A robust approach to estimate relative phytoplankton cell abundance from metagenomic data

Juan José Pierella Karlusich[1,2] *, Eric Pelletier[2,3], Lucie Zinger[1], Fabien Lombard[2,4,5], Adriana Zingone[6], Sébastien Colin[7,8,9], Josep M. Gasol[10], Richard Dorrell[1], Eleonora Scalco[6], Silvia G. Acinas[10], Patrick Wincker[2,3], Colomban de Vargas[2,8], Chris Bowler[1,2] *

[1] Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France
[2] CNRS Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GOSEE, 3 rue Michel-Ange, 75016 Paris, France
[3] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France
[4] Sorbonne Universités, CNRS, Laboratoire d'Océanographie de Villefranche (LOV), 06230 Villefranche-sur-Mer, France
[5] Institut Universitaire de France (IUF), Paris, France
[6] Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy
[7] European Molecular Biology Laboratory, Heidelberg, Germany
[8] Sorbonne Université, CNRS, Station Biologique de Roscoff, UMR 7144, ECOMAP, 29680 Roscoff, France
[9] Max Planck Institute for Developmental Biology, Tübingen, Germany
[10] Department of Marine Biology and Oceanography, Institut de Ciènces del Mar, CSIC, Barcelona, Spain

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

Running title: phytoplankton relative abundance from metagenomes

Keywords: photosynthesis, phytoplankton, *psbO*, V9-18S rRNA, metabarcoding, metagenomics, metatranscriptomics, *Tara* Oceans

Abstract

Phytoplankton account for >45% of primary production on Earth, and hence have an enormous impact on aquatic food webs and on the entire Earth System. Their members are found across many of the major clades of the tree of life, including prokaryotes (cyanobacteria) and multiple eukaryotic lineages. Phytoplankton communities are generally studied by PCR amplification of rRNA marker genes from the DNA extracted from environmental samples and high-throughput sequencing of the amplicons. However, our appreciation of phytoplankton abundance or biomass is limited by PCR-amplification biases, rRNA gene copy number variations across taxa and the fact that rRNA do not provide insights into metabolic traits such as photosynthesis (i.e., it is sometimes difficult to distinguish an rRNA sequence of a phytoplanktonic species from a phylogenetically close non-photosynthetic species). In addition, rRNA marker genes fail to capture both cyanobacteria and photosynthetic eukaryotes simultaneously, thus hampering assessment of the entire phytoplankton community. Here, we propose to use the core single-copy photosynthetic gene *psbO* extracted from metagenomes to circumvent these limitations: the method is PCR-free, and the gene is present in both prokaryotes and eukaryotes, mainly in one copy per genome. We applied and validated this new strategy with the *Tara* Oceans datasets, and showed in most cases improved correlations with flow cytometry and microscopy than when based on rRNA genes. Furthermore, we revealed unexpected features of the ecology of these organisms, such as the high abundance of picocyanobacterial aggregates and symbionts in the ocean, and the decrease in relative abundance of phototrophs towards the larger size classes of marine communities of dinoflagellates. To facilitate the incorporation of *psbO* in future molecular-based surveys, we released a curated database

including >20,000 unique sequences. Overall, *psbO* appears to be a promising new gene marker for improving molecular-based evaluations of phytoplankton communities.

Introduction

Photosynthetic plankton, or phytoplankton, consist of unicellular organisms of diverse evolutionary history and ecology. They are responsible for more than 45% of Earth's primary production (Field et al. 1998), fueling aquatic food chains, microbial decomposition, and the global ocean biological carbon pump (Guidi et al. 2009). They include prokaryotes (cyanobacteria) and multiple eukaryotic lineages that acquired photosynthesis mainly through secondary endosymbiotic events (Pierella Karlusich et al. 2020a). They display a broad body size spectrum, from less than 1 micra to several millimetres, either due to cell size variation, aggregation or symbioses (Beardall et al. 2009). This size variability partly explains their different roles in the food web and the biological carbon pump. For example, cyanobacteria are generally thought to be recycled within the microbial loop, whereas larger eukaryotic phytoplankton are usually considered more important in energy transfer to higher trophic levels (through grazing by small protists, zooplankton, and/or larvae (Ullah et al. 2018)) and in sequestering atmospheric CO_2 to the ocean interior through gravitational sinking of particles (Guidi et al. 2009).

The structure and composition of microbial communities is typically characterized by PCR amplification and sequencing of a fragment of the small subunit of the rRNA gene from an environmental sample (rRNA gene metabarcoding). The fraction of the obtained sequencing reads corresponding to a given taxon is then typically used as a proxy for its relative abundance. Most studies have so far focused on taxonomically informative fragments of the hypervariable regions of the 16S (for prokaryotes) or 18S (for eukaryotes) rRNA genes that are by far the most represented in reference databases (Jan Pawlowski et al. 2012; Guillou

et al. 2013; Quast et al. 2013). These markers are occasionally targeted in both DNA and RNA to exclude inactive microbes and as proxies of metabolic activities (Campbell et al. 2011; Logares et al. 2012), but recent studies have indicated severe limitations of this concept and only mRNA can be considered as an indicator of the metabolic state (Blazewicz et al. 2013).

Although rRNA gene metabarcoding is widely used, it has multiple limitations. Firstly, PCR-amplification bias due to mismatches of universal primers on the target sites of certain taxa can generate differences between the observed and the genuine relative read abundances as large as 10-fold, either when using the 16S (Parada, Needham, and Fuhrman 2016; Wear et al. 2018; Polz and Cavanaugh 1998) or 18S rRNA gene markers (Bradley, Pinto, and Guest 2016). Shotgun sequencing is a PCR-free alternative and consists of the detection of these marker genes in metagenomes (Obiol et al. 2020; Liu et al. 2007; Logares et al. 2014).

Secondly, the copy-number of these marker genes varies greatly among species. While bacterial genomes contain between one and fifteen copies of the 16S rRNA gene (Kembel et al. 2012; Acinas et al. 2004; Větrovský and Baldrian 2013), protists can differ by >5 orders of magnitude in their 18S rRNA gene copy numbers, from 33,000 copies in dinoflagellates to one in some small chlorophytes (Mäki et al. 2017; de Vargas et al. 2015; Zhu et al. 2005; Godhe et al. 2008). Although the copy number remains unknown for most microbial species, abundances could be corrected by assuming that this property is phylogenetically conserved, and thus its assessment in different organisms could lead to the establishment of correction factors. These approaches have been applied for 16S rRNA gene in bacteria, but their accuracy is limited for taxa with no close representatives in reference phylogenies (Louca, Doebeli, and Parfrey 2018; Kembel et al. 2012; Starke, Pylro,

and Morais 2020). In the case of protists, this correction is even more challenging due to intraspecies variation in 18S rRNA gene copy number. For example, it varies almost 10-fold among 14 different strains of the haptophyte *Emiliana huxleyi* (Gong and Marchetti 2019). In addition, there are major difficulties for generating a comprehensive database of 18S rRNA copy numbers (Gong and Marchetti 2019).

Finally, functional traits such as photosynthesis cannot be inferred solely from rRNA genes or other housekeeping markers, whereby their knowledge is limited to a restricted number of taxa known from experts and the literature. Indeed, while photosynthesis occurs in almost all cyanobacteria (except a few symbiotic lineages that have partially or totally lost it, such as UCYN-A (Thompson et al. 2012)), it is not necessarily conserved within protist taxa, such as dinoflagellates - of which only half of known species are photosynthetic (Saldarriaga et al. 2001; Dorrell and Smith 2011) -, chrysophytes (Dorrell and Smith 2011; Dorrell et al. 2019) and apicomplexa (Moore et al. 2008). This is an important issue because we still do not know how extended among related lineages are the independent events of chloroplast gains and losses or the extent of loss of photosynthesis with retention of the plastids. Thus, it is not possible to annotate the photosynthesis trait to those sequences whose taxonomic affiliation is, for example, “unknown dinoflagellate”.

Another disadvantage is the impossibility of making direct comparisons between cyanobacteria and eukaryotic phytoplankton with two different rRNA marker genes. This can be still attempted by targeting the plastidial and cyanobacterial versions of the 16S rRNA gene (Nicholas J. Fuller et al. 2006; N. J. Fuller et al. 2006; McDonald et al. 2007; Lepère, Vaulot, and Scanlan 2009; Kirkham et al. 2013, 2011; Shi et al. 2011). However, dinoflagellates and chromerids are not represented in these surveys because their plastidial 16S rRNA genes are extremely divergent

(Green 2011), and this approach can still capture non-photosynthetic plastids and kleptoplastids (functional plastids temporarily retained from ingested algal prey). Plastid-encoded markers directly involved in photosynthesis have also been used, such as *psbA* and *rbcL* (Paul, Alfreider, and Wawrik 2000; Zeidner, Preston, Delong, Massana, Post, Scanlan, and Beja 2003; Man-Aharonovich et al. 2010a). The *psbA* gene encodes the D1 protein of photosystem II and is also found in cyanophages (viruses) and the used primers target essentially the cyanobacterial and cyanophage sequences (Adriaenssens and Cowan 2014). The *rbcL* gene encodes the large subunit of the ribulose-l,5-diphosphate carboxylase/oxygenase (RuBisCO). There are multiple *rbcL* types, even in non-photosynthetic organisms, and the gene location varies: form I is plastid encoded in plants and most photosynthetic protists (and is present in cyanobacteria) while form II is nuclear-encoded in peridinin dinoflagellates and chromerids (and is also present in proteobacteria) (Tabita et al. 2008). The different *rbcL* variants thus prevent its use for covering the whole phytoplankton community.

Plastid-encoded genes (16S rRNA, *psbA*, *rbcL*, etc) are affected by copy number variability among taxa at different levels of organisation: gene copies, plastid genomes per plastid, and plastids per cell. For example, there are four copies of the 16S rRNA gene in the plastid genome of the euglenophyte *Euglena gracilis* and six in the prasinophyte *Pedinomonas minor* (Decelle et al. 2015). The plastid number per cell varies from one or a few in most microalgal species to more than 100 in many centric diatoms (Decelle et al. 2015). In addition, this varies according to biotic interactions, e.g., the haptophyte *Phaeocystis* has two plastids in a free-living stage but increases up to 30 when present as an endosymbiont of radiolarians (Decelle et al. 2019). Photosynthetic eukaryotes typically maintain 50–100 plastid genome

copies per plastid, but there is a continuous increase throughout development and cell cycle (Oldenburg and Bendich 2004; Armbrust and Virginia Armbrust, n.d.; Coleman and Nerozzi 1999; Koumandou and Howe 2007; Hiramatsu et al. 2006). These limitations of plastid-encoded marker genes can be circumvented by the use of photosynthetic nuclear-encoded genes, which is still an unexplored approach.

In spite of the aforementioned biases, the gene metabarcoding studies, either based on rRNA genes or on alternative marker genes such as *psbA* or *rbcL*, usually assume that the relative abundance of the gene sequences is an accurate measure of the relative abundance of the organisms containing those sequences. However, this assumption can lead to misleading inferences about microbial community structure and diversity, including relative abundance distributions, estimates of the abundance of different taxa, and overall measures of community diversity and similarity (Medinger et al. 2010; Mäki et al. 2017; Bachy et al. 2013; Kembel et al. 2012; Egge et al. 2013; Pinto and Raskin 2012). For example, less than 30% of the variance in true organismal abundance is explained by observed prokaryotic 16S rRNA gene abundance in some simulation analyses (Kembel et al. 2012). In addition, comparative studies between morphological and molecular approaches in environmental samples or in mock communities revealed discrepancies up to several orders of magnitudes among protist taxa with regard to their relative abundances (Medinger et al. 2010; Mäki et al. 2017; Bachy et al. 2013; J. Pawlowski et al. 2016; Egge et al. 2013). Most of these studies focused on the biases generated by primers and copy-number variations, but not on uncertainties in assigning photosynthesis (i. e., the lack of accuracy when discriminating sequences from phytoplankton from those of phylogenetically close non-photosynthetic species).

Therefore, we deemed it important to find more accurate alternative procedures to the most widely-used molecular approaches to make reliable estimations of species abundance, an important measure for inferring community assembly processes. We propose to target nuclear-encoded single-copy core photosynthetic genes obtained from metagenomes to circumvent these limitations: the method is PCR-free, and the genes are present in both prokaryotes and eukaryotes, in one copy per genome. We focused on the *psbO* gene, which encodes a subunit of photosystem II, and has the additional advantage of lacking any non-photosynthetic homologues. We quantified the biases in taxon abundance estimates using rRNA gene markers as compared to optical approaches (flow cytometry, microscopy), and we compared these patterns with those obtained by our proposed method. We also searched for *psbO* within metatranscriptomes to analyse its potential use as a proxy of photosynthetic activity and/or biovolume (due to the higher transcript level requirements in larger cells). Besides finding a more relevant marker gene for phytoplankton, we also propose its combination with single-copy housekeeping genes (e.g., *recA* for bacteria and genes encoding ribosomal proteins in eukaryotes) to estimate the fraction of photosynthetic members in the whole community or in a given taxon. In this context, we also quantified the unknowns in the functional assignation of photosynthetic capacity based on the 18S rRNA gene. Finally, we applied our method to several empirical data sets to illustrate how the approach improves measures of microbial community diversity, structure, and composition as compared to rRNA gene metabarcoding.

Materials and Methods

Search for phytoplankton marker genes

To estimate cell-based relative abundances of the major marine phytoplankton groups, we searched for genes present in all photosynthetic organisms (both prokaryotes and eukaryotes) and with low copy-number variability among taxa. To fulfil the latter requirement, we first excluded the plastid-encoded genes to avoid the variations in number of chloroplasts per cell and in number of chloroplast genomes per organelle. We did this by retrieving sequences from the KEGG (Kanehisa 2000) and MetaCyc (Caspi et al. 2018) databases that are assigned to the photosynthetic electron transport chain, the Calvin Cycle and chlorophyll biosynthesis to be used as queries for sequence similarity searches against >4,100 plastid genomes available at NCBI (https://www.ncbi.nlm.nih.gov/genome/organelle/). For this, BLAST version 2.2.31 ("tBLASTn" and "BLASTp" programs) searches were conducted with an e-value cutoff of 1e-20 (Camacho et al. 2009). To retain only core photosynthetic genes, i.e., those present in all phototrophs, we then made an equivalent BLAST search against cyanobacterial and eukaryotic nuclear genomes from the IMG (Chen et al. 2019) and PhycoCosm (Grigoriev et al. 2021) databases and from the polyA-derived transcriptomes of the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling et al. 2014). To minimize false-negative cases, only completely sequenced genomes were considered for establishing the absence of a gene. This survey was also used for determining gene copy number variation.

This survey resulted in a list of five genes that are core, nuclear-encoded and present in low-copy numbers (Table II). For selecting a gene marker of phytoplankton among them, we carried out a deeper sequence analysis to detect

non-photosynthetic homologes and to see if the phylogeny reflects the evolutionary history of cyanobacteria and endosymbiosis. We first performed a sequence similarity search using HMMer version 3.2.1 with gathering threshold option (http://hmmer.org/) for the corresponding Pfam domain against the translated sequenced genomes and transcriptomes of phytoplankton from PhycoCosm and MMETSP as well as in the whole IMG database (including viruses, archaea, bacteria and eukaryotes). The Pfams used in the search were: MSP (PF01716) for PsbO, Rieske (PF00355) for PetC, PRK (PF00485) for phosphoribulokinase, UbiA (PF01040) for (bacterio)chlorophyll-*a* synthase, and NAD_binding_1 (PF00175) for Ferredoxin: NADP+ reductase. CDHIT version 4.6.4 (W. Li and Godzik 2006) was used at a 80% dentity cut-off to reduce taxonomic bias and redundancy within each of the resulting protein sets. Protein sequences were aligned with MAFFT version 6 using the G-INS-I strategy (Katoh and Toh 2008). Phylogenetic trees were generated with PhyML version 3.0 using the LG substitution model plus gamma-distributed rates and four substitution rate categories (Guindon et al. 2010). The starting tree was a BIONJ tree and the type of tree improvement was subtree pruning and regrafting. Branch support was calculated using the approximate likelihood ratio test (aLRT) with a Shimodaira–Hasegawa-like (SH-like) procedure. These analyses led to focus on *psbO* as a gene marker of phytoplankton.

