

1 Differential global distribution of marine picocyanobacteria gene
2 clusters reveals distinct niche-related adaptive strategies

3 Hugo Doré^{a,1}, Ulysse Guyet^{a,1}, Jade Leconte^a, Gregory K. Farrant^a, Benjamin Alric^a, Morgane
4 Ratin^a, Martin Ostrowski^{b,2}, Mathilde Ferrieux^a, Loraine Brillet-Guéguen^{c,d}, Mark Hoebeke^c, Jukka
5 Siltanen^c, Gildas Le Corguillé^c, Erwan Corre^c, Patrick Wincker^{f,g}, David J. Scanlan^b, Damien
6 Eveillard^{h,g}, Frédéric Partensky^a, and Laurence Garczarek^{a,g,*}

7 ^aSorbonne Université, CNRS, UMR 7144 Adaptation and Diversity in the Marine Environment
8 (AD2M), Station Biologique de Roscoff (SBR), Roscoff, France; ^b School of Life Sciences,
9 University of Warwick, Coventry CV4 7AL, UK; ^cCNRS, FR 2424, ABiMS Platform, Station
10 Biologique de Roscoff (SBR), Roscoff, France; ^dSorbonne Université, CNRS, UMR 8227,
11 Integrative Biology of Marine Models (LBI2M), Station Biologique de Roscoff (SBR), Roscoff,
12 France; ^eGenoscope, Institut de biologie François-Jacob, Commissariat à l’Energie Atomique
13 (CEA), Université Paris-Saclay, Evry, France; ^fGénomique Métabolique, Genoscope, Institut de
14 biologie François Jacob, CEA, CNRS, Université d’Evry, Université Paris-Saclay, Evry, France;
15 ^gResearch Federation (FR2022) Tara Océans GO-SEE, Paris, France; ^hNantes Université, Centrale
16 Nantes, CNRS, LS2N, UMR 6004, Nantes, France.

17 ¹H.D. and U.G. contributed equally to this work

18 ²Current address: Climate Change Cluster, University of Technology, Broadway NSW 2007,
19 Australia

20 *To whom correspondence should be addressed. Email: laurence.garczarek@sb-roscoff.fr.
21 phone number: +33 2 98 29 25 38

22

23 **Author Contributions:** Paste the author contributions here.

24 **Competing Interest Statement:** The authors declare no competing interests.

25 **Classification:** Biological Sciences: Microbiology and Environmental Sciences

26 **Keywords:** *Prochlorococcus*, *Synechococcus*, niche partitioning, Tara Oceans, metagenomics

27 **This PDF file includes:**

28 Main Text

29 Figures 1 to 7

30 **Abstract**

31 The ever-increasing number of available microbial genomes and metagenomes provide new
32 opportunities to investigate the links between niche partitioning and genome evolution in the
33 ocean, notably for the abundant and ubiquitous marine picocyanobacteria *Prochlorococcus* and
34 *Synechococcus*. Here, by combining metagenome analyses of the *Tara Oceans* dataset with
35 comparative genomics, including phyletic patterns and genomic context of individual genes from
36 256 reference genomes, we first showed that picocyanobacterial communities thriving in
37 different niches possess distinct gene repertoires. We then managed to identify clusters of
38 adjacent genes that display specific distribution patterns in the field (CAGs) and are thus
39 potentially involved in the adaptation to particular environmental niches. Several CAGs are likely
40 involved in the uptake or incorporation of complex organic forms of nutrients, such as
41 guanidine, cyanate, cyanide, pyrimidine or phosphonates, which might be either directly used by
42 cells, for e.g. the biosynthesis of proteins or DNA, or degraded into inorganic nitrogen and/or
43 phosphorus forms. We also highlight the frequent presence of CAGs involved in polysaccharide
44 capsule biosynthesis in *Synechococcus* populations thriving in both nitrogen- and phosphorus-
45 depleted areas, which are absent in low-iron regions, suggesting that the complexes they
46 encode may be too energy-consuming for picocyanobacteria thriving in these areas. In contrast,
47 *Prochlorococcus* populations thriving in iron-depleted areas specifically possess an alternative
48 respiratory terminal oxidase, potentially involved in the reduction of Fe(III) into Fe(II). Together,
49 this study provides insights into how these key members of the phytoplankton community might
50 behave in response to ongoing global change.

51 **Significance Statement**

52 Picocyanobacteria face various environmental conditions in the ocean and numerous studies
53 have shown that genetically distinct ecotypes colonize different niches. Yet the functional basis
54 of their adaptation remains poorly known, essentially due to the large number of genes of yet
55 unknown function, many of which have little or no beneficial effect on fitness. Here, by
56 combining comparative genomics and metagenomics approaches, we have identified not only
57 single genes but also entire gene clusters, potentially involved in niche adaptation. Although
58 being sometimes present in only one or a few sequenced strains, they occur in a large part of

59 the population in specific ecological niches and thus constitute precious targets for elucidating
60 the biochemical function of yet unknown niche-related genes.

61

62 **Main Text**

63

64 **Introduction**

65 During the last two decades the sequencing of a large number of microbial genomes (more than
66 425,000 were available in Genbank in July 2022) has allowed tremendous advances in the
67 delineation of core, accessory and unique gene repertoires within closely related organisms by
68 building clusters of likely orthologous genes (CLOGs) based on sequence homology (1–4).
69 Although this approach was tentatively used to identify the genetic basis of niche adaptation,
70 relatively few genes were identified as being specific to particular ecotypes and thus potentially
71 involved in niche adaptation (5–9). Various reasons may underpin this difficulty to identify
72 niche-specific genes by a mere comparative genomics approach. These include the still fairly low
73 number of genomes available given extensive known microbial genomic diversity (10), a lack of
74 ecological representation due to cultivation biases, a limited knowledge of physiological traits of
75 sequenced strains and/or the imprecise delineation of ecotypes and of the limits of their
76 realized environmental niches *sensu* (11), especially for lineages present in low abundance in
77 the field (12–14).

78 An alternative to comparative genomics to better decipher the link between niche
79 partitioning and genome evolution consists of using the rapidly growing number of
80 metagenomes. Besides triggering the generation of numerous metagenome-assembled
81 genomes (MAGs), allowing to fill the gap for yet uncultured microbial taxa and/or ecotypes (15,
82 16), metagenome recruitment analyses using reference genomes have also allowed scientists to
83 identify spatial or temporal niche-specific genes (17–19). In this context, due to their abundance
84 and ubiquity in the field and the numerous available genomes, single amplified genomes (SAGs)
85 and MAGs, marine picocyanobacteria constitute highly pertinent models for these metagenomic
86 recruitment approaches. The *Prochlorococcus* and *Synechococcus* genera are indeed the two
87 most abundant members of the phytoplankton community, *Prochlorococcus* being restricted to
88 the 40°S-50°N latitudinal band, while *Synechococcus* distribution extends from the equator to
89 subpolar waters (20, 21). Furthermore, physiological and environmental studies have allowed

90 scientists to decipher their genetic diversity and their main physiological traits as well as to
91 delineate ecotypes or Ecologically Significant Taxonomic Units (ESTUs), i.e., genetic groups
92 within clades occupying a given ecological niche, notably using *Tara* oceans metagenomic data
93 at the global scale (22). While three major ESTU assemblages were identified for
94 *Prochlorococcus* in surface waters, whose distribution was found to be mainly driven by
95 temperature and iron (Fe) availability, eight distinct assemblages were identified for
96 *Synechococcus* depending on three main environmental parameters (temperature, Fe and
97 phosphate availability). Nevertheless, few studies have so far integrated our wide knowledge of
98 ecotype distributions and the genetic and functional diversity of these organisms to identify
99 niche- and/or ecotype-specific genes based on their relative abundance in the field (12, 23–26).
100 Furthermore, most of these previous studies have focused on the abundance of individual
101 genes, or more rarely, on a few genomic regions with known functions, e.g. involved in nitrogen
102 or phosphorus uptake and assimilation (27, 28).

103 Here, by using a network approach to integrate metagenome analyses of the *Tara* Oceans
104 dataset and synteny of individual accessory genes in 256 reference genomes, MAGs and SAGs,
105 we managed to identify clusters of adjacent genes that display specific distribution patterns
106 along the *Tara* Oceans transect. This led us to the unveil niche- and/or ecotype-specific genomic
107 regions, including several previously unreported and sometimes only present in a few or even
108 single genomes, potentially involved in the adaptation to the main ecological niches occurring in
109 the marine environment (N, P and/or Fe-limited as well as cold vs. warm areas). Delineation of
110 these gene clusters also led us to predict the putative functions of previously uncharacterized
111 genes in these genomic regions based on genes functionally annotated in the same cluster.
112 Altogether, this study provides unique insights into the functional basis of microbial niche
113 partitioning and the molecular bases of fitness in key members of the phytoplankton
114 community.

115

116

117 **Results and Discussion**

118

119 **Different picocyanobacterial communities exhibit distinct gene repertoires**

120 To analyze the distribution of *Prochlorococcus* and *Synechococcus* reads along the *Tara* Oceans
121 transect, metagenomic reads corresponding to the bacterial size fraction were recruited against
122 256 picocyanobacterial reference genomes, including 178 whole genome sequences (WGS), and
123 a selection of 48 SAGs and 30 MAGs, primarily representative of still uncultured lineages (e.g.
124 *Prochlorococcus* HLIII-IV, *Synechococcus* EnvA or EnvB). This yielded a total of 1.07 billion
125 recruited reads, of which 87.7% mapped onto *Prochlorococcus* genomes and 12.3% onto
126 *Synechococcus* ones, which were then functionally assigned by mapping them on the manually
127 curated Cyanorak v2.1 CLOG database (29). In order to identify picocyanobacterial genes
128 potentially involved in niche adaptation, we analyzed the distribution across the oceans of
129 flexible genes (i.e., non-core genes in Cyanorak *Prochlorococcus* and *Synechococcus* reference
130 genomes). *Tara* Oceans stations were first clustered according to the relative abundance of
131 flexible genes. This clustering resulted in three well-defined clusters for *Prochlorococcus* (left
132 tree in Fig. 1A), which matched quite well those obtained when stations were clustered
133 according to the relative abundance of *Prochlorococcus* ESTUs, as assessed using the high-
134 resolution marker gene *petB*, encoding the cytochrome *b₆* (right tree in Fig. 1A; see also (22)).
135 Only a few discrepancies can be observed between the two trees, including stations TARA-070
136 that displayed one of the most disparate ESTU compositions and TARA-094, dominated by the
137 rare HLID ESTU (Fig. 1A). For *Synechococcus*, there was also a good consistency between
138 dendrograms obtained from flexible gene abundance and relative abundance of ESTUs (Fig. 1B).
139 Of the eight assemblages of stations discriminated based on the relative abundance of ESTUs
140 (Fig. 1B), most were retrieved in the clustering based on flexible gene abundance, except for a
141 few intra-assemblage switches between stations, notably those dominated by ESTU IIA (Fig. 1B).
142 Despite these few variations between *Synechococcus* trees, four major clusters can be clearly
143 delineated in both trees, corresponding to four broadly defined ecological niches, namely i) cold,
144 nutrient-rich, pelagic or coastal environments (blue and light red in Fig. 1B), ii) Fe-limited
145 environments (purple and grey), iii) temperate, P-depleted, Fe-replete areas (yellow) and iv)
146 warm, N-depleted, Fe-replete regions (dark red). This correspondence between taxonomic and
147 functional information was also confirmed by the high congruence between distance matrices
148 based on ESTU relative abundance and on CLOG relative abundance (p-value < 10⁻⁴, mantel test

149 $r=0.84$ and $r=0.75$ for *Synechococcus* and *Prochlorococcus*, respectively; dataset 1-4). Altogether,
150 this indicates that distinct picocyanobacterial communities, as assessed based on a single
151 taxonomic marker, also display different gene repertoires. As previously suggested for
152 *Prochlorococcus* (30), this strong correlation between taxonomy and gene content strengthens
153 the idea that, in both genera, the evolution of the accessory genome mainly occurs by vertical
154 transmission, with a relatively low extent of lateral gene transfer.

