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. This comprehensive study sheds light on marine diatom distributions, offering insights to assess impacts of 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 inconsistencies . 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

88 since the 1930s, but is mainly useful for quantifying larger-celled species[20]. The first
89 global ocean database for diatoms was compiled for the MARine Ecosystem DATa
90 (MAREDAT) project, mainly using classical microscopy counts from bottle samples
91 collected in regional studies[21]. Over the last years, genetic surveys have been
92 producing an impressive amount of new data about diatom distributions. These
93 investigations are typically based on PCR amplification and sequencing of a fragment
94 of the small subunit of the rRNA gene from the DNA isolated from an environmental
95 sample (a method hereafter referred to as metabarcoding[22)]. rRNA gene-based
96 metabarcoding has a comparatively deeper taxonomic resolution (including for cryptic
97 species) as well as higher detection power (e.g., for rare species) than routine optical
98 methods[23–25]. Currently, the only available large-scale and standardized molecular
99 survey of marine planktonic diatoms was carried out during the first leg of the *Tara*
100 Oceans expedition (46 sampling sites) by targeting the V9 fragment of the 18S rRNA
101 gene[2]. This study was able to improve diatom diversity estimates, e.g., 43% of the
102 observed genera were not represented in the MAREDAT database. However, there
103 was a poor representation of areas expected to have high diatom abundances, such
104 as polar regions.
105 These taxonomic surveys can be complemented by metagenonomic and
106 metatranscriptomic approaches to infer diatom functional diversity, as well as their
107 adaptation and acclimation mechanisms (i.e., variations in gene and transcript copy
108 numbers, respectively) across environmental gradients. The ecological success of
109 diatoms depends on a range of adaptive attributes in addition to the frustule, such as
110 prominent vacuoles for nutrient storage, ice-binding proteins to withstand harsh polar
111 environments, proton pumping rhodopsins to convert light energy into chemical energy
112 in low-light environments, and a metazoan-like urea cycle for cellular management of

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

130 Results and Discussion

131 Prevalence of diatoms among eukaryotic phytoplankton and 132 silicifiers

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

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

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

168 Diatom ASV diversity

169 We focused on diatoms by selecting only those V9 and V4 ASVs assigned to this
170 group. From 769 samples and almost 385 million V4 reads in total, we obtained a
171 total of >19.5 million reads and 4,998 ASVs for diatom V4, while from 889
172 samples and >1,554 million V9 reads, we obtained >105 million reads and 3,957
173 ASVs for diatom V9. In comparison, the 293 samples of the initial segment of the *Tara*
174 Oceans expedition analysed in Malviya et al.[2] produced less than half the number
175 (1,761) of diatom V9 ASVs.

176 We obtained higher number of diatom ASVs for V4, and accumulation curves
177 and richness estimators indicated that both V4 and V9 ASV diversity approached
178 saturation at a global scale (**Figure S4 and Table S1**). The V4 marker is almost
179 three times longer (385 bp *vs* 130 bp for V9), which increases the potential for
180 genuine variation as well as the likelihood of PCR/sequencing errors that may
181 not be fully detected by the DADA2 algorithm [46,47]. It is important to note that the
182 ASVs can account for variation not only at the species level but also at intraspecific
183 and intragenomic levels.

184 The diatom ASVs represented 34 % and 42 % of the total V4 and V9 ASVs
185 attributed to phytoplankton, respectively (**Figure 1c**). This highlights diatoms as the
186 most diverse group within the global phytoplankton community, followed by

187 dinoflagellates and haptophytes. However, photosynthetic dinoflagellate diversity
188 might have been underestimated because a great number of ASVs cannot be
189 classified as either photosynthetic or non-photosynthetic (e.g., those categorized as
190 'unknown dinoflagellate' or 'unknown Gymnodiniales').

191 Patterns of diatom abundance and diversity

192 When examining diatom reads relative abundance across marine biomes and oceanic
193 regions, the highest values were observed in polar biomes (**Figure S5c-d**).
194 Accordingly, a relative abundance latitudinal gradient towards the poles is clearly
195 observed in all size classes (**Figures 2a-b and S6a-b**). This is consistent with the
196 larger thermal breadths and lower minimal thermal growth of diatoms when compared
197 with cyanobacteria and dinoflagellates[48]. In contrast, diatom diversity follows a
198 latitudinal gradient with a decrease in species richness towards the poles (**Figures 2c**
199 **and S6c**), where diatom communities are likely hyper-dominated by a lower number
200 of species, as observed in previous studies[49].
201 We employed Partial Least Squares (PLS) and correlation analyses to explore
202 environmental factors affecting diatom relative abundance and the Shannon diversity
203 index (**Figure 3a and tables S2 and S3**). Diatom Shannon diversity and relative
204 abundance generally correlate with the same environmental variables, but in opposite
205 directions, e.g., for temperature (Spearman's rho = -0.34 and 0.36, respectively;
206 $p<0.05$). The positive correlation between diatom relative abundance and chlorophyll
207 *a* suggests that diatoms are usually the dominant group in high-biomass phytoplankton
208 communities (Spearman's rho = 0.47, $p<0.05$). We also observed a significant positive
209 association between diatom abundance and nitrate concentration (Spearman's
210 rho = 0.43, $p<0.05$), underscoring nitrate availability as a key factor for diatom growth.
211 In addition, diatom abundance tends to moderately increase with phosphate

212 (Spearman's rho = 0.35, $p<0.05$) and silica (Spearman's rho = 0.25, $p<0.05$), while to
213 decrease with the ratio of ammonia to total dissolved inorganic nitrogen
214 (DIN; e.g. sum of ammonia, nitrite, and nitrate) (Spearman's rho = -0.37, $p<0.05$). The
215 role of biotic interactions is also significant. Diatoms show a negative association with
216 the picocyanobacterium Prochlorococcus (Spearman's rho = -0.35, $p<0.05$),
217 supporting findings that diatoms can be outcompeted in nitrogen-limited environments
218 where cyanobacteria efficiently utilize ammonia[50,51], often leading to low chlorophyll *a*
219 levels. In contrast, grazing organisms such as copepods exhibit a weak positive
220 relationship with diatoms (Spearman's rho = 0.18, $p<0.05$), suggesting grazing top-
221 down control of diatom populations. Lastly, the presence of abundant silicified
222 radiolarians (Nassellaria, Spumellaria, and Phaeodarea; Spearman's rho = -0.52, -
223 0.23, and -0.38; respectively, $p<0.05$)) negatively correlates with diatom abundance,
224 likely reflecting competitive interactions for silica utilization[45].

225 Diatom size patterns

226 As a reflection of their wide size range, diatoms were relatively abundant in the
227 nanoplankton (3/5-20 µm; mean relative abundance of ~9% of eukaryotic reads), but
228 also in the piconanoplankton (0.8-5/2000 µm; ~5% mean abundance) and in the
229 micro- and mesoplankton (20-180 µm and 180-2000 µm; ~5% and 3% mean
230 abundances, respectively) (**Figures 1a, S3a and S5a**). Historically, the abundance of
231 smaller diatoms has been underestimated due to detection challenges using
232 microscopy, but recent morphological and molecular studies support our observations
233 of their prevalence in various marine regions[52–55]. The broad size spectrum of diatoms
234 is further evidenced by previous research on their cell volumes, spanning from 3 μm^3
235 (*Thalassiosira* sp.) to 4.71×10^9 μm^3 [4,21]. It is worth noting that we detected in many
236 cases the same diatom species across all size fractions. Although we cannot discard

237 the possibility of filter clogging or extracellular DNA adsorbed on particles of varying
238 sizes, this can also be attributed to life-cycle variations (e.g., asexually dividing
239 diatoms decreasing in cell size, rare gamete formation), and the physical
240 characteristics of certain species (e.g., long needle-like cells and broken colonies
241 passing through small mesh sizes)[56].
242 There is a clear environmental partition between diatom sizes, with larger diatoms
243 associated with polar regions rich in nitrate, whereas small diatoms are prevalent in
244 oligotrophic and temperate waters with high ratios of ammonia to DIN (**Figure 3b**).
245 This is in line with Bergmann's rule of reduced organism body size in warmer
246 environments[57]. Thus, the energy transfer from diatoms to higher trophic levels may
247 be reduced in temperate environments due to their smaller sizes[58].

248 Environmental distribution of diatom classes

249 Following Medlin and Desdevises 2020[59], diatoms are classified into three
250 monophyletic classes: Coscinodiscophyceae (basal radial centric diatoms),
251 Mediophyceae (polar centric diatoms and radial centric Thalassiosirales), and
252 Bacillariophyceae (pennate diatoms). Pennate diatoms are further divided into raphid
253 (Bacillariophycidae) and araphid genera (i.e., subclasses Urneidiophycida and
254 Fragilariophycidae).
255 For both V4 and V9, Mediophyceae were the most prevalent, followed by raphid
256 pennates and Coscinodiscophyceae (**Figures 4a and S8a)**. Araphid pennates were
257 considerably less abundant and only prevalent in an ice edge station in the AO
258 (TARA_188) (**Figure S9)**. Our observation of low abundance of araphid pennates is
259 consistent with their typical presence in benthic rather than planktonic
260 environments[1,60].

261 Mediophyceae were highly abundant in regions characterized by high chlorophyll *a*
262 concentrations and high latitudes **(Figures 4b-c, S7, S8b, and S9)**. Similarly,
263 Coscinodiscophyceae were also notably abundant in high latitude regions. In contrast,
264 raphid pennates were predominantly found in temperate and stable environments,
265 although they were also present in significant numbers at some high latitude stations.
266 There was also a compositional shift among the different size fractions of plankton.
267 The relative abundance of Mediophyceae was highest among the piconano-, nano-,
268 and microplankton **(Figures 4a and S7a)**. Although their relative abundance
269 decreased among mesoplankton, Mediophyceae continued to be the dominant diatom
270 class in this size range **(Figure S8a)**. In contrast, the relative abundance of raphid
271 pennate diatoms demonstrated a continuous decline as the size fractions increased.
272 Coscinodiscophyceae exhibited a predominant presence particularly within the nano-
273 and to a lesser extent, microplankton size fractions, with their numbers decreasing in
274 the piconano- and, particularly , the mesoplankton. All these biogeographical and
275 size patterns are driven by the patterns of the main genera within each class (see
276 below).

277 Diatom genera composition

278 We found a total of **75** genera for V4 barcodes, and **64** genera for V9 barcodes (**Figure**
279 **5**). The higher number of genera generated by V4 barcodes is probably related to its
280 higher taxonomic resolution and better representation in the reference database (PR2
281 v4.14[39]).
282 High variations across genera were observed in terms of abundance and richness.
283 *Chaetoceros* was the most abundant genus, followed by *Thalassiosira*, *Actinocyclus*,
284 and *Pseudo-nitzschia* (**Figure 5**; see also next section). Intragenus richness ranged
285 from 571 V9 ASVs and 945 V4 ASVs for *Chaetoceros* to as low as 1 ASV for some

286 other genera (**Figure 5**). This result reflects the differences in known species for each
287 genus (e.g., >100 for *Chaetoceros*, a few for *Arcocellulus* and *Lauderia*, and only 1 for
288 *Tenuicylindrus* and *Pseudohimantidium*). Additionally, this result parallels read
289 abundance, indicating a greater potential of capturing intraspecific and intragenomic
290 variations of these multicopy markers, as well as a higher likelihood of
291 PCR/sequencing errors.
292 We found ~65% of cumulative read relative abundances in surface samples (**Figure**
293 **5**), which reflects the sampling bias (513 samples from surface and 267 from DCM;
294 **Figure S5b**). When only focusing on locations with both SRF and DCM samples, we
295 observed that most genera were more abundant in surface waters (e.g, *Eucampia*,
296 *Thalassiothrix*, *Rhizosolenia*, *Proboscia*) or equally abundant in surface and DCM. The
297 only genus enriched in DCM was the centric diatom *Attheya*, which is restricted to AO
298 waters (**Figure S10**).
299 A wide range of genera were identified among the piconanoplankton communities.
300 Besides the expected genera such as *Minidiscus*, *Minutocellus*, *Arcocellulus*,
301 *Brockmanniella*, *Cyclotella*, and *Cocconeis* (**Figure 5**), we observed several typically
302 larger genera traditionally classified as microplankton (20-180 µm). These included
303 the centric genera *Hemiaulus*, *Dactyliosolen*, *Detonula*, as well as the pennate genera
304 *Pseudo-nitzschia*, *Surirella*, *Thalassionema*, and *Triceratium*, which often form chains
305 and are generally larger. Their presence in smaller plankton samples could be due to
306 broken cells, cells passing through the mesh by their shorter dimension, or possibly
307 being gametes.

