Patterns and drivers of diatom diversity and abundance in the global ocean

Juan J. Pierella Karlusich[1,2,3,4]*, Karen Cosnier[1,2], Lucie Zinger[1,2+], Nicolas Henry[2,6], Charlotte Nef[1,2], Guillaume Bernard[1,2], Eleonora Scalco[7], Etienne Dvorak[1,2], *Tara* Oceans Coordinators, Fabio Rocha Jimenez Vieira[1,2], Erwan Delage[8], Samuel Chaffron[2,8], Sergey Ovchinnikov[4,9], Adriana Zingone[7], Chris Bowler[1,2]*

[1] Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France
[2] CNRS Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, Paris, France
[3] FAS Division of Science, Harvard University, Cambridge, MA, USA
[4] Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139
[5] Centre de Recherche sur la Biodiversité et l'Environnement (CRBE), CNRS, Université Toulouse III, IRD, INP, Toulouse, France
[6] CNRS, FR2424, ABiMS, Station Biologique de Roscoff, Sorbonne Université, Roscoff, France
[7] Stazione Zoologica Anton Dohrn, Villa Comunale, Naples, Italy
[8] Nantes Université, CNRS UMR 6004, LS2N, F-44000, Nantes, France
[9] John Harvard Distinguished Science Fellowship Program, Harvard University, Cambridge, MA, USA
[+]Present address: Centre de Recherche sur la Biodiversité et l'Environnement (CRBE), CNRS, Université Toulouse III, IRD, INP, Toulouse, France

*corresponding authors: Juan J. Pierella Karlusich (pierella@mit.edu) and Chris Bowler (cbowler@bio.ens.psl.eu)

ORCID ids:
Juan J. Pierella Karlusich, https://orcid.org/0000-0003-1739-4424
Karen Cosnier,
Lucie Zinger, https://orcid.org/0000-0002-3400-5825
Nicolas Henry, https://orcid.org/0000-0002-7702-1382
Charlotte Nef, https://orcid.org/0000-0002-9446-6367
Guillaume Bernard, https://orcid.org/0000-0001-6251-9500
Eleonora Scalco, https://orcid.org/0000-0001-8561-0614
Etienne Dvorak, http://orcid.org/0000-0002-7347-4544
Fabio Rocha Jimenez Vieira, https://orcid.org/0000-0001-9872-0337
Erwan Delage, https://orcid.org/0000-0002-7653-5111
Samuel Chaffron, https://orcid.org/0000-0001-5903-617X
Sergey Ovchinnikov, https://orcid.org/0000-0003-2774-2744
Adriana Zingone, https://orcid.org/0000-0001-5946-6532
Chris Bowler, https://orcid.org/0000-0003-3835-6187

Abstract

Diatoms constitute one of the most diverse and ecologically important phytoplankton groups, yet their large-scale diversity patterns and drivers of abundance are unclear due to limited observations. Here, we utilize *Tara* Oceans molecular and morphological data, spanning pole to pole, to describe marine diatom diversity, abundance, and environmental adaptation and acclimation strategies. The dominance of diatoms among phytoplankton in terms of relative abundance and diversity was confirmed, and the most prevalent genera were *Chaetoceros*, *Thalassiosira*, *Actinocyclus* and *Pseudo-nitzschia*. We defined 25 distinct diatom communities with varying environmental preferences illustrative of different life strategies. The Arctic Ocean stands out as a diatom hotspot with 6 of the diatom communities being exclusive to it. Light harvesting and photoprotection were among the cellular functions in which natural diatom populations invested the bulk of their transcriptional efforts. The biogeography of diatom lineages was consistent with their copy number for genes related to photoacclimation. This comprehensive study sheds light on marine diatom distributions, offering insights amid global change and oceanic anthropogenic impacts.

Key words: biodiversity; diatoms; metabarcoding; V4 and V9 hypervariable region; metagenomics; metatranscriptomics; *Tara* Oceans

Introduction

Diatoms are the most abundant and diverse group of eukaryotic phytoplankton, represented by tens of thousands of species[1–3]. Diatoms cover a broad size spectrum, which ranges over more than nine orders of magnitude in cell volume and is further expanded by many chain-forming species[4]. They are believed to be responsible for around 20% of the total primary production on the planet, thus constituting the base of aquatic food webs and contributing to the export of carbon from the surface ocean to depth in many marine environments[5–7]. Their cells possess a rigid silica cell wall denoted frustule, which makes them the world's largest contributors to biosilicification[8]. They are also important players in the nitrogen cycle because they have efficient strategies for nitrogen capture and utilization[9,10]. Furthermore, some species even harbor nitrogen-fixing symbionts[11–14]. Given their important role in biogeochemical cycles, food webs and carbon export, understanding what drives diatom diversity and spatial distribution is an important topic, in order to better predict the responses and resilience of marine ecosystems to natural and human-induced perturbations.

Field observations and modeling efforts have shown that marine diatoms are especially common in well-mixed coastal and upwelling regions and at high latitudes[2,15,16]. However, how different diatom populations (defined here broadly in terms of species, genera, size classes, or morphotypes) diverge in their biogeography and niches at global ocean scale is still hindered by limited observations and methodological limitations. For example, while satellite remote sensing now allows to estimate the distribution of diatoms and other phytoplankton coarse groups based on optical properties at the sea surface, it fails to discriminate their taxonomy, to detect subsurface blooms or to determine variable cellular pigment quotas[17–19]. The Continuous Plankton Recorder (CPR) survey has been collecting plankton records

since the 1930s, but is mainly useful for quantifying larger-celled species[20]. The first global ocean database for diatoms was compiled for the MARine Ecosystem DATa (MAREDAT) project, mainly using classical microscopy counts from bottle samples collected in regional studies[21]. Over the last years, genetic surveys have been producing an impressive amount of new data about diatom distributions. These investigations are typically based on PCR amplification and sequencing of a fragment of the small subunit of the rRNA gene from the DNA isolated from an environmental sample (a method hereafter referred to as metabarcoding[22]. rRNA gene-based metabarcoding has a comparatively deeper taxonomic resolution (including for cryptic species) as well as higher detection power (e.g., for rare species) than routine optical methods[23–25]. Currently, the only available large-scale and standardized molecular survey of marine planktonic diatoms was carried out during the first leg of the *Tara* Oceans expedition (46 sampling sites) by targeting the V9 fragment of the 18S rRNA gene[2]. This study was able to improve diatom diversity estimates, e.g., 43% of the observed genera were not represented in the MAREDAT database. However, there was a poor representation of areas expected to have high diatom abundances, such as polar regions.

These taxonomic surveys can be complemented by metagenonomic and metatranscriptomic approaches to infer diatom functional diversity, as well as their adaptation and acclimation mechanisms (i.e., variations in gene and transcript copy numbers, respectively) across environmental gradients. The ecological success of diatoms depends on a range of adaptive attributes in addition to the frustule, such as prominent vacuoles for nutrient storage, ice-binding proteins to withstand harsh polar environments, proton pumping rhodopsins to convert light energy into chemical energy in low-light environments, and a metazoan-like urea cycle for cellular management of

115 nitrogen[9,26–28]. However, between 30 and 60% of diatom genes still have unknown or
116 poorly described functions[29]. In addition, functional studies of diatom populations have
117 been restricted to specific genera in local and regional studies[30–35]. Therefore, many
118 gaps still remain in understanding the molecular mechanisms underpinning diatom
119 success and their distribution in the contemporary ocean.
120 In the current study, we aimed to provide a pole-to-pole view of the diversity,
121 adaptation and acclimation features in natural populations of diatoms. This study
122 expands the previous survey by Malviya et al.[2] from *Tara* Oceans not only by covering
123 a much higher number of sites (from 43 to 144 stations) in all ocean regions but also
124 by the significant incorporation of two new regions: the North Pacific Ocean (NPO)
125 and the Arctic Ocean (AO). We also provide a robust picture of diatom taxonomic
126 diversity and distribution through the comparative analysis of DNA data using two DNA
127 markers (i.e., the V9 and V4 variable regions of the 18S rRNA gene; only V9 had been
128 used in our previous study) together with observations obtained through microcopy.
129 Finally, we mine the metagenomes and metatranscriptomes of the same samples to
130 explore the adaptation and acclimation processes that may underlie diatom distribution
131 patterns.

132 Results and Discussion

133 Prevalence of diatoms among eukaryotic phytoplankton and 134 silicifiers

135 To capture the whole size spectrum of plankton, a combination of filter membranes
136 with different pore sizes were used to serially size-fractionate organisms by cell
137 diameter and aggregation forms (specifically from 0.8-5 µm or 0.8-2000 µm, 3-20 or
138 5-20 µm, 20-180 µm, and 180-2000 µm)[36]. V4 and V9 amplicons were sequenced

139 from these size-fractionated samples collected in epipelagic waters from 144 stations
140 (**Figures S1 and S2**). Amplicon sequence variants (ASVs) were inferred using
141 denoising methods (DADA2)[37] and taxonomically assigned using IDTAXA[38] with the
142 PR2 database version 4.14[39] (see Materials and Methods). We retrieved ASVs
143 assigned to eukaryotic phytoplankton, and normalized their read abundance by the
144 total eukaryotic read abundance of the corresponding sample.
145 As expected, diatoms are overall the most abundant and the second most occurring
146 group (**Figures 1a-b and S3a-b**). In addition, we found that diatoms exhibited the
147 largest variation in relative abundance, from 0.0002% to 97%, while other
148 phytoplankton groups had abundance variations of up to 42% for photosynthetic
149 dinoflagellates (hereafter dinoflagellates), 39% for haptophytes, 25% for
150 pelagophytes, 13% for dictyochophytes (another silicifying group), and 11% for
151 prasinophytes (**Figures 1a and S3a**). Therefore, diatoms constitute an important
152 group due to their dominance and their higher sensitivity to environmental variations,
153 which likely has important consequences for the ecosystem in a context of climate
154 change. Given that diatoms have been used to assess freshwater quality for
155 decades[40], this result also suggests a similar use as health indicators for the ocean.
156 Diatom and dinoflagellate relative abundance values based on 18S rRNA gene
157 metabarcoding were higher compared to those inferred from the single-copy marker
158 gene *psbO* found in metagenomes[41]. This difference could be attributed at least partly
159 to the larger cell size of these organisms. Although cell size is not the only factor
160 influencing cell rRNA copy number[42], larger cells generally require a higher number of
161 copies of the 18S rRNA gene to maintain cell activity[43,44]. Therefore, the prevalence
162 of diatoms among the 18S rRNA gene barcodes could reflect to some extent a higher
163 relative biovolume proportion.

164 Further, diatoms were also found to be the leading group among silicifiers[45], rivaled
165 only by Nassellaria, an order of polycystine rhizarians with silica-based skeletal
166 structures (**Figure 1d-e and S3c-d**). All in all, this survey reaffirms the dominance of
167 diatoms within both phytoplankton and silicifier communities, highlighting their integral
168 role in marine biogeochemistry and their potential as ecological indicators in a
169 changing global climate.

170 Diatom ASV diversity

171 We focused on diatoms by selecting only those V9 and V4 ASVs assigned to this
172 group. From 969 samples and almost 385 million V4 reads in total, we obtained a total
173 of >20 million reads and 3,429 ASVs for diatom V4, while from 932 samples and
174 >1,554 million V9 reads, we obtained >112 million reads and 3,425 ASVs for diatom
175 V9. In comparison, the 293 samples of the initial segment of the *Tara* Oceans
176 expedition analysed in Malviya et al.[2] produced less than half the number (1,510) of
177 diatom V9 ASVs.
178 Despite the ~5.5 times higher sequencing depth for diatom V9, we obtained an almost
179 identical number of diatom ASVs compared to V4, and accumulation curves and
180 species richness estimators indicated that both V4 and V9 ASV diversity approached
181 saturation at a global scale (**Figure S4 and Table S1**). Similar results despite the
182 difference in sequencing depth can be attributed to the fact that the V4 marker is
183 almost three times longer (385 bp *vs* 130 bp for V9), which inherently increases not
184 only the amount of possible genuine variation of the marker, but also the likelihood of
185 including PCR/sequencing errors that may not be detected by the DADA2 algorithm,
186 as reported by previous studies[46,47].
187 The diatom ASVs represented 33% and 41% of the total V4 and V9 ASVs attributed
188 to phytoplankton, respectively (**Figure 1c**). This highlights diatoms as the most diverse

group within the global phytoplankton community, followed by dinoflagellates and haptophytes. However, photosynthetic dinoflagellate diversity might have been underestimated because a great number of ASVs cannot be classified as either photosynthetic or non-photosynthetic (e.g., those categorized as 'unknown dinoflagellate' or 'unknown Gymnodiniales').

Patterns of diatom abundance and diversity

When examining diatom reads relative abundance across marine biomes and oceanic regions, the highest values were observed in polar biomes (**Figure S5c-d**). Accordingly, a relative abundance latitudinal gradient towards the poles is clearly observed in all size classes (**Figures 2a-b and S6a-b**). This is consistent with the larger thermal breadths and lower minimal thermal growth of diatoms when compared with cyanobacteria and dinoflagellates[48]. In contrast, diatom diversity follows a latitudinal gradient with a decrease in species richness towards the poles (**Figures 2c and S6c**), where diatom communities are likely hyper-dominated by a lower number of species, as observed in previous studies[49].

We employed Partial Least Squares (PLS) analysis to explore environmental factors affecting diatom relative abundance and the Shannon diversity index (**Figure 3a**). Diatom Shannon diversity and relative abundance generally correlate with the same environmental variables, but in opposite directions, e.g., for temperature. The positive correlation between diatom relative abundance and chlorophyll *a* suggests that diatoms are usually the dominant group in high-biomass phytoplankton communities. We also observed a significant positive association between diatom abundance and annual maximum nitrate concentration (AM NO_3^-), underscoring nitrate availability as a key factor for diatom growth. Conversely, diatom abundance tends to decrease relative to the ratios of various nutrients (ammonia, iron, or silica) to total dissolved

214 inorganic nitrogen (DIN; e.g. sum of ammonia, nitrite, and nitrate). The role of biotic
215 interactions is also significant. Diatoms show a negative association with
216 picocyanobacteria, supporting findings that diatoms can be outcompeted in nitrogen-
217 limited environments where cyanobacteria efficiently utilize ammonia[50,51], often
218 leading to low chlorophyll *a* levels. In contrast, grazing organisms such as copepods
219 exhibit a substantial positive relationship with diatoms, suggesting grazing top-down
220 control of diatom populations. Lastly, the presence of abundant silicified radiolarians
221 (Nassellaria, Spumellaria, and Phaeodarea) negatively correlates with diatom
222 abundance, likely reflecting competitive interactions over silica[45].

