Classification: BIOLOGICAL SCIENCES

Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria

Gregory K. Farrant[1,2†], Hugo Doré[1†], Francisco M. Cornejo-Castillo[3], Frédéric Partensky[1], Morgane Ratin[1], Martin Ostrowski[4], Frances D. Pitt[5], Patrick Wincker[6], David J. Scanlan[5], Daniele Iudicone[7], Silvia G. Acinas[3] and Laurence Garczarek[1]

[1]Sorbonne Universités, UPMC Université Paris 06, CNRS, UMR 7144, Station Biologique, CS 90074, Roscoff, France. [2]Present address: Matis, Vinlandsleid 12, 113 Reykjavik, Iceland. [3]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM), CSIC, Passeig Marítim de la Barceloneta, 37–49, Barcelona ES-08003, Spain. [4]Macquarie University, Department of Chemistry and Biomolecular Sciences, Sydney, Australia; [5]University of Warwick, School of Life Sciences, Gibbet Hill Road, Coventry CV4 7AL, UK; [6]Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, Genoscope, 2 Rue Gaston Crémieux, 91057 Evry, France. [7]Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy.

†These authors contributed equally to this work

Correspondence to: laurence.garczarek@sb-roscoff.fr

Keywords: biodiversity, next-generation sequencing, *Tara* Oceans, cyanobacteria, *Prochlorococcus*, *Synechococcus*, metagenomics, $_{\text{mi}}$Tags, molecular ecology.

Submitted to: Proceedings of the National Academy of Sciences of the USA

\abstract

Prochlorococcus and *Synechococcus* are the two most abundant and widespread phytoplankton in the global ocean. In order to better understand the factors controlling their biogeography, a reference database of the high resolution taxonomic marker *petB*, encoding cytochrome b_6, was used to recruit reads out of 109 metagenomes from the *Tara* Oceans expedition. An unsuspected novel genetic diversity was unveiled within both genera, even for the most abundant and well-characterized clades, and 136 divergent *petB* sequences were successfully assembled from metagenomic reads, significantly enriching the reference database. We then defined Ecologically Significant Taxonomic Units (ESTUs), i.e. organisms belonging to the same clade and occupying a given oceanic niche. Three major ESTU assemblages were identified along the cruise transect for *Prochlorococcus* and eight for *Synechococcus*. The picocyanobacterial population structure of the Pacific Ocean proved to be quite different from other oceanic areas. While *Prochlorococcus* HLIIIA and HLIVA ESTUs co-dominated in iron-depleted areas, CRD1 and the yet-to-be cultured EnvB were the prevalent clades in Pacific *Synechococcus* populations, with three different CRD1 and EnvB ESTUs occupying distinct ecological niches with regard to iron availability and temperature. Sharp community shifts were also observed over short geographic distances, e.g. around the Marquesas Islands or between southern Indian and Atlantic Oceans, pointing to a tight correlation between ESTU assemblages and specific physico-chemical parameters. Together, this study demonstrates that there is ecologically meaningful fine-scale diversity within currently defined picocyanobacterial clades, bringing novel insights into the ecology, diversity and biology of the two most abundant phototrophs on Earth.

Significance

Metagenomics has become an accessible approach to study complex microbial communities thanks to the advent of high-throughput sequencing technologies. However, molecular ecology studies often face interpretation issues, notably due to the lack of reliable reference databases for assigning reads to the correct taxa and use of fixed cut-offs to delineate taxonomic groups. Here, we considerably refined the phylogeography of marine picocyanobacteria, responsible for about 25% of global marine productivity, by recruiting reads targeting a high resolution marker from *Tara* Oceans metagenomes. By clustering lineages based on their distribution patterns, we showed that there is significant diversity at a finer resolution than the currently defined 'ecotypes', which is tightly controlled by environmental cues.

\body

Introduction

The ubiquitous marine picocyanobacteria *Prochlorococcus* and *Synechococcus* are major contributors to global chlorophyll biomass, together accounting for a quarter of global carbon fixation in marine ecosystems, a contribution predicted to further increase in the context of global change (1-3). Thus, determining how environmental conditions control their global distribution patterns, particularly at a fine taxonomic resolution (i.e., sufficient to identify lineages with distinct traits), is critical for understanding how these organisms populate the oceans, and in turn contribute to global carbon cycling. The availability of numerous strains in culture and sequenced genomes make picocyanobacteria particularly well suited for cross-scale studies from genes to the global ocean (4). Physiological studies of a range of *Prochlorococcus* strains isolated from various depths and geographical regions, notably revealed the occurrence of genetically distinct populations exhibiting different light or temperature growth optima and tolerance ranges (5, 6). These observations are congruent on the one hand, with the well-known depth partitioning of genetically distinct *Prochlorococcus* populations in the ocean, with high light-adapted (hereafter HL) populations in the upper lit layer and low light-adapted (hereafter LL) populations located further down the water column, and on the other hand, with the latitudinal partitioning between *Prochlorococcus* HLI and HLII clades that are adapted to temperate and tropical waters, respectively (5, 7, 8). For *Synechococcus*, although no clear depth partitioning (i.e., phototypes) has been observed so far, the occurrence of different 'thermotypes' has been clearly demonstrated among strains isolated from different latitudes (9, 10). This latter finding agrees well with biogeographical patterns of the most abundant *Synechococcus* lineages, with members of clades I and IV restricted to cold and temperate waters, while clade II populations are mostly found in warm, (sub)tropical areas (11-14). More recently, several studies have shown that iron could also be an important parameter controlling the composition of picocyanobacterial community structure since *Prochlorococcus* HLIII/IV ecotypes (15, 16) and *Synechococcus* clade CRD1 (17, 18) were shown to be dominant within high nutrient-low chlorophyll (HLNC) areas, where iron is limiting. Most of these studies considered members of the same clade —i.e. *Prochlorococcus* clades HLI-VI and LLI-VI or *Synechococcus* clades I-IX, which are congruent between different genetic markers (14, 19-22)— as one ecotype, i.e. a group of phylogenetically related organisms sharing the same ecological niche (4, 23). Yet the use of a high taxonomic resolution marker, the core, single copy *petB* gene encoding cytochrome b_6, has revealed different spatially structured populations (subclades) within the major *Synechococcus* clades that were adapted to distinct niches (12), suggesting that the 'clade' level might not be the most ecologically relevant taxonomic unit. Moreover, the systematic use of probes and/or PCR amplification might have led to overlook some important genetic diversity, a drawback potentially resulting in a poor assessment of the relative proportion of co-occurring populations at any given station. In this context, the occurrence of a huge microdiversity within wild *Prochlorococcus* populations was recently demonstrated by estimating the genomic diversity within coexisting members of the HLII clade using a large-scale single-cell genomics approach (24). Still, the congruency of phylogenies based on whole genome and internally transcribed spacer (ITS) suggests that ITS ribotype clusters coincide, in most cases, with distinct genomic backbones that would have diverged at least a few million years ago and the relative abundance of which vary through temporal and local adjustments (24). Thus, approaches using a single marker gene remain valid but fine spatial, temporal and taxonomic resolution is required to better understand how divergent picocyanobacterial lineages have adapted to different niches in the global ocean.

Here, we analyzed 109 metagenomic samples collected during the 2.5-year *Tara* Oceans circumnavigation (25, 26), a project surveying the diversity of marine plankton that produced nearly eleven times more non-redundant sequences than the previous Global Ocean Survey (GOS) expedition (15). In order to retrieve taxonomically relevant information for picocyanobacteria and to avoid PCR-amplification biases, reads targeting the high resolution *petB* gene (12) were recruited using a $_{\text{mi}}$Tag approach (27). Even though this approach did not give us access to the rare biodiversity, these analyses unveiled a previously unsuspected genetic diversity within both the *Prochlorococcus* and *Synechococcus* genera. Clustering based on the distribution patterns of picocyanobacterial communities allowed us to define Ecologically Significant Taxonomic Units (ESTUs), i.e., genetically related subgroups within clades that co-occur in the field. Analyses of the biogeography of ESTU assemblages showed that they were strongly correlated with specific environmental cues, allowing us to define distinct realized environmental niches for the major ESTUs.

Results

Revealing novel picocyanobacterial diversity using *petB*-$_{\text{mi}}$Tags and newly assembled sequences. To evaluate the taxonomic resolution potential of *petB* $_{\text{mi}}$Tags, for assessing picocyanobacterial genetic diversity, simulated 100 bp reads (i.e., the minimum size of the *Tara* Oceans merged metagenomic reads) were generated by fragmenting sequences from our reference databases (**Datasets 1-3**). This analysis showed that *petB* reads can be assigned reliably at the finest taxonomic level, i.e. subclade (12), over most of the gene length (**Fig. S1**). The *petB*-$_{\text{mi}}$Tags approach was therefore applied to the whole *Tara* Oceans transect (66 stations, 109 metagenomes, 20.2 ± 9.9 Gb of metagenomic data per sample). With the exception of the Southern Ocean and its vicinity (TARA_082 to TARA_085) for which no *petB* reads were recruited, picocyanobacteria were present at all sampled *Tara* Oceans stations. From 119 to 14,139 picocyanobacterial *petB* reads (average: 3,309; median: 2,545; **Dataset 4**) were recruited per sample using a non-redundant reference database of 585 high quality *petB* sequences, representing most of the genetic diversity identified so far among *Prochlorococcus* and *Synechococcus* isolates and environmental clone libraries (**Fig. 1**). Interestingly, most *petB* sequences in our database recruited at least one read from the *Tara* Oceans metagenome as best hit, with the notable exception of some sequences of the cold-water adapted *Synechococcus* clade I, likely due to the limited sampling performed at high latitudes during the *Tara* Oceans expedition (28). This suggests that most genotypes known so far are sufficiently well represented in the marine environment to be detected by this approach. Still, we cannot exclude that this preliminary analysis provides a somewhat biased picture of the diversity toward the 'already known', since most current reference sequence databases are potentially skewed by culture isolation and/or amplification biases.

To search for potential hidden genetic diversity within the *Tara* Oceans picocyanobacterial communities, we examined the percent identity of recruited reads with regard to their best hit in the *petB* database (**Figs. 2A-B and S2**). *Prochlorococcus* and *Synechococcus petB* sequences can be easily differentiated from non-specific signal by selecting reads above 80 % identity to the closest reference *petB* sequence. The diversity within the most abundant *Synechococcus* clades (I-IV) was generally well covered by reference sequences since most reads displayed >94 % identity to their best-hit in the database, a cut-off value previously shown to allow an optimal separation of *Synechococcus* lineages displaying distinct distribution patterns (12). In contrast, for other clades, some of the recruited reads were quite distantly related to reference sequences (i.e., between 80-94% identity), indicating that the *in situ* diversity of these clades was not fully covered by the reference database (**Fig. 2B**, top panels).

