Light color acclimation is a key process in the global ocean distribution of Synechococcus cyanobacteria

Théophile Grébert, Hugo Doré, Frédéric Partensky, Gregory Farrant, Emmanuel Boss, Marc Picheral, Lionel Guidi, Stéphane Pesant, David Scanlan, Patrick Wincker, et al.

To cite this version:

Théophile Grébert, Hugo Doré, Frédéric Partensky, Gregory Farrant, Emmanuel Boss, et al.. Light color acclimation is a key process in the global ocean distribution of Synechococcus cyanobacteria. Proceedings of the National Academy of Sciences of the United States of America, National Academy of Sciences, 2018, 115 (9), pp.E2010-E2019. <10.1073/pnas.1717069115>. <hal-01772758>

HAL Id: hal-01772758
https://hal.sorbonne-universite.fr/hal-01772758
Submitted on 20 Apr 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.
Light color acclimation: a key process in the global ocean distribution of *Synechococcus* cyanobacteria


*Sorbonne Universités-Université Paris 06 & Centre National de la Recherche Scientifique (CNRS), UMR 7144, Marine Phototrophic Prokaryotes Team, Station Biologique, CS 90074, 29688 Roscoff cedex, France; †Maine In-situ Sound and Color Lab, University of Maine, Orono, ME 04469; ªSorbonne Universités-Université Paris 06 & CNRS, UMR 7093, Observatoire océanologique, 06230 Villefranche-sur-mer, France; ‡PANGAEA, Data Publisher for Earth and Environmental Science, University of Bremen, Bremen, Germany; §MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany; †University of Warwick, School of Life Sciences, Coventry CV4 7AL, UK; ¶Commissariat à l’Energie Atomique et aux Energies Alternatives (CEA), Institut de Génomique, Genoscope, 91057 Evry, France; †Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM), Consejo Superior de Investigaciones Científicas (CSIC), Barcelona ES-08003, Spain; ‡Department of Biology, Indiana University, Bloomington, IN 47405.

†Present address: Food Safety, Environment, and Genetics, Matís Ltd., 113 Reykjavík, Iceland.

‡To whom correspondence should be addressed. Email: laurence.garczarek@sb-roscoff.fr.

Classification: Biological Sciences/Environmental Sciences

Keywords: marine cyanobacteria, metagenomics, light quality, phycobilisome, *Tara* Oceans

Short title: Biogeography of *Synechococcus* pigment types
**Abstract**

Marine *Synechococcus* cyanobacteria are major contributors to global oceanic primary production and exhibit a unique diversity of photosynthetic pigments, allowing them to exploit a wide range of light niches. However, the relationship between pigment content and niche partitioning has remained largely undetermined due to the lack of a single-genetic marker resolving all pigment types (PTs). Here, we developed and employed a novel and robust method based on three distinct marker genes (*cpcBA*, *mpeBA* and *mpeW*) to estimate the relative abundance of all known *Synechococcus* PTs from metagenomes. Analysis of the *Tara Oceans* dataset allowed us, for the first time, to reveal the global distribution of *Synechococcus* PTs and to define their environmental niches. Green-light specialists (PT 3a) dominated in warm, green equatorial waters, whereas blue-light specialists (PT 3c) were particularly abundant in oligotrophic areas. Type IV chromatic acclimaters (CA4-A/B), which are able to dynamically modify their light absorption properties to maximally absorb green or blue light, were unexpectedly the most abundant PT in our dataset and predominated at depth and high latitudes. We also identified populations in which CA4 might be nonfunctional due to the lack of specific CA4 genes, notably in warm high-nutrient low-chlorophyll areas. Major ecotypes within clades I-IV and CRD1 were preferentially associated with a particular PT, while others exhibited a wide range of PTs. Altogether, this study provides important insights into the ecology of *Synechococcus* and highlights the complex interactions between vertical phylogeny, pigmentation and environmental parameters that shape *Synechococcus* community structure and evolution.

**Significance Statement**

Understanding the functional diversity of specific microbial groups at the global scale is critical yet poorly developed. By combining the considerable knowledge accumulated through recent years on the molecular bases of photosynthetic pigment diversity in marine *Synechococcus*, a major phytoplanktonic organism, with the wealth of metagenomic data provided by the *Tara Oceans*
expedition, we have been able to reliably quantify all known pigment types along its transect and provide the first global distribution map. Unexpectedly, cells able to dynamically change their pigment content to match the ambient light color were ubiquitous and predominated in many environments. Altogether, our results unveiled the role of adaptation to light quality on niche partitioning in a key primary producer.
Introduction

Marine *Synechococcus* is the second most abundant phytoplankton group in the world’s oceans and constitutes a major contributor to global primary production and carbon cycling (1, 2). This genus displays a wide genetic diversity and several studies have shown that among the ~20 clades defined based on various genetic markers, five (clades I-IV and CRD1) predominate *in situ* and can be broadly associated with distinct sets of physico-chemical parameters (3–5). In a recent study, we further defined Ecologically Significant Taxonomic Units (ESTUs), i.e. organisms belonging to the same clade and co-occurring in the field, and highlighted that the three main parameters affecting the *in situ* distribution of these ESTUs were temperature and availability of iron and phosphorus (6). Yet, marine *Synechococcus* also display a wide pigment diversity, suggesting that light could also influence their ecological distribution, both qualitatively and quantitatively (7, 8).

This pigment diversity comes from differences in the composition of their main light-harvesting antennae, called phycobilisomes (PBS; 7–9). These water-soluble macromolecular complexes consist of a central core anchoring at least six radiating rods made of several distinct phycobiliproteins, i.e. proteins to which specific enzymes (phycobilin lyases) covalently attach chromophores called phycobilins (7, 10). Although the PBS core is conserved in all marine *Synechococcus*, rods have a very variable composition, and three main pigment types (PTs) are usually distinguished (Fig. S1; 7, 11). In PT 1, PBS rods are solely made of phycocyanin (PC, encoded by the *cpcBA* operon) and bear the red-light absorbing phycocyanobilin (PCB; \( A_{\text{max}} = 620 \text{ nm} \)) as the sole chromophore. In PT 2, rods are made of PC and phycoerythrin I (PE-I, encoded by *cpeBA*) and attach both PCB and the green-light absorbing phycoerythrobilin (PEB; \( A_{\text{max}} = 550 \text{ nm} \)). All other marine *Synechococcus* belong to PT 3 and have rods made of PC, PE-I and PE-II (encoded by *mpeBA*) that bind PCB, PEB and the blue-light absorbing phycourobilin (PUB; \( A_{\text{max}} = 495 \text{ nm} \); Fig. S1). Several subtypes can be defined within PT 3, based on the fluorescence excitation ratio at 495 nm and 545 nm (hereafter \( \text{Ex}_{495:545} \); Fig. S1), a proxy for the PUB:PEB ratio. This ratio is low (\( \text{Ex}_{495:545} < 0.6 \)) in subtype 3a (green light specialists),
intermediate in subtype 3b (0.6 ≤ Ex_{495:545} < 1.6) and high (Ex_{495:545} ≥ 1.6) in subtype 3c (blue light specialists; 7, 11). Additionally, strains of subtype 3d are able to change their PUB:PEB ratio depending on ambient light color, a process called type IV chromatic acclimation (hereafter CA4), allowing them to maximally absorb blue or green light (11–14). Comparative genomic analyses showed that genes involved in the synthesis and regulation of PBS rods are gathered into a dedicated genomic region, the content and organization of which correspond to the different PTs (7). Similarly, chromatic acclimation has been correlated with the presence of a small specific genomic island (CA4 genomic island) that exists in two distinct configurations (CA4-A and -B; 11). Both contain two regulators (fciA and fciB) and a phycobilin lyase (mpeZ in CA4-A or mpeW in CA4-B), thus defining two distinct CA4 genotypes: 3dA and 3dB (11, 14, 15). Finally, some strains possess a complete or partial CA4 genomic island but are not able to perform CA4, displaying a fixed Ex_{495:545} corresponding to 3a, 3b or 3c phenotypes (11).