223 Diatom size patterns

224 As a reflection of their wide size range, diatoms were relatively abundant in the
225 nanoplankton (3/5-20 µm; mean relative abundance of ~9% of eukaryotic reads), but
226 also in the piconanoplankton (0.8-5/2000 µm; ~5% mean abundance) and in the
227 micro- and mesoplankton (20-180 µm and 180-2000 µm; ~5% and 3% mean
228 abundances, respectively) (**Figures 1a, S3a and S5a**). Historically, the abundance of
229 smaller diatoms has been underestimated due to detection challenges using
230 microscopy, but recent morphological and molecular studies support our observations
231 of their prevalence in various marine regions[52–55]. The broad size spectrum of diatoms
232 is further evidenced by previous research on their cell volumes, spanning from 3 μm^3
233 (*Thalassiosira* sp.) to 4.71×10^9 μm^3 [4,21]. It is worth noting that we detected in many
234 cases the same diatom species across all size fractions. Although we cannot discard
235 the possibility that extracellular DNA adsorbed on particles of varying sizes, this can
236 also be attributed to life-cycle variations (e.g., asexually dividing diatoms decreasing
237 in cell size, rare gamete formation), and the physical characteristics of certain species
238 (e.g., long needle-like cells and broken colonies passing through small mesh sizes)[56].

239 There is a clear environmental partition between diatom sizes, with larger diatoms
240 associated with polar regions rich in nitrate, whereas small diatoms are prevalent in
241 oligotrophic and temperate waters with high ratios of ammonia to DIN (**Figure 3b**).
242 This is in line with Bergmann's rule of reduced organism body size in warmer
243 environments[57]. Thus, the energy transfer from diatoms to higher trophic levels may
244 be reduced in temperate environments due to their smaller sizes[58].

245 Environmental distribution of diatom classes

246 Following Medlin and Desdevises 2020[59], diatoms are classified into three
247 monophyletic classes: Coscinodiscophyceae (basal radial centric diatoms),
248 Mediophyceae (polar centric diatoms and radial centric Thalassiosirales), and
249 Bacillariophyceae (pennate diatoms). Pennate diatoms are further divided into raphid
250 (Bacillariophycidae) and araphid genera (i.e., subclasses Urneidiophycida and
251 Fragilariophycidae).
252 For both V4 and V9, Mediophyceae were the most prevalent, followed by raphid
253 pennates and Coscinodiscophyceae (**Figures 4a and S8a)**. Araphid pennates were
254 considerably less abundant and only prevalent in an ice edge station in the AO
255 (TARA_188) (**Figure S9)**. Our observation of low abundance of araphid pennates is
256 consistent with their typical presence in benthic rather than planktonic
257 environments[1,60].
258 Mediophyceae were highly abundant in regions characterized by high chlorophyll *a*
259 concentrations and high latitudes **(Figures 4b-c, S7, S8b, and S9)**. Similarly,
260 Coscinodiscophyceae were also notably abundant in high latitude regions. In contrast,
261 raphid pennates were predominantly found in temperate and stable environments,
262 although they were also present in significant numbers at some high latitude stations.
263 There was also a compositional shift among the different size fractions of plankton.

264 The relative abundance of Mediophyceae was highest among the piconano-, nano-,
265 and microplankton **(Figures 4a and S7a)**. Although their relative abundance
266 decreased among mesoplankton, Mediophyceae continued to be the dominant diatom
267 class in this size range **(Figure S8a)**. In contrast, the relative abundance of raphid
268 pennate diatoms demonstrated a continuous decline as the size fractions increased.
269 Coscinodiscophyceae exhibited a predominant presence particularly within the nano-
270 and to a lesser extent, microplankton size fractions, with their numbers decreasing in
271 the piconano- and, specially, the mesoplankton. All these biogeographical and size
272 patterns are driven by the patterns of the main genera within each class (see below).

273 Diatom genera composition

274 We found a total of **76** genera for V4 barcodes, and **63** genera for V9 barcodes (**Figure**
275 **5**). The higher number of genera generated by V4 barcodes is probably related to its
276 higher taxonomic resolution and better representation in the reference database (PR2
277 v4.14[39]).
278 High variations across genera were observed in terms of abundance and richness.
279 *Chaetoceros* was the most abundant genus, followed by *Thalassiosira*, *Actinocyclus*,
280 and *Pseudo-nitzschia* (**Figure 5**; see also next section). Intragenus richness ranged
281 from **502** V9 ASVs and **687** V4 ASVs for *Chaetoceros* to as low as 1 ASV for some
282 other genera (**Figure 5**). This result reflects the differences in known species for each
283 genus (e.g., >100 for *Chaetoceros*, a few for *Arcocellulus* and *Lauderia*, and only 1 for
284 *Tenuicylindrus* and *Pseudohimantidium*). Additionally, this result parallels read
285 abundance, indicating a greater potential of capturing intraspecific and intragenomic
286 variations of these multicopy markers, as well as a higher likelihood of
287 PCR/sequencing errors.

288 We found ~65% of cumulative read relative abundances in surface samples (**Figure**
289 **5**), which reflects the sampling bias (513 samples from surface and 267 from DCM;
290 **Figure S5b**). When only focusing on locations with both SRF and DCM samples, we
291 observed that most genera were more abundant in surface waters (e.g, *Eucampia*,
292 *Thalassiothrix*, *Rhizosolenia*, *Proboscia*) or equally abundant in surface and DCM. The
293 only genus enriched in DCM was the centric diatom *Attheya*, which is restricted to AO
294 waters (**Figure S10**).
295 A wide range of genera were identified among the piconanoplankton communities.
296 Besides the expected genera such as *Minidiscus*, *Minutocellus*, *Arcocellulus*,
297 *Brockmanniella*, *Cyclotella*, and *Cocconeis* (**Figure 5**), we observed several typically
298 larger genera traditionally classified as microplankton (20-180 µm). These included
299 the centric genera *Hemiaulus*, *Dactyliosolen*, *Detonula*, as well as the pennate genera
300 *Pseudo-nitzschia*, *Surirella*, *Thalassionema*, and *Triceratium*, which often form chains
301 and are generally larger. Their presence in smaller plankton samples could be due to
302 broken cells, cells passing through the mesh by their shorter dimension, or possibly
303 being gametes.

304 Patterns for the most abundant diatom genera

305 We focused on the top 20 most abundant genera (**Figure 6**) since they accounted for
306 the majority of the signal from known taxa, with 92% of assigned reads in the entire
307 V9 dataset and 90% using V4. Genera that were exclusively abundant in one large
308 size fraction (20-180 µm or 180-2000 µm; **Figure S11**) were excluded from the
309 analysis as they tend to be less abundant overall as there is a logarithmic negative
310 relationship between size and abundance[36]. Some of the most abundant genera, such
311 as *Porosira*, *Fragilaria*, *Trieres*/*Odontella*, and *Attheya*, were regionally constrained,
312 i.e., highly abundant in a few specific stations (**Figure S12**; further details below).

313 Among the top 20 most abundant diatom genera, 15 were consistently found in both
314 V4 and V9 datasets, including *Chaetoceros*, *Thalassiosira*, *Actinocyclus*, *Pseudo-*
315 *nitzschia*, *Porosira*, *Proboscia*, *Guinardia*, *Corethron*, *Leptocylindrus*, *Eucampia*,
316 *Skeletonema*, *Bacteriastrum*, *Nitzschia*, *Fragilariopsis*, and *Minidiscus* (**Figure 6b**).
317 Six genera appear only in one dataset but still rank high in the other (between positions
318 21 and 28: *Rhizosolenia*, *Planktoniella*, *Navicula*, *Attheya*, *Hemiaulus*, and
319 *Minutocellus*), indicating the robustness of our results. Nonetheless, a few differences
320 were observed, with *Fragilaria*, *Asteromphalus* and *Brockmanniella* found only in V4,
321 and *Trieres*/*Odontella* mainly detected in V9 (**Figure 6b**). These discrepancies could
322 be due to differences in the coverage and quality of reference databases, the resolving
323 power of the DNA barcode, and variations in PCR amplification efficiency among
324 different taxa.
325 We found 6 new genera among the V9 top 20 in comparison with the previous *Tara*
326 Oceans V9 survey (see Figure 6A in Malviya et al.[2]), which did not include AO, NPO
327 and most of NAO (**Figure S1b**). As expected, the list contains genera prevalent in
328 these new regions: *Bacteriastrum* was abundant in NPO, SPO and IO; *Porosira*,
329 *Skeletonema* and *Nitzschia* were abundant in AO (despite *Skeletonema* being
330 previously reported as absent from polar waters[61]) (**Figures 6b and 7**). However, the
331 remaining two genera were more abundant in SO: *Hemiaulus* and *Trieres*/*Odontella*
332 (**Figures 6b and 7**). Although SO was already included in Malviya et al.[2], the higher
333 regional coverage of the present work explains this result for *Trieres*/*Odontella* (6
334 current stations *vs* 2 stations in the previous survey), and the updated reference
335 database for *Hemiaulus*.
336 Given that the 20-180 µm size fraction was also quantified using optical
337 microscopy[2,62,63], we compared the most abundant genera between methods in this

size fraction **(Figure S13**). The comparison resulted in a high degree of similarity: 13 of the top 20 genera based on microscopy are also in the top 20 in both molecular datasets, plus three top 20 genera in the microscopy data that are only in one molecular dataset. Some of the differences are explained by the scarcity or lack of references for some genera (e.g., *Cylindrotheca* and *Thalassionema)* or the difficulty of detecting and/or identifying some genera in microscopy *(*e.g., *Arcocellulus, Minidiscus* and *Haslea).*

We observed a high variability in the biogeography of the 10 most abundant diatom genera (**Figures 7 and S14 and Supplementary File 1**). There was a group of cosmopolitan genera whose abundances were much higher in polar regions: *Chaetoceros*, *Thalassiosira*, *Actinocyclus* and *Fragilariopsis* (this latter case is detected in fewer stations with V4 barcodes). *Pseudo-nitzschia*, another cosmopolitan genus, was more homogeneously distributed in the different ocean regions. Other genera were mainly detected in polar regions: *Porosira* and *Proboscia* were abundant in both AO and SO, *Fragilaria* in the AO (only detected by V4), *Corethron*, *Trieres*/*Odontella*, and *Hemialus* in the SO (the last two cases mainly detected by V9). Finally, two genera were prevalent in a few stations from tropical and subtropical regions: *Guinardia* and *Leptocylindrus*. Overall, these results confirm the importance of the higher geographical coverage of the present work, especially with the inclusion of AO.

Communities of co-occurring diatom lineages and their association to environmental gradients

We applied a weighted gene correlation network analysis (WGCNA)[64] to detect co-occurring diatom sub-communities. For both V4 and V9 datasets, we found a total of 25 modules (and a set of unclustered ASVs grouped into the grey module) with similar

363 environmental variables associated to them (**Figures 8 and S15**). The modules were
364 clustered according to their co-occurrences into 4 main groups: i) one formed by 6
365 modules prevalent in the AO and dominated by Mediophyceae ASVs, ii) another
366 formed by 4-5 modules mainly detected in the SPO (and to a lesser extent in the NPO)
367 and dominated by Mediophyceae and raphid pennate ASVs, iii) another group formed
368 by 2-4 modules abundant in the SO/SPO and with Mediophyceae and raphid pennate
369 diatoms, and iv) a group of different basin-specific modules detected in tropical and
370 subtropical regions, including 2 modules prevalent in the IO, 1-2 in the MS, 2 in the
371 NAO (only in V4), and 2 in the SAO (only V9). Therefore, we found diatom sub-
372 communities with a significant association to each oceanic basin or to basins that are
373 physically connected (e.g., the SPO and IO via the Indonesian Passage). The AO
374 stands out from the rest as it contains a relatively high-number of AO-specific sub-
375 communities. The reasons why the AO is rich in endemic diatom communities might
376 be related to its particular physical oceanography (the influx from large rivers, and the
377 Atlantic and Pacific Ocean waters), long residence times, and the extreme seasonality
378 in day length and sea ice cover, as well as low temperatures and variable salinity[65].
379 Overall, *Chaetoceros*, *Pseudo-nitzschia* and *Thalassiosira*, previously identified as the
380 most prevalent diatom genera, were found within almost all modules, together with
381 unidentified Bacillariophyta, Mediophyceae and raphid pennates (**Figure S15**).
382 For both V4 and V9, the turquoise and grey modules gathered the most ASVs, with
383 the grey modules containing a set of unclustered ASVs that were the most enriched in
384 different taxa. Both grey modules were negatively correlated with iron, oxygen and
385 latitude while positively correlated with temperature (**Figure 8**). Together with the fact
386 that they harbored diatoms from a diverse range of basins and sizes, the results
387 suggest that the grey modules represent a wide group of rather temperate diatoms;

388 while the absence of other significant correlations of these modules with environmental
389 parameters indicates potential generalist species, or that the wide biological diversity
390 of these modules precludes the identification of a general pattern.
391 Exploring the association patterns of diatom sub-communities with environmental
392 parameters revealed 3 modules positively correlated with nutrient (nitrate, silica,
393 phosphate) availability in both datasets. The V4 turquoise module, enriched in
394 *Thalassiosira*, *Porosira* and *Chaetoceros* (~10% each), was positively correlated with
395 nitrate, silica, phosphate and chlorophyll *a*. Conversely, both V4 modules midnightblue
396 – enriched in *Pseudo-nitzschia* (16%) *Actinocyclus* (14%) and *Corethron* (12%) – and
397 purple – enriched in *Actinocyclus* (10%) and *Chaetoceros* (9%) – displayed a similar
398 positive correlation with nitrate and phosphate while the latter was negatively
399 correlated with iron and latitude. This potentially indicates that these diatom sub-
400 communities are the most responsive to nutrient availability while being associated
401 with different oceanic regions – either SO (midnightblue, turquoise) or SPO/NPO
402 (purple).
403 On the other hand, all V9 pink, brown and turquoise modules, which all harboured
404 species representative of SO basins and clustered together, were significantly
405 correlated with the three nutrients but were enriched in a different set of species. While
406 both V9 pink (SO/IO) and brown (SO/NAO) modules contained larger diatoms and
407 showed a high proportion of *Chaetoceros* ASVs (49% and 13% of ASVs, respectively,
408 ranked 2nd and 3rd), the latter also displayed a significant contribution of *Odontella*
409 ASVs (7%) and was positively correlated with chlorophyll *a* and ammonium. In
410 contrast, the V9 turquoise (SO/SPO) module was enriched in *Hemiaulus* and
411 *Thalassiosira* ASVs (~10% and 7%). This therefore suggests the existence of 3 distinct
412 SO-associated diatom sub-communities similarly responding to the availability of

413 major nutrients but differing by their size, taxonomic composition and response to
414 other environmental factors.
415 Both V4 and V9 blue modules were enriched in *Chaetoceros* and *Thalassiosira* genera
416 from the AO, corroborating their negative correlation with temperature, and showed
417 positive correlations with dissolved inorganic carbon and chlorophyll *a*. The V9 blue
418 module was also positively correlated with net primary productivity, as was the V4
419 royalblue module (~15% *Bacteriastrum*), illustrating the diatom taxa associated with
420 the most productive regions. These could correspond to either species outcompeted
421 in conditions of nutrient scarcity, and/or to taxa highly contributing to net primary
422 productivity.
423 Finally, the V4 lightcyan module, which gathered very few ASVs (47) and notably from
424 the genera *Chaetoceros* and *Guinardia*, was the most associated with iron but showed
425 no other significant pattern, which could either indicate taxa specialized in iron uptake
426 or species with elevated iron quotas. Altogether, these analyses provide novel insights
427 into factors regulating the assemblage of diatom communities and highlight modules
428 harboring species indicative of different environmental gradients and, therefore,
429 showing different life strategies. These results could be considered for future studies
430 exploring potential indicator species among plankton.

