSTOS-DB. We used the AGNOSTOS workflow to integrate the protein coding genes predicted from the SMAG into a variant of the AGNOSTOS-DB that contains 1,829 metagenomes from the marine and human microbiomes, 28,941 archaeal and bacterial genomes from the Genome Taxonomy Database (GTDB) and 3,243 nucleocytoplasmic large DNA viruses (NCLDV) metagenome assembled genomes (MAGs)[59].

AGNOSTOS functional aggregation inference. AGNOSTOS partitioned protein coding genes from the SMAGs in groups connected by remote homologies, and categorized those groups as members of the known or unknown coding sequence space based on the workflow described in Vanni et al. 2020[59]. To combine the results from AGNOSTOS and the EggNOG classification we identified those groups of genes in the known space that contain genes annotated with an EggNOG and we inferred a consensus annotation using a quorum majority voting approach. AGNOSTOS produces groups of genes with low functional entropy in terms of EggNOG annotations as shown in Vanni et al. 2020[59] allowing us to combine both sources of information. We merged the groups of genes that shared the same consensus EggNOG annotations and we integrated them with the rest of AGNOSTOS groups of genes, mostly representing the unknown coding sequence space. Finally, we excluded groups of genes occurring in less than 2% of the SMAGs.

Differential occurrence of functions. We performed a Welch's ANOVA test followed by a Games-Howell test for significant ANOVA comparisons to identify EggNog functions occurring differentially between functional groups of SMAGs. All statistics were generated in R 3.5.3.

Functional clustering of SMAGs. We used anvi'o to cluster SMAGs as a function of their functional profile (Euclidean distance with ward's linkage), and the anvi'o interactive interface to visualize the hierarchical clustering in the context of complementary information.

Data availability. All data our study generated are publicly available at http://www.genoscope.cns.fr/tara/. The link provides access to the 11 raw metagenomic co-assemblies, the FASTA files for 713 SMAGs, the ~10 million protein-coding sequences (nucleotides, amino acids and gff format), and the curated DNA-dependent RNA polymerase genes (SMAGs and METdb transcriptomes). This link also provides access to the supplemental figures and the supplemental material.

Contributions

Damien D. Hinsinger, Morgan Gaia, Eric Pelletier, Patrick Wincker, Olivier Jaillon and Tom O. Delmont conducted the study. Tom O. Delmont and Morgan Gaia characterized the SMAGs and RNA polymerase genes, respectively. Damien D. Hinsinger (analysis of the ~10 million genes), Morgan Gaia (phylogenies), Paul Fremont (climate models and world map projections), Eric Pelletier (METdb database, mapping results) and Tom O. Delmont performed the primary analysis of the data. Artem Kourlaiev, Leo d'Agata, Quentin Clayssen and Jean-Marc Aury assembled and annotated the single cell genomes and helped processing metagenomic assemblies. Emilie Villar, Marc Wessner, Benjamin Noel, Corinne Da Silva, Damien D. Hinsinger, Olivier Jaillon and Jean-Marc Aury identified the eukaryotic genes in the MAG assemblies. Antonio Fernandez Guerra and Chiara Vanni characterized the repertoire of functions in the unknown coding sequence space. Tom O. Delmont wrote the manuscript, with critical inputs from the authors.

Acknowledgments

Our survey was made possible by two scientific endeavors: the sampling and sequencing efforts by the *Tara* Oceans Project, and the bioinformatics and visualization capabilities afforded by anvi'o. We are indebted to all who contributed to these efforts, as well as other open-source bioinformatics tools for their commitment to transparency and openness. *Tara* Oceans (which includes the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions) would not exist without the leadership of the *Tara* Ocean Foundation and the continuous support of 23 institutes (https://oceans.taraexpeditions.org/). We thank the commitment of the following people and sponsors who made this singular expedition possible: CNRS (in particular Groupement de Recherche GDR3280 and the Research Federation for the Study of Global Ocean Systems Ecology and Evolution FR2022/Tara GOSEE), the European Molecular Biology Laboratory (EMBL), Genoscope/CEA, the French Ministry of Research and the French Governement 'Investissement d'Avenir' programs Oceanomics (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09), ATIGE Genopole postdoctoral fellowship, HYDROGEN/ANR-14-CE23-0001, MEMO LIFE (ANR-10-LABX-54), PSL Research University (ANR-11-IDEX-0001-02) and EMBRC-France (ANR-10-INBS-02), Fund for Scientific Research—Flanders, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, ANR (projects ALGALVIRUS ANR-17-CE02- 0012, PHYTBACK/ANR-2010-1709-01, POSEIDON/ANR-09-BLAN-0348, PROMETHEUS/ANR-09-PCS-GENM-217, TARA-GIRUS/ANR-09-PCS-GENM-218), EU FP7 (MicroB3/No. 287589, IHMS/HEALTH-F4-2010-261376), Genopole, CEA DRF Impulsion program. The authors also thank agnès b. and E. Bourgois, the Prince Albert II de Monaco Foundation, the Veolia Foundation, the EDF Foundation EDF Diversiterre, Region Bretagne, Lorient Agglomeration, Worldcourier, Illumina, the EDF Foundation EDF Diversiterre, for support and commitment. The global sampling effort was made possible by countless scientists and crew who performed sampling aboard the Tara from 2009 to 2013. The authors are also grateful to the countries that graciously granted sampling permission. Part of the computations were performed using the platine, titane and curie HPC machine provided through GENCI grants (t2011076389, t2012076389, t2013036389, t2014036389, t2015036389 and t2016036389). We also thank Noan Le Bescot (TernogDesign) for artwork on Figures.