As there is no correspondence between pigmentation and core genome phylogeny (7, 16, 17), deciphering the relative abundance and niche partitioning of Synechococcus PTs in the environment requires specific approaches. In the past 30 years, studies have been based either on i) proxies of the PUB:PEB ratio as assessed by flow cytometry (18–20), fluorescence excitation spectra (21–27), epifluorescence microscopy (28), or ii) phylogenetic analyses of cpcBA or cpeBA (17, 29–34). These studies showed that PT 1 is restricted to and dominates in low salinity surface waters and/or estuaries, which are characterized by a high turbidity resulting in a red wavelengths-dominated light field (18, 22, 31–38), whereas PT 2 is found in coastal shelf waters or in the transition zones between brackish and oceanic environments with intermediate optical properties (18, 27, 34, 36–39). Finally, PT 3 with increasing PUB:PEB ratio are found over gradients from onshore mesotrophic waters, characterized by green light dominance, to offshore oligotrophic waters, where blue light penetrates the deepest (19–24, 28, 36, 38, 40). Some authors reported an increase in the PUB:PEB ratio with depth (19, 21, 24), while others observed a constant ratio throughout the water column, a variability
potentially linked to the location, water column features and/or environmental parameters (22, 25, 28).

However, these analyses based on optical properties could only describe the distribution of high- and low-PUB populations without being able to differentiate green (3a) or blue light (3c) specialists from CA4 cells (3d) acclimated to green or blue light, while genetic analysis solely based on cpcBA and/or cpeBA could not differentiate all PTs. For instance, only two studies have reported CA4 populations in situ either in the western English Channel (17) or in sub-polar waters of the western Pacific Ocean (29) but none of them were able to differentiate CA4-B from high PUB (i.e. 3c) populations. As a consequence, the global relative abundance of the different Synechococcus PTs, particularly CA4, and the link between genetic and pigment diversity have remained largely unclear.

Here, we analyzed 109 metagenomic samples collected from all major oceanic basins during the 2.5-yr Tara Oceans (2009-2011) expedition (41) using a bioinformatic pipeline combining a metagenomic read recruitment approach (6, 42) to recruit single reads from multiple PBS gene markers and placement of these reads in reference trees to assign them to a given PT. This pipeline allowed the first description of the worldwide distribution of all known Synechococcus PTs, as well as of their realized environmental niches (sensu 43). This study provides a synoptic view of how a major photosynthetic organism adapts to natural light color gradients in the ocean.
Results

A novel, robust approach for estimating pigment types abundance from metagenomes

We developed a multi-marker approach combining phylogenetic information retrieved from three different genes or operons (cpcBA, mpeBA and mpeW; Fig. 1 and Datasets 1-2) to overcome the issue of fully resolving the whole range of PTs. While cpcBA discriminated PT 1, 2 and 3 (Fig. 1A), only the mpeBA operon, a PT 3 specific marker, was able to distinguish the different PT 3 subtypes (Fig. 1B), though as for cpeBA it could not differentiate PT 3dB (CA4-B) from PT 3c (i.e. blue light specialists; 11, 29). The mpeW marker was thus selected to specifically target PT 3dB and, by subtraction, enumerate PT 3c (Fig. 1C). Using the cpcBA marker, members of PT 2 were split into two well-defined clusters, 2A and 2B (Fig. 1A), the latter corresponding to a purely environmental PT identified from assembled metagenomes of the Baltic Sea (38). Strains KORDI-100 and CC9616 also clustered apart from other strains in the mpeBA phylogeny, suggesting that they have a divergent evolutionary history from other PT 3 members (Fig. 1B). This is supported by the diverged gene content and order of their PBS rod genomic region and these strains were recently referred to as PT 3f, even though they have a similar phenotype as PT 3c (Ex495:545 ratio ≥ 1.6; 30). To investigate the phylogenetic resolution of small fragments of these three markers, sequences were removed one at a time from the reference database, and simulated reads (150 bp long as compared to 164 bp in average for Tara Oceans cleaned/merged reads) generated from this sequence were assigned using our bioinformatic pipeline against a database comprising the remaining sequences. Inferred and known PTs were then compared. The percentage of simulated reads assigned to the correct PT was between 93.2% and 97.0% for all three markers, with less than 2.1-5.6% of reads that could not be classified and an error-rate below 2%, showing that all three markers display a sufficient resolution to reliably assign the different PTs (Fig. S2B, D and F).

To ensure that the different markers could be quantitatively compared in a real dataset, we examined the correlations between estimates of PT abundances using the different markers in the
109 metagenomes analyzed in this study. Total cpcBA counts were highly correlated \( R^2=0.994, n=109; \) Fig. S3A with total Synechococcus counts obtained with the petB gene, which was previously used to study the phylogeography of marine picocyanobacteria (6), and the correlation slope was not significantly different from 1 (slope: 1.040; Wilcoxon’s paired difference test p-value=0.356). cpcBA is thus as good as petB at capturing the total population of Synechococcus reads. Moreover, counts of cpcBA reads assigned to PT 3 and total mpeBA counts (specific for PT 3) were also strongly correlated \( R^2=0.996, n=109; \) Fig. S3B, and not skewed from 1 (slope of 0.991, Wilcoxon’s p-value=0.607), indicating that mpeBA and cpcBA counts can be directly compared. Although no redundant information for PT 3dB is available with the three selected markers, another marker targeting 3dB (fciAB) was tested and produced results similar to mpeW (Fig. S3C). These results demonstrate that our multi-marker approach can be used to reliably and quantitatively infer the different Synechococcus PTs from short metagenomic reads, with PT 1, 2A, 2B abundances being assessed by cpcBA normalized counts, PT 3a, 3f and 3dA by mpeBA normalized counts, PT 3dB by mpeW normalized counts and PT 3c by the difference between mpeBA normalized counts for 3c + 3dB and mpeW normalized counts. We thus used this approach on the Tara Oceans metagenomes, generated from 109 samples collected at 65 stations located in the major oceanic basins (Fig. 2).