155

156 **Distribution of flexible genes is tightly linked to environmental parameters and ESTUs**

157 In order to reduce the amount of data and better interpret the global distribution of
158 picocyanobacterial gene content, a correlation network of genes was built for each genus based
159 on relative abundance profiles of genes across *Tara* Oceans samples. Its analysis emphasized
160 four main modules of genes for *Prochlorococcus* (Fig. S1A) and five main modules for
161 *Synechococcus* (Fig. S1B), each gene module being abundant in a different set of stations. These
162 modules were then associated with the available environmental parameters (Figs. 2A-B) and to
163 the relative abundance of *Prochlorococcus* or *Synechococcus* ESTUs at each station (Figs. 2C-D).
164 For instance, the *Prochlorococcus brown* module was strongly correlated with nutrient
165 concentrations, particularly nitrate and phosphate, and strongly anti-correlated with Fe
166 availability (Fig. 2A). This module thus corresponds to genes preferentially found in Fe-limited
167 high-nutrient low-chlorophyll (HNLC) areas. Indeed, the *brown* module *eigengenes* (Fig. S1A),
168 representative of the abundance profiles of genes of this module at the different *Tara* Oceans
169 stations, showed the highest abundances at stations TARA-100 to 125, localized in the South and
170 North Pacific Ocean, as well as at TARA-052, a station located close to the northern coast of
171 Madagascar and likely influenced by the Indonesian throughflow originating from the tropical
172 Pacific Ocean (22, 31). Furthermore, the correlation of the *Prochlorococcus brown* module with
173 the relative abundance of ESTUs at each station showed that it is also strongly associated with
174 the presence of HLIIIA and HLIVA (Fig. 2C), previously shown to constitute the dominant
175 *Prochlorococcus* ESTUs in low-Fe environments (22, 32, 33) but also the LLIB ESTU, found to
176 dominate the LLI population in these low-Fe areas (22). Altogether, this example and analyses of
177 all other *Prochlorococcus* and *Synechococcus* modules (SI Text1) show that the communities
178 colonizing cold, Fe-, N- and/or P-depleted niches possess specific gene repertoires potentially
179 involved in their adaptation to these peculiar environmental conditions.

180

181 **Identification of individual genes potentially involved in niche partitioning**

182 In order to identify flexible genes related to particular environmental conditions and to specific
183 ESTU assemblages, we correlated relative abundance profiles of each gene to the *eigengene*
184 vector of its corresponding module in order to identify the most representative genes of each
185 module and thus the genes specifically present (or absent) in a given set of stations (Dataset 5,
186 Figs. 3 and S2). Most genes retrieved this way encode proteins of unknown or hypothetical
187 function (85.7% of 7,485 genes). Still, among the genes with a functional annotation (Dataset 6),
188 a large fraction seems to have a function related to their realized environmental niche (Figs. 3
189 and S2). For instance, many genes involved in the transport and assimilation of nitrite and
190 nitrate (*nirA*, *nirX*, *moaA-C*, *moaE*, *mobA*, *moeA*, *narB*, *M*, *nrtP*; all part of the same genomic
191 island: Pro_GI004; (9)) as well as cyanate, an organic form of nitrogen (*cynA*, *B*, *D*, *S*; part of
192 Pro_GI033), are enriched in the *Prochlorococcus blue* module, which is correlated with the
193 HLIIA-D ESTU and to low inorganic N, P and Si levels and anti-correlated with Fe availability (Fig.
194 2A-C). This is consistent with previous studies showing that while few *Prochlorococcus* strains in
195 culture possess the *nirA* gene and even less the *narB* gene, natural *Prochlorococcus* populations
196 inhabiting N-poor areas do possess one or both of these genes (34–36). Similarly, numerous
197 genes among the most representative genes of *Prochlorococcus brown*, *red* and *turquoise*
198 modules are related to adaptation of HLIIIA/IVA, HLIA and LLIA ESTUs to Fe-limited, cold P-
199 limited and cold, mixed waters, respectively (Fig. 3), and comparable results were obtained for
200 *Synechococcus*, although the niche delineation was fuzzier than for *Prochlorococcus* at the
201 module level (Fig. S2). These results therefore constitute a proof of concept that this network
202 analysis was able to retrieve niche-related genes from metagenomics data.

203

204 **Identification of CAGs potentially involved in niche partitioning**

205 In order to better understand the function of niche-related genes, notably the numerous ones
206 encoding conserved hypothetical proteins, we then integrated these data with knowledge on
207 the gene synteny in reference genomes using a network approach (Datasets 7 and 8). This led us
208 to identify clusters of adjacent genes in reference genomes, several not previously reported in
209 the literature, encompassing genes with similar distribution and abundance *in situ* and thus

210 potentially involved in the same metabolic pathway (Figs. 4, S3 and S4; Dataset 6). Hereafter,
211 these ecologically representative clusters of adjacent genes will be called 'CAGs'.

212 Regarding nitrogen, the well-known nitrate/nitrite gene cluster involved in uptake and
213 assimilation of inorganic forms of nitrogen (see above) is present in most *Synechococcus*
214 genomes (Dataset 6) and expectedly not restricted to a particular niche in natural
215 *Synechococcus* populations, as shown by its quasi-absence from Weighted Correlation Network
216 Analysis (WGCNA) modules. In *Prochlorococcus*, this cluster is separated into two CAGs, most
217 genes being included in ProCAG_002, present in only 13 out of 118 *Prochlorococcus* genomes,
218 while *nirA* and *nirX* form an independent CAG (ProCAG_001) due to their presence in many
219 more genomes. Both CAGs are particularly enriched in *Prochlorococcus* populations thriving in
220 low-N areas (Fig. S5A-B), as previously demonstrated by several authors (34–36). In
221 *Prochlorococcus*, the quasi-core *ureA-G/urtB-E* genomic region was also found as a CAG
222 (ProCAG_003) since it was comparatively impoverished in low-Fe compared to other regions
223 (Fig. S5C-D) in agreement with its presence in only two out of six HLIII/IV genomes. In addition,
224 we also uncovered several other *Prochlorococcus* and *Synechococcus* CAGs that seem to be
225 involved in the transport and/or assimilation of more unusual and/or complex forms of
226 nitrogen, including guanidine, cyanate, cyanide and possibly pyrimidine, which might either be
227 degraded into elementary N, P or Fe molecules or possibly directly used by the cells for e.g. the
228 biosynthesis of proteins or DNA. Indeed, we detected in both genera a CAG (ProCAG_004 and
229 SynCAG_001 ; Figs. S6A-B, Dataset 6) that encompasses *speB2*, an ortholog of *Synechocystis* PCC
230 6803 *sl1077*, previously annotated as encoding an agmatinase (23, 37) and which was recently
231 characterized as a guanidinase that degrades guanidine rather than agmatine to urea and
232 ammonium (38). Interestingly *E. coli*, and likely other microorganisms as well, produce guanidine
233 under nutrient-poor conditions, suggesting that guanidine metabolism is biologically significant
234 and prevalent in natural environments (38, 39). Furthermore, the *ykkC* riboswitch candidate,
235 which was shown to specifically sense guanidine and to control the expression of a variety of
236 genes involved in either guanidine metabolism or nitrate, sulfate, or bicarbonate transport, is
237 located immediately upstream of this CAG in *Synechococcus* reference genomes, all genes of this
238 cluster being predicted by RegPrecise 3.0 to be regulated by this riboswitch (Fig. S6C; (39, 40)).
239 The presence of *hypA* and *B* homologs within this CAG furthermore suggests that, in the
240 presence of guanidine, the latter could be involved in the insertion of Ni_2^+ , or another metal

241 cofactor, in the active site of guanidinase. Additionally, we speculate that the next three genes
242 encoding an ABC transporter, similar to the TauABC taurine transporter in *E. coli* (Fig. S6C), could
243 be involved in guanidine transport in low-N areas. Of note, the presence of a gene encoding a
244 putative Rieske Fe-sulfur protein (CK_00002251), downstream of this gene cluster in most
245 *Synechococcus/Cyanobium* genomes possessing this CAG, seems to constitute a specificity
246 compared to its homologs in *Synechocystis* sp. PCC 6803 and might explain why this CAG is
247 absent from picocyanobacteria thriving in low-Fe areas, while it is present in a large proportion
248 of the population in most other oceanic areas (Figs. S6A-B).

249 As concerns compounds containing a cyano radical ($C\equiv N$), the cyanate transporter genes
250 (*cynABD*) are scarce in both *Prochlorococcus* (present only in two HLI and five HLII genomes) and
251 *Synechococcus* genomes (mostly in clade III strains; (9, 41, 42)). In the field, a small proportion of
252 the *Prochlorococcus* community possesses the corresponding CAG (ProCAG_005; Fig. S7A-B),
253 also including the conserved hypothetical gene CK_00055128, in warm, Fe-replete waters, while
254 these genes were not included in a module, and thus not in a CAG, in *Synechococcus* (Dataset 6;
255 Fig. S7C). Interestingly, we also uncovered a 7-gene CAG (ProCAG_006 and SynCAG_002),
256 encompassing a putative nitrilase gene (*nitC*), which also suggests that most *Synechococcus* cells
257 and a more variable proportion of the *Prochlorococcus* population could use nitriles or cyanides
258 in warm, Fe-replete waters and more particularly in low-N areas such as the Indian Ocean (Fig.
259 5A-B). The whole operon (*nitHBCDEFG*; Fig. 5C), called Nit1C, was shown to be upregulated in
260 the presence of cyanide and to trigger an increase in the rate of ammonia accumulation in the
261 heterotrophic bacterium *Pseudomonas fluorescens* (43), suggesting that like cyanate, cyanide
262 could constitute an alternative nitrogen source in marine picocyanobacteria as well. Yet, given
263 the potential toxicity of these $C\equiv N$ -containing compounds, we cannot exclude that these CAGs
264 could also be devoted to cell detoxification (39, 41), as it is the case for arsenate and chromate
265 (44, 45), which act as analogs of phosphate and sulfate respectively, and are toxic to marine
266 phytoplankton (46).