This article is contribution number XX of Tara Oceans.

Supplemental material

--- Available at http://www.genoscope.cns.fr/tara/ --

Supplemental figures

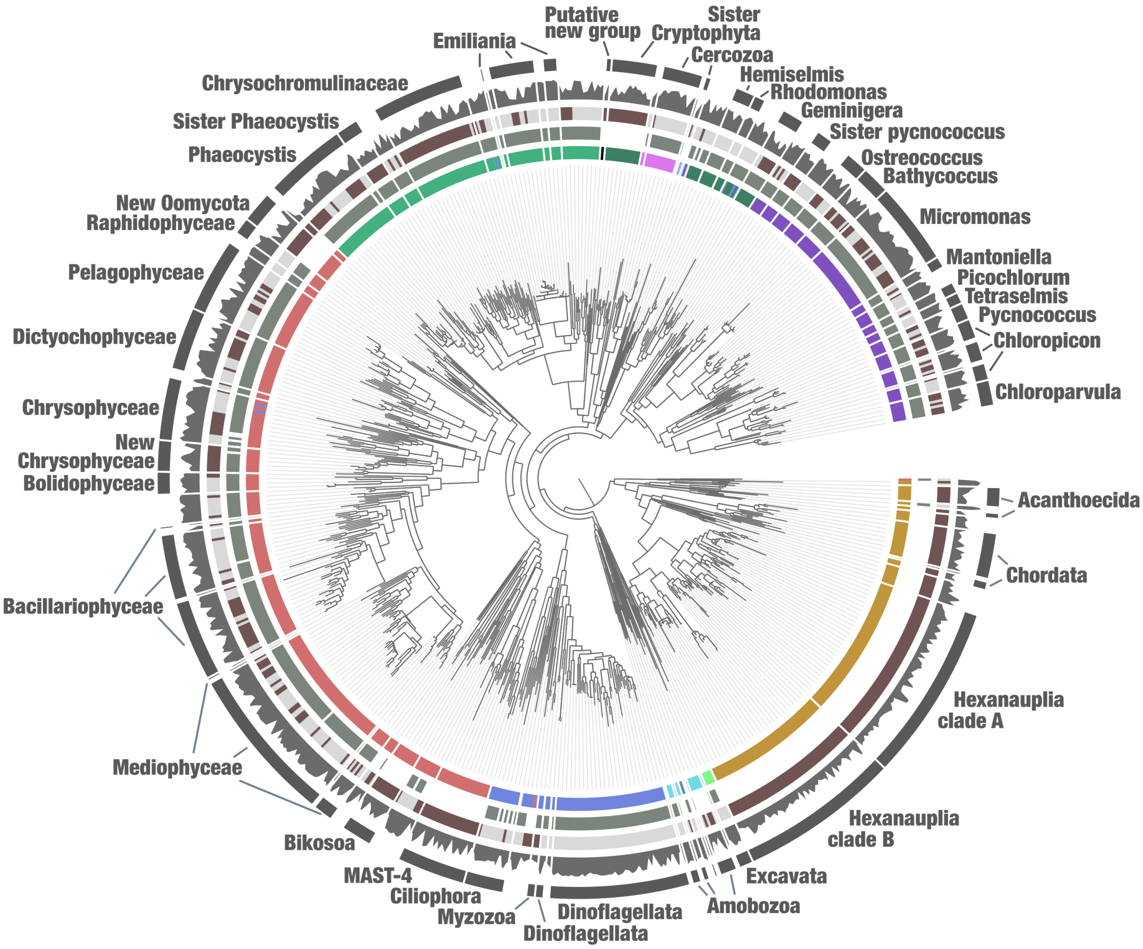


Figure S1. Phylogenomic analysis of the protein sequences of 255 BUSCO genes markers from eukaryotic plankton. The maximum-likelihood phylogenomic tree of the BUSCO gene markers (255 genes) included *Tara* Oceans MAGs and METdb transcriptomes (minimum of 25% of completion) and was generated using a total of 19,785 sites in the alignment and LG+F+R10 model; Opisthokonta was used as the outgroup. The tree was decorated with additional layers using the anvi'o interface. Branches and names in red correspond to lineages lacking representatives in METdb.