To have a more realistic and exhaustive view of this diversity, we assembled 136 distinct nearly complete *petB* sequences from environmental reads (121 *Prochlorococcus* and 15 *Synechococcus*), corresponding to the most divergent genotypes present in the whole *Tara* Oceans dataset**.** By adding these novel sequences to the reference database (see **Dataset 1** and sequences in white in **Fig. 1**), we significantly improved taxonomic assignments of *petB*-$_{\text{mi}}$Tags, since 80.3 % of the *Prochlorococcus* and 90.2 % of the *Synechococcus* environmental *petB* reads were found to display >94 % identity with their best hits in the enriched reference database, an increase of about 11 and 7 % compared to our initial assessment, respectively (**Figs. 2B and S2**). Interestingly, quite a few highly divergent sequences from *Prochlorococcus* HLIII, HLIV and LLI as well as *Synechococcus* CRD1 were assembled from TARA_052, located East of Madagascar, a station exhibiting a picocyanobacterial community atypical for this oceanic area (see below). Although most of these additional sequences fell into known phylogenetic clades, they allowed us to better assess the extent of genetic diversity within both *Prochlorococcus* and *Synechococcus* (**Fig. 1**). While only a few *petB* sequences, all coming from cultured strains, were available for the *Prochlorococcus* HLI and LLI clades prior to this study, we added 43 novel HLI sequences (within-clade nucleotide identity range: 87-99.6%), 29 LLI sequences (within-clade identity range: 85.5-99.6%) as well as 11 sequences of the uncultured HLIII and IV clades, some of which form distinct monophyletic branches comprised entirely of novel sequences (**Fig. 1 and Dataset 1**). Although many HLII sequences were recently obtained by high throughput single cell genomics focused on this clade (24), assembly of *Tara* Oceans reads allowed us to retrieve several divergent HLII sequences (within-clade identity range: 86.2-99.8%) including a new, well-supported group (corresponding to ESTU HLIIC, see below), located at the base of the HLII radiation. Similarly for *Synechococcus*, newly assembled sequences allowed us to refine the taxonomy of several taxa, notably for CRD1 and EnvB clades as well as subcluster 5.3, three ecologically important but previously overlooked phylogenetic lineages.

Using global picocyanobacterial distribution patterns to define ESTUs. As expected from previous literature (1, 2, 5, 29), *Prochlorococcus* was the dominant picocyanobacterium at the global scale, representing ~91% of all *petB* reads from the bacterial size fraction, compared to 9% for *Synechococcus* (**Fig. S3A**). These percentages compare fairly well with the global contribution of *Prochlorococcus* and *Synechococcus* estimated from flow cytometry data as 80.6% (2.9 ± 0.1 × 10^{27}cells) and 19.4 % (7.0 ± 0.3 × 10^{26} cells), respectively (1). The apparent lower contribution of *Synechococcus* in our dataset might be due to the fact that the *Tara* Oceans sampling was not made at random in the ocean, since most stations were located in the inter-tropical zone and/or selected for displaying specific traits of interest (e.g., upwelling, fronts, island proximity, etc.), while Flombaum and coworkers' dataset included many data from temperate stations, where *Synechococcus* is abundant.

To study the global distribution of these organisms at a finer taxonomic resolution, we then examined whether *Prochlorococcus* and *Synechococcus* clades and/or subclades were ecologically meaningful. To do this, we analyzed the distribution patterns along the *Tara* Oceans transect of within-clade Operational Taxonomic Units (OTUs), as defined using a cut-off at 94% nucleotide identity (**Figs. 2C and S4 and Dataset 5**). Although for some clades, OTUs displayed a homogeneous pattern over their geographical distribution area (e.g., *Prochlorococcus* HLIII and IV, **Fig. S4**) or were too scarce to reliably distinguish ESTUs (*Synechococcus* subcluster 5.2 and clades I, V-VIII, WPC1, EnvA, IX, XVI, XX, UC-A, *Prochlorococcus* clades LLII-IV), most of the prevalent clades encompassed several coherent OTU clusters displaying distinct distribution patterns (and thus likely occupying distinct ecological niches) that were gathered into independent ESTUs (**Fig. 2C, Fig. S4**). For instance, OTUs within *Synechococcus* clade CRD1 can be split into 3 ESTUs (CRD1A-C) based on clustering of their abundance per station. Some of these ESTUs correspond to previously described clades (e.g., *Prochlorococcus* HLIIIA and HLIVA) or subclades (e.g., *Synechococcus* IVC), while others gather subclades having similar distribution patterns. For instance, *Synechococcus* ESTU IIA encompasses subclades IIa-d and IIf and ESTU IIB gathers subclades IIe and IIh, as previously defined by Mazard et al. (12). Thus, although most previous field diversity studies on picocyanobacteria focused on clades (5, 14, 18, 21, 22), which were generally considered as distinct 'ecotypes' (*sensu* (19)), our data indicate that ESTUs provide a finer estimate of *Prochlorococcus* and *Synechococcus* ecotypes than do clades. This approach was then used to study the biogeography of marine picocyanobacteria along the *Tara* Oceans transect and stations exhibiting similar ESTU assemblages were clustered together (**Figs. 3A and 4A**).

Biogeographical analyses of *Prochlorococcus* reveals the occurrence of minor ESTUs with unexpected distribution patterns. Most major *Prochlorococcus* clades (HLI, HLII and LLI) could be split into several ESTUs, though for the former two, one ESTU was clearly predominant (**Figs. 3A and S5**). Only three major ESTU assemblages were identified in surface samples: i) dominance of HLIA ESTU in temperate waters (above 35°N and 32°S), ii) dominance of HLIIA in warm and iron-replete waters between 30°S and 30°N, with mixed HLIA-HLIIA profiles at intermediate latitudes and iii) co-occurrence of HLIIIA and IVA at a ratio of ca. 1:2.6 (± 0.7) in warm, high nutrient-low chlorophyll (HNLC) areas. The low abundance of LLII-IV clades in the whole *Tara* Oceans dataset (Fig. S6A-C) is likely due to the fact that they usually thrive below the DCM (5, 30), i.e. at depths not sampled during the expedition. In contrast, most LLI ESTUs were very abundant in subsurface waters (**Figs. S3 and S5b**) and sometimes even reached the surface (e.g., at TARA_066-070, **Figs. 3A**), as expected from the ability of members of the LLI clade to tolerate a strong mixing rate and short-term exposure to high light (5, 8, 30, 31).

HLIIIA and HLIVA ESTUs altogether contributed to 15.5% of the *Prochlorococcus* community in *Tara* Oceans samples, i.e. about as much as HLI (17%) or LLI (15.2%; **Fig. S3A**). This value is slightly higher than the 9% that were previously estimated for HLIII-IV clades from the analysis of GOS samples (11). Consistent with previous studies (11, 16, 32, 33), we show here that their distribution covers most of the warm (>25°C), low-Fe equatorial Pacific zone from 13°S (TARA_100) to 14°N (TARA_137), where they constitute the vast majority of the *Prochlorococcus* community in surface waters. In the Indian Ocean, we only observed them at two stations near the northern coast of Madagascar (TARA_052 and TARA_056), in agreement with a previous report that found them at two sites located further east (32), all these sites likely being influenced by the Indonesian throughflow originating from the tropical Pacific Ocean (34). Thus, HLIII/IV seemingly occurs over a much thinner latitudinal band (centered around 15°S) in the Indian compared to the Pacific Ocean, and they are apparently very scarce in the part of the Atlantic Ocean explored by the *Tara* schooner, even though the area around stations TARA_072 and TARA_070 is known to be iron-depleted (see Fig. S1 in (18)). Altogether, the distribution patterns of the dominant *Prochlorococcus* HL ESTUs seem to be mainly driven by temperature and iron availability, as confirmed by non-metric multidimensional scaling (NMDS) analyses (**Fig. 3C**). These results are globally consistent with previous reports that analyzed *Prochlorococcus* clades (5, 8, 16, 30, 32), indicating that the latter studies actually targeted the dominant ESTUs.

In contrast, a number of minor ESTUs were found to display distribution patterns very different from the major ESTU of the same clade. For instance, the relative contribution of the above mentioned novel HLIIC ESTU was highest at the DCM in the equatorial Indian Ocean (TARA_041-042; **Fig. S5b**), suggesting that members of this ESTU are adapted to mid-depth waters, much like members of the LLI clade (5, 30). Similarly, ESTUs HLIB and D can sometimes take over the prevalent HLIA populations and become abundant in surface waters at specific locations (e.g., at TARA_093 and TARA_094, respectively). In contrast, HLIC, which comprises a complex microdiversity (10 OTUs; **Fig. S4**), was found to exhibit a particularly large niche, co-occurring with HLIA at high latitude but also being present as the major HLI population in warm oligotrophic waters, where HLIIA dominated the *Prochlorococcus* community (e.g., in the Indian Ocean, **Fig. S6A**). This suggests that members of the HLIC ESTU might have a larger tolerance to temperature than the globally dominant HLIA. It is also worth noting that among the four ESTUs defined within the LLI clade, LLIB, which is entirely comprised of newly assembled *petB* sequences, dominates the LLI population in surface iron-limited HNLC areas in both the equatorial/tropical Pacific (TARA_110 to 128) and Indian Ocean (TARA_052, **Fig. S6B**). Thus, adaptation to low iron conditions in *Prochlorococcus* might not be an exclusive trait of HLIIIA and HLIVA.

CRD1 and EnvB ESTUs are the dominant *Synechococcus* lineages in the Pacific Ocean. *Synechococcus* assemblages were much more diverse than *Prochlorococcus* with 8 distinct ESTU clusters observed along the *Tara* Oceans transect (**Fig. 4A-B**). None of these assemblages were specific of a given oceanic region, though cluster 2 was mainly found in the Mediterranean Sea. ESTUs IA and IVA, IVB and/or IVC dominated at most stations within clusters 4, 5 and 8 that were typical of cold, coastal or mixed open ocean waters at high latitude, in agreement with previous reports on the distribution of clades I and IV (11, 12, 14, 18). In contrast, ESTU IIA, dominated by a single OTU (OTU003; **Fig. 2C**), was by far the major component of cluster 1, an assemblage characteristic of most warm, mesotrophic and oligotrophic iron replete waters that encompass the vast majority of the Atlantic and Indian Oceans (**Fig. 4B**). Consistently, NMDS analysis showed that the occurrence of clusters 4, 5, 8 on the one hand, and cluster 1 on the other hand, were associated both with temperature and Chl *a*, but in opposite ways (**Figs. 4C and S7**). Interestingly, while ESTU IIA was typical of warm waters, the minor ESTU IIB was found to be restricted to fairly cold (14.1 to 17.5°C), mixed waters and to co-occur with IVA-B (**Fig. 4**).