CA4 populations are widespread and predominate at depth and high latitudes

The latitudinal distribution of Synechococcus inferred from cpcBA counts is globally consistent with previous studies (2, 6, 44), with Synechococcus being present in most oceanic waters, but quasi absent (< 20 cpcBA counts) beyond 60°S (Southern Ocean stations TARA_082 to TARA_085; Fig. 2B). Overall, the number of recruited cpcBA reads per station was between 0 and 8,151 (n=63, median: 449, mean: 924, sd: 1478) for surface and 0 and 3,200 (n=46, median:170, mean: 446, sd: 664) for deep chlorophyll maximum (DCM) samples, respectively. Stations with less than 30 cpcBA reads were excluded from further analysis.
PT 1 and 2, being both known to be mostly found and abundant in coastal waters (29, 36, 38, 45), were expectedly almost absent from this dataset (total of 15 and 513 cpcBA reads, respectively; Fig. 2A-B) since the Tara cruise sampling was principally performed in oceanic waters. While PT 2A was mostly found at the surface at one station off Panama (TARA_141, 417 out of 6,637 reads at this station; Fig. 2B), PT 2B was virtually absent (total of 3 cpcBA reads) from our dataset and might thus be confined to the Baltic Sea (38). This low abundance of PT 1 and 2 precluded the correlation analysis between their distribution and physico-chemical parameters. PT 3 was by far the most abundant along the Tara Oceans transect, accounting for 99.1 ± 1.4% (mean ± sd) of cpcBA reads at stations with ≥30 cpcBA read counts. Interestingly, several PT 3 subtypes often co-occurred at a given station.

PT 3a (green light specialists) totaled 20.3% of read counts, with similar abundance between surface (20.5%) and DCM (19.4%) samples, and was particularly abundant in intertropical oceanic borders and regional seas, including the Red Sea, the Arabian Sea and the Panama/Gulf of Mexico area (Fig. 2B). Correlation analyses show that this PT is consistently associated with high temperatures but also with greenish (as estimated from a low blue to green downwelling irradiance ratio, $I_{495:545}$), particle-rich waters (high particle backscattering at 470 nm and beam attenuation coefficient at 660 nm; Fig. 3). Still, in contrast with previous studies that reported the distribution of low-PUB populations (19, 21, 23, 24, 26, 27), this PT does not seem to be restricted to coastal waters, explaining its absence of correlation with chlorophyll concentration and colored dissolved organic matter (cDOM).

Blue light specialists (PT 3c) appear to be globally widespread, with the exception of high latitude North Atlantic waters, and accounted for 33.4% of reads, with a higher relative abundance at the surface (36.8%) than at the DCM (23.3%, Fig. 2A). This PT is dominant in transparent, oligotrophic, iron-replete areas such as the Mediterranean Sea as well as South Atlantic and Indian Ocean gyres (Figs. 2B and 4C). In the South Pacific, PT 3c was also found to be predominant in the
Marquesas Islands area (TARA_123 and 124), where the coast proximity induced a local iron enrichment (6). Consistently, PT 3c was found to be positively associated with iron concentration, high temperature and DCM depth and anti-correlated with chlorophyll fluorescence, nitrogen concentrations, net primary production (NPP) as well as other related optical parameters, such as backscattering at 470 nm and beam attenuation coefficient at 660 nm (Fig. 3). Despite its rarity, PT 3f seems to thrive in a similar environment, with the highest relative abundances in the Indian Ocean and Mediterranean Sea (Figs. 2B and 4C). Its occurrence in the latter area might explain its strong anti-correlation with phosphorus availability.

Both CA4 types, 3dA and 3dB, which represented 22.6% and 18.9% of reads respectively, were unexpectedly widespread and could locally account for up to 95% of the total Synechococcus population (Figs. 2, 4C and S4). In contrast to blue and green light specialists, both CA4 types were proportionally less abundant at the surface (19.8% and 17.5%, for 3dA and 3dB, respectively) than at depth (30.9% and 22.9%). Interestingly, PT 3dA and 3dB generally displayed complementary distributions along the Tara Oceans transect (Fig. 2B). PT 3dA was predominant at high latitude in the northern hemisphere as well as in other vertically mixed waters such as in the Chilean upwelling (TARA_093) or in the Agulhas current (TARA_066 and 68; Fig 2B). Accordingly, PT 3dA distribution seems to be driven by low temperature, high nutrient and highly productive waters (high NPP, chlorophyll a and optical parameters), a combination of physico-chemical parameters almost opposite to those observed for blue light specialists (PT 3c; Fig. 3). In contrast, PT 3dB shares a number of characteristics with PT 3c, including the anti-correlation with nitrogen concentration and association with iron availability (as indicated by both a positive correlation with [Fe] and negative correlation with the iron limitation proxy Φsat; Fig. 3), consistent with their widespread occurrence in iron replete oceanic areas. Also noteworthy, PT 3dB was one of the sole PT (with 3f) to be associated with low photosynthetically available radiation (PAR).
Niche partitioning of *Synechococcus* populations rely on a subtle combination of ESTU and PT niches

We previously showed that temperature, iron and phosphorus availability constituted major factors influencing the diversification and niche partitioning of *Synechococcus* ESTUs (i.e. genetically related subgroups within clades that co-occur in the field; 6). Yet, these results cannot be extended to PTs since the pigment content does not follow the vertical phylogeny (7). In order to decipher the respective roles of genetic and pigment diversity in *Synechococcus* community structure, we examined the relationships between ESTUs and PTs *in situ* abundances through correlation and NMDS analyses (Fig. 4A-B) and compared their respective distributions (Figs. 4C and S4).

Interestingly, all PTs are either preferentially associated with or excluded from a subset of ESTUs. PT 2A is found at low abundance at a few stations along the Tara Oceans transect and, when present, it is seemingly associated with the rare ESTU 5.3B (Fig. 4A), an unusual PT/ESTU combination so far only observed in metagenomes from freshwater reservoirs (46). PT 3a is associated with ESTUs EnvBC (occurring in low iron areas) and IIa, the major ESTU in the global ocean (Fig. 4A), a result consistent with NMDS analysis, which shows that PT 3a is found in assemblages dominated by these two ESTUs (indicated by red and grey backgrounds in Fig. 4B), as well as with independent observations on cultured strains (Dataset 3). PT 3c is associated with ESTU IIIA (the dominant ESTU in P-depleted areas), as observed on many isolates (Dataset 3), and is also linked, like PT 3f, with ESTUs IIIB and WPC1A, both present at lower abundance than IIIA in P-poor waters (Fig. 4A). PT 3f is also associated with the newly described and low-abundance ESTU XXA (previously EnvC; Fig. S5; 4, 6). Both PT 3f and ESTU XXA were rare in our dataset but systematically co-occurred, in agreement with the fact that the only culture representative of the latter clade belongs to PT 3f (Dataset 3).