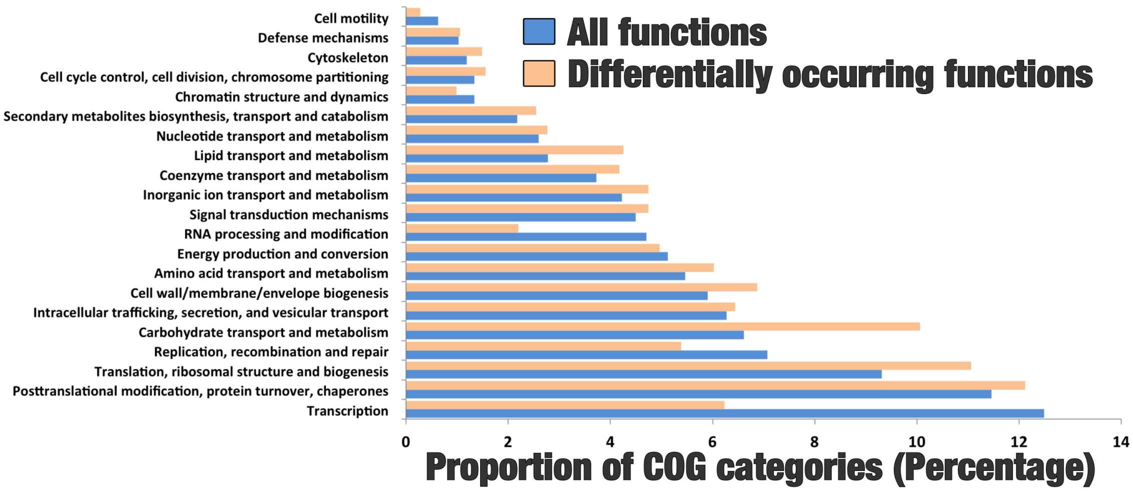


Figure S2. Relative proportion of known COG categories in annotated functions versus those that were significantly differentially occurring between the four functional groups.

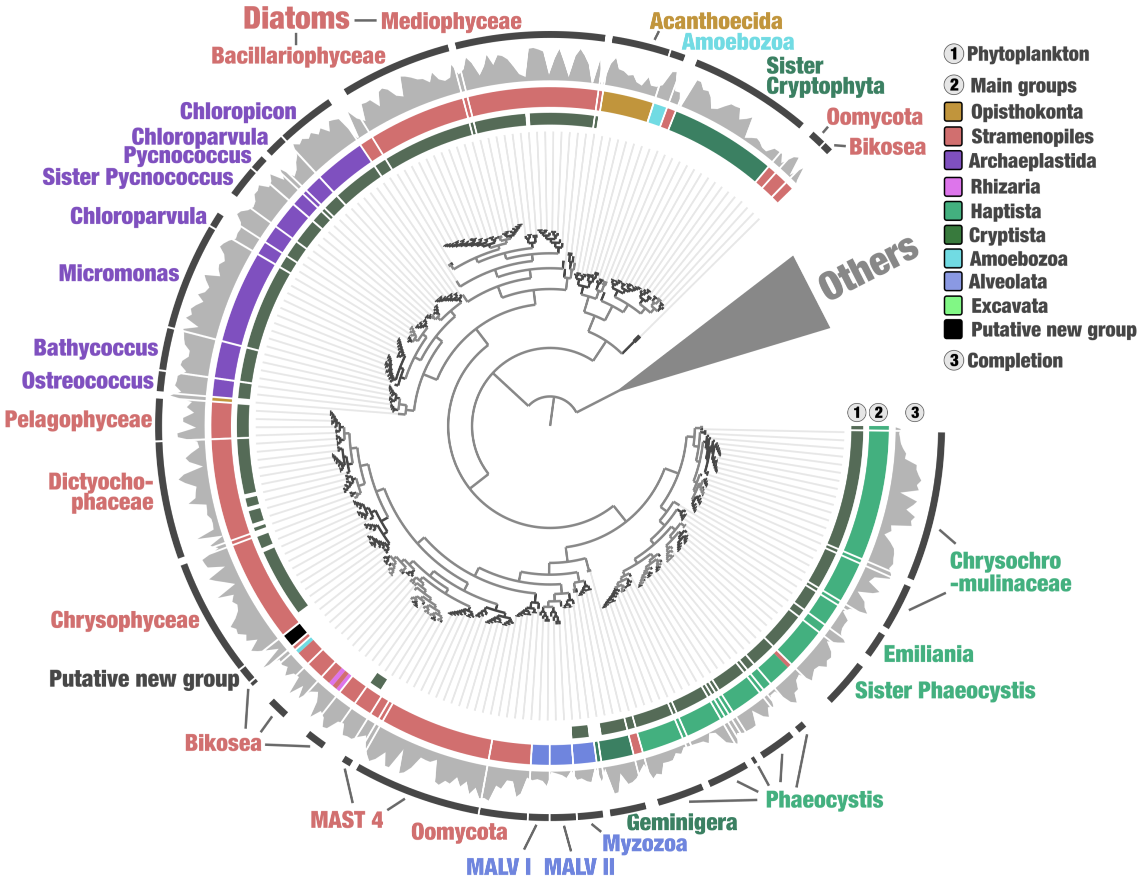


Figure S3. Functional landscape of unicellular eukaryotes in the sunlit ocean by combining EggNOG and Agnostos for gene processing. The figure displays a hierarchical clustering (Euclidean distance with Ward's linkage) of 681 SMAGs based on the occurrence of ~39,705 groups of genes (total of 5,178,829 genes) identified by combining EggNOG[52–54] with Agnostos[109], rooted with MAGs dominated by small animals (Chordata, Crustacea and copepods) and decorated with layers of information using the anvi'o interactive interface. Removed from the analysis were Ciliophora MAGs (gene calling is problematic for this lineage), and functions occurring more than 1,000 times in the

gigabase-scale MAG and linked to retrotransposons connecting otherwise unrelated SMAGs, or occurring in less than 2% of the SMAGs.