Analysis of Tara *Oceans datasets*

Tara Oceans expeditions between 2009 and 2013 performed a worldwide sampling of plankton in the upper layers of the ocean covering the main ocean basins (Sunagawa et al. 2020). To capture the whole size spectrum of plankton, a

combination of filter membranes with different pore sizes (size-fractionation) was used to separate organisms by body size (Pesant et al. 2015). There is an inverse relationship between plankton size and abundance (Belgrano et al. 2002; W. K. W. Li 2002; Pesant et al. 2015), so small size fractions represent the numerically dominant organisms in terms of cell abundance (albeit not necessarily in terms of total biomass). Thus, the protocols consisted in the filtering of higher seawater volumes for the larger size fractions (Pesant et al. 2015). Five major organismal size fractions were collected: picoplankton (0.2-3 μm size fraction), piconanoplankton (0.8-5 μm size fraction), nanoplankton (5-20 μm size fraction), microplankton (20 to 180 μm), and mesoplankton (180 to 2000 μm). These plankton samples were used to generate different molecular and optical datasets that were analysed in the current work (Table I). We exclusively used the datasets corresponding to surface samples (5 m average depth).

Table I: *Tara* Oceans datasets relevant to the current study.

Target	Size fraction	Dataset	Dataset construction	Subset used in the current work	References and link
Prokaryotes and picoeukaryotes	0.2-3 µm	16S miTags (metagenomic Illumina tags)	16S rRNA gene sequences were identified in metagenomes and assembled. OTUs were defined at a 97% identity cut-off	726 OTUs assigned to picophytoplankton (258 cyanobacteria + 468 eukaryotic phytoplankton)	Salazar et al., 2019 https://www.ocean-microbiome.org/
Eukaryotes	5 size fractions (0.8-2000 µm)*	V9-18S rRNA gene fragment metabarcoding	PCR amplification of the V9 region (130 ± 4 base pairs length) of 18S rRNA gene followed by the high-throughput sequencing of the amplicons, which were clustered into OTUs using SWARM	31,930 OTUs assigned to eukaryotic phytoplankton (including photosynthetic dinoflagellates and chrysophytes)	de Vargas et al., 2015; Ibarbalz et al., 2019 https://zenodo.org/record/3768510#.Xraby6gzY2w
Prokaryotes and picoeukaryotes	0.2-3 µm	Ocean Microbial Reference Gene Catalog (OM-RGC-v2)	Unigenes assembled from metagenomes and clustered at 95% identity. Metagenomic and metatranscriptomic reads were then mapped on these unigenes.	307 *psbO* sequences from cyanobacteria and eukaryotic phytoplankton	Salazar et al., 2019 https://www.ocean-microbiome.org/
Eukaryotes	5 size fractions (0.8-2000 µm)*	Marine Atlas of *Tara* Oceans Unigenes (MATOU-v1)	Transcribed sequences assembled from poly-A+ metatranscriptomes and clustered at 95% identity. Metagenomic and metatranscriptomic reads were then mapped on these unigenes.	10,646 *psbO* sequences from eukaryotic phytoplankton	Carradec et al., 2018 http://www.genoscope.cns.fr/tara/
Prokaryotes and eukaryotes	6 size fractions (0.2-2000 µm)**	Metagenomes	Raw metagenomic reads	~3.2 million metagenomic reads aligned to a curated database of *psbO* sequences	EBI accession numbers: PRJEB1787, PRJEB1788, PRJEB4352, PRJEB4419, PRJEB9691, PRJEB9740 and PRJEB9742
Prokaryotes and	<200 µm	Flow cytometry		Abundances and biovolume of picocyanobacteria and	Hingamp et al. 2013; Gasol and Morán 2015 https://

eukaryotes				eukaryotic picophytoplankton	data.mendeley.com/ datasets/p9r9wttjkm/2
Eukaryotes	5-20 µm	Confocal microscopy		Abundance and biovolume of nanophytoplankton	Colin et al., 2017 https://www.ebi.ac.uk/ biostudies/studies/S-BSST51
Eukaryotes	20-180 µm	Light microscopy		Abundance of microphytoplankton	Malviya et al., 2016

*0.8-5 µm, 5-20 µm, 20-180 µm, 180-2000 µm

**0.2-3 µm, 0.8-5 µm, 5-20 µm, 20-180 µm, 180-2000 µm

psbO-*based community data*

To use the metagenomic and metatranscriptomic read abundances of *psbO* as a proxy of phytoplankton relative cell abundance and 'activity', respectively, we carried out a HMMer searched as stated in the previous section against the two *Tara* Oceans gene catalogues: the Ocean Microbial Reference Gene Catalog version 2 (OM-RGC.v2) covering prokaryotic and eukaryotic picoplankton (<3 µm), and the Marine Atlas of *Tara* Oceans Unigenes version 1 (MATOU.v1) covering eukaryotic plankton ranging from 0.8 to 2000 µm (Table I). The metagenomic and metatranscriptomic reads are already mapped onto both catalogs, thus we retrieved these values for those sequences obtained by our HMMer search. For the taxonomic assignment of the *psbO* unigenes, we performed a phylogenetic placement of the translated sequences on the PsbO protein reference phylogenetic tree described in the previous section. For parallelization of the task, a set of 50 unigenes were translated and the PsbO specific Pfam PF01716 region was retrieved for the analysis in the following way. First, they were aligned against the reference alignment described in the previous section using the option --add of MAFFT version 6 with the G-INS-I strategy (Katoh and Toh 2008). The resulting alignment was used for building a phylogeny with PhyML version 3.0 as described above (Guindon et al. 2010). The sequences were classified using the APE library in R (Paradis and Schliep 2019) according to their grouping in monophyletic branches of statistical support >0.7 with reference sequences of the same taxonomic group.

Due to challenges of assembling eukaryotic genomes from complex metagenomes, the MATOU-v1 catalog only contains sequences assembled from

poly-A-tailed RNA (Alberti et al. 2017; Carradec et al. 2018), which biases against prokaryotic sequences. To determine the structure of the whole phytoplankton community (including both cyanobacteria and eukaryotic phytoplankton), we thus aligned all the metagenomic reads from *Tara* Oceans to a curated database of *psbO* sequences (described below; see also Table I). The analysis was carried out using the bwa tool with the following parameters: -minReadSize 70 -identity 80 -alignment 80 -complexityPercent 75 -complexityNumber 30. Abundance values were expressed in rpkm (reads per kilobase covered per million of mapped reads).

In general, the rpkm values for the different taxa under study were converted to percentage of total phytoplankton. However, for a specific analysis the *psbO* rpkm values were normalized by single-copy housekeeping genes. We used the single-copy core bacterial gene *recA* (Pierella Karlusich et al. 2015; Sunagawa et al. 2013) to quantify the bacterial community in each metagenome; thus the read abundance ratio of cyanobacterial *psbO* to bacterial *recA* provides an estimate for the relative contribution of cyanobacteria in the bacterioplankton. In a similar way, we used the average abundance of 25 genes encoding ribosomal proteins in eukaryotes (Ciccarelli et al. 2006; Carradec et al. 2018) to divide the eukaryotic *psbO* abundance. The abundance values for *recA* were retrieved from previous work (Pierella Karlusich et al. 2020b) while the ribosomal proteins were recovered from the MATOU-v1 and OMRGC-v2 abundance tables.

rDNA-based community data

We used two different datasets generated by *Tara* Oceans for “traditional” DNA-based methods: 16S rRNA gene miTags size fraction 0.2-3 µm) and 18S rRNA

gene (V9 region) metabarcoding (sizes fractions 0.8-5, 5,20, 20-180, 180-2000 µm) (Table I). We extracted the relative abundances for the 726 Operational Taxonomic Units (OTUs) assigned to picophytoplankton (cyanobacteria and eukaryotes) from the 16S miTags and the 31,930 OTUs assigned to eukaryotic phytoplankton from the V9-18S metabarcoding. These assignations are based on literature and expert information, including photosynthetic dinoflagellates and chrysophytes when their taxonomic resolution was sufficient to match known photosynthetic lineages from the literature (a full description of the 18S taxonomic classification procedure and reference database is available at http://taraoceans.sb-roscoff.fr/EukDiv/).

The read abundances were expressed as relative abundance (%) in relation to the picophytoplankton community for 16S miTags, and the eukaryotic phytoplankton for V9-18S metabarcoding..

Optical-based community data

We also used quantitative optical data generated by *Tara* Oceans (Table I), where cell abundance is assumed to be more accurate and less biased and additional features such as biovolume (a proxy of biomass) can be determined. The datasets cover: flow cytometry for picoplankton, confocal microscopy for 5–20 µm size fraction, and light microscopy for 20-180 µm size fraction.

Flow cytometry counts were determined on 1ml replicated seawater samples filtered through 200 µm that were fixed with cold 25% glutaraldehyde (final concentration 0.125%) and stored at -80°C until analysis. Details about the proceeding can be found in (Hingamp et al. 2013; Gasol and Morán 2015). The cell

biovolume was calculated using the equation of (Calvo-Díaz and Morán 2006) on the bead-standardized side scatter of the populations and considering cells to be spherical.

Quantitative confocal microscopy was performed using environmental High Content Fluorescence Microscopy (eHCFM) (Colin et al. 2017). Briefly, samples were fixed with 10% monomeric formaldehyde (1 % final concentration) buffered at pH 7.5 and 500 µl EM grade glutaraldehyde (0.25% final concentration) and kept at 4 °C until analysis. Sample collection and preparation as well imaging acquisition is described in (Colin et al. 2017). The 5–20 µm size fraction has been classified at a coarse taxonomic level (with an estimated accuracy of 93.8% at the phylum or class level), into diatoms, dinoflagellates, haptophytes, and other/unclassified eukaryotic phytoplankton (Colin et al. 2017). We used the major and minor axis of every image to calculate their ellipsoidal equivalent biovolume.

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

To be compared with the molecular data, the optical data were expressed as relative abundance (%). In the case of flow cytometry, as % over the total number of cells counted as picophytoplankton (picocyanobacteria + eukaryotic picophytoplankton). In the case of confocal and optical microscopy, as % over the total number of cells counted as eukaryotic phytoplankton.

psbO *database generation*

We curated and generated a database of 20,359 *psbO* sequences covering cyanobacteria, photosynthetic protists, macroalgae and land plants. It includes sequences retrieved from IMG, NCBI, MMETSP and other sequenced genomes and transcriptomes, as well as from the environmental sequence catalogs of the Global Ocean Sampling (Rusch et al. 2007) and *Tara* Oceans expeditions (Salazar et al. 2019; Carradec et al. 2018; Delmont et al. 2018; Delmont et al. 2020). The database can be downloaded from the EMBL-EBI repository BioStudies (www.ebi.ac.uk/biostudies) under accession S-BSST659. We expect to maintain it updated to facilitate its incorporation in molecular-based surveys.

Results

Search for phytoplankton marker genes

We first analysed transcriptomes and nuclear- and plastid- genomes derived from cultured strains to inventory photosynthetic genes in relation to their genome location (nuclear- vs plastid-encoded) and taxonomic prevalence (core vs non-core, i.e., present in all phototrophs or not) (see Methods; Figure 1A). Among the plastid-encoded genes, we identified phytoplankton marker genes previously used in environmental surveys, such as *psbA* (Man-Aharonovich et al. 2010b; Zeidner, Preston, Delong, Massana, Post, Scanlan, and Béjà 2003), *rbcL* (nuclear encoded in dinoflagellates containing peridinin) (Paul, Alfreider, and Wawrik 2000) and *petB* (Farrant et al. 2016) (Figure 1A).

Among the nuclear-encoded genes, we retrieved some which are non-core, such as those encoding flavodoxin (*fld*) and plastocyanin (*petE*), but also five core

genes (Figure 1A; Table II). These five genes are present in low-copy number and encode components of the photosynthetic electron transport chain (*psbO*, *petC* and *petH*), the carbon fixation pathway (*prk*) or the chlorophyll biosynthesis (*chlG*) (Table II). The absence of non-photosynthetic homologues is a unique characteristic of *psbO* (Table II) and a clear advantage for its use as a marker gene for phytoplankton, and in addition its phylogeny reflects the evolutionary history of endosymbiosis (Figure 1B; Pierella Karlusich et al. 2015), so we focused on this gene for the analysis of environmental samples.

Then, we retrieved the *psbO* sequences from the two *Tara* Oceans gene catalogues (the picoplankton catalog OM-RGC.v2 and the eukaryotic catalog MATOU.v1; see Methods and Table I). A total of 307 distinct sequences were identified in OM-RGC.v2 (202 from *Prochlorococcus*, 79 from *Synechoccocus* and 26 from eukaryotic picophytoplankton), with an average length for the conserved coding region of 473 base pairs and a range between 94 and 733 base pairs. A total of 10,646 sequences from eukaryotic phytoplankton were retrieved from MATOU.v1, with an average length for the conserved coding region of 385 base pairs with a range between 66 and 784 base pairs. The analyses of the metagenomic and metatranscriptomic read abundances of these sequences are presented in the following sections.

Table II: List of nuclear-encoded photosynthetic genes present in all cyanobacteria and eukaryotic phytoplankton. These genes are always nuclear-encoded, with the exception of the amoeba of the genus *Paulinella* (Fig. 2A), which has gained its plastid only very recently and independently of the event at the origin of all other known plastids (Ponce-Toledo, López-García, and Moreira 2019).