431 Unassigned sequences

432 Unassigned ASVs, which could not be classified at genus or species level, constituted
433 41.5% and 57.9% of the total diatom V4 and V9 ASVs, and 25.7% and 19.2% of total
434 diatom V4 and V9 reads (**Figure S16a**). These include ASVs identified only at the
435 phylum level (e.g., 'unknown' diatoms), which consist of 229 V4 and 903 V9 ASVs.
436 Additionally, ASVs classified only at the class level included 1,129 V4 and 1,080 V9
437 ASVs. Notably, the number of unclassified ASVs from pennate diatoms exceeded

438 those assigned to specific pennate genera (578 vs 414 ASVs for V4, and 554 vs 200
439 V9 ASVs). Given that pennate diatoms are primarily benthic and only occasionally
440 found in plankton—with exceptions such as *Thalassionema, Fragilariopsis* and
441 *Pseudo-nitzschia* [60]— this distribution aligns with the limited studies from sediments[56].
442 The proportion of unassigned reads was higher in the 3/5-20 µm size fraction and less
443 prevalent in the larger size classes (20-180 µm and 180-2000 µm) (**Figure S16b**). This
444 is an expected result because diatoms belonging to the largest size fractions are those
445 more intensively studied and are hence better represented in genetic reference
446 databases. In addition, there were variations across oceanic regions and sampling
447 sites, with SPO, SAO, AO, and in particular SO, showing higher relative abundances
448 of unassigned ASVs (**Figure S16c**), probably because the southern hemisphere and
449 the poles are the least explored areas. While the V4 and V9 datasets showed similar
450 trends for unassigned sequences overall, a few stations (TARA_085 and TARA_086
451 in the SO, TARA_188 in the AO, and TARA_006 in the MS) had a high abundance of
452 unassigned V9 ASVs compared to V4 ASVs, while the opposite was found in
453 TARA_158 in AO (**Figure S16d**).

454 The biogeography of diatoms is consistent with their genetic 455 and transcriptional features

456 Beyond taxonomic insights from 18S V4 and V9 sequences, complementary functional
457 information can be derived from metatranscriptomics, which provides quantifiable
458 information about mRNA levels. Therefore, we mined polyA RNA sequencing data
459 from the same plankton samples used for metabarcoding. We focused on the top 100
460 nuclear-encoded Pfam families with the highest transcript abundances, which covered
461 56% of the total sum of diatom metatranscriptomic relative abundances in the global
462 ocean, thus representing the cellular functions and processes in which natural diatom

463 populations invest the bulk of their energy and resources, at least as viewed at the
464 level of mRNA abundance (**Figure 9**). To reduce redundancy, Pfams for similar
465 functions were merged (e.g., 137 Pfams coding for different diatom ribosomal subunits
466 were merged into “Ribosomal proteins”, and 47 Pfams coding for different ubiquitin
467 domains were merged into “Ubiquitin domains”).
468 Many photosynthetic proteins are in the list. Indeed, the most abundant single Pfam
469 family is found in the light-harvesting complex (LHC) proteins (PF00504), the group of
470 pigment-binding proteins proposed to be the most abundant membrane proteins on
471 Earth. LHCs comprise almost 3% of the total sum of diatom metatranscriptomic
472 relative abundances, in comparison to 9.2% of the 137 Pfam families for ribosomal
473 proteins, and 4.2% of the 47 ubiquitin Pfam domains. Additionally, our top 100 Pfams
474 include ATP synthase subunit C (PF00137), which contributes to carbon-
475 concentrating mechanisms[66], and flavodoxin (PF00258), with 0.24% and 0.20% of the
476 total metatranscriptomic normalized abundances. Beyond photosynthesis, the
477 maintenance of the frustule also appeared to require a high transcriptional effort,
478 reflected by the observation that the silicon transporters (PF03842)[67] were the 40th
479 most transcriptionally abundant family (0.36% of the total sum of metatranscriptomic
480 relative abundances), and are mainly found in diatoms and dictyochophytes (another
481 silicifying group).
482 Among the top 100 Pfams, only one domain of unknown function (DUF) was present:
483 DUF285 (PF03382). This enigmatic tandem repeat protein is distantly related to
484 leucine rich repeats[68] (**Figure 10a)** and has multiple gene copies in some bacteria,
485 giant viruses and in eukaryotic phytoplankton (**Figure S17**). Genes and transcripts
486 coding for DUF285 are notably abundant at high latitudes (**Figure 10b-c**), and this
487 trend extends to phytoplankton beyond diatoms (**Figure S18**). Interestingly, DUF285

488 exhibits a solenoid conformation (**Figure 10a**) similar to proteins with antifreeze
489 activity[69].
490 Other Pfam families displayed specific transcript abundance patterns based on ocean
491 regions, such 'Cold-shock' DNA-binding domains (PF00313) being predominant in AO
492 and SO, and the opposite trends for Heat Shock Factor-type DNA binding domains
493 (PF00447) and Heat Shock Proteins 20, 70 and 90 (HSP20, HSP70 and HSP90;
494 PF00011, PF00012 and PF00183, respectively) **(Figures 9 and and S18)**. Other
495 Pfams more abundant in temperate waters include many of the diatom prevalent
496 families such as dioxygenase (PF00775), retinal pigment epitherial membrane protein
497 (PF03055), subtilase family (PF00082), YHYH protein (PF14240), Zinc finger C2H2
498 type (PF00096), and zinc-binding domain (PF12907).
499 Given the dominance of LHCs, we conducted a detailed analysis of this family. In
500 addition to the canonical members involved in light-harvesting (LHCf, LHCq, and
501 LHCr), this superfamily includes the diatom-specific LHCx and LHCz subfamilies,
502 engaged in photoprotection[70] **(Figure 10d)**. After annotating over 60,000 diatom LHC
503 sequences into these subfamilies, distinct expression patterns emerged. The ratio of
504 transcripts for light-harvesting subfamilies to those for photoprotection subfamilies was
505 higher in larger size fractions (**Figure 10e-g)**. This increase likely serves as a
506 compensatory response to counteract intracellular light attenuation: as cell volume
507 increases, the path length of light within the cell also increases, reducing the light
508 intensity per unit volume. We also observed spatial distinctions in expression related
509 to depth and latitude. Among the pool of LHC transcripts, photoprotective subfamilies
510 were prevalent at the surface, whereas light-harvesting subfamilies were predominant
511 in the DCM (**Figure 10f)**. In addition, photoprotection subfamilies were slightly more
512 prevalent at higher latitudes, which are areas characterized by photo-physiological

513 stresses such as low temperatures and elongated photoperiods (long days in the
514 summer, and long nights in the winter) (**Figure 10e**).
515 To infer adaptation patterns (e.g., gene copy number variations) in different diatom
516 species, we explored the metagenome-assembled genomes (MAGs) recently
517 reconstructed from the same DNA samples used for amplifying the 18S rRNA gene
518 amplicons[71]. Due to assembly and binning challenges, these MAGs do not contain
519 the 18S rRNA gene, so their taxonomy was inferred using protein-coding marker
520 genes[71]. A total of 52 MAGs were assigned to diatoms, with variable ranges of
521 completeness (3-87% BUSCO score[72]). Focusing on the 18 MAGs with over 70%
522 BUSCO completeness, we observed a tendency towards higher gene copy numbers
523 for LHCx, LHCz and DUF285 in MAGs common in high latitudes **(Figure 10h)**.
524 Therefore, the biogeography of the MAGs aligns with their genetic characteristics,
525 further underlining the intricate relationship between genetic adaptation and
526 environmental factors in diatom populations.

527 Conclusions

528 This study utilizes the *Tara* Oceans datasets to provide an extensive analysis of
529 diatom diversity and distributions across the world's ocean, enhanced by functional
530 profiling. We employed parallel metabarcoding analyses of the V9 and V4
531 hypervariable regions, offering a wide range of taxonomic resolution ideal for genus-
532 level diversity research. Despite differences in taxonomic resolution and reference
533 databases, our results revealed significant similarities between V4 and V9 molecular
534 patterns, with 15 genera common among the top 20 in both datasets. Additionally,
535 these molecular methods were corroborated with light microscopy data. Our findings
536 reinforce the role of diatoms as a key photosynthetic group in marine ecosystems,

537 highlighting prevalent genera such as *Chaetoceros*, *Thalassiosira*, *Actinocyclus*, and
538 *Pseudo-nitzschia*. We defined 25 distinct communities of co-occurring diatom lineages
539 with varying environmental preferences, illustrating diverse ecological adaptations.
540 The AO stands out as a hotspot of diatom abundance and diversity with 6 of the diatom
541 communities exclusive to it. Our results provide insights into diatom life strategies from
542 a wide range of ecosystems and illustrate the ecological convergence of diatom
543 communities from distinct oceanic regions regarding the availability of major nutrients.
544 We envision that these results will help to identify indicator species for ocean
545 monitoring, especially since many diatom lineages were predicted to be vulnerable to
546 climate change scenarios[73].
547 In addition to taxonomic insights from metabarcoding, we obtained functional
548 information from metatranscriptomics by analyzing changes in mRNA abundance in
549 environmental samples. Our data indicate distinct gene expression patterns
550 correlating with biogeographical distribution. For instance, diatoms in higher latitudes
551 exhibited increased transcript levels coding for 'Cold-shock' DNA-binding proteins as
552 well as for the enigmatic DUF285, suggesting a link between metabarcoding patterns
553 and gene expression profiles. Globally, the most expressed gene family was LHC, with
554 varying expression patterns among subfamilies related to light harvesting versus
555 photoprotection, reflecting changes in depth, latitude and cell size. To understand
556 adaptation strategies, such as gene copy number variations among diatom species,
557 we analyzed MAGs derived from the same DNA samples used for 18S rRNA gene
558 amplification. The biogeographical distribution of diatom MAGs aligned with their gene
559 copy numbers related to photoacclimation.
560 In conclusion, our study offers a thorough pole-to-pole examination of marine diatom
561 populations, encompassing the entire plankton size spectrum. This research provides

562 critical insights into the impact of global changes and human-induced disturbances on
563 marine ecosystems.

564 Material and Methods

565 *Tara* Oceans sampling

566 Tara Oceans expeditions between 2009 and 2013 performed a worldwide sampling of
567 plankton in the upper layers of the ocean, including surface (SRF; 5 m depth) and
568 deep chlorophyll maximum (DCM; 17–188 m). To capture the whole size spectrum of
569 plankton, a combination of filter membranes with different pore sizes (size-
570 fractionation) was used[36]. Four major eukaryotic-enriched size fractions were
571 collected: pico-nanoplankton (0.8 to 5 µm or 0.8 to 2000 µm; named here 0.8-5/2000
572 µm), nanoplankton (5 to 20 µm or 3 to 20 µm; named here 3/5-20 µm), microplankton
573 (20 to 180 µm), and mesoplankton (180 to 2000 µm)[3,36]. Given the inverse relationship
574 between plankton size and abundance[36], the seawater volumes filtered was increased
575 for larger size fractions (from 10^2 to 10^5 L; see Table 1 and Fig. 5 in Pesant et al.[36]).
576 These plankton samples were leveraged to generate different molecular and optical
577 datasets analysed in the current work (**Figure S2**). Specific details about them and
578 their analyses are described below.

579 Metabarcoding datasets

580 Nucleic acid extraction methods were applied to obtain DNA from the different size-
581 fractionated samples. In order to target eukaryotic plankton diversity and relative
582 abundance, a metabarcoding approach was performed using the 18S rRNA gene as
583 molecular marker. Detailed information about DNA extraction, PCR amplification and
584 Illumina sequencing of metabarcodes are described by[74,75]. Briefly, DNA samples

585 were amplified by PCR targeting the hypervariable region V9 (130 ± 4 base pairs
586 length; primer pair: 1389F 5'-TTGTACACACCGCCC-3' and 1510R 5'-
587 CCTTCYGCAGGTTCACCTAC-3')[76] or V4 (385 ± 4 base pairs length; primer pair
588 TAReuk454FWD1 5'-CCAGCASCYGCGGTAATTCC-3' and TAReukREV3 5'-
589 ACTTTCGTTCTTGATYRA-3')[77] of the 18S rRNA marker gene followed by the high-
590 throughput sequencing of the amplicons[74]. The details of PCR mixes, thermocycling
591 and sequencing conditions are provided in Alberti et al.[75]
592 Resulting paired-end reads were mixed-oriented meaning that both R1 and R2 files
593 are composed by a mix of forward and reverse reads. Paired-end reads were trimmed
594 to remove PCR primer sequences using Cutadapt v2.7[78] and dispatched into four files,
595 2 files for the classical orientation (forward reads in R1 and reverse reads in R2) and
596 2 others for the other orientation (reverse reads in R1 and forward reads in R2). Paired-
597 end reads without both primers were filtered out using the option --discard-untrimmed.
598 Forward and reverse reads were trimmed at position 80 for V9 and at position 215 for
599 V4 and reads with ambiguous nucleotides or with a maximum number of expected
600 errors (maxEE) superior to 2 were filtered out using the function *filterAndTrim*() from
601 the R package *dada2* [37]. For each run and read orientation, error rates were defined
602 using the function *learnErrors*() and denoised using the *dada*() function with pool =
603 TRUE before being merged using *mergePairs*() with default parameters. Mixed
604 orientated reads from the same sample and sequencing replicates were summed
605 together. Remaining chimeras were removed using the function
606 *removeBimeraDenovo*(). Scripts producing the ASV tables are publicly available here:
607 https://gitlab.univ-nantes.fr/combi-ls2n/taradada. ASVs were taxonomically assigned
608 using IDTAXA (50% confidence threshold, no sequence similarity threshold) [38] with
609 the PR2 database version 4.14[39]. The scripts for the taxonomic assignment are

610 publicly available here: https://gitlab.sb-roscoff.fr/nhenry/abims-metabarcoding-
611 taxonomic-assignment/-/tree/v1.0.1. To facilitate a comparison between diatoms and
612 other photosynthetic groups, we annotated the ASVs as phytoplankton—including
613 photosynthetic dinoflagellates and chrysophytes—whenever their taxonomic
614 resolution was adequate for alignment with known phytoplanktonic lineages. This
615 classification was informed by a literature review and expert insights compiled in the
616 trait reference database accessible at
617 https://zenodo.org/doi/10.5281/zenodo.3768950. To reduce the influence of PCR and
618 sequencing errors not detected by the dada2 algorithm, only sequences seen in at
619 least two different samples with at least three copies were retained.