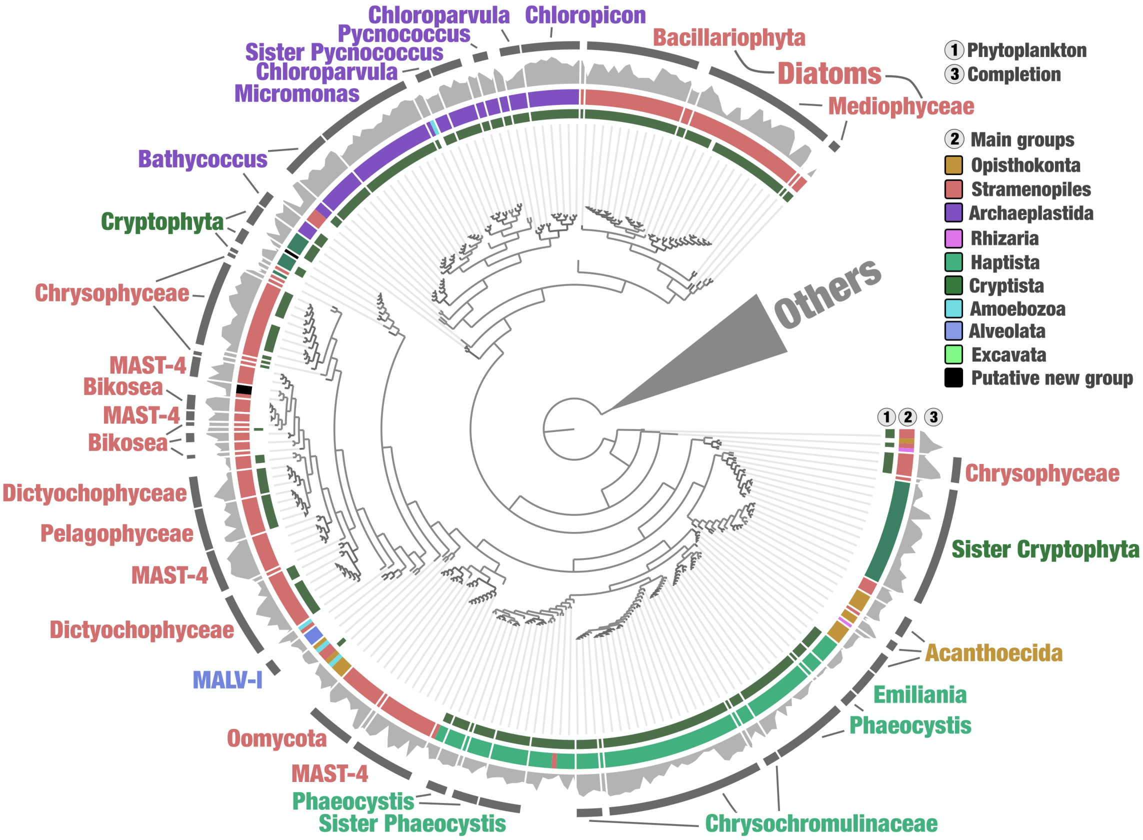


Figure S4. The genomic unknown functional landscape of unicellular eukaryotes in the sunlit ocean. The figure displays a hierarchical clustering (Euclidean distance with Ward's linkage) of 681 SMAGs based on the occurrence of ~28,000 gene clusters of unknown function (total of 1.3 million genes) identified by solely with Agnostos[109] (environmental unknowns plus genomic unknowns), rooted with MAGs dominated by small animals (Chordata, Crustacea and copepods) and decorated with layers of information using the anvi'o interactive interface. Removed from the analysis were Ciliophora MAGs (gene calling is problematic for this lineage), and functions occurring more than 1,000 times in the gigabase-scale MAG and linked to retrotransposons connecting otherwise unrelated SMAGs, or occurring in less than 2% of the SMAGs.

Supplemental tables

--- Available at http://www.genoscope.cns.fr/tara/ ---

Table S1. Summary of 939 Tara Oceans metagenomes that include their station ID, size fraction and number of quality filtered reads.

Table S2. Summary of the genome-resolved metagenomics and single cell genomics outcomes. The table includes statistics for the metagenomic co-assemblies, redundant MAGs and SAGs, and targeted efforts regarding the one giga scale MAG.

Table S3. Statistics of non-redundant SMAGs and METdb transcriptomes. The table includes genomic statistics and taxonomical inferences, the occurrence of RNA polymerase genes, and distribution patterns across stations and size fractions.

Table S4. Mapping result for the non-redundant SMAGs and METdb transcriptomes.

