



HAL
open science

Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean

Tom Delmont, Morgan Gaia, Damien Hinsinger, Paul Frémont, Chiara Vanni, Antonio Fernandez-Guerra, A. Murat Eren, Artem Kourlaiev, Leo d'Agata, Quentin Clayssen, et al.

► To cite this version:

Tom Delmont, Morgan Gaia, Damien Hinsinger, Paul Frémont, Chiara Vanni, et al.. Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genomics*, 2022, 2 (5), pp.100123. 10.1016/j.xgen.2022.100123 . hal-03872935v2

HAL Id: hal-03872935

<https://hal.science/hal-03872935v2>

Submitted on 22 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Functional repertoire convergence of distantly related eukaryotic plankton lineages abundant in the sunlit ocean

Tom O. Delmont^{*1,2}, Morgan Gaia^{1,2}, Damien D. Hingsinger^{1,2}, Paul Fremont^{1,2}, Chiara Vanni³, Antonio Fernandez Guerra^{3,4}, A. Murat Eren^{5,6}, Artem Kourlaiev^{1,2}, Leo d'Agata^{1,2}, Quentin Clayssen^{1,2}, Emilie Villar¹, 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, Colombar 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.

*Lead contact: Tom O. Delmont (Tom.Delmont@genoscope.fr)

Summary: Marine planktonic eukaryotes play critical roles 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 completely 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

Genome-wide functional classification of unicellular eukaryotic plankton

52 Introduction

53

54 Plankton in the sunlit ocean contributes about half of Earth's primary productivity,
55 impacting global biogeochemical cycles and food webs^{1,2}. Plankton biomass appears
56 to be dominated by unicellular eukaryotes and small animals³⁻⁶ including a
57 phenomenal evolutionary and morphological biodiversity^{5,7,8}. The composition of
58 planktonic communities is highly dynamical and shaped by biotic and abiotic
59 variables, some of which are changing abnormally fast in the Anthropocene⁹⁻¹¹. Our
60 understanding of marine eukaryotes has progressed substantially in recent years
61 with the transcriptomic (e.g.,^{12,13}) and genomic (e.g.,¹⁴⁻¹⁶) analyses of organisms
62 isolated in culture, and the emergence of efficient culture-independent surveys
63 (e.g.,^{17,18}). However, most eukaryotic lineages' genomic content remains
64 uncharacterized^{19,20}, limiting our understanding of their evolution, functioning,
65 ecological interactions, and resilience to ongoing environmental changes.

66

67 Over the last decade, the *Tara* Oceans program has generated a homogeneous
68 resource of marine plankton metagenomes and metatranscriptomes from the sunlit
69 zone of all major oceans and two seas²¹. Critically, most of the sequenced plankton
70 size fractions correspond to eukaryotic organismal sizes, providing a prime dataset
71 to survey genomic traits and expression patterns from this domain of life. More than
72 100 million eukaryotic gene clusters have been characterized by the
73 metatranscriptomes, half of which have no similarity to known proteins⁵. Most of
74 them could not be linked to a genomic context²², limiting their usefulness to gene-
75 centric insights. The eukaryotic metagenomic dataset (the equivalent of ~10,000
76 human genomes) on the other hand has been partially used for plankton
77 biogeographies^{23,24}, but remains unexploited for the characterization of genes and
78 genomes due to a lack of robust methodologies to make sense of its diversity.

79

80 Genome-resolved metagenomics²⁵ has been extensively applied to the smallest *Tara*
81 Oceans plankton size fractions, unveiling the ecology and evolution of thousands of
82 viral, bacterial, and archaeal populations abundant in the sunlit ocean²⁶⁻³¹. This
83 approach may thus be appropriate also to characterize the genomes of the most
84 abundant eukaryotic plankton. However, very few eukaryotic genomes have been
85 resolved from metagenomes thus far^{26,32-35}, in part due to their complexity (e.g.,
86 high density of repeats³⁶) and extended size³⁷ that might have convinced many of
87 the unfeasibility of such a methodology. With the notable exception of some
88 photosynthetic eukaryotes^{26,32,35}, metagenomics is lagging far behind cultivation for
89 eukaryote genomics, contrasting with the two other domains of life. Here we fill this
90 critical gap using hundreds of billions of metagenomic reads generated from the
91 eukaryotic plankton size fractions of *Tara* Oceans and demonstrate that genome-
92 resolved metagenomics is well suited for marine eukaryotic genomes of substantial
93 complexity and length exceeding the emblematic gigabase. We used this new
94 genomic resource to place major eukaryotic planktonic lineages in the tree of life
95 and explore their evolutionary history based on both phylogenetic signals from
96 conserved gene markers and present-day genomic functional landscape.

97 Results and discussion

98 A new resource of environmental genomes for eukaryotic plankton from the 99 sunlit ocean

100
101
102 We performed the first comprehensive genome-resolved metagenomic survey of
103 microbial eukaryotes from polar, temperate, and tropical sunlit oceans using 798
104 metagenomes (265 of which were released through the present study) derived from
105 the *Tara* Oceans expeditions. They correspond to the surface and deep chlorophyll
106 maximum layer of 143 stations from the Pacific, Atlantic, Indian, Arctic, and
107 Southern Oceans, as well as the Mediterranean and Red Seas, encompassing eight
108 eukaryote-enriched plankton size fractions ranging from 0.8 μm to 2 mm (Figure 1,
109 Table S1). We used the 280 billion reads as inputs for 11 metagenomic co-
110 assemblies (6-38 billion reads per co-assembly) using geographically bounded
111 samples (Figure 1, Table S2), as previously done for the *Tara* Oceans 0.2–3 μm size
112 fraction enriched in bacterial cells²⁶. We favored co-assemblies to gain in coverage
113 and optimize the recovery of large marine eukaryotic genomes. However, it is likely
114 that other assembly strategies (e.g., from single samples) will provide access to
115 genomic data our complex metagenomic co-assemblies failed to resolve. In addition,
116 we used 158 eukaryotic single cells sorted by flow cytometry from seven *Tara*
117 Oceans stations (Table S2) as input to perform complementary genomic assemblies
118 (STAR Methods).

119
120 We thus created a culture-independent, non-redundant (average nucleotide identity
121 <98%) genomic database for eukaryotic plankton in the sunlit ocean consisting of
122 683 metagenome-assembled genomes (MAGs) and 30 single-cell genomes (SAGs),
123 all containing more than 10 million nucleotides (Table S3). These 713 MAGs and
124 SAGs were manually characterized and curated using a holistic framework within
125 anvio^{38,39} that relied heavily on differential coverage across metagenomes (STAR
126 Methods and Supplemental Material). Nearly half the MAGs did not have vertical
127 coverage >10x in any of the metagenomes, emphasizing the relevance of co-
128 assemblies to gain sufficient coverage for relatively large eukaryotic genomes.
129 Moreover, one-third of the SAGs remained undetected by *Tara* Oceans'
130 metagenomic reads, emphasizing cell sorting's power to target less abundant
131 lineages. Absent from the MAGs and SAGs are DNA molecules physically associated
132 with the focal eukaryotic populations, but that did not necessarily correlate with
133 their nuclear genomes across metagenomes or had distinct sequence composition.
134 They include chloroplasts, mitochondria, and viruses generally present in multi-
135 copy. Finally, some highly conserved multi-copy genes such as the 18S rRNA gene
136 were also missing due to technical issues associated with assembly and binning,
137 following the fate of 16S rRNA genes in marine bacterial MAGs²⁶.

138
139 This new genomic database for eukaryotic plankton has a total size of 25.2 Gbp and
140 contains 10,207,450 genes according to a workflow combining metatranscriptomics,
141 *ab-initio*, and protein-similarity approaches (STAR Methods). Estimated completion

Genome-wide functional classification of unicellular eukaryotic plankton

142 of the *Tara* Oceans MAGs and SAGs averaged to ~40% (redundancy of 0.5%) and
143 ranged from 0.0% (a 15 Mbp long Opisthokonta MAG) to 93.7% (a 47.8 Mbp long
144 Ascomycetes MAG). Genomic lengths averaged to 35.4 Mbp (up to 1.32 Gbp for the
145 first Giga-scale eukaryotic MAG, affiliated to *Odontella weissflogii*), with a GC-content
146 ranging from 18.7% to 72.4% (Table S3). MAGs and SAGs are affiliated to Alveolata
147 (n=44), Amoebozoa (n=4), Archaeplastida (n=64), Cryptista (n=31), Haptista
148 (n=92), Opisthokonta (n=299), Rhizaria (n=2), and Stramenopiles (n=174). Only
149 three closely related MAGs could not be affiliated to any known eukaryotic
150 supergroup (see the phylogenetic section below). Among the 713 MAGs and SAGs,
151 271 contained multiple genes corresponding to chlorophyll *a-b* binding proteins and
152 were considered phytoplankton (Table S3). Genome-wide comparisons with 484
153 reference transcriptomes from isolates of marine eukaryotes (the METdb database⁴⁰
154 which improved data from MMETSP¹² and added new transcriptomes from *Tara*
155 Oceans, see Table S3) linked only 24 of the MAGs and SAGs (~3%) to a eukaryotic
156 population already in culture (average nucleotide identity >98%). These include
157 well-known Archaeplastida populations within the genera *Micromonas*, *Bathycoccus*,
158 *Ostreococcus*, *Pycnococcus*, *Chloropicon* and *Prasinoderma* and a few taxa amongst
159 Stramenopiles (e.g., the diatom *Minutocellus polymorphus*) and Haptista (e.g.,
160 *Phaeocystis cordata*). Among this limited number of matches, MAGs represented a
161 nearly identical subset of the corresponding culture genomes (Figure S1, Table S4).
162 Overall, we found metagenomics, single-cell genomics, and culture highly
163 complementary with very few overlaps for marine eukaryotic plankton's genomic
164 characterization.

165
166 The MAGs and SAGs recruited 39.1 billion reads with >90% identity (average
167 identity of 97.4%) from 939 metagenomes, representing 11.8% of the *Tara* Oceans
168 metagenomic dataset dedicated to unicellular and multicellular organisms ranging
169 from 0.2 μm to 2 mm (Table S5). In contrast, METdb with a total size of ~23 Gbp
170 recruited less than 7 billion reads (average identity of 97%), indicating that the
171 collection of *Tara* Oceans MAGs and SAGs reported herein better represents the
172 diversity of open ocean eukaryotes as compared to transcriptomic data from
173 decades of culture efforts worldwide. The majority of *Tara* Oceans metagenomic
174 reads were still not recruited, which could be explained by eukaryotic genomes that
175 our methods failed to reconstruct, the occurrence of abundant bacterial, archaeal,
176 and viral populations in the large size fractions we considered⁴¹⁻⁴³, and the
177 incompleteness of the MAGs and SAGs. Indeed, with the assumption of correct
178 completion estimates, complete MAGs and SAGs would have recruited ~26% of all
179 metagenomic reads, including >50% of reads for the 20-180 μm size fraction alone
180 due in part to an important contribution of hundreds of large copepod MAGs
181 abundant within this cellular range (see Figure 1 and Table S5).

182 Expanding the genomic representation of the eukaryotic tree of life

184
185 We then determined the phylogenetic distribution of the new ocean MAGs and SAGs
186 in the tree of eukaryotic life. METdb was chosen as a taxonomically curated
187 reference transcriptomic database from culture collections, and the two largest

Genome-wide functional classification of unicellular eukaryotic plankton

188 subunits of the three DNA-dependent RNA polymerases (six multi-kilobase genes
189 found in all modern eukaryotes and hence already present in the Last Eukaryotic
190 Common Ancestor). These genes are highly relevant markers for the phylogenetic
191 inference of distantly related microbial organisms⁴⁴ and contributed to our
192 understanding of eukaryogenesis⁴⁵. They have long been overlooked to study the
193 eukaryotic tree of life, possibly because automatic methods are currently missing to
194 effectively identify each DNA-dependent RNA polymerase type prior to performing
195 the phylogenetic analyses. Here, protein sequences were identified using HMMs
196 dedicated to the two largest subunits for the MAGs and SAGs (n=2,150), and METdb
197 reference transcriptomes (n=2,032). These proteins were manually curated and
198 linked to the corresponding DNA-dependent RNA polymerase types for each subunit
199 using reference proteins and phylogenetic inferences (STAR Methods and
200 Supplemental Material). BLAST results provided a novelty score for each of them
201 (STAR Methods and Table S3), expanding the scope of our analysis to eukaryotic
202 genomes stored in NCBI as of August 2020. Our final phylogenetic analysis included
203 416 reference transcriptomes and 576 environmental MAGs and SAGs that
204 contained at least one of the six marker genes (Figure 2). The concatenated DNA-
205 dependent RNA polymerase protein sequences effectively reconstructed a coherent
206 tree of eukaryotic life, comparable to previous large-scale phylogenetic analyses
207 based on other gene markers⁴⁶, and to a complementary BUSCO-centric
208 phylogenomic analysis using protein sequences corresponding to hundreds of
209 smaller gene markers (Figure S2). As a noticeable difference, the Haptista were most
210 closely related to Archaeplastida, while Cryptista included the phylum Picozoa and
211 was most closely related to the TSAR supergroup (Telonemia not represented here,
212 Stramenopiles, Alveolata and Rhizaria), albeit with weaker supports. This view of
213 the eukaryotic tree of life using a previously underexploited universal marker is by
214 no means conclusive by itself but contributes to ongoing efforts to understand deep
215 evolutionary relationships amongst eukaryotes while providing an effective
216 framework to assess the phylogenetic positions of a large number of the *Tara*
217 Oceans MAGs and SAGs.

218
219 Amongst small planktonic animals, the *Tara* Oceans MAGs recovered one lineage of
220 Chordata related to the Oikopleuridae family, and Crustacea including a wide range
221 of copepods (Figure 2, Table S3). Copepods dominate large size fractions of
222 plankton⁸ and represent some of the most abundant animals on the planet^{47,48}. They
223 actively feed on unicellular plankton and are a significant food source for larger
224 animals such as fish, thus representing a key trophic link within the global carbon
225 cycle⁴⁹. For now, less than ten copepod genomes have been characterized by
226 isolates^{50,51}. The additional 8.4 Gbp of genomic material unveiled herein is split into
227 217 MAGs, and themselves organized into two main phylogenetic clusters that we
228 dubbed marine Hexanauplia clades A and B. The two clades considerably expanded
229 the known genomic diversity of copepods, albeit clade B was linked to few reference
230 genomes (Figure S3). These clades were equally abundant and detected in all
231 oceanic regions. Copepod MAGs typically had broad geographic distributions, being
232 detected on average in 25% of the globally distributed *Tara* Oceans stations. In

Genome-wide functional classification of unicellular eukaryotic plankton

233 comparison, Opisthokonta MAGs affiliated to Chordata and Choanoflagellata
234 (Acanthoecida) were, on average detected in less than 10% of sampling sites.

235

236 Generally occurring in smaller size fractions, MAGs and SAGs corresponding to
237 unicellular eukaryotes considerably expanded our genomic knowledge of known
238 genera within Alveolata, Archaeplastida, Haptista and Stramenopiles (Figure 2,
239 Table S3). Just within the diatoms for instance (Stramenopiles), MAGs were
240 reconstructed for *Fragilariopsis* (n=5), *Pseudo-nitzschia* (n=7), *Chaetoceros* (n=11),
241 *Thalassiosira* (n=5) and seven other genera (including the intriguing >1 Gbp long
242 genome of a blooming *Odontella weissflogii species*), all of which are known to
243 contribute significantly to photosynthesis in the sunlit ocean^{52,53}. Among the
244 Archaeplastida, genome-wide average nucleotide identities and distribution
245 patterns indicated that the large majority of MAGs correspond to distinct
246 populations, many of which have not been characterized by means of culture
247 genomics. Especially, we characterized the genomic content of at least 16
248 *Micromonas* populations (Figure S4), 11 *Chloropicon* populations (Figure S5) and 5
249 *Bathycoccus* populations (Figure S6). Beyond this genomic expansion of known
250 planktonic genera, MAGs and SAGs covered various lineages lacking representatives
251 in METdb. These included (1) Picozoa as a sister clade to Cryptista (SAGs from this
252 phylum were recently linked to the Archaeplastida using different gene markers and
253 databases⁵⁴), to the class Chrysophyceae, and the genera *Phaeocystis* and
254 *Pycnococcus*, (2) basal lineages of Oomycota within Stramenopiles and Myxozoa
255 within Alveolata, (3) multiple branches within the MAST lineages⁵⁵ (Figure S7), (4)
256 and a small cluster possibly at the root of Rhizaria we dubbed “putative new group”
257 (Figure S8). The novelty score of individual DNA-dependent RNA polymerase genes
258 was supportive of the topology of the tree. Significantly, diverse MAST lineages,
259 Picozoa and the putative new group all displayed a deep branching distance from
260 cultures and a high novelty score. In addition, the BUSCO-centric phylogenomic
261 analysis placed the “putative new group” at the root of Haptista (Figure S2),
262 supporting its high novelty while stressing the difficulty placing it accurately in the
263 eukaryotic tree of life. In addition, this alternative phylogenomic analysis confirmed
264 placement for the sister clade to *Phaeocystis* but not for the sister clade to
265 *Pycnococcus*, placing it instead as a stand-alone lineage distinct from any
266 Archaeplastida lineages represented by the MAGs, SAGs and METdb. While different
267 gene markers might provide slightly different evolutionary trends, a well-known
268 phylogenetic phenomenon, here our two approaches concur when it comes to
269 emphasizing the genomic novelty of the MAGs and SAGs as compared to culture
270 references.