267 Also noteworthy is the presence of a CAG encompassing *asnB*, *pyrB2* and *pydC*
268 (ProCAG_007, SynCAG_003, Fig. S8), which could contribute to an alternative pyrimidine
269 biosynthesis pathway and thus provide another way for cells to recycle complex nitrogen forms.
270 While this CAG is found in only one fifth of HLII genomes and in quite specific locations for

271 *Prochlorococcus*, notably in the Red Sea, it is found in most *Synechococcus* cells in warm, Fe-
272 replete, N and P-depleted niches, consistent with its phyletic pattern showing its absence only
273 from most clade I, IV, CRD1 and EnvB genomes (Fig. S8; Dataset 6). More generally, most N-
274 uptake and assimilation genes in both genera were specifically absent from Fe-depleted areas,
275 including the *nirA/narB* CAG for *Prochlorococcus*, as mentioned by Kent et al. (30) as well as
276 guanidinase and nitrilase CAGs. In contrast, picocyanobacterial populations present in low-Fe
277 areas possess, in addition to the core ammonium transporter *amt1*, a second transporter *amt2*,
278 also present in cold areas for *Synechococcus* (Fig. S9). Additionally, *Prochlorococcus* populations
279 thriving in HNLC areas also possess two amino acid-related CAGs that are quasi-core in
280 *Synechococcus*, the first one involved in polar amino acids N-II transport system (ProCAG_008;
281 *natF-G-H-bgtA*; (47); Fig. S10A-B) and the second one (*leuDh*, *soxA*, CK_00001744, ProCAG_009,
282 Fig. S10C-D) that notably encompasses a leucine dehydrogenase, able to produce ammonium
283 from branched-chain amino acids. Thus, the primary nitrogen sources for picocyanobacterial
284 populations dwelling in Fe-limited areas seem to be ammonium and amino acids.

285 Adaptation to phosphorus depletion has been well documented in marine
286 picocyanobacteria showing that while in P-replete waters *Prochlorococcus* and *Synechococcus*
287 essentially rely on inorganic phosphate acquired by core transporters (PstABC), strains isolated
288 from low-P regions and natural populations thriving in these areas additionally contain a
289 number of accessory genes related to P metabolism, located in specific genomic islands (9, 25–
290 28, 48). Here, we indeed found that virtually the whole *Prochlorococcus* population in the
291 Mediterranean Sea, the Gulf of Mexico and the Western North Atlantic Ocean, which are known
292 to be P-limited (26, 49), contained the *phoBR* operon (ProCAG_010, Fig. S11A-B) that encodes a
293 two-component system response regulator, as well as the ProCAG_011, including the alkaline
294 phosphatase *phoA*. By comparison, in *Synechococcus*, we only identified the *phoBR* CAG
295 (SynCAG_005, Fig. S11C) that is systematically present in warm waters whatever their limiting
296 nutrient, in agreement with its phyletic pattern in reference genomes showing its specific
297 absence from cold thermotypes (clades I and IV, Dataset 6). Furthermore, although our analysis
298 did not retrieve them within CAGs due to the variability of the content and order of genes in this
299 genomic region, even within each genus, several other P-related genes were enriched in low-P
300 areas but interestingly partially differed between *Prochlorococcus* and *Synechococcus* (Figs. S11,

301 3, S2 and Dataset 6). While the genes putatively encoding a chromate transporter (ChrA) and an
302 arsenate efflux pump ArsB were present in both genera in different proportions, a putative
303 transcriptional phosphate regulator related to PtrA (CK_00056804; (50)) was specific to
304 *Prochlorococcus*. *Synechococcus* in contrast harbors a large variety of alkaline phosphatases
305 (PhoX, CK_00005263 and CK_00040198) as well as the phosphate transporter SphX (Fig. S11).

306 A second alternative P form are phosphonates, i.e. reduced organophosphorus compounds
307 containing C–P bonds, which constitute up to 25% of the high-molecular-weight dissolved
308 organic P pool in the open ocean (51). Indeed, the quasi-totality of the *Prochlorococcus*
309 population of the most P-limited areas of the ocean possess, additionally to the core
310 phosphonate ABC transporter (*phnD1-C1-E1*), a second previously unreported putative
311 phosphonate transporter (*phnC2-D2-E2-E3*; ProCAG_012; Fig. 6A), while these genes are only
312 present in a few *Prochlorococcus* (including MIT9314) and no *Synechococcus* genomes.
313 Furthermore, as previously mentioned in several studies (52–54), a fairly low proportion of
314 *Prochlorococcus* populations thriving in low-P areas also possess a gene cluster encompassing
315 the *phnYZ* operon, involved in C-P bond cleavage, the putative phosphite dehydrogenase *ptxD*
316 as well as the phosphite and methylphosphonate transporter *ptxABCD* (ProCAG_0013, Dataset 6,
317 and Fig. 6B, (54–56)). Compared to these previous studies that mainly reported the presence of
318 these genes in *Prochlorococcus* cells from the North Atlantic Ocean, here we show that they
319 actually occur in a much larger geographic area, including the Mediterranean Sea, the Gulf of
320 Mexico and the ALOHA station (TARA_132) in the North Pacific, and are also present in an even
321 larger proportion of the *Synechococcus* population (Fig. S12, Dataset 6). Interestingly,
322 *Synechococcus* cells from the Mediterranean Sea, dominated by clade III, seem to lack *phnYZ*, in
323 agreement with the phyletic pattern of these genes in reference genomes, showing the absence
324 of this two-gene operon in the sole clade III strain that possesses the *ptxABDC* gene cluster. In
325 contrast, the presence of the complete gene set (*ptxABDC-phnYZ*) in the North Atlantic and at
326 the entrance of the Mediterranean Sea as well as in several clade II reference genomes rather
327 suggests that it is primarily attributable to this specific clade. Altogether, our data indicate that
328 at least part of the natural populations of both *Prochlorococcus* and *Synechococcus* would be
329 able to assimilate phosphonate and phosphite as alternative P-sources in low-P areas using the
330 *ptxABDC-phnYZ* operon. Yet, the fact that no picocyanobacterial genome except *P. marinus* RS01

331 (Fig. 6C) possesses both *phnC2-D2-E2-E3* and *phnYZ*, raises the question of how the
332 phosphonate taken up by the *phnC2-D2-E2-E3* transporter is metabolized in these cells. Finally,
333 although the Mediterranean Sea is not known to be N-limited, all reference clade III genomes
334 possess a complete set of genes involved in the assimilation of organic nitrogen (Dataset 6),
335 suggesting that at least part of these organic nutrients might also be a source of organic
336 phosphorus.

337 As for macronutrients, it has been hypothesized that the survival of marine
338 picocyanobacteria in low-Fe regions was made possible through several strategies, including the
339 elimination from the genomes of genes encoding proteins that contain Fe as a cofactor, the
340 replacement of Fe by another metal cofactor, and the acquisition of genes involved in Fe uptake
341 and storage (24, 25, 30, 33, 57). Accordingly, several CAGs encompassing genes encoding
342 proteins interacting with Fe were found in the present study to be anti-correlated with HNLC
343 regions in both genera. These include three subunits of the (photo)respiratory complex
344 succinate deshydrogenase (SdhABC, ProCAG_014, SynCAG_006, Fig. S13; (58)) as well as Fe-
345 containing proteins encoded in most of the abovementioned CAGs involved in N or P
346 metabolism, such as the guanidinase CAG (Fig. S6), the NitC1 CAG (Fig. 5), the *pyrB2* CAG (Fig.
347 S8), the phosphonate CAGs (Figs. 6 and S12) and the urea and inorganic nitrogen CAGs (Fig. S5).
348 Most *Synechococcus* cells thriving in Fe-replete areas also possess the *sodT/sodX* CAG
349 (SynCAG_007, Fig. S14A-B) involved in nickel transport and maturation of the Ni-superoxide
350 dismutase (SodN), these three genes being in contrast core in *Prochlorococcus*. Additionally,
351 *Synechococcus* from Fe-replete areas, notably from the Mediterranean Sea and the Indian
352 Ocean, specifically possess two CAGs (Syn CAG_008 and 009; Fig. S14C-D), involved in the
353 biosynthesis of a polysaccharide capsule that appear to be most similar to the *E. coli* groups 2
354 and 3 *kps* loci (59). These extracellular structures, known to provide protection against biotic or
355 abiotic stress, were recently shown in *Klebsiella* to provide a clear fitness advantage in nutrient-
356 poor conditions since they were associated with increased growth rates and population yields
357 (60). Yet, while these authors suggested that capsules may play a role in Fe uptake, the
358 significant reduction of the relative abundance of *kps* genes in low-Fe compared to Fe-replete
359 areas (t-test p-value <0.05 for all genes of the Syn CAG_008 and 009 except CK_00002157; Fig.
360 S14C) and their absence in CRD1 strains (Dataset 6) rather suggests that these capsules may be

361 too energy-consuming for some picocyanobacteria thriving in this peculiar niche, while they may
362 have a meaningful and previously overlooked role in their adaptation to low-P and low-N niches.

363 A number of CAGs were in contrast found to be enriched in populations dwelling in HNLC
364 environments, dominated by *Prochlorococcus* HLIIIA/HLIVA/LLIB and *Synechococcus*
365 CRD1A/EnvBA ESTUs (Fig. 2). For *Prochlorococcus*, this includes the abovementioned *natFGH*
366 (ProCAG_008) and *leudH/soxA* (ProCAG_009) CAGs, involved in amino acid metabolism (Fig.
367 S10), while a large proportion of the *Synechococcus* populations in these areas possess i) a large
368 CAG involved in glycine betaine synthesis and transport (SynCAG_010, Fig. S15A-B; (9, 61)),
369 almost absent from low-N areas, ii) a CAG encompassing a flavodoxin and a thioredoxin
370 reductase (SynCAG_011, Fig. S15C-D), mostly absent from low-P areas, as well as iii) the *nfeD-*
371 *floT1-floT2* CAG (SynCAG_012, Fig. S16A-B) involved in the production of lipid rafts, potentially
372 affecting cell shape and motility (9, 62). Both *Prochlorococcus* and *Synechococcus* thriving in
373 low-Fe waters also possess the TonB-dependent siderophore uptake operon (*fecDCAB-tonB-*
374 *exbBD*, Dataset 6). The latter gene cluster, which is found in a few picocyanobacterial genomes
375 and was previously shown to be anti-correlated with dissolved Fe concentration (24, 25, 57), is
376 indeed systematically present in a significant part of the *Prochlorococcus* and *Synechococcus*
377 population in low-Fe areas (ProCAG_015 and SynCAG_013-014, Fig. S17). However, it is also
378 present in a small fraction of the populations thriving in the Indian Ocean, consistent with its
379 occurrence in two *Prochlorococcus* HLII and one *Synechococcus* clade II genomes (Dataset 6).
380 The most striking result in this category is that the vast majority of *Prochlorococcus* cells thriving
381 in low-Fe regions possess a CAG encompassing the *ctaC2-D2-E2* operon, also found in 85% of all
382 *Synechococcus* reference genomes, including all CRD1 (Fig. 7; Dataset 6). This CAG encodes the
383 alternative respiratory terminal oxidase ARTO, a protein complex that has been suggested to be
384 part of a minimal respiratory chain in the cytoplasmic membrane, potentially providing an
385 additional electron sink under Fe-deprived conditions (63, 64). Furthermore, a *Synechocystis*
386 mutant in which the *ctaD2* and *ctaE2* genes had been inactivated was found to display markedly
387 impaired Fe reduction and uptake rates as compared to wild-type cells, suggesting that ARTO is
388 involved in the reduction of Fe(III) into Fe(II) prior to its transport through the plasma
389 membrane via the Fe(II) transporter FeoB (65). Thus, the presence of the ARTO system appears

390 to represent a major and previously unreported adaptation for *Prochlorococcus* populations
391 thriving in low-Fe areas.