308 Biogeography of the most abundant diatom genera

309 We focused on the top 20 most abundant genera (**Figure 6a**) since they accounted
310 for the majority of the signal from known taxa, with 97 % of assigned reads in the

311 entire V4 dataset and 98 % using V9 . We compared two normalization
312 strategies—normalization by total eukaryotic reads and by total diatom reads—and
313 summed the relative abundances across all size-fractionated samples (**Figure 6a**).
314 The normalization choice presents a trade-off: normalizing by eukaryotic reads tended
315 to over-represent taxa abundant in high-latitude regions dominated by diatoms
316 (*Fragilaria, Stellarima,* and *Attheya*); while diatom-based normalization over-
317 represented taxa from temperate waters where overall diatom abundance was lower
318 (*Haslea, Coscinodiscus, Cyclotella,* and *Pleurosigma*) (**Figures S6a and 7**). Despite
319 these biases, the majority of genera appear consistently in the top 20 list across both
320 normalization approaches.
321 Among the top 20 most abundant genera when based on total eukaryotic reads, 15
322 were consistently found in both V4 and V9 datasets, including *Chaetoceros*,
323 *Thalassiosira*, *Actinocyclus*, *Pseudo-nitzschia*, *Porosira*, *Proboscia*, *Corethron*,
324 *Leptocylindrus*, Guinardia, *Eucampia*, *Skeletonema*, *Bacteriastrum*, Minidiscus ,
325 *Fragilariopsis*, and Planktoniella (**Figure 6a**). Six genera appear only in one
326 dataset but still rank high in the other (between positions 21 and 29 : *Nitzschia*,
327 *Rhizosolenia*, *Hemiaulus*, *Minutocellus, Attheya, and Stellarima*), indicating
328 the robustness of our results. Nonetheless, a few differences were observed, with
329 *Fragilaria*, *Asteromphalus* and *Brockmanniella* found only in V4, and *Trieres*/*Odontella*
330 mainly detected in V9 (**Figure 6a**). These discrepancies could be due to differences
331 in the coverage and quality of reference databases, the resolving power of the DNA
332 barcode, and variations in PCR amplification efficiency among different taxa. The top
333 20 genera based exclusively on analyzing diatom-specific reads were similarly
334 consistent between V4 and V9 datasets (14 shared genera, and seven top 20 genera
335 only in one dataset but still ranking high in the other) (**Figure 6a**).

336 We found 6 new genera among the V9 top 20 in comparison with the previous *Tara*
337 Oceans V9 survey (see Figure 6A in Malviya et al.[2]), which did not include AO, NPO
338 and most of NAO (**Figure S1b**). As expected, the list contains genera prevalent in
339 these new regions: *Bacteriastrum* was abundant in NPO, SPO and IO; *Porosira* and
340 *Skeletonema* were abundant in AO (despite *Skeletonema* being previously reported
341 as absent from polar waters[61]) (**Figures 6a and 7**). However, the remaining three
342 genera were more abundant in SO: *Hemiaulus, Trieres*/*Odontella* and Stellarima
343 (**Figures 6a and 7**). Although SO was already included in Malviya et al.[2], the higher
344 regional coverage of the present work explains this result for *Trieres*/*Odontella* (6
345 current stations *vs* 2 stations in the previous survey), and the updated reference
346 database for *Hemiaulus and Stellarima*.
347 Given that the 20-180 µm size fraction was also quantified using optical
348 microscopy[2,62,63] (98 samples from surface and/or DCM from 66 stations), we
349 compared the most abundant genera between methods in this size fraction **(Figure**
350 **6b**). The comparison resulted in a high degree of similarity: 11 of the top 20
351 genera based on microscopy are also in the top 20 in both molecular datasets, plus
352 six top 20 genera in the microscopy data that are only in one molecular dataset.
353 Some of the differences are explained by the scarcity or lack of references for some
354 genera (e.g., *Cylindrotheca* and *Thalassionema)* or the difficulty of detecting and/or
355 identifying some genera by microscopy *(*e.g., *Arcocellulus, Minidiscus* and *Haslea).*
356 A particular case is *Fragilaria*, whose ASVs were primarily found at an Arctic ice
357 station, where the morphologically similar and closely related *Fossulaphycus*
358 (previously named *Fossula*) was abundant in microscopy observations. While
359 *Fragilaria* is considered freshwater, *Fossulaphycus* is typically associated with Arctic
360 sea ice[91]. Therefore, the detected ASVs assigned to *Fragilaria* may actually be

Fossulaphycus, but we retain the name *Fragilaria* due to the lack of reference sequences of *Fossulaphycus.*

We observed a high variability in the biogeography of the most abundant diatom genera (**Figures 7 and S11 and Supplementary File 1**). Some of them, such as *Porosira*, *Fragilaria*, *Trieres*/*Odontella*, and *Attheya*, were regionally constrained, i.e., highly abundant in a few specific stations (Figures 7 and S11). The top 10 genera, which constituted between 82 and 94% of total diatom reads, tended to be more abundant in high latitudes. There was a group of cosmopolitan genera whose abundances were much higher in polar regions: *Chaetoceros*, *Thalassiosira*, *Actinocyclus* and *Fragilariopsis* . *Pseudo-nitzschia* and *Nitzschia*, other cosmopolitan genera , were 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 , *Corethron*, *Trieres*/*Odontella*, and *Hemialus* in the SO . Finally, some genera were prevalent in a few stations from tropical and subtropical regions: *Guinardia Leptocylindrus, Bacteriastrum, and Asteromphalus*. The top 11-20 diatom genera, which represented roughly 5-10% of total diatom reads, also included cosmopolitan taxa (*Eucampia*, *Skeletonema*, *Rhizosolenia*) and those with more regionally specific distributions (*Attheya* in AO, *Stellarima* in SO), but also a noticeable increase in genera that were abundant in temperate waters (*Cyclotella*, *Haslea, Planktoniella, Brockmaniella, Coscinodiscus, Minutocellus, Pleurosignma*). 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 environmental variables associated to them (**Figures 8 and S11**). The modules were clustered according to their co-occurrences into 4 main groups: i) one formed by 6 modules prevalent in the AO and dominated by Mediophyceae ASVs, ii) another formed by 4-5 modules mainly detected in the SPO (and to a lesser extent in the NPO) and dominated by Mediophyceae and raphid pennate ASVs, iii) another group formed by 2-4 modules abundant in the SO/SPO and with Mediophyceae and raphid pennate diatoms, and iv) a group of different basin-specific modules detected in tropical and subtropical regions, including 2 modules prevalent in the IO, 1-2 in the MS, 2 in the NAO (only in V4), and 2 in the SAO (only V9). Therefore, we found diatom sub-communities with a significant association to each oceanic basin or to basins that are physically connected (e.g., the SPO and IO via the Indonesian Passage). The AO stands out from the rest as it contains a relatively high-number of AO-specific sub-communities. The reasons why the AO is rich in endemic diatom communities might be related to its particular physical oceanography (the influx from large rivers, and the Atlantic and Pacific Ocean waters), long residence times, and the extreme seasonality in day length and sea ice cover, as well as low temperatures and variable salinity[65].

Overall, *Chaetoceros*, *Pseudo-nitzschia* and *Thalassiosira*, previously identified as the most prevalent diatom genera, were found within almost all modules, together with unidentified Bacillariophyta, Mediophyceae and raphid pennates (**Figure S15**).

For both V4 and V9, the turquoise and grey modules gathered the most ASVs, with the grey modules containing a set of unclustered ASVs that were the most enriched in

410 different taxa. Both grey modules were negatively correlated with iron, oxygen and
411 latitude while positively correlated with temperature (**Figure 8**). Together with the fact
412 that they harbored diatoms from a diverse range of basins and sizes, the results
413 suggest that the grey modules represent a wide group of rather temperate diatoms;
414 while the absence of other significant correlations of these modules with environmental
415 parameters indicates potential generalist species, or that the wide biological diversity
416 of these modules precludes the identification of a general pattern.
417 Exploring the association patterns of diatom sub-communities with environmental
418 parameters revealed 3 modules positively correlated with nutrient (nitrate, silica,
419 phosphate) availability in both datasets. The V4 turquoise module, enriched in
420 *Thalassiosira*, *Porosira* and *Chaetoceros* (~10% each), was positively correlated with
421 nitrate, silica, phosphate and chlorophyll *a*. Conversely, both V4 modules midnightblue
422 – enriched in *Pseudo-nitzschia* (16%) *Actinocyclus* (14%) and *Corethron* (12%) – and
423 purple – enriched in *Actinocyclus* (10%) and *Chaetoceros* (9%) – displayed a similar
424 positive correlation with nitrate and phosphate while the latter was negatively
425 correlated with iron and latitude. This potentially indicates that these diatom sub-
426 communities are the most responsive to nutrient availability while being associated
427 with different oceanic regions – either SO (midnightblue, turquoise) or SPO/NPO
428 (purple).
429 On the other hand, all V9 pink, brown and turquoise modules, which all harboured
430 species representative of SO basins and clustered together, were significantly
431 correlated with the three nutrients but were enriched in a different set of species. While
432 both V9 pink (SO/IO) and brown (SO/NAO) modules contained larger diatoms and
433 showed a high proportion of *Chaetoceros* ASVs (49% and 13% of ASVs, respectively,
434 ranked 2nd and 3rd), the latter also displayed a significant contribution of *Odontella*

435 ASVs (7%) and was positively correlated with chlorophyll *a* and ammonium. In
436 contrast, the V9 turquoise (SO/SPO) module was enriched in *Hemiaulus* and
437 *Thalassiosira* ASVs (~10% and 7%). This therefore suggests the existence of 3 distinct
438 SO-associated diatom sub-communities similarly responding to the availability of
439 major nutrients but differing by their size, taxonomic composition and response to
440 other environmental factors.
441 Both V4 and V9 blue modules were enriched in *Chaetoceros* and *Thalassiosira* genera
442 from the AO, corroborating their negative correlation with temperature, and showed
443 positive correlations with dissolved inorganic carbon and chlorophyll *a*. The V9 blue
444 module was also positively correlated with net primary productivity, as was the V4
445 royalblue module (~15% *Bacteriastrum*), illustrating the diatom taxa associated with
446 the most productive regions. These could correspond to either species outcompeted
447 in conditions of nutrient scarcity, and/or to taxa highly contributing to net primary
448 productivity.
449 Finally, the V4 lightcyan module, which gathered very few ASVs (47) and notably from
450 the genera *Chaetoceros* and *Guinardia*, was the most associated with iron but showed
451 no other significant pattern, which could either indicate taxa specialized in iron uptake
452 or species with elevated iron quotas. Altogether, these analyses provide novel insights
453 into factors regulating the assemblage of diatom communities and highlight modules
454 harboring species indicative of different environmental gradients and, therefore,
455 showing different life strategies. These results could be considered for future studies
456 exploring potential indicator species among plankton.

457 Unassigned sequences

458 Unassigned ASVs, which could not be classified at genus or species level, constituted
459 42.7 % and 58.8 % of the total diatom V4 and V9 ASVs, and 24 % and 20.7

460 % of total diatom V4 and V9 reads (**Figure S13 a**). These include ASVs identified
461 only at the phylum level (e.g., 'unknown' diatoms), which consist of 438 V4 and
462 1056 V9 ASVs. Additionally, ASVs classified only at the class level included 1,694
463 V4 and 1,270 V9 ASVs. Notably, the number of unclassified ASVs from pennate
464 diatoms exceeded those assigned to specific pennate genera/species (904 vs 666
465 ASVs for V4, and 634 vs 213 V9 ASVs). While planktonic pennate genera
466 such as *Thalassionema, Fragilariopsis*, and *Pseudo-nitzschia* were abundant,
467 pennate diatoms are overall more diverse and abundant in benthic environments[60].
468 The high number of unclassified pennate diatom ASVs in our dataset is consistent with
469 the limited number of sediment studies, which are crucial to contribute to sequence
470 diversity in reference datasets[56].
471 The proportion of unassigned reads was higher in the 3/5-20 µm size fraction and less
472 prevalent in the larger size classes (20-180 µm and 180-2000 µm) (**Figure S13 b**).
473 This is an expected result because diatoms belonging to the largest size fractions are
474 those more intensively studied and are hence better represented in genetic reference
475 databases. In addition, there were variations across oceanic regions and sampling
476 sites, with SPO, SAO, AO, and in particular SO, showing higher relative abundances
477 of unassigned ASVs (**Figure S13 c**), probably because the southern hemisphere
478 and the poles are the least explored areas. While the V4 and V9 datasets showed
479 similar trends for unassigned sequences overall, a few stations (TARA_085 and
480 TARA_086 in the SO, TARA_188 in the AO, and TARA_006 in the MS) had a high
481 abundance of unassigned V9 ASVs compared to V4 ASVs, while the opposite was
482 found in TARA_158 in AO (**Figure S13 d**). Some of these unknown ASVs are
483 particularly abundant and should be prioritized in the future for further study.