Table S5. Functional profiling of the non-redundant SMAGs based on EggNOG.

Table S6. Functional profiling of the non-redundant SMAGs based on EggNOG and Agnostos.

Table S7. Functional profiling of the non-redundant SMAGs solely based on Agnostos genomic and environmental unknowns not covered by EggNOG.

Table S8. Niche partitioning and world map projection statistics for 374 SMAGs

References

1. Sanders, R. *et al.* The Biological Carbon Pump in the North Atlantic. *Prog. Oceanogr.* (2014). doi:10.1016/j.pocean.2014.05.005
2. Boyd, P. W. Toward quantifying the response of the oceans' biological pump to climate change. *Front. Mar. Sci.* (2015). doi:10.3389/fmars.2015.00077
3. Dortch, Q. & Packard, T. T. Differences in biomass structure between oligotrophic and eutrophic marine ecosystems. *Deep Sea Res. Part A, Oceanogr. Res. Pap.* (1989). doi:10.1016/0198-0149(89)90135-0
4. Gasol, J. M., Del Giorgio, P. A. & Duarte, C. M. Biomass distribution in marine planktonic communities. *Limnol. Oceanogr.* (1997). doi:10.4319/lo.1997.42.6.1353
5. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nat. Commun.* (2018). doi:10.1038/s41467-017-02342-1
6. Caron, D. A., Countway, P. D., Jones, A. C., Kim, D. Y. & Schnetzer, A. Marine Protistan Diversity. *Ann. Rev. Mar. Sci.* (2012). doi:10.1146/annurev-marine-120709-142802
7. Leray, M. & Knowlton, N. Censusing marine eukaryotic diversity in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2016). doi:10.1098/rstb.2015.0331
8. De Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science (80-.).* (2015). doi:10.1126/science.1261605
9. Sieracki, M. E. *et al.* Optical Plankton Imaging and Analysis Systems for Ocean Observation. in (2010). doi:10.5270/oceanobs09.cwp.81
10. Jonkers, L., Hillebrand, H. & Kucera, M. Global change drives modern plankton communities away from the pre-industrial state. *Nature* (2019). doi:10.1038/s41586-019-1230-3
11. Hays, G. C., Richardson, A. J. & Robinson, C. Climate change and marine plankton. *Trends in Ecology and Evolution* (2005).

doi:10.1016/j.tree.2005.03.004

12. Hutchins, D. A. & Fu, F. Microorganisms and ocean global change. *Nat. Microbiol.* (2017). doi:10.1038/nmicrobiol.2017.58
13. Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.* (2014). doi:10.1371/journal.pbio.1001889
14. Johnson, L. K., Alexander, H. & Brown, C. T. Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *bioRxiv* (2018). doi:10.1101/323576
15. Palenik, B. *et al.* The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. U. S. A.* (2007). doi:10.1073/pnas.0611046104
16. Bowler, C. *et al.* The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* (2008). doi:10.1038/nature07410
17. Worden, A. Z. *et al.* Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes micromonas. *Science (80-.).* (2009). doi:10.1126/science.1167222
18. Sunagawa, S. *et al.* Ocean plankton. Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
19. Sieracki, M. E. *et al.* Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Sci. Rep.* (2019). doi:10.1038/s41598-019-42487-1
20. Sibbald, S. J. & Archibald, J. M. More protist genomes needed. *Nature Ecology and Evolution* (2017). doi:10.1038/s41559-017-0145
21. Del Campo, J. *et al.* The others: Our biased perspective of eukaryotic genomes. *Trends in Ecology and Evolution* (2014). doi:10.1016/j.tree.2014.03.006
22. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* 1–18 (2020). doi:10.1038/s41579-020-0364-5
23. Vorobev, A. *et al.* Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *Genome Res.* (2020). doi:10.1101/gr.253070.119
24. Richter, D. *et al.* Genomic evidence for global ocean plankton biogeography shaped by large-scale current systems. *bioRxiv doi:https://doi.org/10.1101/867739* **23**, 31 (2019).
25. Frémont, P. *et al.* Restructuring of genomic provinces of surface ocean plankton under climate change 2. *bioRxiv* 2020.10.20.347237 (2020). doi:10.1101/2020.10.20.347237
26. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43 (2004).
27. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* 1 (2018). doi:10.1038/s41564-018-0176-9
28. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **5**, (2018).