PT 3dA appears to be associated with all ESTUs from clades CRD1 (specific to iron-depleted areas) as well as with those representative of coastal and cold waters (IA, IVA, IVC), but is anti-correlated with most other major ESTUs (IIA, IIIA and ~B, WPC1A and 5.3B; Fig. 4A). This pattern is
opposite to PT 3dB that is preferentially found associated with ESTU IIA, IIB and 5.3A, but not in CRD1A or –C (Fig. 4A). Thus, it seems that the two types of CA4 are found in distinct and complementary sets of ESTUs. Interestingly, our analysis might suggest the occurrence of additional PTs not isolated so far, since a number of reads (0.7% and 2.7% of cpcBA and mpeBA counts, respectively, Fig. 2A) could not be assigned to any known PTs. For instance, while most CRD1C seem preferentially associated with PT 3dA, a fraction of the population could only be assigned at the PT 3 level (Fig. 4A). Similarly, a number of reads could not be assigned to any known PT in stations rich in ESTU 5.3A and XXA, although one cannot exclude that this observation might be due to a low number of representative strains, and thus PT reference sequences, for these ESTUs.

The preferred association of PTs with specific ESTUs is also well illustrated by some concomitant shifts of PTs and ESTU assemblages. For instance, in the wintertime North Atlantic Ocean, the shift from 3dB-dominated stations on the western side (TARA_142 and TARA_146-149) to 3dA-dominated stations near European coasts (TARA_150 to 152) and North of Gulf stream (TARA_145) is probably related to the shift in ESTU assemblages occurring along this transect, with ESTU IIA being gradually replaced by ESTU IVA (Fig. 4C; see also 6). Similarly, the takeover of CRD1C by IIA in the Marquesas Island area (TARA_123 to 125), which is iron-enriched with regard to surrounding high-nutrient low-chlorophyll (HNLC) waters (TARA_122 and 128), perfectly matched the corresponding replacement of PT 3dA by 3c. However, in several other cases, PT shifts were not associated with a concomitant ESTU shift or vice versa. One of clearest examples of these dissociations is the transect from the Mediterranean Sea to the Indian Ocean, where the entry in the northern Red Sea through the Suez Canal triggered a sharp shift from a IIIA- to a IIA-dominated community (TARA_030 and 031), which was not accompanied by any obvious change in PTs. Conversely, a sharp rise in the relative abundance of PT 3a was observed in the southern Red Sea/northeastern Indian Ocean (TARA_033 to 038) without changes in the large dominance of ESTU IIA. Altogether, this strongly suggests that a subtle combination of ESTUs and PTs respective niche occupancy is responsible for the observed niche partitioning of *Synechococcus* populations.
Deficient chromatic acclimaters are dominant in HNLC areas

Although our results clearly indicate that CA4 cells represent a large proportion of the *Synechococcus* community in a wide range of ecological niches, this must be somewhat tempered by the fact that, in culture, about 30% of the strains possessing a CA4-A or B genomic island are not able to chromatically acclimate (Dataset 3; 11). Some of these natural mutants have an incomplete CA4 genomic island (Fig. S6K). For example, strains WH8016 (ESTU IA) and KORDI-49 (WPC1A) both lack the CA4-A specific lyase-isomerase MpeZ, an enzyme shown to bind a PUB molecule on PE-II (14), and display a green light specialist phenotype (PT 3a, $Ex_{495:545} \sim 0.4$) whatever the ambient light color (11). However, since they possess a PT 3a *mpeBA* allele, reads from field WH8016- or KORDI-49-like cells are adequately counted as PT 3a (Fig. S6K). Another CA4-deficient strain, BIOS-E4-1 (ESTU CRD1C), possesses *mpeZ* and a 3dA *mpeBA* allele but lacks the CA4 regulators FciA and FciB as well as the putative lyase MpeY and exhibits a fixed blue light specialist phenotype (PT 3c, $Ex_{495:545} \sim 1.7$; Fig. S6K; 11, 15). Thus, reads from such natural *Synechococcus* CA4-incapable mutants in the field are counted as 3dA using the *mpeBA* marker. Lastly, the strain MVIR-18-1 possesses a complete CA4-A island and a 3dA *mpeBA* allele but lacks *mpeU*, a gene necessary for blue light acclimation (Fig. S6K; 47). While MVIR-18-1 displays a fixed green light phenotype, reads from such *Synechococcus* are also erroneously counted as 3dA.

To assess the significance of these genotypes in the field, we compared the normalized read counts obtained for 3dA with *mpeBA, fciAB, mpeZ, mpeU* and *mpeY* (Fig. S6A-J). Overall this analysis revealed a high consistency between these different markers ($0.860 < R^2 < 0.986$), indicating that most *mpeZ*-containing populations also contained 3dA alleles for *fciAB, mpeY, mpeU* and *mpeBA* and are therefore likely able to perform CA4. However, a number of stations, all located in HNLC areas (TARA_094, 111 and 122 to 128 in the Pacific Ocean and TARA_052 located northwest of Madagascar, Fig. 2B), displayed more than 10-fold higher *mpeBA, mpeU* and *mpeZ* counts than *fciAB*...
and mpeY counts (Fig. S6A, B, E, F, H, I). This indicates that a large proportion or even the whole population (TARA_122 and 124) of 3dA in these HNLC areas is probably lacking the FciA/B regulators and MpeY and, like strain BIOS-E4-1 (Fig. S6K), might thus be stuck in the blue light specialist phenotype (PT 3c; 11). Conversely, station TARA_067 exhibited consistently more than twice the fciAB and mpeZ counts compared to mpeBA, mpeY or mpeU (Fig. S6B-E, G, H) and was a clear outlier when comparing pigment type and clade composition (Fig. S7). This suggests that the proportion of PT 3dA might have been underestimated at this station, as a significant proportion of this population probably corresponds to PT 3a genotypes that have acquired a CA4-A island by lateral gene transfer, as is seemingly the case for strains WH8016 and KORDI-49. Finally, no station exhibited markedly lower mpeU counts compared to all other genes, indicating that the genotype of strain MVIR-18-1 is probably rare in the oceans.

It must be noted that two out of the six sequenced CA4-B strains (WH8103 and WH8109) also have a deficient CA4 phenotype and display a constant, intermediate Ex495:545 ratio (0.7 and 1, respectively), despite any obvious PBS- or CA4-related gene deletion (11). Accordingly, the plot of 3dB normalized read counts obtained with mpeW vs. fciAB shows no clear outlier (Fig. S3C).

**Discussion**

Marine *Synechococcus* display a large pigment diversity, with different PTs preferentially harvesting distinct regions of the light spectrum. Previous studies based on optical properties or on a single genetic marker could not differentiate all PTs (17, 29–31), and thus neither assess their respective realized environmental niches (43) nor the role of light quality on the relative abundance of each PT. Here, we showed that a metagenomic read recruitment approach combining three genetic markers can be used to reliably predict all major PTs. Applied to the extensive *Tara* Oceans dataset, this original approach, which avoids PCR amplification and cloning biases, allowed us to describe for the
first time the distribution of the different *Synechococcus* PTs at the global scale and to refine our understanding of their ecology.