392 Besides genes involved in nutrient acquisition and metabolism, several *Prochlorococcus* and
393 *Synechococcus* CAGs were found to be correlated with low-temperature waters. A closer
394 examination of *Prochlorococcus* CAGs however, shows that their occurrence is not directly
395 related to temperature adaptation but mainly explained by the prevalence at high latitude of
396 either i) the HLIA ESTU (Fig. 2A, C and Fig. 4), the *red* module encompassing most of the above-
397 mentioned CAGs involved in P-uptake and assimilation pathways, or ii) the LLIA ESTU, present in
398 surface waters at vertically-mixed stations, the *turquoise* module mainly gathering
399 *Prochlorococcus* LL-specific genes, such as Pro_CAG_017, involved in phycoerythrin-III
400 biosynthesis (*ppeA*, *cpeFTZY*, *unk13*) or ProCAG_018, encoding the two subunits of
401 exodeoxyribonuclease VII (XseA-B). As concerns *Synechococcus*, although a fairly high number of
402 CAGs were identified in the *tan* module associated with ESTUs IA and IVA-C (Fig. 2B, D and Fig.
403 S4), only very few are conserved in more than two reference strains and/or have a characterized
404 function (Dataset 6). Among these, at least one CAG is clearly related to adaptation to cold
405 waters, the orange carotenoid protein (OCP) operon (*ocp-crtW-frp*; SynCAG_016). Indeed, this
406 operon is involved in a photoprotective process (66) and was recently shown to provide cells
407 with the ability to deal with oxidative stress under cold temperatures (67). In agreement with
408 the latter study, our data shows that *Synechococcus* populations colonizing mixed waters at high
409 latitudes or in upwelling areas all possess the *ocp* CAG (Fig. S18), highlighting the importance of
410 this photoprotection system in *Synechococcus* populations colonizing cold and temperate areas.
411 *Synechococcus* populations thriving in cold waters also appear to be enriched in CAGs involved
412 in gene regulation. This includes transcriptional regulators involved in the regulation of the CA4-
413 A form of the type IV chromatic acclimation process (*fciA-B*; SynCAG_017), consistent with the
414 predominance of *Synechococcus* CA4-A cells in temperate or cold environments (68–70)(Dataset
415 6) as well as the *hidABC* operon (SynCAG_018), involved in the synthesis of a secondary
416 metabolite (hierridin C; (71)). Altogether, the fairly low number of ‘strong’ CAGs associated with
417 temperature supports the hypothesis that adaptation to cold temperature is not mediated by
418 evolution of gene content but rather of protein sequences (8, 9, 30, 72).

419 In conclusion, our analysis of *Prochlorococcus* and *Synechococcus* gene distributions at the
420 global scale using the deeply sequenced metagenomes collected along the *Tara* Oceans
421 expedition transect revealed that each community has a specific gene repertoire, with different
422 sets of accessory genes being highly correlated with distinct ESTUs and physicochemical
423 parameters. As previously suggested for *Prochlorococcus* (30), this strong correlation between
424 taxonomy and gene content strengthens the idea that, in both genera, genome evolution mainly
425 occurs by vertical transmission and selective gene retention, with a fairly low extent of lateral
426 gene transfer between clades. By combining information about gene synteny in 256 reference
427 genomes with the distribution and abundance of these genes in the field, we further managed
428 to delineate suites of adjacent genes likely involved in the same metabolic pathways that may
429 have a crucial role in adaptation to specific niches. These analyses confirmed previous
430 observations about the niche partitioning of individual genes and a few genomic regions
431 involved in nutrient uptake and assimilation (24, 25, 27, 30, 34, 36). Most importantly, this
432 network approach unveiled several novel genomic regions that could confer cells a fitness
433 benefit in particular niches and also highlighted that some previously detected individual genes
434 are part of larger genomic regions. It notably revealed the potential importance of the uptake
435 and assimilation of organic forms of limiting nutrients, which might either be directly used by
436 the cells, e.g. for the biosynthesis of proteins or DNA, or be degraded into inorganic N and/or P
437 forms. Consistently, many CAGs potentially involved in the uptake and assimilation of complex
438 compounds, such as guanidine, C₄N-containing compounds or pyrimidine were present in both
439 N- and P-depleted waters, and might constitute an advantage in areas of the world ocean co-
440 limited in these nutrients (26). In contrast, most of these CAGs were specifically absent from N
441 and/or P-rich, Fe-poor areas ((30); this study). Adaptation to Fe-limitation seemingly relies on
442 specific adaptation mechanisms including reduction of Fe³⁺ to Fe²⁺ using ARTO, Fe storage, Fe
443 scavenging using siderophores as well as reduction of the iron quota and of energy-consuming
444 adaptation mechanisms, such as polysaccharide capsule biosynthesis. Altogether, this study
445 provides unique insights into the functional basis of microbial niche partitioning and the
446 molecular bases of fitness in key members of the phytoplankton community. A future challenge
447 will clearly consist of biochemically characterizing the function of the different genes, including
448 many unknown, gathered in the above-mentioned CAGs (Datasets 5 and 6), which are
449 sometimes present only in a few or even a single strain but can occur in a large part or even the

450 whole *Prochlorococcus* and/or *Synechococcus* population *in situ*, and which likely all contribute
451 to the same complex and/or metabolic pathway.

452

453

454 **Materials and Methods**

455

456 **Tara Oceans dataset**

457 A total of 131 bacterial-size metagenomes (0.2-1.6 μm for stations TARA_004 to TARA_052 and
458 0.2-3 μm for TARA_056 to TARA_152), collected in surface from 83 stations along the *Tara*
459 Oceans expedition transect (73), were used in this study. Briefly, all metagenomes were
460 sequenced as Illumina overlapping paired reads of 100-108 bp and paired reads were merged
461 and trimmed based on quality, resulting in 100-215 bp fragments, as previously described (22).
462 All metagenomes and corresponding environmental parameters were retrieved from PANGAEA
463 (www.pangaea.de/) except for Fe and ammonium concentrations that were modeled and the Fe
464 limitation index Φ_{sat} that was calculated from satellite data, as previously described (22).

465

466 **Recruitment and taxonomic and functional assignment of metagenomic reads**

467 Metagenomic reads were first recruited against 256 reference genomes, including the 97
468 genomes available in the information system Cyanorak v2.1 (www.sb-roscoff.fr/cyanorak; (28))
469 as well as 84 additional WGS, 27 MAGs and 48 SAGs retrieved from Genbank (Dataset 9).
470 Recruitment was made using MMseqs2 Release 11-e1a1c (76) with maximum sensitivity
471 (mmseqs search -s 7.5) and limiting the results to one target sequence (mmseqs filterdb --
472 extract-lines 1). Using the same MMseqs2 options, the resulting reads were then mapped to an
473 extended database of 978 genomes, including all picocyanobacterial reference genomes
474 complemented with 722 outgroup cyanobacterial genomes downloaded from NCBI. Reads
475 mapping to outgroup sequences or having less than 90% of their sequence aligned were filtered
476 out and the remaining reads were taxonomically assigned to either *Prochlorococcus* or
477 *Synechococcus* according to their best hit. Reads were then functionally assigned to a cluster of
478 likely orthologous genes (CLOGs) from the information system Cyanorak v2.1 based on the
479 position of their MMseqs2 match on the genome, the coordinates of which correspond to a
480 particular gene in the database. More precisely, a read was functionally assigned to a gene if at
481 least 75% of its size was aligned to the reading frame of this gene and if the percentage identity

482 of the blast alignment was over 80%. Finally, read counts were aggregated by CLOG and
483 normalized by dividing read counts by $L+1$, where L represents the average gene length of the
484 CLOG and l the mean length of recruited reads. Only environmental samples that contained at
485 least 2,500 and 1,700 distinct CLOGs for *Synechococcus* and *Prochlorococcus*, respectively, were
486 kept, corresponding roughly to the average number of genes in a *Synechococcus* and a
487 *Prochlorococcus* HL genome, respectively. After this filtration step, a CLOG was kept if it showed
488 a gene-length normalized abundance higher than 1 (i.e., a gene coverage of 1) in at least 2 of the
489 selected environmental samples. Then, large-core genes, as previously defined (9), were
490 removed to keep only accessory genes. The resulting abundance profiles were used to perform
491 co-occurrence analyses by weighted genes correlation network analysis, as detailed below
492 (WGCNA, (74)).

493

494 **Station clustering and ESTU analyses**

495 In order to cluster stations displaying similar CLOG abundance patterns, the abundance of a
496 given CLOG in a sample was divided by the total CLOG abundance in this sample to obtain
497 relative abundance profiles per sample. Bray-Curtis similarities were calculated from these
498 profiles and used to cluster *Tara* Oceans stations with the Ward's minimum variance method
499 (75). The same normalization method was applied to picocyanobacterial ESTUs that were
500 defined based on the *petB* marker gene at each station using a similar approach as in Farrant *et*
501 *al.* (2016) but using a Ward's minimum variance method (75) to be consistent with the clustering
502 of CLOG profiles. In order to check whether the Bray-Curtis distances between stations based on
503 *petB* picocyanobacterial communities and based on gene content were significantly correlated,
504 a mantel test was performed between the distance matrices, as implemented in the R package
505 *vegan* v2.5 with 9,999 permutations (76).

506

507 **Gene co-occurrence network analysis**

508 A data-reduction approach based on WGCNA, as implemented in the R package WGCNA v1.51
509 (77), was used to build a co-occurrence network of CLOGs based on their relative abundance in
510 *Tara* Oceans stations and to delineate modules of CLOGs (i.e., subnetworks). The WGCNA
511 adjacency matrix was calculated in 'signed' mode (i.e., considering correlated and anti-
512 correlated CLOGs separately), by using the *Pearson* correlation between pairs of CLOGs (based

513 on their relative abundance in every sample) and raising it to the power 12, which allowed to
514 obtain a scale-free topology of the network. Modules were identified by setting the minimum
515 number of genes in each module to 100 and 50 for *Synechococcus* and *Prochlorococcus*,
516 respectively, and by forcing every gene to be included in a module. The *eigengene* of each
517 module (representative of the relative abundance of genes of a given module at each *Tara*
518 Oceans station) was then correlated to environmental parameters and to the relative
519 abundance of ESTUs. Finally, genes in each module with the highest correlation to the *eigengene*
520 (a measurement called 'membership'), were extracted in order to identify the most
521 representative genes of each module.

522

523 **Identification of differentially distributed clusters of adjacent genes (CAGs)**

524 Results on individual niche-related genes identified by WGCNA were then integrated with
525 knowledge on gene synteny in reference genomes (Datasets 7 and 8). For each WGCNA module,
526 we defined CAGs by searching adjacent genes of the module in the 256 reference genomes. In
527 order to be considered as belonging to the same CAG, two genes of the same module must be
528 less than 6 genes apart in 80% of the genomes in which these two genes are present. This led us
529 to identify clusters of adjacent genes in reference genomes, comprising genes displaying a
530 similar distribution pattern, called CAGs. A network of CAGs was then built for each WGCNA
531 module, taking into account the number of genomes in which these genes are adjacent (Figs. 4,
532 S3 and S4). An unweighted, undirected graph was drawn for each module according to the
533 Fruchterman-Reingold layout algorithm implemented in the R package igraph. This is a force-
534 directed algorithm, meaning that node layout is determined by the forces pulling nodes
535 together and pushing them apart. In other words, its purpose is to position the nodes of a graph
536 so that the edges of more or less equal length are gathered together and to avoid as many
537 crossing edges as possible.

