Mixotrophic protists display contrasted biogeographies in the global ocean

Emile Faure, Fabrice Not, Anne-Sophie Benoiston, Karine Labadie, Lucie Bittner, Sakina-Dorothée Ayata

Title: Mixotrophic protists display contrasted biogeographies in the global ocean

Running title: Global biogeography of marine mixotrophic protists

Authors : Emile Faure [1,2,*], Fabrice Not [3], Anne-Sophie Benoiston [2], Karine Labadie[4], Lucie Bittner[2,†], and Sakina-Dorothée Ayata [1,†]

[1]*Sorbonne Université, CNRS, Laboratoire d'océanographie de Villefranche, LOV, F-06230 Villefranche-sur-Mer, France.*

[2]*Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, CNRS, Sorbonne Université, EPHE, CP 50, 57 rue Cuvier, 75005 Paris, France*

[3]*Sorbonne Université, CNRS, UMR7144 Adaptation and Diversity in Marine Environment (AD2M) laboratory, Ecology of Marine Plankton team, Station Biologique de Roscoff, Place Georges Teissier, 29680 Roscoff, France*

[4]*Commissariat à l'Energie Atomique (CEA), Institut de Biologie François Jacob, Genoscope, F-91057 Evry, France.*

[*]*Corresponding author* [†]*Co-senior authors*

Competing interests: The authors declare that they have no competing interests.

Abstract

Mixotrophy, or the ability to acquire carbon from both auto- and heterotrophy, is a widespread ecological trait in marine protists. Using a metabarcoding dataset of marine plankton from the global ocean, 318 054 mixotrophic metabarcodes represented by 89 951 866 sequences and belonging to 133 taxonomic lineages were identified and classified into four mixotrophic functional types: constitutive mixotrophs (CM), generalist non-constitutive mixotrophs (GNCM), endo-symbiotic specialist non-constitutive mixotrophs (eSNCM) and plastidic specialist non-constitutive mixotrophs (pSNCM). Mixotrophy appeared ubiquitous, and the distributions of the four mixotypes were analyzed to identify the abiotic factors shaping their biogeographies. Kleptoplastidic mixotrophs (GNCM & pSNCM) were detected in new zones compared to previous morphological studies. Constitutive and non-constitutive mixotrophs had similar ranges of distributions. Most lineages were evenly found in the samples, yet some of them displayed strongly contrasted distributions, both across and within mixotypes. Particularly divergent biogeographies were found within endo-symbiotic mixotrophs, depending on the ability to form colonies or the mode of symbiosis. We showed how metabarcoding can be used in a complementary way with previous morphological observations to study the biogeography of mixotrophic protists and

to identify key drivers of their biogeography.

Introduction

Marine unicellular eukaryotes, or protists, have a tremendous range of life styles, sizes and forms [1], showing a taxonomic and functional diversity that remains hard to define [2, 3]. This variety of organisms is having an impact on major biogeochemical cycles such as carbon, oxygen, nitrogen, sulfur, silica, or iron, while being at the base of marine trophic networks [4–8]. Hence, they are key actors of the global functioning of the ocean.

Historically, marine protists have been classified into two groups depending on their trophic strategy: the photosynthetic plankton (phytoplankton) and the heterotrophic plankton (zooplankton). It is now clear that mixotrophy, *i.e.* the ability to combine autotrophy and heterotrophy, has been largely underestimated and is commonly found in planktonic protists [6, 9–13]. Instead of a dichotomy between two trophic types, their trophic regime should be regarded as a continuum between full phototrophy and full heterotrophy, with species from many planktonic lineages lying between these two extremes [10]. Mitra et al. [11] have proposed a classification of marine mixotrophic protists into four functional groups, or mixotypes. The constitutive mixotrophs, or CM, are photosynthetic organisms that

60 are capable of phagotrophy, also called ”phytoplankton that eat” [11]. They include
61 most mixotrophic nanoflagellates (*e.g. Prymnesium parvum*, *Karlodinium micrum*).
62 On the opposite, the non-constitutive mixotrophs, or “photosynthetic zooplankton”,
63 are heterotrophic organisms that have developed the ability to acquire energy
64 through photosynthesis [9]. This ability can be acquired in three different ways: the
65 generalist non-constitutive mixotrophs (GNCM) steal the chloroplasts of their prey,
66 such as most plastid-retaining oligotrich ciliates (*e.g. Laboea strobila*), the plastidic
67 specialist non-constitutive mixotrophs (pSNCM) steal the chloroplasts of a specific
68 type of prey (*e.g. Mesodinium rubrum* or *Dinophysis* spp.), and finally the endo-
69 symbiotic specialist non-constitutive mixotrophs (eSNCM) are bearing
70 photosynthetically active endo-symbionts (most mixotrophic Rhizaria from
71 Collodaria, Acantharea, Polycystinea, and Foraminifera, as well as dinoflagellates
72 like *Noctiluca scintillans*).
73 As drivers of biogeochemical cycles in the global ocean, and particularly of the
74 biological carbon pump [5, 14, 15], marine protists are a key part of ocean
75 biogeochemical models [7, 16–18]. However, physiological details of mixotrophic
76 energy acquisition strategies have only been studied in a restricted number of
77 lineages [9, 19, 20]. They appear to be quite complex and greatly differ across
78 mixotypes, which makes mixotrophy hard to include in a simple model structure
79 [21–25]. Hence at this time, mixotrophy is not included in most biogeochemical

models, neglecting the amount of carbon fixed by non-constitutive mixotrophs through photosynthesis, and missing the population dynamics of photosynthetically active constitutive mixotrophs that can still grow under nutrient limitation [23, 26]. This is most probably skewing climatic models predictions [11, 26], as well as our ability to understand and prevent future effects of global change.

A better understanding of the environmental diversity of marine mixotrophic protists, as well as a description of the abiotic factors driving their biogeography at global scale are still needed, in particular to integrate them in biogeochemical models. Leles et al. [27] attempted to tackle this problem by reviewing about 110 000 morphological identification records of a set of more than 60 mixotrophic protists species in the ocean, taken from the Ocean Biogeographic Information System (OBIS) database. They found distinctive patterns in the biogeography of the three different non-constitutive mixotypes (GNCM, pSNCM and eSNCM), highlighting the need to better understand such diverging distributions [27]. Environmental molecular biodiversity surveys through metabarcoding have been widely used in the past fifteen years to decipher planktonic taxonomic diversity [2, 28–30]. Here we exploited the global *Tara* Oceans datasets [31–33], and identified 133 mixotrophic lineages, that we classified into the four mixotypes defined by Mitra et al. [11]. This first ever set of mixotrophic metabarcodes allowed us to investigate the global biogeography of both constitutive and non-constitutive

100 mixotrophs, in relation with *in-situ* abiotic measurements. We tested (i) if new
101 information on marine mixotrophic protists distribution can be gained in
102 comparison with previous morphological identifications [27]; (ii) if the constitutive
103 mixotrophs, which are not addressed in Leles et al. [27], and the non-constitutive
104 mixotrophs diverge in terms of biogeography; (iii) if the study of diversity and
105 abundance of environmental metabarcodes could lead to the definition of key
106 environmental factors shaping mixotrophic communities.
107
108 **Materials and Methods**
109
110 *Samples collection and dataset creation*
111 Metabarcoding datasets from the worldwide *Tara* Oceans sampling campaigns that
112 took place between 2009 and 2013 [31, 33] (data published in open access at the
113 European Nucleotide Archive under project accession number PRJEB6610) were
114 investigated. We analyzed 659 samples from 122 distinct stations, and for each
115 sample, the V9-18S ribosomal DNA region was sequenced through Illumina HiSeq
116 [32]. Assembled and filtered V9 metabarcodes (cf. details in de Vargas et al. [2])
117 were assigned to the lowest taxonomic rank possible *via* the Protist Ribosomal
118 Reference (PR2) database [34]. To limit false positives, we chose to only analyze
119 the metabarcodes (*i.e.* unique versions of V9 sequences) for which the assignment

to a reference sequence had been achieved with a similarity of 95% or higher. This represents 65% of the total dataset in terms of metabarcodes and 84% in terms of total sequences. Our dataset involved 1,492,912,215 sequences, distributed into 4,099,567 metabarcodes assigned to 5,071 different taxonomic assignations, going from species to kingdom level precision.

Defining a set of mixotrophic organisms

Among these 5,071 taxonomic assignations, we searched for mixotrophic protist lineages, taking into account the 4 mixotypes described by Mitra et al. [11]: constitutive mixotrophs (CM), generalist non-constitutive mixotrophs (GNCM), endo-symbiotic specialist non-constitutive mixotrophs (eSNCM), and plastidic specialist non-constitutive mixotrophs (pSNCM). We used the table S2 from Leles et al. [27] which is referencing 71 species or genera belonging to three non-constitutive mixotypes (GNCM, pSNCM and eSNCM), as well as multiple other sources coming from the recent literature on mixotrophy [6, 9–12, 35–47], and inputs from mixotrophic protists' taxonomy specialists (cf. Acknowledgments section). Within the 5,071 taxonomic assignations of variable precisions, we identified 5 GNCM, 9 pSNCM, 77 eSNCM, and 42 CM lineages (detailed list available publicly under the DOI 10.6084/m9.figshare.6715754, and all metabarcodes were tagged with their mixotypes in the PR2 database). Among these

133 taxonomic assignations that we will call “lineages”, 92 were defined at the species level, 119 at the genus level, and the last 14 at higher taxonomic levels where mixotrophy is always present (mostly eSNCM groups like Collodaria). In the Chrysophyceae family, metabarcodes assigned to clades B2, E, G, H and I were included even though we couldn’t find a general proof that all species included in these clades have mixotrophic capabilities. However, if we exclude the photolithophic Synurophyceae and genera like *Paraphysomonas* and *Spumella*, which we did, a vast majority of Chrysophyceae are considered mixotrophic [10]. The final dataset included 318 054 metabarcodes assigned to the 133 mixotrophic lineages selected, as well as their sequence abundance in 659 samples (table available publicly under the DOI 10.6084/m9.figshare.6715754).