PT 3 was found to be largely dominant over PT 1 and 2 along the oceanic *Tara* Oceans transect, in agreement with the coastal-restricted distribution of the latter PTs (18, 22, 27, 31–34, 37–39). Biogeography and correlation analyses with environmental parameters provided several novel and important insights concerning niche partitioning of PT 3 subtypes. Green (PT 3a) and blue (PT 3c) light specialists were both shown to dominate in warm areas but display clearly distinct niches, with 3a dominating in *Synechococcus*-rich stations located on oceanic borders, while 3c predominated in purely oceanic areas where the global abundance of *Synechococcus* is low. These results are in agreement with the prevailing view of an increase in the PUB:PEB ratio from green onshore mesotrophic waters to blue offshore oligotrophic waters (19–24, 26–29, 40, 48). Similarly, we showed that PT 3dB, which could not be distinguished from PT 3c in previous studies (17, 29–31), prevails in more coastal and/or mixed temperate waters than do 3c populations. The realized environmental niche of the second type of CA4 (PT 3dA) is the best defined of all PTs as it is clearly associated with nutrient-rich waters and with the coldest stations of our dataset, occurring at high latitude, at depth and/or in vertically mixed waters (e.g., TARA_068, 093 and 133). This result is consistent with a recent study demonstrating the dominance of 3dA in sub-Arctic waters of the Northwest Pacific Ocean (29), suggesting that the prevalence of 3dA at high latitude can be generalized. The decrease of PT 3c (blue light specialists) with depth is unexpected given previous reports of a constant (22, 25, 28, 49) or increasing (19, 21, 24) PUB:PEB ratio throughout the water column. However, the high abundance of CA4 can reconcile these observations with the decreased abundance of PT 3c, as cells capable of CA4 likely have a blue-light phenotype at depth. Altogether, while little was previously known about the abundance and distribution of CA4 populations in the field, here we show that they are ubiquitous, dominate in a wide range of niches, are present both in coastal and oceanic mixed waters, and overall are the most abundant *Synechococcus* PT.
The relationship between ESTUs and PTs shows that some ESTUs are preferentially associated with only one PT, while others present a much larger pigment diversity. ESTU IIA, the most abundant and ubiquitous ESTU in the field (5, 6), displays the widest PT diversity (Fig. 4B), a finding confirmed by clade II isolates spanning the largest diversity of pigment content, with representative strains of PT 2, 3a, 3c and 3dB within this clade (Dataset 3; see also 7, 11, 50–52). This suggests that this ESTU can colonize all light color niches, an ability which might be partially responsible for its global ecological success. Our current results do not support the previously observed correlation between clade III and PT 3a (29) since the two ESTUs defined within this clade (IIIA and B) were associated with PT 3c and/or 3f. This discrepancy could be due either to the different methods used in these studies or to the occurrence of genetically distinct clade III populations in coastal areas of the northwestern Pacific Ocean and along the Tara Oceans transect. However, the pigment phenotype of strains isolated to date is more consistent with our findings (Dataset 3; 16, 36).

In contrast to most other PTs, the association between PT 3dA and ESTUs was found to be nearly exclusive in the field, as ESTUs from clades I, IV, CRD1 and EnvA were not associated with any other PT, and reciprocally PT 3dA is only associated with these clades (Fig. 4A). An interesting exception to this general rule was observed in the Benguela upwelling (TARA_067), where the dominant ESTU IA population both displays a 3a mpeBA allele and possesses fciA/B and mpeZ genes (Figs. S6K and S7), suggesting that cells, which were initially green light specialists (PT 3a), have inherited a complete CA4-A island through lateral gene transfer at this station. Interestingly, among the seven clade I strains sequenced to date, three possess a 3a mpeBA allele, among which WH8016 also has a CA4-A island but only partial (lacking mpeZ) and therefore not functional (11). It is thus difficult to conclude whether the lateral transfer of this island, likely a rare event since it was only observed in populations of the Benguela upwelling, has conferred these populations the ability to perform CA4.

Another important result of this study was the unsuspected importance of populations that have likely lost the ability to chromatically acclimate, specifically in warm HNLC areas, which cover
wide expanses of the South Pacific Ocean (53). Interestingly, populations living in these ultra-
oligotrophic environments have a different genetic basis for their consistently elevated PUB
phenotype than do typical blue light specialists (i.e. PT 3c), since they have lost the CA4 regulators
fciA/B and accumulated mutations in mpeY, a yet uncharacterized member of the phycobilin lyase
family, as observed in strain BIOS-E4-1 (Fig. S6K; 11). This finding, consistent with the previous
observation that the south Pacific is dominated by high-PUB Synechococcus (22), is further supported
by the recent sequencing of three isolates from the Equatorial Pacific, strains MITS9504, MITS9509
(both CRD1C) and MITS9508 (CRD1A; 54), all of which contain, like BIOS-E4-1, a 3dA mpeBA allele, a
CA4-A island lacking fciA/B and a partial (MITS9508) or highly degenerated (2 other MIT strains)
mpeY gene sequence (Fig. S6K). Thus, these natural CA4-A mutants seem to have adapted to blue,
ultra-oligotrophic waters by inactivating a likely energetically costly acclimation mechanism (positive
selection), although we cannot exclude that it might be a consequence of the lower selection
efficiency associated to the reduced effective population size of Synechococcus in such an extreme
environment (genetic drift). If, as we hypothesize, all Synechococcus cells counted as 3dA at these
stations are CA4-deficient, these natural mutants would represent about 15% of the total 3dA
population. In contrast, CRD1-A populations of the eastern border of the Pacific Ocean (TARA_102,
109-110, 137) are likely true CA4 populations as they possess all CA4 genes (Fig. S6K).

In conclusion, our study provided novel insights into the distribution, ecology and adaptive value
of all known Synechococcus PTs. Unexpectedly, the sum of 3dA and 3dB constituted about 40% of the
total Synechococcus counts in the Tara Oceans dataset, making chromatic acclimatizers (PT 3d) the
most globally abundant PT, even when taking into account potential CA4-deficient natural mutants.
In addition, this PT made up 95% of the Synechococcus population at high latitudes and was present
in every one of the five major clades in the field (I, II, III, IV and CRD1). This suggests that chromatic
acclimation likely confers a strong adaptive advantage compared to strains with a fixed
pigmentation, particularly in vertically mixed environments and at depth at stations with a stratified
water column. The occurrence of natural CA4 mutants and evidence for lateral transfer of the CA4
Genomic island further support previous hypotheses that not only temperature and nutrient availability (3, 5, 6) but also light quality (7, 52) co-exert selective pressures affecting marine Synechococcus evolution. Thus, changes in pigment diversity could occur in response to changes in light niches by acquisition or loss of specific PBS synthesis and/or regulation genes, as previously observed for phosphorus and nitrogen transport genes in Prochlorococcus (55–57). Still, the complex interactions between PTs, vertical phylogeny and environmental parameters remain unclear and more work is needed to refine our understanding of the balance between the forces shaping community composition and Synechococcus evolution. At the boundaries of Synechococcus environmental niche(s), where the harshest conditions are encountered, both pigment and clade diversity are drastically reduced, and this concomitant reduction tends to support a co-selection by light quality and other environmental parameters. On the contrary, the diverse PTs occurring within some clades, as well as the co-occurrence of different PTs at most stations compared to more clear-cut clade shifts (e.g., in the Red Sea/Indian Ocean) might indicate that light quality is not the strongest selective force or that light changes are too transient to allow the dominance and fixation of a particular PT in a population. Future experimental work exploring the fitness of distinct ESTU/PT combinations under different controlled environmental conditions (including temperature, nutrients and light) might help clarifying the respective effects of these parameters on the diversification of this ecologically important photosynthetic organism.