538

539 **Data sharing plans:** All genomic and metagenomic data used in this study are publicly available

540

541 **Acknowledgments**

542

543 This work was supported by the French “Agence Nationale de la Recherche” Programs SAMOSA
544 (ANR-13-ADAP-0010), CINNAMON (ANR-17-CE02-0014-01), EFFICACY (ANR-19-CE02-0019) and

545 France Génomique (ANR-10-INBS-09) as well as the European Union program Assemble+
546 (Horizon 2020, under grant agreement number 287589). We acknowledge Christophe Six for his
547 help with cloning some of the *Synechococcus* strains used in this study and Francisco M.
548 Cornejo-Castillo for useful discussions. We also thank the support and commitment of the *Tara*
549 Oceans coordinators and consortium, Agnès b. and E. Bourgois, the Veolia Environment
550 Foundation, Région Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF
551 Foundation, FRB, the Prince Albert II de Monaco Foundation, the *Tara* schooner, and its captains
552 and crew. *Tara* Oceans would not exist without continuous support from 23 institutes
553 (<http://oceans.taraexpeditions.org>).

554

555

556 References

557

558 1. H. Tettelin, *et al.*, Genome analysis of multiple pathogenic isolates of *Streptococcus*
559 *agalactiae*: Implications for the microbial “pan-genome.” *Proc Natl Acad Sci USA* **102**,
560 13950–13955 (2005).

561 2. M. López-Pérez, F. Rodríguez-Valera, Pangenome evolution in the marine bacterium
562 *Alteromonas*. *Genome Biol Evol* **8**, 1556–1570 (2016).

563 3. T. D. Read, *et al.*, The genome sequence of *Bacillus anthracis* Ames and comparison to
564 closely related bacteria. *Nature* **423**, 81–86 (2003).

565 4. C. Zhu, T. O. Delmont, T. M. Vogel, Y. Bromberg, Functional basis of microorganism
566 classification. *PLOS Comput Biol* **11**, e1004472 (2015).

567 5. M. L. Reno, N. L. Held, C. J. Fields, P. V. Burke, R. J. Whitaker, Biogeography of the
568 *Sulfolobus islandicus* pan-genome. *Proc Natl Acad Sci USA* **106**, 8605–8610 (2009).

569 6. S. S. Porter, P. L. Chang, C. A. Conow, J. P. Dunham, M. L. Friesen, Association mapping
570 reveals novel serpentine adaptation gene clusters in a population of symbiotic
571 *Mesorhizobium*. *ISME J* **11**, 248–262 (2017).

572 7. S. Kellner, *et al.*, Genome size evolution in the Archaea. *Emerg Top Life Sci* **2**, 595–605
573 (2018).

574 8. G. C. Kettler, *et al.*, Patterns and implications of gene gain and loss in the evolution of
575 *Prochlorococcus*. *PLOS Genet* **3**, e231 (2007).

576 9. H. Doré, *et al.*, Evolutionary mechanisms of long-term genome diversification associated
577 with niche partitioning in marine picocyanobacteria. *Front Microbiol* **11** (2020).

578 10. N. Kashtan, *et al.*, Single-cell genomics reveals hundreds of coexisting subpopulations in
579 wild *Prochlorococcus*. *Science* **344**, 416–420 (2014).

580 11. P. B. Pearman, A. Guisan, O. Broennimann, C. F. Randin, Niche dynamics in space and

- 581 time. *Trends Ecol Evol* **23**, 149–158 (2008).
- 582 12. T. O. Delmont, A. M. Eren, Linking pangenomes and metagenomes: the *Prochlorococcus*
583 metapangenome. *PeerJ* **6**, e4320 (2018).
- 584 13. J.-H. Hehemann, *et al.*, Adaptive radiation by waves of gene transfer leads to fine-scale
585 resource partitioning in marine microbes. *Nat Commun* **7**, 12860 (2016).
- 586 14. H. Koch, *et al.*, Genomic, metabolic and phenotypic variability shapes ecological
587 differentiation and intraspecies interactions of *Alteromonas macleodii*. *Sci Rep* **10** (2020).
- 588 15. J. P. Engelberts, *et al.*, Characterization of a sponge microbiome using an integrative
589 genome-centric approach. *ISME J* **14**, 1100–1110 (2020).
- 590 16. B. J. Tully, E. D. Graham, J. F. Heidelberg, The reconstruction of 2,631 draft metagenome-
591 assembled genomes from the global oceans. *Sci Data* **5**, 170203 (2018).
- 592 17. B. L. Hurwitz, A. H. Westveld, J. R. Brum, M. B. Sullivan, Modeling ecological drivers in
593 marine viral communities using comparative metagenomics and network analyses. *Proc Natl*
594 *Acad Sci USA* **111**, 10714–10719 (2014).
- 595 18. A. Meziti, *et al.*, Quantifying the changes in genetic diversity within sequence-discrete
596 bacterial populations across a spatial and temporal riverine gradient. *ISME J* **13**, 767–779
597 (2019).
- 598 19. I. Raimundo, *et al.*, Functional metagenomics reveals differential chitin degradation and
599 utilization features across free-living and host-associated marine microbiomes. *Microbiome*
600 **9**, 43 (2021).
- 601 20. P. Flombaum, *et al.*, Present and future global distributions of the marine Cyanobacteria
602 *Prochlorococcus* and *Synechococcus*. *Proc Natl Acad Sci USA* **110**, 9824–9 (2013).
- 603 21. N. Visintini, A. C. Martiny, P. Flombaum, *Prochlorococcus*, *Synechococcus*, and
604 picoeukaryotic phytoplankton abundances in the global ocean. *Limnol Oceanogr* **6**, 207–215
605 (2021).
- 606 22. G. K. Farrant, *et al.*, Delineating ecologically significant taxonomic units from global
607 patterns of marine picocyanobacteria. *Proc Natl Acad Sci USA* **113**, E3365–E3374 (2016).
- 608 23. A. G. Kent, *et al.*, Parallel phylogeography of *Prochlorococcus* and *Synechococcus*. *ISME J*
609 **13**, 430–441 (2019).
- 610 24. N. A. Ahlgren, B. S. Belisle, M. D. Lee, Genomic mosaicism underlies the adaptation of
611 marine *Synechococcus* ecotypes to distinct oceanic iron niches. *Environ Microbiol* **22**, 1801–
612 1815 (2020).
- 613 25. C. A. Garcia, *et al.*, Linking regional shifts in microbial genome adaptation with surface
614 ocean biogeochemistry. *Phil Trans Roy Soc B Biol Sci* **375**, 20190254 (2020).
- 615 26. L. J. Ustick, *et al.*, Metagenomic analysis reveals global-scale patterns of ocean nutrient
616 limitation. *Science* **372**, 287–291 (2021).

- 617 27. A. C. Martiny, M. L. Coleman, S. W. Chisholm, Phosphate acquisition genes in
618 *Prochlorococcus* ecotypes: Evidence for genome-wide adaptation. *Proc Natl Acad Sci USA*
619 **103**, 12552–12557 (2006).
- 620 28. A. C. Martiny, Y. Huang, W. Li, Occurrence of phosphate acquisition genes in
621 *Prochlorococcus* cells from different ocean regions. *Environmental Microbiology* **11**, 1340–
622 1347 (2009).
- 623 29. L. Garczarek, *et al.*, Cyanorak v2.1: a scalable information system dedicated to the
624 visualization and expert curation of marine and brackish picocyanobacteria genomes. *Nucl*
625 *Acids Res* **49**, D667–D676 (2021).
- 626 30. A. G. Kent, C. L. Dupont, S. Yooseph, A. C. Martiny, Global biogeography of
627 *Prochlorococcus* genome diversity in the surface ocean. *The ISME Journal* **10**, 1856–1865
628 (2016).
- 629 31. Q. Song, A. L. Gordon, M. Visbeck, Spreading of the Indonesian throughflow in the Indian
630 Ocean. *J Phys Oceanogr* **34**, 772–792 (2004).
- 631 32. N. J. West, P. Lebaron, P. G. Strutton, M. T. Suzuki, A novel clade of *Prochlorococcus*
632 found in high nutrient low chlorophyll waters in the South and Equatorial Pacific Ocean.
633 *ISME J* **5**, 933–944 (2011).
- 634 33. D. B. Rusch, A. C. Martiny, C. L. Dupont, A. L. Halpern, J. C. Venter, Characterization of
635 *Prochlorococcus* clades from iron-depleted oceanic regions. *Proc Natl Acad Sci USA* **107**,
636 16184–16189 (2010).
- 637 34. A. C. Martiny, S. Kathuria, P. M. Berube, Widespread metabolic potential for nitrite and
638 nitrate assimilation among *Prochlorococcus* ecotypes. *Proc Natl Acad Sci USA* **106**, 10787–
639 10792 (2009).
- 640 35. P. M. Berube, A. Rasmussen, R. Braakman, R. Stepanauskas, S. W. Chisholm, Emergence
641 of trait variability through the lens of nitrogen assimilation in *Prochlorococcus*. *eLife* **8**,
642 e41043–e41043 (2019).
- 643 36. P. M. Berube, *et al.*, Physiology and evolution of nitrate acquisition in *Prochlorococcus*.
644 *ISME J* (2015) <https://doi.org/10.1038/ismej.2014.211>.
- 645 37. M. Burnat, B. Li, S. H. Kim, A. J. Michael, E. Flores, Homospermidine biosynthesis in the
646 cyanobacterium *Anabaena* requires a deoxyhypusine synthase homologue and is essential for
647 normal diazotrophic growth. *Mol Microbiol* **109**, 763–780 (2018).
- 648 38. B. Wang, *et al.*, A guanidine-degrading enzyme controls genomic stability of ethylene-
649 producing cyanobacteria. *Nat Commun* **12**, 5150 (2021).
- 650 39. J. W. Nelson, R. M. Atilho, M. E. Sherlock, R. B. Stockbridge, R. R. Breaker, Metabolism
651 of free guanidine in Bacteria is regulated by a widespread riboswitch class. *Mol Cell* **65**,
652 220–230 (2017).
- 653 40. P. S. Novichkov, *et al.*, RegPrecise 3.0 – A resource for genome-scale exploration of
654 transcriptional regulation in bacteria. *BMC Genomics* **14**, 745 (2013).