271

272 One of the most conspicuous lineages lacking any MAGs and SAGs was the
273 Dinoflagellata, a prominent and extremely diverse phylum in small and large
274 eukaryotic size fractions of Tara Oceans⁸. These organisms harbor very large and
275 complex genomes⁵⁶ that likely require much deeper sequencing efforts to be
276 recovered by genome-resolved metagenomics. Besides, many other important
277 lineages are also missing in MAGs and SAGs (e.g., within Radiolaria and Excavata),
278 possibly due to a lack of abundant populations despite their diversity.

Genome-wide functional classification of unicellular eukaryotic plankton

A complex interplay between the evolution and functioning of marine eukaryotes

MAGs and SAGs 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 EggNOG⁵⁷⁻⁵⁹, we identified orthologous groups corresponding to known (n=15,870) and unknown functions (n=12,567, orthologous groups with no assigned function at <http://eggno5.embl.de/>) for 4.7 million genes (nearly 50% of the genes, STAR Methods). Among them, functional redundancy (i.e., a function detected multiple time in the same MAG or SAG) encompassed 46.6% to 96.8% of the gene repertoires (average of 75.2% of functionally redundant genes). We then used these gene annotations to classify the MAGs and SAGs based on their functional profiles (Table S6). Our hierarchical clustering analysis using Euclidean distance and Ward linkage (an approach to organize genomes based on pangenomic traits⁶⁰) first split the MAGs and SAGs 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, MAGs affiliated to the coccolithophore *Emiliana* (completion ranging from 7.8% to 32.2%), Dictyochophaceae family (completion ranging from 8.6% to 76.9%) and the sister clade to *Phaeocystis* (completion ranging from 18.4% to 60.4%) formed distinct clusters. The phylum Picozoa (completion ranging from 1.6% to 75.7%) 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 MAGs 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 putative sister clade to *Pycnococcus*, but also the sensitivity of our framework to draw the functional landscape of unicellular marine eukaryotes. Clearly, the important functional redundancy of MAGs and SAGs minimized the effect of genomic incompleteness in our efforts assessing the functional profile of unicellular marine eukaryotes.

Four major functional groups of unicellular eukaryotes emerged from the hierarchical clustering (Figure 3), which was perfectly recapitulated when incorporating the standard culture genomes matching to a MAG (Figure S9), and when clustering only the MAGs and SAGs >25% complete (Figure S10). 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 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

Genome-wide functional classification of unicellular eukaryotic plankton

325 plankton. Group C was mostly photosynthetic and covered the diatoms
326 (Stramenopiles of various genomic size) and Archaeplastida (small genomes) as
327 sister clusters. This finding likely reflects that diatoms are the only group with an
328 obligatory photoautotrophic lifestyle within the Stramenopiles, like the
329 Archaeplastida. Finally, Group D encompassed three distantly related lineages of
330 heterotrophs (those systematically lacked gene markers for photosynthesis)
331 exhibiting rather large genomes: Oomycota, Acanthoecida choanoflagellates, and
332 Picozoa. Those four functional groups have similar amounts of detected functions
333 and contained both cosmopolite and rarely detected MAGs and SAGs across the *Tara*
334 Oceans stations. While attempts to classify marine eukaryotes based on genomic
335 functional traits have been made in the past (e.g., using a few SAGs⁶¹), our resource
336 therefore provided a broad enough spectrum of genomic material for a first
337 genome-wide functional classification of abundant lineages of unicellular eukaryotic
338 plankton in the upper layer of the ocean.
339

340 A total of 2,588 known and 680 unknown functions covering 1.94 million genes
341 (~40% of the annotated genes) were significantly differentially occurring between
342 the four functional groups (Welch's ANOVA tests, p-value $<1.e^{-05}$, Table S6). We
343 displayed the occurrence of the 100 functions with lowest p-values in the
344 hierarchical clustering presented in Figure 3 to illustrate and help convey the strong
345 signal between groups. However, more than 3,000 functions contributed to the basic
346 partitioning of MAGs and SAGs. They cover all high-level functional categories
347 identified in the 4.7 million genes with similar proportions (Figure S11), indicating
348 that a wide range of functions related to information storage and processing,
349 cellular processes and signaling, and metabolism contribute to the partitioning of
350 the groups. As a notable difference, functions related to transcription (-50%) and
351 RNA processing and modification (-47%) were less represented, while those related
352 to carbohydrate transport and metabolism were enriched (+43%) in the
353 differentially occurring functions. Interestingly, we noticed within Group C a
354 scarcity of various functions otherwise occurring in high abundance among
355 unicellular eukaryotes. These included functions related to ion channels (e.g.,
356 extracellular ligand-gated ion channel activity, intracellular chloride channel
357 activity, magnesium ion transmembrane transporter activity, calcium ion
358 transmembrane transport, calcium sodium antiporter activity) that may be linked to
359 flagellar motility and the response to external stimuli⁶², reflecting the lifestyle of
360 true autotrophs. Group D, on the other hand, had significant enrichment of various
361 functions associated with carbohydrate transport and metabolism (e.g., alpha and
362 beta-galactosidase activities, glycosyl hydrolase families, glycogen debranching
363 enzyme, alpha-L-fucosidase), denoting a distinct carbon acquisition strategy.
364 Overall, the properties of thousands of differentially occurring functions suggest
365 that eukaryotic plankton's complex functional diversity is vastly intertwined within
366 the tree of life, as inferred from phylogenies. This reflects the complex nature of the
367 genomic structure and phenotypic evolution of organisms, which rarely fit their
368 evolutionary relationships.
369

Genome-wide functional classification of unicellular eukaryotic plankton

370 To this point, our analysis focused on the 4.4 million genes that were functionally
371 annotated to EggNOG, which discarded more than half of the genes we identified in
372 the MAGs and SAGs. Our current lack of understanding of many eukaryotic
373 functional genes even within the scope of model organisms⁶³ can explain the limits
374 of reference-based approaches to study the gene content of eukaryotic plankton.
375 Thus, to gain further insights and overcome these limitations, we partitioned and
376 categorized the eukaryotic gene content with AGNOSTOS⁶⁴. AGNOSTOS grouped 5.4
377 million genes in 424,837 groups of genes sharing remote homologies, adding 2.3
378 million genes left uncharacterized by the EggNOG annotation. AGNOSTOS applies a
379 strict set of parameters for the grouping of genes discarding 575,053 genes by its
380 quality controls and 4,264,489 genes in singletons. The integration of the EggNOG
381 annotations into AGNOSTOS resulted in a combined dataset of 25,703 EggNOG
382 orthologous groups (singletons and gene clusters) and 271,464 AGNOSTOS groups
383 of genes, encompassing 6.4 million genes, 45% more genes than the original dataset
384 (STAR Methods). The genome-wide functional classification of MAGs and SAGs
385 based on this extended set of genes supported most trends previously observed
386 with EggNOG annotation alone (Figure S12; Table S7), reinforcing our observations.
387 But most interestingly, classification based solely on 23,674 newly identified groups
388 of genes of unknown function (Table S8, a total of 1.3 million genes discarded by
389 EggNOG) were also supportive of the overall trends, including notable links between
390 diatoms and green algae and between Picozoa and Acanthoecida (Figure S13). Thus,
391 we identified a functional repertoire convergence of distantly related eukaryotic
392 plankton lineages in both the known and unknown coding sequence space, the latter
393 representing a substantial amount of biologically relevant gene diversity.

394

395 Niche and biogeography of individual eukaryotic populations

396

397 Besides insights into organismal evolution and genomic functions, the MAGs and
398 SAGs provided an opportunity to evaluate the present and future geographical
399 distribution of eukaryotic planktonic populations (close to species-level resolution)
400 using the genome-wide metagenomic read recruitments. Here, we determined the
401 niche characteristics (e.g., temperature range) of 374 MAGs and SAGs (~50% of the
402 resource) detected in at least five stations (Table S9) and used climate models to
403 project world map distributions (<https://gigaplankton.shinyapps.io/TOENDB/>)
404 based on climatologies for the periods of 2006-2015 and 2090-2099²⁴ (STAR
405 Methods and Supplemental Material).

406

407 Each of these MAGs and SAGs was estimated to occur in a surface averaging 42 and
408 39 million km² for the first and second period, respectively, corresponding to ~12%
409 of the surface of the ocean. Our data suggest that most eukaryotic populations in the
410 database will remain widespread for decades to come. However, many changes in
411 biogeography are projected to occur. For instance, the most widespread population
412 in the first period (a MAST MAG) would still be ranked first at the end of the century
413 but with a surface area increasing from 37% to 46% (Figure 4), a gain of 28 million
414 km² corresponding to the surface of North America. Its expansion from the tropics

Genome-wide functional classification of unicellular eukaryotic plankton

415 towards more temperate oceanic regions regardless of longitude is mostly explained
416 by temperature and reflects the expansion of tropical niches due to global warming,
417 echoing recent predictions made with amplicon surveys and imaging data⁶⁵. As an
418 extreme case, the MAG benefiting the most between the two periods (a copepod)
419 could experience a gain of 55 million km² (Figure 4), more than the surface of Asia
420 and Europe combined. On the other hand, the MAG losing most ground (also a
421 copepod) could undergo a decrease of 47 million km². Projected changes in these
422 two examples correlated with various variables (including a notable contribution of
423 silicate), an important reminder that temperature alone cannot explain plankton's
424 biogeography in the ocean. Our integration of genomics, metagenomics, and climate
425 models provided the resolution needed to project individual eukaryotic population
426 niche trajectories in the sunlit ocean.

427

428 Limitations of the study:

429

430 Genome-resolved metagenomics applied to the considerable environmental DNA
431 sequencing legacy of the *Tara* Oceans large cellular size fractions proved effective at
432 complementing our culture portfolio of marine eukaryotes. Nevertheless, the
433 approach failed to cover lineages (1) containing very large genomes (e.g., the
434 Dinoflagelates⁵⁶), (2) only found in low abundance, (3) or found to be abundant but
435 with unusually high levels of microdiversity, challenging metagenomic assemblies
436 (e.g., the prominent *Pelagomonas* genus⁶⁶ for which we only recovered high latitude
437 MAG representatives). Deeper sequencing efforts coupled with long read
438 sequencing technologies will likely overcome many of these limitations in years to
439 come.

440

441 Our functional clustering of marine eukaryotes took advantage of a wide range of
442 genomes manually characterized with the platform *anvi'o*, and also considered
443 numerous gene clusters of unknown function using the AGNOSTOS framework.
444 However, this methodology also contains noticeable limitations. For instance,
445 clustering methodologies can influence the observed trends. Furthermore,
446 integration of additional taxonomic groups that currently lack genomic
447 characterizations might impact functional clustering, similar to what is often
448 observed with phylogenomic analyses. Thus, we anticipate that follow-up
449 investigations might identify functional clusters slightly differing from the four
450 major groups we have identified, refining our understanding of the functional
451 convergence of distantly related eukaryotic lineages identified in our study.

452

453 Conclusion

454

455 Similar to recent advances that elucidated viral, bacterial and archaeal lineages,
456 microbiology is experiencing a shift from cultivation to metagenomics for the
457 genomic characterization of marine eukaryotes *en masse*. Indeed, our culture-
458 independent and manually curated genomic characterization of abundant
459 unicellular eukaryotic populations and microscopic animals in the sunlit ocean

Genome-wide functional classification of unicellular eukaryotic plankton

460 covers a wide range of poorly characterized lineages from multiple trophic levels
461 (e.g., copepods and their prey, mixotrophs, autotrophs, and parasites) and provided
462 the first gigabase-scale metagenome-assembled genome. Our genome-resolved
463 survey and parallel efforts by others^{67,68} are not only different from past
464 transcriptomic surveys of isolated marine organisms but also better represent
465 eukaryotic plankton in the open photic ocean. They represent innovative steps
466 towards using genomics to explore in concert the ecological and evolutionary
467 underpinnings of environmentally relevant eukaryotic organisms, using
468 metagenomics to fill critical gaps in our remarkable culture portfolio²¹.

469
470 Phylogenetic gene markers such as the DNA-dependent RNA polymerases (the basis
471 of our phylogenetic analysis) provide a critical understanding of the origin of
472 eukaryotic lineages and allowed us to place most environmental genomes in a
473 comprehensible evolutionary framework. However, this framework is based on
474 sequence variations within core genes that in theory are inherited from the last
475 eukaryotic common ancestor representing the vertical evolution of eukaryotes,
476 disconnected from the structure of genomes. As such, it does not recapitulate the
477 functional evolutionary journey of plankton, as demonstrated in our genome-wide
478 functional classification of unicellular eukaryotes in both the known and unknown
479 coding sequence space. The dichotomy between phylogeny and function was
480 already well described with morphological and other phenotypic traits and could be
481 explained in part by secondary endosymbiosis events that have spread plastids and
482 genes for their photosynthetic capabilities across the eukaryotic tree of life⁶⁹⁻⁷².
483 Here we moved beyond morphological inferences and disentangled the phylogeny
484 of gene markers and broad genomic functional repertoire of a comprehensive
485 collection of marine eukaryotic lineages. We identified four major genomic
486 functional groups of unicellular eukaryotes made of distantly related lineages. The
487 Stramenopiles proved particularly effective in terms of genomic functional
488 diversification, possibly explaining part of their remarkable success in this biome^{8,73}.

489
490 The topology of phylogenetic trees compared to the functional clustering of a wide
491 range of eukaryotic lineages has revealed contrasting evolutionary journeys for
492 widely scrutinized gene markers of evolution and less studied genomic functions of
493 plankton. The apparent functional convergence of distantly related lineages that
494 coexisted in the same biome for millions of years could not be explained by neither a
495 vertical evolutionary history of unicellular eukaryotes nor their trophic modes
496 (phytoplankton versus heterotrophs), shedding new lights into the complex
497 functional dynamics of plankton over evolutionary time scales. Convergent
498 evolution is a well-known phenomenon of independent origin of biological traits
499 such as molecules and behaviors^{74,75} that has been observed in the morphology of
500 microbial eukaryotes⁷⁶ and is often driven by common selective pressures within
501 similar environmental conditions. However, an independent origin of similar
502 functional profiles is not the only possible explanation for organisms sharing the
503 same habitat. Indeed, one could wonder if lateral gene transfers between
504 eukaryotes^{77,78} have played a central role in these processes, as previously observed
505 between eukaryotic plant pathogens⁷⁹ or grasses⁸⁰. As a case in point, secondary

Genome-wide functional classification of unicellular eukaryotic plankton

506 endosymbiosis events are known to have resulted in massive gene transfers
507 between endosymbionts and their hosts in the oceans^{69,70}. In particular, these
508 events involved transfers of genes from green algae to diatoms⁸¹, two lineages
509 clustering together in our genomic functional classification of eukaryotic plankton.
510 However, lineages sharing the same secondary endosymbiotic history did not
511 always fall in the same functional group. This was the case for diatoms, Haptista and
512 Cryptista that have different functional trends yet originate from a common
513 ancestor that likely acquired its plastid from red and green algae^{69,70,82}. Surveying
514 phylogenetic trends for functions derived from the ~10 million genes identified
515 here will likely contribute to new insights regarding the extent of lateral gene
516 transfers between eukaryotes^{83,84}, the independent emergence of functional traits
517 (convergent evolution), as well as functional losses between lineages⁸⁵, that
518 altogether might have driven the functional convergences of distantly related
519 eukaryotic lineages abundant in the sunlit ocean.

520
521 Regardless of the mechanisms involved, the functional repertoire convergences we
522 observed likely highlight primary organismal functioning, which have fundamental
523 impacts on plankton ecology, and their functions within marine ecosystems and
524 biogeochemical cycles. Thus, the apparent dichotomy between phylogenies (a
525 vertical evolutionary framework) and genome-wide functional repertoires (genome
526 structure evolution) depicted here should be viewed as a fundamental attribute of
527 marine unicellular eukaryotes that we suggest warrants a new rationale for
528 studying the structure and state of plankton, a rationale also based on present-day
529 genomic functions rather than phylogenetic and morphological surveys alone.