Several other salient features arose from analyses of the *Tara* Oceans metagenomes. First, ESTU IIIA, the major contributor of cluster 2, was found only in the Mediterranean Sea (TARA_007 to 030) and the Gulf of Mexico (TARA_142; **Fig. 4A-B**). Both areas are known to be P-depleted (35, 36), suggesting that the dominance of this ESTU could be linked to a specific adaptation to P limitation, as confirmed by the inverse correlation of cluster 2 with P concentrations (**Fig. 4C**) and correlation analyses between IIIA and individual physico-chemical parameters (**Fig. S7**). The differential availability of this nutrient on both sides of the Suez Canal is therefore probably responsible for the strong community shift from a IIIA- to a IIA-dominated assemblage between the Mediterranean and Red Sea (**Fig. S5a**), although one cannot exclude that other specific characteristics of the Mediterranean Sea, such as the presence in the eastern basin of copper, a trace metal toxic to a number of phytoplankton species (37), might also be involved. While the dominance of clade III in the Mediterranean Sea is consistent with previous studies (14, 38), it was also reported in fair abundance along a N-S transect in the northern Atlantic Ocean in fall 2004 (AMT15) as well as in sub-tropical waters of the Pacific and Atlantic oceans (12, 14), whereas we found it only as a minor component of the *Synechococcus* community in these areas. It is possible that the relative contribution of clade III might have been overestimated using PCR-based or dot-blot hybridization approaches. A more likely explanation is that this clade is subject to seasonality, as suggested by a year-round survey in the Red Sea, showing that clade III abundance peaks occur during summer, stratified conditions, and remains at low concentrations over the rest of the year (20, 39). In this context, it is important to note that during *Tara* Oceans, the north and south Atlantic as well as the southern Indian Ocean were all sampled during winter or early spring, while the Mediterranean Sea was sampled in fall (**Dataset 4**). Hence, this warrants future global metagenomic studies at various seasons as well as finer-scale studies looking at seasonal variations in community structure.

Also unexpected was the large global abundance (6% of total *Synechococcus* reads, Fig. S3) of subcluster 5.3 (formerly clade X; (40)). Members of ESTU 5.3A (mostly co-occurring with ESTU IIIA) were found mostly along the transect from Panama to Bermuda (TARA_140-149), in the Mozambique Channel (TARA_057 and TARA_062) as well as at all stations of the Red Sea and Mediterranean Sea, where they contributed up to ca. 30 % of the local *Synechococcus* community, e.g., at the Gibraltar strait (TARA_007, Fig. 4A-B). In contrast, ESTU 5.3B (co-occurring with ESTU IIA) was always present in low relative abundance. Members of subcluster 5.3 have only been sporadically detected in previous studies mostly in open-ocean habitats in the northwestern Atlantic and Pacific Ocean and in the Mediterranean Sea (11, 12, 14, 17, 21, 38), reaching significant abundances only in transitional waters, such as the Amazon plume or the Benguela upwelling (18).

These specific localizations might explain why only a few sequences of this subcluster were previously detected in the GOS database (11).

Another striking result of this study was the strong global contribution of the co-occurring clades CRD1 and EnvB (8.4% and 5.4% of total *Synechococcus* reads, respectively; **Fig. S3D-E**). Recently, low Fe regions of the western equatorial Pacific (5°S-10°N) and southeastern Atlantic Oceans (15-20°S) were shown to be dominated by CRD1 (17, 18), a clade that was previously thought to be specific to the Costa Rica dome, where *Synechococcus* cell densities are known to be the highest worldwide (41, 42). Here, we show that CRD1 and EnvB ESTUs actually co-dominate the *Synechococcus* community over most of the Pacific Ocean from 33°S to 35°N and can also be prevalent in both the South (TARA_068-072) and North Atlantic (TARA_150-152) as well as in the Indian Ocean (TARA_052) but are virtually absent from the Mediterranean Sea (**Fig. 4A-B**). So, it seems that, in contrast to *Prochlorococcus* HLIII/IV, the distribution of CRD1 in the Pacific Ocean extends way beyond HNLC areas. Furthermore, we show here that both the CRD1 and EnvB clades actually encompassed 3 distinct ESTUs, displaying partially overlapping niches and falling into five clusters (3, 5-8; **Fig. 4A**) that were also split far apart by NMDS analyses (**Fig. 4C**). CRD1B and EnvBB were restricted to high latitude, cold, mixed waters (cluster 8), where they systematically co-dominated with ESTU IA, IVA and IVC. This includes TARA_093 located in the Chilean upwelling, TARA_152 in North Atlantic as well as TARA_068 in South Atlantic corresponding to a young Agulhas ring (43). In contrast, CRD1C and EnvBC preferentially thrived in warm HNLC regions (cluster 3 and the warmest stations of cluster 6), with CRD1C largely dominating the *Synechococcus* population in the Pacific inter-tropical area as well as at the Indian Ocean station TARA_052. Comparatively, CRD1A and EnvBA that were found in both kinds of environments, appear to be much more ubiquitous and to tolerate a much wider temperature range, not only than other CRD1 and EnvB ESTUs, but also more generally than all other *Synechococcus* strains characterized so far in culture (9, 10). Several previous studies also reported the presence of CRD2, co-occurring with CRD1 mainly in the Costa Rica dome area and in equatorial waters and generally constituting around 10-15 % of the total *Synechococcus* surface population (17, 18). It is tempting to speculate that the *petB*-defined EnvB clade, which had so far only been reported at one station in the middle of the North Atlantic basin (12), corresponds to the ITS-defined CRD2 clade. However, the different proportions of EnvB and CRD2 relative to CRD1 strongly suggests that the qPCR primers used in these studies targeted only a fraction of the CRD2/EnvB population, possibly corresponding to EnvBC, which like CRD2, is positively correlated with temperature ((18) and **Fig. S7**). Alternatively, seasonal variations might also explain the differences observed between these two datasets.

Discussion

The comprehensive nature of the *Tara* Oceans dataset, analyzed here at high taxonomic resolution, has markedly improved our current knowledge of the global phylogeography of marine picocyanobacteria, and highlighted the key role of environmental parameters in shaping their distribution patterns. Indeed, by assigning *petB*-$_{\mathrm{mi}}$Tags recruited for each clade to narrow OTUs, then clustering those sharing a similar ecological distribution into the same ESTU, we showed that despite a wide genetic diversity, *Prochlorococcus* and *Synechococcus* communities can be split into a fairly limited number of characteristic ESTU assemblages, often dominated by one or two major ESTU(s).This includes the co-dominating *Prochlorococcus* HLIIIA-HLIVA, which co-dominated at a fairly constant ratio (1:2.6) all over low Fe regions (Fig. 3A), *Synechococcus* IIIA that was abundant all over the Mediterranean Sea or CRD1 and EnvB ESTUs, co-dominating the *Synechococcus* community in vast expanses of the Pacific Ocean (Fig. 4A). Interestingly, we also showed that most picocyanobacterial clades encompass minor ESTUs that occupy niches distinct from dominant ones. This indicates that there is ecologically meaningful fine-scale diversity within currently defined *Synechococcus* or *Prochlorococcus* clades, even though the latter have often be referred to as 'ecotypes' (5, 30). In this context, it is important to note that the *Prochlorococcus* genus is thought to have occurred concomitantly to the major diversification event that also led to the splitting of *Synechococcus* subcluster 5.1 into about fifteen distinct clades (21, 44, 45), suggesting that, from a phylogenetic point of view, the whole *Prochlorococcus* genus is actually equivalent to a single *Synechococcus* clade, explaining why linking clades to a given ecological niche is trickier for the latter genus. In *Prochlorococcus*, several physico-chemical parameters have seemingly played a decisive role in the genetic diversification of this genus, at distinct periods of its evolutionary history, starting with light (split between LL and HL lineages), then iron availability (HLIII/IV vs. other HL) and temperature (HLI *vs.* HLII; (19, 22, 46)). In contrast, nitrogen and phosphorus availability influenced genetic diversification only in the 'leaves' of the *Prochlorococcus* radiation, through lateral transfers of gene cassettes conferring on populations the ability to adapt to local N or P-depleted niches (47, 48). Despite this apparent solid relationship between *Prochlorococcus* phylogeny and community structure, a recent study looking at the genomic diversity of individual *Prochlorococcus* cells in a single water sample highlighted a huge microdiversity within the HLII clade (24). This microdiversity seemingly allows cells to adapt to slightly different selective pressures, such as biotic factors (phages, grazing, etc). Here, we also observed a large microdiversity within the HLII lineage, with 25 OTUs comprising 4 ESTUs, but in agreement with a recent study (49), there were only subtle differences between the distribution patterns of these intra-clade groups (except for ESTU HLIIC, represented by a single OTU; **Fig. 2C**), confirming that abiotic factors have only marginally affected the genetic diversification within this clade. In contrast, the microdiversity that we identified within HLI and LLI has seemingly allowed members of these clades to colonize ecological niches clearly different from that of the dominant ESTUs, extending the global niche occupied by these lineages. This includes LLIB, which seems to be adapted to Fe-limited surface waters, much like HLIIIA-IVA, as well as HLIC, which thrives not only in cold temperate waters, as do the more typical HLIA, but also in warm sub-tropical waters, where it co-occurs with the dominant HLIIA (**Fig. S6**). This is consistent with the recent finding that HLI sub-clades are driven by distinct environmental traits (49) and that even in HLII-dominated waters, HLI is never competed to extinction (7).

Similarly, splitting *Synechococcus* clades into ESTUs revealed that this genus comprises a number of specialists, mostly characterized by their respective temperature and Fe requirements (**Fig. 5**). While CRD1B/EnvBB, CRD1A/EnvBA/EnvAA and CRD1C/EnvBC were found in cold, intermediate and warm waters respectively with various degrees of Fe limitation, other ESTUs preferentially thrive in regions where this nutrient is not limiting in either cold (IA, IVA, IIB), intermediate (IIIA, 5.3A) or warm (IIA) waters. The third most discriminating parameter appears to be P-limitation that only ESTUs IIIA and 5.3A can stand, but only in Fe-replete conditions. It is also worth noting that several ESTUs, such as those classified as 'temperature intermediate', display a larger tolerance range with regard to temperature than their 'cold' and 'warm' counterparts (**Fig. 5**). Altogether, these results temper the paradigm of *Synechococcus* being a generalist and physiologically more plastic than *Prochlorococcus*, which mainly relied on the ability of the former to colonize much wider ecological niches than the latter and on the apparent absence of genome streamlining in *Synechococcus* compared to *Prochlorococcus* (19, 50-52). Thus, our results demonstrate that the observed ubiquity of the *Synechococcus* genus as a whole (1, 2) in fact rests on a complex suite of specialists adapted to fairly narrow niches, as is the case for *Prochlorococcus*.