620 Metatranscriptomic and metagenomic datasets

621 We mined the *Tara* Oceans catalog of eukaryotic expressed genes (Marine Atlas of
622 *Tara* Oceans Unigenes version 2 (MATOU.v2[79,80]). Briefly, the catalog was built in the
623 following steps. Nucleic acid extraction methods were applied to obtain eukaryotic
624 polyA RNA from the different size-fractionated samples and subsequent Illumina
625 sequencing was carried out[75,79]. The corresponding metatranscriptomic reads were
626 assembled and clustered at 95% identity[79,80]. The metatranscriptomic reads were
627 mapped onto the sequence catalogue to provide the abundance of each transcript
628 sequence in every sample[79]. Similarly, metagenomic reads from the same seawater
629 samples were mapped onto the sequence catalog to ascertain the abundance of each
630 gene in every sample[79]. The taxonomic assignation of the unigene catalog was carried
631 out by sequence similarity against a custom reference database and functional
632 assignation was performed using the Pfam database[79].
633 For our analysis, we focused on sequences assigned to diatoms and other eukaryotic
634 phytoplankton groups for comparative purposes. It is important to note that we could

635 not exclude heterotrophic dinoflagellate species due to the limited number of reference
636 sequences available for this group. We then retrieved the metatranscriptomic read
637 abundances (rpkm values) of the selected sequences and normalized them to the total
638 read abundance for transcripts of the corresponding phytoplankton taxon in each
639 sample. Additionally, for specific cases of interest such as LHC, DUF285, and HSP90,
640 we also retrieved the metagenomic read abundances and performed the equivalent
641 abundance normalization.

642 Metagenome-assembled genomes (MAGs) dataset

643 MAGs were reconstructed and manually curated from *Tara* Oceans metagenomic
644 reads by Delmont et al.[71] (available at https://www.genoscope.cns.fr/tara/#SMAGs).
645 The MAGs received a geographical assignment based on their read recruitment in the
646 *Tara* Oceans sampling stations after mapping each MAG onto the *Tara* Oceans
647 metagenomic dataset[71]. The estimation of genome completion was performed by
648 retrieving the Benchmarking Universal Single-Copy Orthologs (BUSCO)[72]. The
649 taxonomic annotation was carried out by phylogenetic analyses of eukaryotic core
650 gene markers[71]. For the current analysis, we selected the 18 MAGs assigned to
651 diatoms and with over 70% BUSCO completeness, and we determined their gene
652 copy number for LHCs and DUF285 as described below.

653 Optical microscopy datasets

654 For light microscopy, three ml of each sample (from 20-180 µm size fractions) were
655 placed in an Utermöhl chamber with a drop of calcofluor dye (1:100,000) which stains
656 cellulose, thus allowing to better detect and identify dinoflagellates. Cells falling in 2 or
657 4 transects of the chamber were identified and enumerated using an inverted light
658 microscope (Carl Zeiss Axiophot200) at 400x magnification.

659 ## WGCNA of diatom metabarcoding data

660 We used the R package WGCNA (Weighted Gene Co-Expression Network Analysis)
661 [64] to infer the ASVs co-occurence networks for both V4 and V9 metabarcoding data.
662 To remove redundancy, for every station we pooled the reads for the four size fractions
663 into a single aggregated sample. We discarded the samples without all the four size
664 fractions. A Pearson correlation matrix was computed using the ASVs abundances.
665 We raised the Pearson correlation to a power of 12 and 20, for V4 and V9 respectively,
666 to compute the adjacency matrices and we used these matrices to generate a
667 topological overlap measure (TOM). A hierarchical clustering was finally performed on
668 the TOM matrices to identify 25 modules and a set of unclustered ASVs (grey module)
669 (**Figure 8**). The module eigengenes were also computed for each module identified
670 (grey module included) and then used to generate correlations (and associated *p*-
671 values) between each module and 48 environmental variables retrieved from
672 https://doi.org/10.1594/PANGAEA.875582 (**Figure S19**) (see also below).

673 ## Analysis of LHC and DUF285 in metagenomes and
674 metatranscriptomes

675 We analyzed the gene and transcript abundances of LHCs and DUF285 according to
676 environmental variables by mining the metatranscriptomes and metagenomes
677 generated by *Tara* Oceans. In particular, we searched the MATOU-v2[79,80] catalog for
678 sequences coding for LHCs (PF00504) and DUF285 (PF03382) by running HMMer
679 (version 3.2.1 with the gathering threshold option; http://hmmer.org/) over the
680 translated sequences. We retained only sequences assigned to diatoms, and retrieved
681 the corresponding metagenomic and metatranscriptomic read abundances (rpkm
682 values) and normalized them to the total read abundance for diatom transcripts in each

683 metatranscriptomic sample. We did a similar HMMer search among the gene catalog
684 from diatom MAGs.
685 For the functional classification of LHCs into the main subfamilies (LHCf, LHCq, LHCr,
686 LHCx and LHCz), we performed a phylogenetic placement of the translated
687 sequences on the reference phylogeny described in Kumazawa et al[70]. Protein
688 sequences were aligned with mafft version 6 using the G-INS-I strategy[81]. The output
689 alignment was trimmed in both N- and C-terminal regions to maintain the reference
690 alignment limits. The phylogenetic inference was made using approximate maximum
691 likelihood with FastTree[82]. The sequences were classified according to their grouping
692 in monophyletic branches with statistical support of >0.7 with reference sequences of
693 the same functional group.

694 Plotting and statistical analysis

695 Analyses were carried out in R language (http://www.r-project.org/). Exponentiated
696 Shannon Diversity Index (expH), a diversity index that accounts for both species
697 richness and evenness, was calculated using R library *vegan* [83]. Estimates of total
698 ASV richness were done with the *specpool()* function from the *vegan* library. Graphical
699 analyses were carried out using R library *ggplot2*[84] and treemaps were generated with
700 *treemap* [85]. Maps were generated with *borders()* function in *ggplot2*[84] and
701 *geom_point()* function for bubbles or *scatterpie* package [86] for pie charts. Loess
702 smooth plots with 95% confidence windows were plotted with *geom_smooth*() function
703 in *ggplot2*. Spearman's rho correlation coefficients and *p*-values were calculated using
704 the *cor.test()* function of the *stats* package. Metric multidimensional scaling (NMDS)
705 analysis to visualize Bray–Curtis distances was carried out with the *metaMDS()*
706 command in the R package *vegan*[87], and the influence of environmental variables on
707 sample ordination was evaluated with the function *envfit*() in the same R package,

708 whereas the pie charts were plotted with *scatterpie* package. Partial least square
709 analyses were implemented with the R package *plsdepot* [88].

710 Physico-chemical and biotic parameters

711 Physicochemical parameters were retrieved from
712 https://doi.org/10.1594/PANGAEA.875582 and a full description of the performed
713 measurements is described in Pesant et al[36]. Measurements of temperature,
714 conductivity, salinity, depth, pressure, and oxygen were carried out with a vertical
715 profile sampling system (CTD-rosette) and Niskin bottles. Chlorophyll *a*
716 concentrations, a proxy for total phytoplankton biomass, were measured using high-
717 performance liquid chromatography. Dissolved nutrients (NO_3^-, PO_4^{3-}, Si) were
718 analyzed according to previous methods[89,90]. We complemented the *in situ*
719 measurements with (i) the average intra-annual variation of sea surface temperature
720 (IAV SST) as indicator of environmental instability, which we obtained from the
721 Extended Reconstructed Sea Surface Temperature v5 for years 1997-2017[91], (ii) the
722 annual maximum of nitrate concentration (AM NO_3^-) retrieved from the World Ocean
723 Atlas 2009[92], (iii) silica, nitrate, nitrite, ammonium, and iron levels derived from the
724 ECCO2-DARWIN ocean model[93].
725 For biological predictors, we included the abundances of picocyanobacteria
726 (*Synechococcus* and *Prochlorococcus*), copepods (grazers), and silicifiers. Copepod
727 counts were collected with the WP2 net (mesh size of 200 µm) towed vertically from
728 500 m to the surface (ecotaxa.obs-vlfr.fr/prj/377, ecotaxa.obs-vlfr.fr/prj/378).
729 Picocyanobacteria counts were determined by flow cytometry[12]
730 (data.mendeley.com/datasets/p9r9wttjkm/2). The relative abundances of silicifiers
731 were based on the percentage of reads among V4 and V9 barcode datasets as
732 described previously.

Acknowledgements

We would like to thank all colleagues from the *Tara* Oceans consortium as well as the Tara Ocean Foundation for their inspirational vision. We are grateful to Shruti Malviya and Federico Ibarbalz for providing initial details about the metabarcoding data. This work was granted access to the HPC resources of IDRIS, Jean Zay supercomputer, under the allocation 2020–DARI A0080111497 made by GENCI, and has been supported by the FFEM - French Facility for Global Environment, French Government 'Investissements d'Avenir' programs OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), and PSL Research University (ANR-11-IDEX-0001-02). C.B. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (Diatomic; grant agreement No. 835067), the Horizon Europe projects Marco-Bolo (grant agreement no. 101082021) and BlueRemediomics (grant agreement no. 101082304), Agence Nationale de la Recherche "Phytomet" (ANR-16-CE01-0008) and BrownCut (ANR-19-CE20-0020), and the BNP Paribas Foundation Climate & Biodiversity Initiative. J.J.P.K. acknowledges postdoctoral funding from the Fonds Français pour l'Environnement Mondial, and by Moore–Simons Project on the Origin of the Eukaryotic Cell, Simons Foundation (735929LPI). This article is contribution number xxx of *Tara* Oceans.

Contributions

CB and JJPK designed the study and supervised the project. JJPK and CB wrote the paper with substantial input from LZ, AZ, KC and all co-authors. LZ and AZ helped with new ideas and result interpretation. AZ and ES assisted with the optical microscopy data and interpretation. SC and Erwan Delage built the ASV tables, and

NH performed the taxonomic annotation. KC runned initial exploratory molecular analyses. GB runned the WGCNA calculations, JJPK plotted the results, and CN and JJPK performed manual analysis and provided interpretations. FRJV and Etienne Dvorak runned initial exploratory analysis of metatranscriptomics, and SO helped with result interpretation. JJPK performed the formal analysis and visualization.

Competing interests

The authors declare no competing interests.

Data availability

The authors declare that all the data supporting the findings of this study are publicly available in the following repositories and in the supplementary information files of this paper. The contextual data are available in Pangaea (www.pangaea.de) with the identifier https://doi.org/10.1594/PANGAEA.875582. rDNA 18S metabarcoding data are deposited at the European Nucleotide Archive under accession numbers PRJEB6610 and PRJEB9737, and metagenomic and metatranscriptomic data under PRJEB402, PRJEB9738 and PRJEB9739. The ASV table and the taxonomic annotation for the V4 marker are available in https://doi.org/10.5281/zenodo.7551644 and for the V9 marker in https://doi.org/10.5281/zenodo.7548227. The manually curated MAGs by Delmont et al.[71] are available at https://www.genoscope.cns.fr/tara/#SMAGs. Flow cytometry data are available at https://data.mendeley.com/datasets/p9r9wttjkm/2.

Main Figures

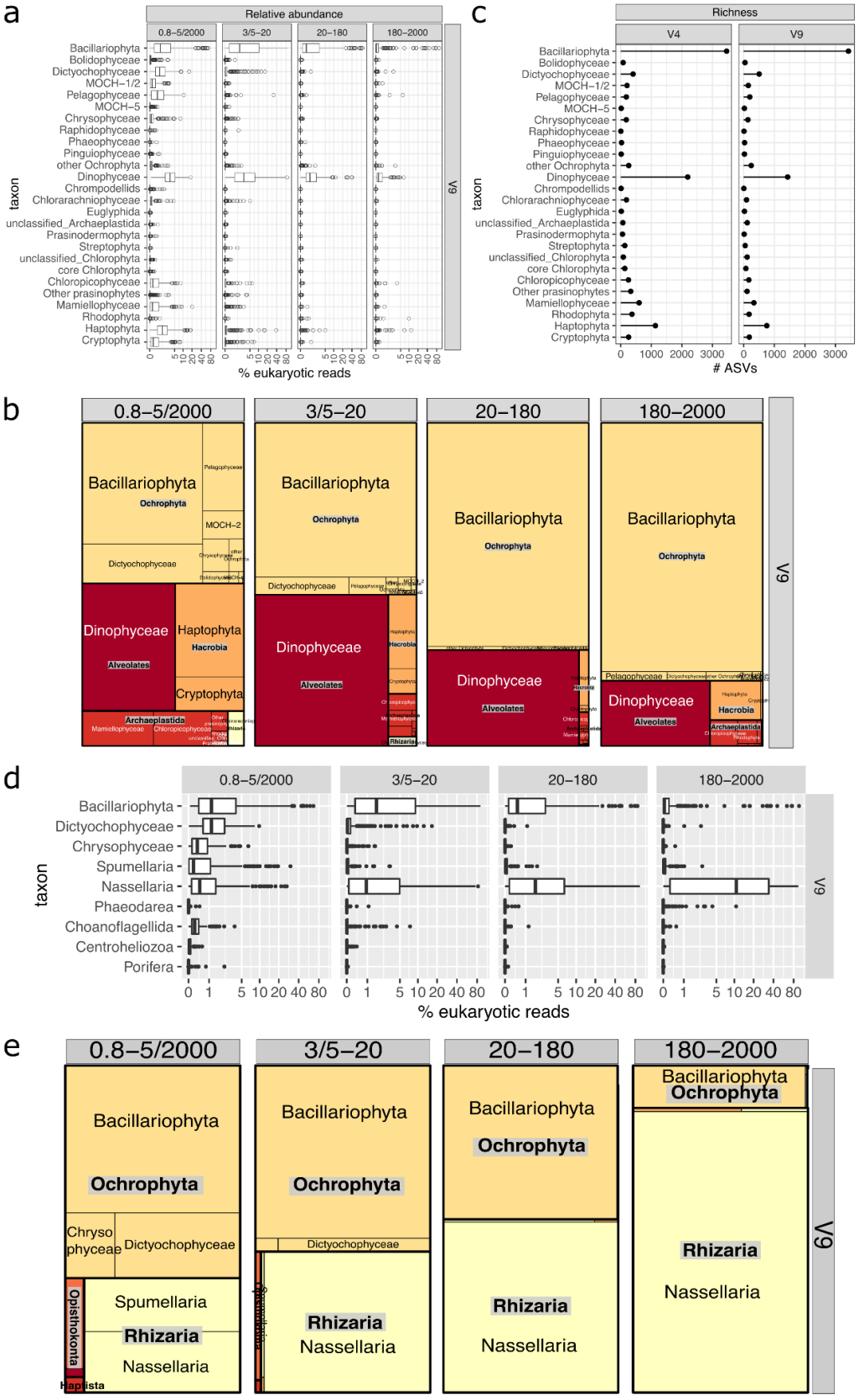


Figure 1: Contribution of diatoms to eukaryotic phytoplankton (a-c) and silicifiers (d-e) in epipelagic waters of the global ocean sampled by *Tara* Oceans using 18S rDNA metabarcoding data obtained from different size-fractions. a-c) Phytoplankton. We focused exclusively on the phytoplankton signal of these data sets based on a functional database (https://zenodo.org/record/3768951#.YM4odnUzbuE), including dinoflagellates and chrysophytes though we acknowledge there are uncertainties in assigning photosynthesis capacity in these groups. a) Relative abundances for V9 marker (log scale). Each point is a size-fractionated sample. b) Sum of normalized reads for V9 marker in the overall dataset. The equivalent panels a and b for the V4 marker are displayed in Fig S3. c) ASV richness for both V4 and V9 marker regions in the overall dataset. d-e) Silificiers. a) Relative abundances for V9 marker (log scale). Each point is a size-fractionated sample. b) Sum of normalized reads for V9 marker in the overall dataset. The equivalent panels a and b for the V4 marker are displayed in Fig S3.