530

531 Acknowledgments

532

533 Our survey was made possible by two scientific endeavors: the sampling and
534 sequencing efforts by the *Tara* Oceans Project, and the bioinformatics and
535 visualization capabilities afforded by *anvi'o*. We are indebted to all who contributed
536 to these efforts, as well as other open-source bioinformatics tools for their
537 commitment to transparency and openness. *Tara* Oceans (which includes the *Tara*
538 Oceans and *Tara* Oceans Polar Circle expeditions) would not exist without the
539 leadership of the *Tara* Ocean Foundation and the continuous support of 23
540 institutes (<https://oceans.taraexpeditions.org/>). We thank the commitment of the
541 following people and sponsors who made this singular expedition possible: CNRS
542 (in particular Groupement de Recherche GDR3280 and the Research Federation for
543 the Study of Global Ocean Systems Ecology and Evolution FR2022/Tara GOSEE), the
544 European Molecular Biology Laboratory (EMBL), Genoscope/CEA, the French
545 Ministry of Research and the French Government 'Investissement d'Avenir'
546 programs Oceanomics (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-
547 09), ATIGE Genopole postdoctoral fellowship, HYDROGEN/ANR-14-CE23-0001,
548 MEMO LIFE (ANR-10-LABX-54), PSL Research University (ANR-11-IDEX-0001-02)
549 and EMBRC-France (ANR-10-INBS-02), Fund for Scientific Research—Flanders, VIB,
550 Stazione Zoologica Anton Dohrn, UNIMIB, ANR (projects ALGALVIRUS ANR-17-

Genome-wide functional classification of unicellular eukaryotic plankton

551 CE02- 0012, PHYTBACK/ANR-2010-1709-01, POSEIDON/ANR-09-BLAN-0348,
552 PROMETHEUS/ANR-09-PCS-GENM-217, TARA-GIRUS/ANR-09-PCS-GENM-218), EU
553 FP7 (MicroB3/No. 287589, IHMS/HEALTH-F4-2010-261376), Genopole, CEA DRF
554 Impulsion program. The authors also thank agnès b. and E. Bourgois, the Prince
555 Albert II de Monaco Foundation, the Veolia Foundation, the EDF Foundation EDF
556 Diversiterre, Region Bretagne, Lorient Agglomeration, Worldcourier, Illumina, the
557 EDF Foundation EDF Diversiterre, for support and commitment. The global
558 sampling effort was made possible by countless scientists and crew who performed
559 sampling aboard the Tara from 2009 to 2013. The authors are also grateful to the
560 countries that graciously granted sampling permission. Part of the computations
561 were performed using the platine, titane and curie HPC machine provided through
562 GENCI grants (t2011076389, t2012076389, t2013036389, t2014036389,
563 t2015036389 and t2016036389). We also thank Noan Le Bescot (TernogDesign) for
564 artwork on Figures.

565

566 This article is contribution number XX of *Tara Oceans*.

567

568 Author contributions

569

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

583

584 Main figure titles and legends

585

586 **Figure 1. A genome-resolved metagenomic survey dedicated to eukaryotes in the sunlit ocean.**

587 The map displays *Tara Oceans* stations used to perform genome-resolved metagenomics,
588 summarizes the number of metagenomes, contigs longer than 2,500 nucleotides, and eukaryotic
589 MAGs characterized from each co-assembly and outlines the stations used for single-cell genomics.
590 ARC: Arctic Ocean; MED: Mediterranean Sea; RED: Red Sea, ION: Indian Ocean North; IOS: Indian
591 Ocean South; SOC: Southern Ocean; AON: Atlantic Ocean North; AOS: Atlantic Ocean South; PON:
592 Pacific Ocean North; PSE: Pacific South East; PSW: Pacific South West. The bottom panel summarizes
593 mapping results from the MAGs and SAGs across 939 metagenomes organized into four size
594 fractions. The mapping projection of complete MAGs and SAGs is described in the STAR Methods and
595 Supplemental Material.

596

Genome-wide functional classification of unicellular eukaryotic plankton

597 **Figure 2: Phylogenetic analysis of concatenated DNA-dependent RNA polymerase protein**
598 **sequences from eukaryotic plankton.** The maximum-likelihood phylogenetic tree of the
599 concatenated two largest subunits from the three DNA-dependent RNA polymerases (six genes in
600 total) included *Tara* Oceans MAGs and SAGs and METdb transcriptomes and was generated using a
601 total of 7,243 sites in the alignment and LG+F+R10 model; Opisthokonta was used as the outgroup.
602 Supports for selected clades are displayed. Phylogenetic supports were considered high (aLRT \geq 80
603 and UFBoot \geq 95), medium (aLRT \geq 80 or UFBoot \geq 95) or low (aLRT $<$ 80 and UFBoot $<$ 95) (STAR
604 Methods). The tree was decorated with additional layers using the anvi'o interface. The novelty score
605 layer (STAR Methods) was set with a minimum of 30 (i.e., 70% similarity) and a maximum of 60 (i.e.,
606 40% similarity). Branches and names in red correspond to main lineages lacking representatives in
607 METdb.

608
609 **Figure 3. The genomic functional landscape of unicellular eukaryotes in the sunlit ocean.** The
610 figure displays a hierarchical clustering (Euclidean distance with Ward's linkage) of 681 MAGs and
611 SAGs based on the occurrence of \sim 28,000 functions identified with EggNOG⁵⁷⁻⁵⁹, rooted with small
612 animals (Chordata, Crustacea and copepods) and decorated with layers of information using the
613 anvi'o interactive interface. Layers include the occurrence in log 10 of 100 functions with lowest p-
614 value when performing Welch's ANOVA between the functional groups A, B, C and D (see nodes in the
615 tree). Removed from the analysis were Ciliophora MAGs (gene calling is problematic for this lineage),
616 two less complete MAGs affiliated to Opisthokonta, and functions occurring more than 500 times in
617 the gigabase-scale MAG and linked to retrotransposons connecting otherwise unrelated MAGs and
618 SAGs.

619
620 **Figure 4. World map distribution projections for three eukaryotic MAGs during the periods of**
621 **2006-2015 and 2090-2099.** The probability of presence ranges from 0 (purple) to 1 (red), with
622 green corresponding to a probability of 0.5. The bottom row displays first-rank region-dependent
623 environmental parameters driving the projected shifts of distribution (in regions where $|\Delta P| > 0.1$).
624 Noticeably, projected decreases of silicate in equatorial regions drive 34% of the expansion of
625 TARA_PSW_MAG_00299 while driving 34% of the reduction of TARA_PSE_93_MAG_00246, possibly
626 reflecting different life strategies of these copepods (e.g., grazing). In contrast, the expansion of
627 TARA_IOS_50_MAG_00098 is mostly driven by temperature (74%).

628

629 STAR Methods

630

631 RESOURCE AVAILABILITY

632

633 **Lead contact:**

634

- 635 • Further information and requests for resources and analyses should be
636 directed to and will be fulfilled by the lead contact, Tom O. Delmont
637 (Tom.Delmont@genoscope.fr).

638

639 **Materials availability:**

640

- 641 • This study did not generate new materials.

642

643 **Data and code availability:**

644

645 All data our study generated are publicly available at
646 <http://www.genoscope.cns.fr/tara/>. The link provides access to the 11 raw

Genome-wide functional classification of unicellular eukaryotic plankton

647 metagenomic co-assemblies, the FASTA files for 713 MAGs and SAGs, the ~10
648 million protein-coding sequences (nucleotides, amino acids and gff format),
649 and the curated DNA-dependent RNA polymerase genes (MAGs and SAGs and
650 METdb transcriptomes). This link also provides access to the supplemental
651 figures and the supplemental material. Finally, code development within
652 anvi'o for the BUSCO single copy core genes is available at
653 <https://github.com/merenlab/anvio>.

654

655 • Original code has been deposited at Zenodo and is publicly available. The
656 accession number is listed in the key resources table.

657

658 • Any additional information required to reanalyze the data reported in this
659 paper is available from the lead contact upon request.

660

661

METHOD DETAILS

662

663 • **Tara Oceans metagenomes.** We analyzed a total of 939 *Tara Oceans*
664 metagenomes available at the EBI under project PRJEB402
665 (<https://www.ebi.ac.uk/ena/browser/view/PRJEB402>). 265 of these
666 metagenomes have been released through this study. Table S1 reports
667 accession numbers and additional information (including the number of
668 reads and environmental metadata) for each metagenome.

669

670 • **Genome-resolved metagenomics.** We organized the 798 metagenomes
671 corresponding to size fractions ranging from 0.8 μm to 2 mm into 11
672 'metagenomic sets' based upon their geographic coordinates. We used those
673 0.28 trillion reads as inputs for 11 metagenomic co-assemblies using
674 MEGAHIT⁸⁶ v1.1.1, and simplified the scaffold header names in the resulting
675 assembly outputs using anvi'o^{38,39} v.6.1 (available from
676 <http://merenlab.org/software/anvio>). Co-assemblies yielded 78 million
677 scaffolds longer than 1,000 nucleotides for a total volume of 150.7 Gbp. We
678 performed a combination of automatic and manual binning on each co-
679 assembly output, focusing only on the 11.9 million scaffolds longer than
680 2,500 nucleotides, which resulted in 837 manually curated eukaryotic
681 metagenome-assembled genomes (MAGs) longer than 10 million nucleotides.
682 Briefly, (1) anvi'o profiled the scaffolds using Prodigal⁸⁷ v2.6.3 with default
683 parameters to identify an initial set of genes, and HMMER⁸⁸ v3.1b2 to detect
684 genes matching to 83 single-copy core gene markers from BUSCO⁸⁹
685 (benchmarking is described in a dedicated blog post⁹⁰), (2) we used a
686 customized database including both NCBI's NT database and METdb to infer
687 the taxonomy of genes with a Last Common Ancestor strategy⁵ (results were
688 imported as described in [http://merenlab.org/2016/06/18/importing-](http://merenlab.org/2016/06/18/importing-taxonomy)
689 [taxonomy](http://merenlab.org/2016/06/18/importing-taxonomy)), (3) we mapped short reads from the metagenomic set to the
690 scaffolds using BWA v0.7.15⁹¹ (minimum identity of 95%) and stored the
691 recruited reads as BAM files using samtools⁹², (4) anvi'o profiled each BAM

Genome-wide functional classification of unicellular eukaryotic plankton

692 file to estimate the coverage and detection statistics of each scaffold, and
693 combined mapping profiles into a merged profile database for each
694 metagenomic set. We then clustered scaffolds with the automatic binning
695 algorithm CONCOCT⁹³ by constraining the number of clusters (thereafter
696 dubbed metabins) per metagenomic set to a number ranging from 50 to 400
697 depending on the set. Each metabin (n=2,550, ~12 million scaffolds) was
698 manually binned using the anvi'o interactive interface. The interface
699 considers the sequence composition, differential coverage, GC-content, and
700 taxonomic signal of each scaffold. Finally, we individually refined each
701 eukaryotic MAG >10 Mbp as outlined in Delmont and Eren⁹⁴, and renamed
702 scaffolds they contained according to their MAG ID. Table S2 reports the
703 genomic features (including completion and redundancy values) of the
704 eukaryotic MAGs. For details on our protocol used for binning and curation of
705 metabins, see Methods S1, supplemental methods, Related to the STAR
706 Methods.

- 707
708 • **A first Gigabase scale eukaryotic MAG.** We performed targeted genome-
709 resolved metagenomics to confirm the biological relevance and improve
710 statistics of the single MAG longer than 1 Gbp with an additional co-assembly
711 (five Southern Ocean metagenomes for which this MAG had average vertical
712 coverage >1x) and by considering contigs longer than 1,000 nucleotides,
713 leading to a gain of 181,8 million nucleotides. To our knowledge, we describe
714 here the first successful characterization of a Gigabase-scale MAG (1.32 Gbp
715 with 419,520 scaffolds), which we could identify using two distinct
716 metagenomic co-assemblies.
- 717
718 • **MAGs from the 0.2–3 µm size fraction.** We incorporated into our database
719 20 eukaryotic MAGs longer than 10 million nucleotides previously
720 characterized from the 0.2–3 µm size fraction²⁶, providing a set of MAGs
721 corresponding to eukaryotic cells ranging from 0.2 µm (picoeukaryotes) to 2
722 mm (small animals).
- 723
724 • **Single-cell genomics:** We used 158 eukaryotic single cells sorted by flow
725 cytometry from seven *Tara* Oceans stations as input to perform genomic
726 assemblies (up to 18 cells with identical 18S rRNA genes per assembly to
727 optimize completion statistics, see Supplementary Table 2), providing 34
728 single-cell genomes (SAGs) longer than 10 million nucleotides. Cell sorting,
729 DNA amplification, sequencing and assembly were performed as described
730 elsewhere¹⁸. In addition, manual curation was performed using sequence
731 composition and differential coverage across 100 metagenomes in which the
732 SAGs were most detected, following the methodology described in the
733 genome-resolved metagenomics section. For SAGs with no detection in *Tara*
734 Oceans metagenomes, only sequence composition and taxonomical signal
735 could be used, limiting this curation effort's scope. Notably, manual curation
736 of SAGs using the genome-resolved metagenomic workflow turned out to be

Genome-wide functional classification of unicellular eukaryotic plankton

- 737 highly valuable, leading to the removal of more than one hundred thousand
738 scaffolds for a total volume of 193.1 million nucleotides. This metagenomic-
739 guided decontamination effort contributes to previous efforts characterizing
740 eukaryotic SAGs from the same cell sorting material^{18,61,95-97} and provides
741 new marine eukaryotic guidelines SAGs. For details on our protocol used for
742 curation of eukaryotic SAGs, see Methods S1, supplemental methods, Related
743 to the STAR Methods.
744
- 745 • **Characterization of a non-redundant database of MAGs and SAGs.** We
746 determined the average nucleotide identity (ANI) of each pair of MAGs and
747 SAGs using the dnadiff tool from the MUMmer package⁹⁸ v.4.0b2. MAGs and
748 SAGs were considered redundant when their ANI was >98% (minimum
749 alignment of >25% of the smaller MAG or SAG in each comparison). We then
750 selected the longest MAG or SAG to represent a group of redundant MAGs
751 and SAGs. This analysis provided a non-redundant genomic database of 713
752 MAGs and SAGs.
753
 - 754 • **Taxonomical inference of MAGs and SAGs.** We manually determined the
755 taxonomy of MAGs and SAGs using a combination of approaches: (1)
756 taxonomical signal from the initial gene calling (Prodigal), (2) phylogenetic
757 approaches using the RNA polymerase and METdb, (3) ANI within the MAGs
758 and SAGs and between MAGs and SAGs and METdb, (4) local blasts using
759 BUSCO gene markers, (5) and lastly the functional clustering of MAGs and
760 SAGs to gain knowledge into very few MAGs and SAGs lacking gene markers
761 and ANI signal. In addition, Picozoa SAGs⁵⁴ were used to identify MAGs from
762 this phylum lacking representatives in METdb. For details on METdb, see
763 Methods S1, supplemental methods, Related to the STAR Methods.
764
 - 765 • **Protein coding genes.** Protein coding genes for the MAGs and SAGs were
766 characterized using three complementary approaches: protein alignments
767 using reference databases, metatranscriptomic mapping from *Tara* Oceans
768 and *ab-initio* gene predictions. While the overall framework was highly
769 similar for MAGs and SAGs, the methodology slightly differed to take the best
770 advantage of those two databases when they were processed (see the two
771 following sections).
772
 - 773 • **Protein-coding genes for the MAGs. Protein alignments:** Since the
774 alignment of a large protein database on all the MAG assemblies is time
775 greedy, we first detected the potential proteins of Uniref90 + METdb that
776 could be aligned to the assembly by using MetaEuk⁹⁹ with default
777 parameters. This subset of proteins was aligned using BLAT with default
778 parameters, which localized each protein on the MAG assembly. The
779 exon/intron structure was refined using genewise¹⁰⁰ with default
780 parameters to detect splice sites accurately. Each MAG's GeneWise
781 alignments were converted into a standard GFF file and given as input to

Genome-wide functional classification of unicellular eukaryotic plankton

782 gmove. **Metatranscriptomic mapping from *Tara Oceans*:** A total of 905
783 individual *Tara Oceans* metatranscriptomic assemblies (mostly from large
784 planktonic size fractions) were aligned on each MAG assembly using
785 Minimap2¹⁰¹ (version 2.15-r905) with the “-ax splice” flag. BAM files were
786 filtered as follows: low complexity alignments were removed and only
787 alignments covering at least 80% of a given metatranscriptomic contig with
788 at least 95% of identity were retained. The BAM files were converted into a
789 standard GFF file and given as input to gmove. ***Ab-initio* gene predictions:** A
790 first gene prediction for each MAG was performed using gmove and the GFF
791 file generated from metatranscriptomic alignments. From these preliminary
792 gene models, 300 gene models with a start and a stop codon were randomly
793 selected and used to train AUGUSTUS¹⁰² (version 3.3.3). A second time,
794 AUGUSTUS was launched on each MAG assembly using the dedicated
795 calibration file, and output files were converted into standard GFF files and
796 given as input to gmove. Each individual line of evidence was used as input
797 for gmove (<http://www.genoscope.cns.fr/externe/gmove/>) with default
798 parameters to generate the final protein-coding genes annotations.
799

800 • **Protein coding genes for the SAGs. Protein alignments:** The Uniref90 +
801 METdb database of proteins was aligned using BLAT¹⁰³ with default
802 parameters, which localized protein on each SAG assembly. The exon/intron
803 structure was refined using GeneWise¹⁰⁰ and default parameters to detect
804 splice sites accurately. The GeneWise alignments of each SAG were converted
805 into a standard GFF file and given as input to gmove. **Metatranscriptomic
806 mapping from *Tara Oceans*:** The 905 *Tara Oceans* metatranscriptomic
807 individual fastq files were filtered with kfir
808 (<http://www.genoscope.cns.fr/kfir>) using a k-mer approach to select only
809 reads that shared 25-mer with the input SAG assembly. This subset of reads
810 was aligned on the corresponding SAG assembly using STAR¹⁰⁴ (version
811 2.5.2.b) with default parameters. BAM files were filtered as follows: low
812 complexity alignments were removed and only alignments covering at least
813 80% of the metatranscriptomic reads with at least 90% of identity were
814 retained. Candidate introns and exons were extracted from the BAM files and
815 given as input to gmorse¹⁰⁵. ***Ab-initio* gene predictions:** *Ab-initio* models
816 were predicted using SNAP¹⁰⁶ (v2013-02-16) trained on complete protein
817 matches and gmorse models, and output files were converted into standard
818 GFF files and given as input to gmove. Each line of evidence was used as
819 input for gmove (<http://www.genoscope.cns.fr/externe/gmove/>) with
820 default parameters to generate the final protein-coding genes annotations.
821