29. Parks, D. H. *et al.* Author Correction: Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* 1 (2017). doi:10.1038/s41564-017-0083-5
30. Tully, B. J. Metabolic diversity within the globally abundant Marine Group II Euryarchaea offers insight into ecological patterns. *Nat. Commun.* (2019). doi:10.1038/s41467-018-07840-4
31. Gregory, A. C. *et al.* Marine DNA Viral Macro- and Microdiversity from Pole to Pole. *Cell* (2019). doi:10.1016/j.cell.2019.03.040
32. Moniruzzaman, M., Martinez-Gutierrez, C. A., Weinheimer, A. R. & Aylward, F. O. Dynamic genome evolution and complex virocell metabolism of globally-distributed giant viruses. *Nat. Commun.* (2020). doi:10.1038/s41467-020-15507-2
33. Delmont, T. O., Murat Eren, A., Vineis, J. H. & Post, A. F. Genome reconstructions indicate the partitioning of ecological functions inside a phytoplankton bloom in the Amundsen Sea, Antarctica. *Front. Microbiol.* **6**, (2015).
34. Olm, M. R. *et al.* Genome-resolved metagenomics of eukaryotic populations during early colonization of premature infants and in hospital rooms. *Microbiome* (2019). doi:10.1186/s40168-019-0638-1
35. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* (2018). doi:10.1101/gr.228429.117
36. Duncan, A. *et al.* Metagenome-assembled genomes of phytoplankton communities across the Arctic Circle. *bioRxiv Microbiol.* (2020). doi:10.1101/2020.06.16.154583
37. Biscotti, M. A., Olmo, E. & Heslop-Harrison, J. S. (Pat. Repetitive DNA in eukaryotic genomes. *Chromosom. Res.* (2015). doi:10.1007/s10577-015-9499-z
38. Gregory, T. R. Synergy between sequence and size in large-scale genomics. *Nature Reviews Genetics* (2005). doi:10.1038/nrg1674
39. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
40. Niang, G. *et al.* METdb, an extended reference resource for Marine Eukaryote Transcriptomes. *http://metdb.sb-roscoff.fr/metdb/ (unpublished)*
41. Capone, D. G. Trichodesmium, a Globally Significant Marine Cyanobacterium. *Science (80-.).* **276**, 1221–1229 (1997).
42. Guglielmini, J., Woo, A. C., Krupovic, M., Forterre, P. & Gaia, M. Diversification of giant and large eukaryotic dsDNA viruses predated the origin of modern eukaryotes. *Proc. Natl. Acad. Sci.* (2019). doi:10.1073/pnas.1912006116
43. Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. The New Tree of Eukaryotes. *Trends in Ecology and Evolution* (2020). doi:10.1016/j.tree.2019.08.008
44. Humes, A. G. How many copepods? *Hydrobiologia* (1994). doi:10.1007/BF00229916
45. Kiørboe, T. What makes pelagic copepods so successful? *Journal of Plankton Research* (2011). doi:10.1093/plankt/fbq159

46. Steinberg, D. K. & Landry, M. R. Zooplankton and the Ocean Carbon Cycle. *Ann. Rev. Mar. Sci.* (2017). doi:10.1146/annurev-marine-010814-015924
47. Jørgensen, T. S. *et al.* The whole genome sequence and mRNA transcriptome of the tropical cyclopoid copepod Apocyclops royi. *G3 Genes, Genomes, Genet.* (2019). doi:10.1534/g3.119.400085
48. Jørgensen, T. S. *et al.* The genome and mRNA transcriptome of the cosmopolitan calanoid copepod acartia tonsa dana improve the understanding of copepod genome size evolution. *Genome Biol. Evol.* (2019). doi:10.1093/gbe/evz067
49. Malviya, S. *et al.* Insights into global diatom distribution and diversity in the world's ocean. *Proc. Natl. Acad. Sci. U. S. A.* (2016). doi:10.1073/pnas.1509523113
50. Okamoto, N. & Inouye, I. The katablepharids are a distant sister group of the Cryptophyta: A proposal for Katablepharidophyta divisio nova/Kathablepharida phylum novum based on SSU rDNA and beta-tubulin phylogeny. *Protist* (2005). doi:10.1016/j.protis.2004.12.003
51. Song, B., Chen, S. & Chen, W. Dinoflagellates, a unique lineage for retrogene research. *Frontiers in Microbiology* (2018). doi:10.3389/fmicb.2018.01556
52. Huerta-Cepas, J. *et al.* EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gky1085
53. Jensen, L. J. *et al.* eggNOG: Automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res.* **36**, D250–D254 (2008).
54. Huerta-Cepas, J. *et al.* Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol. Biol. Evol.* (2017). doi:10.1093/molbev/msx148
55. Delmont, T. O. & Eren, A. M. Linking pangenomes and metagenomes: The Prochlorococcus metapangenome. *PeerJ* **2018**, (2018).
56. Seeleuthner, Y. *et al.* Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nat. Commun.* (2018). doi:10.1038/s41467-017-02235-3
57. Hill, K. *et al.* A Ca2+- and voltage-modulated flagellar ion channel is a component of the mechanoshock response in the unicellular green alga Spermatozopsis similis. *Biochim. Biophys. Acta - Biomembr.* (2000). doi:10.1016/S0005-2736(00)00200-5
58. Wood, V. *et al.* Hidden in plain sight: what remains to be discovered in the eukaryotic proteome? *Open Biol.* **9**, 180241 (2019).
59. Vanni, C. *et al.* Unifying the global coding sequence space enables the study of genes with unknown function across biomes. *bioRxiv* 2020.06.30.180448 (2020). doi:10.1101/2020.06.30.180448
60. Ibarbalz, F. M. *et al.* Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell* (2019). doi:10.1016/j.cell.2019.10.008
61. Archibald, J. M. & Keeling, P. J. Recycled plastids: A 'green movement' in eukaryotic evolution. *Trends in Genetics* (2002). doi:10.1016/S0168-9525(02)02777-4
62. Deschamps, P. & Moreira, D. Reevaluating the green contribution to diatom