Environmental dataset

We built a corresponding contextual dataset using the environmental variables available in the PANGAEA repository from the *Tara* Oceans expeditions [33, 49]. The set of 235 environmental variables was reduced to 57 due to several selection steps (Data available publicly under the DOI 10.6084/m9.figshare.6715754; see the details of variable selection in section 1 of Supp. Mat.).

Distribution and diversity of mixotrophic protists

For each mixotype, the number of metabarcodes, the total sequence abundance and the mean sequence abundance by metabarcode was computed (Table 1). Also, we measured each metabarcode's station occupancy, *i.e.* the number of stations in which it was found, and station evenness, *i.e.* the homogeneity of its distribution among the stations in which it was detected (Figure 2). Diversity of mixotrophic protists was investigated through mixotype-specific metabarcode richness per station (Table 1). As the number of samples taken per station can impact the abundance and diversity of detected metabarcodes, richness was computed only at stations for which the maximum number of 8 samples were available (40 stations over 122).

Global biogeography of mixotrophic protists

Two statistical analyses were performed to investigate mixotrophic protists biogeography. One at the metabarcode level, and one at the lineage level, *i.e.* merging the sequence abundance of metabarcodes sharing the same taxonomical assignation. The metabarcodes abundance table was composed of 318054 rows/metabarcodes, and 659 columns/samples, whereas the lineage abundance table was composed of 133 rows/lineages and 659 columns/samples (both datasets are available publicly under the DOI 10.6084/m9.figshare.6715754). The two analyses led to very similar conclusions, but the biogeography of lineages appeared

easier to visually represent and interpret than the one of metabarcodes. Hence, we only present here the results of the lineage-based analysis (See section 3 of Sup. Mat. for metabarcode-level analysis results and discussion).

Our statistical model was designed to identify lineages (or metabarcodes) with contrasted biogeographies, and relate their presence to the environmental context. We normalized the sequence counts from the lineage abundance matrix using a Hellinger transformation [51]. We used the environmental dataset and the mixotrophic lineages' abundance matrix as explanatory and response matrices, respectively, to conduct a redundancy analysis (RDA) [51]. For that, we made a species pre-selection using Escoufier's vectors [52] which allowed to keep only the 62 most significant mixotrophic lineages. This method selects lineages according to a principal component analysis (PCA), sorting them based on their correlation to the principal axes. We then used a maximum model (Y~X) and a null model (Y~1) to conduct a two directional stepwise model selection based on the Akaike information criterion (AIC) [53]. The resulting model contained 28 environmental response variables. More details about statistical analyses are available in section 2 and 3 of the supplementary materials. Analyses and graphs were realized with the R software version 3.4.3 [54]. All scripts are available on GitHub platform (https://github.com/upmcgenomics/MixoBioGeo).

Results

Global distribution and diversity of marine mixotrophic protists

Mixotrophic protists metabarcodes were detected in all the 659 samples with a total sequence abundance of 89 951 866, representing 12.56% of the total sequence abundance in the 659 samples studied. They represented a mean of 12.64% of the total sequence abundance per sample, with a maximum of 96.96% and a minimum of 0.01%. To avoid any potential overestimation of mixotrophic lineages presence in the following results, we marked all records of less than a hundred sequences as questionable. We found both eSNCM and CM in each of the 122 stations studied (Table 1, Figure 1). In only two occasions the number of sequences belonging to CM was questionable, at stations for which only one sample was sequenced. GNCM were found absent in only 2 stations and their presence was questionable in 39 stations (Figure 1). pSNCM were absent at 5 stations (3 in the Indian Ocean, and 2 in the Pacific Ocean) and detected with questionable presence in 54 additional stations, which were mostly located in the central Pacific and the Indian Ocean (Figure 1). We found significant amounts of sequences corresponding to GNCM in the Central Pacific, Southern subtropical Atlantic, and Indian Ocean. The presence of GNCM in these areas has not yet been recorded through morphological identifications during field expeditions [27]. Also, we detected more than 100

220 sequences of pSNCM metabarcodes at 11 stations belonging to biogeographical
221 provinces in which no morphological identifications had been published [27, 55],
222 mostly in offshore areas of the Atlantic and Pacific Ocean (Figure 1).
223
224 The mean evenness of mixotrophic metabarcodes across stations was of 0.87, and
225 82.3% of the metabarcodes had a station evenness above 0.5 (Figure 2). Station
226 occupancy varied a lot depending on the metabarcodes, with a high density of rare
227 metabarcodes leading to a mean of 5.14 stations over a maximum of 122, and a
228 standard deviation of 7.7. However, three eSNCM metabarcodes were found in all
229 the 122 stations, and three CM metabarcodes were detected in 121 stations. The
230 maximum occupancy for a GNCM metabarcode was of 111 stations, while 92
231 stations was the maximum for a pSNCM metabarcode. CM and GNCM
232 metabarcodes showed a strong tendency towards high evenness values (Figure 2,
233 means of 0.90 and 0.95, respectively), even for the most sequence abundant
234 metabarcodes. Many eSNCM metabarcodes had high evenness values, but below
235 average values were detected for the most abundant ones (Figure 2, global mean of
236 0.87). pSNCM metabarcodes had a similar mean of evenness values (0.87), but a
237 different distribution compared to other mixotypes (Figure 2). Among the 50 most
238 abundant metabarcodes, 43 corresponded to Collodaria lineages, 47 were eSNCM

239 and 3 were CM, all three assigned to *Gonyaulax polygramma*. GNCM and pSNCM
240 metabarcodes had homogeneously low sequence abundances (Figure 2, Table 1).
241
242 *Main factors affecting the biogeography of mixotrophic protists*
243 The redundancy analysis helped to investigate further the environmental variables
244 responsible for the mixotrophic protists' biogeography. The 62 lineages selected
245 with the Escoufier's vector method corresponded to 20 CM, 34 eSNCM, 3 GNCM
246 and 5 pSNCM. Even after selection, a significant part of the lineages did not show
247 any response to environmental data in their distribution (Figure 3, ***e.g.*** 19 of the 62
248 lineages were found between -0.01 and 0.01 on both RDA1 and RDA2). The
249 adjusted R-squared of the RDA was of 34.89% (41.43% unadjusted), with 24.01%
250 of variance explained on the two first axes (Figure 3). The first RDA axis (14.96%)
251 marks an opposition between samples from oligotrophic waters with low
252 productivity (RDA1>0) and samples from eutrophic and productive water masses
253 (RDA1<0). This axis is negatively correlated to chlorophyll concentration, particles
254 density, ammonium concentration, absorption coefficient of colored dissolved
255 organic matter (acCDOM), duration of daylight, silica, CO3, oxygen, and PO4
256 concentration, as well as longitude. It is positively correlated to bathymetry, deep
257 euphotic zone, deep oxygen maximum, deep mixed layer, as well as to the distance
258 to coast. The second RDA axis (9.05%) is opposing offshore and subpolar samples

(RDA2>0) to coastal and subtropical ones (RDA2<0). The axis is positively correlated to the depth of the mixed layer, the distance to coast, the bathymetry, high maximum Lyapunov exponents as well as high concentrations of PO4, oxygen, CO3 and silica. It is negatively correlated to temperature, salinity and photosynthetically active radiations (PAR).