**Materials and Methods**

**Metagenomic samples**

This study focused on 109 metagenomic samples corresponding to 65 stations from the worldwide oceans collected during the 2.5-yr Tara Oceans circumnavigation (2009-2011). Water sample and sequence processing are the same than in (6). Dataset 4 describes all metagenomic samples with
location and sequencing effort. Sequencing depths ranged from $16 \times 10^6$ to $258 \times 10^6$ reads per sample after quality control and paired-reads merging, and corresponding fragments lengths averaged $164 \pm 20$ bp (median: $168$ bp).

Databases: reference and outgroup sequences

A reference database comprising the full-length gene or operon nucleotide sequences was generated for each marker used in this study (cpcBA, mpeBA and mpeW) based on culture isolates with characterized pigment type (Dataset 1). These databases comprised 83 cpcBA sequences (64 unique), including 18 PT 1, 5 PT 2A, 19 PT 2B and 39 PT 3, 41 mpeBA sequences (all unique), including 11 PT 3a, 2 PT 3f, 11 PT 3dA and 17 PT 3dB and 5 unique mpeW sequences. For each marker, a reference alignment was generated with MAFFT L-INS-i v6.953b (58), and a reference phylogenetic tree was inferred with PhyML v. 20120412 (GTR+I+G, 10 random starting trees, best of SPR and NNI moves, 500 bootstraps; (59) and drawn using the ETE Toolkit (60).

A database of outgroups was also built, comprising paralogous sequences from marine Synechococcus or Prochlorococcus as well as orthologous sequences from other marine and freshwater organisms retrieved from public databases. For cpcBA and mpeBA, the outgroup databases comprised apcA, apcB, apcD, apcF and cpeBA from marine Synechococcus, ppeBA from Prochlorococcus, cpcBA and cpeBA from other non-picocyanobacterial organisms as well as either mpeBA or cpcBA from marine Synechococcus, respectively (Datasets 1-2). For mpeW, the outgroup database was made of paralogous genes (mpeZ, mpeY and cpeY) from marine Synechococcus or Prochlorococcus, as no ortholog could be identified in public databases. Similarly, for mpeY and mpeZ, the outgroup database comprised cpeY, mpeW as well as mpeZ or mpeY, respectively. The outgroup database for mpeU comprised cpeF paralogous sequences from marine Synechococcus and Prochlorococcus. No outgroup database was used for fciAB, as no paralogs or other distantly related sequences were found either in marine Synechococcus and Prochlorococcus or in public databases.
Read assignment and estimation of PT abundance

Reads were preselected using BLAST+ (61) with relaxed parameters (blastn, maximum E-value of 1e-5, minimum percent identity 60%, minimum 75% of read length aligned), using reference sequences as subjects; the selection was then refined by a second BLAST+ round against databases of outgroups: reads with a best-hit to outgroup sequences were excluded from downstream analysis. Selected reads were then aligned to the marker reference alignment with MAFFT v.7.299b (--addfragments --adjustdirectionaccurately) and placed in the marker reference phylogenetic tree with pplacer (62). For each read, pplacer returns a list of possible positions (referred to as placements) at which it can be placed in the tree and their associated “likelihood weight ratio” (LWR, proxy for the probability of the placement; see pplacer publication and documentation for more details). Reads were then assigned to a pigment type using a custom classifier written in Python. Briefly, internal nodes of the reference tree were assigned a pigment type based on the pigmentation of descending nodes (PT of child reference sequences if the same for all of them, “unclassified” otherwise). For each read, placements were assigned to their nearest ascending or descending node based on their relative position on the edge, and the lowest common ancestor (LCA) of the set of nodes for which the cumulated LWR was greater than 0.95 (LCA of possible placements at 95% probability) was then computed. Finally, the read was assigned to the pigment type of this LCA. Different combinations of read assignment parameters (LCA at 90%, 95% or 100%; assignation of placements to the ascending, descending or nearest node) were also assessed, and resulted either in higher rates of unassigned reads or of wrongly assigned reads (Fig. S2).

Read counts were normalized by adjusted marker length: for each marker and each sequence file, counts were normalized by \((L - \ell + 1)\), with \(L\) the length of the marker gene (cpcBA mean length: 1053.7 bp; mpeBA mean length: 1054.6 bp; mpeW mean length: 1193.3 bp) and \(\ell\) the mean length of reads in the sequence file. Finally, the abundance of PT 1, 2A and 2B was defined as the normalized...
cpcBA read counts of these PT, the abundance of PT 3a, 3f and 3dA as the normalized mpeBA read counts of these PT, 3dB as the normalized mpeW count and 3c as the difference between the normalized mpeBA (3c + 3dB) read count and the PT 3dB count assessed with mpeW. The abundance of unclassified sequences was also taken into account. Detailed petB counts for clade and ESTU abundances were obtained from (6).

Read assignment simulations

For each marker, simulated reads were generated from one reference sequence at a time using a sliding window of 100, 125 or 150 bp (Tara Oceans mean read length: 164.2 bp; median 169 bp) and steps of 5 bp. Simulated reads were then assigned to a pigment type with the aforementioned bioinformatic pipeline, using all reference sequences except the one used to simulate reads (“leave one out” cross-validation scheme). Inferred PTs of simulated fragments were then compared to known PTs of reference sequences.

Statistical analyses

All environmental parameters used for statistical analyses are the same as in (6), except the blue to green irradiance ratio that was modeled as described in the supplementary materials and methods. Hierarchical clustering and NMDS analyses of stations were performed using R (63) packages cluster v1.14.4 (64) and MASS v7.3–29 (65), respectively. PT contingency tables were filtered by considering only stations with more than 30 cpcBA reads and 30 mpeBA reads, and only PT appearing in at least two stations and with more than 150 reads in the whole dataset. Contingency tables were normalized using Hellinger transformation that gives lower weights to rare PT. The Bray–Curtis distance was then used for ordination (isoMDS function; maxit, 100; k, 2). Correlations were
performed with R package Hmisc_3.17-4 with Benjamini & Hochberg multiple comparison adjusted p-value (66).