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

Beyond taxonomic insights from 18S V4 and V9 sequences, complementary functional information can be derived from metatranscriptomics, which quantifies mRNA levels. Therefore, we mined polyA RNA sequencing data from the same plankton samples used for metabarcoding. We focused on the top 100 nuclear-encoded Pfam families with the highest transcript abundances, accounting for 56% of the total metatranscriptomic abundance for diatoms in the global ocean. Thus, this represents the cellular functions and processes in which natural diatom populations invest the bulk of their energy and resources, at least as viewed at the level of mRNA abundance (**Figure 9**).

To reduce redundancy, Pfams with similar functions were merged. Specifically, 137 Pfams coding for different ribosomal subunits were grouped under "Ribosomal proteins", and 47 Pfams coding for various ubiquitin domains were combined into "Ubiquitin domains". As a result of this merging process, these two functional categories rank higher, representing 9.2% and 4.2% of total diatom transcript abundance, respectively. However, at individual level, the most abundant Pfam family was the light-harvesting chlorophyll a/b-binding complex (LHC) proteins (PF00504). LHCs, proposed to be the most abundant membrane proteins on Earth, accounted for almost 3% of diatom metatranscriptomic abundance. Additional photosynthetic families within our top 100 Pfams include ATP synthase subunit C (PF00137), which plays a role in carbon-concentrating mechanisms[66], and flavodoxin (PF00258), contributing 0.24% and 0.20% of diatom metatranscriptomic abundance. Beyond photosynthesis, the maintenance of the frustule also appeared to require a high transcriptional effort, reflected by the observation that the silicon transporters (PF03842)[67] were the 40th most

510 transcriptionally abundant family (0.36% of diatom metatranscriptomic abundance
511), and are mainly found in diatoms and dictyochophytes (another silicifying group).
512 Among the top 100 Pfams, only one domain of unknown function (DUF) was present:
513 DUF285 (PF03382). This enigmatic tandem repeat protein is distantly related to
514 leucine rich repeats[68] (**Figure 10a)** and has multiple gene copies in some bacteria,
515 giant viruses and in eukaryotic phytoplankton (**Figure S14**). Genes and transcripts
516 coding for DUF285 are notably abundant at high latitudes (**Figure 10b-c**), and this
517 trend extends to phytoplankton beyond diatoms (**Figure S15**). Interestingly,
518 DUF285 exhibits a solenoid conformation (**Figure 10a**) similar to proteins with
519 antifreeze activity[69].
520 Other Pfam families displayed specific transcript abundance patterns based on ocean
521 regions, such 'Cold-shock' DNA-binding domains (PF00313) being predominant in AO
522 and SO, and the opposite trends for Heat Shock Factor-type DNA binding domains
523 (PF00447) and Heat Shock Proteins 20, 70 and 90 (HSP20, HSP70 and HSP90;
524 PF00011, PF00012 and PF00183, respectively) **(Figures 9 and and S15)**. Other
525 Pfams more abundant in temperate waters include many of the diatom prevalent
526 families such as dioxygenase (PF00775), retinal pigment epitherial membrane protein
527 (PF03055), subtilase family (PF00082), YHYH protein (PF14240), Zinc finger C2H2
528 type (PF00096), and zinc-binding domain (PF12907).
529 Given the dominance of LHCs, we conducted a detailed analysis of this family. In
530 addition to the canonical members involved in light-harvesting (LHCf, LHCq (also
531 known as LHCy), and LHCr), this superfamily includes the diatom-specific LHCx and
532 LHCz subfamilies, engaged in photoprotection[70] **(Figure 10d)**. After annotating over
533 52 ,000 diatom LHC sequences into these subfamilies, distinct expression patterns
534 emerged. The ratio of transcripts for light-harvesting subfamilies to those for

535 photoprotection subfamilies was higher in larger size fractions (**Figure 10e-g)**. This
536 increase likely serves as a compensatory response to counteract intracellular light
537 attenuation: as cell volume increases, the path length of light within the cell also
538 increases, reducing the light intensity per unit volume. We also observed spatial
539 distinctions in expression related to depth and latitude. Among the pool of LHC
540 transcripts, photoprotective subfamilies were prevalent at the surface, whereas light-
541 harvesting subfamilies were predominant in the DCM (**Figure 10f)**. In addition,
542 photoprotection subfamilies were slightly more prevalent at higher latitudes, which are
543 areas characterized by photo-physiological stresses such as low temperatures and
544 elongated photoperiods (long days in the summer, and long nights in the winter)
545 (**Figure 10e**).
546 To infer adaptation patterns (e.g., gene copy number variations) in different diatom
547 species, we explored the metagenome-assembled genomes (MAGs) recently
548 reconstructed from the same DNA samples used for amplifying the 18S rRNA gene
549 amplicons[71]. Due to assembly and binning challenges, these MAGs lack the 18S
550 rRNA gene, so their taxonomy was inferred using protein-coding marker genes[71]. A
551 total of 52 MAGs were assigned to diatoms, with variable ranges of completeness (3-
552 87% BUSCO score[72]). Focusing on the 18 MAGs with over 70% BUSCO
553 completeness, we observed a slight tendency towards higher copy numbers of
554 genes encoding LHCx, LHCz and DUF285 in MAGs common in high latitudes
555 **(Figure 10h)**. While additional data—more MAGs with higher completeness—are
556 needed to confirm these trends , the observed biogeography of the MAGs seem
557 to align with these genetic characteristics, further underlining the intricate
558 relationship between genetic adaptation and environmental factors in diatom
559 populations.

Conclusions

This study utilizes the *Tara* Oceans datasets to provide an extensive analysis of diatom diversity and distributions across the world's ocean, enhanced by functional profiling. We employed parallel metabarcoding analyses of the V9 and V4 hypervariable regions, offering a wide range of taxonomic resolution ideal for genus-level diversity research. Despite differences in taxonomic resolution and reference databases, our results revealed significant similarities between V4 and V9 molecular patterns, with 15 genera common among the top 20 in both datasets. Additionally, these molecular methods were corroborated with light microscopy data. Our findings reinforce the role of diatoms as a key photosynthetic group in marine ecosystems, highlighting prevalent genera such as *Chaetoceros*, *Thalassiosira*, *Actinocyclus*, and *Pseudo-nitzschia*. Correlation analyses indicated that nutrient limitations and biotic interactions significantly shape diatom relative abundance, particularly through competition with radiolarians for silicon and niche differentiation with picocyanobacteria over nitrogen forms. We defined 25 distinct communities of co-occurring diatom lineages with varying environmental preferences, illustrating diverse ecological adaptations. The AO stands out as a hotspot of diatom abundance and diversity with 6 of the diatom communities exclusive to it. Our results provide insights into diatom life strategies from a wide range of ecosystems and illustrate the ecological convergence of diatom communities from distinct oceanic regions regarding the availability of major nutrients. We envision that these results will help to identify indicator species for ocean monitoring, especially since many diatom lineages were predicted to be vulnerable to climate change scenarios[73].

In addition to taxonomic insights from metabarcoding, we obtained functional information from metatranscriptomics by analyzing changes in mRNA abundance in

585 environmental samples. Our data indicate distinct gene expression patterns
586 correlating with biogeographical distribution. For instance, diatoms in higher latitudes
587 exhibited increased transcript levels coding for 'Cold-shock' DNA-binding proteins as
588 well as for the enigmatic DUF285, suggesting a link between metabarcoding patterns
589 and gene expression profiles. Globally, the most expressed gene family was LHC, with
590 varying expression patterns among subfamilies related to light harvesting versus
591 photoprotection, reflecting changes in depth, latitude and cell size. To understand
592 adaptation strategies, such as gene copy number variations among diatom species,
593 we analyzed MAGs derived from the same DNA samples used for 18S rRNA gene
594 amplification. The biogeographical distribution of diatom MAGs showed some
595 alignement with their gene copy numbers related to photoacclimation, but more
596 MAGs with higher completeness are needed to confirm these trends.
597 In conclusion, our study offers a thorough pole-to-pole examination of marine diatom
598 populations, encompassing the entire plankton size spectrum. This research can
599 reveal important insights into the impact of global changes and human-induced
600 disturbances on marine ecosystems.

601 Material and Methods

602 *Tara* Oceans sampling

603 Tara Oceans expeditions between 2009 and 2013 performed a worldwide sampling of
604 plankton in the upper layers of the ocean, including surface (SRF; 5 m depth) and
605 deep chlorophyll maximum (DCM; 17–188 m). To capture the whole size spectrum of
606 plankton, a combination of filter membranes with different pore sizes (size-
607 fractionation) was used[36]. Four major eukaryotic-enriched size fractions were
608 collected: pico-nanoplankton (0.8 to 5 µm or 0.8 to 2000 µm; named here 0.8-5/2000

609 µm), nanoplankton (5 to 20 µm or 3 to 20 µm; named here 3/5-20 µm), microplankton
610 (20 to 180 µm), and mesoplankton (180 to 2000 µm)[3,36]. Given the inverse relationship
611 between plankton size and abundance[36], the seawater volumes filtered was increased
612 for larger size fractions (from 10^2 to 10^5 L; see Table 1 and Fig. 5 in Pesant et al.[36]).
613 These plankton samples were leveraged to generate different molecular and optical
614 datasets analysed in the current work (**Figure S2**). Specific details about them and
615 their analyses are described below.

616 Metabarcoding datasets

617 Nucleic acid extraction methods were applied to obtain DNA from the different size-
618 fractionated samples. In order to target eukaryotic plankton diversity and relative
619 abundance, a metabarcoding approach was performed using the 18S rRNA gene as
620 molecular marker. Detailed information about DNA extraction, PCR amplification and
621 Illumina sequencing of metabarcodes are described by[74,75]. Briefly, DNA samples
622 were amplified by PCR targeting the hypervariable region V9 (130 ± 4 base pairs
623 length; primer pair: 1389F 5'-TTGTACACACCGCCC-3' and 1510R 5'-
624 CCTTCYGCAGGTTCACCTAC-3')[76] or V4 (385 ± 4 base pairs length; primer pair
625 TAReuk454FWD1 5'-CCAGCASCYGCGGTAATTCC-3' and TAReukREV3 5'-
626 ACTTTCGTTCTTGATYRA-3')[77] of the 18S rRNA marker gene followed by the
627 Illumina sequencing of the amplicons[74]. The details of PCR mixes, thermocycling
628 and sequencing conditions are provided in Alberti et al.[75]
629 Resulting paired-end reads were mixed-oriented meaning that both R1 and R2 files
630 are composed by a mix of forward and reverse reads. Paired-end reads were trimmed
631 to remove PCR primer sequences using Cutadapt v2.7[78] and dispatched into four files,
632 2 files for the classical orientation (forward reads in R1 and reverse reads in R2) and
633 2 others for the other orientation (reverse reads in R1 and forward reads in R2). Paired-

634 end reads without both primers were filtered out using the option --discard-untrimmed.
635 Forward and reverse reads were trimmed at position 80 for V9 and at position 215 for
636 V4 and reads with ambiguous nucleotides or with a maximum number of expected
637 errors (maxEE) superior to 2 were filtered out using the function *filterAndTrim*() from
638 the R package *dada2* [37]. For each run and read orientation, error rates were defined
639 using the function *learnErrors*() and denoised using the *dada*() function with pool =
640 TRUE before being merged using *mergePairs*() with default parameters. Mixed
641 orientated reads from the same sample and sequencing replicates were summed
642 together. Remaining chimeras were removed using the function
643 *removeBimeraDenovo*(). Scripts producing the ASV tables are publicly available here:
644 https://gitlab.univ-nantes.fr/combi-ls2n/taradada. ASVs were taxonomically assigned
645 using IDTAXA (50% confidence threshold, no sequence similarity threshold) [38] with
646 the PR2 database version 4.14[39]. The scripts for the taxonomic assignment are
647 publicly available here: https://gitlab.sb-roscoff.fr/nhenry/abims-metabarcoding-
648 taxonomic-assignment/-/tree/v1.0.1. To facilitate a comparison between diatoms and
649 other photosynthetic groups, we annotated the ASVs as phytoplankton—including
650 photosynthetic dinoflagellates and chrysophytes—whenever their taxonomic
651 resolution was adequate for alignment with known phytoplanktonic lineages. This
652 classification was informed by a literature review and expert insights compiled in the
653 trait reference database accessible at
654 https://zenodo.org/doi/10.5281/zenodo.3768950. To reduce the influence of PCR and
655 sequencing errors not detected by the dada2 algorithm, only ASVs seen in at least
656 two different samples with at least three copies were retained.