Focusing on shifts in community composition associated to changes in local environmental conditions or to physical barriers (**Fig. S5a-b**) provided additional insights into this global picture and revealed that some ESTUs behave as opportunists. For instance, this is the case off the Marquesas Islands, where the proximity of the coast induced an iron enrichment at TARA_123 and 124 as compared to a typical HNLC situation at TARA_122 and TARA-128. While CRD1-C dominated at the latter stations, ESTU IIA took over this local population in these iron-replete patches (with an intermediate situation at TARA_125; **Fig S5a**). By comparison, the *Prochlorococcus* abundance drastically dropped at TARA_123 but without any significant change in the community structure, suggesting that the minor HLIIA component of this assemblage was not responsive enough to local iron enrichment to outcompete the dominant HLIIIA/IVA population. Another abrupt shift in community composition occurred at the Agulhas choke point off the southern tip of Africa, where huge anticyclonic rings (i.e., Agulhas rings) are formed in the Indian Ocean and then drift across the South Atlantic (43, 53). The strong drop in temperature, occurring within the youngest ring (TARA_068), was likely responsible for a large part in the shift from a typical subtropical ESTU assemblage in the Indian Ocean, dominated by *Prochlorococcus* HLIIA-C and *Synechococcus* IIA (TARA_064-065), to a cold water ESTU assemblage (HLIA-C, LLIA and C, CRD1A, EnvBA and IVA-B) at TARA_068 (**Fig. S5a**), suggesting that the latter ESTUs might also have an opportunistic behavior with regard to their warm waters counterparts. Although these two examples correspond to biogeochemical processes likely occurring at different time scales, the observed ESTU assemblage changes likely result from differences in the intrinsic dynamics of ESTUs within both genera, the most adapted one outcompeting others in favorable ecological conditions, with *Synechococcus* displaying a more opportunistic behavior than *Prochlorococcus*.

Our results also raise several questions that can only be addressed in the laboratory or in *silico*. From a physiological point of view, the fact that some ESTUs seemingly get counter-selected in response to nutrient enrichment (e.g., iron in the case of CRD1C) suggests that their growth capacity in nutrient replete conditions is lower than that of opportunistic ESTUs (e.g. IIA) and this could be checked by comparing representative strains of these two lifestyles in single or co-cultures. It is also unclear yet whether differences between these two behaviors is due to the loss of genes costly to maintain for the cells, to a better affinity of core enzymes (e.g., for nutrient scavenging) and/or to the acquisition of specific gene sets by lateral gene transfer, as reported for *Prochlorococcus* regarding phosphate and nitrogen uptake and assimilation (47, 48). Adaptation to low iron is particularly striking in this context since our study showed that this ability, previously thought to be specific to *Prochlorococcus* HLIII and IV (16, 32), seems to have appeared several times during evolution in quite distantly related ESTUs, namely *Prochlorococcus* HLIIIA/HLIVA —that likely occurred via a single diversification event— and LLIB as well as *Synechococcus* CRD1A, CRD1C, EnvBA, EnvBC and EnvAA (**Fig. 5**). Although no *Prochlorococcus* isolates of HLIIIA/IVA are available in culture yet, sequencing of single amplified genomes suggested that these organisms have adapted to iron-limited environments by lowering their cellular iron requirement through loss of genes encoding iron-rich proteins and by acquiring siderophore transporters for efficient scavenging of organic-bound forms of this element (32, 33). Genomic comparison of *Synechococcus* strains, including representatives of the different CRD1 ESTUs, as well as whole genome recruitment of metagenomic data should allow to check whether a similar adaptation process has occurred in this genus.

In conclusion, although very few studies have so far combined information from high resolution phylogenetic markers and geographical distribution to detect ecologically coherent taxonomic groups (e.g., (49, 54)), we show here that this approach can bring invaluable insights for deciphering the links between genetic diversity and niche occupancy. Indeed, the definition of within-clade ESTUs using a reference *petB* database enriched with ecologically relevant and distantly related sequences assembled from *Tara* Oceans reads, has allowed us to obtain clear-cut spatial distribution patterns for taxa within both *Prochlorococcus* and *Synechococcus* genera, indicating that we explored the diversity of the picocyanobacterial community at the right taxonomic resolution. Additionally, in contrast to other phytoplankton groups, such as diatoms (55), these biogeographical patterns were found to be tightly controlled by environmental factors. Besides helping to refine models of picocyanobacterial distributions and predicting their behavior in response to ongoing climate change, knowledge of the oceanic areas where poorly characterized ESTUs predominate, will also guide future strain isolation (e.g., for the yet uncultured EnvA and EnvB) and sequencing efforts. Characterizing and comparing such ecologically representative strains will help further unveil the basis of niche partitioning.

Materials and methods

Genomic material. This study focused on 109 *Tara* Oceans metagenomes corresponding to 66 stations along the *Tara* Oceans transect for which a ‘bacterial size fraction’ was available (i.e. 0.2-1.6 µm for TARA_004 to TARA_052 and 0.2-3 µm for TARA_056 to TARA_152). Water samples were collected at two depths, surface (SUR) and deep chlorophyll maximum (DCM), the latter sample sometimes being merely collected in the upper mixed layer, when the DCM was not clearly delineated (**Dataset 4**). Metagenomes were sequenced using the Illumina® technology as overlapping paired reads of ~100/108 bp with various sequencing depths, ranging from 16 x 10^6 to 258 x 10^6 reads after quality control, corresponding to an average 20.2 ± 9.9 Gb of sequence data per sample. Reads were merged using FLASH v1.2.7 with default parameters (56) and cleaned based on quality using CLC QualityTrim v4.10.86742 (CLC Bio, Aarhus, Denmark), resulting in 100 to 215 bp fragments. **Dataset 4** describes all metagenomic samples with location and sequencing effort. All metagenomes and corresponding environmental parameters measured during the *Tara* Oceans expedition are available at www.pangea.de, except for the iron and ammonium data that were simulated with the ECCO2-Darwin model and the iron limitation index Φsat (57) and are available in Dataset 4.

Building of the PetB-DB database. To recruit and taxonomically assign metagenomics reads targeting the high resolution *petB* gene marker, we analyzed 1,091 sequences of the *petB* gene from cultured isolates and environmental samples and built a reference database including all non-redundant high quality sequences of this marker available for the marine picocyanobacteria *Prochlorococcus* (69 sequences covering 7 clades) and *Synechococcus* (399 sequences covering 3 subclusters, 22 clades and 30 subclades). The dataset also includes outgroup sequences from publicly available cyanobacteria, including marine (13 sequences) and freshwater isolates (40 sequences), as well as representatives of the main marine eukaryotic phytoplankton taxa and eukaryotic cyanobionts (64 plastid *petB* sequences), raising the number of *petB* sequences to 585 (**Tables S1 and S2**). To avoid differential alignment effects at the edge of the reference sequences, all sequences were aligned and trimmed to 557 bp. This database was secondarily complemented by 136 *petB* sequences assembled from selected *Tara* Oceans stations and displaying less than 94 % identity with previously known *petB* sequences (yet some of these new sequences could exhibit more than 94 % identity with one another).

Read recruitments. Targeted *petB* fragment recruitments were performed using a two-step protocol. In order to maximize the diversity while reducing the weight of the resulting tabulated files, translated sequences of the non-redundant *petB* database were used to recruit candidate *petB* gene fragments by BLASTX (v2.2.28+) using default parameters but by limiting the results to 1 target sequence. These *petB* candidates were then compared to the full reference *petB* database using BLASTN (v2.2.28+) with sensitive configuration (–task blastn –gapopen 8 –gapextend 6 –reward 5 –penalty -4 –word_size 8) and cut-offs to reduce the weight of resulting tabulated files (–perc_identity 50 –evalue 0.0001).

Reads with more than 90 % of their sequence aligned and with more than 80 % sequence identity to their best-hit (see result section for the determination of this cut-off) were selected as genuine picocyanobacterial *petB*, taxonomically assigned to their best-hit and subsequently used to build per-strain read counts tables. Counts were then aggregated by clade or ESTU and subsequently used to build pie charts or community structure profiles.

Phylogenetic and statistical analyses. Phylogenetic reconstructions were based on multiple alignments of *petB* nucleotide sequences generated using MAFFT v7.164b with default parameters (58). A maximum likelihood tree was inferred using PHYML v3.0 – 20120412, (59) with the HKY + G substitution model, as determined using jModeltest v2.1.4 (60), and the estimation of the gamma distribution parameter of the substitution rates among sites and of the proportion of invariables sites. Confidence of branch points was determined by performing bootstrap analyses including 1000 replicate data sets. Phylogenetic trees were edited using the Archaeopteryx v0.9901 beta program (61) and drawn using iTOL (http://itol.embl.de; (62)). Operational taxonomical units (OTUs) for the *petB* reference data set at 94% were defined by nucleotide identity using Mothur v1.34.4 (63).

In each clade, ESTUs were defined using a type 3 SIMPROF approach (54) by considering: i) for *Prochlorococcus*, stations with more than 100 reads and OTUs recruiting more than 150 reads and ii) for *Synechococcus,* stations with more than 20 reads and OTUs recruiting more than 25 reads. Hierarchical clustering was performed on the remaining stations and OTUs using the Bray-Curtis distance between relative abundance profiles using *heatmap.3* function in GMD v0.3.1.1 R package (ward algorithm; (64)). Statistical significance of the difference between clusters was first assessed by a permutation analysis using the *clustsig* v1.1 R package (alpha=0.05, Bray-Curtis distance, otherwise default parameters). ESTU delineation was then manually refined, e.g. ESTUs were sometimes defined from single OTUs if the Bray-Curtis distance was >0.65 or if pairs of OTUs were not defined as coherent groups because all OTUs within a clade were equally distant from each other. In contrast, some potential ESTUs were not considered as reliable, e.g. if high Bray-Curtis distances were due to differences in abundance and not in distribution.

Hierarchical clustering and NMDS analyses of stations were performed using R packages *cluster* v1.14.4 (65) and *MASS* v7.3-29 (66), respectively. *petB*-$_{\text{mi}}$Tag contingency tables aggregated at the ESTU level were filtered as above and normalized using Hellinger transformation that gives lower rates to rare ESTUs. Bray-Curtis distance was then used for both clustering (*agnes* function, default parameters) and ordination (*isoMDS* function, maxit=100, k=2). All displayed clusters were significant ($p < 0.01$, permutation tests). Fitting of environmental parameters on NMDS ordination was performed with function *envfit* in vegan v2.2-1 package and p-value based on 999 permutations was used to assess the significance of the fit and only environmental parameters showing a p-value below 0.05 were used.

Visualization of realized environmental niches. In order to visualize the tolerance range of each ESTU with regard to physico-chemical parameters, values were scaled and reduced before analysis. For each ESTU, *Tara* Oceans stations were sorted by order of abundance, and stations gathering 80% of all reads of the given ESTU were kept. A boxplot was then computed for each parameter taking into account the values of this parameter in the kept stations.

Acknowledgements

We warmly thank Mick Follows and Oliver Jahn for providing us with ECCO2-Darwin simulation values for iron and Sabrina Speich for fruitful discussions on oceanographic context. This work was supported by the French “Agence Nationale de la Recherche” Programs BioAdapt SAMOSA (ANR-13-ADAP-0010) and France Génomique (ANR-10-INBS-09), the French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008), UK Natural Environment Research Council grants NE/I00985X/1 and NE/J02273X/1, and the European Union's Seventh Framework Programmes FP7 MicroB3 (grant agreement 287589; http://www.microb3.eu) and MaCuMBA (grant agreement 311975, http://www.macumbaproject.eu). We also thank the support and commitment of the *Tara* Oceans coordinators and consortium, Agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the EDF Foundation, FRB, the Prince Albert II de Monaco Foundation, the Tara schooner and its captains and crew. Tara Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org).