822 • **BUSCO completion scores for protein-coding genes in MAGs and SAGs.**
823 BUSCO⁸⁹ v.3.0.4 with the set of eukaryotic single-copy core gene markers
824 (n=255). Completion and redundancy (number of duplicated gene markers)
825 of MAGs and SAGs were computed from this analysis.
826

Genome-wide functional classification of unicellular eukaryotic plankton

- 827
- 828
- 829
- 830
- 831
- 832
- 833
- 834
- 835
- 836
- 837
- 838
- 839
- 840
- 841
- 842
- 843
- 844
- 845
- 846
- 847
- 848
- 849
- 850
- 851
- 852
- 853
- 854
- 855
- 856
- 857
- 858
- 859
- 860
- 861
- 862
- 863
- 864
- 865
- 866
- 867
- 868
- 869
- 870
- **Biogeography of MAGs and SAGs.** We performed a final mapping of all metagenomes to calculate the mean coverage and detection of the MAGs and SAGs (Table S5). Briefly, we used BWA v0.7.15 (minimum identity of 90%) and a FASTA file containing the 713 non-redundant MAGs and SAGs to recruit short reads from all 939 metagenomes. We considered MAGs and SAGs were detected in a given filter when >25% of their length was covered by reads to minimize non-specific read recruitments²⁶. 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 MAGs and SAGs 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 MAGs and SAGs with completion <10% to avoid substantial errors to be made in the projections.
 - **Identifying the environmental niche of MAGs and SAGs.** Seven physicochemical parameters were used to define environmental niches: sea surface temperature (SST), salinity (Sal), dissolved silica (Si), nitrate (NO₃), phosphate (PO₄), iron (Fe), and a seasonality index of nitrate (SI NO₃). Except for Fe and SI NO₃, these parameters were extracted from the gridded World Ocean Atlas 2013 (WOA13)¹⁰⁷. Climatological Fe fields were provided by the biogeochemical model PISCES-v2¹⁰⁸. 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. For details on the environmental niches, see Methods S1, supplemental methods, Related to the STAR Methods.
 - **Cosmopolitan score.** Using metagenomes from the Station subset 1 (n=757), MAGs and SAGs were assigned a “cosmopolitan score” based on their detection across 119 stations. For details on metagenomic subsets, see Methods S1, supplemental methods, Related to the STAR Methods.
 - **A database of manually curated DNA-dependent RNA polymerase genes.** A eukaryotic dataset¹⁰⁹ 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 (Prodigal⁸⁷ annotation) in the MAGs and SAGs 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 MAGs and SAGs (n=2,150) and METdb (n=2,032) with no duplicates. For details on this

Genome-wide functional classification of unicellular eukaryotic plankton

- 871 protocol, see Methods S1, supplemental methods, Related to the STAR
872 Methods.
873
- 874 • **Novelty score for the DNA-dependent RNA polymerase genes.** We
875 compared both the RNA-Pol A and RNA-Pol B peptides sequences identified
876 in MAGs and SAGs and MetDB to the nr database (retrieved on October 25,
877 2019) using blastp, as implemented in blast+¹¹⁰ v.2.10.0 (e-value of $1e^{-10}$).
878 We kept the best hit and considered it as the closest sequence present in the
879 public database. For each MAG and SAG, we computed the average percent
880 identity across RNA polymerase genes (up to six genes) and defined the
881 novelty score by subtracting this number from 100. For example, with an
882 average percent identity is 64%, the novelty score would be 36%.
883
 - 884 • **Phylogenetic analyses of MAGs and SAGs.** The protein sequences included
885 for the phylogenetic analyses (either the **DNA-dependent RNA polymerase**
886 **genes** we recovered manually or the **BUSCO set of 255 eukaryotic single-**
887 **copy core gene markers** we recovered automatically from the ~10 million
888 protein coding genes) were aligned with MAFFT¹¹¹ v.7.64 and the FFT-NS-i
889 algorithm with default parameters. Sites with more than 50% of gaps were
890 trimmed using Goalign v0.3.0-alpha5
891 (<http://www.github.com/evolbioinfo/goalign>). The phylogenetic trees were
892 reconstructed with IQ-TREE¹¹² v1.6.12, and the model of evolution was
893 estimated with the ModelFinder¹¹³ Plus option: for the concatenated tree, the
894 LG+F+R10 model was selected. Supports were computed from 1,000
895 replicates for the Shimodaira-Hasegawa (SH)-like approximation likelihood
896 ratio (aLRT)¹¹⁴ and ultrafast bootstrap approximation (UFBoot)¹¹⁵. As per IQ-
897 TREE manual, we deemed the supports good when SH-aLRT $\geq 80\%$ and
898 UFBoot $\geq 95\%$. Anvi'o v.6.1 was used to visualize and root the phylogenetic
899 trees.
900
 - 901 • **EggNOG functional inference of MAGs and SAGs.** We performed the
902 functional annotation of protein-coding genes using the EggNog-mapper^{58,59}
903 v2.0.0 and the EggNog5 database⁵⁷. We used Diamond¹¹⁶ v0.9.25 to align
904 proteins to the database. We refined the functional annotations by selecting
905 the orthologous group within the lowest taxonomic level predicted by
906 EggNog-mapper.
907
 - 908 • **Eukaryotic MAGs and SAGs integration in the AGNOSTOS-DB.** We used
909 the AGNOSTOS workflow to integrate the protein coding genes predicted
910 from the MAGs and SAGs into a variant of the AGNOSTOS-DB that contains
911 1,829 metagenomes from the marine and human microbiomes, 28,941
912 archaeal and bacterial genomes from the Genome Taxonomy Database
913 (GTDB) and 3,243 nucleocytoplasmic large DNA viruses (NCLDV)
914 metagenome assembled genomes (MAGs)⁶⁴.
915

Genome-wide functional classification of unicellular eukaryotic plankton

- 916
- **AGNOSTOS functional aggregation inference.** AGNOSTOS partitioned protein coding genes from the MAGs and SAGs 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. 2020⁶⁴. 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. 2020⁶⁴ 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 MAGs and SAGs.
- 931
- **Functional clustering of MAGs and SAGs.** We used *anvi'o* to cluster MAGs and SAGs 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.

QUANTIFICATION AND STATISTICAL ANALYSIS

- 937
- **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 MAGs and SAGs. All statistics were generated in R 3.5.3. Results are available in the table S6.

Declaration of interest

944

945

946 The authors declare no competing interests.

References

- 947
- 948
- 949
1. Boyd, P.W. (2015). Toward quantifying the response of the oceans' biological pump to climate change. *Front. Mar. Sci.* 2, 77.
 2. Sanders, R., Henson, S.A., Koski, M., De La Rocha, C.L., Painter, S.C., Poulton, A.J., Riley, J., Salihoglu, B., Visser, A., Yool, A., et al. (2014). The Biological Carbon Pump in the North Atlantic. *Prog. Oceanogr.* 129, 200–218.
 3. Dortch, Q., and Packard, T.T. (1989). Differences in biomass structure between oligotrophic and eutrophic marine ecosystems. *Deep Sea Res. Part A. Oceanogr. Res. Pap.* 36, 223–240.
 4. Gasol, J.M., Del Giorgio, P.A., and Duarte, C.M. (1997). Biomass distribution in marine planktonic communities. *Limnol. Oceanogr.* 42, 1353–1363.
 5. Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-

Genome-wide functional classification of unicellular eukaryotic plankton

- 961 Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., et al. (2018).
962 A global ocean atlas of eukaryotic genes. *Nat. Commun.* 2018 91 9, 1–13.
- 963 6. Caron, D.A., Countway, P.D., Jones, A.C., Kim, D.Y., and Schnetzer, A. (2011).
964 Marine Protistan Diversity. *Ann. Rev. Mar. Sci.* 4, 467–493.
- 965 7. Leray, M., and Knowlton, N. (2016). Censusing marine eukaryotic diversity in
966 the twenty-first century. *Philos. Trans. R. Soc. B Biol. Sci.* 371.
- 967 8. De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E.,
968 Berney, C., Le Bescot, N., Probert, I., et al. (2015). Eukaryotic plankton
969 diversity in the sunlit ocean. *Science* (80-.). 348.
- 970 9. Jonkers, L., Hillebrand, H., and Kucera, M. (2019). Global change drives
971 modern plankton communities away from the pre-industrial state. *Nat.* 2019
972 5707761 570, 372–375.
- 973 10. Hays, G.C., Richardson, A.J., and Robinson, C. (2005). Climate change and
974 marine plankton. *Trends Ecol. Evol.* 20, 337–344.
- 975 11. Hutchins, D.A., and Fu, F. (2017). Microorganisms and ocean global change.
976 *Nat. Microbiol.* 2017 26 2, 1–11.
- 977 12. Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A.,
978 Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., et al. (2014). The Marine
979 Microbial Eukaryote Transcriptome Sequencing Project (MMETSP):
980 Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through
981 Transcriptome Sequencing. *PLOS Biol.* 12, e1001889.
- 982 13. Johnson, L.K., Alexander, H., and Brown, C.T. (2019). Re-assembly, quality
983 evaluation, and annotation of 678 microbial eukaryotic reference
984 transcriptomes. *Gigascience* 8, 1–12.
- 985 14. Palenik, B., Grimwood, J., Aerts, A., Rouzé, P., Salamov, A., Putnam, N., Dupont,
986 C., Jorgensen, R., Derelle, E., Rombauts, S., et al. (2007). The tiny eukaryote
987 *Ostreococcus* provides genomic insights into the paradox of plankton
988 speciation. *Proc. Natl. Acad. Sci. U. S. A.* 104, 7705–7710.
- 989 15. Bowler, C., Allen, A.E., Badger, J.H., Grimwood, J., Jabbari, K., Kuo, A.,
990 Maheswari, U., Martens, C., Maumus, F., Otiillar, R.P., et al. (2008). The
991 *Phaeodactylum* genome reveals the evolutionary history of diatom genomes.
992 *Nat.* 2008 4567219 456, 239–244.
- 993 16. Worden, A.Z., Lee, J.H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., Allen,
994 A.E., Cuvelier, M.L., Derelle, E., Everett, M. V., et al. (2009). Green evolution and
995 dynamic adaptations revealed by genomes of the marine picoeukaryotes
996 *Micromonas*. *Science* 324, 268–272.
- 997 17. Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G.,
998 Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Ocean
999 plankton. Structure and function of the global ocean microbiome. *Science* 348,
1000 1261359.
- 1001 18. Sieracki, M.E., Poulton, N.J., Jaillon, O., Wincker, P., de Vargas, C., Rubinat-
1002 Ripoll, L., Stepanauskas, R., Logares, R., and Massana, R. (2019). Single cell
1003 genomics yields a wide diversity of small planktonic protists across major
1004 ocean ecosystems. *Sci. Reports* 2019 91 9, 1–11.
- 1005 19. Sibbald, S.J., and Archibald, J.M. (2017). More protist genomes needed. *Nat.*
1006 *Ecol. Evol.* 1.

Genome-wide functional classification of unicellular eukaryotic plankton

- 1007 20. Del Campo, J., Sieracki, M.E., Molestina, R., Keeling, P., Massana, R., and Ruiz-
1008 Trillo, I. (2014). The others: our biased perspective of eukaryotic genomes.
1009 Trends Ecol. Evol. 29, 252.
- 1010 21. Sunagawa, S., Acinas, S.G., Bork, P., Bowler, C., Eveillard, D., Gorsky, G., Guidi,
1011 L., Iudicone, D., Karsenti, E., Lombard, F., et al. (2020). Tara Oceans: towards
1012 global ocean ecosystems biology. Nat. Rev. Microbiol., 1–18.
- 1013 22. Vorobev, A., Dupouy, M., Carradec, Q., Delmont, T.O., Annamalé, A., Wincker, P.,
1014 and Pelletier, E. (2020). Transcriptome reconstruction and functional analysis
1015 of eukaryotic marine plankton communities via high-throughput
1016 metagenomics and metatranscriptomics. Genome Res. 30, 647–659.
- 1017 23. Richter, D., Watteaux, R., Vannier, T., Leconte, J., Frémont, P., Reygondeau, G.,
1018 Maillet, N., Henry, N., Benoit, G., Fernández-Guerra, A., et al. (2019). Genomic
1019 evidence for global ocean plankton biogeography shaped by large-scale
1020 current systems. bioRxiv doi:https://doi.org/10.1101/867739 23, 31.
- 1021 24. Frémont, P., Gehlen, M., Vrac, M., Leconte, J., Wincker, P., Iudicone, D., and
1022 Jaillon, O. (2020). Restructuring of genomic provinces of surface ocean
1023 plankton under climate change. bioRxiv, 2020.10.20.347237.
- 1024 25. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson,
1025 P.M., Solovyev, V. V., Rubin, E.M., Rokhsar, D.S., and Banfield, J.F. (2004).
1026 Community structure and metabolism through reconstruction of microbial
1027 genomes from the environment. Nature 428, 37–43.
- 1028 26. Delmont, T.O., Quince, C., Shaiber, A., Esen, Ö.C., Lee, S.T., Rappé, M.S.,
1029 MacLellan, S.L., Lückner, S., and Eren, A.M. (2018). Nitrogen-fixing populations
1030 of Planctomycetes and Proteobacteria are abundant in surface ocean
1031 metagenomes. Nat. Microbiol. 2018 37 3, 804–813.
- 1032 27. Tully, B.J., Graham, E.D., and Heidelberg, J.F. (2018). The reconstruction of
1033 2,631 draft metagenome-assembled genomes from the global oceans. Sci. Data
1034 2018 51 5, 1–8.
- 1035 28. Tully, B.J. (2019). Metabolic diversity within the globally abundant Marine
1036 Group II Euryarchaea offers insight into ecological patterns. Nat. Commun.
1037 2019 101 10, 1–12.
- 1038 29. Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B.,
1039 Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., et al. (2019).
1040 Marine DNA Viral Macro- and Microdiversity from Pole to Pole. Cell 177,
1041 1109-1123.e14.
- 1042 30. Moniruzzaman, M., Martinez-Gutierrez, C.A., Weinheimer, A.R., and Aylward,
1043 F.O. (2020). Dynamic genome evolution and complex virocell metabolism of
1044 globally-distributed giant viruses. Nat. Commun. 2020 111 11, 1–11.
- 1045 31. Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans,
1046 P.N., Hugenholtz, P., and Tyson, G.W. (2017). Recovery of nearly 8,000
1047 metagenome-assembled genomes substantially expands the tree of life. Nat.
1048 Microbiol. 2017 211 2, 1533–1542.
- 1049 32. Delmont, T.O., Murat Eren, A., Vineis, J.H., and Post, A.F. (2015). Genome
1050 reconstructions indicate the partitioning of ecological functions inside a
1051 phytoplankton bloom in the Amundsen Sea, Antarctica. Front. Microbiol. 6.
- 1052 33. Olm, M.R., West, P.T., Brooks, B., Firek, B.A., Baker, R., Morowitz, M.J., and

Genome-wide functional classification of unicellular eukaryotic plankton

- 1053 Banfield, J.F. (2019). Genome-resolved metagenomics of eukaryotic
1054 populations during early colonization of premature infants and in hospital
1055 rooms. *Microbiome* 7, 1–16.
- 1056 34. West, P.T., Probst, A.J., Grigoriev, I. V., Thomas, B.C., and Banfield, J.F. (2018).
1057 Genome-reconstruction for eukaryotes from complex natural microbial
1058 communities. *Genome Res.* 28, gr.228429.117.
- 1059 35. Duncan, A., Barry, K., Daum, C., Eloë-Fadrosh, E., Roux, S., Tringe, S.G., Schmidt,
1060 K., Valentin, K.U., Varghese, N., Grigoriev, I. V., et al. (2020). Metagenome-
1061 assembled genomes of phytoplankton communities across the Arctic Circle.
1062 *bioRxiv*, 2020.06.16.154583.
- 1063 36. Biscotti, M.A., Olmo, E., and Heslop-Harrison, J.S. (Pat. (2015). Repetitive DNA
1064 in eukaryotic genomes. *Chromosome Res.* 23, 415–420.
- 1065 37. Gregory, T.R. (2005). Synergy between sequence and size in Large-scale
1066 genomics. *Nat. Rev. Genet.* 2005 69 6, 699–708.
- 1067 38. Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., and
1068 Delmont, T.O. (2015). Anvi'o: an advanced analysis and visualization platform
1069 for 'omics data. *PeerJ* 3, e1319.
- 1070 39. Eren, A.M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S.E., Schechter, M.S., Fink, I.,
1071 Pan, J.N., Yousef, M., Fogarty, E.C., et al. (2020). Community-led, integrated,
1072 reproducible multi-omics with anvi'o. *Nat. Microbiol.* 2020 61 6, 3–6.
- 1073 40. Niang, G., Hoebeke, M., Meng, A., Liu, X., Scheremetjew, M., Finn, R., Pelletier,
1074 E., and Erwan, C. METdb, an extended reference resource for Marine
1075 Eukaryote Transcriptomes. <http://metdb.sb-roscoff.fr/metdb/> (unpublished).
- 1076 41. Delmont, T.O., Pierella Karlusich, J.J., Veseli, I., Fuessel, J., Eren, A.M., Foster,
1077 R.A., Bowler, C., Wincker, P., and Pelletier, E. (2021). Heterotrophic bacterial
1078 diazotrophs are more abundant than their cyanobacterial counterparts in
1079 metagenomes covering most of the sunlit ocean. *ISME J.* 2021, 1–10.
- 1080 42. Delmont, T.O. (2021). Discovery of nondiazotrophic *Trichodesmium* species
1081 abundant and widespread in the open ocean. *Proc. Natl. Acad. Sci.* 118,
1082 e2112355118.
- 1083 43. Gaia, M., Meng, L., Pelletier, E., Forterre, P., Vanni, C., Fernandez-Guerra, A.,
1084 Jaillon, O., Wincker, P., Ogata, H., and Delmont, T.O. (2021). Discovery of a
1085 class of giant virus relatives displaying unusual functional traits and prevalent
1086 within plankton: the Mirusviricetes. *bioRxiv*, 2021.12.27.474232.
- 1087 44. Martinez-Gutierrez, C.A., and Aylward, F.O. (2021). Phylogenetic Signal,
1088 Congruence, and Uncertainty across Bacteria and Archaea. *Mol. Biol. Evol.* 38,
1089 5514–5527.
- 1090 45. Guglielmini, J., Woo, A.C., Krupovic, M., Forterre, P., and Gaia, M. (2019).
1091 Diversification of giant and large eukaryotic dsDNA viruses predated the
1092 origin of modern eukaryotes. *Proc. Natl. Acad. Sci. U. S. A.* 116, 19585–19592.
- 1093 46. Burki, F., Roger, A.J., Brown, M.W., and Simpson, A.G.B. (2020). The New Tree
1094 of Eukaryotes. *Trends Ecol. Evol.* 35, 43–55.
- 1095 47. Humes, A.G. (1994). How many copepods? *Ecol. Morphol. Copepods*, 1–7.
- 1096 48. Kiørboe, T. (2011). What makes pelagic copepods so successful? *J. Plankton*
1097 *Res.* 33, 677–685.
- 1098 49. Steinberg, D.K., and Landry, M.R. (2017). Zooplankton and the Ocean Carbon