Among the 20 CM lineages, 7 clearly emerged from the redundancy analysis (Figure 3) and showed distinct biogeographies related to environmental variables. *Gonyaulax polygramma*, *Alexandrium tamarense* and *Fragilidium mexicanum*, three Dinophyceae belonging to the Gonyaulacales order, were mainly found in oligotrophic waters with a deep euphotic zone, warm temperature, high salinity and PAR (RDA1>0, RDA2<0). The four other CMs (involving all the Chrysophyceae included in the analysis as well as one Dinophyceae from the Kareniaceae family, *Karlodinium micrum*) were found mostly in productive water masses (RDA1<0).

eSNCMs can be divided in three groups in the RDA space. The first group (RDA1<0) corresponds to eSNCM species dominating rich and productive environments. It includes mainly Acantharia and Spumellaria species. The second group (RDA1>0) dominates oligotrophic environments, and includes multiple Collodaria as well as one Dinophyceae genus (*Ornithocercus)*. Within this group,

279 *Ornithocercus* spp. is found mainly in coastal subtropical environments (RDA2<0),
280 as opposed to *Sphaerozoum punctatum* that is found mainly in offshore subpolar
281 regions (RDA2>0). *Siphonosphaera cyathina* lies between these two trends as it is
282 found only in oligotrophic samples, but isn't influenced by temperature or
283 bathymetry (Figure 3 and 4). The third group corresponds to the eSNCM lineages
284 that can be interpreted as distributed homogeneously in regards of the
285 environmental data we are using (e.g. lineages with the shortest arrows in Figure
286 3). These notably include the 12 Foraminifera lineages present in the RDA.
287 Looking at filters centroids in the RDA space (Figure 3), we can suppose that
288 eSNCM lineages dominating eutrophic systems (RDA1<0) are smaller in size than
289 those dominating oligotrophic ones (RDA1>0).
290
291 Out of the five pSNCM included in the RDA, only *Mesodinium rubrum*, the most
292 abundant one, is distinctively represented in the RDA space. This suggests that the
293 other pSNCM have homogeneous distributions in response to our environmental
294 variables. *Mesodinium rubrum* dominates eutrophic environments, independently
295 from the bathymetry or the temperature (RDA1<0, RDA2 ≈ 0). We find a similar
296 pattern for GNCM, with only *Pseudotontonia simplicidens* well represented in the
297 RDA space out of the three species included in the analysis. Like *M. rubrum*,

298 *Pseudotontonia simplicidens* is the most abundant species in its group and it is
299 mainly found in eutrophic waters (RDA1<0).
300
301 **Discussion**
302
303 *Mixotrophy occurs everywhere in the global ocean*
304 Our metabarcoding survey confirms that marine mixotrophic protists are ubiquitous
305 in the global ocean [27], possibly extending the known range of distribution of two
306 mixotypes (Figure 1 and 2). Mixotrophic organisms represented more than 12% of
307 the sequences in the complete *Tara* Oceans metabarcoding dataset, showing that
308 they should not be understated. We found contrasted biogeographies among
309 metabarcodes and their corresponding lineages, both within and across mixotypes
310 (Figure 2, 3, 4 and S1, Sup. Mat. section 3). We found constitutive mixotrophs
311 (CM) and endo-symbiotic specialist non-constitutive mixotrophs (eSNCM)
312 metabarcodes at all the 122 stations included in this global study (Table 1 and
313 Figure 2), verifying that these two mixotypes are the most abundant in the ocean
314 [27, 47, 54, 55]. This dominance of eSNCM and CM in our data is also linked to
315 the relatively high number of metabarcodes available for these two mixotypes in
316 databases. Using 1 360 generalist non-constitutive mixotrophs (GNCM)
317 metabarcodes corresponding to only 5 lineages, we detected them in 10

biogeographical provinces [55] where no morphological identification had been recorded before [27]. GNCM metabarcodes had consistently high evenness values, and some had station occupancy records comparable to the most abundant eSNCM and CM metabarcodes (Figure 2). These results support the hypothesis of a globally ubiquitous distribution of GNCM. Plastidic specialist non-constitutive mixotrophs (pSNCM) were found in 5 provinces in which no record existed so far from morphological identification field studies [27]. However, these observations were often in a questionable range in terms of sequence abundance (Figure 1), and the overall distribution of pSNCM in our data appears as very concordant with morphological observations [27]. pSNCM metabarcodes had dominantly low station evenness values, which again supports the conclusions of Leles et al. [27] that identified pSNCM as highly seasonal and spatially restricted in their distribution.

While building our set of mixotrophic lineages, some widespread and potentially mixotrophic genera did not appear, such as *Ceratium* spp., *Tontonia* spp., *Amphisolenia* spp., *Triposolenia* spp. or *Citharistes* spp., mainly because of a poor representation in the PR2 database. Also, we decided to only consider metabarcodes with more than 95% similarity to a reference sequence. This threshold could be too selective for some species and not enough for some others, as single similarity threshold are hardly efficient when studying whole eukaryotic

populations [56, 57]. For example, some species appeared with low sequence abundance in the data even though they couldn't have been sampled, such as three lacustrine species, *e.g. Poteriospumella lacustris*. Considering these biases and the sometimes relatively low sequence counts (marked as questionable in Figure 1), some of the new GNCM and pSNCM records we observed should be considered with care, as they could be over-estimated or even sometimes artefactual. However, the low number of lineages found for these two mixotypes in PR2 and in our dataset are leading us to think that we were unable to capture the whole GNCM and pSNCM communities. This supposes a global underestimation of GNCM and pSNCM abundances in our results.

Tara Oceans metabarcoding dataset is built on snapshot samples taken irregularly during a three-year cruise, hence allowing no proper seasonal variations investigations. However, morphological identifications of mixotrophic protists revealed seasonal variations in their abundance, with *Mesodinium* biomass blooming in spring in coastal seas for example [27]. As metabarcoding datasets have been successfully applied on time series to detect species successions across gradients of time and space [58–60], it would be interesting to similarly investigate seasonal trends in mixotrophic communities. Our set of mixotrophic lineages and metabarcodes being publicly available, our method will be applicable to any other

357 metabarcoding dataset, including time-series. It will also be open to inputs and
358 updates from the global scientific community.
359
360 *The contrasted biogeographies of marine mixotypes*
361 • *Constitutive Mixotrophs*
362 Constitutive mixotrophs (CM) have very diverse feeding behaviours, with some
363 species requiring phototrophy to grow, others phagotrophy, and some being
364 obligate mixotrophs [9]. They were described in all waters of the global ocean [61–
365 65]. We found them distributed in a range of conditions almost as wide as non-
366 constitutive mixotrophs (Figure 1 and 3). Among highly abundant lineages, most
367 were dominantly found in eutrophic and shallow habitats. However, a few
368 dinoflagellates were found to be highly dominant in oligotrophic, subtropical
369 waters, showing how wide of a range of conditions constitutive mixotrophs can
370 grow in (Figure 3). This illustrates how mixotrophy can allow organisms to
371 dominate ecosystems even when environmental conditions are poorly adapted to
372 purely phototrophic or heterotrophic organisms. When taken explicitly into account
373 in biogeochemical models, marine mixotrophs increase carbon export by up to 30%
374 [22]. Hence, their global ubiquity supposes that the carbon export of the biological
375 carbon pump could be underestimated in both oligotrophic and eutrophic areas
376 [26].

- *Plastidic-Specialist and Generalist Non-Constitutive mixotrophs (pSNCM & GNCM)*

Like Leles et al. [27], we found pSNCM and GNCM to have quite similar biogeographies (Figure 3, section 3 of Sup. Mat.). Sequence abundance of most of the metabarcodes for these two mixotypes was homogeneously low (Table 1), but the two most abundant species, *Mesodinium rubrum* (pSNCM) and *Pseudotontonia simplicidens* (GNCM), were found mostly in coastal and eutrophic waters, consistently with Leles et al. [27]'s morphological observations (Figure 3, section 3 of Sup. Mat.). No species-level barcode is available in the PR2 database for the *Tontonia* genus, and only one can be found for *Pseudotontonia* and *Laboea* genera, even though morphological records of these GNCM are numerous [27]. Experiments using meso- and microcosms combined with individual counts and morphological identification have found that GNCM ciliates can represent up to half of the individuals in ciliate communities of the photic zone [11, 66, 67]. A proportion we would have trouble to reach with the 5 lineages we were able to consider, knowing that there are 8,686 different ciliate lineages available in PR2. This highlights the urgent need for supplementing 18S reference databases for mixotrophic ciliates.

- *Endo-symbiotic Specialist Non-Constitutive Mixotrophs (eSNCM)*

Endo-symbiotic specialist non-constitutive mixotrophs (eSNCM) is by far the most widespread and abundant non-constitutive mixotype in the global ocean (Figure 1 and 2) [27, 47, 54]. Their biogeography stands out, with a lot of highly abundant ubiquitous lineages, and some other specialized towards certain types of ecosystems (Figure 3). They represent 95.7% of the sequence counts in our study and correspond to 90.7% of the metabarcodes (Table 1), which highlights their abundance and diversity. The very high number of rDNA copies present in Rhizaria orders such as Collodaria [47] might lead the eSNCM to appear more abundant in metabarcoding datasets than they ecologically are. However, in oligotrophic open oceans the Rhizaria biomass is estimated to be equivalent to that of all other mesozooplankton [68], and positively correlated to the carbon export [15], showing how ecologically important they can be.

Investigating the divergent biogeographies of Collodaria and Acantharia

Collodaria are living either as solitary large cells or as colonies [47], which explains why they are predominantly found in macro-sized (180-2000 μm) filter samples (Figure 3). All described Collodaria species so far harbour photosynthetic endo-symbionts, mostly identified as the dinoflagellate species *Brandtodinium nutricula* [47, 69]. These dinoflagellates are able to get in and out of their symbiotic state, which implies a light and/or reversible effect of the Collodarian

host on its symbiont metabolism [69]. Based on the same metabarcoding dataset, Collodaria were described as particularly abundant and diverse in the oligotrophic open ocean [47]. In our results, Collodaria dominate oligotrophic, relatively deep waters (Figure 3 and 4a). These Collodaria appear opposed to another set of Rhizaria (Acantharia and Spumellaria) linked to eutrophic and shallow waters (Figure 3 and 4b, section 3 of Sup. Mat.). Acantharia are found ubiquitously in the global ocean, but display particularly high sequence abundances in some specific regions [54]. Mixotrophic Acantharia live in symbiosis with the cosmopolitan haptophyte *Phaeocystis*, which is highly abundant and ecologically active in its free-living phase [54]. Unlike the one of Collodaria, this symbiosis is irreversible : an algal symbiont can not go back to its free-living phase [54]. Our results suppose that these specific symbiotic modes could enable Acantharia and Collodaria to dominate different ecosystems (Figure 3 and 4). Moreover, living in colonies as Collodaria could help to dominate oligotrophic systems, *e.g.* by accumulating more food and nutrients through their gelatinous extra-cellular matrix [47]. Experiments and modeling studies should help to investigate the contribution of this assumption, comparing food acquisition capacity and growth rates of free individuals *versus* in colony.