657 ## Metatranscriptomic and metagenomic datasets

658 We mined the *Tara* Oceans catalog of eukaryotic expressed genes (Marine Atlas of
659 *Tara* Oceans Unigenes; MATOU [79,80]). Version 1.5 of MATOU, released alongside
660 this publication (https://www.genoscope.cns.fr/tara/#MATOU-1.5), includes as an
661 important update the incorporation of samples from the Arctic Ocean.

662 Briefly, the catalog was built in the following steps. Nucleic acid extraction methods
663 were applied to obtain eukaryotic polyA RNA from the different size-fractionated
664 samples and subsequent Illumina sequencing was carried out[75,79]. The corresponding
665 metatranscriptomic reads were assembled and clustered at 95% identity[79,80]. The
666 metatranscriptomic reads were mapped onto the sequence catalogue to provide the
667 abundance of each transcript sequence in every sample[79]. Similarly, metagenomic
668 reads from the same seawater samples were mapped onto the sequence catalog to
669 ascertain the abundance of each gene in every sample[79]. The taxonomic classification
670 of the unigene catalog was carried out by sequence similarity against a custom
671 reference database comprising METdb[92], UniRef90 (version 20190810), and single-
672 cell amplified genomes (SAGs) from Tara Oceans[93]. Functional assignation was
673 performed using the Pfam database[79].

674 For our analysis, we focused on sequences assigned to diatoms and other eukaryotic
675 phytoplankton groups for comparative purposes. It is important to note that we could
676 not exclude heterotrophic dinoflagellate species due to the limited number of reference
677 sequences available for this group. We then retrieved the metatranscriptomic read
678 abundances (rpkm values) of the selected sequences and normalized them to the total
679 read abundance for transcripts of the corresponding phytoplankton taxon in each
680 sample. Additionally, for specific cases of interest such as LHC, DUF285, and HSP90,

681 we also retrieved the metagenomic read abundances and performed the equivalent
682 abundance normalization.

683 Metagenome-assembled genomes (MAGs) dataset

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

694 Optical microscopy datasets

695 For light microscopy, three ml of each sample (from 20-180 µm size fractions) were
696 placed in an Utermöhl chamber . Cells falling in 2 or 4 transects of the chamber
697 were identified and enumerated using an inverted light microscope (Carl Zeiss
698 Axiophot200) at 400x magnification. To be compared with the molecular data, the
699 optical microscopy counts of diatoms were expressed as percentages of the genera
700 (or of each genus) over the total diatom community (%).
701

WGCNA of diatom metabarcoding data

We used the R package WGCNA (Weighted Gene Co-Expression Network Analysis) [64] to infer the ASVs co-occurence networks for both V4 and V9 metabarcoding data. To remove redundancy, for every station we pooled the reads for the four size fractions into a single aggregated sample. We discarded the samples without all the four size fractions. A Pearson correlation matrix was computed using the ASVs abundances. We raised the Pearson correlation to a power of 12 and 20, for V4 and V9 respectively, to compute the adjacency matrices and we used these matrices to generate a topological overlap measure (TOM). A hierarchical clustering was finally performed on the TOM matrices to identify 25 modules and a set of unclustered ASVs (grey module) (**Figure 8**). The module eigengenes were also computed for each module identified (grey module included) and then used to generate correlations (and associated *p*-values) between each module and 48 environmental variables retrieved from https://doi.org/10.1594/PANGAEA.875582 (**Figure S16**) (see also below).

Analysis of LHC and DUF285 in metagenomes and metatranscriptomes

We analyzed the gene and transcript abundances of LHCs and DUF285 according to environmental variables by mining the metatranscriptomes and metagenomes generated by *Tara* Oceans. In particular, we searched the MATOU-v1.5 catalog (https://www.genoscope.cns.fr/tara/#MATOU-1.5) for sequences coding for LHCs (PF00504) and DUF285 (PF03382) by running HMMer (version 3.2.1 with the gathering threshold option; http://hmmer.org/) over the translated sequences. We retained only sequences assigned to diatoms, and retrieved the corresponding metagenomic and metatranscriptomic read abundances (rpkm values) and normalized

726 them to the total read abundance for diatom transcripts in each metatranscriptomic
727 sample. We did a similar HMMer search among the gene catalog from diatom MAGs.
728 For the functional classification of LHCs into the main subfamilies (LHCf, LHCq, LHCr,
729 LHCx and LHCz), we performed a phylogenetic placement of the translated
730 sequences on the reference phylogeny described in Kumazawa et al[70]. Protein
731 sequences were aligned with mafft version 6 using the G-INS-I strategy[81]. The output
732 alignment was trimmed in both N- and C-terminal regions to maintain the reference
733 alignment limits. The alignment was also processed using trimAl v1.4[95] with the gap
734 threshold option *-gt* 0.1 to keep columns where at most 10% of sequences contain a
735 gap. The phylogenetic inference was made using approximate maximum likelihood
736 with FastTree[82]. The sequences were classified according to their grouping in
737 monophyletic branches with statistical support of >0.7 with reference sequences of the
738 same functional group.

739 Plotting and statistical analysis

740 Analyses were carried out in R language (http://www.r-project.org/). Exponentiated
741 Shannon Diversity Index (expH), a diversity index that accounts for both species
742 richness and evenness, was calculated using R library *vegan* [83]. Estimates of total
743 ASV richness were done with the *specpool()* function from the *vegan* library. Graphical
744 analyses were carried out using R library *ggplot2*[84] and treemaps were generated with
745 *treemap* [85]. Maps were generated with *borders()* function in *ggplot2*[84] and
746 *geom_point()* function for bubbles or *scatterpie* package [86] for pie charts. Loess
747 smooth plots with 95% confidence windows were plotted with *geom_smooth*() function
748 in *ggplot2*. Spearman's rho correlation coefficients and *p*-values were calculated using
749 the *cor.test()* function of the *stats* package. Metric multidimensional scaling (NMDS)
750 analysis to visualize Bray–Curtis distances was carried out with the *metaMDS()*

751 command in the R package *vegan*[87], and the influence of environmental variables on
752 sample ordination was evaluated with the function *envfit*() in the same R package,
753 whereas the pie charts were plotted with *scatterpie* package. Partial least square
754 analyses were implemented with the R package *plsdepot* [88] to explore the
755 correlations between diatom relative abundance and Shannon index with the
756 physicochemical context (range-transformed median values of nutrient
757 concentrations, chlorophyl a, absolute latitude, and temperature). A second set of
758 partial least square analyses were performed to explore the correlation of diatom
759 relative abundance to the relative abundance of other silicifiers, and the abundances
760 of picocyanobacteria, copepods and total Rhizaria (see below).

761 Physico-chemical and biotic parameters

762 Physicochemical parameters were retrieved from
763 https://doi.org/10.1594/PANGAEA.875582 and a full description of the performed
764 measurements is described in Pesant et al[36]. Measurements of temperature,
765 conductivity, salinity, depth, pressure, and oxygen were carried out with a vertical
766 profile sampling system (CTD-rosette) and Niskin bottles. Chlorophyll *a*
767 concentrations, a proxy for total phytoplankton biomass, were measured using high-
768 performance liquid chromatography. Dissolved nutrients (NO_3^-, PO_4^{3-}, Si) were
769 analyzed according to previous methods[89,90]. We complemented the *in situ*
770 measurements with silica, nitrate, nitrite, ammonium, and iron levels derived from
771 the ECCO2-DARWIN ocean model [94].
772 For biological predictors, we included the abundances of picocyanobacteria
773 (*Synechococcus* and *Prochlorococcus*), copepods (grazers), and silicifiers. Copepod
774 counts were collected with the WP2 net (mesh size of 200 µm) towed vertically from
775 500 m to the surface (ecotaxa.obs-vlfr.fr/prj/377, ecotaxa.obs-vlfr.fr/prj/378).

776 Picocyanobacteria counts were determined by flow cytometry[12]
777 (data.mendeley.com/datasets/p9r9wttjkm/2). The relative abundances of silicifiers
778 were based on the percentage of reads among V4 and V9 barcode datasets as
779 described previously.

780 Acknowledgements

781 We would like to thank all colleagues from the *Tara* Oceans consortium as well as the
782 Tara Ocean Foundation for their inspirational vision. We are grateful to Shruti Malviya
783 and Federico Ibarbalz for providing initial details about the metabarcoding data. This
784 work was granted access to the HPC resources of IDRIS, Jean Zay supercomputer,
785 under the allocation 2020–DARI A0080111497 made by GENCI, and has been
786 supported by the FFEM - French Facility for Global Environment, French Government
787 'Investissements d'Avenir' programs OCEANOMICS (ANR-11-BTBR-0008), FRANCE
788 GENOMIQUE (ANR-10-INBS-09-08), MEMO LIFE (ANR-10-LABX-54), and PSL
789 Research University (ANR-11-IDEX-0001-02). C.B. acknowledges funding from the
790 European Research Council (ERC) under the European Union's Horizon 2020
791 research and innovation program (Diatomic; grant agreement No. 835067), the
792 Horizon Europe projects Marco-Bolo (grant agreement no. 101082021) and
793 BlueRemediomics (grant agreement no. 101082304), Agence Nationale de la
794 Recherche "Phytomet" (ANR-16-CE01-0008) and BrownCut (ANR-19-CE20-0020),
795 and the BNP Paribas Foundation Climate & Biodiversity Initiative. J.J.P.K.
796 acknowledges postdoctoral funding from the Fonds Français pour l'Environnement
797 Mondial, and by Moore–Simons Project on the Origin of the Eukaryotic Cell, Simons
798 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