Genome-wide functional classification of unicellular eukaryotic plankton

- 1099 Cycle. *Ann. Rev. Mar. Sci.* *9*, 413–444.
- 1100 50. Jørgensen, T.S., Nielsen, B.L.H., Petersen, B., Browne, P.D., Hansen, B.W., and
1101 Hansen, L.H. (2019). The Whole Genome Sequence and mRNA Transcriptome
1102 of the Tropical Cyclopoid Copepod *Apocyclops royi*. *G3*
1103 *Genes|Genomes|Genetics* *9*, 1295–1302.
- 1104 51. Jørgensen, T.S., Petersen, B., Petersen, H.C.B., Browne, P.D., Prost, S., Stillman,
1105 J.H., Hansen, L.H., and Hansen, B.W. (2019). The Genome and mRNA
1106 Transcriptome of the Cosmopolitan Calanoid Copepod *Acartia tonsa* Dana
1107 Improve the Understanding of Copepod Genome Size Evolution. *Genome Biol.*
1108 *Evol.* *11*, 1440–1450.
- 1109 52. Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., Wincker,
1110 P., Iudicone, D., De Vargas, C., Bittner, L., et al. (2016). Insights into global
1111 diatom distribution and diversity in the world's ocean. *Proc. Natl. Acad. Sci. U.*
1112 *S. A.* *113*, E1516–E1525.
- 1113 53. Costa, R.R., Mendes, C.R.B., Tavano, V.M., Dotto, T.S., Kerr, R., Monteiro, T.,
1114 Odebrecht, C., and Secchi, E.R. (2020). Dynamics of an intense diatom bloom
1115 in the Northern Antarctic Peninsula, February 2016. *Limnol. Oceanogr.* *65*,
1116 2056–2075.
- 1117 54. Schön, M.E., Zlatogursky, V. V., Singh, R.P., Poirier, C., Wilken, S., Mathur, V.,
1118 Strassert, J.F.H., Pinhassi, J., Worden, A.Z., Keeling, P.J., et al. (2021). Single cell
1119 genomics reveals plastid-lacking Picozoa are close relatives of red algae. *Nat.*
1120 *Commun.* *2021* *12*, 1–10.
- 1121 55. Massana, R., Del Campo, J., Sieracki, M.E., Audic, S., and Logares, R. (2014).
1122 Exploring the uncultured microeukaryote majority in the oceans: reevaluation
1123 of ribogroups within stramenopiles. *ISME J.* *8*, 854.
- 1124 56. Song, B., Chen, S., and Chen, W. (2018). Dinoflagellates, a Unique Lineage for
1125 Retrogene Research. *Front. Microbiol.* *9*.
- 1126 57. Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S.K.,
1127 Cook, H., Mende, D.R., Letunic, I., Rattei, T., Jensen, L.J., et al. (2019). eggNOG
1128 5.0: a hierarchical, functionally and phylogenetically annotated orthology
1129 resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* *47*,
1130 D309–D314.
- 1131 58. Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Von
1132 Mering, C., and Bork, P. (2017). Fast Genome-Wide Functional Annotation
1133 through Orthology Assignment by eggNOG-Mapper. *Mol. Biol. Evol.* *34*, 2115–
1134 2122.
- 1135 59. Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., and Bork,
1136 P. (2008). eggNOG: Automated construction and annotation of orthologous
1137 groups of genes. *Nucleic Acids Res.* *36*, D250–D254.
- 1138 60. Delmont, T.O., and Eren, E.M. (2018). Linking pangenomes and metagenomes:
1139 The *Prochlorococcus* metapangenome. *PeerJ* *2018*.
- 1140 61. Seeleuthner, Y., Mondy, S., Lombard, V., Carradec, Q., Pelletier, E., Wessner, M.,
1141 Leconte, J., Mangot, J.F., Poulain, J., Labadie, K., et al. (2018). Single-cell
1142 genomics of multiple uncultured stramenopiles reveals underestimated
1143 functional diversity across oceans. *Nat. Commun.* *2018* *9*, 1–10.
- 1144 62. Hill, K., Hemmler, R., Kovermann, P., Calenberg, M., Kreimer, G., and Wagner, R.

Genome-wide functional classification of unicellular eukaryotic plankton

- 1145 (2000). A Ca(2+)- and voltage-modulated flagellar ion channel is a component
1146 of the mechanoshock response in the unicellular green alga *Spermatozopsis*
1147 *similis*. *Biochim. Biophys. Acta* *1466*, 187–204.
- 1148 63. Wood, V., Lock, A., Harris, M.A., Rutherford, K., Bähler, J., and Oliver, S.G.
1149 (2019). Hidden in plain sight: what remains to be discovered in the eukaryotic
1150 proteome? *Open Biol.* *9*, 180241.
- 1151 64. Vanni, C., Schechter, M.S., Acinas, S.G., Barberán, A., Buttigieg, P.L., Casamayor,
1152 E.O., Delmont, T.O., Duarte, C.M., Eren, A.M., Finn, R.D., et al. (2021). Unifying
1153 the known and unknown microbial coding sequence space. *bioRxiv*,
1154 2020.06.30.180448.
- 1155 65. Ibarbalz, F.M., Henry, N., Brandão, M.C., Martini, S., Busseni, G., Byrne, H.,
1156 Coelho, L.P., Endo, H., Gasol, J.M., Gregory, A.C., et al. (2019). Global Trends in
1157 Marine Plankton Diversity across Kingdoms of Life. *Cell* *179*, 1084-1097.e21.
- 1158 66. Guérin, N., Ciccarella, M., Flamant, E., Frémont, P., Mangenot, S., Istace, B., Noel,
1159 B., Romac, S., Bachy, C., Gachenot, M., et al. (2021). Genomic adaptation of the
1160 picoeukaryote *Pelagomonas calceolata* to iron-poor oceans revealed by a
1161 chromosome-scale genome sequence. *bioRxiv*, 2021.10.25.465678.
- 1162 67. Saary, P., Mitchell, A.L., and Finn, R.D. (2020). Estimating the quality of
1163 eukaryotic genomes recovered from metagenomic analysis with EukCC.
1164 *Genome Biol.* *21*, 1–21.
- 1165 68. Alexander, H., Hu, S.K., Krinos, A.I., Pachiadaki, M., Tully, B.J., Neely, C.J., and
1166 Reiter, T. (2021). Eukaryotic genomes from a global metagenomic dataset
1167 illuminate trophic modes and biogeography of ocean plankton. *bioRxiv*,
1168 2021.07.25.453713.
- 1169 69. Archibald, J.M., and Keeling, P.J. (2002). Recycled plastids: a “green
1170 movement” in eukaryotic evolution. *Trends Genet.* *18*, 577–584.
- 1171 70. Deschamps, P., and Moreira, D. (2012). Reevaluating the Green Contribution
1172 to Diatom Genomes. *Genome Biol. Evol.* *4*, 683–688.
- 1173 71. Keeling, P.J. (2013). The number, speed, and impact of plastid endosymbioses
1174 in eukaryotic evolution. *Annu. Rev. Plant Biol.* *64*, 583–607.
- 1175 72. Reyes-Prieto, A., Weber, A.P.M., and Bhattacharya, D. (2007). The Origin and
1176 Establishment of the Plastid in Algae and Plants. *Annu. Rev. Genet.* *41*, 147–
1177 168.
- 1178 73. Derelle, R., López-García, P., Timpano, H., and Moreira, D. (2016). A
1179 Phylogenomic Framework to Study the Diversity and Evolution of
1180 Stramenopiles (=Heterokonts). *Mol. Biol. Evol.* *33*, 2890–2898.
- 1181 74. Emery, N.J., and Clayton, N.S. (2004). The Mentality of Crows: Convergent
1182 Evolution of Intelligence in Corvids and Apes. *Science* (80-). *306*, 1903–1907.
- 1183 75. Zakon, H.H. (2002). Convergent evolution on the molecular level. *Brain.*
1184 *Behav. Evol.* *59*, 250–261.
- 1185 76. Leander, B.S. (2008). A hierarchical view of convergent evolution in microbial
1186 eukaryotes. *J. Eukaryot. Microbiol.* *55*, 59–68.
- 1187 77. Keeling, P.J., and Palmer, J.D. (2008). Horizontal gene transfer in eukaryotic
1188 evolution. *Nat. Rev. Genet.* 2008 *9*, 605–618.
- 1189 78. Danchin, E.G.J. (2016). Lateral gene transfer in eukaryotes: Tip of the iceberg
1190 or of the ice cube. *BMC Biol.* *14*, 1–3.

Genome-wide functional classification of unicellular eukaryotic plankton

- 1191 79. Andersson, J.O. (2006). Convergent evolution: gene sharing by eukaryotic
1192 plant pathogens. *Curr. Biol.* *16*.
- 1193 80. Dunning, L.T., Olofsson, J.K., Parisod, C., Choudhury, R.R., Moreno-Villena, J.J.,
1194 Yang, Y., Dionora, J., Paul Quick, W., Park, M., Bennetzen, J.L., et al. (2019).
1195 Lateral transfers of large DNA fragments spread functional genes among
1196 grasses. *Proc. Natl. Acad. Sci. U. S. A.* *116*, 4416–4425.
- 1197 81. Chan, C.X., Bhattacharya, D., and Reyes-Prieto, A. (2012). Endosymbiotic and
1198 horizontal gene transfer in microbial eukaryotes: Impacts on cell evolution
1199 and the tree of life. *Mob. Genet. Elements* *2*, 101.
- 1200 82. Dorrell, R.G., Gile, G., McCallum, G., Méheust, R., Baptiste, E.P., Klinger, C.M.,
1201 Brillet-Guéguen, L., Freeman, K.D., Richter, D.J., and Bowler, C. (2017).
1202 Chimeric origins of ochrophytes and haptophytes revealed through an ancient
1203 plastid proteome. *Elife* *6*.
- 1204 83. Martin, W.F. (2017). Too Much Eukaryote LGT. *BioEssays* *39*, 1700115.
- 1205 84. Leger, M.M., Eme, L., Stairs, C.W., and Roger, A.J. (2018). Demystifying
1206 Eukaryote Lateral Gene Transfer (Response to Martin 2017 DOI:
1207 10.1002/bies.201700115). *BioEssays* *40*, 1700242.
- 1208 85. Zmasek, C.M., and Godzik, A. (2011). Strong functional patterns in the
1209 evolution of eukaryotic genomes revealed by the reconstruction of ancestral
1210 protein domain repertoires. *Genome Biol.* *12*, 1–13.
- 1211 86. Li, D., Liu, C.M., Luo, R., Sadakane, K., and Lam, T.W. (2014). MEGAHIT: An
1212 ultra-fast single-node solution for large and complex metagenomics assembly
1213 via succinct de Bruijn graph. *Bioinformatics* *31*, 1674–1676.
- 1214 87. Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J.
1215 (2010). Prodigal: prokaryotic gene recognition and translation initiation site
1216 identification. *BMC Bioinformatics* *11*, 119.
- 1217 88. Eddy, S.R. (2011). Accelerated Profile HMM Searches. *PLoS Comput. Biol.* *7*,
1218 e1002195.
- 1219 89. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E. V., and Zdobnov,
1220 E.M. (2015). BUSCO: assessing genome assembly and annotation
1221 completeness with single-copy orthologs. *Bioinformatics* *31*, 3210–3212.
- 1222 90. Delmont, T.O. (2018). Assessing the completion of eukaryotic bins with anvio.
1223 Blog post. <http://merenlab.org/2018/05/05/eukaryotic-single-copy-core-genes/>.
1224
- 1225 91. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with
1226 Burrows-Wheeler transform. *Bioinformatics* *25*, 1754–1760.
- 1227 92. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G.,
1228 Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and
1229 SAMtools. *Bioinformatics* *25*, 2078–2079.
- 1230 93. Alneberg, J., Bjarnason, B.S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U.Z., Lahti,
1231 L., Loman, N.J., Andersson, A.F., and Quince, C. (2014). Binning metagenomic
1232 contigs by coverage and composition. *Nat. Methods* *11*, 1144–1146.
- 1233 94. Delmont, T.O., and Eren, A.M. (2016). Identifying contamination with
1234 advanced visualization and analysis practices: metagenomic approaches for
1235 eukaryotic genome assemblies. *PeerJ* *4*, e1839.
- 1236 95. Mangot, J.F., Logares, R., Sánchez, P., Latorre, F., Seeleuthner, Y., Mondy, S.,

Genome-wide functional classification of unicellular eukaryotic plankton

- 1237 Sieracki, M.E., Jaillon, O., Wincker, P., Vargas, C. De, et al. (2017). Accessing the
1238 genomic information of unculturable oceanic picoeukaryotes by combining
1239 multiple single cells. *Sci. Reports* 2017 7, 1–12.
- 1240 96. López-Escardó, D., Grau-Bové, X., Guillaumet-Adkins, A., Gut, M., Sieracki, M.E.,
1241 and Ruiz-Trillo, I. (2017). Evaluation of single-cell genomics to address
1242 evolutionary questions using three SAGs of the choanoflagellate *Monosiga*
1243 *brevicollis*. *Sci. Reports* 2017 7, 1–14.
- 1244 97. Vannier, T., Leconte, J., Seeleuthner, Y., Mondy, S., Pelletier, E., Aury, J.M., De
1245 Vargas, C., Sieracki, M., Iudicone, D., Vaulot, D., et al. (2016). Survey of the
1246 green picoalga *Bathycoccus* genomes in the global ocean. *Sci. Reports* 2016 6
1247 6, 1–11.
- 1248 98. Delcher, A.L., Phillippy, A., Carlton, J., and Salzberg, S.L. (2002). Fast
1249 algorithms for large-scale genome alignment and comparison. *Nucleic Acids*
1250 *Res.* 30, 2478–2483.
- 1251 99. Levy Karin, E., Mirdita, M., and Söding, J. (2020). MetaEuk-sensitive, high-
1252 throughput gene discovery, and annotation for large-scale eukaryotic
1253 metagenomics. *Microbiome* 8, 1–15.
- 1254 100. Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise.
1255 *Genome Res.* 14, 988–995.
- 1256 101. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences.
1257 *Bioinformatics* 34, 3094–3100.
- 1258 102. Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B.
1259 (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic*
1260 *Acids Res.* 34.
- 1261 103. Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res.* 12, 656–
1262 664.
- 1263 104. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P.,
1264 Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq
1265 aligner. *Bioinformatics* 29, 15–21.
- 1266 105. Denoeud, F., Aury, J.M., Da Silva, C., Noel, B., Rogier, O., Delledonne, M.,
1267 Morgante, M., Valle, G., Wincker, P., Scarpelli, C., et al. (2008). Annotating
1268 genomes with massive-scale RNA sequencing. *Genome Biol.* 9, 1–12.
- 1269 106. Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics* 5, 1–9.
- 1270 107. Boyer, T.P., Antonov, J.I., Baranova, O.K., Coleman, C., Garcia, H.E., Grodsky, A.,
1271 Johnson, D.R., Locarnini, R. a, Mishonov, A. V, O'Brien, T.D., et al. (2013).
1272 WORLD OCEAN DATABASE 2013, NOAA Atlas NESDIS 72. Sydney Levitus, Ed.;
1273 Alexey Mishonoc, Tech. Ed.
- 1274 108. Aumont, O., Ethé, C., Tagliabue, A., Bopp, L., and Gehlen, M. (2015). PISCES-v2:
1275 An ocean biogeochemical model for carbon and ecosystem studies. *Geosci.*
1276 *Model Dev.* 8, 2465–2513.
- 1277 109. Da Cunha, V., Gaia, M., Nasir, A., and Forterre, P. (2018). Asgard archaea do not
1278 close the debate about the universal tree of life topology. *PLOS Genet.* 14,
1279 e1007215.
- 1280 110. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and
1281 Madden, T.L. (2009). BLAST+: Architecture and applications. *BMC*
1282 *Bioinformatics* 10, 1–9.