Towards an integration of mixotrophic diversity into marine ecosystem models

436 The future of marine communities' modeling lies in the integration of omics
437 datasets into modeling frameworks [18, 70–73]. The use of metabolic networks and
438 gene-centric methods has already shown very promising results in modeling
439 prokaryotic ecological dynamics [18, 73]. However, eukaryotic metabolic
440 complexity makes these methods hard to apply on protists for now, and we still
441 lack a universal theoretical framework on how to integrate such methods into
442 concrete modeling [70]. Mixotrophic protists are physiologically complex, and
443 their feeding behaviour can vary drastically on short time scales [9]. It will then
444 take a few more years of comparative genomics and transcriptomics studies before
445 being able to model their physiology with purely gene-based approaches. Still,
446 mechanistic models of mixotrophy exist and are quite complex [21, 23], even if the
447 one from Ghyoot et al. [23] could be implemented in a global biogeochemical
448 model [74]. Most models make the choice to represent either one or two (NCM and
449 CM) types of organisms able to play the role of all mixotypes depending on
450 parameterization. However, no global agreement has been reached on to what
451 extent the different mixotypes should be modeled. This is mainly due to a lack of
452 quantitative and comparative data on the global impact of grazing and carbon
453 fixation by the different mixotypes [75]. With our study, we show how meta-omics
454 data can be used to identify groups of organisms distributed differently in response
455 to the environment. It also allows the identification of ecological traits and

environmental factors potentially responsible for these divergences. This information can be used to identify key species or lineages, and design controlled experiments with variations of targeted environmental factors to produce the quantitative data needed by modelers. Considering our results, we propose that host-symbiont dynamics of eSNCM should be investigated as a trait playing a potential role on Rhizaria ability to thrive in oligotrophic conditions. Particularly, the mechanisms behind holobiont formation and its potential reversibility could play major roles on eSNCM carbon fixation in various nutrient conditions. Future experiments comparing responses of Collodaria and Acantharia holobionts to different stresses in terms of grazing and carbon fixation could lead to a better understanding of the physiological differences between their two modes of symbiosis. Also, our results suggest that the metabolic flexibility of CM should allow this mixotype to grow in almost any conditions, with individual species probably spanning continuously between complete autotrophy and complete heterotrophy. The risk is then to create a ”perfect beast” mixotroph dominating all systems [21]. To avoid that, we need more comparative data on grazing and carbon fixation of obligate phototrophs *versus* obligate heterotrophs in response to nutrient depletion and environmental fluctuation. Here again, meta-omics data could help to identify candidates for efficient experiment designs. Finally, the small number of lineages of GNCM and pSNCM in our study makes it hard to come up with

476 strongly supported conclusions on whether they should be differentiated in models
477 or not. They seem to share similar biogeographies using snapshot data (Figure 3,
478 section 3 of Sup. Mat.), but considering that they have different abilities for
479 conserving stolen chloroplasts over time, it might not be the case when looking at a
480 time series analysis [20, 76, 77].
481
482 Our study uses meta-omics data to investigate the global distribution and
483 biogeography of mixotrophic protists in the ocean. Our results, currently based on
484 metabarcoding data, complement morphological records and will be complemented
485 in the near future by metagenomics and metatranscriptomics studies. The latter will
486 allow to distinguish the protists with mixotrophic capabilities from the ones with
487 ongoing mixotrophic activity. This could lead to quantitative estimations of
488 mixotrophic rates in environmental samples, allowing a sharpened study of
489 mixotrophic protists ecology in the global ocean. It could also lead to a metabolic
490 description of complex processes like kleptoplasty and endo-symbiosis, hence
491 facilitating the modeling of mixotrophic behaviours and its incorporation in ocean
492 biogeochemical models.
493
494 **Acknowledgements**
495

We would like to particularly thank Stéphane Pesant and Stéphane Audic for their work on making *Tara* Oceans datasets available. We also thank John Dolan (CNRS, LOV, Villefranche-sur-mer, France), Miguel Mendez-Sandin (Sorbonne Université, Station Biologique de Roscoff, France) and Wei-Ting Chen (National Taiwan Ocean University, Taiwan) for their essential help during the construction of the mixotrophic lineages set. We also thank Florentin Constancias for his help on the metabarcodes clustering tests conducted. Finally, we thank the three anonymous reviewers for their very constructive comments. This article is contribution number #XX of *Tara* Oceans. For the *Tara* Oceans expedition, we thank the commitment of the CNRS (in particular, Groupement de Recherche GDR3280), European Molecular Biology Laboratory (EMBL), Genoscope/CEA, VIB, Stazione Zoologica Anton Dohrn, UNIMIB, Fund for Scientific Research—Flanders, Rega Institute, KU Leuven, The French Ministry of Research. We also thank the support and commitment of Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Région Bretagne, Lorient Agglomération, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, the *Tara* schooner and its captains and crew. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who graciously granted sampling permissions. *Tara* Oceans would not

exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org).

This work was funded by the FunOmics project of the French national programme EC2CO-LEFE of CNRS and by the ModelOmics project of the Émergence programme of Sorbonne Université. Emile Faure acknowledges a 3-year Ph.D. grant from the "Interface Pour le Vivant" (IPV) doctoral program of Sorbonne Université.

Competing interests

The authors declare that they have no competing interests.

Supplementary information is available at ISME's journal website.

References

1. Caron DA, Countway PD, Jones AC, Kim DY, Schnetzer A. Marine Protistan Diversity. *Annu Rev Mar Sci* 2012; **4**: 467–493.

2. de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, et al. Eukaryotic plankton diversity in the sunlit ocean. *Science* 2015; **348**: 1261605–1261605.

3. Pawlowski J, Audic S, Adl S, Bass D, Belbahri L, Berney C, et al. CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms. *PLOS Biol* 2012; **10**: e1001419.

4. Caron DA, Alexander H, Allen AE, Archibald JM, Armbrust EV, Bachy C, et al. Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nat Rev Microbiol* 2017; **15**: 6–20.

5. Keeling PJ, Campo J del. Marine Protists Are Not Just Big Bacteria. *Curr Biol* 2017; **27**: R541–R549.

6. Caron DA. Mixotrophy stirs up our understanding of marine food webs. *Proc Natl Acad Sci* 2016; **113**: 2806–2808.

7. Le Quéré C, Harrison SP, Colin Prentice I, Buitenhuis ET, Aumont O, Bopp L, et al. Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Glob Change Biol* 2005; **11**: 2016–2040.

8. Amacher J, Neuer S, Anderson I, Massana R. Molecular approach to determine contributions of the protist community to particle flux. *Deep Sea Res Part Oceanogr Res Pap* 2009; **56**: 2206–2215.

9. Stoecker DK, Hansen PJ, Caron DA, Mitra A. Mixotrophy in the Marine Plankton. *Annu Rev Mar Sci* 2017; **9**: 311–335.

551 10. Flynn KJ, Stoecker DK, Mitra A, Raven JA, Glibert PM, Hansen PJ, et al. Misuse of the
552 phytoplankton-zooplankton dichotomy: the need to assign organisms as mixotrophs within
553 plankton functional types. *J Plankton Res* 2013; **35**: 3–11.
554 11. Mitra A, Flynn KJ, Tillmann U, Raven JA, Caron D, Stoecker DK, et al. Defining
555 Planktonic Protist Functional Groups on Mechanisms for Energy and Nutrient Acquisition:
556 Incorporation of Diverse Mixotrophic Strategies. *Protist* 2016; **167**: 106–120.
557 12. Esteban GF, Fenchel T, Finlay BJ. Mixotrophy in Ciliates. *Protist* 2010; **161**: 621–641.
558 13. Selosse M-A, Charpin M, Not F, Jeyasingh P. Mixotrophy everywhere on land and in
559 water: the grand écart hypothesis. *Ecol Lett* 2017; **20**: 246–263.
560 14. Ducklow HW, Steinberg DK, Buesseler KO. Upper ocean carbon export and the
561 biological pump. *Oceanogr-Wash DC-Oceanogr Soc-* 2001; **14**: 50–58.
562 15. Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, et al. Plankton networks
563 driving carbon export in the oligotrophic ocean. *Nature* 2016; **532**: 465–470.
564 16. Aumont O, Ethé C, Tagliabue A, Bopp L, Gehlen M. PISCES-v2: an ocean
565 biogeochemical model for carbon and ecosystem studies. *Geosci Model Dev* 2015; **8**: 2465–2513.
566 17. Follows MJ, Dutkiewicz S, Grant S, Chisholm SW. Emergent Biogeography of Microbial
567 Communities in a Model Ocean. *Science* 2007; **315**: 1843–1846.
568 18. Reed DC, Algar CK, Huber JA, Dick GJ. Gene-centric approach to integrating
569 environmental genomics and biogeochemical models. *Proc Natl Acad Sci* 2014; **111**: 1879–1884.
570 19. Johnson MD. Acquired Phototrophy in Ciliates: A Review of Cellular Interactions and
571 Structural Adaptations. *J Eukaryot Microbiol* ; **58**: 185–195.
572 20. Stoecker DK, Johnson MD, Vargas C de, Not F. Acquired phototrophy in aquatic protists.
573 *Aquat Microb Ecol* 2009; **57**: 279–310.
574 21. Flynn KJ, Mitra A. Building the 'perfect beast': modelling mixotrophic plankton. *J*