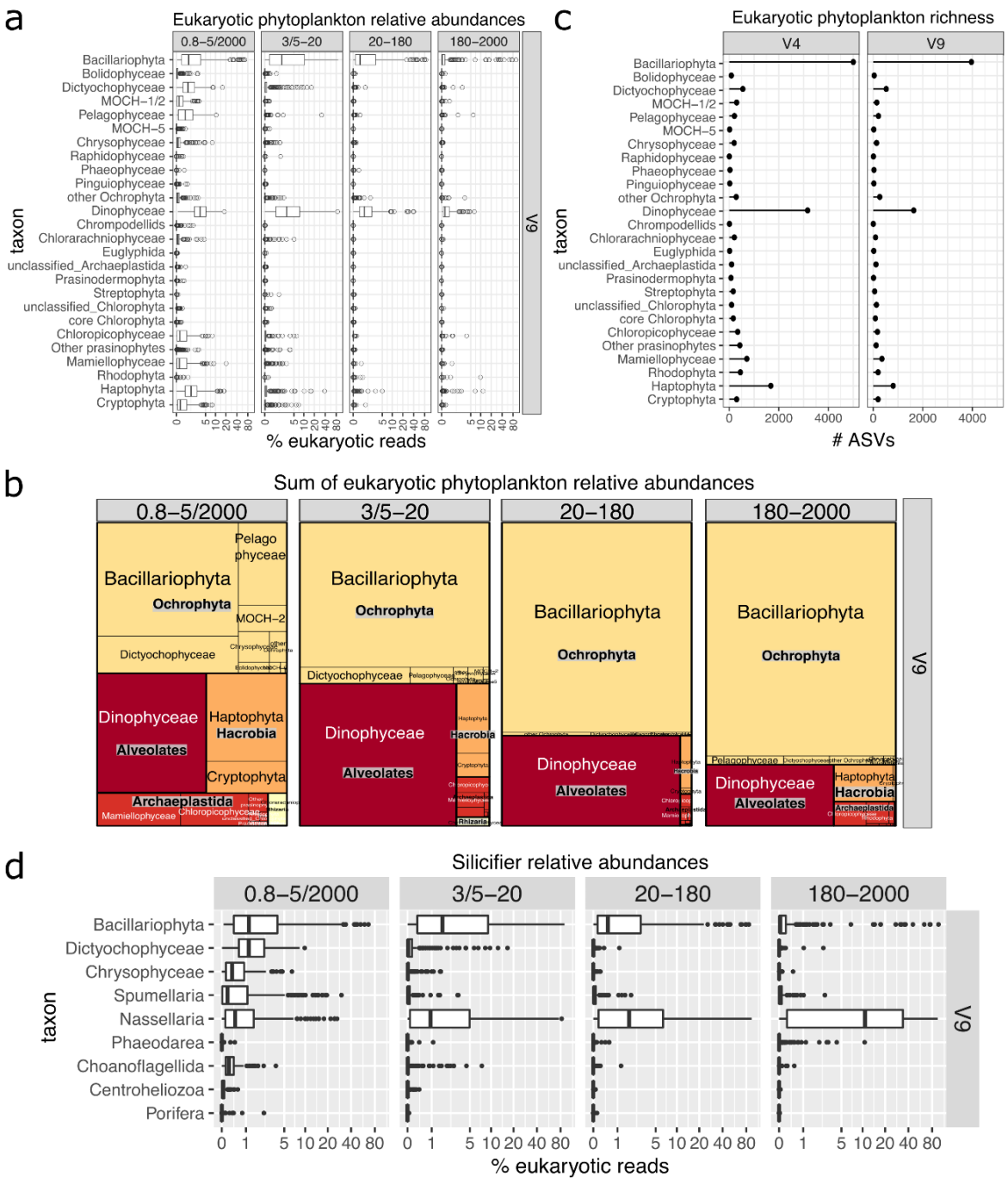
The authors declare that all the data supporting the findings of this study are publicly available in the following repositories and in the supplementary information files of this paper. The contextual data are available in Pangaea (www.pangaea.de) with the identifier https://doi.org/10.1594/PANGAEA.875582, and a simplified version is available in https://zenodo.org/records/7229815. rDNA 18S metabarcoding data are deposited at the European Nucleotide Archive under accession numbers PRJEB6610 and PRJEB9737, and metagenomic and metatranscriptomic data under PRJEB402, PRJEB9691, PRJEB9738 and PRJEB9739. The ASV table and the taxonomic annotation for the V4 marker are available in https://zenodo.org/records/13881376 and for the V9 marker in https://zenodo.org/records/13881418. The files for the

Marine Atlas of *Tara* Oceans Unigenes version 1.5 (MATOU-v1.5) are available in https://www.genoscope.cns.fr/tara/#MATOU-1.5, including the FASTA sequences, taxonomic and functional annotation tables, as well as metagenomic and metatranscriptomic abundance tables. 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.

Code Availability

Scripts producing the ASV tables are publicly available here: https://gitlab.univ-nantes.fr/combi-ls2n/taradada. The scripts for the ASV taxonomic assignment are publicly available here: https://gitlab.sb-roscoff.fr/nhenry/abims-metabarcoding-taxonomic-assignment/-/tree/v1.0.1. Scripts for producing the figures in this paper are in https://github.com/JJPierellaKarlusich/Diatom_patters/tree/main.

Main Figures


a
Eukaryotic phytoplankton relative abundances
0.8–5/2000
3/5–20
20–180
180–2000
V9
taxon
Bacillariophyta
Bolidophyceae
Dictyochophyceae
MOCH–1/2
Pelagophyceae
MOCH–5
Chrysophyceae
Raphidophyceae
Phaeophyceae
Pinguiophyceae
other Ochrophyta
Dinophyceae
Chrompodellids
Chlorarachniophyceae
Euglyphida
unclassified_Archaeplastida
Prasinodermophyta
Streptophyta
unclassified_Chlorophyta
core Chlorophyta
Chloropicophyceae
Other prasinophytes
Mamiellophyceae
Rhodophyta
Haptophyta
Cryptophyta
% eukaryotic reads
c
Eukaryotic phytoplankton richness
V4
V9
ASVs
b
Sum of eukaryotic phytoplankton relative abundances
Pelago phyceae
Bacillariophyta
Ochrophyta
MOCH–2
Dictyochophyceae
Dinophyceae
Alveolates
Haptophyta
Hacrobia
Cryptophyta
Archaeplastida
Mamiellophyceae
Chloropicophyceae
Rhizaria
d
Silicifier relative abundances
Bacillariophyta
Dictyochophyceae
Chrysophyceae
Spumellaria
Nassellaria
Phaeodarea
Choanoflagellida
Centroheliozoa
Porifera
% eukaryotic reads
e
Sum of silicifier relative abundances
Chryso phyceae
Dictyochophyceae
Spumellaria
Rhizaria
Nassellaria
Opisthokonta
Haptista

837

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.

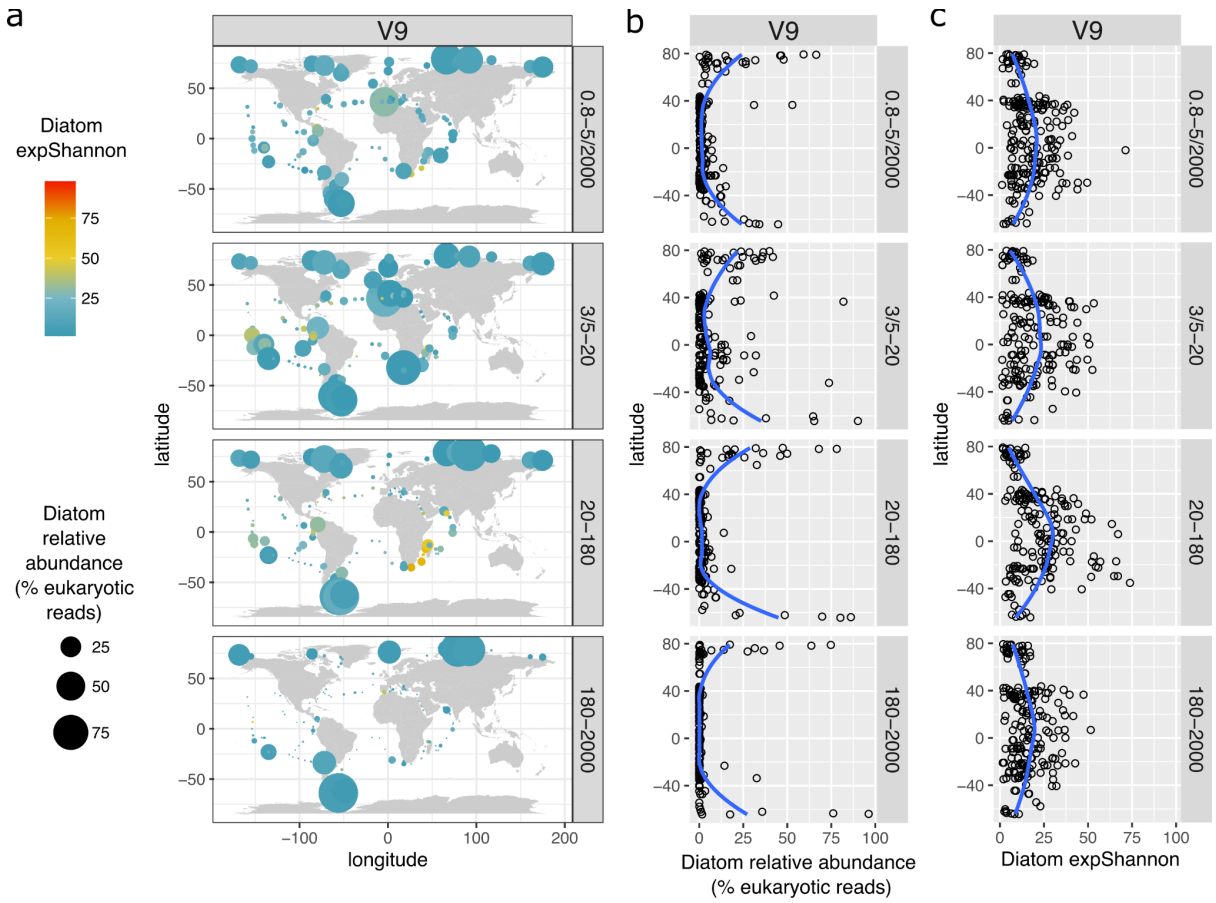


Figure 2: Latitudinal gradients of diatom relative abundance and diversity in surface waters using V9 metabarcoding data obtained from different size-fractions. a) Distribution maps. b) Latitudinal gradient for relative abundance. c) Latitudinal gradient for the exponentiated Shannon Diversity Index.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

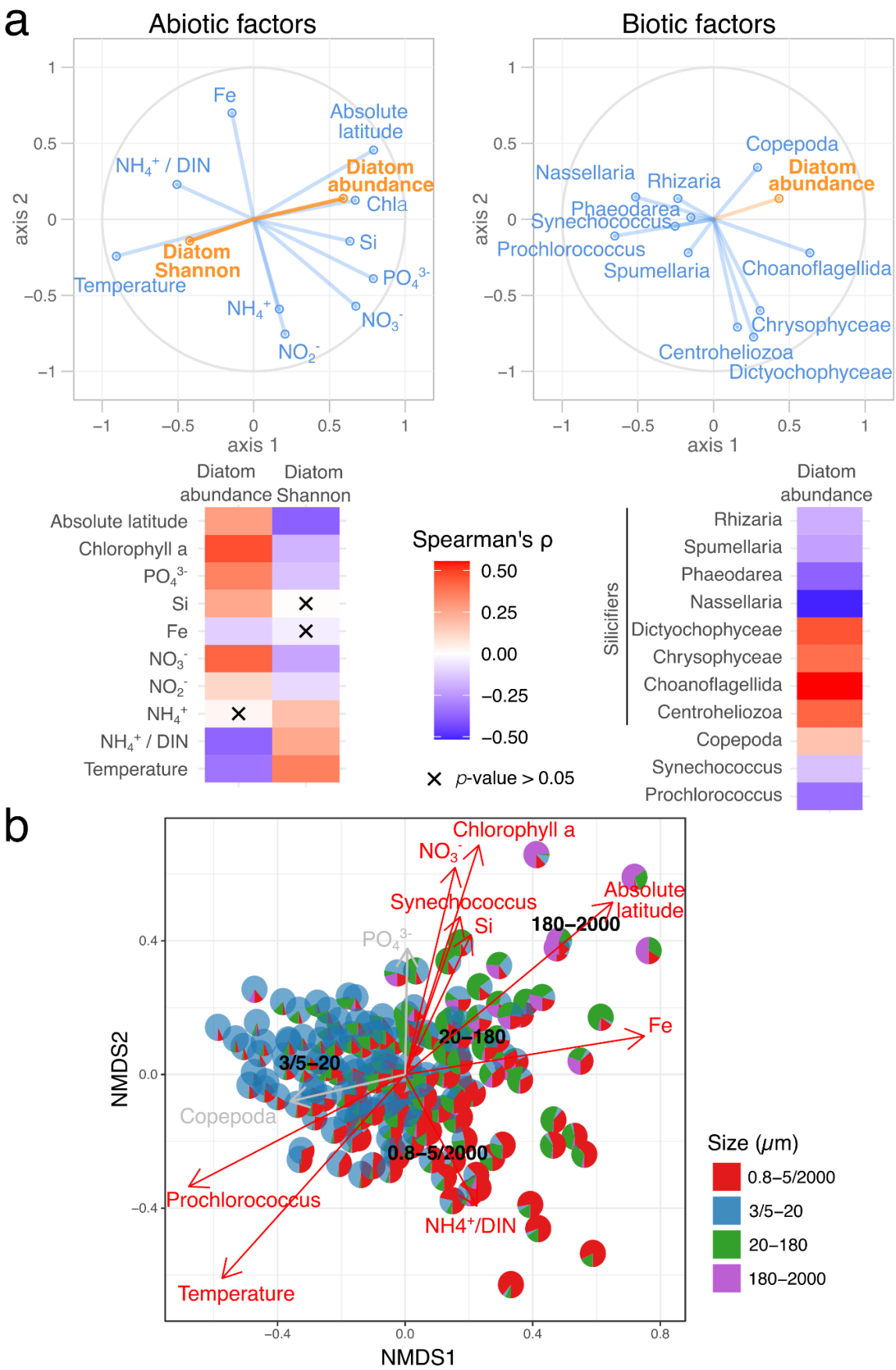


Figure 3: Diatom abundance and diversity in relation to environmental parameters. a) Correlation circle based on the partial least square analysis of V4 and V9 diatom relative abundance and Shannon diversity as responses and a diverse set of abiotic (left) and biotic

(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 biotic and abiotic factors . 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.