Gene	Pathway	Function	Copies	Non-photosynthetic homologues	References
prk (phosphoribulokinase)	Calvin-Benson-Bassham cycle	phosphorylation of ribulose-5-phosphate to ribulose-1,5-bisphosphate, the RuBisCO substrate	1	PRKs from archaea and bacteria	(Kono et al. 2017; Jaffe et al. 2019)
chlG ((bacterio)chlorophyll-*a* synthase)	chlorophyll-*a* biosynthesis	last step of chlorophyll-*a* biosynthesis	1	Prenyltransferases with UbiA domain	(Wang et al. 2015)
petH (ferredoxin-NADP$^+$ oxidoreductase)	Photosynthetic electron transport chain	last step of the linear electron flow (NADP$^+$ reduction using oxidized ferredoxin or flavodoxin)	1-3	-FNRs involved in nitrogen metabolism -FNRs from non-photosynthetic plastids -C-terminal region of benzoyl-CoA oxygenase component A (BoxA) from bacteria	(Mohamed et al. 2001; Pierella Karlusich and Carrillo 2017)
petC	Photosynthetic electron transport chain	Rieske subunit of the chloroplast Cyt b_6f complex	2-3	Rieske proteins from mitochondria, bacteria and archaea	(Veit et al. 2016; Lebrun et al. 2006)
psbO	Photosynthetic electron transport chain	Manganese-stabilizing protein of photosystem II	1-2	No	(Pierella Karlusich et al. 2015)

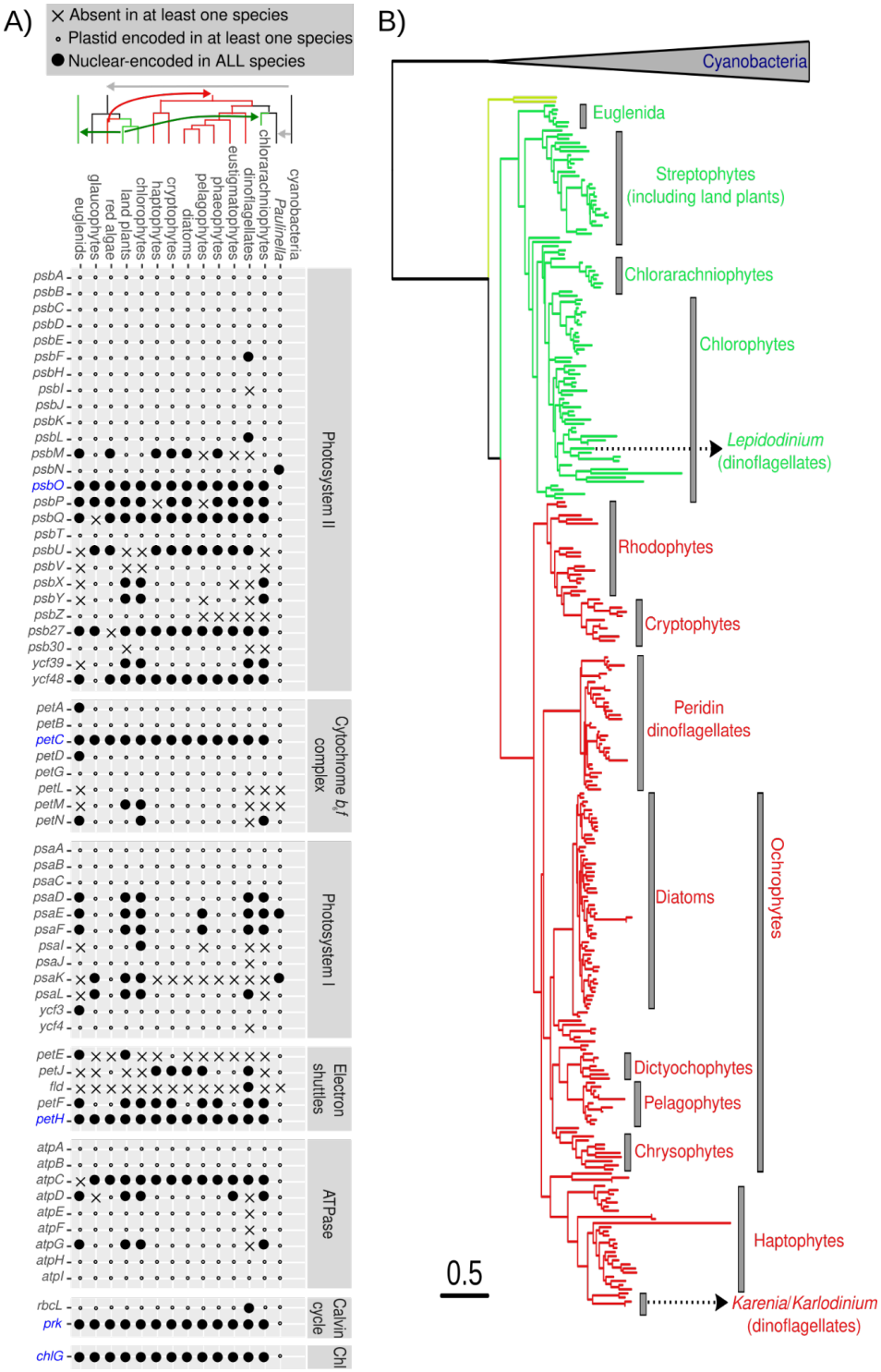

Figure 1: Search for nuclear-encoded core photosynthetic gene marker candidates. A) Presence and location of the genes encoding proteins involved in photosynthesis. The genes found to be core and nuclear-encoded are labelled in blue. B) Phylogeny of PsbO protein. Translated sequences from genomes and transcriptomes of cultured phytoplanktonic species were used for the phylogeny reconstruction. The scale bar indicates the number of expected amino acid substitutions per site per unit of branch length.

Marine phytoplankton community structure based on psbO *showed remarkable differences with the traditional molecular approaches*

The abundance and diversity of phytoplankton was first examined in *Tara* Oceans samples by focusing on the traditional marker genes coding for the small subunit of rRNA (16S for prokaryotes, 18S for eukaryotes) in the different size-fractionated samples. We focused exclusively on the phytoplankton signal of these datasets, despite the uncertainties in assigning photosynthesis capacity in groups such as dinoflagellates and chrysophytes (this is evaluated in one of the next sections).

Based on 16S miTag read abundance among picophytoplankton (0.2-3 µm), as expected *Prochlorococcus* and *Synechococcus* dominated by far the picocyanobacteria, while ~60% of the reads were attributed to eukaryotic photosynthetic taxa such as haptophytes, chlorophytes, pelagophytes, dictyochophytes, chrysophytes, cryptophytes and diatoms (Figure 2A). In the larger size fractions, based on the V9-18S region metabarcoding reads, diatoms and dinoflagellates were the most frequent among eukaryotic phototrophs, especially in the 5–20 µm and 20–180 µm size fractions (Figure 2B). In the 180-2000 µm fraction, diatoms and dinoflagellates were still abundant, due to the presence of large diameter cells (*Tripos*, *Pyrocystis*), chain-forming (e.g., *Chaetoceros*, *Fragilaria*) or epizoic (e.g., *Pseudohimantidium*) species, without discarding that smaller species may be retained in samples of this size fraction due to net clogging or within herbivorous guts and faecal pellets. Relative abundance in the smaller 0.8–5 µm size fraction was much more homogeneously distributed between the different groups.

For *psbO*-based methods, we observed that metagenomic and metatranscriptomic reads from *Synechococcus*, *Prochlorococcus*, pelagophytes, chlorophytes and haptophytes to be dominant among picophytoplankton (0.2-3 µm), along with dictyochophytes and chrysophytes (Figure 2A). In the larger size fractions, haptophytes, chlorophytes and pelagophytes clearly dominated the 0.8-5 µm size fraction, whereas diatoms and dinoflagellates were more abundant in the three larger size ranges (5-20 µm, 20-180 µm, 180-2000 µm), but haptophytes, chlorophytes and pelagophytes were also detected in large quantities (Figure 2B). The potential cyanobacteria present in these large size fractions are presented later in another section, as we need to bypass the use of sequences assembled from poly-A-tailed RNA for analysing prokaryotes (see also Methods and Table I).

We noted some differences in *psbO* read counts between metagenomic and metatranscriptomic datasets. In the case of picophytoplankton, *Prochlorococcus* was enriched in metagenomes in comparison to metatranscriptomes, the opposite occurred for pelagophytes and haptophytes, whereas no major changes were observed for *Synechococcus* and chlorophytes (Figures 2A and S1A). In the case of larger photosynthetic protists, dinoflagellates were highly abundant at the transcript level in comparison to gene abundance, the opposite was observed for pelagophytes and chlorophytes (in this latter taxon only in the 20-180 and 180-2000 µm size ranges), whereas no major shifts were apparent for diatoms and haptophytes (Figures 2B and S1B).

Interestingly, the taxonomic abundance patterns based on *psbO* showed some differences from those from 16S miTags of 0.2-3 µm size fraction, but exhibited remarkable differences with those based on V9-18S metabarcoding of the large size fractions (Figures 2A-B and S2). When compared with the 16S miTags, no major

changes were detected for *Prochlorococcus*, whereas the average *psbO* metagenomic contribution increased for *Synechococcus* (from ~8% to ~14%), at the expense of decreasing eukaryotic picoplankton contribution (from 57% to ~50%), which is expected due to the fact that the 16S rRNA is a plastid-encoded gene in eukaryotes. When we compared *psbO* with V9-18S metabarcoding, the differences were very significant. In the 0.8-5 µm size fraction, diatoms and dinoflagellates accounted for just ~6% of average *psbO* metagenomic read abundance but for ~44% of 18S reads assigned to phytoplankton. In the three larger size ranges (5-20 µm, 20-180 µm and 180-2000 µm), they accounted for 37-47% of average *psbO* metagenomic read abundance, but for >90% of average 18S read abundance. The 18S read abundance was extremely low for haptophytes, chlorophytes and pelagophytes in these three size fractions (<7% average 18S read abundance). When we compared the metatranscriptomic profile, it was more similar to the profile obtained with metagenomes than to the obtained with V9-18S metabarcoding (Figure 2B).

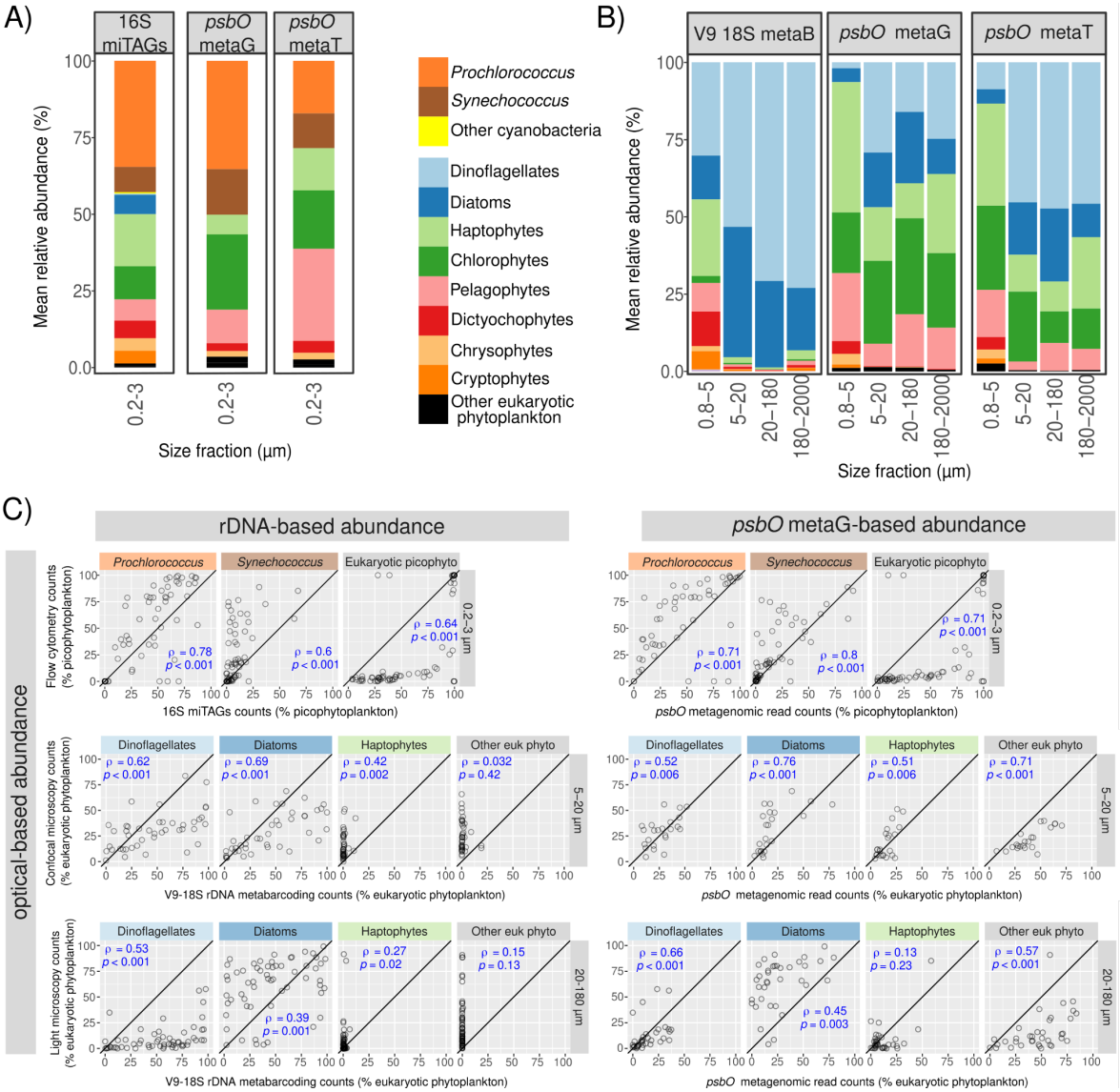

Figure 2: Congruence in relative abundances of the main phytoplankton groups based on different gene markers. A-B) Average relative abundances for all surface samples of each size fraction using different marker genes. In A, picocyanobacteria and eukaryotic picophytoplankton (0.2-3 µm) were analysed using 16S rRNA gene miTags and the metagenomic and metatranscriptomic read abundances for *psbO*. In B, eukaryotic phytoplankton was analysed in larger size fractions using V9-18S rRNA gene amplicons and the metagenomic and (polyA-derived) metatranscriptomic *psbO* read abundances. C) Correlations between relative abundances of different phytoplankton groups obtained with optical versus DNA-based methodologies. In the upper panel, 16S rRNA gene miTags and *psbO*-based relative abundances in picophytoplankton were compared with flow cytometry counts (values displayed as % total abundance of picophytoplankton). In the middle and lower panels, V9-18S rRNA gene metabarcoding and metagenomic *psbO* relative abundances were compared with confocal microscopy counts from size fraction 5-20 µm and light microscopy counts from size fraction 20-180 µm (values displayed as % total abundance of eukaryotic phytoplankton). It is worth mentioning that the molecular and microscopy data were generated from the same samples, while there were differences between molecular data of 0.2-3 µm size fraction and flow cytometry data (see Methods). Axes are in the same scale and the diagonal line corresponds to a 1:1 slope. Spearman's rho correlation coefficients and p-values are displayed. The correlations of relative abundances between metatranscriptomic *psbO* reads and optical methods are shown in Figure S3.

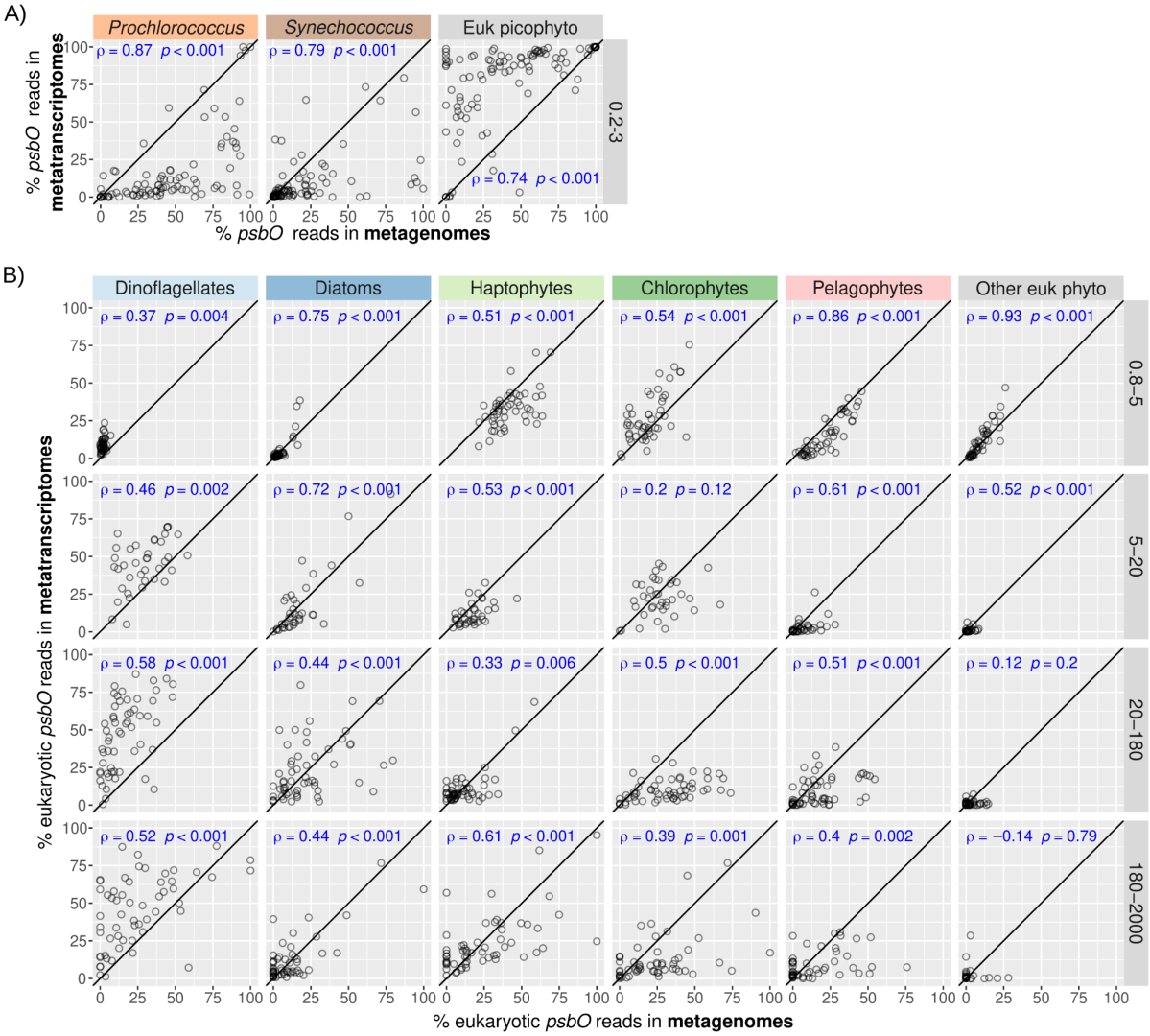

Figure S1: Comparison of relative abundances of *psbO* reads between metagenomes and metatranscriptomes of size fractionated samples. A) Picocyanobacteria and eukaryotic picophytoplankton (size fraction 0.2-3 µm). B) Eukaryotic phytoplankton in the large size fractions (0.8-5 µm, 5-20 µm, 20-180 µm, 180-2000 µm), with metagenomes compared to metatranscriptomes derived from poly-A RNA. Axis are in the same scale and the diagonal line corresponds to a 1:1 slope. Spearman's rho correlation coefficients and p-values are displayed in blue.