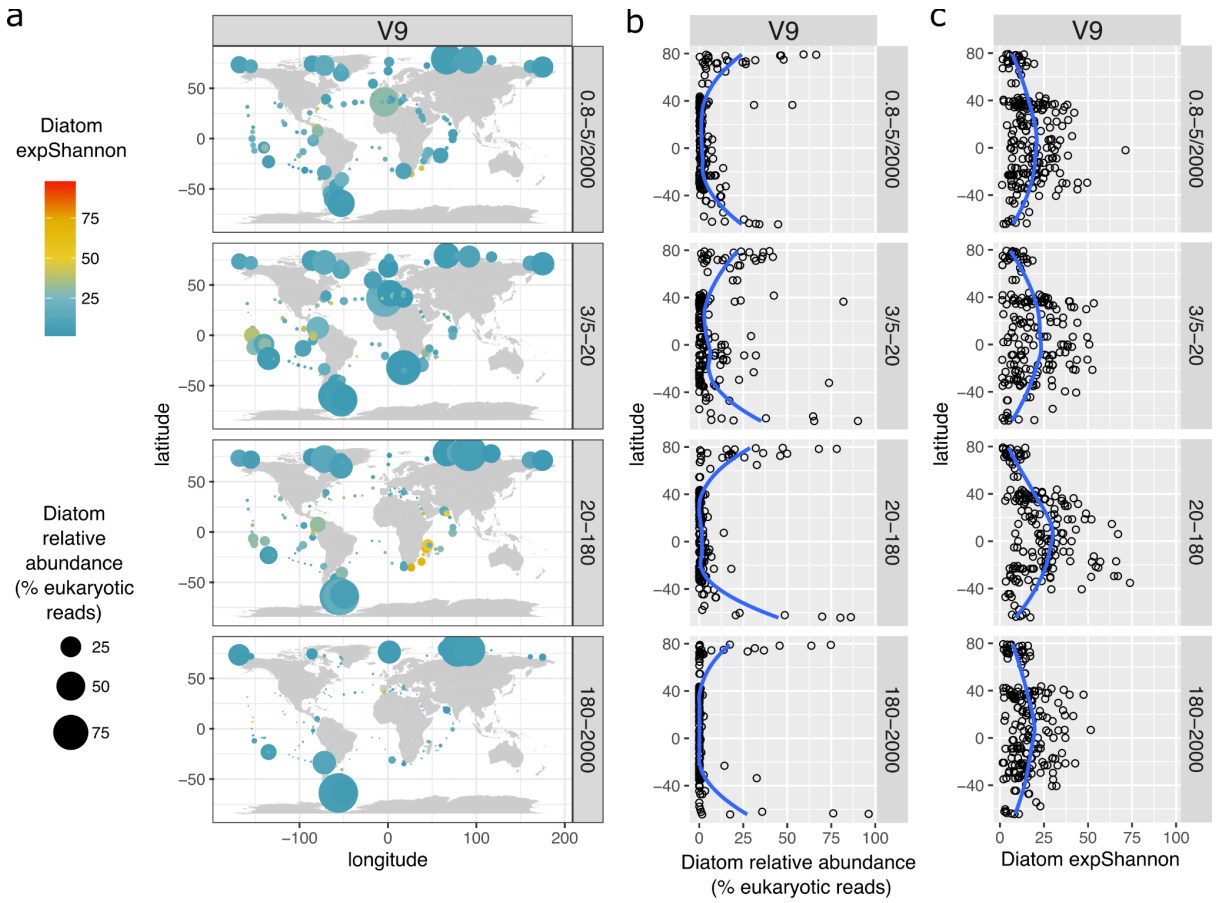


Figure 2: Latitudinal gradients of diatom relative abundance and diversity in surface waters using V9 metabarcoding data obtained from different size-fractions. a) Distribution maps. b) Latitudinal gradient for relative abundance. c) Latitudinal gradient for the exponentiated Shannon Diversity Index. The blue lines correspond to Loess smoothings. The equivalent plots for the V4 marker are displayed in Fig S6.

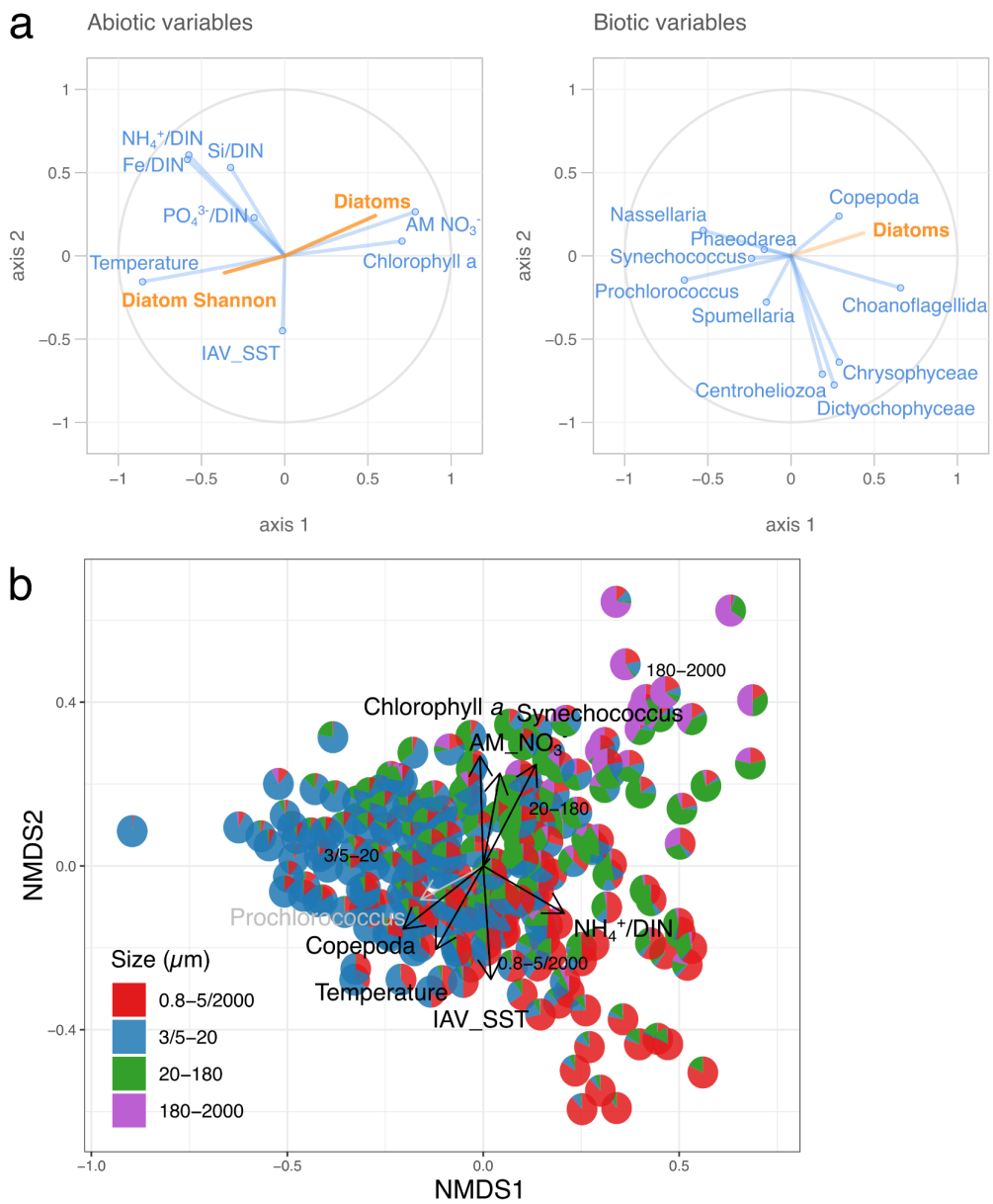


Figure 3: Diatom abundance and diversity in relation to environmental parameters. a) Correlation circle based on the partial least square analysis of V4 and V9 diatom relative abundance and Shannon diversity as responses and a diverse set of abiotic (left) and biotic (right) variables as predictors: annual maximum nitrate concentration (AM NO_3^-), ratios of specific nutrients (ammonia, phosphate, iron, silica) to total dissolved inorganic nitrogen (DIN), temperature, intra-annual variability in sea surface temperature (IAV SST), chlorophyll a concentration, grazers (copepods), picocyanobacteria (*Synechococcus* and *Prochlorococcus*), and silicifiers (the radiolarians of the orders *Nassellaria*, *Spumellaria*, and *Phaeodarea*, the ochrophytes *Chrysophyceae* and *Dictyochophyceae*, choanoflagellates, and centroheliozoa). b) Pattern of the contribution of diatom reads across size fractions in relation with physico-chemical parameters. Non-metric multidimensional scaling (NMDS) analysis of stations according to Bray–Curtis distance between size assemblages, with fitted statistically significant physico-chemical parameters (black, adjusted *p*-value < 0.05; grey, 0.05 < adjusted *p*-value < 0.10). Each pie chart is the V4 or V9 data from a single station. NMDS stress value: 0.1055151.

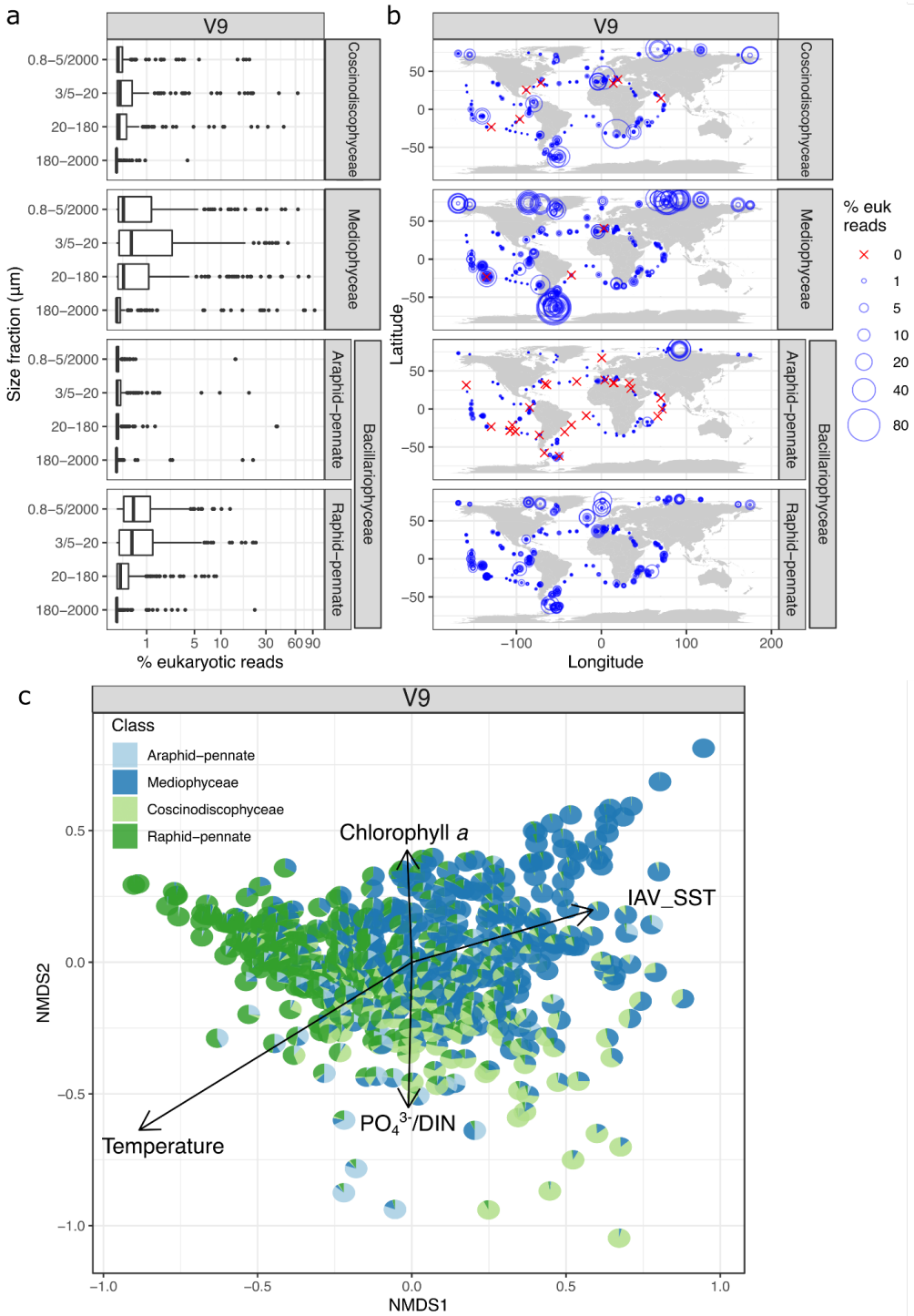


Figure 4: Abundance patterns of diatom classes. a) Relative abundances for V9 marker (log scale). Each point is a size-fractionated sample. b) Biogeography. Each circle is a size-fractionated sample. c) NMDS analysis of stations according to Bray–Curtis distance. Fitted statistically significant physico-chemical parameters (adjusted P value < 0.05) are displayed: ratio of phosphate to total dissolved inorganic nitrogen (DIN), temperature, intra-annual variability in sea surface temperature (IAV SST), and chlorophyll a concentration. Each pie chart is a size-fractionated sample. NMDS stress value: 0.13359. Comparison between V4 and V9 patterns are displayed in **Figure S7**. The maps separated by diatom classes are shown in **Figure S9**.