genomes. *Genome Biol. Evol.* (2012). doi:10.1093/gbe/evs053

63. Keeling, P. J. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annual Review of Plant Biology* (2013). doi:10.1146/annurev-arplant-050312-120144
64. Reyes-Prieto, A., Weber, A. P. M. & Bhattacharya, D. The origin and establishment of the plastid in algae and plants. *Annual Review of Genetics* (2007). doi:10.1146/annurev.genet.41.110306.130134
65. Derelle, R., López-García, P., Timpano, H. & Moreira, D. A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). *Mol. Biol. Evol.* (2016). doi:10.1093/molbev/msw168
66. Emery, N. J. & Clayton, N. S. The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science* (2004). doi:10.1126/science.1098410
67. Zakon, H. H. Convergent evolution on the molecular level. in *Brain, Behavior and Evolution* (2002). doi:10.1159/000063562
68. Leander, B. S. A hierarchical view of convergent evolution in microbial eukaryotes. in *Journal of Eukaryotic Microbiology* (2008). doi:10.1111/j.1550-7408.2008.00308.x
69. Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* (2008). doi:10.1038/nrg2386
70. Danchin, E. G. J. Lateral gene transfer in eukaryotes: Tip of the iceberg or of the ice cube. *BMC Biology* (2016). doi:10.1186/s12915-016-0330-x
71. Andersson, J. O. Convergent Evolution: Gene Sharing by Eukaryotic Plant Pathogens. *Current Biology* (2006). doi:10.1016/j.cub.2006.08.042
72. Dunning, L. T. *et al.* Lateral transfers of large DNA fragments spread functional genes among grasses. *Proc. Natl. Acad. Sci. U. S. A.* (2019). doi:10.1073/pnas.1810031116
73. Chan, C. X., Bhattacharya, D. & Reyes-Prieto, A. Endosymbiotic and horizontal gene transfer in microbial eukaryotes. *Mob. Genet. Elements* (2012). doi:10.4161/mge.20110
74. Dorrell, R. G. *et al.* Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *Elife* (2017). doi:10.7554/eLife.23717
75. Leger, M. M., Eme, L., Stairs, C. W. & Roger, A. J. Demystifying Eukaryote Lateral Gene Transfer. *BioEssays* (2018). doi:10.1002/bies.201700242
76. Martin, W. F. Too Much Eukaryote LGT. *BioEssays* (2017). doi:10.1002/bies.201700115
77. Zmasek, C. M. & Godzik, A. Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. *Genome Biol.* (2011). doi:10.1186/gb-2011-12-1-r4
78. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2014).
79. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
80. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

81. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* (2015). doi:10.1093/bioinformatics/btv351
82. Delmont, T. O. Assessing the completion of eukaryotic bins with anvi'o. *Blog post* (2018). Available at: http://merenlab.org/2018/05/05/eukaryotic-single-copy-core-genes/.
83. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
84. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
85. Alneberg, J. *et al.* Binning metagenomic contigs by coverage and composition. *Nat. Methods* **11**, 1144–1146 (2014).
86. Delmont, T. O. & Eren, A. M. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ* **4**, e1839 (2016).
87. Mangot, J. F. *et al.* Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci. Rep.* (2017). doi:10.1038/srep41498
88. López-Escardó, D. *et al.* Evaluation of single-cell genomics to address evolutionary questions using three SAGs of the choanoflagellate Monosiga brevicollis. *Sci. Rep.* (2017). doi:10.1038/s41598-017-11466-9
89. Vannier, T. *et al.* Survey of the green picoalga Bathycoccus genomes in the global ocean. *Sci. Rep.* (2016). doi:10.1038/srep37900
90. Delcher, A. L., Phillippy, A., Carlton, J. & Salzberg, S. L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res.* **30**, 2478–2483 (2002).
91. Levy Karin, E., Mirdita, M. & Söding, J. MetaEuk-sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* (2020). doi:10.1186/s40168-020-00808-x
92. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* (2004). doi:10.1101/gr.1865504
93. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* (2018). doi:10.1093/bioinformatics/bty191
94. Stanke, M. *et al.* AUGUSTUS: A b initio prediction of alternative transcripts. *Nucleic Acids Res.* (2006). doi:10.1093/nar/gkl200
95. Kent, W. J. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* (2002). doi:10.1101/gr.229202
96. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* (2013). doi:10.1093/bioinformatics/bts635
97. Denoeud, F. *et al.* Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* (2008). doi:10.1186/gb-2008-9-12-r175
98. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* (2004). doi:10.1186/1471-2105-5-59
99. Boyer, T. P. *et al.* WORLD OCEAN DATABASE 2013, NOAA Atlas NESDIS 72. *Sydney Levitus, Ed.; Alexey Mishonoc, Tech. Ed.* (2013).