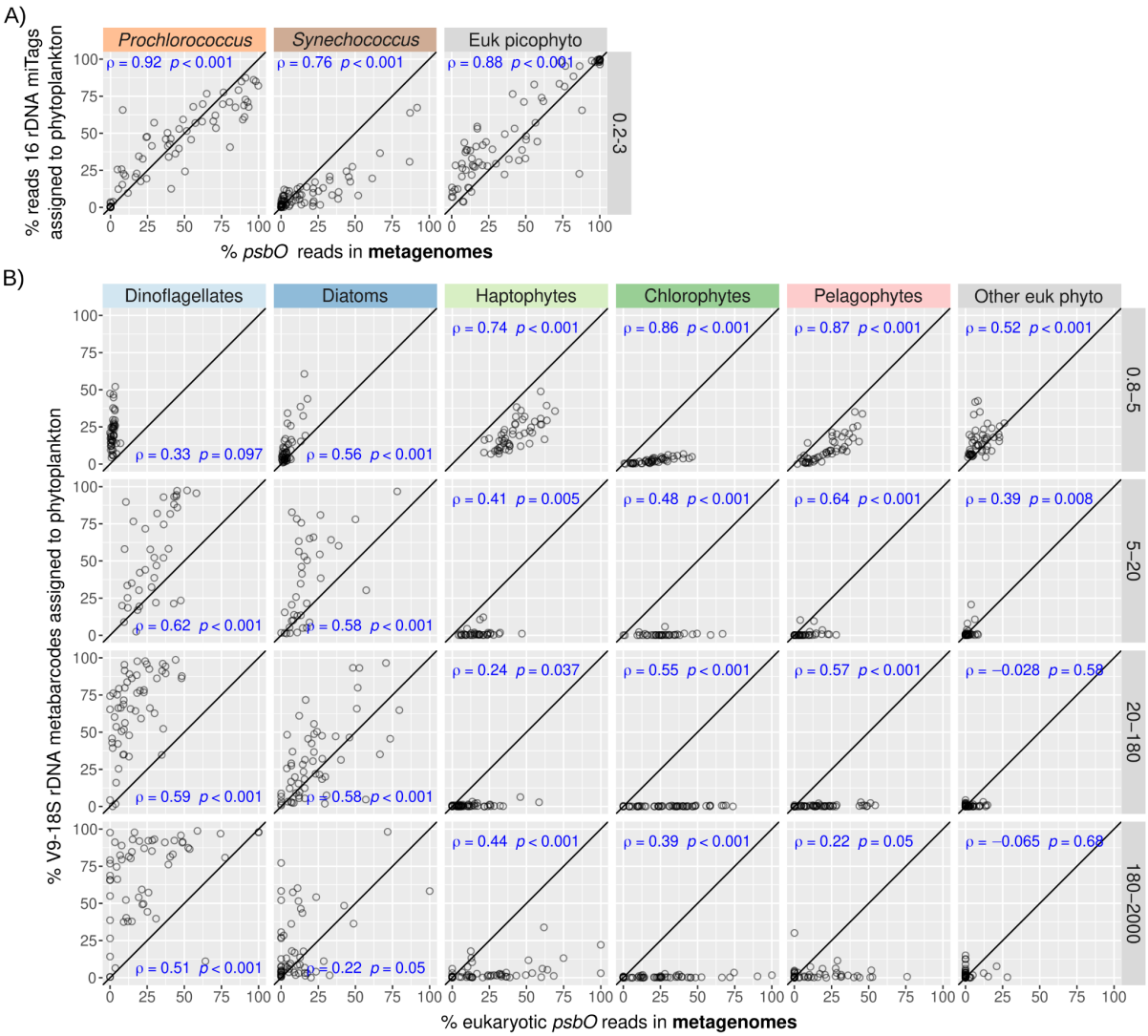

Figure S2: Comparison of relative read abundances of *psbO* and rRNA genes of size fractionated samples. A) Picocyanobacteria and eukaryotic picophytoplankton (0.2-3 µm) were analysed using the relative abundances for 16S rRNA gene miTags and for *psbO* metagenomic reads. B) Eukaryotic phytoplankton were analysed in the large size fractions (0.8-5 µm, 5-20 µm, 20-180 µm, 180-2000 µm) using the relative abundances for V9-18S rRNA gene amplicons and *psbO* metagenomic reads. Axis are in the same scale and the diagonal line corresponds to a 1:1 slope. Spearman's rho correlation coefficients and p-values are displayed in blue.

Comparison with imaging dataset suggested that psbO *is a robust marker gene for estimating relative cell abundance of phytoplankton from metagenomes*

To assess the accuracy of the *psbO* gene for determining phytoplankton cell relative abundances, we carried out comparative analyses with imaging datasets. For the 0.2-3 µm size fraction, we compared relative abundances based on 16S and

psbO counts with those inferred from flow cytometry determinations (Figure 2C). Both genes were found to correlate well with flow cytometry. Although the correlations for eukaryotic picophytoplankton were strong (Spearman's rho=0.64-0.71, p-value<<<0.001), the relationships were not linear and picoeukaryotes appeared at much higher relative abundances in metagenomes than in flow cytometry. This is consistent by the fact that flow cytometry can count cells of up to 10-20 µm diameter and was performed on seawater aliquots pre-filtered through a 200-µm mesh (see Methods), whereas DNA isolation of picoplankton was carried out on seawater volumes mainly filtered through 3-µm pore sizes. When we discarded eukaryotes to focus only on the ratio *Synechococcus* / (*Synechococcus* + *Prochlorococcus*) (Figure S3), flow cytometry data shows a linear relashiopship with *psbO* metagenomic reads, while 16S miTags reads underestimated *Synechococcus* and the opposite occured for *psbO* metatranscriptomic reads. In addition, the highest correlation with flow cytometry data occurred with the *psbO* metagenomic counts (Spearman's rho=0.92, 0.90 and 0.75, p<<<0.001, for *psbO* metagenomic reads, *psbO* metatranscriptomic reads and 16S miTags, respectively).

For the 5–20 µm size fraction, the relative abundance of eukaryotic photosynthetic organisms was determined by cell counts using high-throughput confocal microscopy. We compared these results with the proportions based on V9-18S metabarcoding and *psbO* metagenomic reads (Figure 2D). The metabarcoding data for dinoflagellates and diatoms were in good agreement with the microscopy but it clearly underestimated the relative abundance of haptophytes and other eukaryotic phytoplankton. Regarding *psbO*, the metagenomic relative abundances were in stronger agreement with the microscopy counts for the four defined phytoplankton groups (Figure 2D). Therefore, in the 5–20 µm size fraction, diatoms and

dinoflagellates displayed robust patterns of relative abundance using either V9-18S metabarcoding or *psbO* metagenomic counts, while haptophytes and the rest of groups were better described by *psbO*.

In the 20-180 µm size fraction, the relative abundance of eukaryotic phytoplankton was determined by light microscopy. Again, the metabarcoding data for dinoflagellates and diatoms were in good agreement with the microscopy data but clearly underestimated the relative abundance of haptophytes and other eukaryotic phytoplankton groups (Figure 2D). The metagenomic read relative abundances of *psbO* were in stronger agreement with the microscopy counts for the four defined phytoplankton groups, although the correlation with haptophytes was weaker (Figure 2D). Therefore, in the 20-180 µm size fraction, diatoms and dinoflagellates displayed robust patterns of relative abundance using either V9-18S metabarcoding or *psbO* metagenomic counts, while haptophytes were weakly described by both methods and the rest of the groups were much better described by *psbO*.

We also compared the relative abundances based on optical methods against those based on *psbO* metatranscriptomic reads, and in general we observed good agreement (Figure S4). Some phytoplankton groups displayed stronger correlations against optical methods using metatranscriptomic *psbO* counts than 16S miTAGs (e.g. *Synechococcus*) or V9-18S metabarcoding (e.g. other eukaryotic phytoplankton in the 5-20 µm) (Figures 2D and S4). However, the *psbO* relative abundance agreement with optical methods was always better for metagenomes than for metatranscriptomes (Figures 2D and S4).

Finally, we also compared the relative read abundances of the different marker genes against the proportional biovolume for each taxon (Figure 3). Although the

copy number of rRNA marker genes was previously proposed as proxy of cell biovolume (due to a positive association between rDNA content and cell size in some marine protist taxa (Lavrinienko et al. 2021)), the correlation of biovolume against rRNA gene relative abundances were not stronger than those against *psbO* (Figures 3 and S5). The relative read abundances for *Prochlorococcus* and eukaryotic picophytoplankton based either on 16S rDNA or *psbO* were higher than their proportional biovolume in the same samples, while the opposite occurred for *Synechococcus.* In the 5-20 µm size fraction, the biovolume proportion for haptophytes was clearly described by their *psbO* relative abundance, while their V9-18S rDNA reads were very low in relation to their biovolume. Both V9-18S rDNA and *psbO* reads were correlated with the relative biovolume for diatoms and dinoflagellates, but for V9-18S rDNA the data points were somewhat scattered and for *psbO* the relative abundances for the reads were higher in relation to their biovolume. As the biovolume of other taxa was very low, their proportion of *psbO* reads were much higher than the corresponding biovolume fraction, whereas there was no correlation between V9-18S and biovolume.

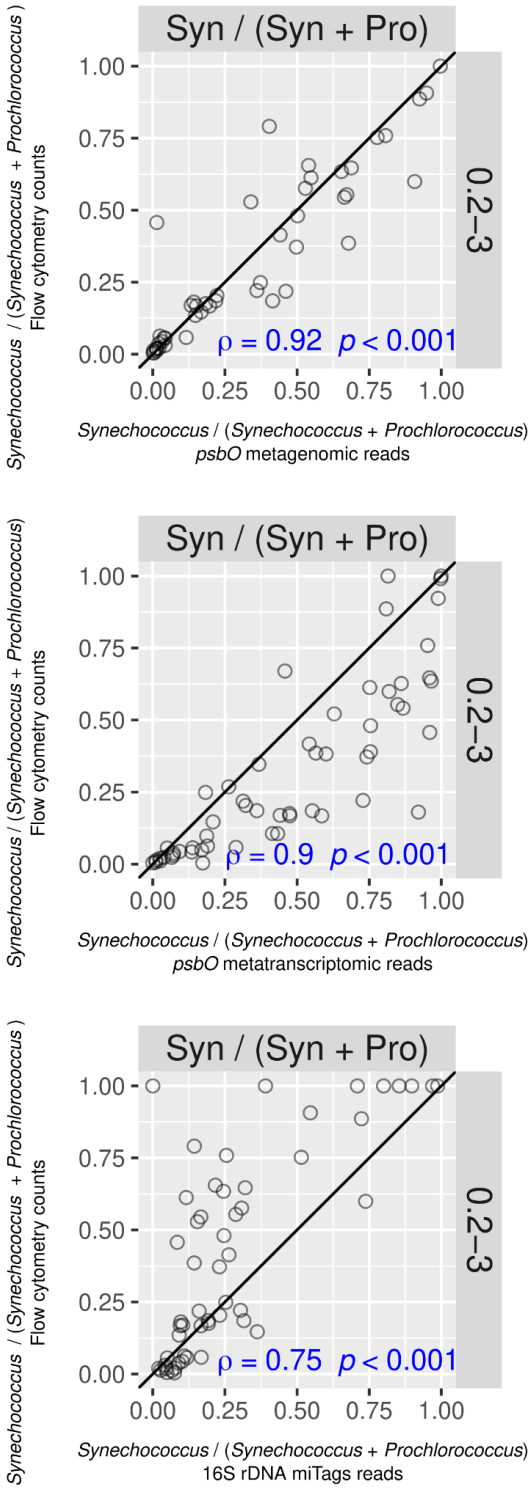

Figure S3: Correlation between the abundance ratio of the picocyanobacteria *Synechococcus* and *Prochlorococcus* obtained with different methodologies. The vertical axis corresponds to the ratio based on flow cytometry while the horizontal axis corresponds to the ratio based on *psbO* metagenomic reads (upper plot) or *psbO* metatranscriptomic reads (middle plot) or 16S miTAGs reads (bottom plot). Axis are in the same scale and the diagonal line corresponds to a 1:1 slope. Spearman's rho correlation coefficients and p-values are displayed in blue.