References

1. Flombaum P*, et al.* (2013) Present and future global distributions of the marine cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc Natl Acad Sci USA* 110(24):9824-9829.
2. Partensky F, Hess WR, & Vaulot D (1999) *Prochlorococcus*, a marine photosynthetic prokaryote of global significance. *Microbiol Mol Biol Rev* 63(1):106-127.
3. Dutkiewicz S*, et al.* (2015) Impact of ocean acidification on the structure of future phytoplankton communities. *Nat Clim Change*. 5: 1002-1009.

4. Coleman ML & Chisholm SW (2007) Code and context: *Prochlorococcus* as a model for cross-scale biology. *Trends Microbiol* 15(9):398-407.

5. Johnson ZI*, et al.* (2006) Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. *Science* 311(5768):1737-1740.

6. Moore LR, Rocap G, & Chisholm SW (1998) Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* 393(6684):464-467.

7. Chandler JW*, et al.* (2016) Variable but persistent coexistence of *Prochlorococcus* ecotypes along temperature gradients in the ocean's surface mixed layer. *Environ Microbiol Rep*. 8(2): 272-284.

8. Zinser ER*, et al.* (2007) Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. *Limnol Oceanogr* 52(5):2205-2220.

9. Mackey KR*, et al.* (2013) Effect of temperature on photosynthesis and growth in marine *Synechococcus* spp. *Plant Physiol* 163(2):815-829.

10. Pittera J*, et al.* (2014) Connecting thermal physiology and latitudinal niche partitioning in marine *Synechococcus*. *The ISME J* 8(6):1221-1236.

11. Huang S*, et al.* (2012) Novel lineages of *Prochlorococcus* and *Synechococcus* in the global oceans. *The ISME J* 6(2):285-297.

12. Mazard S, Ostrowski M, Partensky F, & Scanlan DJ (2012) Multi-locus sequence analysis, taxonomic resolution and biogeography of marine *Synechococcus*. *Environ Microbiol* 14(2):372-386.

13. Zwirglmaier K*, et al.* (2007) Basin-scale distribution patterns of picocyanobacterial lineages in the Atlantic Ocean. *Environ Microbiol* 9(5):1278-1290.

14. Zwirglmaier K*, et al.* (2008) Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages among oceanic biomes. *Environ Microbiol* 10(1):147-161.

15. Rusch DB*, et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* 5(3):398-431.

16. West NJ, Lebaron P, Strutton PG, & Suzuki MT (2010) A novel clade of *Prochlorococcus* found in high nutrient low chlorophyll waters in the South and Equatorial Pacific Ocean. *The ISME J* 5(6):933-944.

17. Ahlgren NA*, et al.* (2014) The unique trace metal and mixed layer conditions of the Costa Rica upwelling dome support a distinct and dense community of *Synechococcus*. *Limnol Oceanogr* 59:2166–2218.

18. Sohm JA*, et al.* (2016) Co-occurring *Synechococcus* ecotypes occupy four major oceanic regimes defined by temperature, macronutrients and iron. *The ISME J* 10: 333-345.

19. Kettler G*, et al.* (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* 3:e231.

20. Post AF*, et al.* (2011) Long term seasonal dynamics of *Synechococcus* population structure in the Gulf of Aqaba, northern Red Sea. *Front Microbiol* 2(2):131.

21. Ahlgren NA & Rocap G (2012) Diversity and distribution of marine *Synechococcus*: Multiple gene phylogenies for consensus classification and development of qPCR Assays for sensitive measurement of clades in the ocean. *Front Microbiol* 3:213.

22. Biller SJ, Berube PM, Lindell D, & Chisholm SW (2015) *Prochlorococcus*: the structure and function of collective diversity. *Nat Rev Microbiol* 13(1):13-27.

23. Koeppel AF*, et al.* (2013) Speedy speciation in a bacterial microcosm: new species can arise as frequently as adaptations within a species. *The ISME J* 7(6):1080-1091.

24. Kashtan N*, et al.* (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. *Science* 344(6182):416-420.

25. Armbrust EV & Palumbi SR (2015) Marine biology. Uncovering hidden worlds of ocean biodiversity. *Science* 348(6237):865-867.

26. Karsenti E*, et al.* (2011) A holistic approach to marine eco-systems biology. *Plos Biol* 9(10).

27. Logares R*, et al.* (2014) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ Microbiol* 16(9):2659-2671.

28. Sunagawa S*, et al.* (2015) Ocean plankton. Structure and function of the global ocean microbiome. *Science* 348(6237):1261359.

29. Bouman HA*, et al.* (2006) Oceanographic basis of the global surface distribution of *Prochlorococcus* ecotypes. *Science* 312(5775):918-921.

30. Malmstrom RR*, et al.* (2010) Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *The ISME J* 4(10):1252-1264.

31. Partensky F & Garczarek L (2010) *Prochlorococcus*: advantages and limits of minimalism. *Ann Rev Mar Sci* 2:305-331.

32. Rusch DB, Martiny AC, Dupont CL, Halpern AL, & Venter JC (2010) Characterization of *Prochlorococcus* clades from iron-depleted oceanic regions. *Proc Natl Acad Sci USA* 107(37):16184-16189.

33. Malmstrom RR*, et al.* (2013) Ecology of uncultured *Prochlorococcus* clades revealed through single-cell genomics and biogeographic analysis. *The ISME J* 7(1):184-198.

34. Song Q, Gordon AL, & Visbeck M (2004) Spreading of the Indonesian Throughflow in the Indian Ocean. *J Phys Oceanogr* 34(4):772–792.

35. Moutin T, *et al.* (2002) Does competition for nanomolar phosphate supply explain the predominance of the cyanobacterium *Synechococcus*? *Limnol Oceanogr* 47(5):1562-1567.

36. Popendorf KJ & Duhamel S (2015) Variable phosphorus uptake rates and allocation across microbial groups in the oligotrophic Gulf of Mexico. *Environ Microbiol* 17(10):3992-4006.

37. Paytan A, *et al.* (2009) Toxicity of atmospheric aerosols on marine phytoplankton. *Proc Natl Acad Sci USA* 106:4601-4605.

38. Mella-Flores D, *et al.* (2011) Is the distribution of *Prochlorococcus* and *Synechococcus* ecotypes in the Mediterranean Sea affected by global warming? *Biogeosciences* 8:2785–2804.

39. Fuller NJ, *et al.* (2005) Dynamics of community structure and phosphate status of picocyanobacterial populations in the Gulf of Aqaba, Red Sea. *Limnol Oceanogr* 50(1):363-375.

40. Dufresne A, *et al.* (2008) Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol* 9(5):R90.

41. Saito MA, Rocap G, & Moffett JW (2005) Production of cobalt binding ligands in a *Synechococcus* feature at the Costa Rica upwelling dome. *Limnol Oceanogr* 50(1):279-290.

42. Gutierrez-Rodrıguez A, *et al.* (2014) Fine spatial structure of genetically distinct picocyanobacterial populations across environmental gradients in the Costa Rica Dome. *Limnol Oceanogr* 59(3):705–723.

43. Villar E, *et al.* (2015) Ocean plankton. Environmental characteristics of Agulhas rings affect interocean plankton transport. *Science* 348(6237):1261447.

44. Urbach E & Chisholm SW (1998) Genetic diversity in *Prochlorococcus* populations flow cytometrically sorted from the Sargasso Sea and Gulf Stream. *Limnol Oceanogr* 43(7):1615-1630.

45. Fuller NJ, *et al.* (2003) Clade-specific 16S ribosomal DNA oligonucleotides reveal the predominance of a single marine *Synechococcus* clade throughout a stratified water column in the Red Sea. *Appl Environ Microbiol* 69(5):2430-2443.

46. Martiny JB, Jones SE, Lennon JT, & Martiny AC (2015) Microbiomes in light of traits: A phylogenetic perspective. *Science* 350(6261):aac9323.

47. Martiny AC, Huang Y, & Li W (2009) Occurrence of phosphate acquisition genes in *Prochlorococcus* cells from different ocean regions. *Environ Microbiol* 11(6):1340-1347.

48. Martiny AC, Kathuria S, & Berube PM (2009) Widespread metabolic potential for nitrite and nitrate assimilation among *Prochlorococcus* ecotypes. *Proc Natl Acad Sci USA* 106(26):10787-10792.

49. Larkin AA*, et al.* (2016) Niche partitioning and biogeography of high light adapted *Prochlorococcus* across taxonomic ranks in the North Pacific. *The ISME J*. doi:10.1038/ismej.2015.244

50. Palenik B*, et al.* (2003) The genome of a motile marine Synechococcus. *Nature* 424(6952):1037-1042.

51. Scanlan DJ*, et al.* (2009) Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* 73(2):249-299.

52. Dufresne A, Garczarek L, & Partensky F (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol* 6(2):R14.

53. Biastoch A, Boning CW, & Lutjeharms JR (2008) Agulhas leakage dynamics affects decadal variability in Atlantic overturning circulation. *Nature* 456(7221):489-492.

54. Somerfield PJ & Clarke KR (2013) Inverse analysis in non-parametric multivariate analyses: distinguishing of groups of associated species which covary coherently across samples. *J Exp Mar Biol Ecol* 449:261 – 273

55. Malviya S*, et al.* (2016) Insights into global diatom distribution and diversity in the world's ocean. *Proc Natl Acad Sci USA* E1516-E1525.

56. Magoc T & Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27(21):2957-2963.

57. Behrenfeld MJ*, et al.* (2009) Satellite-detected fluorescence reveals global physiology of ocean phytoplankton. *Biogeosciences* 6(5):779-794.

58. Katoh K & Standley DM (2014) MAFFT: iterative refinement and additional methods. *Methods Mol Biol* 1079:131-146.

59. Guindon S & Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696-704.

60. Darriba D, Taboada GL, Doallo R, & Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 9(8):772.

61. Han MV & Zmasek CM (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinfo* 10:356.

62. Letunic I & Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* 23(1):127-128.

63. Schloss PD*, et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75(23):7537-7541.

64. Zhao X, Valen E, Parker BJ, & Sandelin A (2011) Systematic clustering of transcription start site landscapes. *PLoS One* 6(8):e23409.

65. Maechler M, Rousseeuw P, Struyf A, Hubert M, & Hornik K (2015) Cluster: cluster analysis basics and extensions. R package version 2.0.3), 2.0.3.

66. Venables WN & Ripley BD (2002) *Modern applied statistics with S* (Springer, New York) 4th Ed Ed p 495 pp.

67. Biller SJ*, et al.* (2014) Genomes of diverse isolates of the marine cyanobacterium *Prochlorococcus*. *Nature Scient Data* 1:140034.

68. Choi DH & Noh JH (2009) Phylogenetic diversity of *Synechococcus* strains isolated from the East China Sea and the East Sea. *FEMS Microbiol Ecol* 69(3):439-448.