575 *Plankton Res* 2009; **31**: 965–992.
576 22. Ward BA, Follows MJ. Marine mixotrophy increases trophic transfer efficiency, mean
577 organism size, and vertical carbon flux. *Proc Natl Acad Sci* 2016; **113**: 2958–2963.
578 23. Ghyoot C, Flynn KJ, Mitra A, Lancelot C, Gypens N. Modeling Plankton Mixotrophy: A
579 Mechanistic Model Consistent with the Shuter-Type Biochemical Approach. *Front Ecol Evol*
580 2017; **5**.
581 24. Ward BA, Dutkiewicz S, Barton AD, Follows MJ. Biophysical Aspects of Resource
582 Acquisition and Competition in Algal Mixotrophs. *Am Nat* 2011; **178**: 98–112.
583 25. Berge T, Chakraborty S, Hansen PJ, Andersen KH. Modeling succession of key resource-
584 harvesting traits of mixotrophic plankton. *ISME J* 2017; **11**: 212–223.
585 26. Mitra A, Flynn KJ, Burkholder JM, Berge T, Calbet A, Raven JA, et al. The role of
586 mixotrophic protists in the biological carbon pump. *Biogeosciences* 2014; **11**: 995–1005.
587 27. Leles SG, Mitra A, Flynn KJ, Stoecker DK, Hansen PJ, Calbet A, et al. Oceanic protists
588 with different forms of acquired phototrophy display contrasting biogeographies and abundance.
589 *Proc R Soc B Biol Sci* 2017; **284**: 20170664.
590 28. Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner H-W, et al. Multiple marker
591 parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in
592 marine anoxic water. *Mol Ecol* 2010; **19**: 21–31.
593 29. Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK. Sequencing our
594 way towards understanding global eukaryotic biodiversity. *Trends Ecol Evol* 2012; **27**: 233–243.
595 30. Bittner L, Gobet A, Audic S, Romac S, Egge ES, Santini S, et al. Diversity patterns of
596 uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. *Mol Ecol* 2013; **22**: 87–
597 101.
598 31. Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, et al. A Holistic

599 Approach to Marine Eco-Systems Biology. *PLoS Biol* 2011; **9**: e1001177.
600 32. Alberti A, Poulain J, Engelen S, Labadie K, Romac S, Ferrera I, et al. Viral to metazoan
601 marine plankton nucleotide sequences from the *Tara* Oceans expedition. *Sci Data* 2017; **4**:
602 170093.
603 33. Pesant S, Not F, Picheral M, Kandels-Lewis S, Bescot NL, Gorsky G, et al. Open science
604 resources for the discovery and analysis of Tara Oceans data. *Sci Data* 2015; **2**: 150023.
605 34. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, et al. The Protist Ribosomal
606 Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences
607 with curated taxonomy. *Nucleic Acids Res* 2013; **41**: D597–D604.
608 35. Granéli E, Edvardsen B, Roelke DL, Hagström JA. The ecophysiology and bloom
609 dynamics of Prymnesium spp. *Harmful Algae* 2012; **14**: 260–270.
610 36. Liu H, Aris-Brosou S, Probert I, de Vargas C. A Time line of the Environmental Genetics
611 of the Haptophytes. *Mol Biol Evol* 2010; **27**: 161–176.
612 37. Hansen P, Moldrup M, Tarangkoon W, Garcia-Cuetos L, Moestrup ø. Direct evidence for
613 symbiont sequestration in the marine red tide ciliate Mesodinium rubrum. *Aquat Microb Ecol*
614 2012; **66**: 63–75.
615 38. Agatha S, Strüder-Kypke MC, Beran A, Lynn DH. Pelagostrobilidium neptuni
616 (Montagnes and Taylor, 1994) and Strombidium biarmatum nov. spec. (Ciliophora,
617 Oligotrichea): phylogenetic position inferred from morphology, ontogenesis, and gene sequence
618 data. *Eur J Protistol* 2005; **41**: 65–83.
619 39. Jones HLJ, Leadbeater BSC, Green JC. Mixotrophy in marine species of
620 Chrysochromulina (Prymnesiophyceae): ingestion and digestion of a small green flagellate. *J*
621 *Mar Biol Assoc U K* 1993; **73**: 283.
622 40. Johnsen G, Dalløkken R, Eikrem W, Legrand C, Aure J, Skjoldal HR. Eco-physiology,

623 bio-optics and toxicity of the ichtyotoxic Chrysochromulina leadbeateri (Prymnesiophyceae). *J*
624 *Phycol* 1999; **35**: 1465–1476.
625 41. Rhodes L, Burke B. Morphology and growth characteristics of *Chrysochromulina* species
626 (Haptophyceae = Prymnesiophyceae) isolated from New Zealand coastal waters. *N Z J Mar*
627 *Freshw Res* 1996; **30**: 91–103.
628 42. Hemleben C, Be AWH, Anderson OR, Tuntivate S. Test morphology, organic layers and
629 chamber formation of the planktonic foraminifer Globorotalia menardii (d'Orbigny). *J*
630 *Foraminifer Res* 1977; **7**: 1–25.
631 43. Fehrenbacher JS, Spero HJ, Russell AD. Observations of living non-spinose planktic
632 foraminifers Neogloboquadrina dutertrei and N. pachyderma from specimens grown in culture.
633 *AGU Fall Meet Abstr* 2011; **41**.
634 44. Spero HJ, Parker SL. Photosynthesis in the symbiotic planktonic foraminifer Orbulina
635 universa, and its potential contribution to oceanic primary productivity. *J Foraminifer Res* 1985;
636 **15**: 273–281.
637 45. Faber WW, Anderson OR, Caron DA. Algal-foraminiferal symbiosis in the planktonic
638 foraminifer Globigerinella aequilateralis; II, Effects of two symbiont species on foraminiferal
639 growth and longevity. *J Foraminifer Res* 1989; **19**: 185–193.
640 46. Kuile B ter, Erez J. In situ growth rate experiments on the symbiont-bearing foraminifera
641 Amphistegina lobifera and Amphisorus hemprichii. *J Foraminifer Res* 1984; **14**: 262–276.
642 47. Biard T, Bigeard E, Audic S, Poulain J, Gutierrez-Rodriguez A, Pesant S, et al.
643 Biogeography and diversity of Collodaria (Radiolaria) in the global ocean. *ISME J* 2017; **11**:
644 1331–1344.
645 48. Ardyna M, Ovidio F, Speich S, Leconte J, Chaffron S, Audic S, et al. Environmental
646 context of all samples from the Tara Oceans Expedition (2009-2013), about mesoscale features at

647 the sampling location. 2017. PANGAEA.
648 49. Legendre P, Legendre LFJ. Numerical Ecology. 1998. Elsevier Science.
649 50. Escoufier Y. Le Traitement des Variables Vectorielles. *Biometrics* 1973; **29**: 751.
650 51. Borcard D, Gillet F, Legendre P. Numerical ecology with R. 2011. Springer.
651 52. R Core Team. R: A Language and Environment for Statistical Computing. 2017. R
652 Foundation for Statistical Computing, Vienna, Austria.
653 53. Longhurst AR. Ecological Geography of the Sea. 1998. Academic Press.
654 54. Decelle J, Probert I, Bittner L, Desdevises Y, Colin S, de Vargas C, et al. An original
655 mode of symbiosis in open ocean plankton. *Proc Natl Acad Sci* 2012; **109**: 18000–18005.
656 55. Le Bescot N, Mahé F, Audic S, Dimier C, Garet M-J, Poulain J, et al. Global patterns of
657 pelagic dinoflagellate diversity across protist size classes unveiled by metabarcoding. *Environ*
658 *Microbiol* 2016; **18**: 609–626.
659 56. Wu S, Xiong J, Yu Y. Taxonomic Resolutions Based on 18S rRNA Genes: A Case Study
660 of Subclass Copepoda. *PLOS ONE* 2015; **10**: e0131498.
661 57. Brown EA, Chain FJJ, Crease TJ, MacIsaac HJ, Cristescu ME. Divergence thresholds and
662 divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities?
663 *Ecol Evol* 2015; **5**: 2234–2251.
664 58. Egge E, Bittner L, Andersen T, Audic S, de Vargas C, Edvardsen B. 454 pyrosequencing
665 to describe microbial eukaryotic community composition, diversity and relative abundance: a test
666 for marine haptophytes. *PloS One* 2013; **8**: e74371.
667 59. Gilbert JA, Field D, Swift P, Thomas S, Cummings D, Temperton B, et al. The
668 Taxonomic and Functional Diversity of Microbes at a Temperate Coastal Site: A 'Multi-Omic'
669 Study of Seasonal and Diel Temporal Variation. *PLoS ONE* 2010; **5**: e15545.
670 60. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, et al. Community