Acknowledgements

We warmly thank Dr. Annick Bricaud for fruitful discussions on biooptics, members of the ABiMS platform (Roscoff) for providing us an efficient storage and computing facility for our bioinformatics analyses as well as the NERC Biomolecular Analysis Facility (NBAF, Centre for Genomic Research, University of Liverpool) for sequencing some *Synechococcus* genomes used in this study. This work was supported by the French “Agence Nationale de la Recherche” Programs SAMOSA (ANR-13-ADAP-0010) and France Génomique (ANR-10-INBS-09), the French Government “Investissements d’Avenir” programs OCEANOMICS (ANR-11-BTBR-0008), the European Union’s Seventh Framework Programs FP7 MicroB3 (grant agreement 287589) and MaCuMBA (grant agreement 311975), UK Natural Environment Research Council Grant NE/I00985X/1 and the Spanish Ministry of Science and Innovation grant MicroOcean PANGENOMICS (GL2011-26848/BOS). We also thank the support and commitment of the *Tara* Oceans coordinators and consortium, Agnès b. and E. Bourgois, the Veolia Environment Foundation, Région 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). This article is contribution number XXXX of *Tara* Oceans.
References


Legends of Figures

Fig. 1: Maximum likelihood phylogenetic trees of (A) cpcBA operon, (B) mpeBA operon and (C) the mpeW/Y/Z gene family. The cpcBA tree includes both strains with characterized pigment type (PT) and environmental sequences (prefixed with GS) assembled from metagenomes of the Baltic Sea (38). Circles at nodes indicate bootstrap support (black: > 90%; white: > 70%). Note that for PT 2B clade, only environmental sequences are available. The PT associated with each sequence is indicated as a colored square. The scale bar represents the number of substitutions per nucleotide position.

Fig. 2: Distribution of *Synechococcus* pigment types (PT). (A) Relative abundance of each PT in the whole dataset (Total), in surface and at the DCM (Deep Chlorophyll Maximum). (B) Map showing the global distribution of all *Synechococcus* PTs in surface waters along the *Tara* Oceans transect. Diameters of pies are proportional to the number of cpcBA reads normalized by the sequencing effort. Stations with less than 30 cpcBA or mpeBA reads are indicated by open circles and those with no cpcBA reads by black dots. Numbers next to pies correspond to *Tara* Oceans stations.

Fig. 3: Correlation analysis between *Synechococcus* pigment types (PT) and environmental parameters measured along the *Tara* Oceans transect for all sampled depths. The scale shows the degree of correlation (blue) or anti-correlation (red) between two variables. Non-significant correlations (adjusted $P$ value > 0.05) are indicated by crosses. Number of observations for each environmental parameter is indicated at the bottom. Abbreviations: MLD, mixed layer depth; DCM, deep chlorophyll maximum; IS, in situ; Backscatt., backscattering; part., particulate; cDOM fluo, colored dissolved organic matter fluorescence; BAC, beam attenuation coefficient; $\Phi$sat, satellite-based non-photochemical quenching (NPQ)-corrected quantum yield of fluorescence (proxy for iron limitation; 6); PAR, photosynthetically active radiation; NPP, net primary production; $I_{R495:545}$, ratio of downwelling irradiance at 495 nm and 545 nm.
Fig. 4: Relationship between *Synechococcus* pigment types (PT) and Ecologically Significant Taxonomic Units (ESTUs, as defined in 6). (A) Correlation analysis between *Synechococcus* PTs and the most abundant ESTUs (>1% relative abundance) for all sampled depths (the complete dataset is shown in Fig. S5). Non-significant correlations (adjusted P value > 0.05) are indicated by crosses. The surface of station TARA_067, identified as an outlier (see Fig. S7), was removed for this analysis. (B) NMDS analysis of stations according to Bray–Curtis distance between PT assemblages. Samples that belong to the same ESTU assemblage have been contoured with a background color according to the color code used in (6), namely: red, assemblage 1 dominated by ESTU IIA; yellow, assemblage 2 dominated by ESTU IIIA; dark blue, assemblage 4 dominated by ESTUs IA and IVA/B; pink, assemblage 5 co-dominated by ESTUs IIB and IVA/B; grey, assemblage 6 co-dominated by ESTUs CRD1C and EnvBC; light blue, assemblage 8 co-dominated by ESTUs IVA/B, EnvBB and CRD1A/B. (C) PT and ESTU relative abundance at each surface station along the *Tara* Oceans transect. Oceanic provinces are indicated in the top gray panels. NAO, North Atlantic Ocean; MS, Mediterranean Sea; RS, Red Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean.
**Supplementary materials and methods**

**Modeling of the blue to green irradiance ratio ($I_{rr}^{495:545}$) at Tara Oceans stations.**

We used the clear sky surface irradiance model of Frouin and McPherson in Fortran and translated to Matlab by Werdell (see Frouin et al., 1989 and Tanre et al., 1979 for the analytical formula used) using the date, latitude and longitude of each station, assuming sunny sky and at noon.

The spectral light distribution averaged over the mixed layer was computed from:

$$< I_r(\lambda) > = \frac{\int_{0}^{MLD} I_{0-k(\lambda,chl)z} dz}{MLD} = \frac{I_{0-k(\lambda,chl)MLD}}{MLD k(\lambda,chl)} \{1 - e^{-k(chl)MLD}\}$$

where:

- $chl$ denotes the average chlorophyll value in the mixed layer. $[chl]$ was based on a fluorometer that was calibrated against HPLC data and corrected for non-photochemical quenching.
- MLD is the mixed layer depth that was computed based on a temperature threshold criterion
- $k(\lambda,chl)$ is the diffuse attenuation coefficient at wavelength $\lambda$ (495 or 545 using a 10 nm bandwidth). This parameter was computed using Morel and Maritorena (2001)’s equation:

$$k(\lambda,chl) = k_w(\lambda) + \chi(\lambda)[chl]e(\lambda)$$

$k_w$, $\chi$ and $e$ are provided in Table 2 of Morel and Maritorena (2001) and have the following values for the wavelengths of interest:

<table>
<thead>
<tr>
<th>Wavelength [nm]</th>
<th>$k_w(\lambda)$ [m$^{-1}$]</th>
<th>$\chi(\lambda)$</th>
<th>$e(\lambda)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>495</td>
<td>0.01885</td>
<td>0.06907</td>
<td>0.68947</td>
</tr>
<tr>
<td>545</td>
<td>0.05212</td>
<td>0.04253</td>
<td>0.65591</td>
</tr>
</tbody>
</table>

If the sampling depth was below the MLD, the irradiance was computed as follows:

$$I_r(\lambda, \text{sampling depth}) = (\lambda, 0^-)e^{-k(\lambda,chl)\text{sampling depth}}.$$  

The ratio was then computed as $I_{rr}^{495:545}$.