Genome-wide functional classification of unicellular eukaryotic plankton

- 1283 111. Katoh, K., and Standley, D.M. (2013). MAFFT Multiple Sequence Alignment
1284 Software Version 7: Improvements in Performance and Usability. *Mol. Biol.*
1285 *Evol.* 30, 772–780.
- 1286 112. Nguyen, L.T., Schmidt, H.A., Von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE:
1287 A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood
1288 Phylogenies. *Mol. Biol. Evol.* 32, 268–274.
- 1289 113. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., Von Haeseler, A., and Jermini,
1290 L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic
1291 estimates. *Nat. Methods* 2017 146 14, 587–589.
- 1292 114. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel,
1293 O. (2010). New algorithms and methods to estimate maximum-likelihood
1294 phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* 59, 307–321.
- 1295 115. Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q., and Vinh, L.S. (2018).
1296 UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Mol. Biol. Evol.*
1297 35, 518–522.
- 1298 116. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein
1299 alignment using DIAMOND. *Nat. Methods* 12, 59–60.
- 1300

Tara Oceans Coordinators

1301
1302
1303 Shinichi Sunagawa¹, Silvia G. Acinas², Peer Bork^{3,4,5}, Eric Karsenti^{6,7,11}, Chris Bowler^{6,7},
1304 Christian Sardet^{7,9}, Lars Stemmann^{7,9}, Colomban de Vargas^{7,19}, Patrick Wincker^{7,18}, Magali
1305 Lescot^{7,26}, Marcel Babin^{7,20}, Gabriel Gorsky^{7,9}, Nigel Grimsley^{7,24,25}, Lionel Guidi^{7,9}, Pascal
1306 Hingamp^{7,26}, Olivier Jaillon^{7,18}, Stefanie Kandels^{3,7}, Daniele Iudicone¹⁰, Hiroyuki Ogata¹²,
1307 Stéphane Pesant^{13,14}, Matthew B. Sullivan^{15,16,17}, Fabrice Not¹⁹, Lee Karp- Boss²¹, Emmanuel
1308 Boss²¹, Guy Cochrane²², Michael Follows²³, Nicole Poulton²⁷, Jeroen Raes^{28,29,30}, Mike
1309 Sieracki²⁷ and Sabrina Speich^{31,32}.

1310

1311 ¹ Department of Biology, institute of Microbiology and swiss institute of Bioinformatics, etH
1312 Zürich, Zürich, switzerland.

1313 ² Department of Marine Biology and Oceanography, institute of Marine sciences–CsiC,
1314 Barcelona, spain.

1315 ³ Structural and Computational Biology, european Molecular Biology Laboratory,
1316 Heidelberg, Germany.

1317 ⁴ Max Delbrück Center for Molecular Medicine, Berlin, Germany.

1318 ⁵ Department of Bioinformatics, Biocenter, university of würzburg, würzburg, Germany.

1319 ⁶ Institut de Biologie de l'ENS, Département de Biologie, École Normale supérieure,
1320 CNRS, INSERM, Université PSL, Paris, France.

1321 ⁷ Research Federation for the study of Global Ocean systems ecology and evolution,
1322 Fr2022/tara G0see, Paris, France.

1323 ⁸ Université de Nantes, CNRS, uMr6004, Ls2N, Nantes, France.

1324 ⁹ Sorbonne université, CNRS, Laboratoire d'Océanographie de Villefranche,
1325 villefranche- sur- Mer, France.

1326 ¹⁰ Stazione Zoologica anton Dohrn, Naples, Italy.

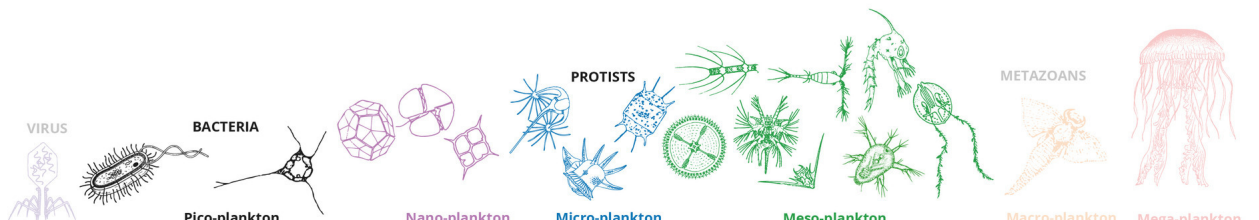
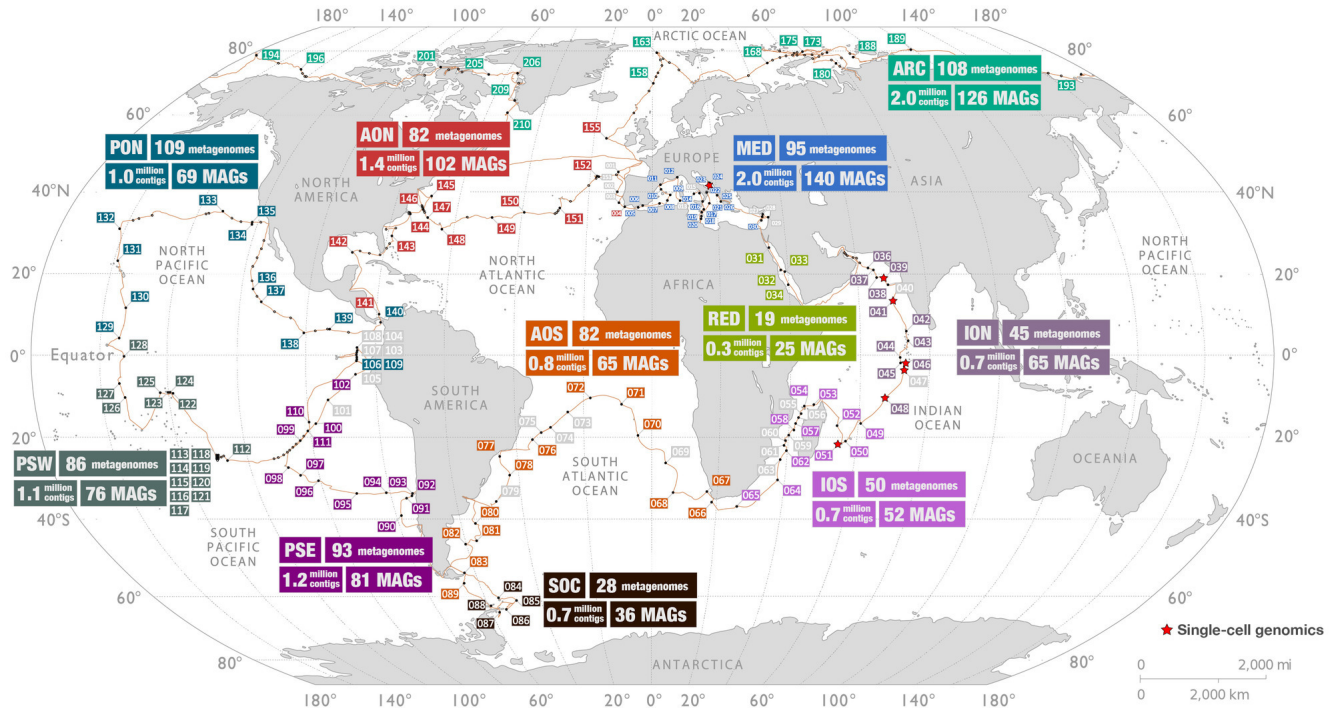
1327 ¹¹ Directors' research, European Molecular Biology Laboratory, Heidelberg, Germany.

1328 ¹² institute for Chemical research, Kyoto university, Kyoto, Japan.

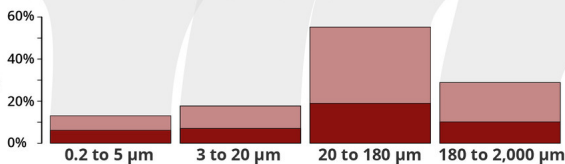
1329 ¹³ PaNGaea, university of Bremen, Bremen, Germany.

Genome-wide functional classification of unicellular eukaryotic plankton

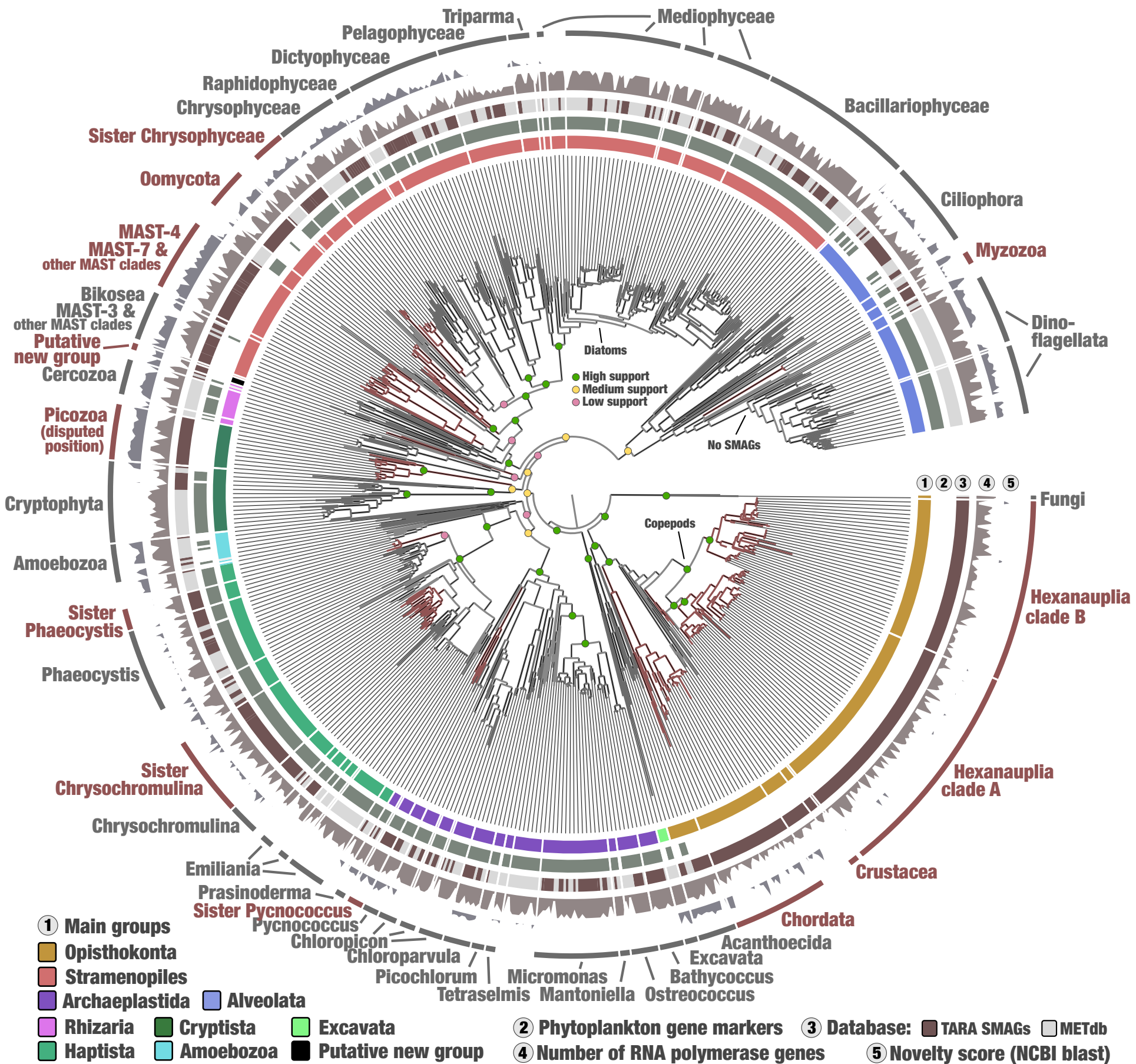
- 1330 ¹⁴ MaruM, Center for Marine environmental sciences, university of Bremen, Bremen,
1331 Germany.
- 1332 ¹⁵ Department of Microbiology, the Ohio state university, Columbus, OH, USA.
- 1333 ¹⁶ Department of Civil, environmental and Geodetic engineering, the Ohio state
1334 university, Columbus, OH, USA.
- 1335 ¹⁷ Center for RNA Biology, the Ohio state university, Columbus, OH, USA.
- 1336 ¹⁸ Génomique Métabolique, Genoscope, institut de Biologie Francois Jacob, Commissariat à
1337 l'Énergie atomique, CNRS, université evry, université Paris- saclay, evry, France.
- 1338 ¹⁹ Sorbonne université and CNRS, UMR 7144 (AD2M), ECOMAP, station Biologique
1339 de Roscoff, Roscoff, France.
- 1340 ²⁰ Département de Biologie, Québec Océan and Takuvik Joint International Laboratory (UMI
1341 3376), Université Laval (Canada)–CNRS (France), Université Laval, Quebec, QC, Canada.
- 1342 ²¹ School of Marine Sciences, University of Maine, Orono, ME, USA. ²²European Molecular
1343 Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus,
1344 Hinxton, Cambridge, UK.
- 1345 ²³ Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of
1346 Technology, Cambridge, MA, USA.
- 1347 ²⁴ CNRS UMR 7232, Biologie Intégrative des Organismes Marins, Banyuls- sur- Mer, France.
- 1348 ²⁵ Sorbonne Universités Paris 06, OOB UPMC, Banyuls- sur- Mer, France.
- 1349 ²⁶ Aix Marseille Universit/e, Université de Toulon, CNRS, IRD, MIO UM 110, Marseille,
1350 France.
- 1351 ²⁷ Bigelow Laboratory for Ocean Sciences, East Boothbay, ME, USA.
- 1352 ²⁸ Department of Microbiology and Immunology, Rega Institute, KU Leuven, Leuven,
1353 Belgium.
- 1354 ²⁹ Center for the Biology of Disease, VIB KU Leuven, Leuven, Belgium.
- 1355 ³⁰ Department of Applied Biological Sciences, Vrije Universiteit Brussel, Brussels, Belgium.
- 1356 ³¹ Department of Geosciences, Laboratoire de Météorologie Dynamique, École Normale
1357 Supérieure, Paris, France.
- 1358 ³² Ocean Physics Laboratory, University of Western Brittany, Brest, France.