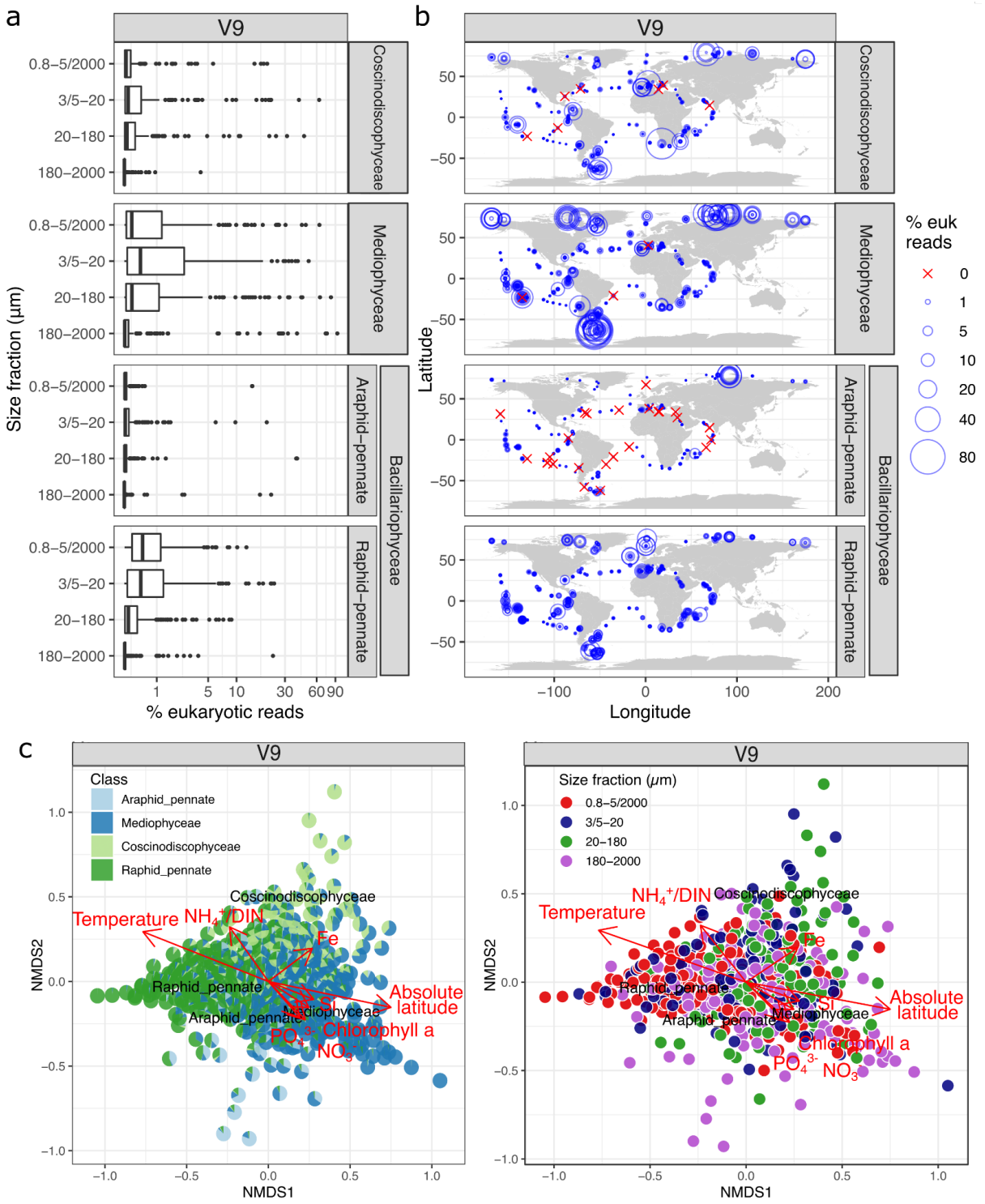


Figure 4: Abundance patterns of diatom classes. a) Relative abundances for V9 marker (log scale). Each point is a size-fractionated sample. b) Biogeography. Each circle is a size-fractionated sample. c) NMDS analysis of stations according to Bray–Curtis distance. Fitted statistically significant physico-chemical parameters (adjusted P value < 0.05) are displayed: nitrate, phosphate, silicon, iron, ratio of ammonium to total dissolved inorganic nitrogen (DIN), temperature, absolute latitude, , and chlorophyll a concentration. Each pie chart (left) or each circle (right) is a size-fractionated sample. NMDS stress value: 0.1470158 . Comparison between V4 and V9 patterns are displayed in **Figure S7**. The maps separated by diatom classes are shown in **Figure S9**.

Genera
a. Richness
b. Abundance
c. Size fraction
d. Depth
e. Occupancy
V4
V9
Actinocyclus
Actinoptychus
Amphora
Arcocellulus
Ardissonea
Asterionella
Asterionellopsis
Asteromphalus
Asteroplanus
Attheya
Bacillaria
Bacteriastrum
Bellerochea
Biddulphiopsis
Bleakeleya
Brockmanniella
Caloneis
Campylodiscus
Cerataulina
Chaetoceros
Climacosphenia
Cocconeis
Corethron
Coscinodiscus
Cyclotella
Cylindrotheca
Cymbella
Dactyliosolen
Delphineis
Detonula
Ditylum
Entomoneis
Eucampia
Florella
Fragilaria
Fragilariopsis
Guinardia
Gyrosigma
Haslea
Helicotheca
Hemiaulus
Hyalosira
Hyalosynedra
Lampriscus
Lauderia
Leptocylindrus
Licmophora
Licmosphenia
Lithodesmium
Melosira
Meuniera
Minidiscus
Minutocellus
Navicula
Nitzschia
Paralia
Planktoniella
Pleurosigma
Porosira
Prestauroneis
Proboscia
Psammodictyon
Pseudo–nitzschia
Pseudogomphonema
Pseudohimantidium
Rhizosolenia
Rossia
Seminavis
Skeletonema
Stellarima
Stephanodiscus
Stephanopyxis
Striatella
Surirella
Synedra
Talaroneis
Tenuicylindrus
Thalassionema
Thalassiosira
Thalassiothrix
Toxarium
Triceratium
Trieres/Odontella
ASVs
Sum normalized reads
Abundance (%)
Abundance (%)
stations
Coscinodiscophyceae
Mediophyceae
Raphid pennate
Araphid pennate
0.8-5/2000 μm
3/5–20 μm
20–180 μm
180–2000 μm
Surface
DCM

Figure 5: Summary of diatom genera detected in the V9 and V4 metabarcoding datasets. A total of 75 genera were defined for V4 barcodes and 64 for V9 barcodes. (Column a) Richness expressed as number of ASVs (log scale). (Column b) Abundances expressed as the sum across samples of the relative abundances of rDNA reads (log scale). (Column c) Percentage distribution of rDNA read relative abundances per size fraction. (Column d) Percentage distribution of rDNA read relative abundances per depth. (Column e) Occupancy expressed as the number of stations in which the genus was observed.

Figure 6: Abundance, diversity and distribution of the top 20 most abundant diatom genera in V4 and V9 datasets. These genera accounted for 97 % of assigned reads in the entire V4 dataset and 98 % in the V9 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. Comparison between microscopy and molecular methods for detecting the top 20 most abundant diatoms in the 20-

180 µm size fraction. a. Abundance rank of each genus within each dataset (in black when ranking as top 20, otherwise in grey). b. Sum of the normalized reads (% for the genus among total diatoms) across all samples. Only samples with both microscopy and molecular methods were analysed (67 samples from surface and DCM from 59 stations). Due to similar morphologies, we merged the counts for *Actinocyclus*, *Coscinodiscus* and centric diatoms >30 µm in length. The microscopy-based identification of *Fossulaphycus* (previously *Fossula*, and neither cultivated nor sequenced) was reclassified as *Fragilaria* due to similar morphologies and co occurrence (microscopy detection of *Fossulaphycus* only at one Arctic station, where V4 detected a high abundance of *Fragilaria* reads).

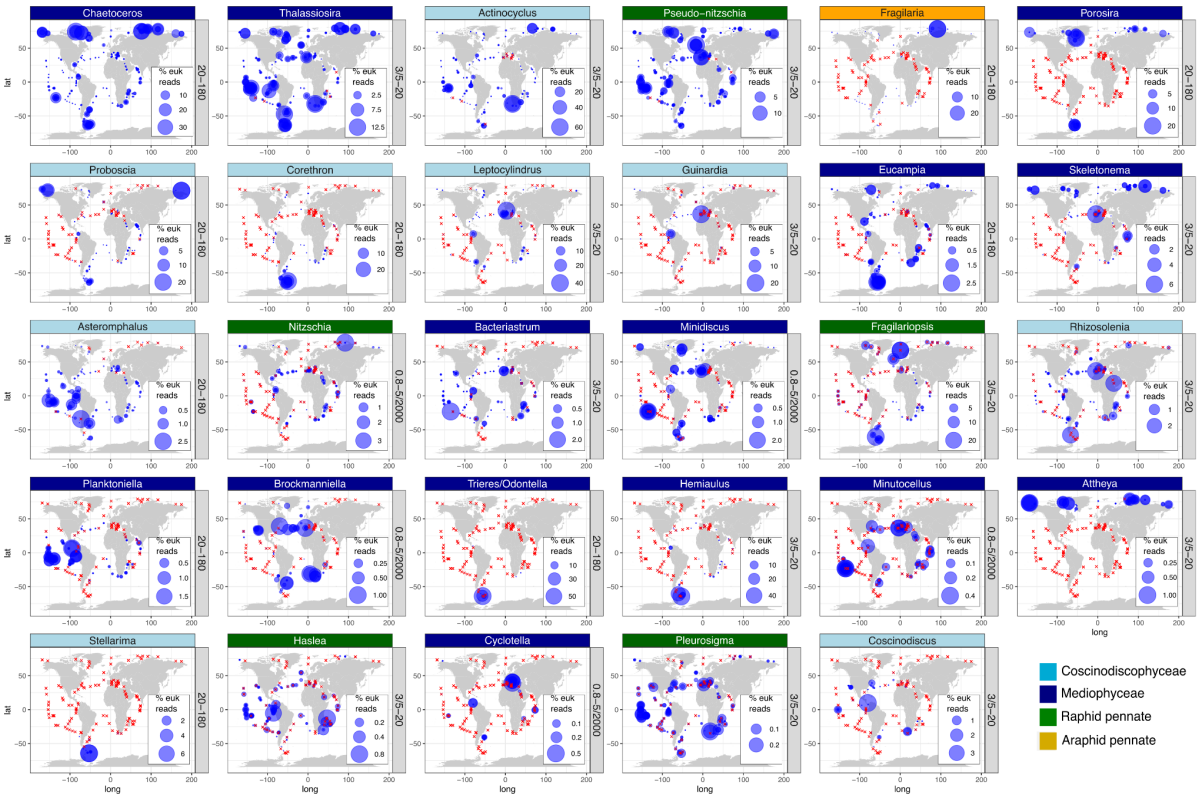


Figure 7: Global distribution of the 20 most abundant genera in the V9 and V4 datasets. All genera in the list of Fig 6a are displayed. Bubble areas are scaled to the % reads for the genus among eukaryotic reads at each station location, whereas crosses indicate absence of detection. The maps correspond to samples from the size fraction where the genus was most prevalent, indicated alongside each map . 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 genera can be found in Supplementary File 1.