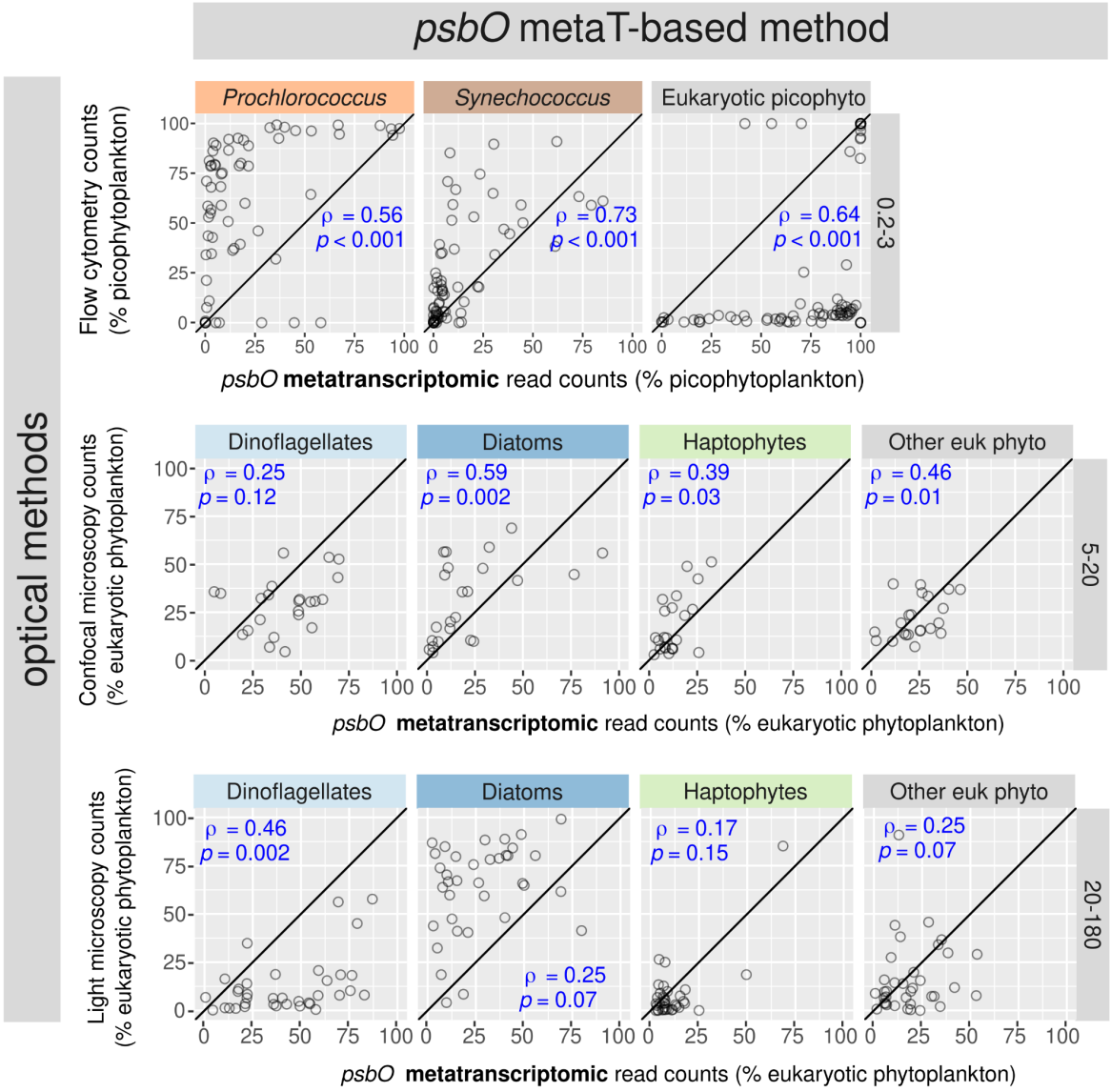

Figure S4: Correlation between relative abundances based on *psbO* metatranscriptomic reads against those based on optical quantifications of different phytoplankton groups. Upper panel: metatranscriptomic *psbO* relative abundances from picophytoplankton (size fraction 0.2-3 µm) were compared with flow cytometry counts (values displayed as % total abundance of picophytoplankton). Middle and lower panels: metatranscriptomic *psbO* relative abundances of eukaryotic phytoplankton were compared with confocal microscopy counts from size fraction 5-20 µm and light microscopy counts from size fraction 20-180 µm (values displayed as % total abundance of eukaryotic phytoplankton). Spearman correlation coefficients and p-values are displayed in blue. Axis are in the same scale and the diagonal line corresponds to a 1:1 slope.

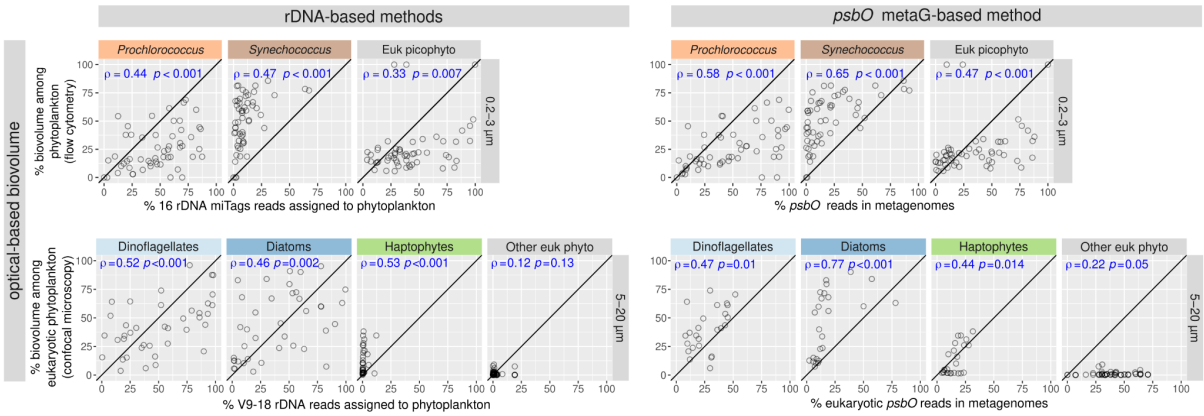

Figure 3: Correlation between relative biovolume (based on optical methods) and relative abundances based on different molecular methodologies. The upper panels show the correlations for picophytoplankton (size fraction 0.2-3 µm). The vertical axis corresponds to the relative biovolume based on flow cytometry (values displayed as % total biovolume of picophytoplankton), while the horizontal axis corresponds to relative read abundance based on molecular methods: 16S miTAGs (left upper panel) and *psbO* metagenomic counts (right upper panel). The lower panels show the correlations for nanophytoplankton (size fraction 5-20 µm). The vertical axis corresponds to the relative biovolume based on confocal microscopy quantification (values displayed as % total abundance of eukaryotic phytoplankton), while the horizontal axis corresponds to relative read abundance based on molecular methods: V9-18S rDNA metabarcoding (left lower panel) and eukaryotic *psbO* metagenomic counts (right bottom panel). Spearman correlation coefficients and p-values are displayed in blue. Axis are in the same scale and the diagonal line corresponds to a 1:1 slope.

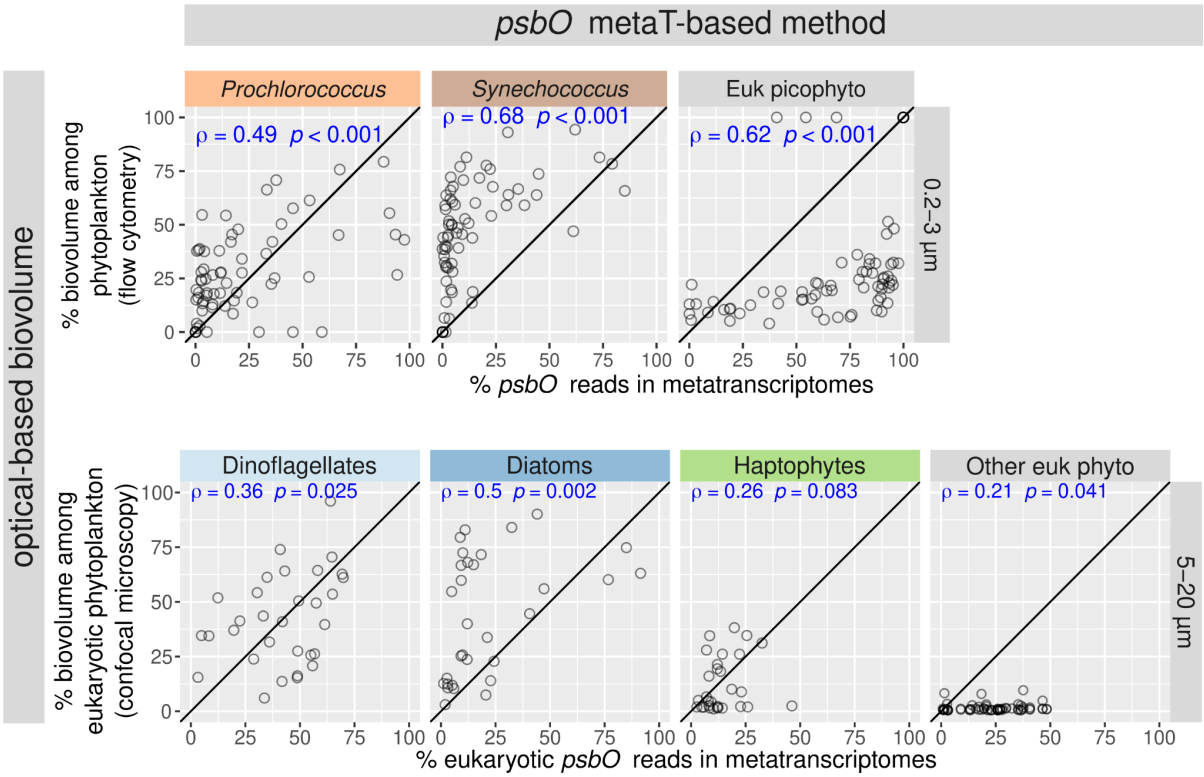

Figure S5: Correlation between relative biovolume (based on optical methods) and relative abundances based on *psbO* metatranscriptomic reads. The upper panel shows the correlations for picophytoplankton (size fraction 0.2-3 µm). The vertical axis corresponds to the relative biovolume based on flow cytometry (values displayed as % total biovolume of picophytoplankton), while the horizontal axis corresponds to relative read abundance based on *psbO* metatranscriptomic reads. The lower panel shows the correlations for nanophytoplankton (size fraction 5-20 µm). The vertical axis corresponds to the relative biovolume based on confocal microscopy quantification (values displayed as % total abundance of eukaryotic phytoplankton), while the horizontal axis corresponds to relative read abundance based on eukaryotic *psbO* metatranscriptomic reads. Spearman correlation coefficients and p-values are displayed in blue. Axis are in the same scale and the diagonal line corresponds to a 1:1 slope.

Diversity analysis: Shannon-index was more robust to the biases introduced by the traditional molecular methods

We further analysed whether our method improved the widely used Shannon index, a diversity index that relates monotonically to species richness but differs in that it downweights rare species, whose numbers are highly sensitive to undersampling and molecular artefacts (Calderón-Sanou et al. 2020). We found a good correlation between Shannon values for eukaryotic phytoplankton defined

either by 18S rRNA gene metabarcoding or by *psbO* metagenomics or metatranscriptomics (Figure 5). This is in agreement with previous reports showing no major effects of 16S rRNA gene copy number variation on the Shannon index of bacterial communities (Ibarbalz et al. 2019; Milanese et al. 2019). However, the largest size fraction did not display any correlation.

All in all, the robust patterns of Shannon index illustrates that not all subsequent analyses can be affected by the biases introduced by traditional molecular methods.

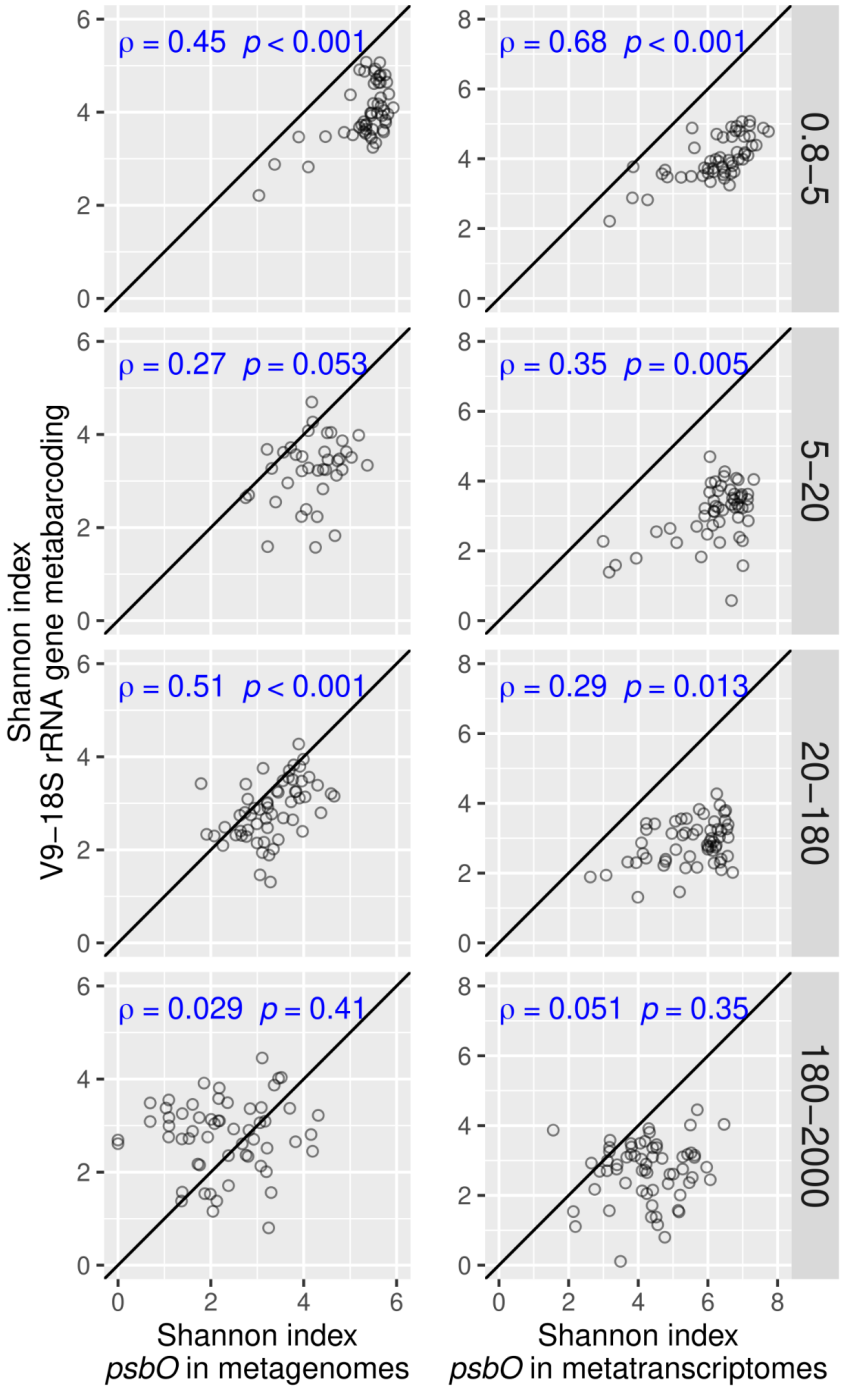

Figure 5: Correlation between the Shannon values derived from different molecular methods for eukaryotic phytoplankton communities. The values derived from *psbO* metagenomics (left) and metatranscriptomics (right) were compared with those derived from V9-18S rDNA metabarcoding. Spearman correlation coefficients and p-values are displayed in blue. Axis are in the same scale and the diagonal line corresponds to a 1:1 slope.

Combining housekeeping and photosynthetic marker genes improved estimates of the distribution and abundance of phototrophs in a given taxonomic group

To evaluate the uncertainties when inferring the photosynthesis trait using the taxonomy obtained from a non-photosynthetic marker gene, we analysed the 18S V9 OTUs assigned to dinoflagellates and found that most of their reads cannot be reliably classified as corresponding or not to a photosynthetic taxon (Figure 6A), especially for those OTUs whose taxonomic affiliation is “unknown dinoflagellate”. The uncertainty was especially significant in the 0.8-5 µm size fraction, where ~80% of the total dinoflagellate read abundance remained unclassified (Figure 6A).

Therefore, besides finding a more relevant marker gene for phytoplankton, we also propose combining it with established single-copy housekeeping genes (i.e., *recA* for bacteria (Pierella Karlusich et al. 2015; Sunagawa et al. 2013) and genes encoding ribosomal proteins for eukaryotes (Ciccarelli et al. 2006; Carradec et al. 2018)), to estimate the fraction of photosynthetic members in a given community or within a specific clade. In the case of eukaryotes, a set of genes of interest for this aim are *petC* and its mitochondrial homologues (i.e., the nuclear-encoded genes for the Rieske subunits of the Cyt *bc*-type complexes from chloroplasts and mitochondria) (Table II). As an example, we analysed the distribution of phototrophy across size fractions among the eukaryotic groups under study. As expected, it did not reveal any differences for diatoms, haptophytes, chlorophytes or pelagophytes (Figure S6), as the loss of photosynthesis is very rare in these groups (but see for example: (Kamikawa et al. 2017)). Instead, for dinoflagellates we observed a significant proportion of non-photosynthetic lineages in the 0.8-5 size-fraction in comparison with the other sizes ranges, which were also shown by the V9-18S rDNA

metabarcoding method (Figure 6B). However, whereas the metabarcoding data showed a dramatic increase in phototrophs towards the larger size classes of dinoflagellates, the metagenomic analysis showed similar levels between the three larger size fractions (5-20 µm, 20-180 µm, 180-2000 µm) (Figure 6B). These different patterns between the two marker genes might be explained by differences in the unknown trait assignment of the 18S rDNA barcodes and/or by differences in the 18S rDNA copy number in photosynthetic versus non-photosynthetic dinoflagellates.