Figure Legends

Figure 1. Maximum likelihood tree of *Synechococcus* and *Prochlorococcus* lineages based on *petB* gene sequences from both isolates and environmental sequences. Diamonds at nodes indicate bootstrap support over 70%. Taxonomic assignments are given by the color codes at clade level for *Prochlorococcus* (top left) and clade (e.g. V, 5.2a) or subclade (e.g. VIIa) for *Synechococcus* (bottom right). Sequences were named after ID_subcluster_clade_subclade_ESTU for *Synechococcus* ID_LL or HL_clade_ESTU for *Prochlorococcus*. The outer pink ring indicates that the corresponding sequence in the tree was the best-hit of at least one *Tara* Oceans picocyanobacterial read and the inner blue bar plot shows the $\log_2$ of the number of metagenomic reads recruited for this sequence (range: 1-10.84). Sequences in black letters correspond to the initial reference database and those in white or light grey letters to newly assembled *petB* sequences from *Tara* Oceans metagenome reads. The scale bar represents the number of substitutions per nucleotide position. For improved readability, the length of three *Prochlorococcus* branches was reduced, as indicated by double slashes. *Prochlorococcus* clade assignment is as in (67), while for *Synechococcus* subcluster 5.1, subclade assignments are as in (68) for WPC1 and WPC2 and as in (12) for all other clades

Figure 2. **Percent identity of *Tara* Oceans *petB* $_{\text{mi}}$tags vs. sequences of the reference database and abundance at different stations along the transect of operational taxonomic units (OTUs) clustered into ESTUs.** (**A**) Distribution of the percent identity of best-hits of all *petB* candidate reads recruited from the *Tara* Oceans bacterial-size fraction metagenomes against the *petB* reference database. Populations 1 and 2 correspond respectively to genuine *petB* reads and to non-specific signal, due either to *petB* reads from organisms not included in the reference database or to *petB*-related genes. The grey part in population 1 corresponds to *petB* reads attributable to photosynthetic organisms of the reference database other than *Prochlorococcus* and *Synechococcus*. The red arrow shows the 80% cut-off used to separate the *petB* signal from noise. The top and bottom panels correspond to recruitments made before and after addition of the 136 newly assembled environmental *petB* sequences, respectively. (**B)** Same as above but for some selected *Synechococcus* taxa (see **Fig. S2** for all other picocyanobacterial taxa). (**C)** Determination of ESTUs based on the distribution patterns of within-clade 94% OTUs. At each station, the number of reads assigned to a given OTU is normalized by the total number of reads assigned to the clade in this station. Stations and OTUs are filtered based on the number of reads recruited and hierarchically clustered (Bray-Curtis distance) according to distribution pattern. Only *Synechococcus* clades split into different ESTUs are shown (see Fig. S4 for *Prochlorococcus*). Stars indicate nodes supported by p-value < 0.05 (test not applicable to pair comparisons).

Figure 3. Biogeography of *Prochlorococcus* ESTUs in surface *Tara* Oceans metagenomes and relation to physico-chemical parameters. (**A)** Histograms of the relative abundance of *Prochlorococcus* ESTUs at each station sorted by similarity, as determined by hierarchical clustering (Bray-Curtis distance). Left panel indicates seawater temperature (°C) at each station. **(B)** Distribution of the ESTU assemblages, color-coded as in A, along the *Tara* Oceans transect. **(C)** NMDS analysis of stations according to Bray-Curtis dissimilarity between *Prochlorococcus* assemblages, with fitted statistically significant physico-chemical parameters. Samples that belong to the same ESTU assemblage have been colored according to the color-code defined in A and contours of the same color gather all samples comprised within each cluster. NMDS stress value: 9.852.

Figure 4. Same as Fig. 3 but for *Synechococcus*. NMDS stress value: 13.694.

Fig. 5: Realized environmental niche of the major *Synechococcus* ESTUs in surface waters. For each ESTU, stations were sorted by order of normalized abundance and only those stations cumulating 80% of the total abundance were used to draw the graph. Boxplots represent the range of each parameter (in relative units) tolerated by any given ESTU and the median is indicated by a yellow line. ESTUs are organized according to their relative temperature range (cold, intermediate or warm), tolerance to iron limitation (-Fe, +Fe) and tolerance to phosphate limitation (-PO4). Please note that the two proxies used to estimate Fe-limitation ([Fe] derived from the ECCO2-Darwin model and the Φsat index; the red line indicates the 1.4 % value above which iron is considered limiting; (57)) are sometimes contradictory e.g., for CRD1B and EnvBB.

Supporting information

Figure S1: Variation of the assignment ability of each individual 100 bp gene fragment along the sequence of *petB* gene using reference databases for *Prochlorococcus* (A) or *Synechococcus* (B). Simulated reads were generated by 100 bp sliding windows along the marker sequences and the lowest taxonomic level at which they could be assigned is shown by a different blue tone (as indicated in the insert; for *Prochlorococcus*, the subcluster level actually corresponds to a LL or HL assignment, while the clade level corresponds to HLI-IV and LLI-IV, the lowest taxonomic level available for this genus).

Figure S2a: Distribution of the percent identity of *petB*-$_{\text{mi}}$tags recruited from the bacterial-size fraction of the *Tara* Oceans metagenomes with regard to their best-hits in the reference database for each *Prochlorococcus* clade (top 9 graphs) and *Synechococcus* subclade (bottom 18 graphs) before addition of the 136 newly assembled environmental *petB* sequences. Note that clade XX was formerly called EnvC (12) but the name was changed here because there is at least one representative isolate (i.e., strain CC9616).

Figure S2b: Same as Fig. S2a but after addition of the 136 newly assembled environmental *petB* sequences.

Figure S3: Global recruitments of marine picocyanobacteria *petB* $_{\text{mi}}$tags in the bacterial size fraction of the *Tara* Oceans metagenomes. (A) All picocyanobacterial clades at both sampled depths; (B-C) percentage of each *Prochlorococcus* clade in surface (B) and at the deep chlorophyll maximum (DCM; C). (D-E) percentage of each *Synechococcus* clade in surface (D) and at the DCM (E). Note that clade XX was formerly called EnvC (12) but the name was changed here because there is now at least one representative isolate (i.e., strain CC9616).

Figure S4: *Prochlorococcus* ESTUs based on the distribution patterns of within-clade 94% OTUs. At each station, the number of reads assigned to a given OTU is normalized by the total number of reads assigned to the clade in this station. Stations and OTUs are filtered based on the number of reads recruited. OTUs are hierarchically clustered (Bray-Curtis distance) according to their distribution pattern. Stars indicate nodes supported by p-value < 0.05 (test not applicable to pair comparisons).

Figure S5a: Marine picocyanobacteria community structure in *Tara* Oceans surface metagenomes based on *petB*-$_{\mathrm{mi}}$Tags recruitments. (A) Surface water temperature along the *Tara* Oceans transect. (B) Relative abundances of *Prochlorococcus* and *Synechococcus* normalized to the total number of reads at each station. (C-D) Relative abundances of *Prochlorococcus* and *Synechococcus* ESTUs, respectively. White, grey and black dots indicate the number of reads used to build the profile, as detailed in the insert. For readability, temperature for stations TARA_082 (7.3°C), TARA_084 (1.8°C) and TARA_085 (0.7°C) are not shown on graph A. Abbreviations: IO, Indian Ocean; MS; Mediterranean Sea; NAO: North Atlantic Ocean; NPO, North Pacific Ocean; RS, Red Sea; SAO, South Atlantic Ocean; SO, Southern Ocean.

Figure S5b: Same as Fig. S5a but at the DCM. A depth profile along the Tara Oceans transect was added. For readability, temperature for stations TARA_082 (7.0°C) and TARA_085 (-0.8°C) are not shown on graph A, while temperatures for stations TARA_007 and TARA_084 are missing.

Figure S6: Distribution of minor *Prochlorococcus* ESTUs with regard to major ESTUs in the *Tara* Oceans metagenomes. Relative abundance normalized to the total number of reads per ESTU of (A) ESTUs HLIA and HLIC with regard to HLIIA in surface waters and (B-C) ESTUs LLIA-C with regard to HLIIIA in surface waters and the DCM, respectively. For graph A, stations were sorted from the lowest to highest temperatures and for graph B by sampling date.

Figure S7: Correlation analysis between marine picocyanobacterial ESTUs and environmental parameters measured along the *Tara* Oceans transect for all sampled depths. (A) *Prochlorococcus* ESTUs, (B) *Synechococcus* ESTUs. The scale shows the degree of correlation (blue) or anti-correlation (red) between the two sets of data. Correlations with p-value > 0.05 are indicated by grey crosses. Abbreviations: Sal, salinity; Temp, temperature; fCDOM, fluorescence, colored dissolved organic matter; MLD, mixed layer depth; DCM, deep chlorophyll maximum; Φsat, satellite-based NPQ-corrected quantum yield of fluorescence.

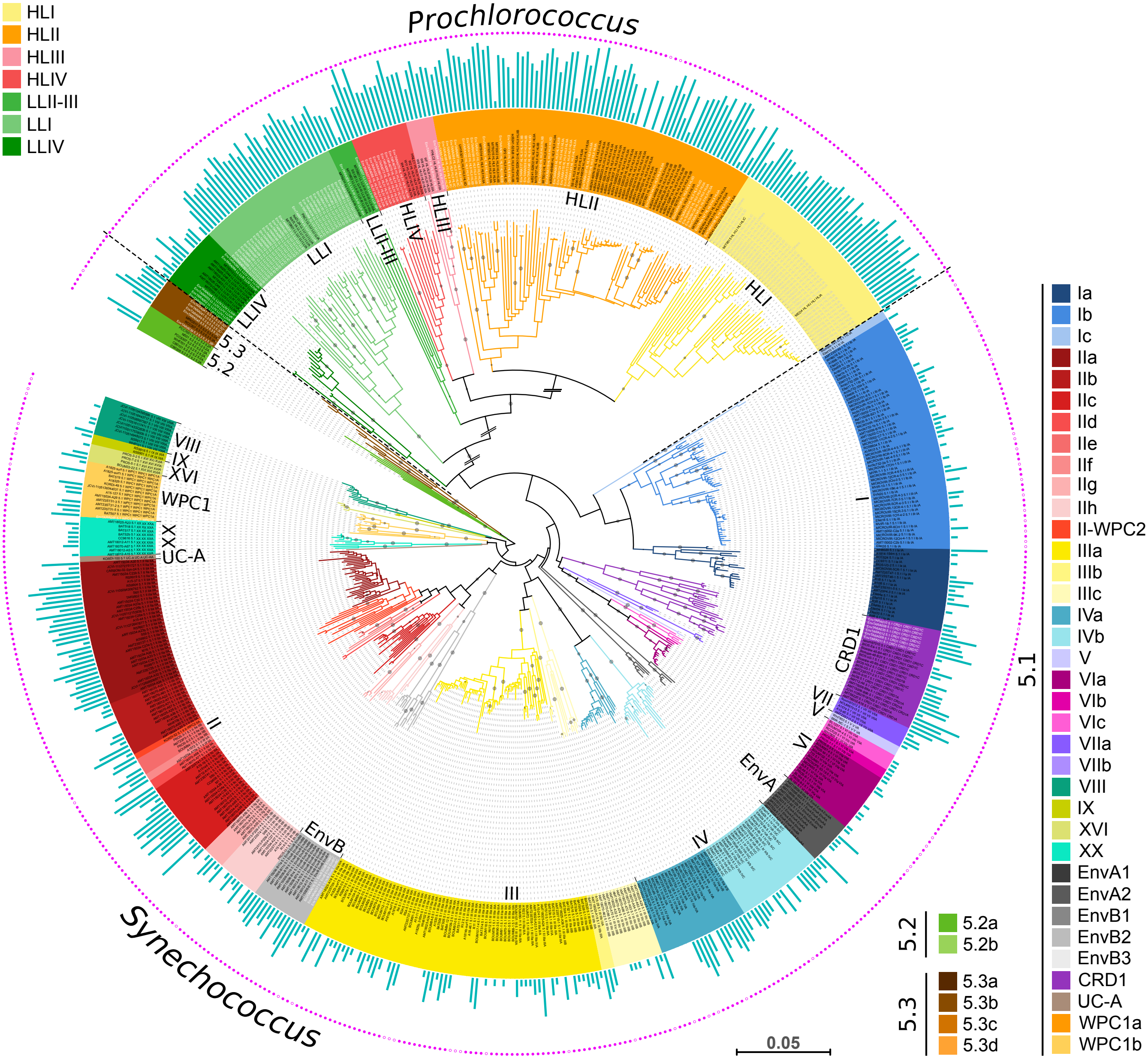
Dataset 1: Summary data for picocyanobacterial *petB* reference sequences used in this study, including newly assembled sequences. The table includes subclade designation based on (12).