671 Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science* 2006; **311**:
672 496–503.
673 61. Arenovski AL, Lim EL, Caron DA. Mixotrophic nanoplankton in oligotrophic surface
674 waters of the Sargasso Sea may employ phagotrophy to obtain major nutrients. *J Plankton Res*
675 1995; **17**: 801–820.
676 62. Safi KA, Hall JA. Mixotrophic and heterotrophic nanoflagellate grazing in the
677 convergence zone east of New Zealand. *Aquat Microb Ecol* 1999; **20**: 83–93.
678 63. Moorthi S, Caron DA, Gast RJ, Sanders RW. Mixotrophy: a widespread and important
679 ecological strategy for planktonic and sea-ice nanoflagellates in the Ross Sea, Antarctica. *Aquat*
680 *Microb Ecol* 2009; **54**: 269–277.
681 64. Unrein F, Gasol JM, Massana R. Dinobryon faculiferum (Chrysophyta) in coastal
682 Mediterranean seawater: presence and grazing impact on bacteria. *J Plankton Res* 2010; **32**: 559–
683 564.
684 65. Sanders RW, Gast RJ. Bacterivory by phototrophic picoplankton and nanoplankton in
685 Arctic waters. *FEMS Microbiol Ecol* 2012; **82**: 242–253.
686 66. Calbet A, Martínez RA, Isari S, Zervoudaki S, Nejstgaard JC, Pitta P, et al. Effects of
687 light availability on mixotrophy and microzooplankton grazing in an oligotrophic plankton food
688 web: Evidences from a mesocosm study in Eastern Mediterranean waters. *J Exp Mar Biol Ecol*
689 2012; **424**–**425**: 66–77.
690 67. Dolan JR, PÉrez MT. Costs, benefits and characteristics of mixotrophy in marine
691 oligotrichs. *Freshw Biol* 2000; **45**: 227–238.
692 68. Biard T, Stemmann L, Picheral M, Mayot N, Vandromme P, Hauss H, et al. *In situ*
693 imaging reveals the biomass of giant protists in the global ocean. *Nature* 2016; **532**: 504–507.
694 69. Probert I, Siano R, Poirier C, Decelle J, Biard T, Tuji A, et al. Brandtodinium gen. nov.

695 and B. nutricula comb. Nov. (Dinophyceae), a dinoflagellate commonly found in symbiosis with
696 polycystine radiolarians. *J Phycol* 2014; **50**: 388–399.
697 70. Stec KF, Caputi L, Buttigieg PL, D'Alelio D, Ibarbalz FM, Sullivan MB, et al. Modelling
698 plankton ecosystems in the meta-omics era. Are we ready? *Mar Genomics* 2017; **32**: 1–17.
699 71. Dick GJ. Embracing the mantra of modellers and synthesizing omics, experiments and
700 models. *Environ Microbiol Rep* 2017; **9**: 18–20.
701 72. Mock T, Daines SJ, Geider R, Collins S, Metodiev M, Millar AJ, et al. Bridging the gap
702 between omics and earth system science to better understand how environmental change impacts
703 marine microbes. *Glob Change Biol* 2016; **22**: 61–75.
704 73. Coles VJ, Stukel MR, Brooks MT, Burd A, Crump BC, Moran MA, et al. Ocean
705 biogeochemistry modeled with emergent trait-based genomics. *Science* 2017; **358**: 1149–1154.
706 74. Shuter B. A model of physiological adaptation in unicellular algae. *J Theor Biol* 1979; **78**:
707 519–552.
708 75. Millette NC, Grosse J, Johnson WM, Jungbluth MJ, Suter EA. Hidden in plain sight: The
709 importance of cryptic interactions in marine plankton. *Limnol Oceanogr Lett* 2018; **3**: 341–356.
710 76. Johnson MD, Oldach D, Delwiche CF, Stoecker DK. Retention of transcriptionally active
711 cryptophyte nuclei by the ciliate Myrionecta rubra. *Nature* 2007; **445**: 426–428.
712 77. Schoener DM, McManus GB. Plastid retention, use, and replacement in a kleptoplastidic
713 ciliate. *Aquat Microb Ecol* 2012; **67**: 177–187.
714
715

FIGURES & TABLES LEGENDS

Figure 1:

Global distribution of mixotypes from metabarcoding data. Maps showing for each station the proportion of sequences (in %) belonging to each mixotype over the total number of mixotrophic sequences. Stations in which no sequence was found were marked as absent, ones with less than 100 sequences marked as questionable. Each Longhurst biogeographical provinces [53] is coloured in the background if more than 100 sequences are detected in at least one of its stations.

Figure 2:

Sequence abundance, occupancy and spatial evenness of each mixotrophic metabarcode across sampled stations. Each metabarcode is plotted as a bubble, with its station occupancy, *i.e.* the number of stations in which it was found, and its station evenness, *i.e.* the homogeneity of its distribution among the stations in which it was detected, as coordinates. Violin plots were drawn for each mixotype on both the x and y axes. The size of each bubble is scaled to the sequence abundance found globally for the corresponding metabarcode.

Figure 3:

Impact of environmental variables on the distribution of marine mixotrophs. Triplot of the redundancy analysis (RDA) computed on the 62 Escoufier-selected lineages, after model selection. The adjusted R-squared of the analysis is of 34.89% (41.43% unadjusted). Each grey dot corresponds to a sample, *i.e.* one filter at one depth at one station. The blue dashed arrows correspond to the quantitative environmental variables. Abbreviations are as follows: *MLD* = mixed layer depth, *O2MaxD* = O2 maximum depth, *EuphzoneD* = euphotic zone depth, *PAR* = photosynthetically active radiations, *Calcite Sat. St.* = Calcite Saturation State, *c_660* = optical beam attenuation coefficient at 660 nm, *c_part* = beam attenuation coefficient of particles, *acCDOM* = absorption coefficient of colored dissolved organic matter. Plain arrows correspond to mixotrophic lineages, colors indicating mixotypes. For more readability, we do not represent all qualitative variables included in the model. That is why only the filter centroids are appearing, even though the sampling depth, season, season moment, *i.e.* early, middle or late, and biogeographical province were used in the RDA calculation.

Figure 4:

Contrasted global distributions of metabarcodes corresponding to two eSNCM lineages. Maps of Hellinger-transformed sequence count abundances for metabarcodes assigned to the Collodaria *Siphonosphaera cyathina* (A) and the

Acantharia Acanthrometridae F3 spp. (B). These two lineages are opposed on the first RDA axis (Figure 3 and S1). Size and color both illustrate abundance for better readability. Ellipses were drawn to highlight high abundance zones, and reveal the differences in lineages distribution.

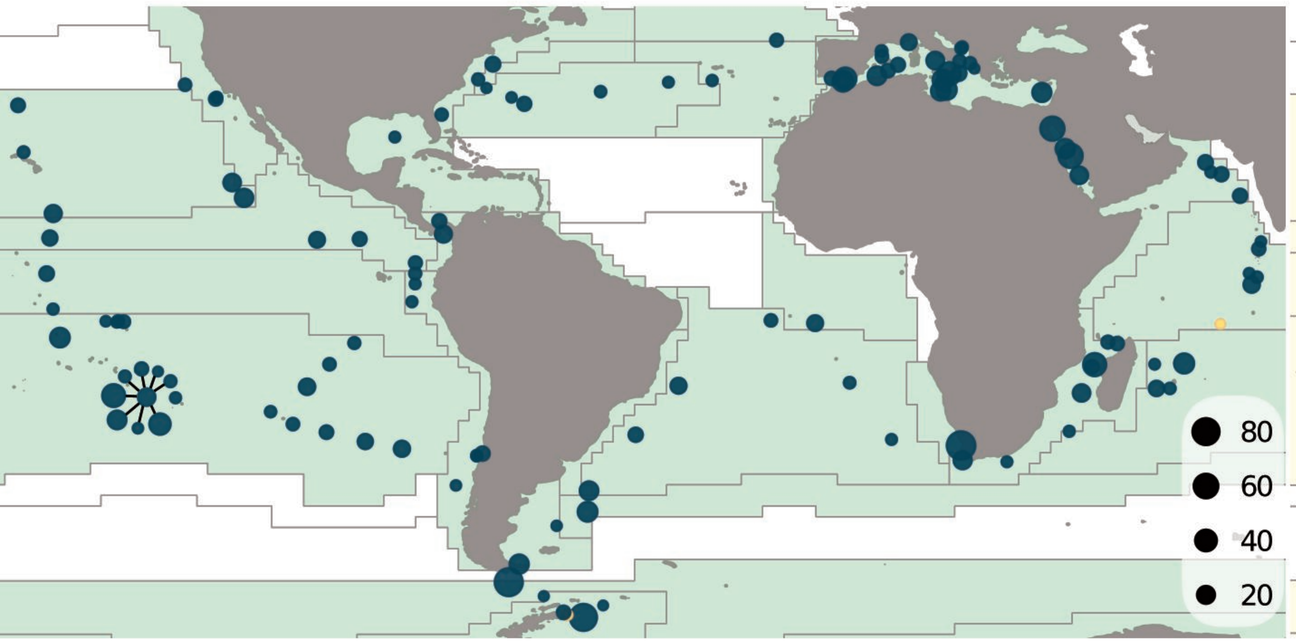
Table 1:

Detailed number of lineages found for each mixotype, as well as the number of metabarcodes, the corresponding total sequence counts over all stations, the mean sequence abundance by metabarcode, and mean metabarcode richness. The richness was computed as the number of different metabarcodes present at each station. It was calculated for each mixotype and means are indicated in the fifth column. Absences correspond to the number of stations in which no sequences were detected for the corresponding mixotype.