The approach suggested can be applied to unveil variation of phototrophs in whole plankton communities, including both bacteria and eukaryotes. In order to do so across the whole plankton size spectrum, we mapped the metagenomic reads against a comprehensive catalogue of *psbO* sequences (see Methods). The highest proportion of phytoplankton among eukaryotes was observed in the 0.8-5 µm size fraction, followed by the 5-20 µm size-fraction, while the lowest value was found in the 180-2000 µm size range (Figure 6C), where copepods are prevalent (considered one of the most abundant animals on the planet). Surprisingly, the percentage of phototrophs among bacterioplankton did not vary across size fractions (10-15 % on average). In the 0.2-3 µm size fraction, very similar values were detected by 16S miTags, but when comparing both molecular methods with flow cytometry, the *psbO*/ *recA* ratio was better correlated to flow cytometry (Spearman's rho of 0.82 vs 0.91, $p<0.001$, and a closer 1:1 relationship) (Figure S7).

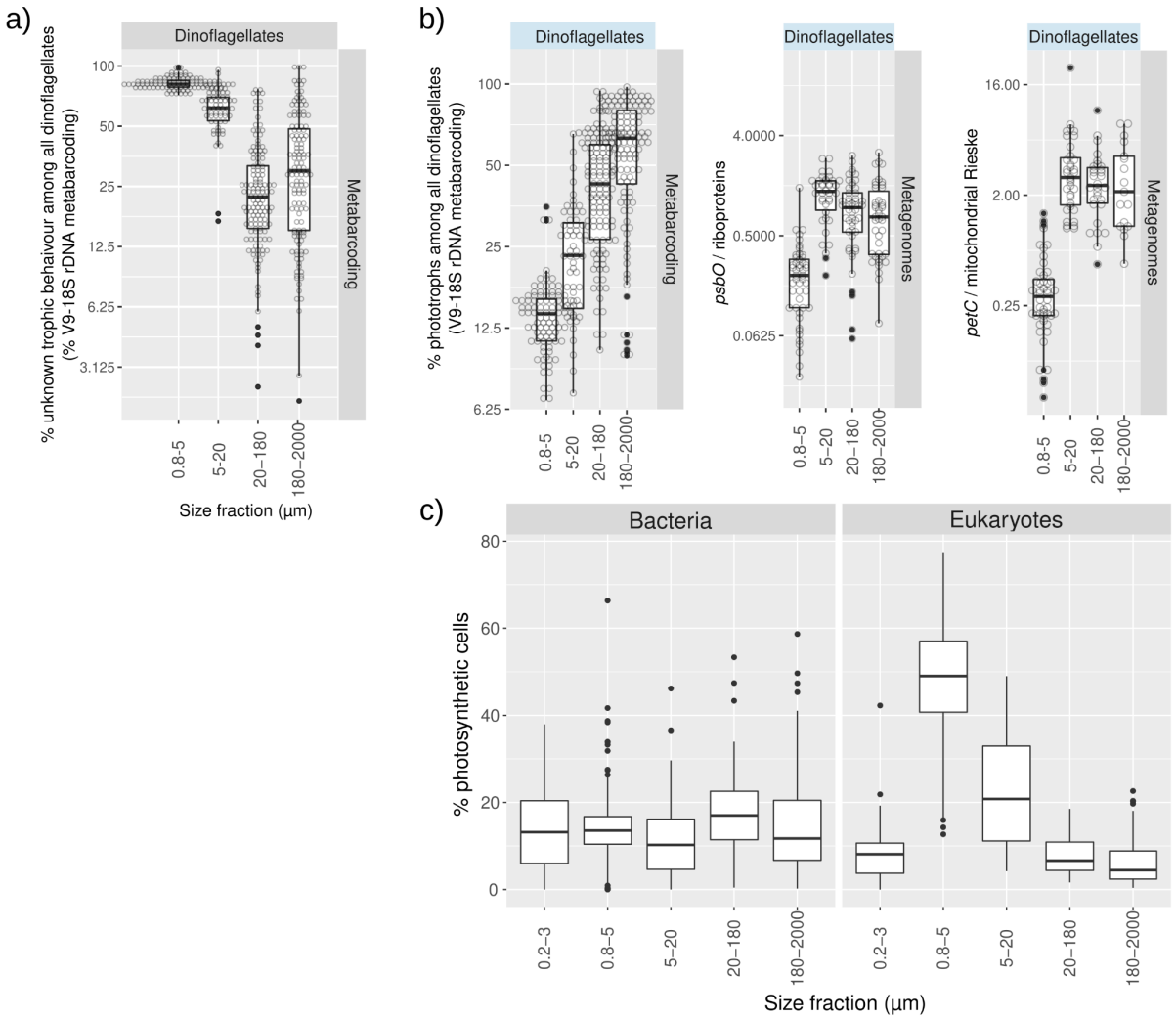

Figure 6: Variations in the abundance of phototrophs *vs* heterotrophs across size fractions. A) Read abundance of V9-18S rDNA metabarcoding assigned to dinoflagellates of unknown capacity for photosynthesis. B) Relative abundance of phototrophs among dinoflagellates based on different molecular methods. The first panel corresponds to the trait classification of V9-18S rDNA metabarcodes based on the literature (a description of the taxonomic classification is available at http://taraoceans.sb-roscoff.fr/EukDiv/). The second and third panels correspond to the ratio of metagenomic counts of photosynthetic *vs* housekeeping single-copy nuclear-encoded genes: *psbO* vs genes coding for ribosomal proteins, and the genes coding for the Rieske subunits of the Cyt bc-type complexes from chloroplasts and mitochondria (i.e., *petC* and its mitochondrial homologue). C) Relative abundance of phototrophs among bacterial and eukaryotic plankton across size fractions. The values were determined by the ratio of metagenomic counts of the single-copy marker genes of photosynthesis (i.e., *psbO*) and housekeeping metabolism (i.e., *recA* for bacteria and genes encoding ribosomal proteins for eukaryotes).

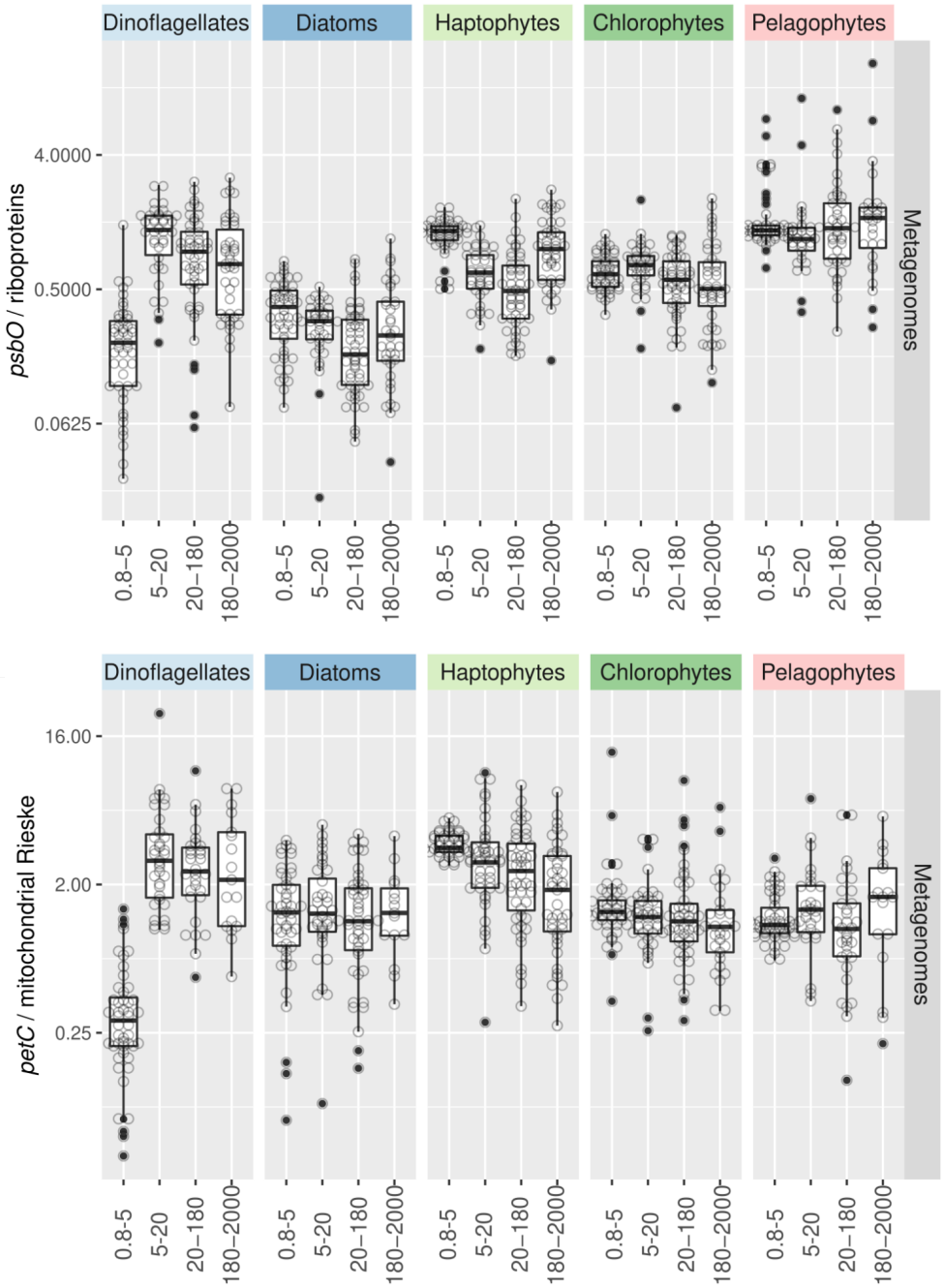

Figure S6: Variations in the abundance of phototrophs *vs* heterotrophs across size fractions based on different marker genes in metagenomes. The estimation were based on the ratio of metagenomic reads of photosynthetic *vs* housekeeping single-copy nuclear-encoded genes: *psbO* vs genes coding for ribosomal protein (upper panel), and the genes coding for the Rieske subunits of the Cyt bc-type complexes from chloroplasts and mitochondria (i.e., *petC* and its mitochondrial homologue) (lower panel).

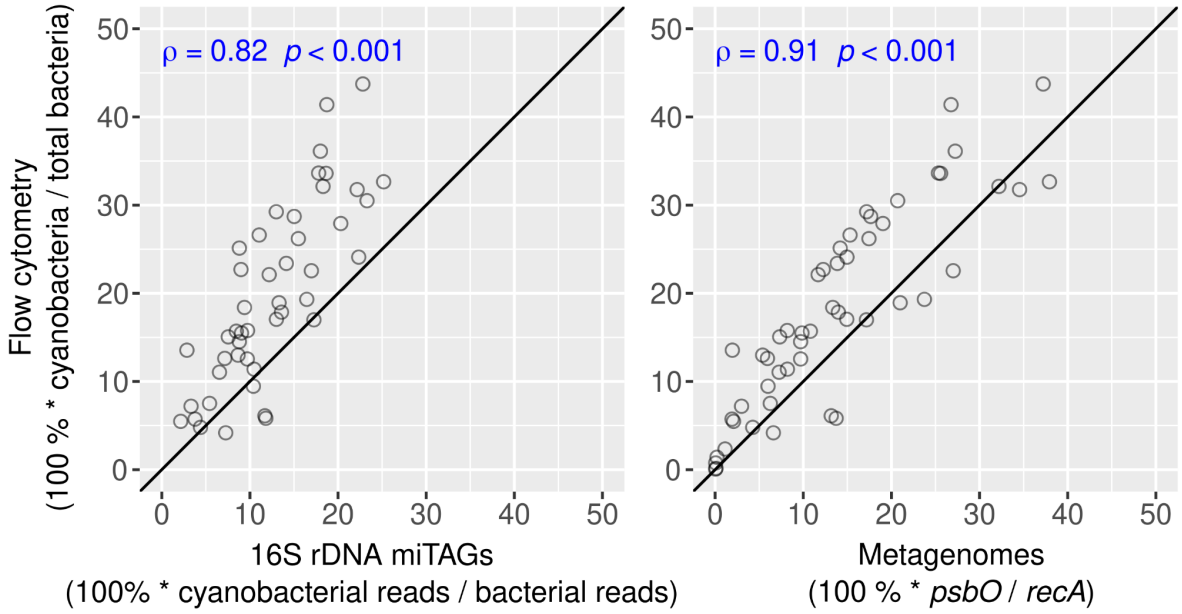

Figure S7: Congruence in the relative abundance of phototrophs among bacterioplankton in 0.2-3 size fraction based on different methods. The vertical axis corresponds to the relative cell abundance based on flow cytometry while the horizontal axis corresponds to the relative read abundance based on 16S miTags (left) or on the ratio of *psbO* to *recA* in metagenomes (right). Spearman correlation coefficients and p-values are displayed in blue. Axis are in the same scale and the diagonal line corresponds to a 1:1 slope.

Trans-domain comparison revealed unexpected abundance of picocyanobacteria in large size fractions

To further examine the distribution of both prokaryotic and eukaryotic phytoplankton across the whole size spectrum, we continued the analysis of the mapped metagenomic reads against our catalogue of *psbO* sequences. We observed a high abundance of cyanobacteria in the large size fractions in relation to the eukaryotic phytoplankton (Figure 7A). The nitrogen-fixers *Trichodesmium* and *Richelia*/*Calothrix* were found principally in the 20-180 and 180-2000 µm size fractions (Figure 7A), which is expected as the former forms filaments and colonies while the second group are symbionts of certain diatoms (Figures 7B-D). These genera were recently quantified in the high-throughput confocal microscopy dataset from the 20-180 µm size fraction (Pierella Karlusich et al., 2020b), therefore we

checked the correlations of these data with the *psbO* determinations and we found them to be very strongly related (Figure 7G). Another nitrogen-fixing cyanobacterium that is known to be abundant is UCYN-A, which lacks the ability to perform photosynthesis due to the absence of photosystem II, and thus it was not found in this analysis.

To our surprise, we also detected a high abundance of both *Prochlorococcus* and, in particular, *Synechococcus*, in the large size fractions (Figure 7A). Picocyanobacteria have small cell diameters (<1 µm), and therefore should readily pass through the filters with pore sizes of 5, 20 or 180 µm. Although smaller cells can get caught on larger filters, their abundance should be limited and hence not responsible for the values observed. The reason why a substantial fraction of picocyanobacteria were found in the largest size fractions may be colony formation, symbiosis, attachment to particles, or their grazing by protists, copepods and/or suspension feeders. We examined these possibilities by looking at the *Tara* Oceans confocal microscopy dataset, and found many microscopy images evidencing colony formation and symbiosis in the 20-180 µm size fraction (Figure 7E-F). In addition, this is in agreement with the mapping of the *Tara* Oceans metagenomes against a recently sequenced single cell genome of a *Synechococcus* living as a dinoflagellate endosymbiont (Nakayama et al. 2019). These results suggest that we should move from the traditional view of *Synechococcus*/*Prochlorococcus* as being exclusive part of picoplankton communities, and instead we should consider them as part of a broader range of the plankton size spectrum (in a similar way as occurs with other small-celled phytoplankton such as the haptophyte *Phaeocystis*; (Decelle et al. 2019; Beardall et al. 2009)). However, it should be borne in mind that these results correspond to estimates of relative cell abundance, and thus the picture can be very

different when translated to biomass (especially when considering the large differences in size). All in all, our approach allows us to make trans-domain comparisons, which can reveal photosymbiosis and aggregations.