Figure 5: Summary of diatom genera detected in the V9 and V4 metabarcoding datasets. A total of 76 genera were defined for V4 barcodes and 63 for V9 barcodes. (Column a) Richness

838 expressed as number of ASVs (log scale). (Column b) Abundances expressed as the sum
839 across samples of the relative abundances of rDNA reads (log scale). (Column c) Percentage
840 distribution of rDNA read relative abundances per size fraction. (Column d) Percentage
841 distribution of rDNA read relative abundances per depth. (Column e) Occupancy expressed
842 as the number of stations in which the genus was observed.
843

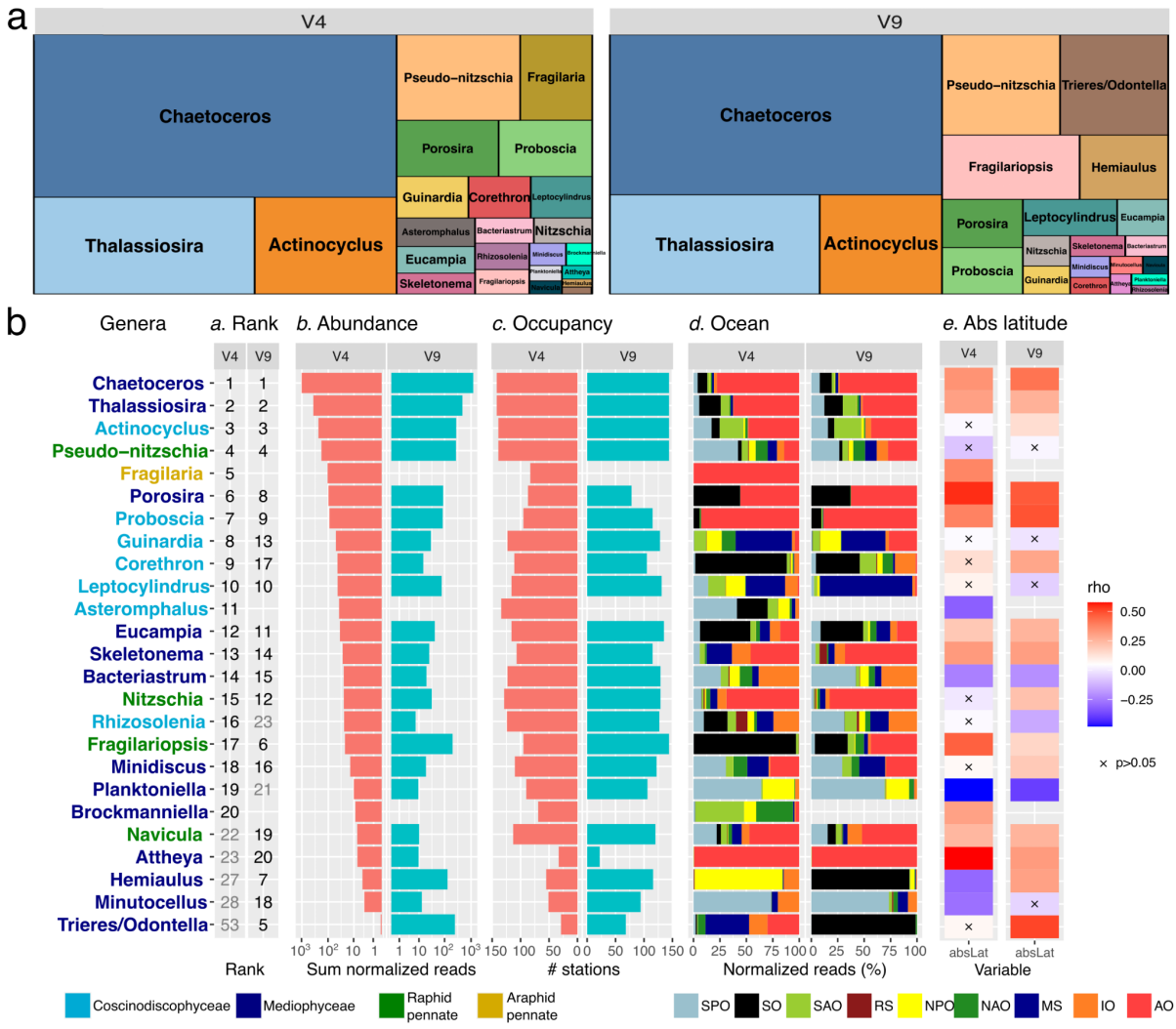


Figure 6: Abundance, diversity and distribution of the top 20 most abundant diatom genera in V4 and V9 datasets. These genera accounted for 92% of assigned reads in the entire V9 dataset and 90% in the V4 dataset when pooling together all size fractions and depths. A) Read taxonomic distribution. B) Diversity and spatial distribution. Genus names are color coded according to diatom class. *a*. Abundance rank of each genus within each dataset (in black when ranking as top 20, otherwise in gray). *b*. Sum of the percentage of reads across all samples (log scale). *c*. Number of stations in which the genera are detected. *d*. Sum of the percentage of reads across oceans. Bars are colour-coded by ocean. *e*. Spearman correlations against absolute latitude. Crosses indicate absence of statistical significance (p>0.05).

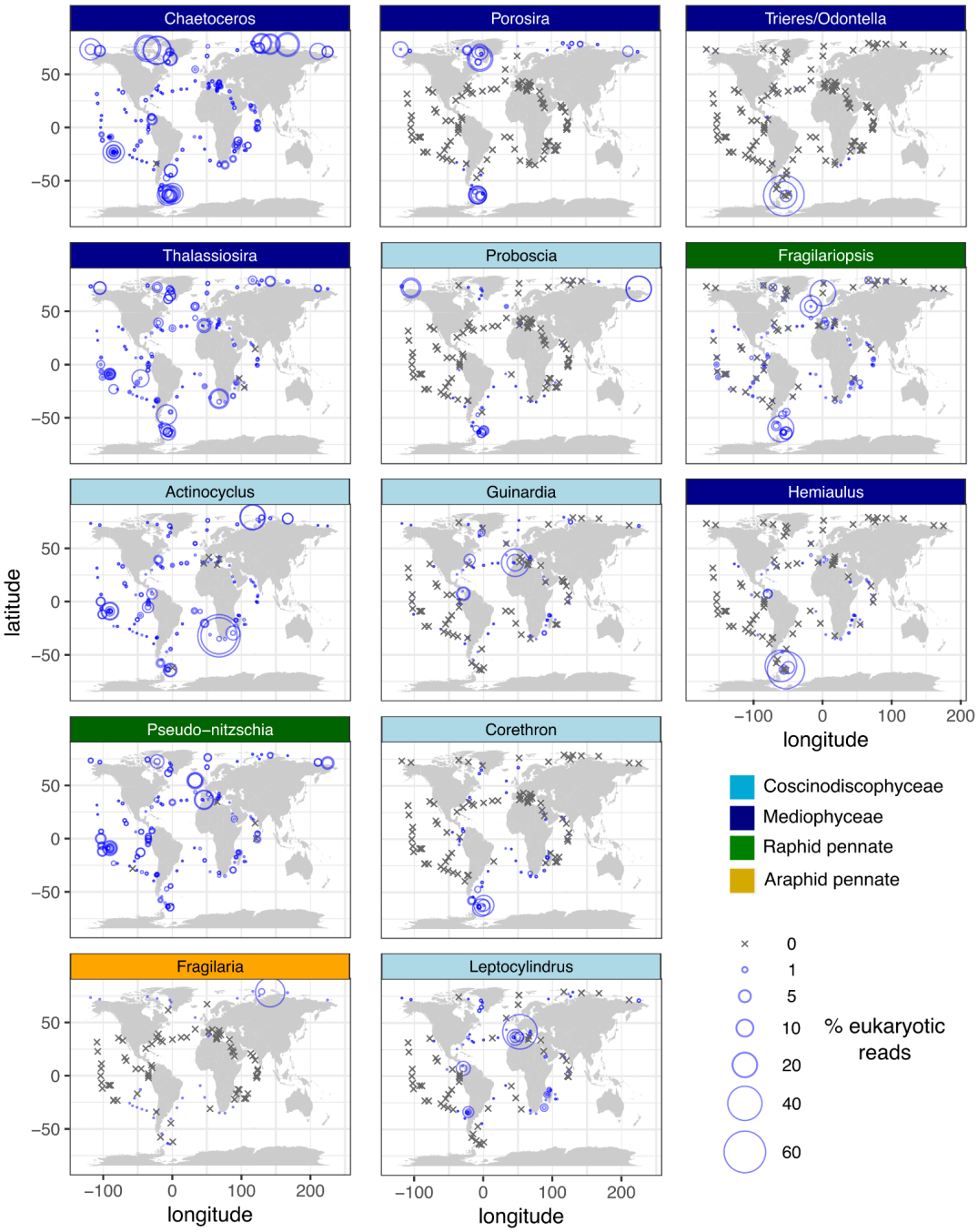


Figure 7: Global distribution of the 10 most abundant genera in each V9 and V4 datasets. These genera account for 88.2% and 82.4% of the assigned reads in the entire V9 and V4 datasets, respectively. Bubble areas are scaled to the % reads for the genus among eukaryotic reads at each station location, whereas crosses indicate absence of detection. The maps correspond to surface samples from the size fraction where the genus was most prevalent: 20-180 µm for *Chaetoceros*, *Fragilaria*, *Porosira*, *Proboscia*, *Corethron*, and *Trieres*/*Odontella*; 3/5-20 µm for *Thalassiosira*, *Actinocyclus*, *Pseudo-nitzschia*, *Guinardia*, *Leptocylindrus*, *Fragilariopsis*, and *Hemiaulus*. The maps include the circles for both V4 and V9 datasets, while separated maps by marker region are displayed in Figure S14. The maps including all size fractions and covering all detected genera can be found in Supplementary File 1.

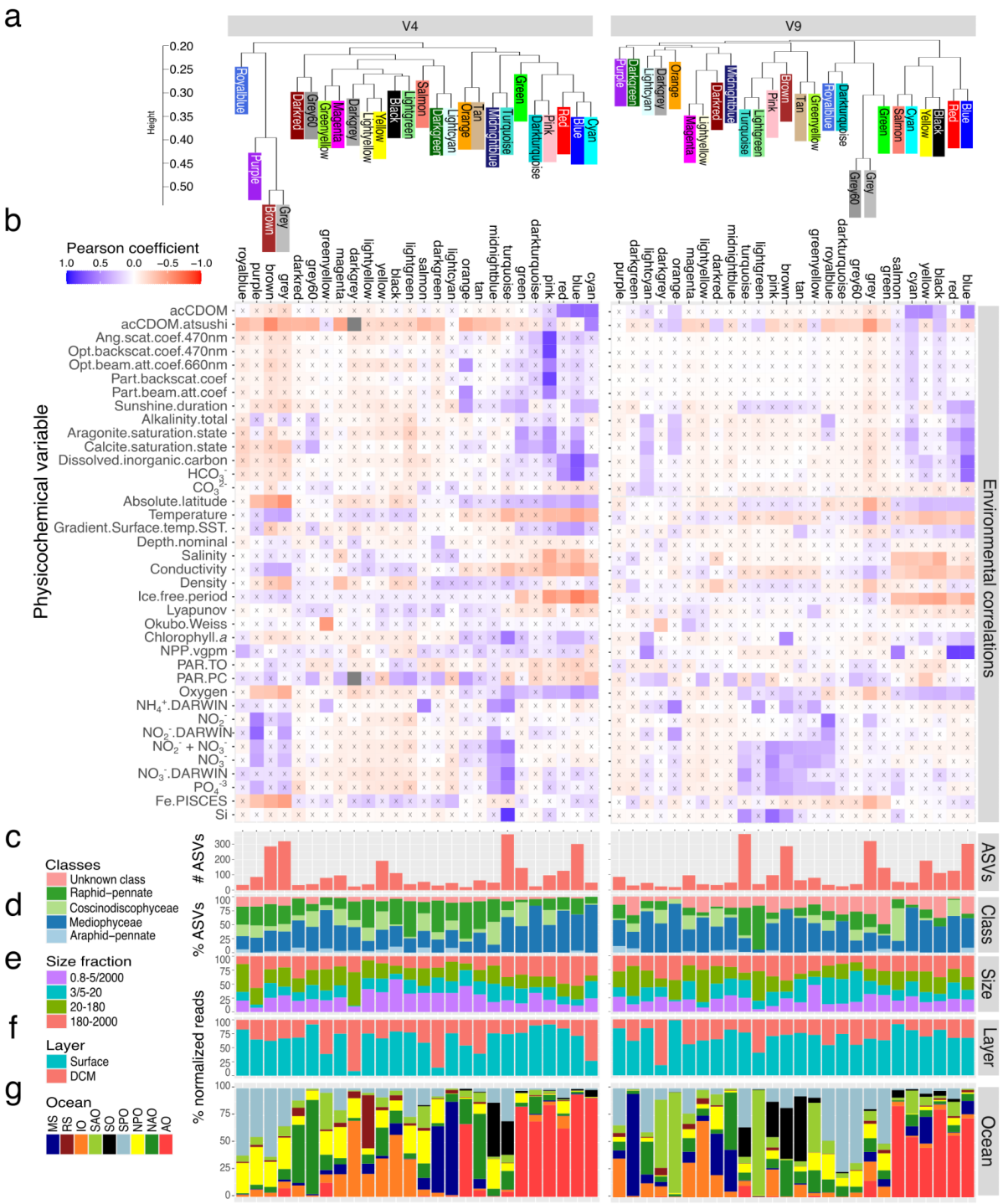


Figure 8. Weighted gene correlation network analysis (WGCNA) of diatom communities for the V4 (left) and V9 datasets (right). *a*. Module clusteringd according to their co-occurrences. *b*. Pearson correlation between module eigenvalues and physico-chemical parameters. Crosses indicate adjusted p-value>0.05. *c*. Total number of ASVs. *d*. Percentage of ASVs for each diatom class. *e*. Sum of the percentage of reads across size fractions. *f*. Sum of the percentage of reads across oceans. *g*. Sum of the percentage of reads across water layers.