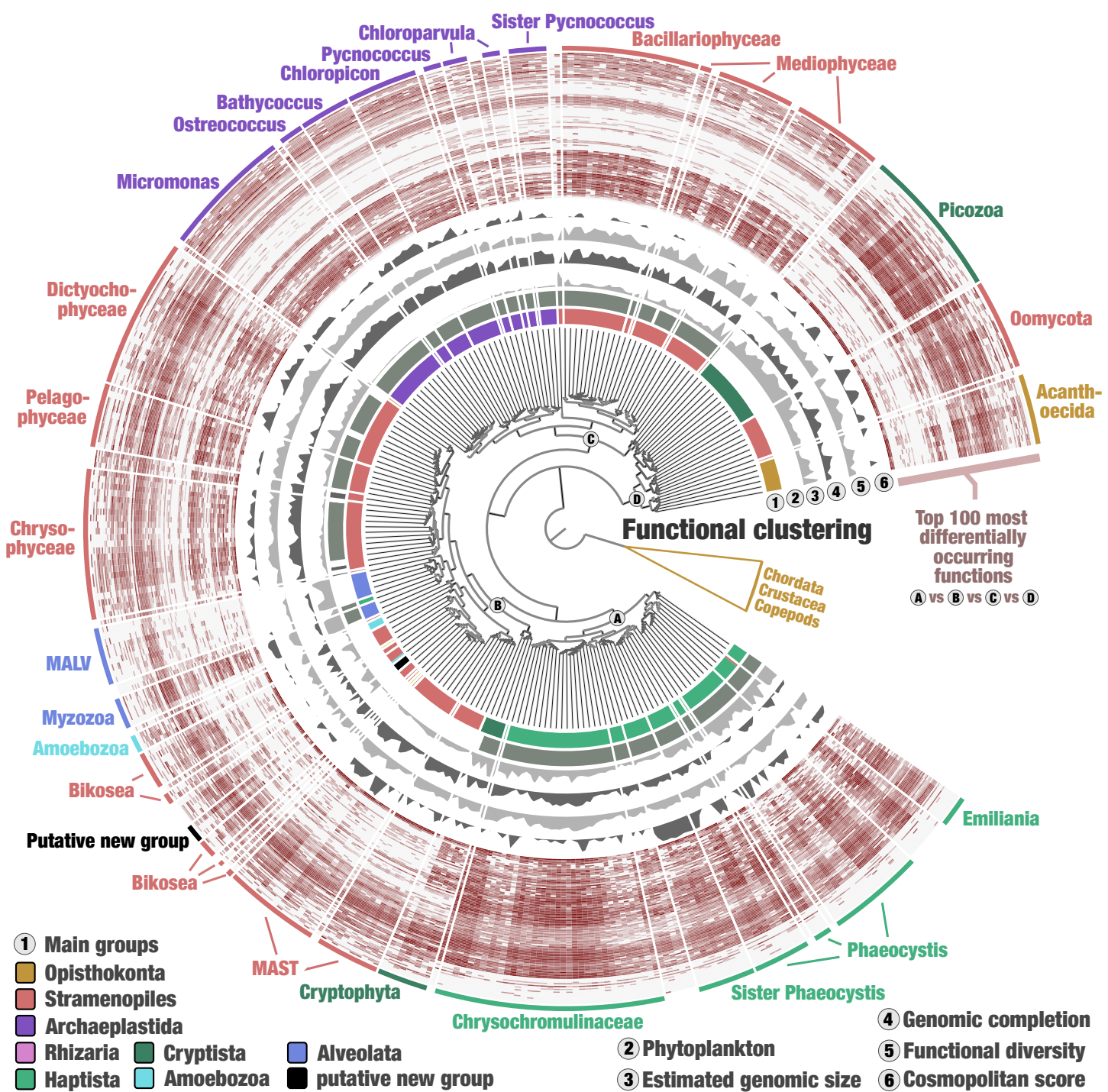


Percentage of mapped reads



■ SMAGs (mapped reads)
■ Complete SMAGs



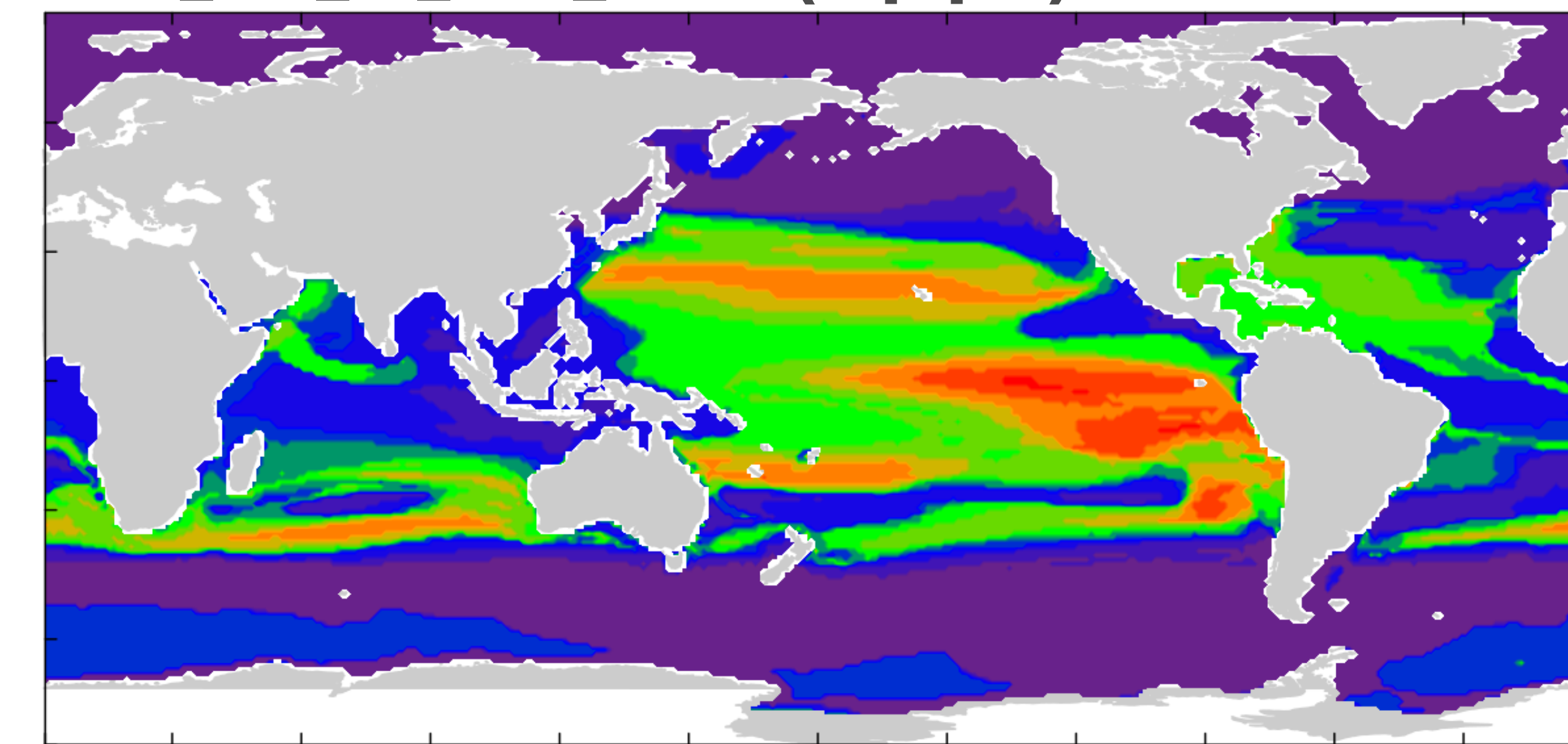
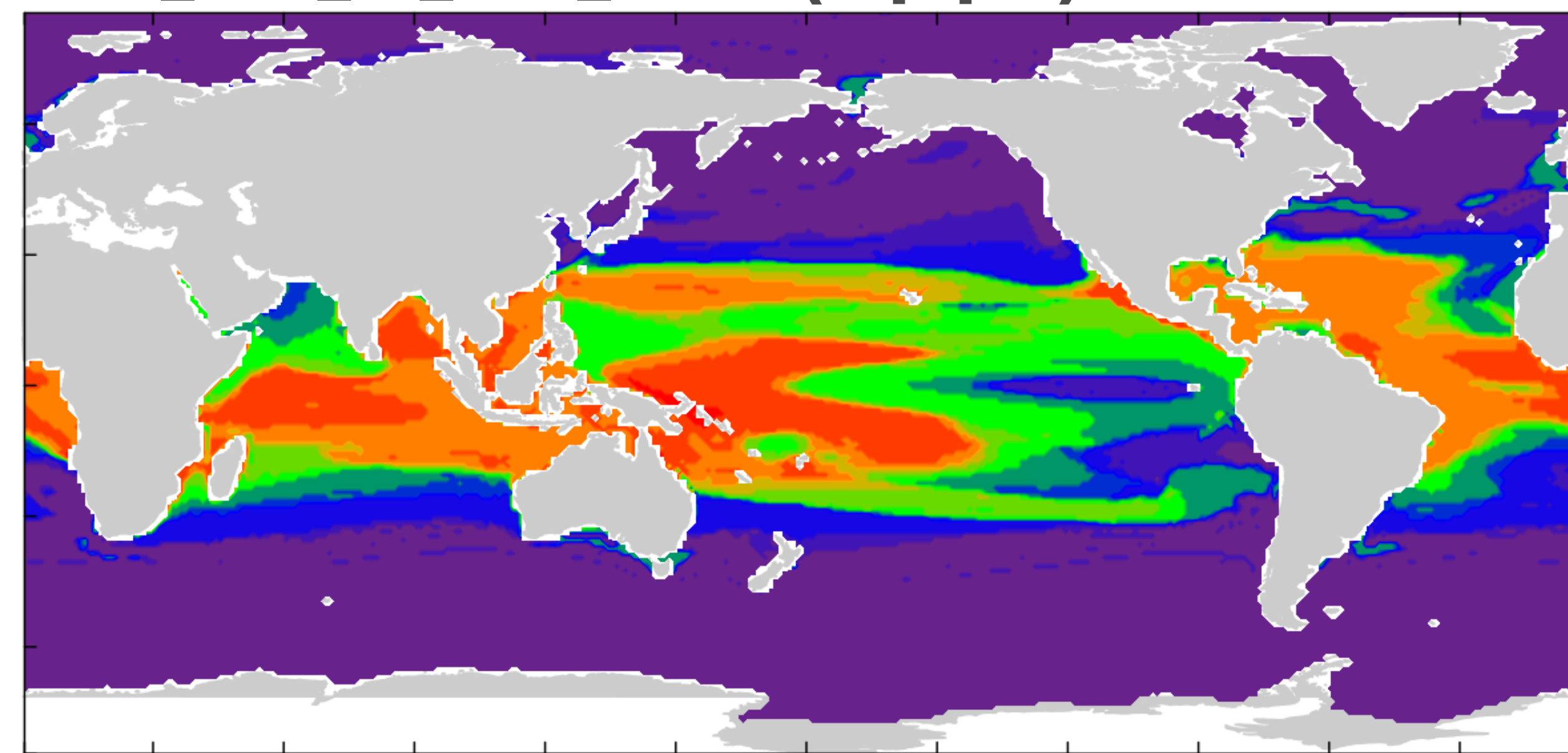
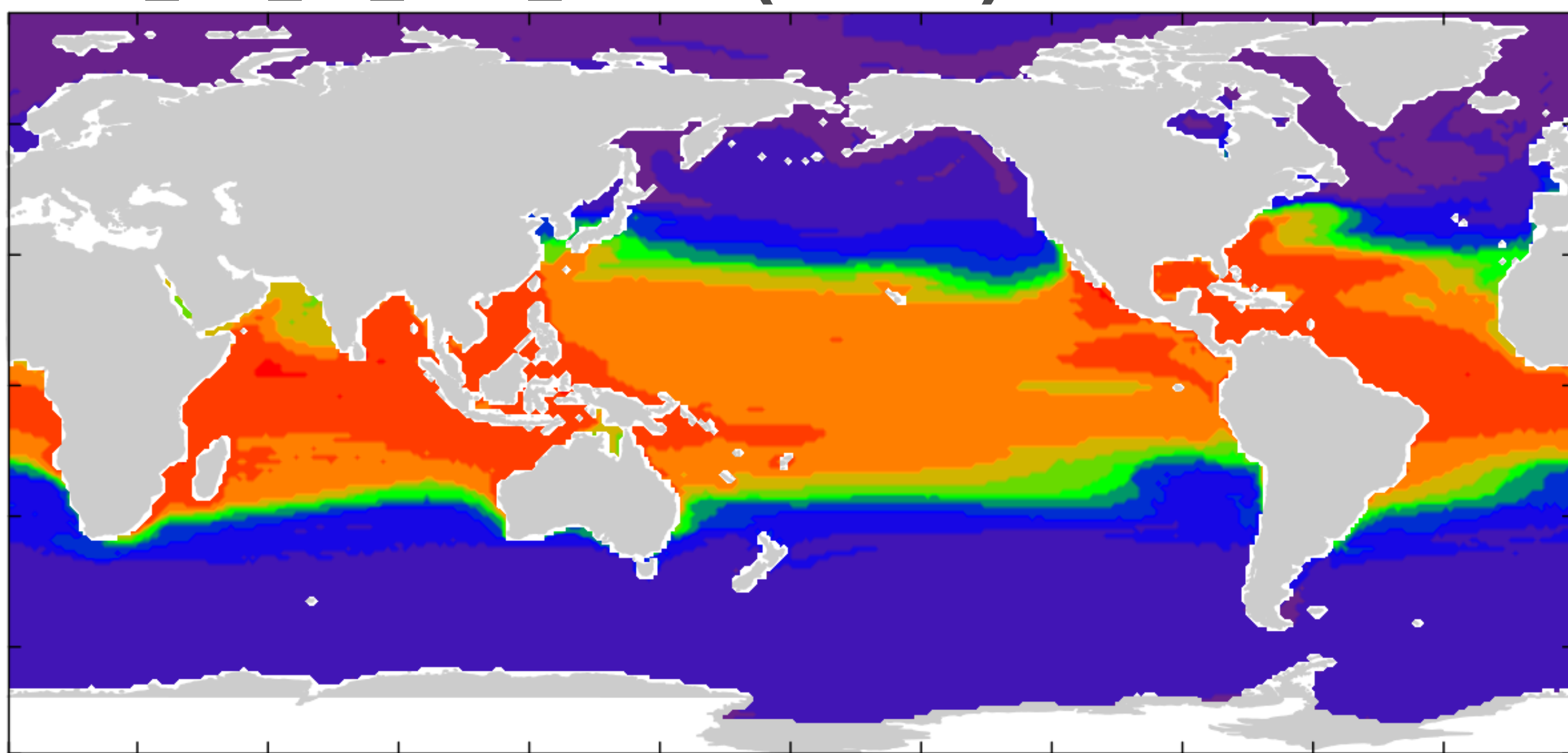


TARA IOS 50_MAG_00098 (MAST-4)