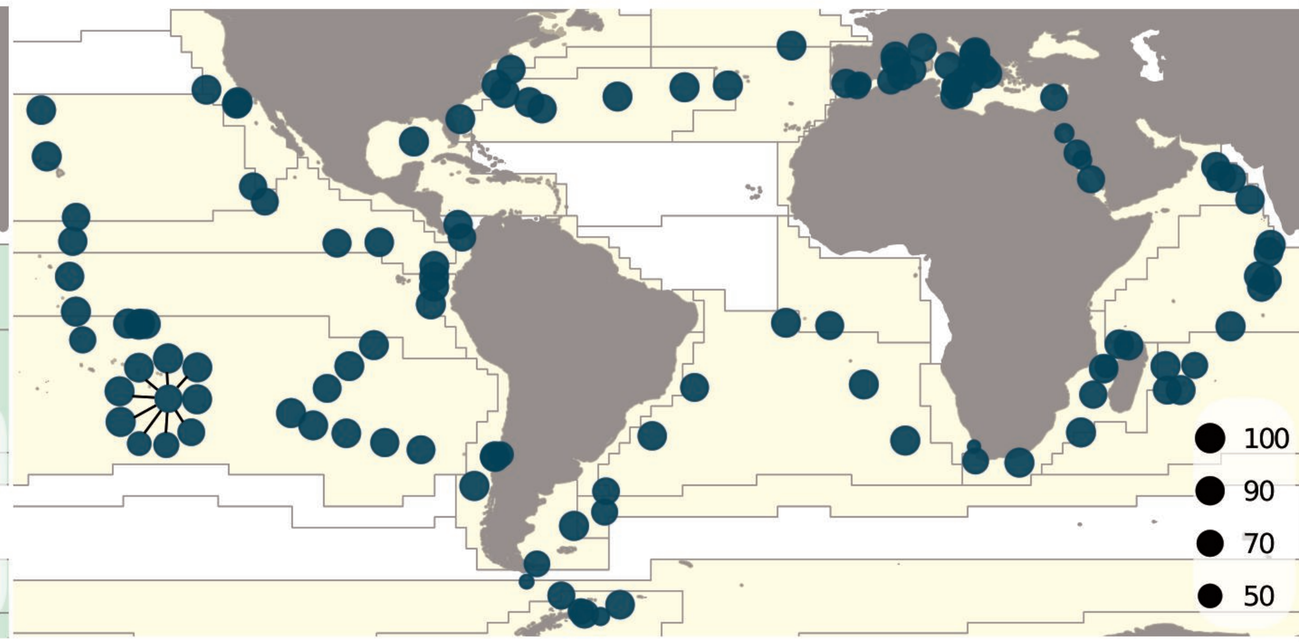
*The mean indicated here was calculated using only stations having the maximum number of samples (see main text).

Mixotypes	CM	eSNCM	pSNCM	GNCM
Number of lineages used in this study	42	77	9	5
Number of V9 metabarcodes	26015	288536	2143	1360
Total sequence abundance	3,581,751	86,098,397	208,096	63,622
Mean sequence counts per metabarcode	137.7	298.4	97.1	46.8
Mean metabarcode richness per station* (Std Dev)	2162 (1115)	18502 (9238)	67 (102)	84 (111)
Number of absences/station	0/122	0/122	5/122	3/122