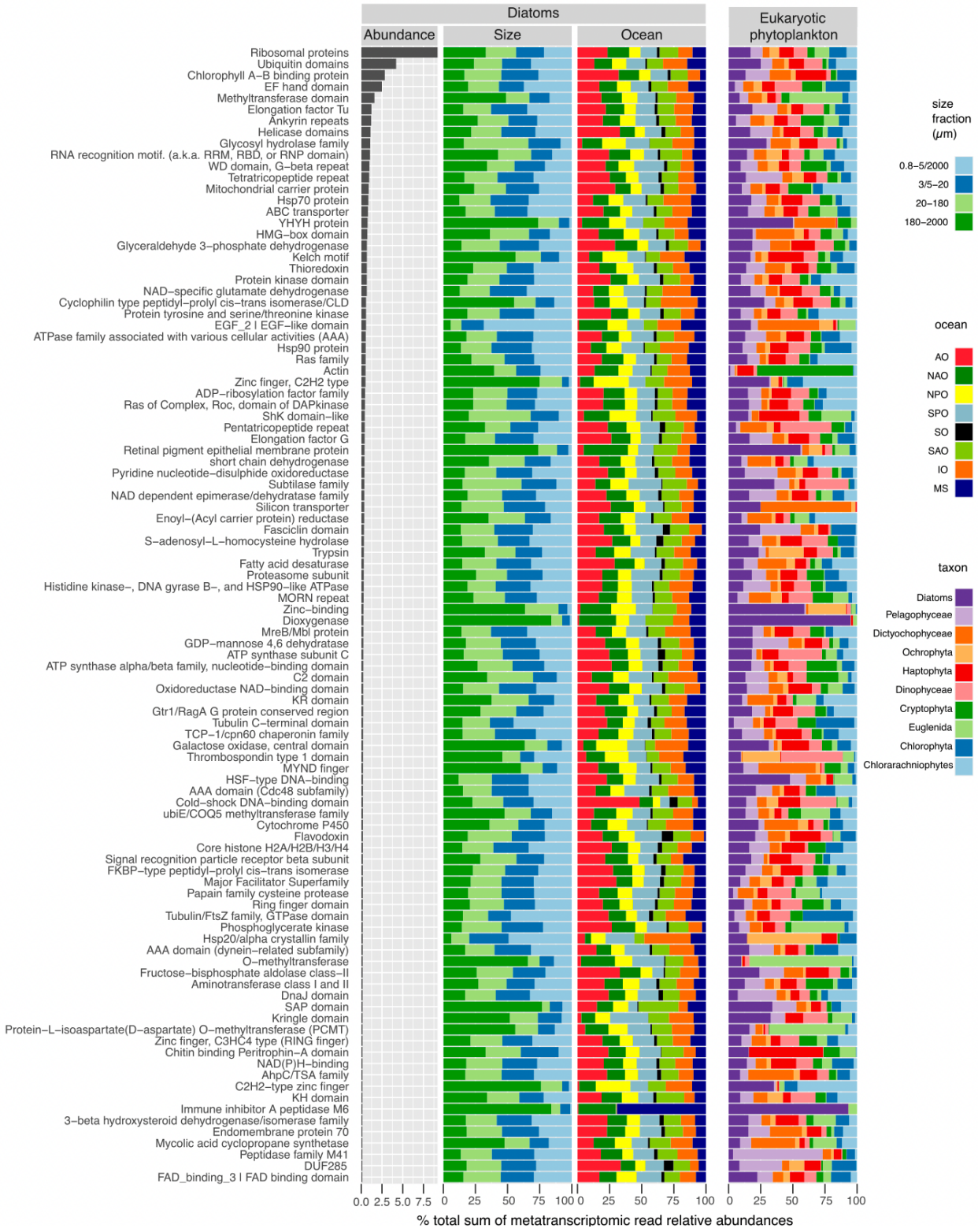
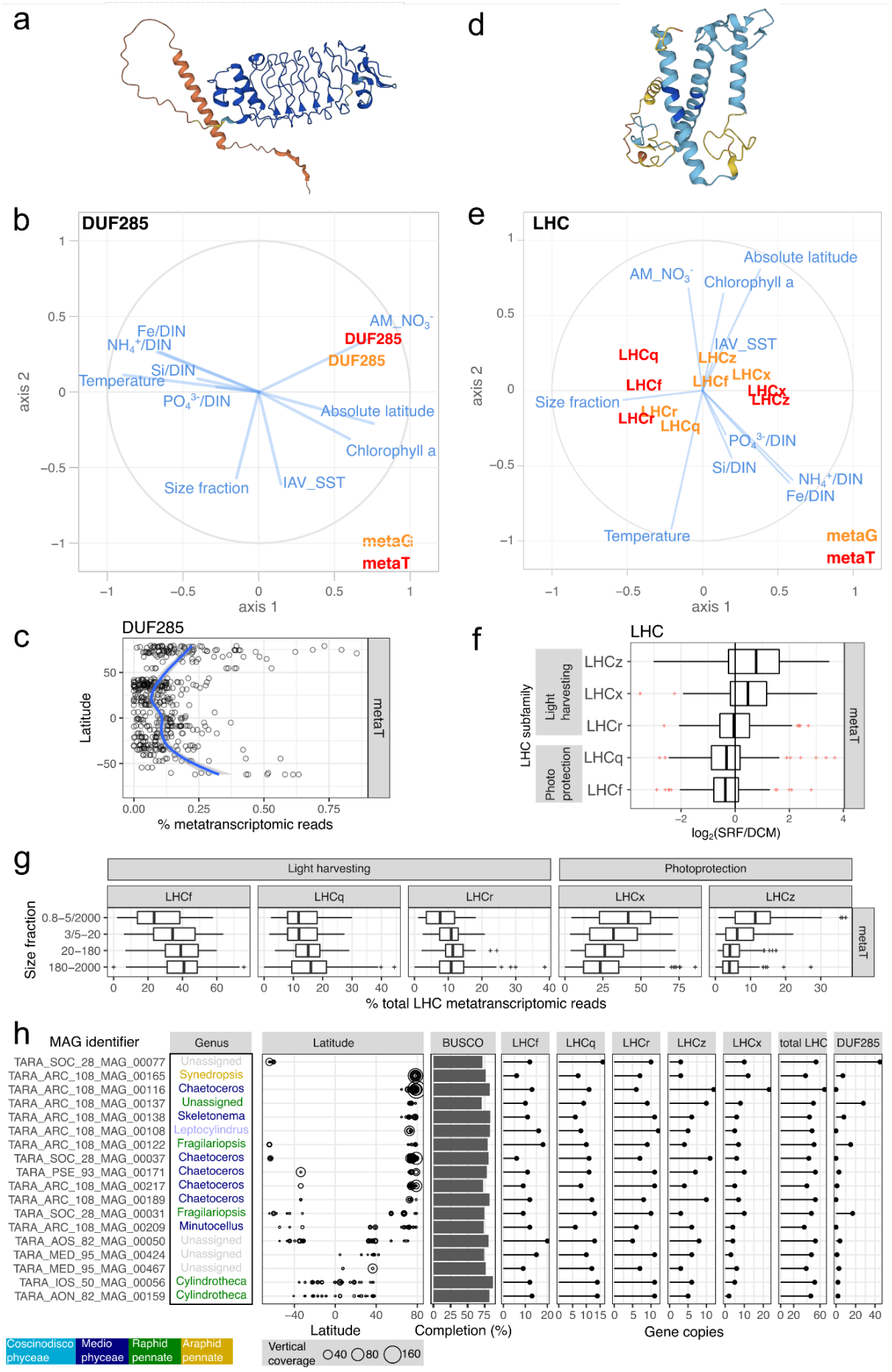


Figure 9: Top 100 most expressed Pfam families among planktonic diatom communities of the global ocean. To reduce redundancy, Pfams for similar functions were merged (e.g., 137 Pfams coding for different diatom ribosomal subunits were merged into "Ribosomal proteins", and 47 Pfams coding for ubiquitin domains were merged into "Ubiquitin domains"). (Column a) Abundance ordered by rank. (Column b) Abundance distribution across size fractions. (Column c) Abundance distribution among ocean regions (AO, Arctic Ocean; NAO, North Atlantic Ocean; NPO, North Pacific Ocean; SPO, South Pacific Ocean; SO, Southern

Ocean; SAO, South Atlantic Ocean; IO, Indian Ocean; MS, Mediterranean Sea). (Column d) Taxonomic distribution across eukaryotic phytoplankton. Note that we were not able to discard heterotrophic species from dinoflagellates due to the small number of reference gene sequences for this group. For each taxon, the relative metatranscriptomic read abundance of each Pfam was summed across all samples. For each sample, the abundance of a given Pfam was calculated as the percentage of total metatranscriptomic reads assigned to diatoms in the corresponding sample.


a
d
b
DUF285
axis 2
axis 1
Fe/DIN
NH4+/DIN
Si/DIN
Temperature
PO43-/DIN
AM_NO3-
DUF285
DUF285
Absolute latitude
Chlorophyll a
IAV_SST
Size fraction
metaG
metaT
e
LHC
Absolute latitude
AM_NO3-
Chlorophyll a
IAV_SST
LHCq
LHCz
LHCf
LHCf
LHCx
LHCx
LHCz
Size fraction
LHCr
LHCf
LHCq
PO43-/DIN
Si/DIN
NH4+/DIN
Fe/DIN
Temperature
metaG
metaT
c
DUF285
Latitude
% metatranscriptomic reads
metaT
f
LHC
LHC subfamily
Light harvesting
Photo protection
LHCz
LHCx
LHCr
LHCq
LHCf
log2(SRF/DCM)
metaT
g
Light harvesting
Photoprotection
LHCf
LHCq
LHCr
LHCx
LHCz
Size fraction
0.8–5/2000
3/5–20
20–180
180–2000
% total LHC metatranscriptomic reads
metaT
h
MAG identifier
Genus
Latitude
BUSCO
LHCf
LHCq
LHCr
LHCz
LHCx
total LHC
DUF285
TARA_SOC_28_MAG_00077 Unassigned
TARA_ARC_108_MAG_00165 Synedropsis
TARA_ARC_108_MAG_00116 Chaetoceros
TARA_ARC_108_MAG_00137 Unassigned
TARA_ARC_108_MAG_00138 Skeletonema
TARA_ARC_108_MAG_00108 Leptocylindrus
TARA_ARC_108_MAG_00122 Fragilariopsis
TARA_SOC_28_MAG_00037 Chaetoceros
TARA_PSE_93_MAG_00171 Chaetoceros
TARA_ARC_108_MAG_00217 Chaetoceros
TARA_ARC_108_MAG_00189 Chaetoceros
TARA_SOC_28_MAG_00031 Fragilariopsis
TARA_ARC_108_MAG_00209 Minutocellus
TARA_AOS_82_MAG_00050 Unassigned
TARA_MED_95_MAG_00424 Unassigned
TARA_MED_95_MAG_00467 Unassigned
TARA_IOS_50_MAG_00056 Cylindrotheca
TARA_AON_82_MAG_00159 Cylindrotheca
Latitude
Completion (%)
Gene copies
Coscinodiscophyceae
Mediophyceae
Raphidpennate
Araphidpennate
Vertical coverage
40
80
160

890

Figure 10: Gene and transcript patterns for the Pfam families DUF285 and LHC among diatom populations. a-c) DUF285 (PF03382). a) Structural prediction (Uniprot id B7FV04). b) PLS analysis for metagenomic (orange) and metatranscriptomic (red) abundances. c) Metatranscriptomic abundance latitudinal gradient. d-g) LHC (PF00504). d) Structural prediction (Uniprot id B7G6Y1). e) PLS analysis for metagenomic (orange) and metatranscriptomic (red) abundances. f) Depth partition between surface (SRF) and deep chlorophyll maximum (DCM). g) Partition across size fractions. h) Gene copy patterns for LHC and DUF285 in metagenomic-assembled genomes (MAGs) assigned to diatoms. The columns correspond to: i) MAG identifier. ii) Class and genus. iii) Latitudinal distribution. Bubbles area is proportional to vertical read coverage. iv) Completeness based on BUSCO[72]. Only MAGs with >70% completeness are displayed. v) Gene copies for LHC subfamilies and for DUF285.

References

1. Mann, D. G. & Vanormelingen, P. An inordinate fondness? The number, distributions, and origins of diatom species. *J. Eukaryot. Microbiol.* **60**, 414–420 (2013).
2. Malviya, S. *et al.* Insights into global diatom distribution and diversity in the world's ocean. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E1516–25 (2016).
3. Pierella Karlusich, J. J., Ibarbalz, F. M. & Bowler, C. Phytoplankton in the Ocean. *Ann. Rev. Mar. Sci.* **12**, 233–265 (2020).
4. Litchman, E., Klausmeier, C. A. & Yoshiyama, K. Contrasting size evolution in marine and freshwater diatoms. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 2665–2670 (2009).
5. Falkowski, P. G., Barber, R. T. & Smetacek, V., V. Biogeochemical Controls and Feedbacks on Ocean Primary Production. *Science* **281**, 200–207 (1998).
6. Rousseaux, C. & Gregg, W. Interannual variation in phytoplankton primary production at A global scale. *Remote Sens. (Basel)* **6**, 1–19 (2013).
7. Tréguer, P. *et al.* Influence of diatom diversity on the ocean biological carbon pump. *Nat. Geosci.* **11**, 27–37 (2018).
8. Tréguer, P. J. *et al.* Reviews and syntheses: The biogeochemical cycle of silicon in the modern ocean. *Biogeosciences* **18**, 1269–1289 (2021).
9. Allen, A. E. *et al.* Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**, 203–207 (2011).
10. Smith, S. R. & Allen, A. E. Comparative and functional genomics of macronutrient utilization in marine diatoms. in *The Molecular Life of Diatoms* 529–566 (Springer International Publishing, Cham, 2022).