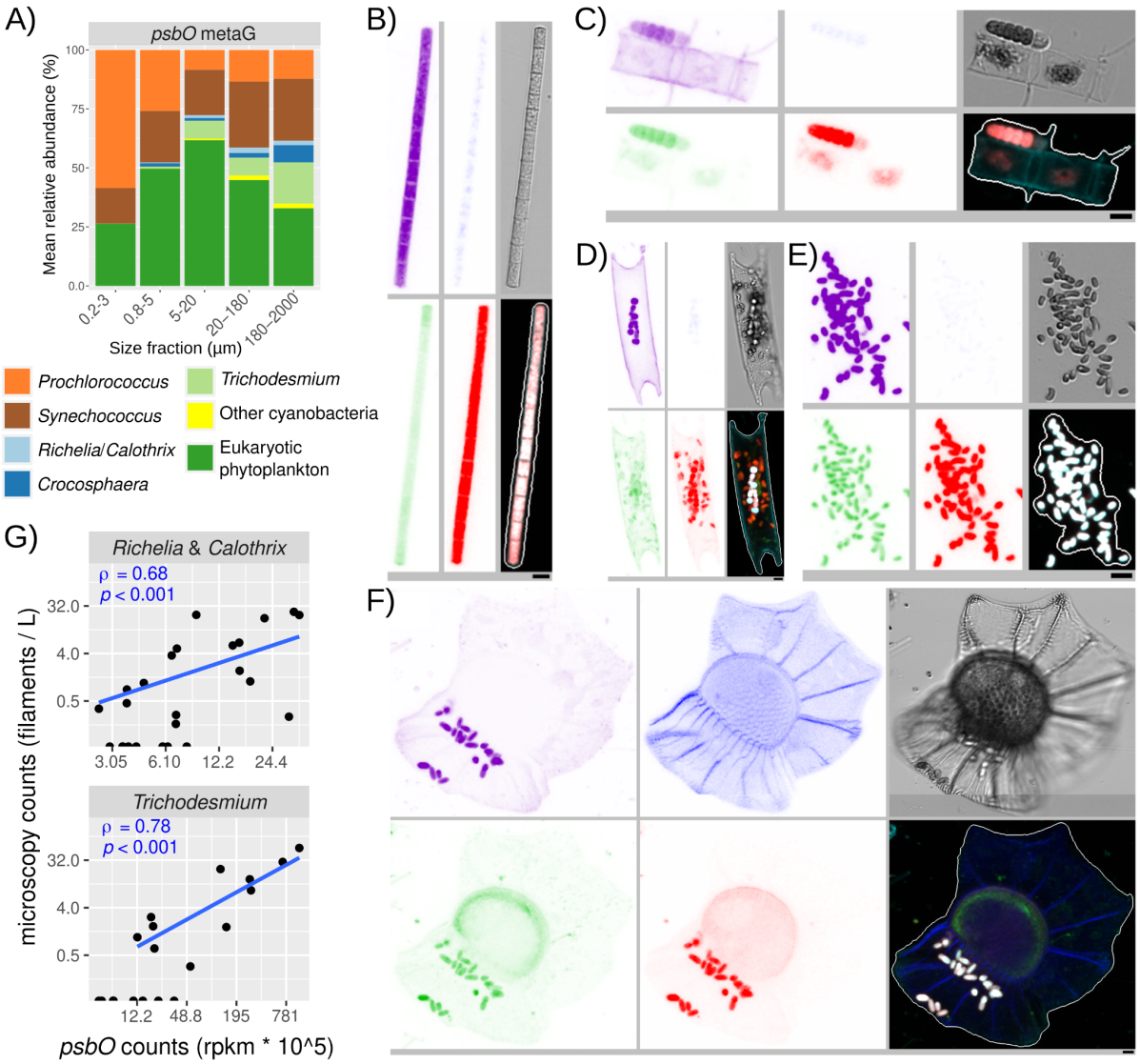

Figure 7: Prokaryotic and eukaryotic phytoplankton community structure across the entire plankton size spectrum. A) Average relative cell abundance of phototrophs across all metagenomes based on *psbO* metagenomic reads. (B-F) Examples of confocal microscopy detection of cyanobacteria in the 20-180 µm size fraction. From top left to bottom right, the displayed channels for each micrograph correspond to cell surface (cyan, AlexaFluor 546 dye), DNA (and the theca in dinoflagellates) (blue, Hoechst dye), cellular membranes (green, DiOC6 dye), chlorophyll autofluorescence (red), bright field, and all merged channels. The size bar at the bottom left of each microscopy image corresponds to 2.5 µm. B) *Trichodesmium* filament. C) *Calothrix* filament outside a chain of the diatom *Chaetoceros* sp. D) *Richelia* filaments inside the diatom *Eucampia cornuta*. E) Picocyanobacterial aggregate. F) Picocyanobacterial symbionts in the dinoflagellate *Ornithocercus thumii*. (G) Correlation analysis between *Trichodesmium* and *Richelia/Calothrix* quantifications by confocal microscopy and *psbO* metagenomic reads (values as rpkm: reads per kilobase per million mapped reads). Spearman rho's correlations coefficients and p-values are indicated.

Biogeography of the entire phytoplankton community using psbO *read abundance from metagenomes*

The use of *psbO* allowed us to examine the biogeography of both prokaryotic and eukaryotic phytoplankton simultaneously (Figure 8). For the sake of clarity, we also plotted the distribution of the main phytoplankton groups in the size fraction in which they were most prevalent (Figure S8). Among picoplankton, *Prochlorococcus* and *Synechococcus* were dominant at low-latitude, with the first most prevalent in oligotrophic regions and the second in coastal areas. Pelagophytes were also more prevalent in low latitudes, while chlorophytes and haptophytes were more widely distributed. At the other extreme in terms of distribution and size, diatoms were particularly abundant at high latitudes and in upwelling environments, while dinoflagellates were in general more ubiquitous. Finally, *Trichodesmium* showed a restricted distribution, identical to previous reported patterns based on *Tara* Oceans imaging datasets (Pierella Karlusich et al., 2020b).

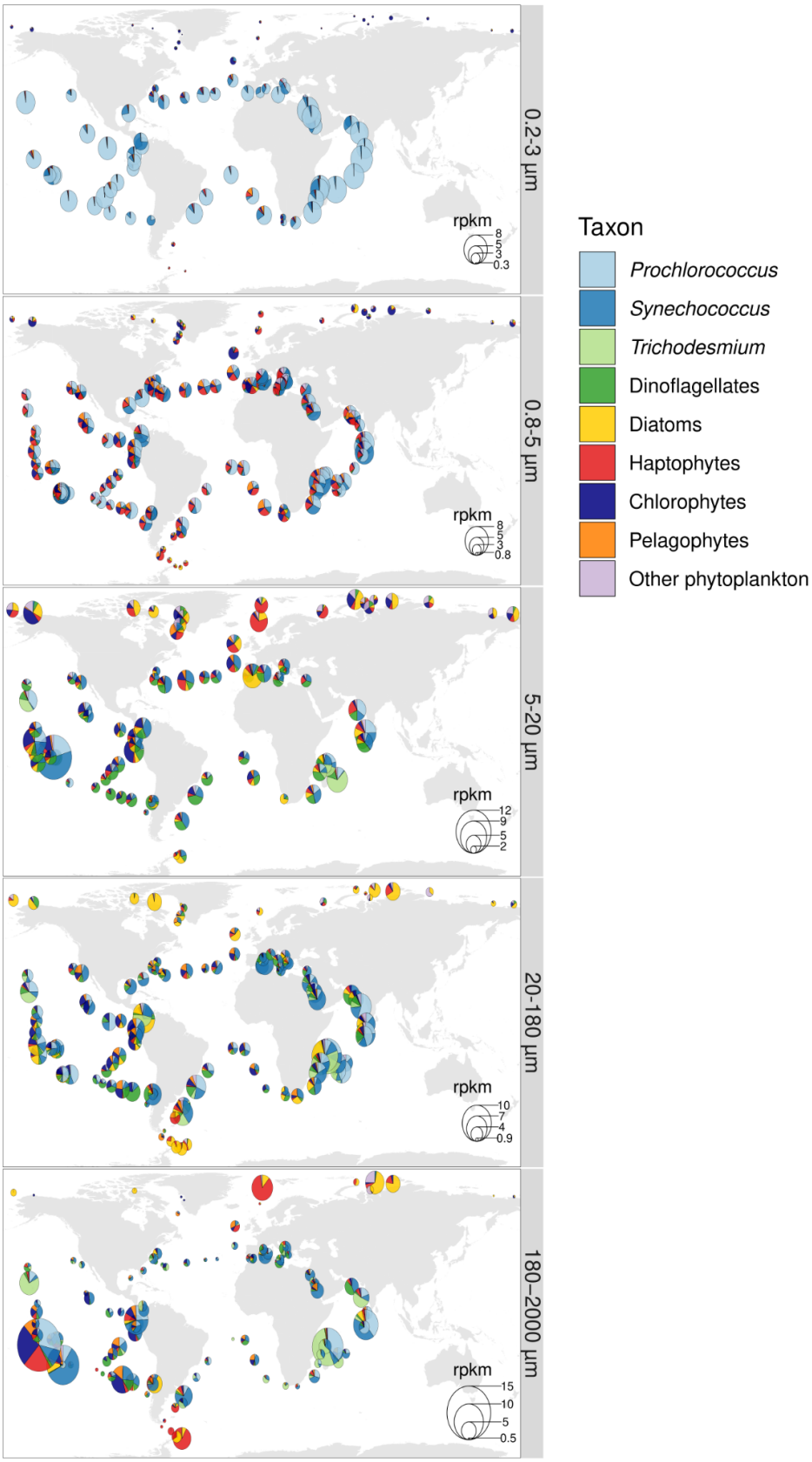

Figure 8: Global biogeographical patterns of marine phytoplankton in surface waters. The pie charts show the *psbO* relative abundance of the main cyanobacteria and eukaryotic phytoplankton in metagenomes derived from different size-fractionated samples. Values are displayed as rpkm (reads per kilobase per million mapped reads).

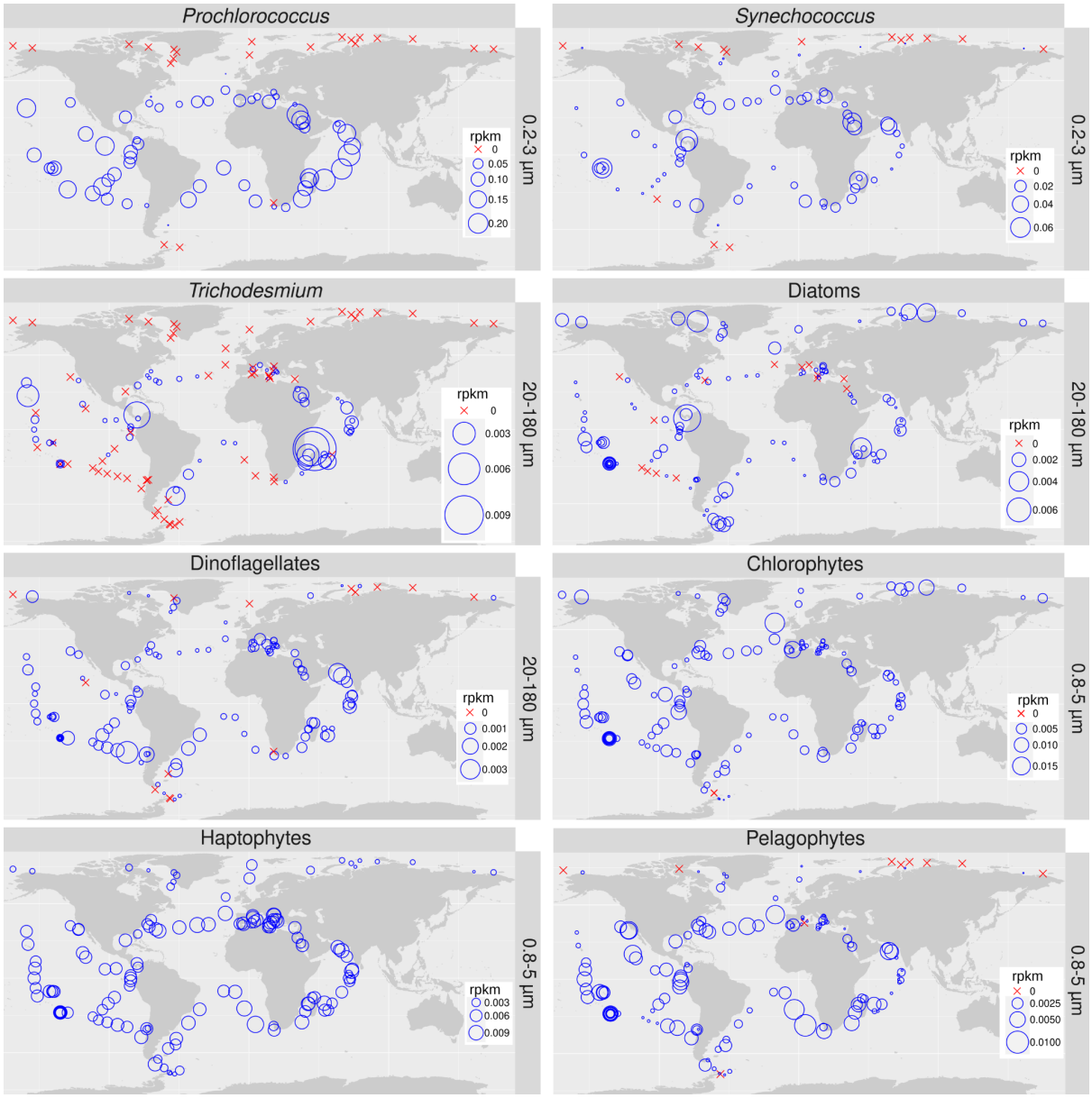

Figure S8: Global biogeographical patterns for main groups of marine phytoplankton in surface waters. The bubbles show the *psbO* relative abundance of the main cyanobacteria and eukaryotic phytoplankton in metagenomes. Values are displayed as rpkm (reads per kilobase per million mapped reads). Only the size fraction where the corresponding taxon was prevalent is shown: 0.2-3 µm for picocyanobacteria, 20-180 µm for diatoms and dinoflagellates, and 0.8-5 µm for chlorophytes, haptophytes and pelagophytes. The corresponding analysis for the whole phytoplankton community in each size fraction is displayed in Figure 8.

Discussion

In the current work, we searched for core photosynthetic, single-copy, nuclear genes in genomes and transcriptomes of cultured phytoplankton strains for their use

as marker genes. Of the five resulting candidates that we found, we decided to use *psbO* due to its lack of non-photosynthetic homologous. We applied this new approach by retrieving *psbO* sequences from the metagenomes generated by *Tara* Oceans, and we successfully validated it using the optical determinations obtained during the same expedition.

We also quantified the biases of “traditional” molecular approaches when compared to the optical methods. The 16S miTags approach avoids the PCR bias and it seems to be little affected by copy variability of the 16S gene, the plastid genome and the chloroplast, probably because only the 0.2-3 size fraction has been analysed in the current work, where most picoplankton cells certainly have a single chloroplast. It would be a future scope to analyse larger size fractions where abundant taxa have multiple plastids (e. g., centric diatoms) or divergent 16S genes difficult to align (e. g., dinoflagellates). When compared with 18S metabarcoding data, our approach yields lower abundances for diatoms and dinoflagellates at the expense of higher abundances of haptophytes, chlorophytes and pelagophytes. These results were remarkably consistent with those obtained by microscopy. To disentangle which is the effect of PCR-bias versus copy number in the patterns of 18S metabarcoding, the next step will be to generate 18S miTags from the analysed metagenomes.

Although our method improved the estimations of the relative abundances of phytoplankton, it is important to take into account that not all downstream analyses are affected by the biases introduced by traditional molecular methods, as we showed for the Shannon index. Besides cell abundance, we analysed correlations against biovolume. Based on an apparent relationship between cell size and rRNA gene copy number, it has been postulated that the relative abundance of rRNA

genes better reflect the relative biovolume of the taxa. However, we did not observe major differences between rDNA or *psbO* when correlated against optical-based biovolume. In addition to metagenomes, we also analysed *psbO* in metatranscriptomes, where dinoflagellates standed out from the rest due to their much higher *psbO* abundance ratio of mRNA to gene, and it would be interest to a analyse if this reflect higher 'photosynthetic activity'.