doi:10.7289/V5NZ85MT

100. Aumont, O., Ethé, C., Tagliabue, A., Bopp, L. & Gehlen, M. PISCES-v2: An ocean biogeochemical model for carbon and ecosystem studies. *Geosci. Model Dev.* (2015). doi:10.5194/gmd-8-2465-2015
101. Da Cunha, V., Gaia, M., Nasir, A. & Forterre, P. Asgard archaea do not close the debate about the universal tree of life topology. *PLOS Genet.* (2018). doi:10.1371/journal.pgen.1007215
102. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC Bioinformatics* (2009). doi:10.1186/1471-2105-10-421
103. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* (2013). doi:10.1093/molbev/mst010
104. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* (2015). doi:10.1093/molbev/msu300
105. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* (2017). doi:10.1038/nmeth.4285
106. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
107. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* (2018). doi:10.1093/molbev/msx281
108. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
109. Vanni, C. *et al.* Light into the darkness: Unifying the known and unknown coding sequence space in microbiome analyses. *bioRxiv* 2020.06.30.180448 (2020). doi:10.1101/2020.06.30.180448

Tara Oceans Coordinators

Shinichi Sunagawa[1], Silvia G. Acinas[2], Peer Bork[3,4,5], Eric Karsenti[6,7,11], Chris Bowler[6,7], Christian Sardet[7,9], Lars Stemmann[7,9], Colomban de Vargas[7,19], Patrick Wincker[7,18], Magali Lescot[7,26], Marcel Babin[7,20], Gabriel Gorsky[7,9], Nigel Grimsley[7,24,25], Lionel Guidi[7,9], Pascal Hingamp[7,26], Olivier Jaillon[7,18], Stefanie Kandels[3,7], Daniele Iudicone[10], Hiroyuki Ogata[12], Stéphane Pesant[13,14], Matthew B. Sullivan[15,16,17], Fabrice Not[19], Lee Karp- Boss[21], Emmanuel Boss[21], Guy Cochrane[22], Michael Follows[23], Nicole Poulton[27], Jeroen Raes[28,29,30], Mike Sieracki[27] and Sabrina Speich[31,32].

[1] Department of Biology, institute of Microbiology and swiss institute of Bioinformatics, etH Zürich, Zürich, switzerland.
[2] Department of Marine Biology and Oceanography, institute of Marine sciences–CsiC, Barcelona, spain.
[3] Structural and Computational Biology, european Molecular Biology Laboratory, Heidelberg, Germany.

[4] Max Delbrück Center for Molecular Medicine, Berlin, Germany.
[5] Department of Bioinformatics, Biocenter, university of würzburg, würzburg, Germany.
[6] Institut de Biologie de l'ENS, Département de Biologie, École Normale supérieure, CNRS, INSERM, Université PSL, Paris, France.
[7] Research Federation for the study of Global Ocean systems ecology and evolution, Fr2022/tara GOsee, Paris, France.
[8] Université de Nantes, CNRS, uMr6004, Ls2N, Nantes, France.
[9] Sorbonne université, CNRS, Laboratoire d'Océanographie de Villefranche, villefranche- sur- Mer, France.
[10] Stazione Zoologica anton Dohrn, Naples, Italy.
[11] Directors' research, European Molecular Biology Laboratory, Heidelberg, Germany.
[12] institute for Chemical research, Kyoto university, Kyoto, Japan.
[13] PaNGaea, university of Bremen, Bremen, Germany.
[14] MaruM, Center for Marine environmental sciences, university of Bremen, Bremen, Germany.
[15] Department of Microbiology, the Ohio state university, Columbus, OH, USA.
[16] Department of Civil, environmental and Geodetic engineering, the Ohio state university, Columbus, OH, USA.
[17] Center for RNA Biology, the Ohio state university, Columbus, OH, USA.
[18] Génomique Métabolique, Genoscope, institut de Biologie Francois Jacob, Commissariat à l'Énergie atomique, CNrs, université evry, université Paris- saclay, evry, France.
[19] Sorbonne université and CNRS, UMR 7144 (AD2M), ECOMAP, station Biologique de Roscoff, Roscoff, France.
[20] Département de Biologie, Québec Océan and Takuvik Joint International Laboratory (UMI 3376), Université Laval (Canada)–CNRS (France), Université Laval, Quebec, QC, Canada.
[21] School of Marine Sciences, University of Maine, Orono, ME, USA. 22European Molecular Biology Laboratory, European Bioinformatics Institute, Welcome Trust Genome Campus, Hinxton, Cambridge, UK.
[23] Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA.
[24] CNRS UMR 7232, Biologie Intégrative des Organismes Marins, Banyuls- sur- Mer, France.
[25] Sorbonne Universités Paris 06, OOB UPMC, Banyuls- sur- Mer, France.
[26] Aix Marseille Universit/e, Université de Toulon, CNRS, IRD, MIO UM 110, Marseille, France.
[27] Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA.
[28] Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven, Belgium.
[29] Center for the Biology of Disease, VIB KU Leuven, Leuven, Belgium.
[30] Department of Applied Biological Sciences, Vrije Universiteit Brussel, Brussels, Belgium.
[31] Department of Geosciences, Laboratoire de Météorologie Dynamique, École Normale Supérieure, Paris, France.
[32] Ocean Physics Laboratory, University of Western Brittany, Brest, France.