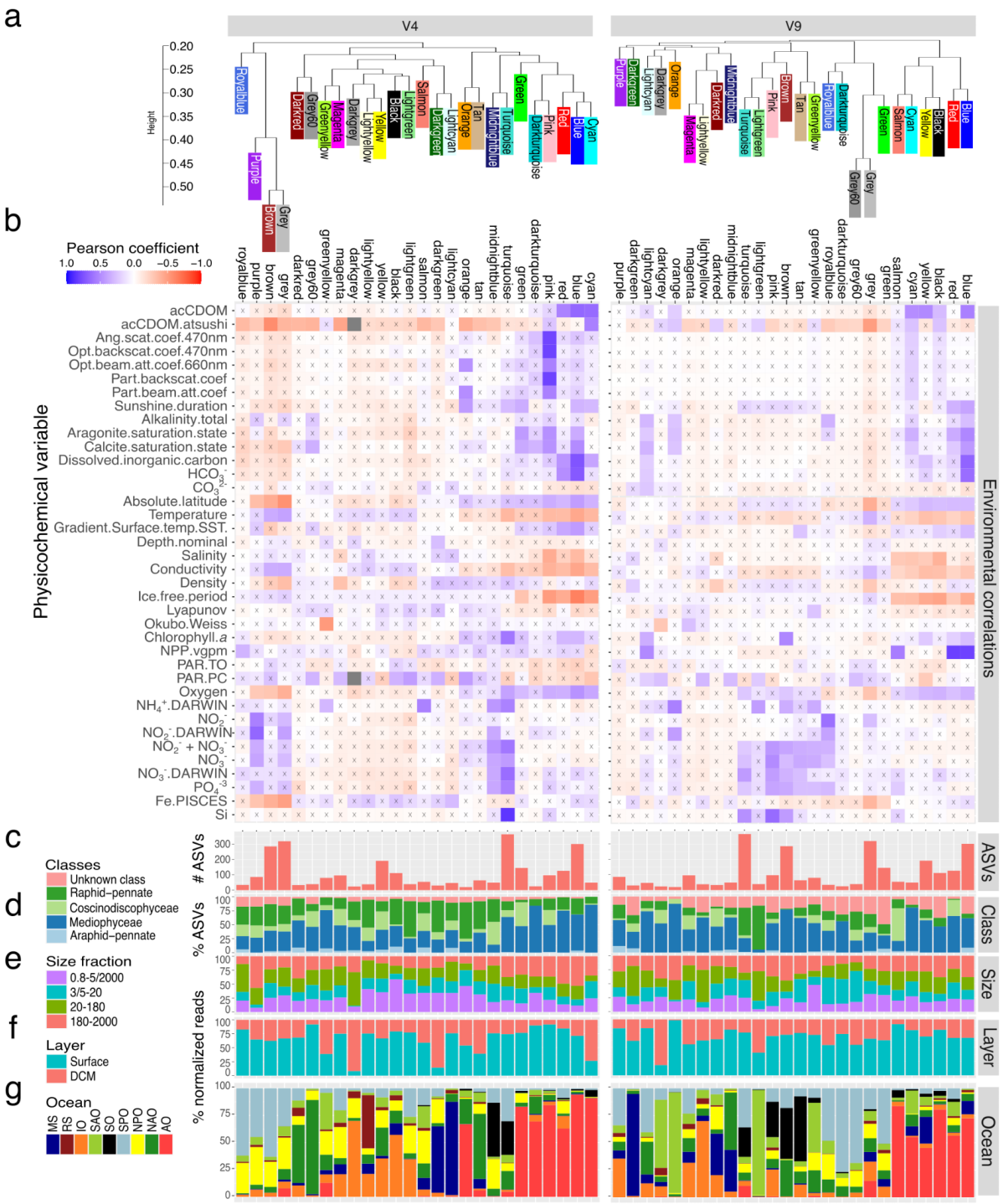


Figure 8. Weighted gene correlation network analysis (WGCNA) of diatom communities for the V4 (left) and V9 datasets (right). *a*. Module clustering according to their co-occurrences. *b*. Pearson correlation between module eigenvalues and physico-chemical parameters. Crosses indicate adjusted p-value>0.05. *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 water layers . *g*. Sum of the percentage of reads across oceans. Abbreviations: acCDOM, absorption coefficient of colored dissolved organic matter; Ang.scat.coef.470nm, Angular scattering coefficient 470 nm; opt.backscat.coef.470nm, Optical backscattering coefficient 470 nm; Opt.beam.att.coef.660nm, Optical beam attenuation coefficient 660 nm; Part.backscat.coef, Backscattering coefficient of particles;

Part.beam.att.coef, Beam attenuation coefficient of particles; Gradient.Surface.Temp.SST, Sea surface temperature gradient; Lyapunov, maximum Lyapunov exponent; Okubo-Weiss, Okubo-Weiss parameter; NPP.vgpm, Net primary production from VGPM product; PAR, photosynthetic active radiation; NH_4^+.DARWIN, modelled ammonium concentration; NO_2^-.DARWIN, modelled nitrite concentration; NO_3^-.DARWIN, modelled nitrate concentration; NH_4^+, NO_2^-+NO_3^-, measured concentration of nitrate plus nitrite; NO_3^-, measured nitrate concentration; Fe.PISCES, Fe concentration from PISCES model.

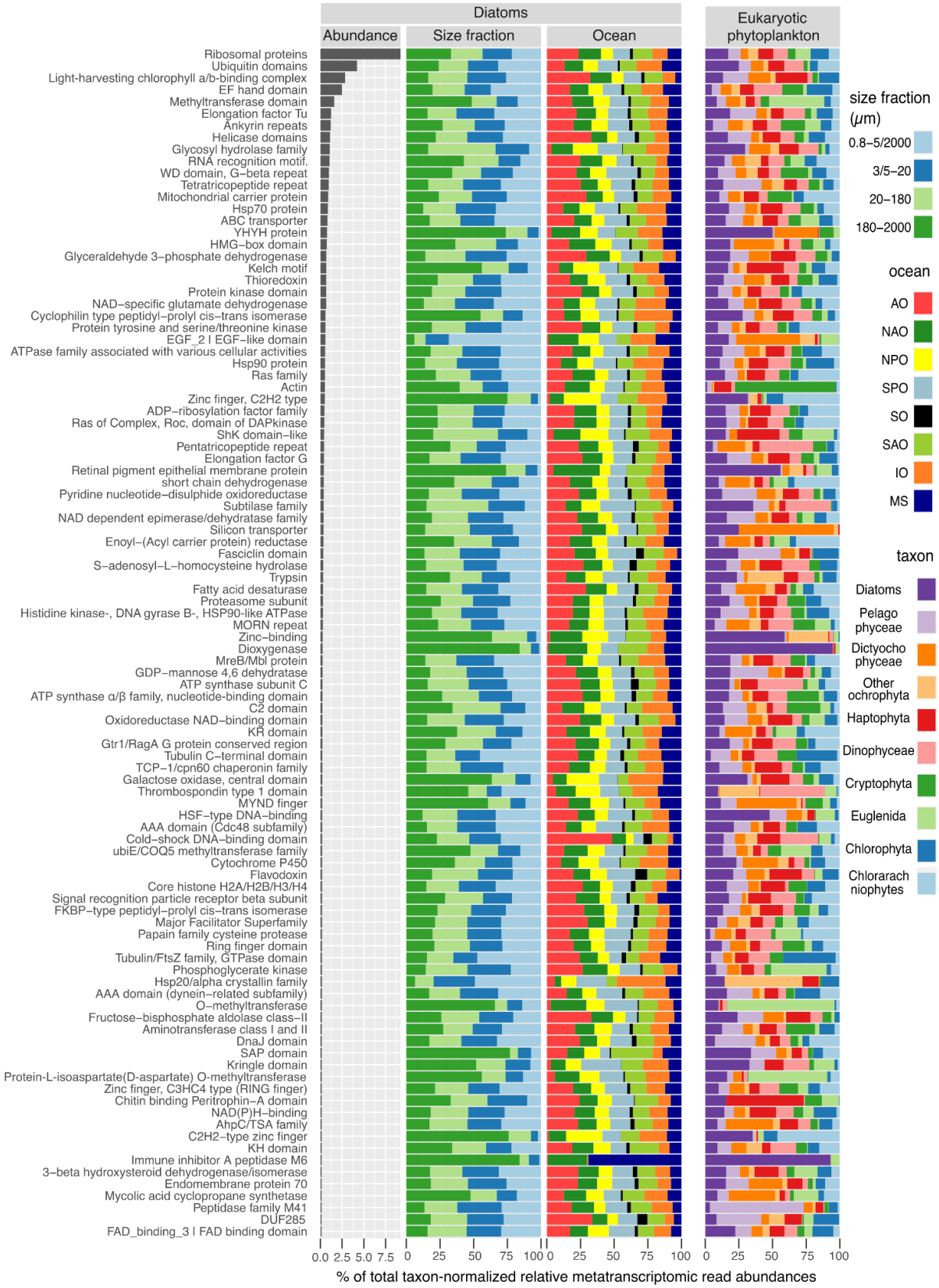
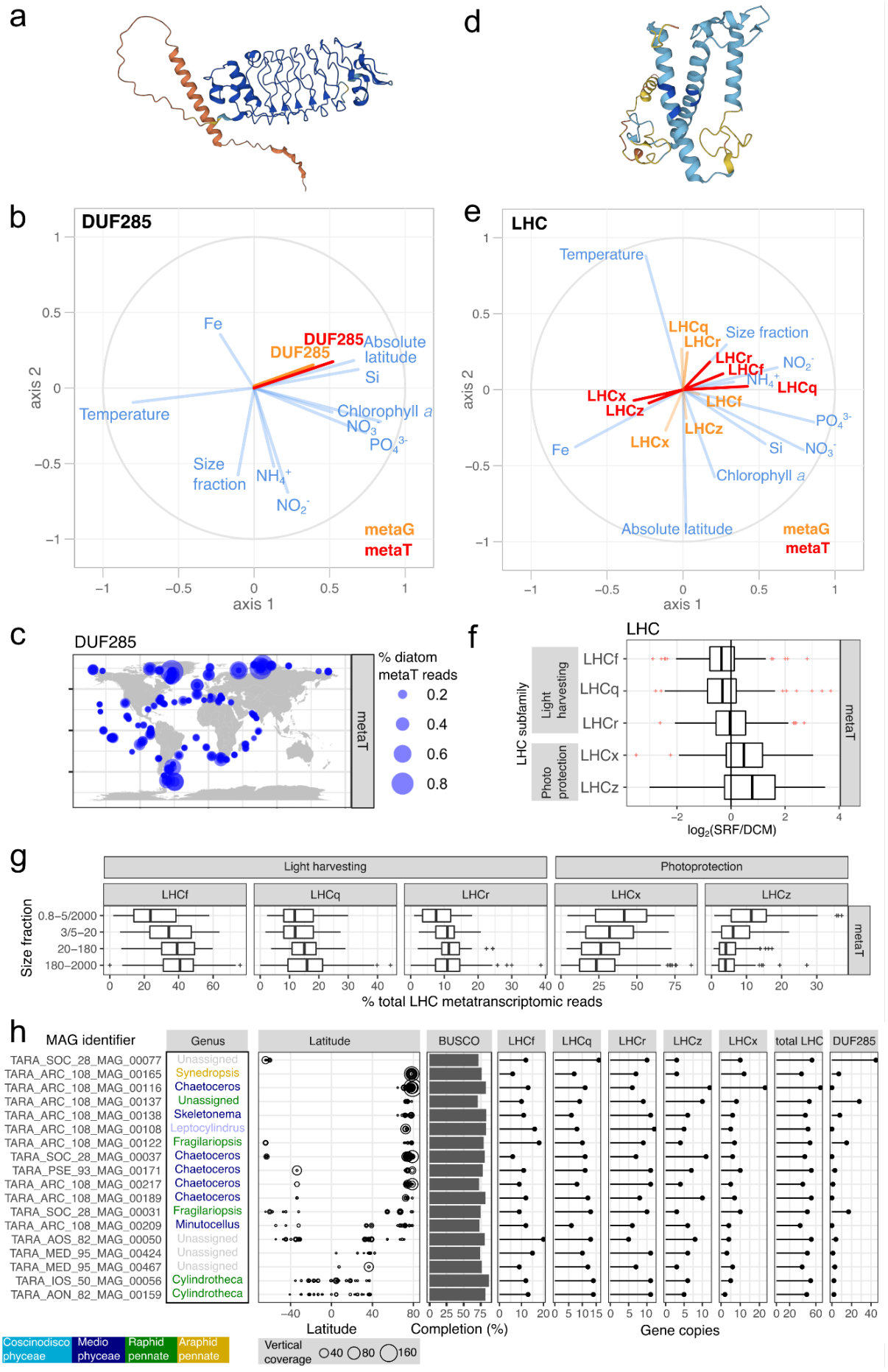


Figure 9: Top 100 most expressed diatom Pfam families in the global ocean. (Column a) Abundance ordered by rank. (Column b) Abundance across size fractions. (Column c) Abundance among ocean regions (AO, Arctic Ocean; NAO, North Atlantic Ocean; NPO, North Pacific Ocean; SPO, South Pacific Ocean; SO, Southern Ocean; SAO,