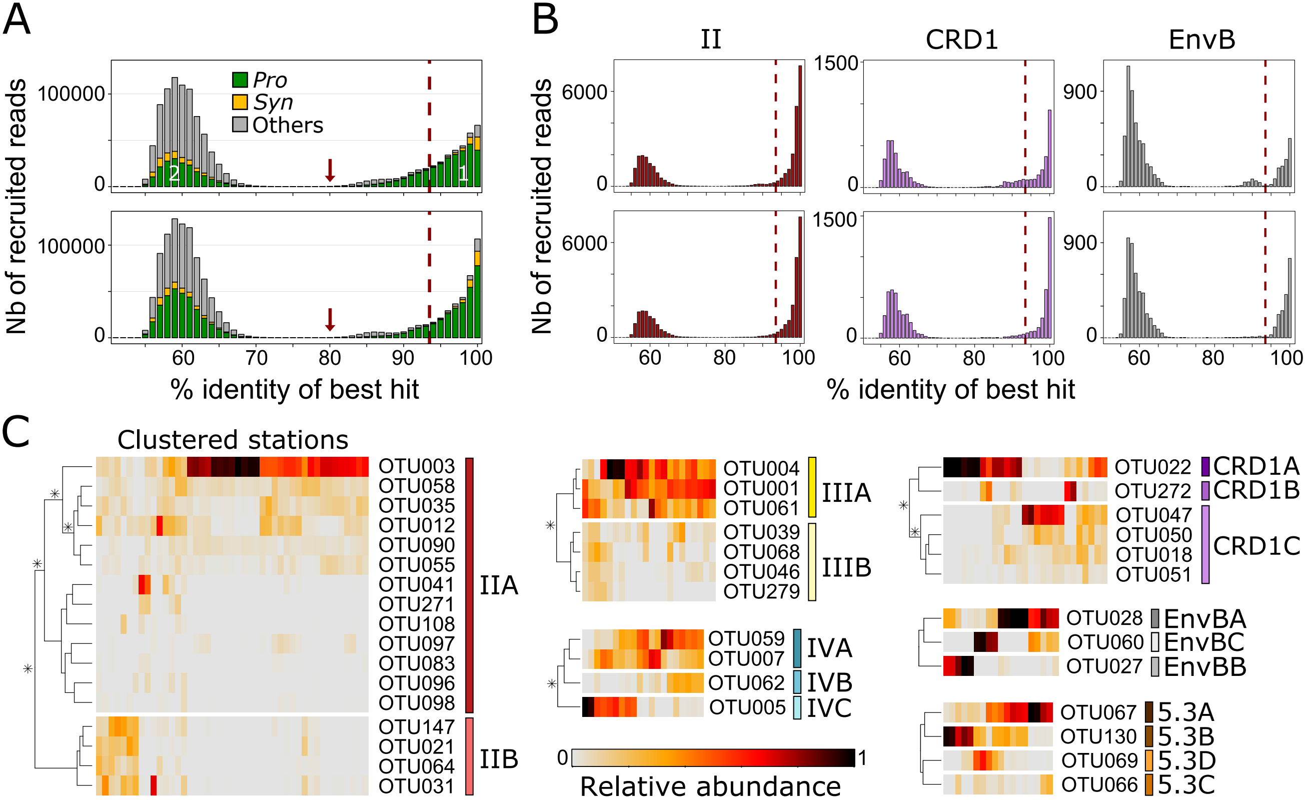
Dataset 2: Summary data for *petB* reference sequences for photosynthetic organisms other than marine picocyanobacteria used in this study.

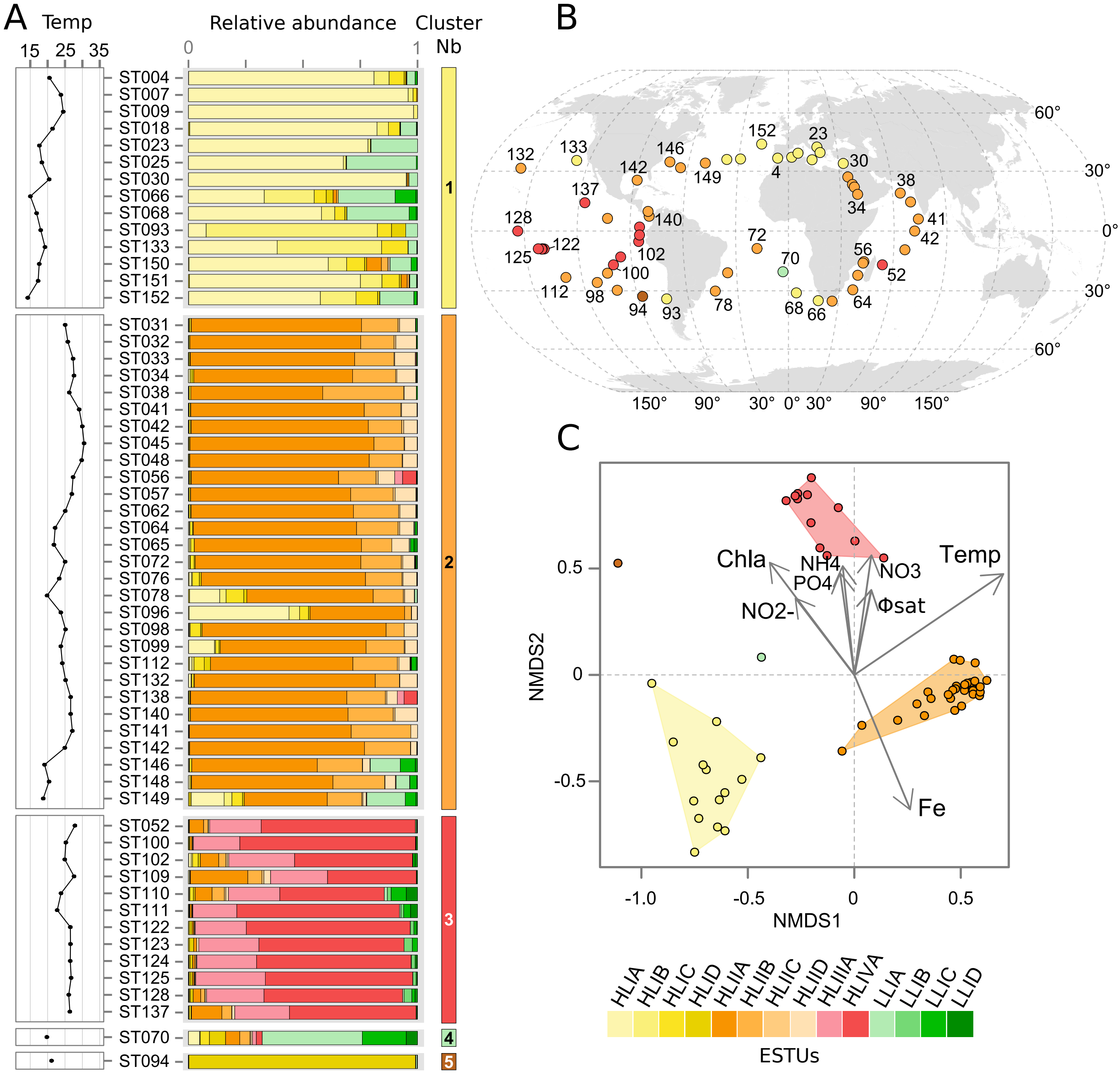
Dataset 3: Summary data for 16S rRNA reference sequences used in this study.

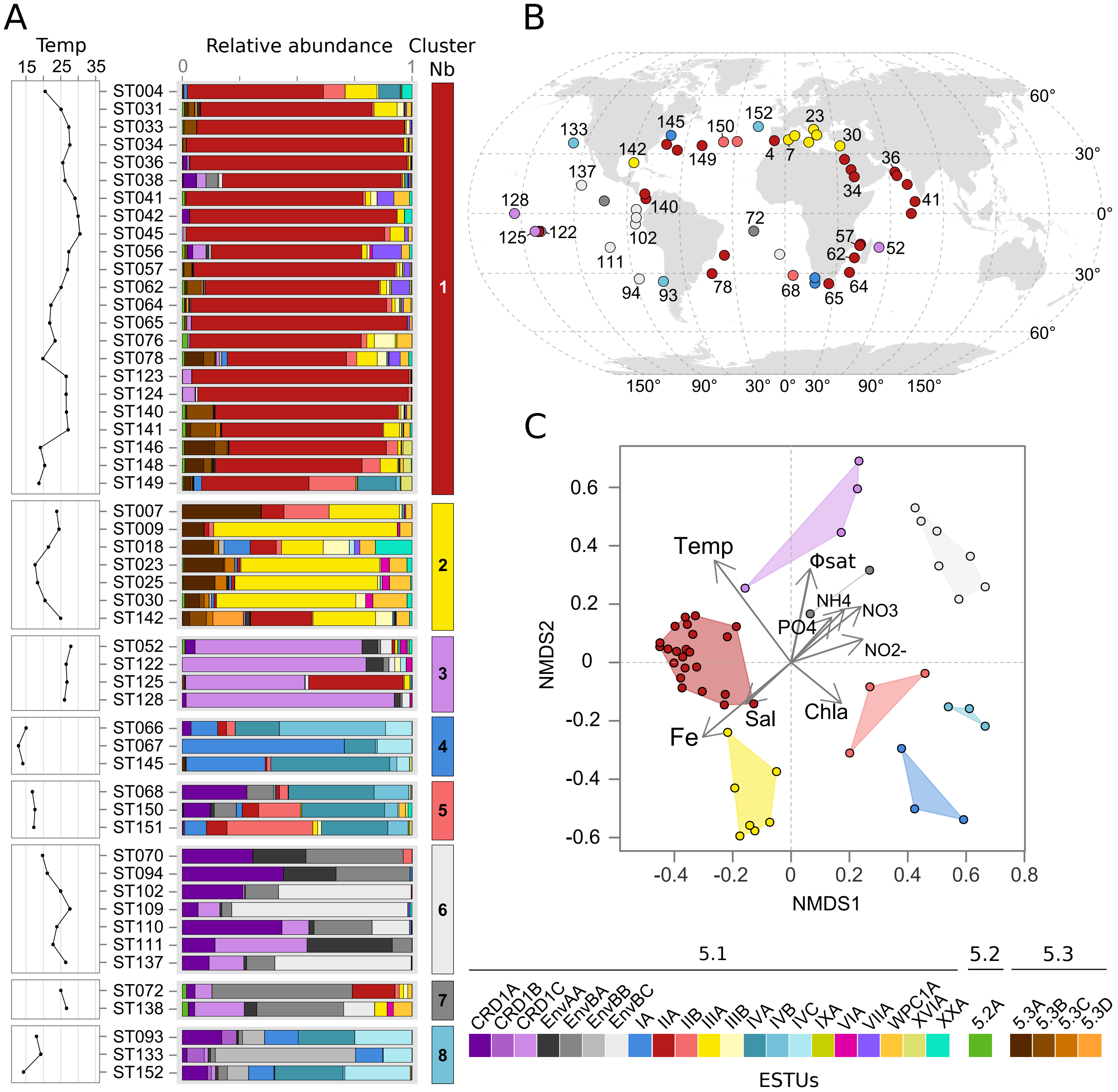
Dataset 4: *Tara* Oceans sample description including the number of recruited *petB* reads per station. Iron and ammonium concentrations were simulated using the ECCO2-Darwin model and an independent parameter to assess iron limitation (Φsat) was obtained using Behrenfeld et al.'s formula (57) applied to monthly averaged satellite data (AMODIS chl_ocx, nflh and ipar) retrieved from the NASA website (http://oceandata.sci.gsfc.nasa.gov/) for each station and corresponding sampling date. Other environmental parameters measured during the Tara Oceans expedition and the methods used to acquire them are available at www.pangea.de.