11. Caputo, A., Nylander, J. A. A. & Foster, R. A. The genetic diversity and evolution of diatom-diazotroph associations highlights traits favoring symbiont integration. *FEMS Microbiol. Lett.* **366**, (2019).
12. Pierella Karlusich, J. J. *et al.* Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods. *Nat. Commun.* **12**, 4160 (2021).
13. Schvarcz, C. R. *et al.* Overlooked and widespread pennate diatom-diazotroph symbioses in the sea. *Nat. Commun.* **13**, 799 (2022).
14. Tschitschko, B. *et al.* Rhizobia-diatom symbiosis fixes missing nitrogen in the ocean. *Nature* (2024) doi:10.1038/s41586-024-07495-w.
15. Bowler, C., Vardi, A. & Allen, A. E. Oceanographic and biogeochemical insights from diatom genomes. *Ann. Rev. Mar. Sci.* **2**, 333–365 (2010).
16. Brun, P. *et al.* Ecological niches of open ocean phytoplankton taxa. *Limnol. Oceanogr.* **60**, 1020–1038 (2015).
17. Sathyendranath, S. *et al.* Discrimination of diatoms from other phytoplankton using ocean-colour data. *Mar. Ecol. Prog. Ser.* **272**, 59–68 (2004).
18. Alvain, S., Moulin, C., Dandonneau, Y. & Loisel, H. Seasonal distribution and succession of dominant phytoplankton groups in the global ocean: A satellite view. *Global Biogeochem. Cycles* **22**, (2008).
19. El Hourany, R. *et al.* Linking satellites to genes with machine learning to estimate phytoplankton community structure from space. *Ocean Sci.* **20**, 217–239 (2024).
20. Hinder, S. L. *et al.* Changes in marine dinoflagellate and diatom abundance under climate change. *Nat. Clim. Chang.* **2**, 271–275 (2012).
21. Leblanc, K. *et al.* A global diatom database – abundance, biovolume and biomass in the world ocean. *Earth Syst. Sci. Data* **4**, 149–165 (2012).

954 22. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E.
955 Towards next-generation biodiversity assessment using DNA metabarcoding.
956 *Mol. Ecol.* **21**, 2045–2050 (2012).
957 23. Canesi, K. L. & Rynearson, T. A. Temporal variation of Skeletonema community
958 composition from a long-term time series in Narragansett Bay identified using
959 high-throughput DNA sequencing. *Mar. Ecol. Prog. Ser.* **556**, 1–16 (2016).
960 24. Ruggiero, M. V. *et al.* Temporal changes of genetic structure and diversity in a
961 marine diatom genus discovered via metabarcoding. *Environ. DNA* **4**, 763–775
962 (2022).
963 25. Pierella Karlusich, J. J., Lombard, F., Irisson, J.-O., Bowler, C. & Foster, R. A.
964 Coupling imaging and omics in plankton surveys: State-of-the-art, challenges,
965 and future directions. *Front. Mar. Sci.* **9**, (2022).
966 26. Bowler, C. *et al.* The Phaeodactylum genome reveals the evolutionary history of
967 diatom genomes. *Nature* **456**, 239–244 (2008).
968 27. Smith, S. R. *et al.* Evolution and regulation of nitrogen flux through
969 compartmentalized metabolic networks in a marine diatom. *Nat. Commun.* **10**,
970 4552 (2019).
971 28. Falciatore, A., Jaubert, M., Bouly, J.-P., Bailleul, B. & Mock, T. Diatom Molecular
972 Research Comes of Age: Model Species for Studying Phytoplankton Biology
973 and Diversity. *Plant Cell* **32**, 547–572 (2020).
974 29. Grigoriev, I. V. *et al.* PhycoCosm, a comparative algal genomics resource.
975 *Nucleic Acids Res.* **49**, D1004–D1011 (2021).
976 30. Durkin, C. A. *et al.* Frustule-related gene transcription and the influence of
977 diatom community composition on silica precipitation in an iron-limited
978 environment. *Limnol. Oceanogr.* **57**, 1619–1633 (2012).

979 31. Alexander, H. *et al.* Functional group-specific traits drive phytoplankton
980 dynamics in the oligotrophic ocean. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5972–9
981 (2015).
982 32. Pearson, G. A. *et al.* Metatranscriptomes reveal functional variation in diatom
983 communities from the Antarctic Peninsula. *ISME J.* **9**, 2275–2289 (2015).
984 33. Cohen, N. R. *et al.* Diatom transcriptional and physiological responses to
985 changes in iron bioavailability across ocean provinces. *Front. Mar. Sci.* **4**,
986 (2017).
987 34. Lampe, R. H. *et al.* Divergent gene expression among phytoplankton taxa in
988 response to upwelling. *Environ. Microbiol.* **20**, 3069–3082 (2018).
989 35. Kolody, B. C. *et al.* Nitrogen and Iron Availability Drive Metabolic Remodeling
990 and Natural Selection of Diverse Phytoplankton during Experimental Upwelling.
991 *mSystems* **7**, e0072922 (2022).
992 36. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara
993 Oceans data. *Sci Data* **2**, 150023 (2015).
994 37. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina
995 amplicon data. *Nat. Methods* **13**, 581–583 (2016).
996 38. Murali, A., Bhargava, A. & Wright, E. S. IDTAXA: a novel approach for accurate
997 taxonomic classification of microbiome sequences. *Microbiome* **6**, 140 (2018).
998 39. Guillou, L. *et al.* The Protist Ribosomal Reference database (PR2): a catalog of
999 unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy.
1000 *Nucleic Acids Res.* **41**, D597–604 (2013).
1001 40. Morin, S., Gómez, N., Tornés, E., Licursi, M. & Rosebery, J. Benthic diatom
1002 monitoring and assessment of freshwater environments: Standard methods and
1003 future challenges. in *Aquatic Biofilms: Ecology, Water Quality and Wastewater*

1004 *Treatment* 111–124 (Caister Academic Press, 2016).
1005 41. Pierella Karlusich, J. J. *et al.* A robust approach to estimate relative
1006 phytoplankton cell abundances from metagenomes. *Mol. Ecol. Resour.* **23**, 16–
1007 40 (2023).
1008 42. Lavrinienko, A., Jernfors, T., Koskimäki, J. J., Pirttilä, A. M. & Watts, P. C. Does
1009 Intraspecific Variation in rDNA Copy Number Affect Analysis of Microbial
1010 Communities? *Trends Microbiol.* **29**, 19–27 (2021).
1011 43. Fu, R. & Gong, J. Single Cell Analysis Linking Ribosomal (r)DNA and rRNA
1012 Copy Numbers to Cell Size and Growth Rate Provides Insights into Molecular
1013 Protistan Ecology. *J. Eukaryot. Microbiol.* **64**, 885–896 (2017).
1014 44. Pérez-Ortín, J. E. *et al.* Cell volume homeostatically controls the rDNA repeat
1015 copy number and rRNA synthesis rate in yeast. *PLoS Genet.* **17**, e1009520
1016 (2021).
1017 45. Hendry, K. R. *et al.* Competition between silicifiers and non-silicifiers in the past
1018 and present ocean and its evolutionary impacts. *Front. Mar. Sci.* **5**, (2018).
1019 46. Piredda, R. *et al.* Diversity and temporal patterns of planktonic protist
1020 assemblages at a Mediterranean Long Term Ecological Research site. *FEMS*
1021 *Microbiol. Ecol.* **93**, (2017).
1022 47. Qin, Y. *et al.* Effects of error, chimera, bias, and GC content on the accuracy of
1023 amplicon sequencing. *mSystems* **8**, e0102523 (2023).
1024 48. Chen, B. Patterns of thermal limits of phytoplankton. *J. Plankton Res.* **37**, 285–
1025 292 (2015).
1026 49. Olguín Salinas, H. F., Alder, V. A., Puig, A. & Boltovskoy, D. Latitudinal diversity
1027 patterns of diatoms in the Southwestern Atlantic and Antarctic waters. *J.*
1028 *Plankton Res.* **37**, 659–665 (2015).

1029 50. Litchman, E. Resource competition and the ecological success of
1030 phytoplankton. in *Evolution of Primary Producers in the Sea* 351–375 (Elsevier,
1031 2007).
1032 51. Buchanan, P. J. *et al.* Enrichment of ammonium in the future ocean threatens
1033 diatom productivity. (2024) doi:10.22541/essoar.171224451.18675433/v1.
1034 52. Percopo, I., Siano, R., Cerino, F., Sarno, D. & Zingone, A. Phytoplankton
1035 diversity during the spring bloom in the northwestern Mediterranean Sea.
1036 *Botanica Marina* **54**, (2011).
1037 53. Zingone, A., Sarno, D., Siano, R. & Marino, D. The importance and
1038 distinctiveness of small-sized phytoplankton in the Magellan Straits. *Polar Biol.*
1039 **34**, 1269–1284 (2011).
1040 54. Leblanc, K. *et al.* Nanoplanktonic diatoms are globally overlooked but play a role
1041 in spring blooms and carbon export. *Nat. Commun.* **9**, 953 (2018).
1042 55. Arsenieff, L. *et al.* Diversity and dynamics of relevant nanoplanktonic diatoms in
1043 the Western English Channel. *ISME J.* **14**, 1966–1981 (2020).
1044 56. Piredda, R. *et al.* Diatom diversity through HTS-metabarcoding in coastal
1045 European seas. *Sci. Rep.* **8**, 18059 (2018).
1046 57. Sommer, U., Peter, K. H., Genitsaris, S. & Moustaka-Gouni, M. Do marine
1047 phytoplankton follow Bergmann's rule sensu lato? *Biol. Rev. Camb. Philos. Soc.*
1048 **92**, 1011–1026 (2017).
1049 58. Beaugrand, G., Edwards, M., Brander, K., Luczak, C. & Ibanez, F. Causes and
1050 projections of abrupt climate-driven ecosystem shifts in the North Atlantic. *Ecol.*
1051 *Lett.* **11**, 1157–1168 (2008).
1052 59. Medlin, L. K. & Desdevises, Y. Phylogenetic reconstruction of diatoms using a
1053 seven-gene dataset, multiple outgroups, and morphological data for a total

1054 evidence approach. *Phycologia* **59**, 422–436 (2020).
1055 60. Kooistra, W. H. C. F., Forlani, G. & De Stefano, M. Adaptations of araphid
1056 pennate diatoms to a planktonic existence. *Mar. Ecol.* **30**, 1–15 (2009).
1057 61. Kooistra, W. H. C. F. *et al.* Global diversity and biogeography of Skeletonema
1058 species (bacillariophyta). *Protist* **159**, 177–193 (2008).
1059 62. Busseni, G. *et al.* Large scale patterns of marine diatom richness: Drivers and
1060 trends in a changing ocean. *Glob. Ecol. Biogeogr.* (2020)
1061 doi:10.1111/geb.13161.
1062 63. Veselá-Strejcová, J. *et al.* Diverse eukaryotic phytoplankton from around the
1063 Marquesas Islands documented by combined microscopy and molecular
1064 techniques. *Protist* **174**, 125965 (2023).
1065 64. Langfelder, P. & Horvath, S. Fast R functions for robust correlations and
1066 hierarchical clustering. *J. Stat. Softw.* **46**, (2012).
1067 65. Wassmann, P. *et al.* The contiguous domains of Arctic Ocean advection: Trails
1068 of life and death. *Prog. Oceanogr.* **139**, 42–65 (2015).
1069 66. Yee, D. P. *et al.* The V-type ATPase enhances photosynthesis in marine
1070 phytoplankton and further links phagocytosis to symbiogenesis. *Curr. Biol.* **33**,
1071 2541–2547.e5 (2023).
1072 67. Durkin, C. A., Koester, J. A., Bender, S. J. & Armbrust, E. V. The evolution of
1073 silicon transporters in diatoms. *J. Phycol.* **52**, 716–731 (2016).
1074 68. Röske, K. *et al.* A versatile palindromic amphipathic repeat coding sequence
1075 horizontally distributed among diverse bacterial and eucaryotic microbes. *BMC*
1076 *Genomics* **11**, 430 (2010).
1077 69. Banach, M., Konieczny, L. & Roterman, I. Why do antifreeze proteins require a
1078 solenoid? *Biochimie* **144**, 74–84 (2018).

1079 70. Kumazawa, M. *et al.* Molecular phylogeny of fucoxanthin-chlorophyll a/c proteins
1080 from Chaetoceros gracilis and Lhcq/Lhcf diversity. *Physiol. Plant.* **174**, e13598
1081 (2022).
1082 71. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related
1083 eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genom* **2**,
1084 100123 (2022).
1085 72. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E.
1086 M. BUSCO: assessing genome assembly and annotation completeness with
1087 single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
1088 73. Chaffron, S. *et al.* Environmental vulnerability of the global ocean epipelagic
1089 plankton community interactome. *Sci Adv* **7**, (2021).
1090 74. de Vargas, C. *et al.* Ocean plankton. Eukaryotic plankton diversity in the sunlit
1091 ocean. *Science* **348**, 1261605 (2015).
1092 75. Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from
1093 the Tara Oceans expedition. *Sci Data* **4**, 170093 (2017).
1094 76. Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W. & Huse, S. M. A method
1095 for studying protistan diversity using massively parallel sequencing of V9
1096 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One* **4**,
1097 e6372 (2009).
1098 77. Stoeck, T. *et al.* Multiple marker parallel tag environmental DNA sequencing
1099 reveals a highly complex eukaryotic community in marine anoxic water. *Mol.*
1100 *Ecol.* **19 Suppl 1**, 21–31 (2010).
1101 78. Martin, M. Cutadapt removes adapter sequences from high-throughput
1102 sequencing reads. *EMBnet J.* **17**, 10 (2011).
1103 79. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nat. Commun.* **9**,

373 (2018).

80. Zayed, A. A. *et al.* Cryptic and abundant marine viruses at the evolutionary origins of Earth's RNA virome. *Science* **376**, 156–162 (2022).
81. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).
82. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
83. Community Ecology Package [R package vegan version 2.6-6.1]. (2024).
84. Wickham, H. Data Analysis. in *Use R!* 189–201 (Springer International Publishing, Cham, 2016).
85. Tennekes, M. Treemap Visualization [R package treemap version 2.4-4]. (2023).
86. Yu, G. Scatter Pie Plot [R package scatterpie version 0.2.3]. (2024).
87. Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R, Solymos P, et al. vegan: Community Ecology Package. https://CRAN.R-project.org/package=vegan. (2022).
88. Sanchez, G. https://github.com/gastonstat/plsdepot (2015).
89. Bendschneider, K. & Robinson, R. J. *Journal of Marine Research* 11, 87–96 (1952).
90. Murphy, J. & Riley, J. P. *Analytica Chimica Acta* 27, 31–36 (1962).
91. Huang, B. *et al.* Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Clim.* **30**, 8179–8205 (2017).
92. Garcia, H.E., Locarnini, R.A., Boyer, T.P., Antonov, J.I., Zweng, M.M., Bara-nova, O.K., and Johnson, D.R. In World Ocean Atlas, S. Levitus, ed. (U.S. Government Printing Office), p. 398 (2010).

93. Menemenlis, B. D., Campin, J., Heimbach, P., Hill, C., & Lee, T. Mercator Ocean Quarterly Newsletter, 31(October), 13–21 (2008).