South Atlantic Ocean; IO, Indian Ocean; MS, Mediterranean Sea). (Column d) Abundance in diatoms vs other eukaryotic phytoplankton. For each sample, the abundance of a given diatom Pfam was calculated as the percentage of total metatranscriptomic reads assigned to diatoms in that sample. The same calculation was applied to other phytoplankton for column d. Note that we were unable to exclude non-photosynthetic species from dinoflagellates due to the limited number of reference gene sequences available for this group. 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”).


a
d
b
DUF285
axis 2
axis 1
Fe
DUF285
DUF285
Absolute latitude
Si
Temperature
Chlorophyll a
NO3-
PO43-
Size fraction
NH4+
NO2-
metaG
metaT
e
LHC
Temperature
LHCq
LHCr
Size fraction
LHCr
LHCf
NO2-
NH4+
LHCq
LHCx
LHCz
LHCf
LHCz
PO43-
Fe
LHCx
Si
NO3-
Chlorophyll a
Absolute latitude
metaG
metaT
c
DUF285
% diatom metaT reads
0.2
0.4
0.6
0.8
metaT
f
LHC
LHC subfamily
Light harvesting
Photo protection
LHCf
LHCq
LHCr
LHCx
LHCz
metaT
log2(SRF/DCM)
g
Light harvesting
Photoprotection
LHCf
LHCq
LHCr
LHCx
LHCz
Size fraction
0.8–5/2000
3/5–20
20–180
180–2000
metaT
% total LHC metatranscriptomic reads
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

966

967 **Figure 10: Gene and transcript patterns for the Pfam families DUF285 and LHC among**
968 **diatom populations.** a-c) DUF285 (PF03382). a) Structural prediction (Uniprot id B7FV04).
969 b) PLS analysis for metagenomic (orange) and metatranscriptomic (red) abundances. c)
970 Metatranscriptomic abundance biogeography . d-g) LHC (PF00504). d) Structural prediction
971 (Uniprot id B7G6Y1). e) PLS analysis for metagenomic (orange) and metatranscriptomic (red)
972 abundances. f) Depth partition between surface (SRF) and deep chlorophyll maximum (DCM).
973 g) Partition across size fractions. h) Gene copy patterns for LHC and DUF285 in metagenomic-
974 assembled genomes (MAGs) assigned to diatoms. The columns correspond to: i) MAG
975 identifier. ii) Class and genus. iii) Latitudinal distribution. Bubbles area is proportional to vertical
976 read coverage. iv) Completeness based on BUSCO[72]. Only MAGs with >70% completeness
977 are displayed. v) Gene copies for LHC subfamilies and for DUF285.
978

979

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).

1029 22. Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E.
1030 Towards next-generation biodiversity assessment using DNA metabarcoding.
1031 *Mol. Ecol.* **21**, 2045–2050 (2012).
1032 23. Canesi, K. L. & Rynearson, T. A. Temporal variation of Skeletonema community
1033 composition from a long-term time series in Narragansett Bay identified using
1034 high-throughput DNA sequencing. *Mar. Ecol. Prog. Ser.* **556**, 1–16 (2016).
1035 24. Ruggiero, M. V. *et al.* Temporal changes of genetic structure and diversity in a
1036 marine diatom genus discovered via metabarcoding. *Environ. DNA* **4**, 763–775
1037 (2022).
1038 25. Pierella Karlusich, J. J., Lombard, F., Irisson, J.-O., Bowler, C. & Foster, R. A.
1039 Coupling imaging and omics in plankton surveys: State-of-the-art, challenges,
1040 and future directions. *Front. Mar. Sci.* **9**, (2022).
1041 26. Bowler, C. *et al.* The Phaeodactylum genome reveals the evolutionary history of
1042 diatom genomes. *Nature* **456**, 239–244 (2008).
1043 27. Smith, S. R. *et al.* Evolution and regulation of nitrogen flux through
1044 compartmentalized metabolic networks in a marine diatom. *Nat. Commun.* **10**,
1045 4552 (2019).
1046 28. Falciatore, A., Jaubert, M., Bouly, J.-P., Bailleul, B. & Mock, T. Diatom Molecular
1047 Research Comes of Age: Model Species for Studying Phytoplankton Biology
1048 and Diversity. *Plant Cell* **32**, 547–572 (2020).
1049 29. Grigoriev, I. V. *et al.* PhycoCosm, a comparative algal genomics resource.
1050 *Nucleic Acids Res.* **49**, D1004–D1011 (2021).
1051 30. Durkin, C. A. *et al.* Frustule-related gene transcription and the influence of
1052 diatom community composition on silica precipitation in an iron-limited
1053 environment. *Limnol. Oceanogr.* **57**, 1619–1633 (2012).

1054 31. Alexander, H. *et al.* Functional group-specific traits drive phytoplankton
1055 dynamics in the oligotrophic ocean. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E5972–9
1056 (2015).
1057 32. Pearson, G. A. *et al.* Metatranscriptomes reveal functional variation in diatom
1058 communities from the Antarctic Peninsula. *ISME J.* **9**, 2275–2289 (2015).
1059 33. Cohen, N. R. *et al.* Diatom transcriptional and physiological responses to
1060 changes in iron bioavailability across ocean provinces. *Front. Mar. Sci.* **4**,
1061 (2017).
1062 34. Lampe, R. H. *et al.* Divergent gene expression among phytoplankton taxa in
1063 response to upwelling. *Environ. Microbiol.* **20**, 3069–3082 (2018).
1064 35. Kolody, B. C. *et al.* Nitrogen and Iron Availability Drive Metabolic Remodeling
1065 and Natural Selection of Diverse Phytoplankton during Experimental Upwelling.
1066 *mSystems* **7**, e0072922 (2022).
1067 36. Pesant, S. *et al.* Open science resources for the discovery and analysis of Tara
1068 Oceans data. *Sci Data* **2**, 150023 (2015).
1069 37. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina
1070 amplicon data. *Nat. Methods* **13**, 581–583 (2016).
1071 38. Murali, A., Bhargava, A. & Wright, E. S. IDTAXA: a novel approach for accurate
1072 taxonomic classification of microbiome sequences. *Microbiome* **6**, 140 (2018).
1073 39. Guillou, L. *et al.* The Protist Ribosomal Reference database (PR2): a catalog of
1074 unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy.
1075 *Nucleic Acids Res.* **41**, D597–604 (2013).
1076 40. Morin, S., Gómez, N., Tornés, E., Licursi, M. & Rosebery, J. Benthic diatom
1077 monitoring and assessment of freshwater environments: Standard methods and
1078 future challenges. in *Aquatic Biofilms: Ecology, Water Quality and Wastewater*

1079 *Treatment* 111–124 (Caister Academic Press, 2016).
1080 41. Pierella Karlusich, J. J. *et al.* A robust approach to estimate relative
1081 phytoplankton cell abundances from metagenomes. *Mol. Ecol. Resour.* **23**, 16–
1082 40 (2023).
1083 42. Lavrinienko, A., Jernfors, T., Koskimäki, J. J., Pirttilä, A. M. & Watts, P. C. Does
1084 Intraspecific Variation in rDNA Copy Number Affect Analysis of Microbial
1085 Communities? *Trends Microbiol.* **29**, 19–27 (2021).
1086 43. Fu, R. & Gong, J. Single Cell Analysis Linking Ribosomal (r)DNA and rRNA
1087 Copy Numbers to Cell Size and Growth Rate Provides Insights into Molecular
1088 Protistan Ecology. *J. Eukaryot. Microbiol.* **64**, 885–896 (2017).
1089 44. Pérez-Ortín, J. E. *et al.* Cell volume homeostatically controls the rDNA repeat
1090 copy number and rRNA synthesis rate in yeast. *PLoS Genet.* **17**, e1009520
1091 (2021).
1092 45. Hendry, K. R. *et al.* Competition between silicifiers and non-silicifiers in the past
1093 and present ocean and its evolutionary impacts. *Front. Mar. Sci.* **5**, (2018).
1094 46. Piredda, R. *et al.* Diversity and temporal patterns of planktonic protist
1095 assemblages at a Mediterranean Long Term Ecological Research site. *FEMS*
1096 *Microbiol. Ecol.* **93**, (2017).
1097 47. Qin, Y. *et al.* Effects of error, chimera, bias, and GC content on the accuracy of
1098 amplicon sequencing. *mSystems* **8**, e0102523 (2023).
1099 48. Chen, B. Patterns of thermal limits of phytoplankton. *J. Plankton Res.* **37**, 285–
1100 292 (2015).
1101 49. Olguín Salinas, H. F., Alder, V. A., Puig, A. & Boltovskoy, D. Latitudinal diversity
1102 patterns of diatoms in the Southwestern Atlantic and Antarctic waters. *J.*
1103 *Plankton Res.* **37**, 659–665 (2015).

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

evidence approach. *Phycologia* **59**, 422–436 (2020).

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

1154 70. Kumazawa, M. *et al.* Molecular phylogeny of fucoxanthin-chlorophyll a/c proteins
1155 from Chaetoceros gracilis and Lhcq/Lhcf diversity. *Physiol. Plant.* **174**, e13598
1156 (2022).
1157 71. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related
1158 eukaryotic plankton lineages abundant in the sunlit ocean. *Cell Genom* **2**,
1159 100123 (2022).
1160 72. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E.
1161 M. BUSCO: assessing genome assembly and annotation completeness with
1162 single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
1163 73. Chaffron, S. *et al.* Environmental vulnerability of the global ocean epipelagic
1164 plankton community interactome. *Sci Adv* **7**, (2021).
1165 74. de Vargas, C. *et al.* Ocean plankton. Eukaryotic plankton diversity in the sunlit
1166 ocean. *Science* **348**, 1261605 (2015).
1167 75. Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from
1168 the Tara Oceans expedition. *Sci Data* **4**, 170093 (2017).
1169 76. Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W. & Huse, S. M. A method
1170 for studying protistan diversity using massively parallel sequencing of V9
1171 hypervariable regions of small-subunit ribosomal RNA genes. *PLoS One* **4**,
1172 e6372 (2009).
1173 77. Stoeck, T. *et al.* Multiple marker parallel tag environmental DNA sequencing
1174 reveals a highly complex eukaryotic community in marine anoxic water. *Mol.*
1175 *Ecol.* **19 Suppl 1**, 21–31 (2010).
1176 78. Martin, M. Cutadapt removes adapter sequences from high-throughput
1177 sequencing reads. *EMBnet J.* **17**, 10 (2011).
1178 79. Carradec, Q. *et al.* A global ocean atlas of eukaryotic genes. *Nat. Commun.* **9**,

1179 373 (2018).
1180 80. Zayed, A. A. *et al.* Cryptic and abundant marine viruses at the evolutionary
1181 origins of Earth's RNA virome. *Science* **376**, 156–162 (2022).
1182 81. Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence
1183 alignment program. *Brief. Bioinform.* **9**, 286–298 (2008).
1184 82. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-
1185 likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
1186 83. Community Ecology Package [R package vegan version 2.6-6.1]. (2024).
1187 84. Wickham, H. Data Analysis. in *Use R!* 189–201 (Springer International
1188 Publishing, Cham, 2016).
1189 85. Tennekes, M. Treemap Visualization [R package treemap version 2.4-4]. (2023).
1190 86. Yu, G. Scatter Pie Plot [R package scatterpie version 0.2.3]. (2024).
1191 87. Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, O'Hara R,
1192 Solymos P, et al. vegan: Community Ecology Package. https://CRAN.R-
1193 project.org/package=vegan. (2022).
1194 88. Sanchez, G. https://github.com/gastonstat/plsdepot (2015).
1195 89. Bendschneider, K. & Robinson, R. J. *Journal of Marine Research* 11, 87–96
1196 (1952).
1197 90. Murphy, J. & Riley, J. P. *Analytica Chimica Acta* 27, 31–36 (1962).
1198 91. Hasle, Grethe R., Erik E. Syvertsen, & Cecilie H. von Quillfeldt. *Fossula*
1199 *arctica* gen. nov., spec. nov., a marine Arctic araphid diatom. *Diatom Research*
1200 11.2, 261-272 (1996).
1201 92. Niang, G. et al. METdb: a genomic reference database for marine species.
1202 Preprint at https://doi.org/10.7490/F1000RESEARCH.1118000.1 (2020).
1203 93. Seeleuthner, Y., et al. Single-cell genomics of multiple uncultured stramenopiles

reveals underestimated functional diversity across oceans. *Nature Communications*, 9(1), 310 (2018).

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

95. Capella-Gutierrez, S., Silla-Martinez, S.M., Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25:1972-1973 (2009).