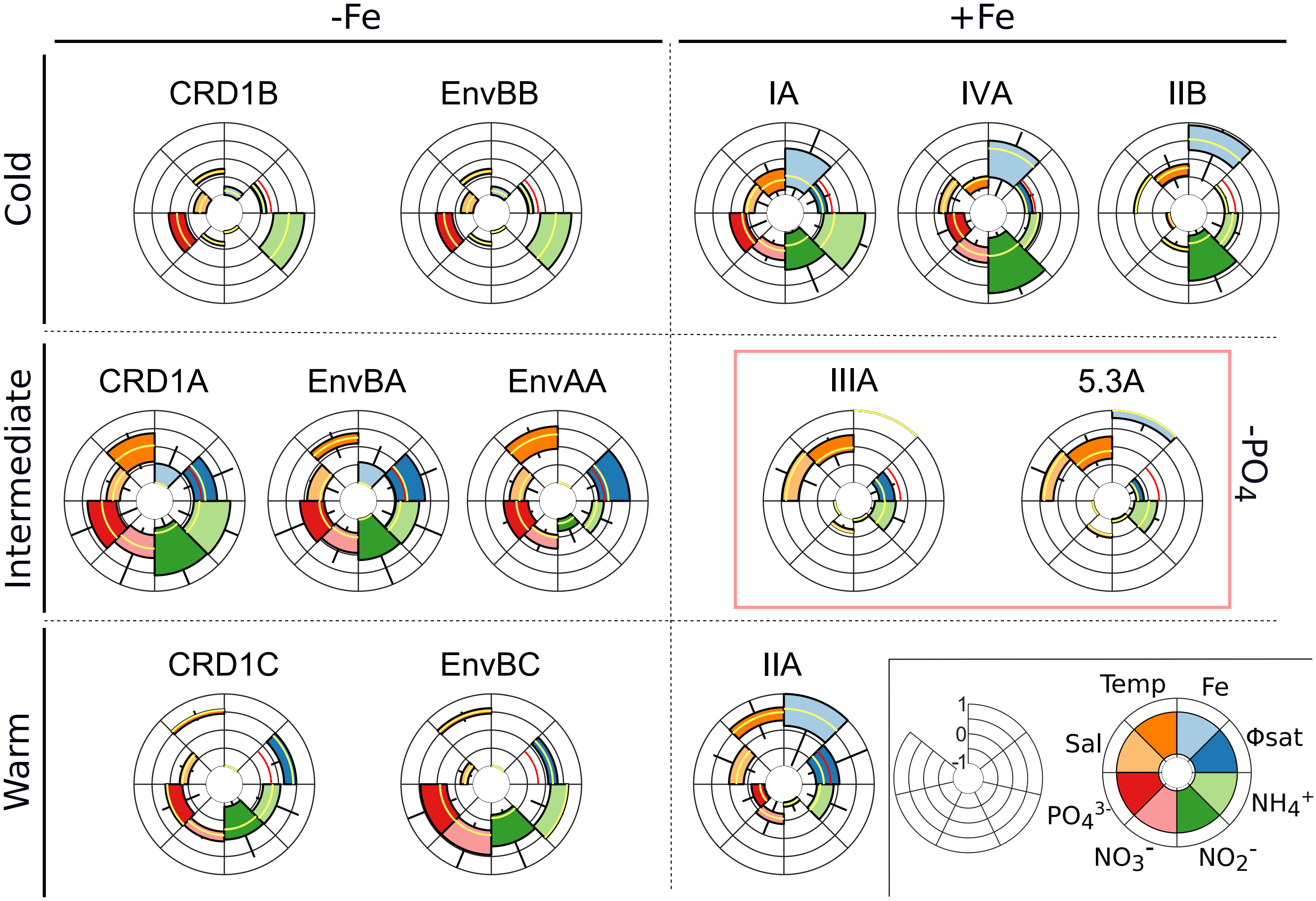
Dataset 5: Sequence names of the members of each Operational Taxonomical Unit (OTU) defined for *petB* at 94% nucleotide sequence identity.


Prochlorococcus
Synechococcus
HLI
HLII
HLIII
HLIV
LLII-III
LLI
LLIV
HLII
HLI
HLIII
HLIV
LLII-III
LLI
LLIV
5.3
5.2
VIII
IX
XVI
WPC1
XX
UC-A
II
EnvB
III
IV
EnvA
VI
V
VII
CRD1
I
Ia
Ib
Ic
IIa
IIb
IIc
IId
IIe
IIf
IIg
IIh
II-WPC2
IIIa
IIIb
IIIc
IVa
IVb
V
VIa
VIb
VIc
VIIa
VIIb
VIII
IX
XVI
XX
EnvA1
EnvA2
EnvB1
EnvB2
EnvB3
CRD1
UC-A
WPC1a
WPC1b
5.1
5.2
5.2a
5.2b
5.3
5.3a
5.3b
5.3c
5.3d
0.05


A
Nb of recruited reads
100000
0
Pro
Syn
Others
2
1
60
70
80
90
100
% identity of best hit
B
II
CRD1
EnvB
6000
1500
900
0
60
80
100
C
Clustered stations
OTU003
OTU058
OTU035
OTU012
OTU090
OTU055
OTU041
OTU271
OTU108
OTU097
OTU083
OTU096
OTU098
IIA
OTU147
OTU021
OTU064
OTU031
IIB
OTU004
OTU001
OTU061
IIIA
OTU039
OTU068
OTU046
OTU279
IIIB
OTU059
OTU007
IVA
OTU062
IVB
OTU005
IVC
0
1
Relative abundance
OTU022
CRD1A
OTU272
CRD1B
OTU047
OTU050
OTU018
OTU051
CRD1C
OTU028
EnvBA
OTU060
EnvBC
OTU027
EnvBB
OTU067
5.3A
OTU130
5.3B
OTU069
5.3D
OTU066
5.3C


A
Temp
15
25
35
Relative abundance
0
1
Cluster
Nb
ST004
ST007
ST009
ST018
ST023
ST025
ST030
ST066
ST068
ST093
ST133
ST150
ST151
ST152
1
ST031
ST032
ST033
ST034
ST038
ST041
ST042
ST045
ST048
ST056
ST057
ST062
ST064
ST065
ST072
ST076
ST078
ST096
ST098
ST099
ST112
ST132
ST138
ST140
ST141
ST142
ST146
ST148
ST149
2
ST052
ST100
ST102
ST109
ST110
ST111
ST122
ST123
ST124
ST125
ST128
ST137
3
ST070
4
ST094
5
B
60°
30°
0°
30°
60°
150°
90°
30°
0°
30°
90°
150°
132
133
146
142
152
23
4
149
30
137
38
34
128
140
41
0°
42
122
125
102
72
56
70
52
100
112
98
94
93
78
68
66
64
C
NMDS2
NMDS1
0.5
0
-0.5
-1.0
-0.5
0
0.5
Chla
NH4
PO4
NO3
Temp
NO2-
Φsat
Fe
HLIA
HLIB
HLIC
HLID
HLIIA
HLIIB
HLIIC
HLIID
HLIIIA
HLIVA
LLIA
LLIB
LLIC
LLID
ESTUs


A
Temp
15 25 35
Relative abundance
0
1
Cluster Nb
ST004
ST031
ST033
ST034
ST036
ST038
ST041
ST042
ST045
ST056
ST057
ST062
ST064
ST065
ST076
ST078
ST123
ST124
ST140
ST141
ST146
ST148
ST149
1
ST007
ST009
ST018
ST023
ST025
ST030
ST142
2
ST052
ST122
ST125
ST128
3
ST066
ST067
ST145
4
ST068
ST150
ST151
5
ST070
ST094
ST102
ST109
ST110
ST111
ST137
6
ST072
ST138
7
ST093
ST133
ST152
8
B
60°
30°
0°
30°
60°
150° 90° 30° 0° 30° 90° 150°
133
145
150
152
23
30
142
149
4
7
36
137
34
128
140
41
72
125
122
102
57
52
111
62
94
93
78
68
64
65
C
Temp
Φsat
NH4
NO3
PO4
NO2-
Sal
Chla
Fe
NMDS2
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.4 -0.2 0 0.2 0.4 0.6 0.8
NMDS1
5.1
5.2
5.3
CRD1A
CRD1B
CRD1C
EnvAA
EnvBA
EnvBB
EnvBC
IA
IIA
IIB
IIIA
IIIB
IVA
IVB
IVC
IXA
VIA
VIIA
WPC1A
XVIA
XXA
5.2A
5.3A
5.3B
5.3C
5.3D
ESTUs


-Fe
+Fe
Cold
Intermediate
Warm
CRD1B
EnvBB
IA
IVA
IIB
CRD1A
EnvBA
EnvAA
IIIA
5.3A
-PO4
CRD1C
EnvBC
IIA
1
0
-1
Temp
Fe
Sal
Φsat
NH4+
PO43-
NO3-
NO2-