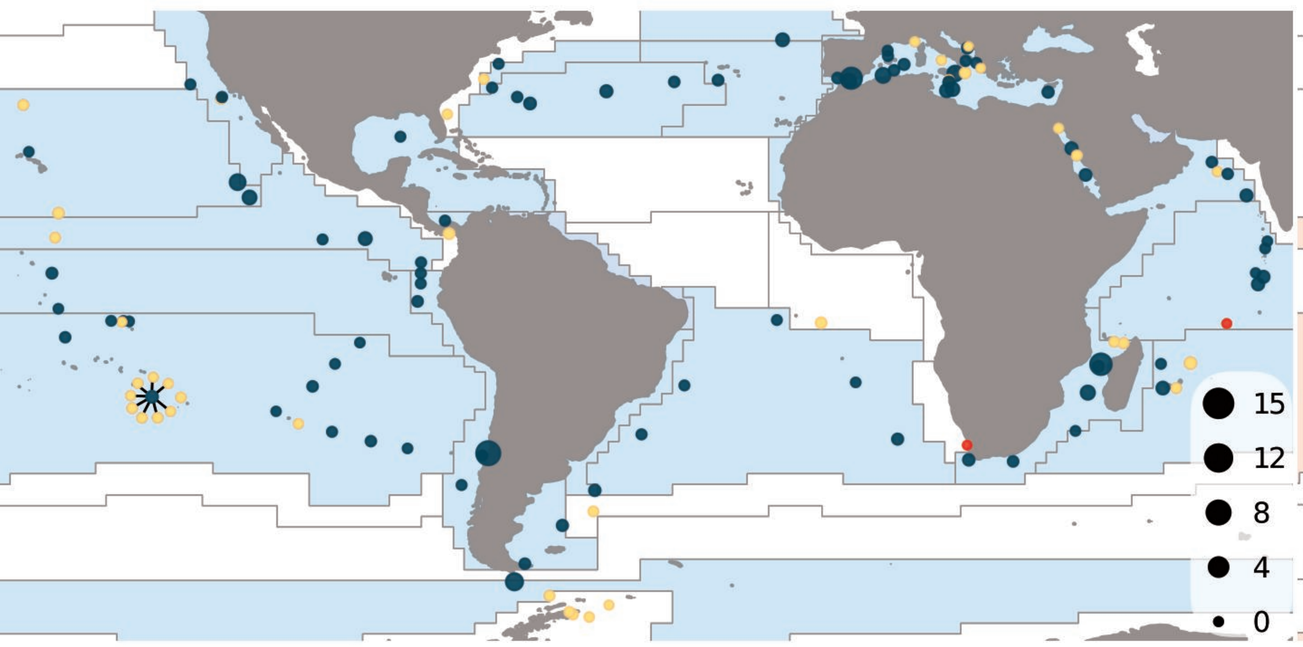
CM



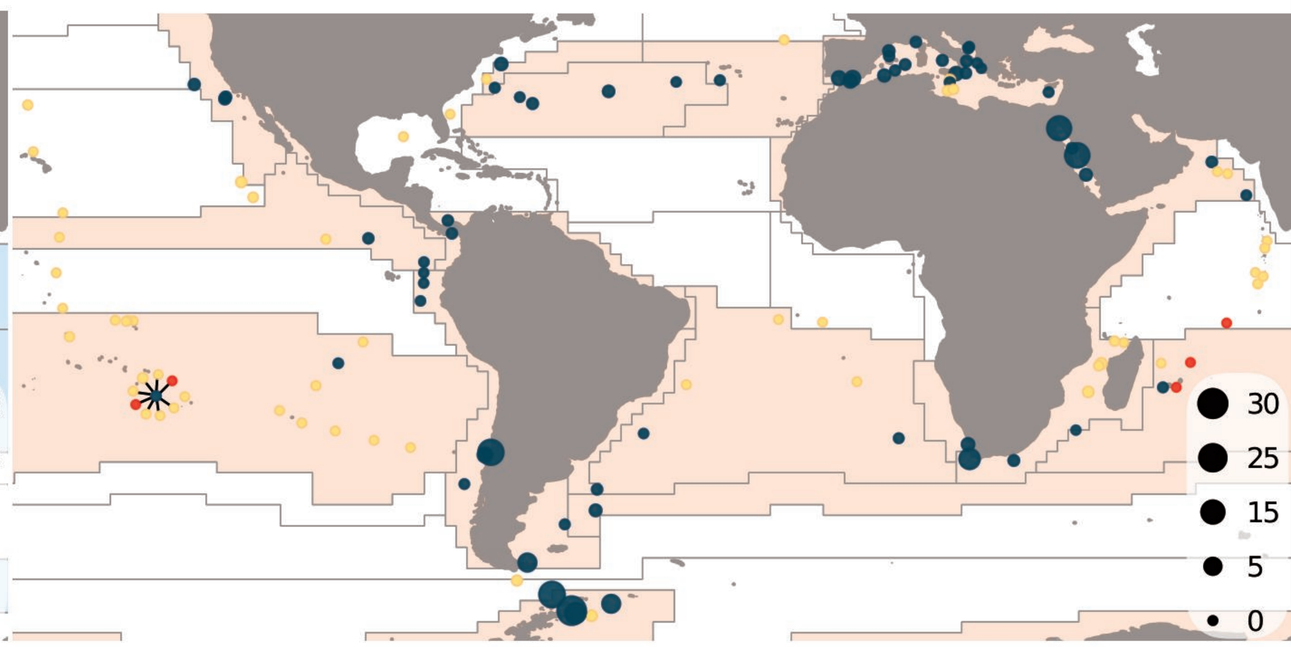
eSNCM



GNCM



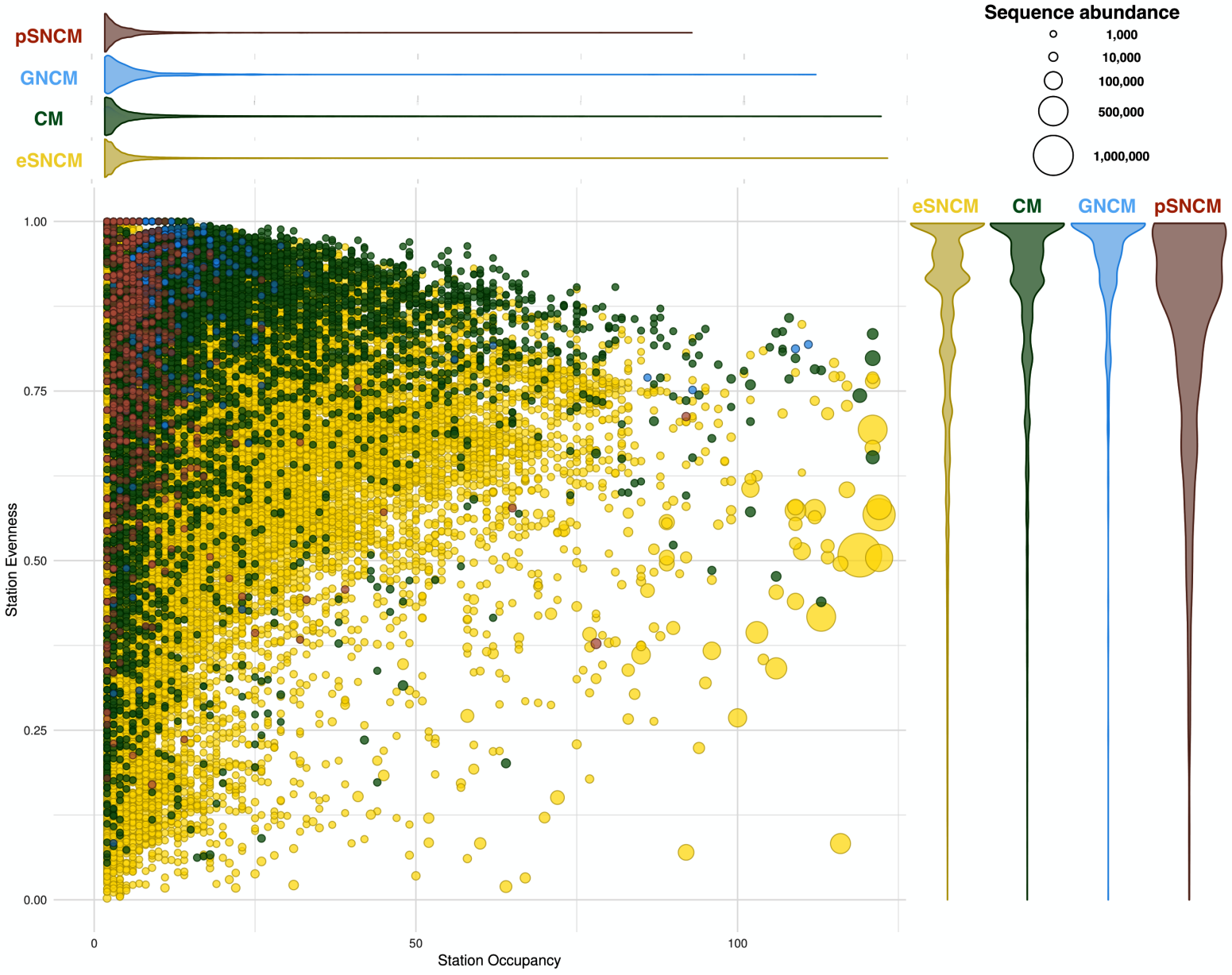
pSNCM



Absent

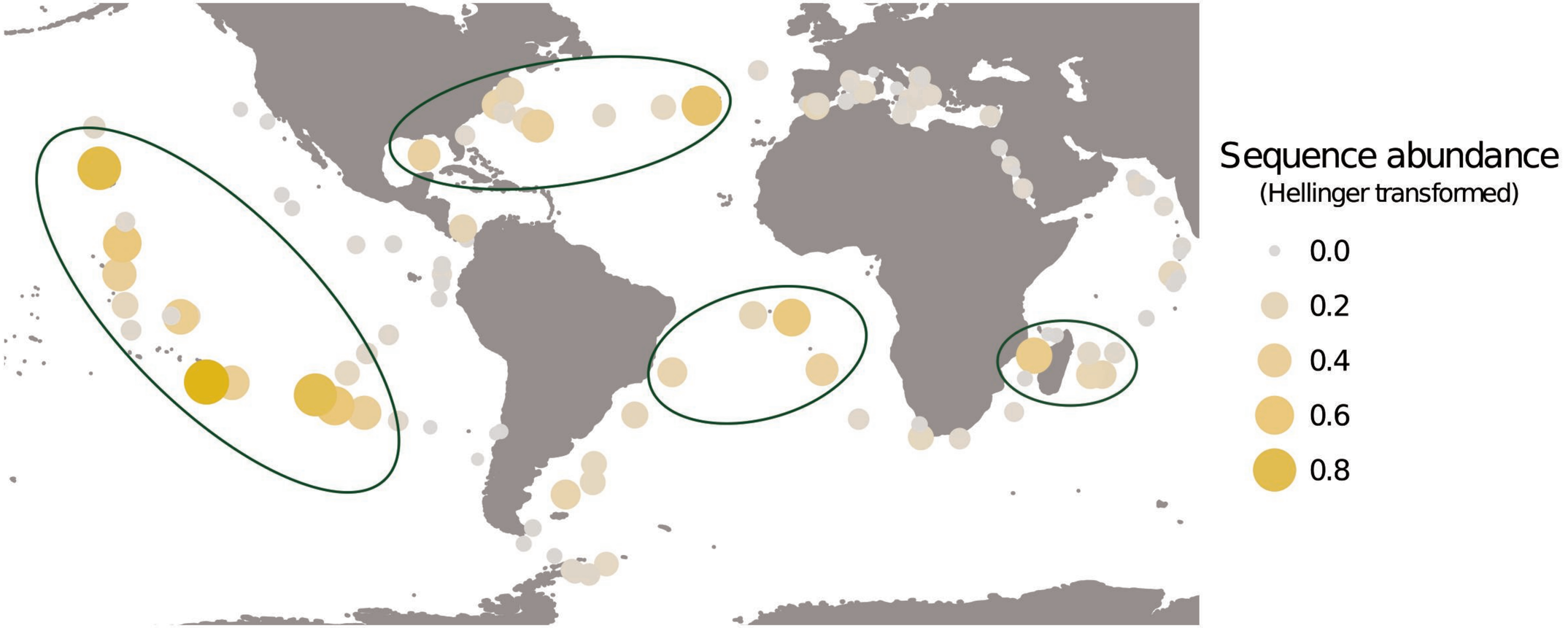
Present (>100 sequences)

Questionable (<100 sequences)


pSNCM
GNCM
CM
eSNCM
Sequence abundance
1,000
10,000
100,000
500,000
1,000,000
eSNCM
CM
GNCM
pSNCM
1.00
0.75
0.50
0.25
0.00
Station Evenness
0
50
100
Station Occupancy


Sphaerozoum punctatum
MLD
Bathymetry
PO4
Oxygen
Distance to coast
Lyapunov
Collozoum inerme
CO3
Collodaria
acCDOM
ChlorophyllA
NH4+
Si
HCO3
O2MaxD
Day length
c_part
Mesodinium rubrum
c_660
Siphonosphaera cyathina
Chrysochromulina acantha
P. simplicidens
Okubo Weiss
EuphZoneD
Chrysophyceae clade G spp
Amphibelone spp.
Acantharea & Spumellaria
Calcite
Alexandrium tamarense
Chrysophyceae clade H spp.
Sat.St
PAR
Karlodinium micrum
Acanthometridae F3
Other Acantharea
Longitude
Temperature
Fragilidium mexicanum
Salinity
Ornithocercus spp.
Gonyaulax polygramma
Mixotype
CM
eSNCM
GNCM
pSNCM
Filter size (µm)
0.8–5
5–20
20–180
180–2000
0.5
0.0
−0.5
−0.50
−0.25
0.00
0.25
RDA2 (9.05%)
RDA1 (14.96%)

A **Distribution of the *Siphonosphaera cyathina* barcodes**



B **Distribution of the Acanthrometridae F3 spp. barcodes**