- 655 41. N. A. Kamennaya, A. F. Post, Characterization of cyanate metabolism in marine
656 *Synechococcus* and *Prochlorococcus* spp. *Appl Environ Microbiol* **77**, 291–301 (2011).
- 657 42. N. A. Kamennaya, A. F. Post, Distribution and expression of the cyanate acquisition
658 potential among cyanobacterial populations in oligotrophic marine waters. *Limnol Oceanogr*
659 **58**, 1959–1971 (2013).
- 660 43. L. B. Jones, P. Ghosh, J.-H. Lee, C.-N. Chou, D. A. Y. 2018 Kunz, Linkage of the Nit1C
661 gene cluster to bacterial cyanide assimilation as a nitrogen source. *Microbiol* **164**, 956–968
662 (2018).
- 663 44. J. K. Saunders, G. Rocap, Genomic potential for arsenic efflux and methylation varies
664 among global *Prochlorococcus* populations. *ISME J* **10**, 197–209 (2016).
- 665 45. G. F. Riedel, Influence of salinity and sulfate on the toxicity of chromium(vi) to the estuarine
666 diatom *Thalassiosira Pseudonana*. *Journal of Phycology* **20**, 496–500 (1984).
- 667 46. F. Pablo, J. L. Stauber, R. T. Buckney, Toxicity of cyanide and cyanide complexes to the
668 marine diatom *Nitzschia closterium*. *Water Res* **31**, 2435–2442 (1997).
- 669 47. R. Pernil, S. Picossi, V. Mariscal, A. Herrero, E. Flores, ABC-type amino acid uptake
670 transporters Bgt and N-II of *Anabaena* sp. strain PCC 7120 share an ATPase subunit and are
671 expressed in vegetative cells and heterocysts. *Mol Microbiol* **67**, 1067–1080 (2008).
- 672 48. M. L. Coleman, *et al.*, Genomic islands and the ecology and evolution of *Prochlorococcus*.
673 *Science* **311**, 1768–1770 (2006).
- 674 49. C. M. Moore, *et al.*, Processes and patterns of oceanic nutrient limitation. *Nature Geosci* **6**,
675 701–710 (2013).
- 676 50. S. G. Tetu, *et al.*, Microarray analysis of phosphate regulation in the marine cyanobacterium
677 *Synechococcus* sp. WH8102. *ISME J* **3**, 835–849 (2009).
- 678 51. L. L. Clark, E. D. Ingall, R. Benner, Marine phosphorus is selectively remineralized. *Nature*
679 **393**, 426–426 (1998).
- 680 52. R. Feingersch, *et al.*, Potential for phosphite and phosphonate utilization by
681 *Prochlorococcus*. *ISME J* **6**, 827–834 (2012).
- 682 53. A. Martinez, G. W. Tyson, E. F. Delong, Widespread known and novel phosphonate
683 utilization pathways in marine bacteria revealed by functional screening and metagenomic
684 analyses. *Environ Microbiol* **12**, 222–238 (2010).
- 685 54. O. A. Sosa, J. R. Casey, D. M. Karl, Methylphosphonate oxidation in *Prochlorococcus* strain
686 MIT9301 supports phosphate acquisition, formate excretion, and carbon assimilation into
687 purines. *Appl Environ Microbiol* **85**, e00289-19 (2019).
- 688 55. A. Martínez, M. S. Osburne, A. K. Sharma, E. F. DeLong, S. W. Chisholm, Phosphite
689 utilization by the marine picocyanobacterium *Prochlorococcus* MIT9301. *Environ Microbiol*
690 **14**, 1363–1377 (2012).
- 691 56. F. R. McSorley, *et al.*, PhnY and PhnZ comprise a new oxidative pathway for enzymatic

- 692 cleavage of a carbon–phosphorus bond. *J Am Chem Soc* **134**, 8364–8367 (2012).
- 693 57. R. R. Malmstrom, *et al.*, Ecology of uncultured *Prochlorococcus* clades revealed through
694 single-cell genomics and biogeographic analysis. *ISME J* **7**, 184–198 (2013).
- 695 58. J. W. Cooley, W. F. J. Vermaas, Succinate dehydrogenase and other respiratory pathways in
696 thylakoid membranes of *Synechocystis* sp. strain PCC 6803: capacity comparisons and
697 physiological function. *J Bacteriol* (2001) (January 27, 2022).
- 698 59. C. Whitfield, Biosynthesis and assembly of capsular polysaccharides in *Escherichia coli*.
699 *Annu Rev Biochem* **75**, 39–68 (2006).
- 700 60. A. Buffet, E. P. C. Rocha, O. Rendueles, Nutrient conditions are primary drivers of bacterial
701 capsule maintenance in *Klebsiella*. *Proc Roy Soc B Biol Sci* **288**, 20202876 (2021).
- 702 61. D. J. Scanlan, *et al.*, Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol*
703 *Rev* **73**, 249–299 (2009).
- 704 62. F. Dempwolff, H. M. Wischhusen, M. Specht, P. L. Graumann, The deletion of bacterial
705 dynamin and flotillin genes results in pleiotrophic effects on cell division, cell growth and in
706 cell shape maintenance. *BMC Microbiol* **12**, 298 (2012).
- 707 63. D. J. Lea-Smith, *et al.*, Thylakoid terminal oxidases are essential for the cyanobacterium
708 *Synechocystis* sp. PCC 6803 to survive rapidly changing light intensities. *Plant Physiol* **162**,
709 484–495 (2013).
- 710 64. D. J. Lea-Smith, P. Bombelli, R. Vasudevan, C. J. Howe, Photosynthetic, respiratory and
711 extracellular electron transport pathways in cyanobacteria. *Biochim Biophys Acta Bioenerget*
712 **1857**, 247–255 (2016).
- 713 65. C. Kranzler, *et al.*, Coordinated transporter activity shapes high-affinity iron acquisition in
714 cyanobacteria. *ISME J* **8**, 409–417 (2014).
- 715 66. C. Boulay, A. Wilson, S. D’Haene, D. Kirilovsky, Identification of a protein required for
716 recovery of full antenna capacity in OCP-related photoprotective mechanism in
717 cyanobacteria. *Proc Natl Acad Sci USA* **107**, 11620–11625 (2010).
- 718 67. C. Six, M. Ratin, D. Marie, E. Corre, Marine *Synechococcus* picocyanobacteria: Light
719 utilization across latitudes. *Proc Natl Acad Sci USA* **118** (2021).
- 720 68. X. Xia, *et al.*, Phylogeography and pigment type diversity of *Synechococcus* cyanobacteria
721 in surface waters of the northwestern Pacific Ocean. *Environ Microbiol* **19**, 142–158 (2017).
- 722 69. T. Grébert, *et al.*, Light color acclimation is a key process in the global ocean distribution of
723 *Synechococcus* cyanobacteria. *Proc Natl Acad Sci USA* **115**, E2010–E2019 (2018).
- 724 70. J. E. Sanfilippo, *et al.*, Self-regulating genomic island encoding tandem regulators confers
725 chromatic acclimation to marine *Synechococcus*. *Proc Natl Acad Sci USA* **113**, 6077–6082
726 (2016).
- 727 71. M. Costa, *et al.*, Structure of Hierridin C, synthesis of hierridins B and C, and evidence for
728 prevalent alkylresorcinol biosynthesis in picocyanobacteria. *J. Nat. Prod.* **82**, 393–402

- 729 (2019).
- 730 72. A. A. Larkin, A. C. Martiny, Microdiversity shapes the traits, niche space, and biogeography
731 of microbial taxa: The ecological function of microdiversity. *Environ Microbiol Rep* **9**, 55–
732 70 (2017).
- 733 73. S. Sunagawa, *et al.*, Structure and function of the global ocean microbiome. *Science* **348**,
734 1261359–1261359 (2015).
- 735 74. B. Zhang, S. Horvath, A general framework for weighted gene co-expression network
736 analysis. *Stat Appl Genet Mol Biol* **4**, Article17 (2005).
- 737 75. B. Szmrecsanyi, *Grammatical Variation in British English Dialects: A Study in Corpus-*
738 *Based Dialectometry* (Cambridge University Press, 2012)
739 <https://doi.org/10.1017/CBO9780511763380>.
- 740 76. Oksanen, J., *et al.*, *Vegan: Community Ecology Package. R package Version 2.4-3* (2017).
- 741 77. P. Langfelder, S. Horvath, WGCNA: an R package for weighted correlation network
742 analysis. *BMC Bioinfo* **9**, 559 (2008).
- 743
- 744

745 **Figure Legends**

746

747 **Figure 1. Comparison of clustering based on relative abundance profiles of ecologically**

748 **significant taxonomic units (ESTUs) and of flexible genes for both picocyanobacteria. A.**

749 *Prochlorococcus*. B. *Synechococcus*. Leaves of the trees correspond to stations along the Tara

750 Oceans transect that are colored according to the code shown at the bottom of the trees,

751 corresponding to ESTU assemblages as determined by Farrant et al. (2016) by clustering

752 stations exhibiting similar ESTU relative abundance profiles shown here on the right of each tree.

753 ESTUs were colored according to the palette below each panel. Dotted lines in dendrograms

754 indicate discrepancies between tree topologies. Accessory genes correspond to all genes except

755 those defined as large-core genes in a previous study (9). Of note, due to a slightly different

756 clustering method (cf. materials and methods), assemblage 7 (dark grey stations in 1B), which

757 was discriminated from assemblage 6 in the Farrant et al. (2016) now clusters with this

758 assemblage. Abbreviations: IO, Indian Ocean; MS, Mediterranean Sea; NAO, North Atlantic

759 Ocean; NPO, North Pacific Ocean; RS, Red Sea; SAO, South Atlantic Ocean; SO, Southern

760 Ocean.

761

762 **Figure 2. Correlation of picocyanobacterial module eigengenes to physico-chemical**
763 **parameters and ESTU abundance.** A, B. Correlation of module eigengenes to physico-chemical
764 parameters for *Prochlorococcus* (A) and *Synechococcus* (B). C, D. Correlation of module
765 eigengenes to relative abundance profiles of ESTUs *sensu* (Farrant et al., 2016). Pearson (A, B)
766 and Spearman (B, D) correlation coefficient (R^2) is indicated by the color scale. Each module is
767 identified by a specific color and the number between brackets specifies the number of genes in
768 each module. The *eigengene* is representative of the relative abundance of genes of a given
769 module at each *Tara* Oceans station. Non-significant correlations (Student asymptotic p-value >
770 0.01) are marked by a cross. Φ sat: index of iron limitation derived from satellite data. PAR30:
771 satellite-derived photosynthetically available radiation at the surface, averaged on 30 days. DCM:
772 depth of the deep chlorophyll maximum.

773

774 **Figure 3. Violin plots highlighting the most representative genes of each *Prochlorococcus***
775 **module.** For each module, each gene is represented as a dot positioned according to its
776 correlation with the eigengene for each module, the most representative genes being localized on
777 top of each violin plot. Genes mentioned in the text and/or in Dataset 6 have been colored
778 according to the color of the corresponding module, indicated by a colored bar above each
779 module. The text above violin plots indicates the most significant environmental parameter(s)
780 and/or ESTU(s) for each module, as derived from Fig. 2.

781

782 **Figure 4. Delineation of *Prochlorococcus* CAGs, defined as a set of genes that are both**
783 **adjacent in reference genomes and share a similar *in situ* distribution.** Nodes correspond to
784 individual genes with their gene name (or significant numbers of the CK number, e.g. 1234 for
785 CK_00001234) and are colored according to their WGCNA module. A link between two nodes
786 indicates that these two genes are less than 5 genes apart in at least one genome. The bottom
787 insert shows the most significant environmental parameter(s) and/or ESTU(s) for each module,
788 as derived from Fig. 2.