**References:**


Legends to supplementary figures

**Figure S1:** Biochemical composition and biooptical properties of phycobilisomes (PBS) of the main *Synechococcus* pigment types (PT). (A) Models of PBS structure, highlighting the conserved core and variable rods of increasing complexity from PT1 to PT3 (Redrawn after Six *et al.*, 2007). (B) Whole cell absorption spectra of the different PTs (Reproduced after Six *et al.*, 2007). Chromophores responsible of each absorption peaks are indicated. (C) Whole cell fluorescence excitation spectra with emission at 680 nm. Note that for chromatic acclimaters (PT 3d), the PBS structure is similar to other PT 3 but that the excitation ratio at 495 nm and 545 nm (Ex$_{495}$:545) varies from 0.6 in green light to 1.6 in blue light (not shown).

**Figure S2:** Evaluation of the assignment pipeline and the resolution power of the different markers used in this study. Simulated reads were generated from the reference dataset and assigned using a custom-designed pipeline (see materials and methods). (A, C, E) Evaluation of different sets of parameters tested for read assignment for the different markers: *cpcBA* (A), *mpeBA* (C) and *mpeW* (E). 100 (yellow), 125 (pink) and 150 bp (dark red) long reads were simulated. For each read, pplacer returns a list of possible positions in the tree, each associated with a likelihood weight. From these placements, we considered only those that reached a summed likelihood weight of either 90% (circle), 95% (square) or 100% (triangle). The assignment was then performed based on the phenotype of either the nearest node (solid symbol) in the tree or the descending (child) node (empty symbol). (B, D, F) Evaluation the resolution power along *cpcBA* (B), *mpeBA* (D) or *mpeW* (F) for 150 bp simulated reads assigned using the parameters selected for *Tara* Oceans metagenomic read assignment (i.e., nearest node assignment and summed weight of 95%). Note that *Tara* Oceans reads had a mean length of 164 bp.
**Figure S3:** Correlations between the number of reads recruited using the main markers used in this study. (A) Correlation between *petB* (vertical phylogeny) and *cpcBA* counts used to discriminate pigment types (PT) 1, 2 and 3. (B) Correlation between PT 3 counts using *cpcBA* and total *mpeBA* counts. Note that *mpeBA* is a PT3 specific marker and is used to discriminate PTs 3a, 3dA, 3f and 3c+3dB. (C) Correlation between PT 3dB counts using *fciAB*, a PT 3dB- and 3dA-specific marker and total *mpeW* counts, a PT 3dB-specific marker.

**Figure S4:** Distribution of *Synechococcus* pigment types (PTs) at depth (Deep Chlorophyll Maximum). (A) Map showing the global distribution of all *Synechococcus* PTs at depth along the *Tara Oceans* transect. Diameters of pies are proportional to the number of *cpcBA* reads normalized by the sequencing effort. Stations with less than 30 *cpcBA* or *mpeBA* reads are indicated by open circles and those with no *cpcBA* reads by black dots. Numbers next to pies correspond to *Tara Oceans* stations. (B) PTs and ESTU relative abundance at depth for sampling station along the *Tara Oceans* transect. Oceanic provinces are indicated in the top gray panels. NAO, North Atlantic Ocean; MS, Mediterranean Sea; RS, Red Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean.

**Figure S5:** Same as Fig. 4A but for all ESTUs. Unclass., unclassified.

**Figure S6:** Focus on pigment type (PT) 3dA natural mutants, exhibiting an altered gene content with regard to typical PT 3dA. (A-J) Correlation between the number of reads assigned as PT 3dA using different markers (all present in single gene copy in typical 3dA). Each circle corresponds to a *Tara Oceans* station and depth. Orange circles: stations with at least 20 *mpeBA* reads assigned to PT 3dA and at least twice more 3dA counted with *mpeBA* than with *fciAB*, corresponding to the surface sample of stations TARA_070, TARA_110 and TARA_137 and the
DCM of stations TARA_038, TARA_058 and TARA_110. Red circles: same but with more than
10-fold 3dA counted with mpeBA than fciAB, corresponding to the surface sample of stations
TARA_052, TARA_094, TARA_111 and TARA_122 to TARA_128, and DCM of stations
TARA_052, TARA_100, TARA_111 and TARA_128. Green circle: surface of station
TARA_067. (K) CA4-A genomic island and fragment of the phycobilisome (PBS) genomic
region for a typical, CA4-able 3dA strain (strain BL107), and 3 CA4-deficient strains, which are
stuck either in blue light phenotype (similar to strain BIOS-E4-1), or green light phenotype (as
strains MVIR-18-1 and WH8016). Note that KORDI-49 and WH8016 strains have identical PBS
gene complement and genomic arrangement. The complete PBS genomic region of the BL107
strain can be found in Six et al., 2007. Note that for readability, surface of station TARA_093 has
been omitted since it has the highest normalized counts (2.7-3.2) for all markers and exhibited a
good agreement between markers (ratio close to 1:1).

**Figure S7:** Correlation between the proportion of clades I, IV and CRD1, as assessed with petB,
and the proportion of pigment type 3dA, as assessed with mpeBA, at each station.
Phycobiliproteins

- **APC**  Allophycocyanin
- **PC**  Phycocyanin
- **PE-I**  Phycoerythrin-I
- **PE-II**  Phycoerythrin-II

Phycobilins

- **PCB**  Phycocyanobilin ($A_{\text{max}} = 620 \text{ nm}$)
- **PEB**  Phycoerythrobilin ($A_{\text{max}} = 545-560 \text{ nm}$)
- **PUB**  Phycourobilin ($A_{\text{max}} = 495 \text{ nm}$)

Other chromophore

- **Chl a**  Chlorophyll a
A

\[ n = 109 \]
\[ y = 1.04x - 0.0163 \]
\[ R^2 = 0.994 \]
\[ p-value = 0.113 \]

B

\[ n = 109 \]
\[ y = 0.991x + 0.00274 \]
\[ R^2 = 0.996 \]
\[ p-value = 0.607 \]

C

\[ n = 109 \]
\[ y = 1.02x + 0.00847 \]
\[ R^2 = 0.989 \]
\[ p-value = 1.57 \times 10^{-5} \]
<table>
<thead>
<tr>
<th></th>
<th>CRD1A</th>
<th>CRD1B</th>
<th>CRD1C</th>
<th>EnvAA</th>
<th>EnvBA</th>
<th>EnvBB</th>
<th>EnvBC</th>
<th>IA</th>
<th>IIA</th>
<th>IIB</th>
<th>IVA</th>
<th>IVB</th>
<th>IVC</th>
<th>VIA</th>
<th>VIIA</th>
<th>XIA</th>
<th>XIB</th>
<th>XIC</th>
<th>XIA</th>
<th>XIB</th>
<th>XIC</th>
<th>5.2A</th>
<th>5.3A</th>
<th>5.3B</th>
<th>5.3C</th>
<th>5.3D</th>
</tr>
</thead>
<tbody>
<tr>
<td>2A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3c</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3f</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3dA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3dB</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclassified</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unclass. PT3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
\[ n = 83 \]
\[ y = 0.86x + 0.0458 \]
\[ R^2 = 0.86 \]