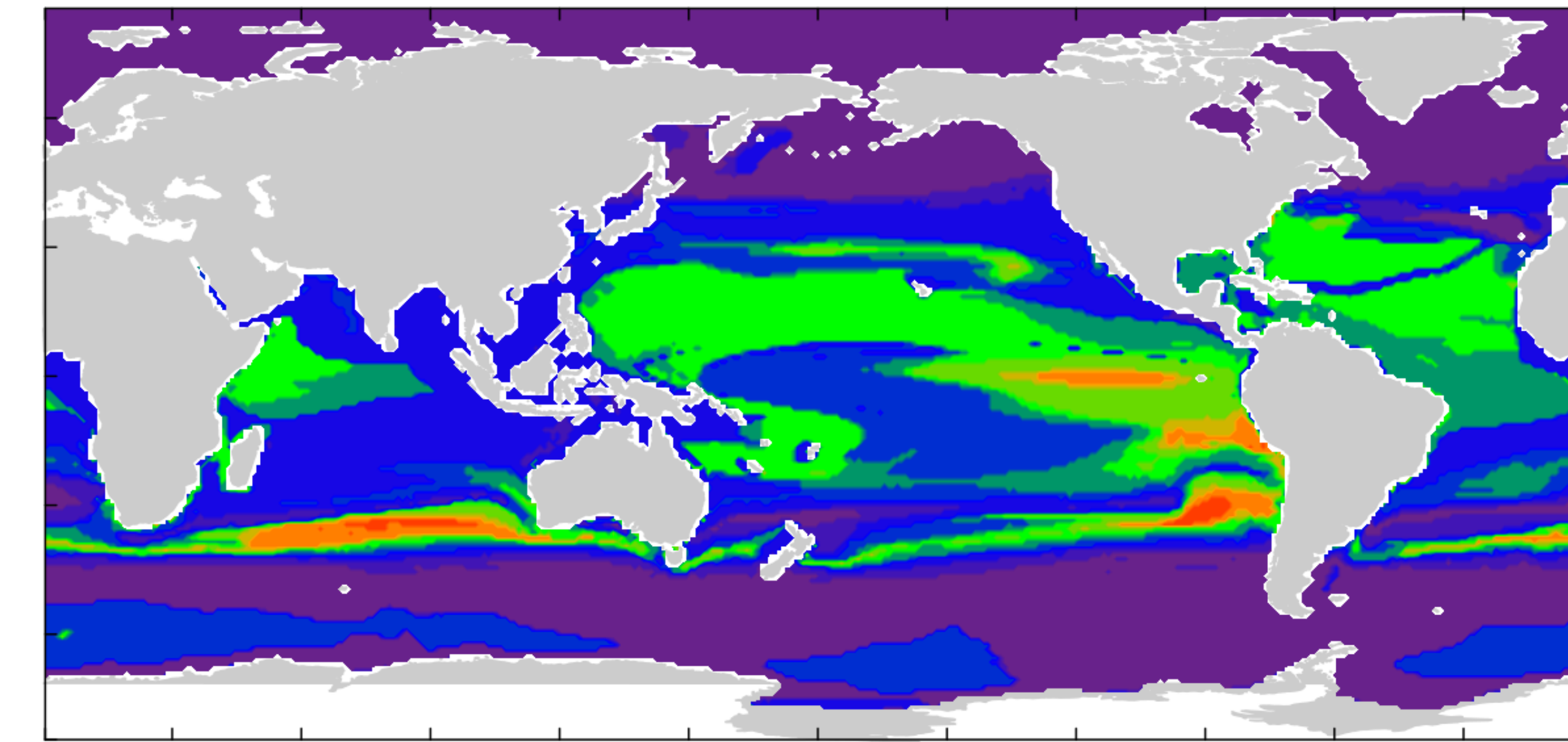
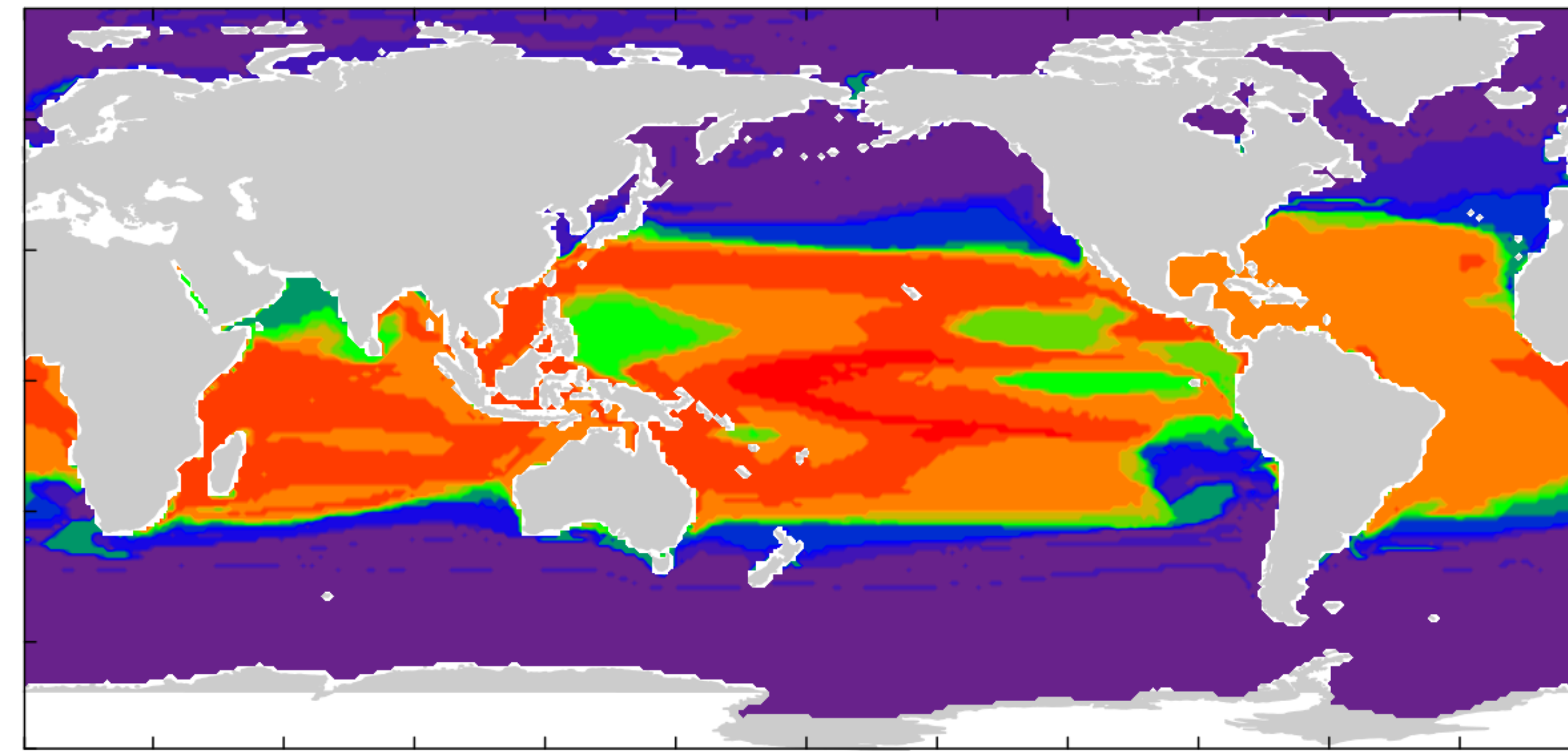
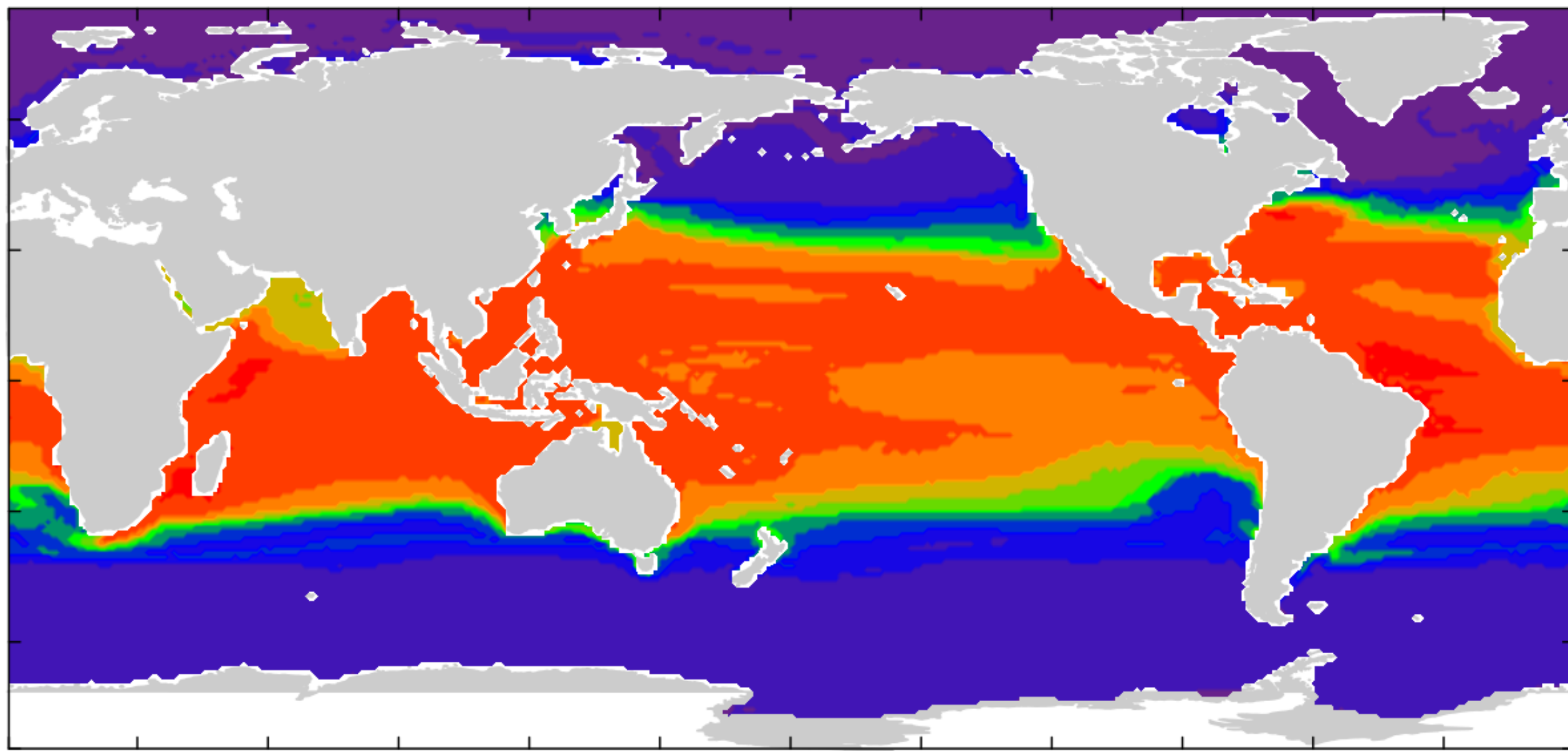
TARA PSW_86_MAG_00299 (Copepod)

TARA PSE_93_MAG_00246 (Copepod)

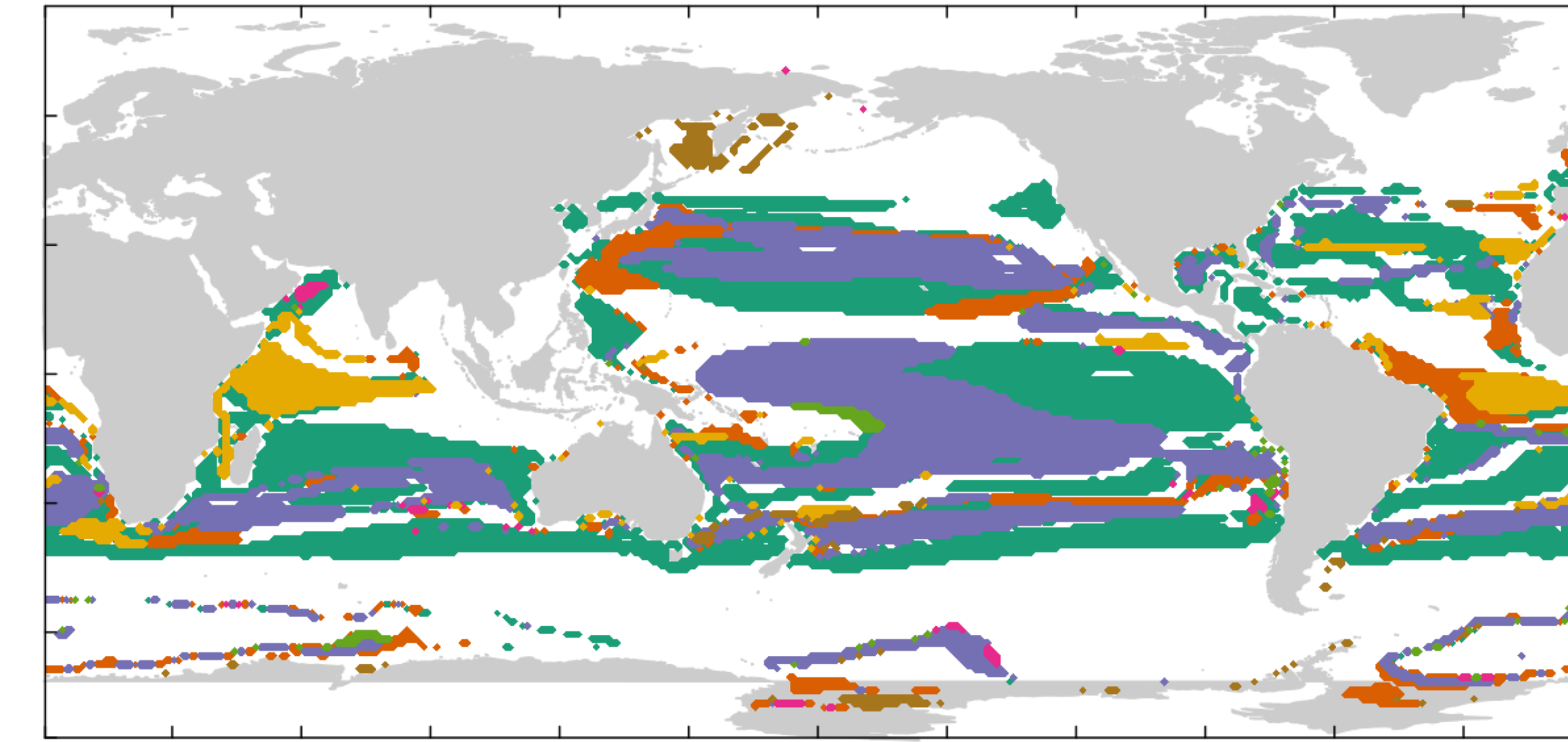
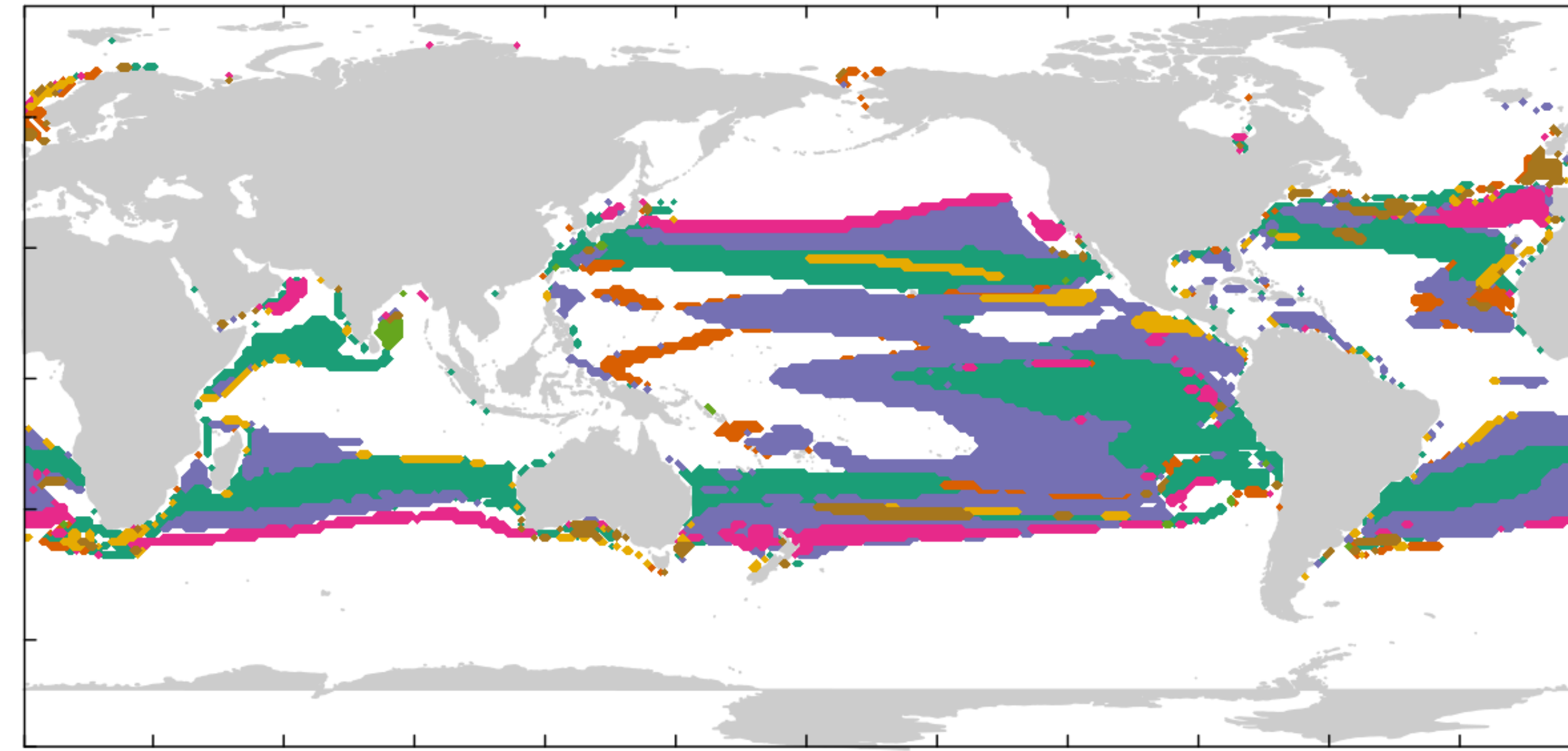
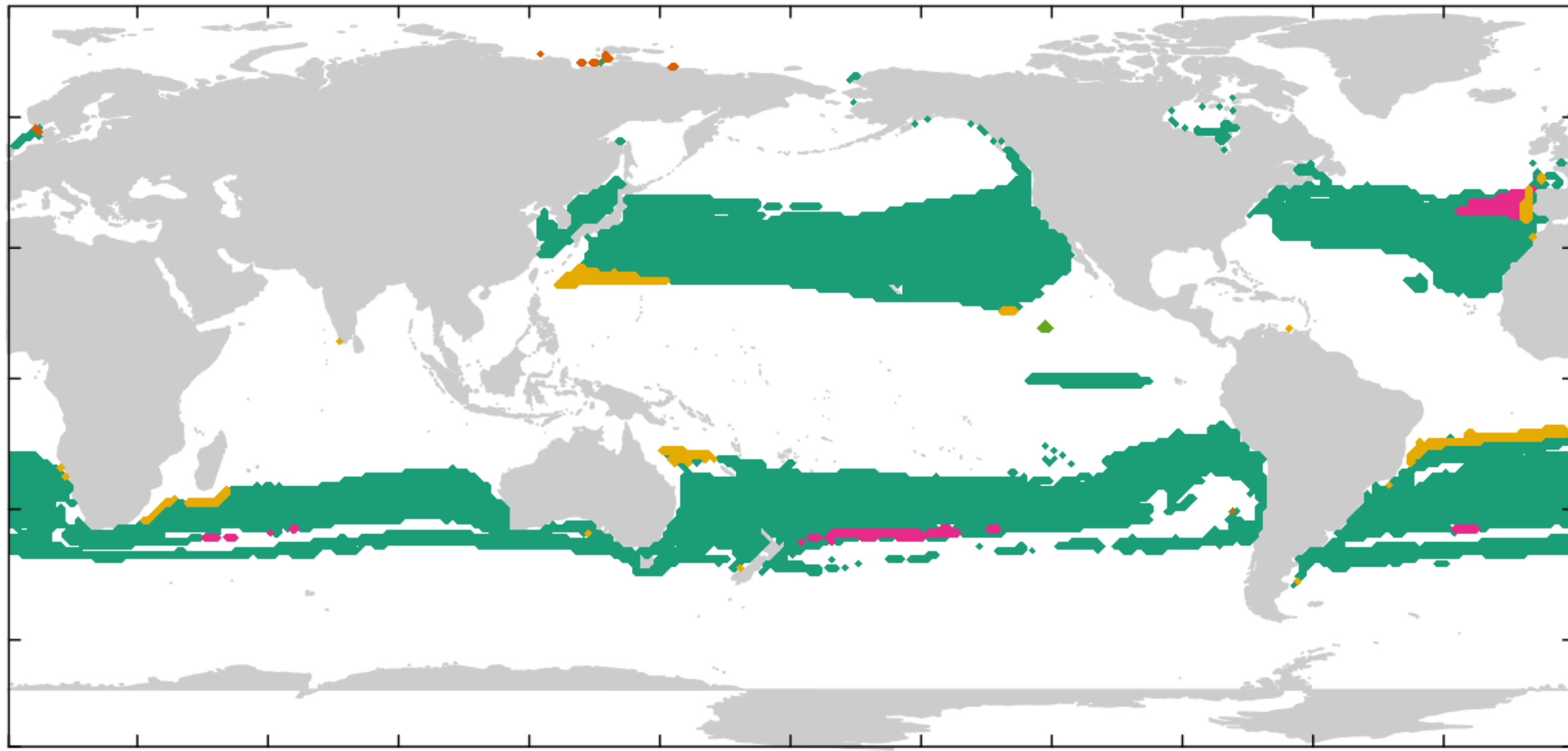
Years 2006 - 2015



Years 2090 - 2099



Environmental parameters driving the projected shift



SI NO₃
 Fe μmol/L
 PO₄ μmol/L
 NO₃ μmol/L
 SI μmol/L
 Sal g/kg
 Temperature °C

1 Genome-resolved metagenomics applied to marine eukaryotes

2 ~700 environmental genome up to 1.3 Gbp in length

3 10 million eukaryotic genes characterized

4 Functional convergences between distantly related lineages

Functional clustering of marine eukaryotes