789

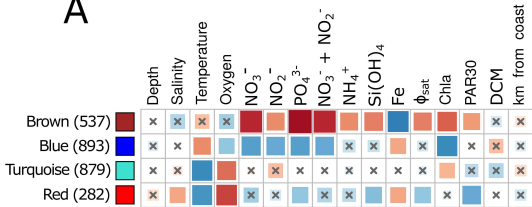
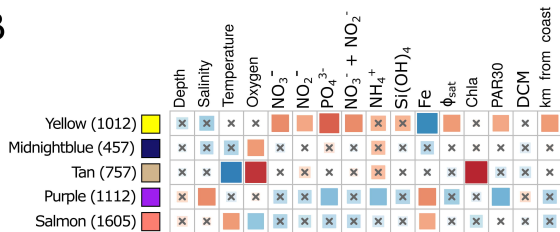
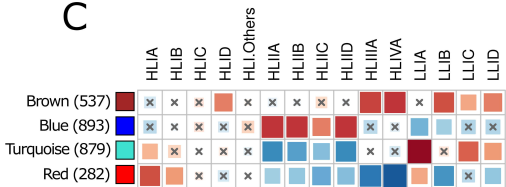
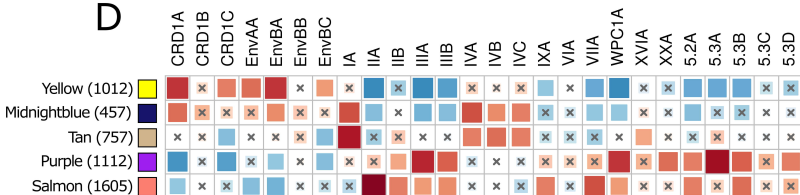
790 **Figure 5. Global distribution map of CAG involved in nitriles or cyanides transport and**
791 **assimilation.** (A) *Prochlorococcus* (ProCAG_006) and (B) *Synechococcus* SynCAG_002. (C)
792 Genomic region in *Prochlorococcus marinus* MIT9301. The size of the circle is proportional to
793 relative abundance of *Prochlorococcus* as estimated based on the single-copy core gene *petB*
794 gene and this gene was also used to estimate the relative abundance of other genes in the
795 population.

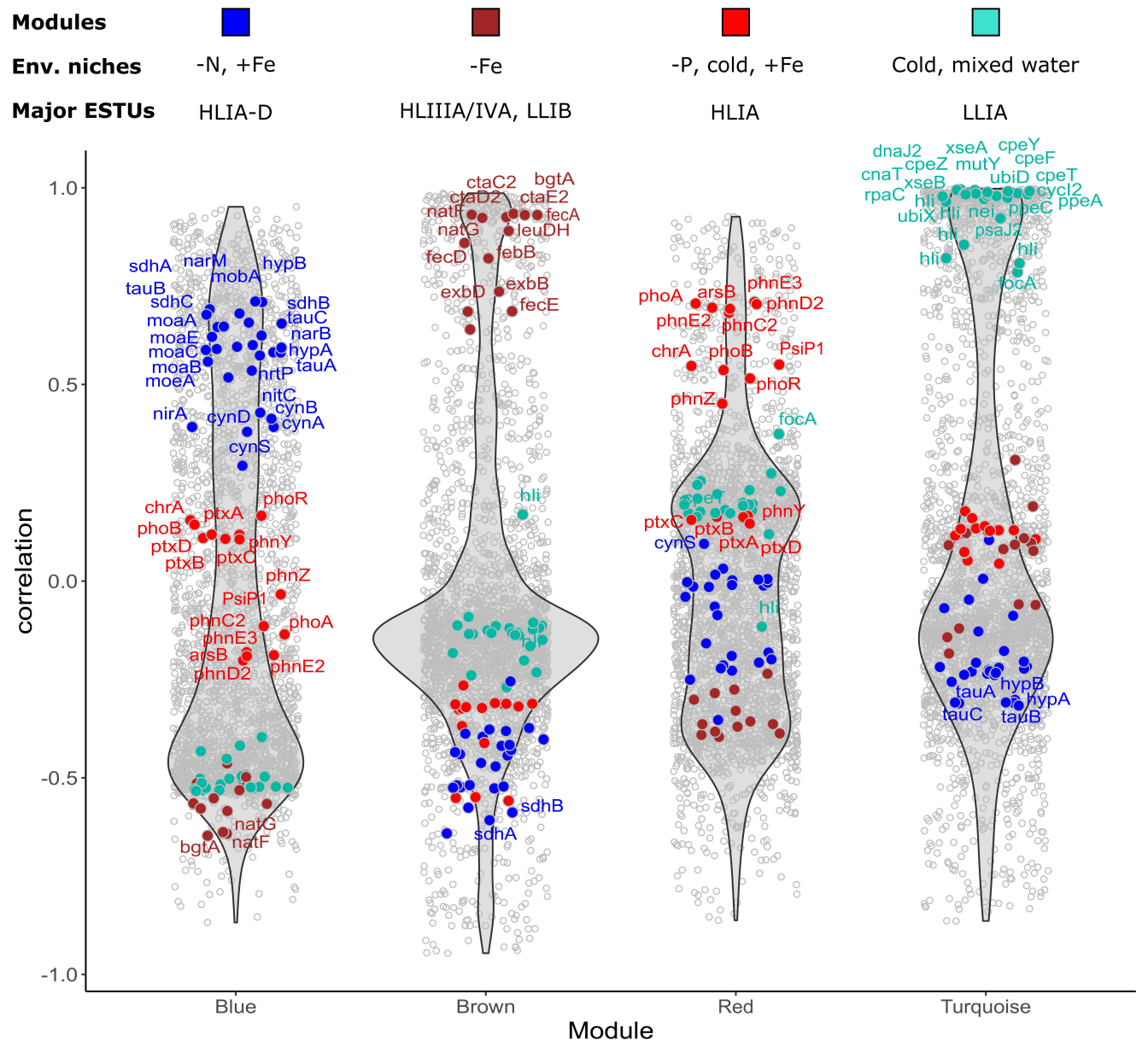
796

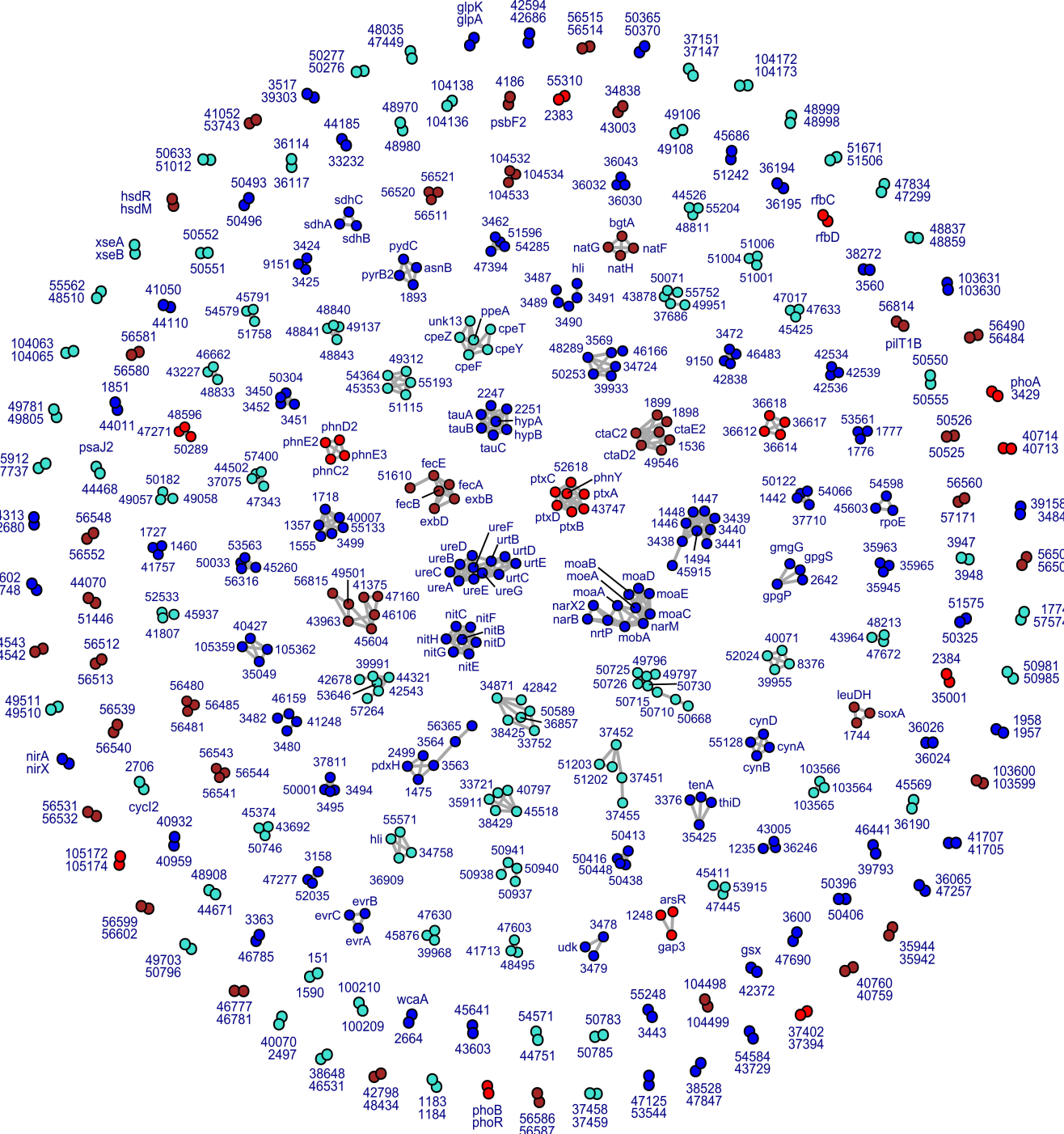
797 **Figure 6. Global distribution map of CAGs putatively involved in phosphonate and**
798 **phosphite transport and assimilation.** *Prochlorococcus* (A) ProCAG_012 putatively involved in
799 phosphonate transport, (B) ProCAG_013, involved in phosphonate/phosphite uptake and
800 assimilation and phosphonate C-P bond cleavage, (C) The genomic region encompassing both
801 *phnC2-D2-E2-E3* and *ptxABDC-phnYZ* specific to *P. marinus* RS01. The size of the circle is
802 proportional to relative abundance of *Prochlorococcus* as estimated based on the single-copy
803 core gene *petB* and this gene was also used to estimate the relative abundance of other genes in
804 the population.

805

806 **Figure 7. Global distribution map of the *Prochlorococcus* CAGs involved in the**
807 **biosynthesis of an alternative respiratory terminal oxidase (ARTO).** (A) *Prochlorococcus*
808 ProCAG_016, (B) *Synechococcus* SynCAG_015. The size of the circle is proportional to relative
809 abundance of *Prochlorococcus* as estimated based on the single-copy core gene *petB* and this
810 gene was also used to estimate the relative abundance of other genes in the population.

A**B****C****D**





Modules



Env. niches

-N, +Fe

-Fe

-P, cold, +Fe

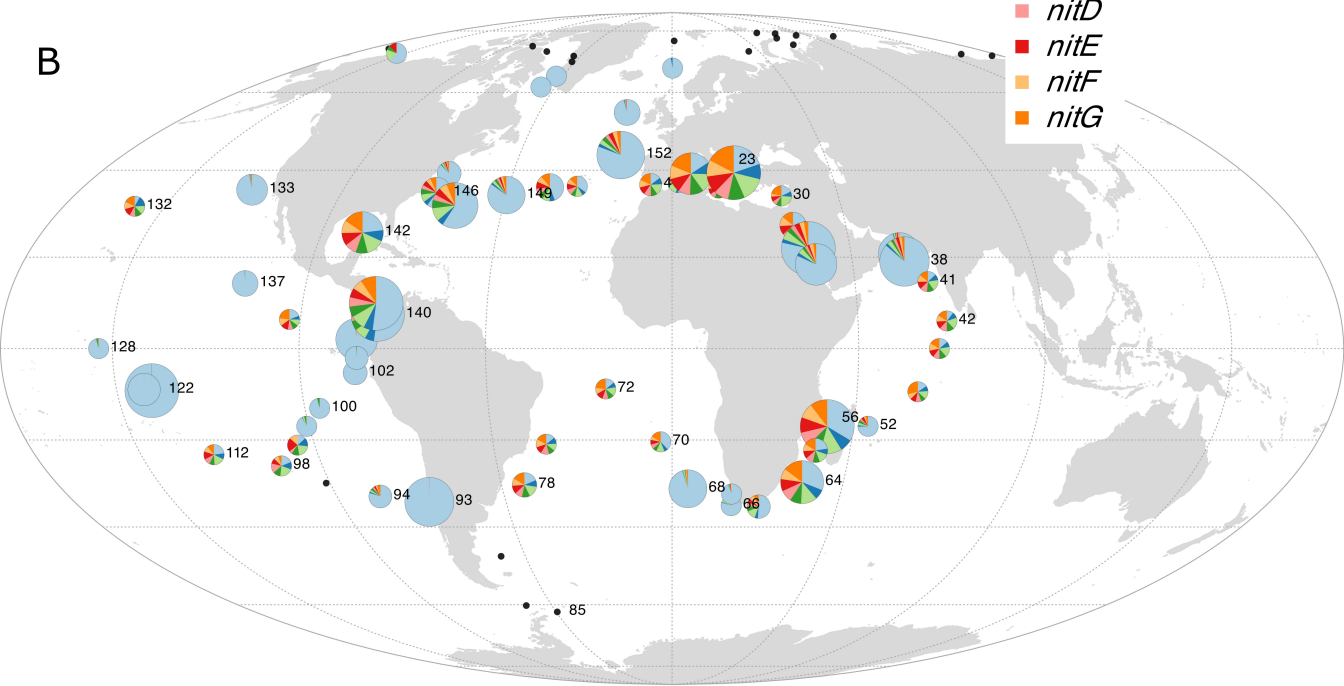
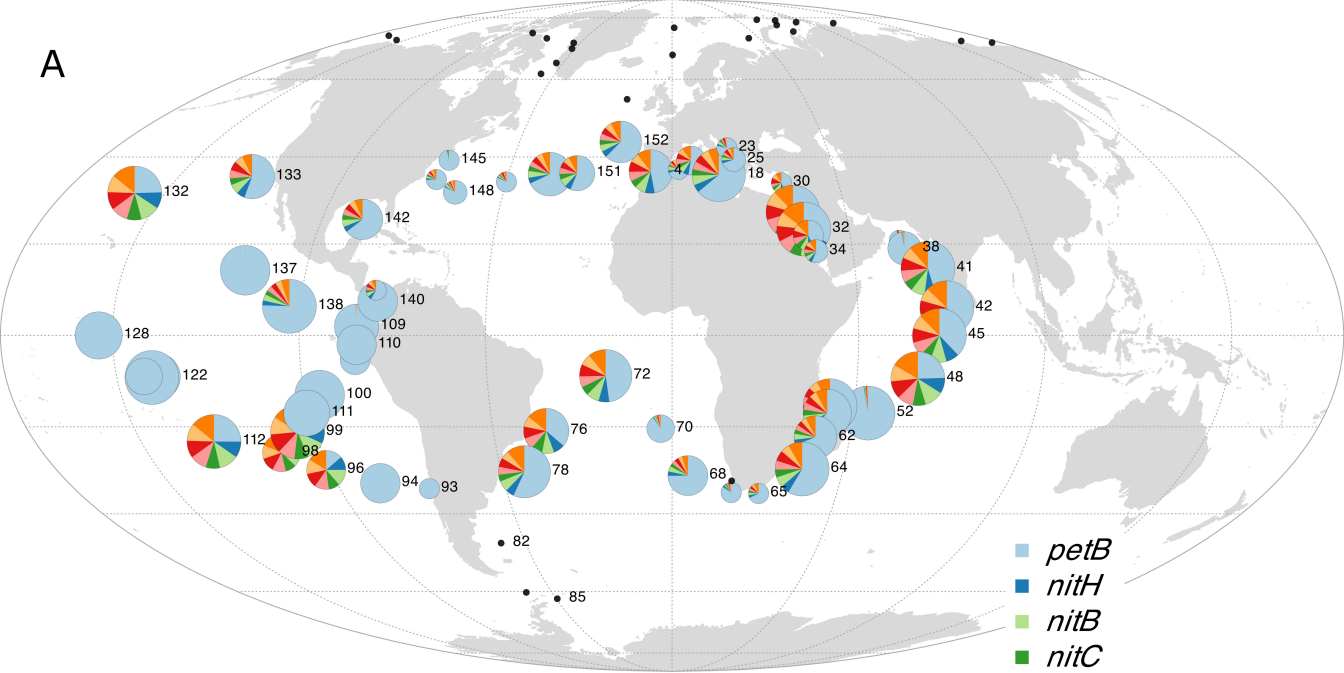
Cold, mixed water

Major ESTUs

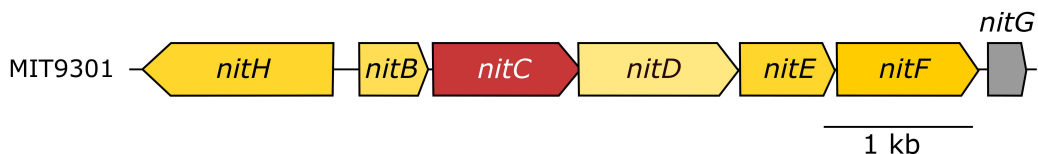
HLIIIA/IVA, LLIB

HLIA

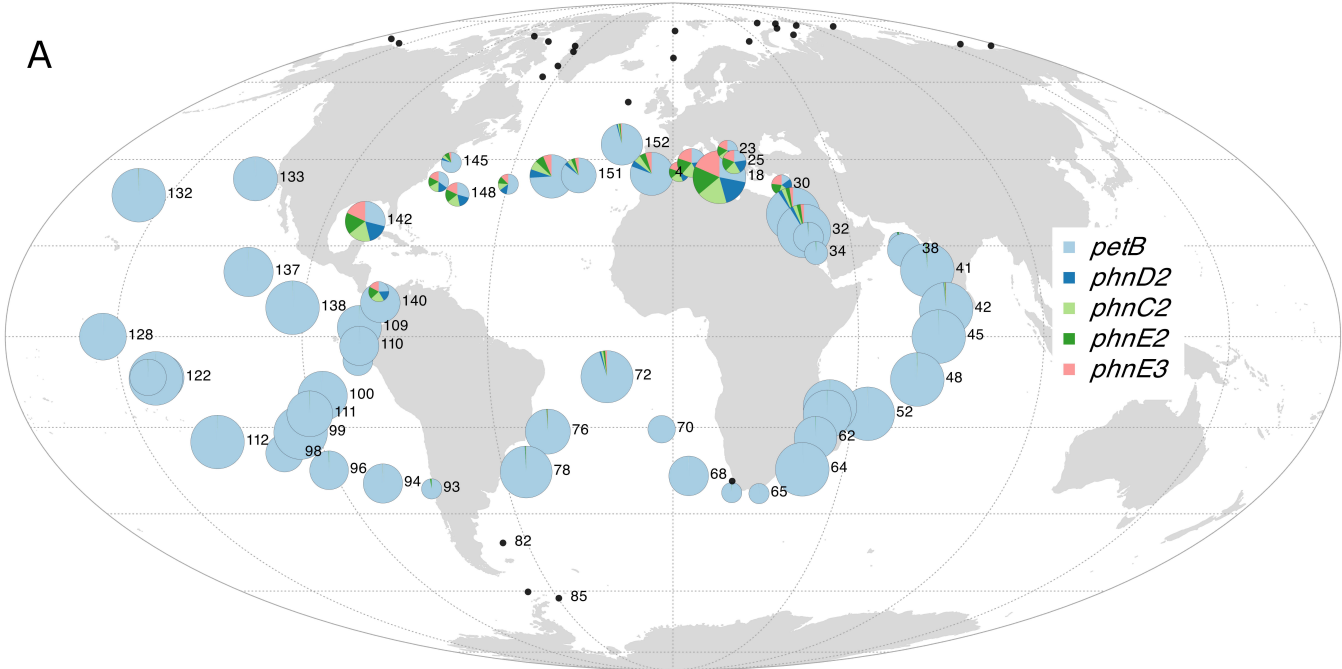
LLIA



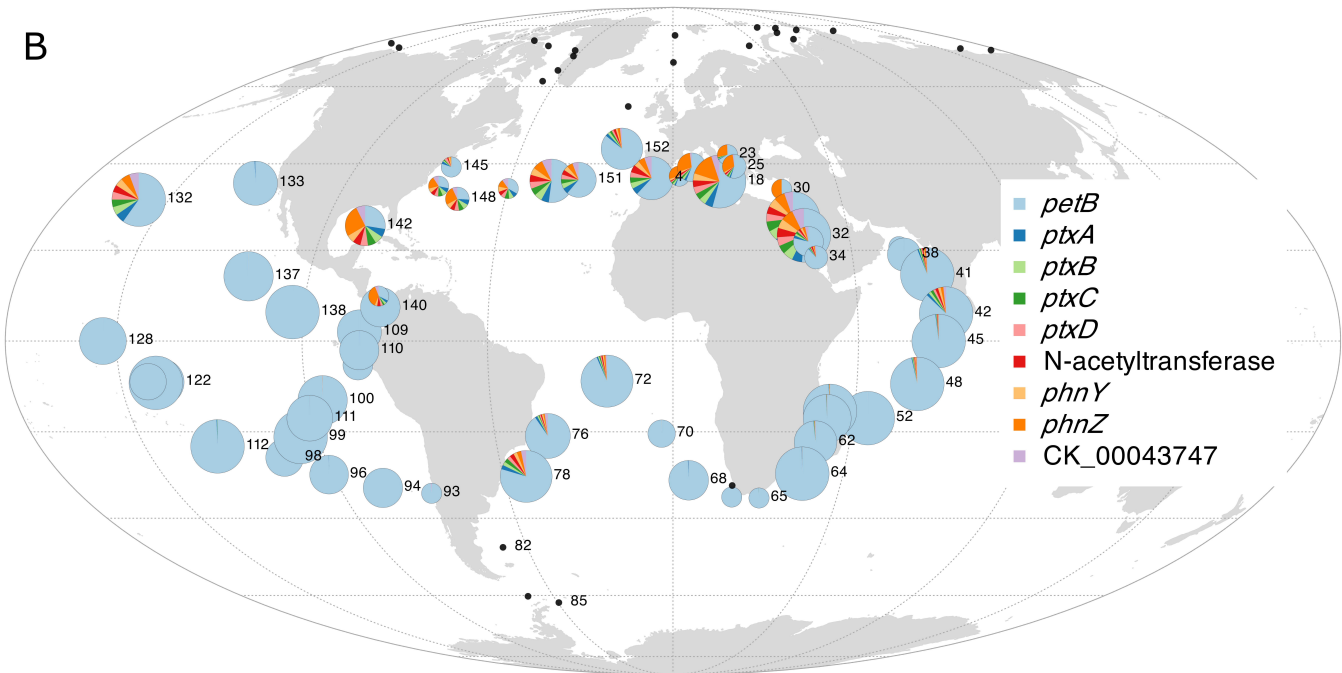
C



A



B



C