In addition to correcting the mentioned biases, we were also able to reveal unexpected ecological features missed by 18S metabarcoding. For example, we were able to make a trans-domain comparison, detecting picocyanobacteria in high numbers in the large size fractions, which was supported by the detection of numerous images of picocyanobacterial aggregates and endosymbionts in the *Tara* Oceans imaging dataset. Moreover, when the analysis of metagenomes includes housekeeping marker genes (in addition to photosynthetic genes), we observed small dinoflagellates (0.8-5 µm) to be mainly heterotrophic, while those in the larger size communities (>5 µm) to be mainly photosynthetic.

The very deep sequencing of the *Tara* Oceans metagenomes (between ~10^8 and ~10^9 total reads per sample) allowed us to carry out the taxonomic analysis based on a unique gene, in spite of dilution of the signal. As reduced DNA sequencing costs are leading to the replacement of amplicon-based methods by metagenome sequencing, we expect the utility of our method to increase in future years. In the short term, a barcode approach using *psbO* primers is a promising cheap alternative, although it will be subject to PCR biases.

The use of functional genes as taxonomic markers for phytoplankton has been restricted to some surveys (using plastid-encoded genes) (Man-Aharonovich et al. 2010a; Zeidner, Preston, Delong, Massana, Post, Scanlan, and Beja 2003; Paul,

Alfreider, and Wawrik 2000; Farrant et al. 2016). This is not the case for other functional groups, such as nitrogen-fixers, which are studied by targeting a gene encoding a subunit of the nitrogenase enzymatic complex (Zehr and Paerl 1998) and for which extensive reference sequence databases are now available (https://www.jzehrlab.com; (Heller et al. 2014)). To facilitate the incorporation of *psbO* into future molecular-based surveys, we released a database of >20,000 annotated p*sbO* sequences.

Overall, this work opens new perspectives for improving molecular-based evaluations of phytoplankton communities.

Acknowledgments

We would like to thank all colleagues from the *Tara* Oceans consortium as well as the *Tara* Ocean Foundation for their inspirational vision. We also acknowledge Quentin Carradec for his help with genes coding for ribosomal proteins. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Diatomic; grant agreement No. 835067). Additional funding is acknowledged from the FFEM - French Facility for Global Environment (Fonds Français pour l'Environnement Mondial), and the French Government "Investissements d'Avenir" Programmes MEMO LIFE (Grant ANR-10-LABX-54), Université de Recherche Paris Sciences et Lettres (PSL) (Grant ANR-1253 11-IDEX-0001-02), France Genomique (ANR-10-INBS-09), and OCEANOMICS (Grant ANR-11-BTBR-0008). JJPK acknowledges postdoctoral funding from the Fonds Français pour l'Environnement Mondial. This article is contribution number *** of *Tara* Oceans.

Author contributions

JJPK and CB designed the project. JJPK conducted the study and performed the primary data analysis and visualization. JJPK compiled the *psbO* gene reference catalog and EP performed the metagenomic mapping on it. RD carried out the phylogenetic-based annotation of 16S rDNA OTUs. FL, SC and CdV helped with the confocal microscopy dataset, AZ with the optical microscopy and JMG and SGA with the flow cytometry. All authors helped interpret the data. JJPK and CB wrote the paper with substantial input from all authors.

Data availability statement

All datasets analyzed for this study are of public access (Table I).

References

Acinas, S. G., L. A. Marcelino, V. Klepac-Ceraj, and M. F. Polz. 2004. "Divergence and Redundancy of 16S rRNA Sequences in Genomes with Multiple Rrn Operons." *Journal of Bacteriology*. https://doi.org/10.1128/jb.186.9.2629-2635.2004.

Adriaenssens, Evelien M., and Don A. Cowan. 2014. "Using Signature Genes as Tools to Assess Environmental Viral Ecology and Diversity." *Applied and Environmental Microbiology* 80 (15): 4470–80.

Alberti, Adriana, Julie Poulain, Stefan Engelen, Karine Labadie, Sarah Romac, Isabel Ferrera, Guillaume Albini, et al. 2017. "Viral to Metazoan Marine Plankton Nucleotide Sequences from the Tara Oceans Expedition." *Scientific Data* 4 (August): 170093.

Armbrust, E. Virginia, and E. Virginia Armbrust. n.d. "Uniparental Inheritance of Chloroplast Genomes." *The Molecular Biology of Chloroplasts and Mitochondria in Chlamydomonas*. https://doi.org/10.1007/0-306-48204-5_6.

Bachy, Charles, John R. Dolan, Purificación López-García, Philippe Deschamps, and David Moreira. 2013. "Accuracy of Protist Diversity Assessments: Morphology Compared with Cloning and Direct Pyrosequencing of 18S rRNA Genes and ITS Regions Using the Conspicuous Tintinnid Ciliates as a Case Study." *The ISME Journal* 7 (2): 244–55.

Beardall, John, Drew Allen, Jason Bragg, Zoe V. Finkel, Kevin J. Flynn, Antonietta Quigg, T. Alwyn V. Rees, Anthony Richardson, and John A. Raven. 2009. "Allometry and Stoichiometry of Unicellular, Colonial and Multicellular Phytoplankton." *The New Phytologist* 181 (2): 295–309.

Belgrano, Andrea, Andrew P. Allen, Brian J. Enquist, and James F. Gillooly. 2002. "Allometric Scaling of Maximum Population Density: A Common Rule for Marine Phytoplankton and Terrestrial Plants." *Ecology Letters*. https://doi.org/10.1046/j.1461-0248.2002.00364.x.

Blazewicz, Steven J., Romain L. Barnard, Rebecca A. Daly, and Mary K. Firestone. 2013.

"Evaluating rRNA as an Indicator of Microbial Activity in Environmental Communities: Limitations and Uses." *The ISME Journal* 7 (11): 2061–68.
Bradley, Ian M., Ameet J. Pinto, and Jeremy S. Guest. 2016. "Design and Evaluation of Illumina MiSeq-Compatible, 18S rRNA Gene-Specific Primers for Improved Characterization of Mixed Phototrophic Communities." *Applied and Environmental Microbiology*. https://doi.org/10.1128/aem.01630-16.
Calderón-Sanou, Irene, Tamara Münkemüller, Frédéric Boyer, Lucie Zinger, and Wilfried Thuiller. 2020. "From Environmental DNA Sequences to Ecological Conclusions: How Strong Is the Influence of Methodological Choices?" *Journal of Biogeography* 47 (1): 193–206.
Calvo-Díaz, A., and X. A. G. Morán. 2006. "Seasonal Dynamics of Picoplankton in Shelf Waters of the Southern Bay of Biscay." *Aquatic Microbial Ecology: International Journal* 42: 159–74.
Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST : Architecture and Applications." *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-10-421.
Campbell, B. J., L. Yu, J. F. Heidelberg, and D. L. Kirchman. 2011. "Activity of Abundant and Rare Bacteria in a Coastal Ocean." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1101405108.
Carradec, Quentin, Eric Pelletier, Corinne Da Silva, Adriana Alberti, Yoann Seeleuthner, Romain Blanc-Mathieu, Gipsi Lima-Mendez, et al. 2018. "A Global Ocean Atlas of Eukaryotic Genes." *Nature Communications* 9 (1): 373.
Caspi, Ron, Richard Billington, Carol A. Fulcher, Ingrid M. Keseler, Anamika Kothari, Markus Krummenacker, Mario Latendresse, et al. 2018. "The MetaCyc Database of Metabolic Pathways and Enzymes." *Nucleic Acids Research* 46 (D1): D633–39.
Chen, I-Min A., Ken Chu, Krishna Palaniappan, Manoj Pillay, Anna Ratner, Jinghua Huang, Marcel Huntemann, et al. 2019. "IMG/M v.5.0: An Integrated Data Management and Comparative Analysis System for Microbial Genomes and Microbiomes." *Nucleic Acids Research* 47 (D1): D666–77.
Ciccarelli, Francesca D., Tobias Doerks, Christian von Mering, Christopher J. Creevey, Berend Snel, and Peer Bork. 2006. "Toward Automatic Reconstruction of a Highly Resolved Tree of Life." *Science* 311 (5765): 1283–87.
Coleman, Annette W., and Andrea M. Nerozzi. 1999. "Temporal and Spatial Coordination of Cells with Their Plastid Component." *International Review of Cytology*. https://doi.org/10.1016/s0074-7696(08)61780-5.
Colin, Sebastien, Luis Pedro Coelho, Shinichi Sunagawa, Chris Bowler, Eric Karsenti, Peer Bork, Rainer Pepperkok, and Colomban de Vargas. 2017. "Quantitative 3D-Imaging for Cell Biology and Ecology of Environmental Microbial Eukaryotes." *eLife* 6 (October). https://doi.org/10.7554/eLife.26066.
Decelle, Johan, Sarah Romac, Rowena F. Stern, El Mahdi Bendif, Adriana Zingone, Stéphane Audic, Michael D. Guiry, et al. 2015. "PhytoREF: A Reference Database of the Plastidial 16S rRNA Gene of Photosynthetic Eukaryotes with Curated Taxonomy." *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.12401.
Decelle, Johan, Hryhoriy Stryhanyuk, Benoit Gallet, Giulia Veronesi, Matthias Schmidt, Sergio Balzano, Sophie Marro, et al. 2019. "Algal Remodeling in a Ubiquitous Planktonic Photosymbiosis." *Current Biology*. https://doi.org/10.1016/j.cub.2019.01.073.
Delmont, T. O., M. Gaia, and D. D. Hinsinger. 2020. "Functional Repertoire Convergence of Distantly Related Eukaryotic Plankton Lineages Revealed by Genome-Resolved Metagenomics." *bioRxiv*.
Delmont, Tom O., Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny Tm Lee, Michael S. Rappé, Sandra L. McLellan, Sebastian Lücker, and A. Murat Eren. 2018. "Nitrogen-Fixing Populations of Planctomycetes and Proteobacteria Are Abundant in Surface Ocean Metagenomes." *Nature Microbiology* 3 (7): 804–13.
Dorrell, Richard G., Tomonori Azuma, Mami Nomura, Guillemette Audren de Kerdrel, Lucas Paoli, Shanshan Yang, Chris Bowler, et al. 2019. "Principles of Plastid Reductive

Evolution Illuminated by Nonphotosynthetic Chrysophytes." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1819976116.
Dorrell, Richard G., and Alison G. Smith. 2011. "Do Red and Green Make Brown?: Perspectives on Plastid Acquisitions within Chromalveolates." *Eukaryotic Cell* 10 (7): 856–68.
Egge, Elianne, Lucie Bittner, Tom Andersen, Stéphane Audic, Colomban de Vargas, and Bente Edvardsen. 2013. "454 Pyrosequencing to Describe Microbial Eukaryotic Community Composition, Diversity and Relative Abundance: A Test for Marine Haptophytes." *PloS One* 8 (9): e74371.
Farrant, Gregory K., Hugo Doré, Francisco M. Cornejo-Castillo, Frédéric Partensky, Morgane Ratin, Martin Ostrowski, Frances D. Pitt, et al. 2016. "Delineating Ecologically Significant Taxonomic Units from Global Patterns of Marine Picocyanobacteria." *Proceedings of the National Academy of Sciences of the United States of America* 113 (24): E3365–74.
Field, C. B., M. J. Behrenfeld, J. T. Randerson, and P. Falkowski. 1998. "Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components." *Science* 281 (5374): 237–40.
Fuller, Nicholas J., Glen A. Tarran, Denise G. Cummings, E. Malcolm S. Woodward, Karen M. Orcutt, Marian Yallop, Florence Le Gall, and David J. Scanlan. 2006. "Molecular Analysis of Photosynthetic Picoeukaryote Community Structure along an Arabian Sea Transect." *Limnology and Oceanography*. https://doi.org/10.4319/lo.2006.51.6.2502.
Fuller, N. J., C. Campbell, D. J. Allen, F. D. Pitt, K. Zwirglmaier, F. Le Gall, D. Vaulot, and D. J. Scanlan. 2006. "Analysis of Photosynthetic Picoeukaryote Diversity at Open Ocean Sites in the Arabian Sea Using a PCR Biased towards Marine Algal Plastids." *Aquatic Microbial Ecology*. https://doi.org/10.3354/ame043079.
Gasol, Josep M., and Xosé Anxelu G. Morán. 2015. "Flow Cytometric Determination of Microbial Abundances and Its Use to Obtain Indices of Community Structure and Relative Activity." In *Springer Protocols Handbooks*, 159–87. Berlin, Heidelberg: Springer Berlin Heidelberg.
Godhe, A., M. E. Asplund, K. Harnstrom, V. Saravanan, A. Tyagi, and I. Karunasagar. 2008. "Quantification of Diatom and Dinoflagellate Biomasses in Coastal Marine Seawater Samples by Real-Time PCR." *Applied and Environmental Microbiology*. https://doi.org/10.1128/aem.01298-08.
Gong, Weida, and Adrian Marchetti. 2019. "Estimation of 18S Gene Copy Number in Marine Eukaryotic Plankton Using a Next-Generation Sequencing Approach." *Frontiers in Marine Science*. https://doi.org/10.3389/fmars.2019.00219.
Green, Beverley R. 2011. "Chloroplast Genomes of Photosynthetic Eukaryotes." *The Plant Journal*. https://doi.org/10.1111/j.1365-313x.2011.04541.x.
Grigoriev, Igor V., Richard D. Hayes, Sara Calhoun, Bishoy Kamel, Alice Wang, Steven Ahrendt, Sergey Dusheyko, et al. 2021. "PhycoCosm, a Comparative Algal Genomics Resource." *Nucleic Acids Research* 49 (D1): D1004–11.
Guidi, Lionel, Lars Stemmann, George A. Jackson, Frédéric Ibanez, Hervé Claustre, Louis Legendre, Marc Picheral, and Gabriel Gorskya. 2009. "Effects of Phytoplankton Community on Production, Size, and Export of Large Aggregates: A World-Ocean Analysis." *Limnology and Oceanography* 54 (6): 1951–63.
Guillou, Laure, Dipankar Bachar, Stéphane Audic, David Bass, Cédric Berney, Lucie Bittner, Christophe Boutte, et al. 2013. "The Protist Ribosomal Reference Database (PR2): A Catalog of Unicellular Eukaryote Small Sub-Unit rRNA Sequences with Curated Taxonomy." *Nucleic Acids Research* 41 (Database issue): D597–604.
Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and Olivier Gascuel. 2010. "New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0." *Systematic Biology* 59 (3): 307–21.
Heller, Philip, H. James Tripp, Kendra Turk-Kubo, and Jonathan P. Zehr. 2014. "ARBitrator: A Software Pipeline for on-Demand Retrieval of Auto-Curated nifH Sequences from

GenBank." *Bioinformatics* 30 (20): 2883–90.
Hingamp, Pascal, Nigel Grimsley, Silvia G. Acinas, Camille Clerissi, Lucie Subirana, Julie Poulain, Isabel Ferrera, et al. 2013. "Exploring Nucleo-Cytoplasmic Large DNA Viruses in Tara Oceans Microbial Metagenomes." *The ISME Journal* 7 (9): 1678–95.
Hiramatsu, Takayoshi, Soichi Nakamura, Osami Misumi, Tsuneyoshi Kuroiwa, and Soichi Nakamura. 2006. "MORPHOLOGICAL CHANGES IN MITOCHONDRIAL AND CHLOROPLAST NUCLEOIDS AND MITOCHONDRIA DURING THE CHLAMYDOMONAS REINHARDTII (CHLOROPHYCEAE) CELL CYCLE." *Journal of Phycology*. https://doi.org/10.1111/j.1529-8817.2006.00259.x.
Ibarbalz, Federico M., Nicolas Henry, Manoela C. Brandão, Séverine Martini, Greta Busseni, Hannah Byrne, Luis Pedro Coelho, et al. 2019. "Global Trends in Marine Plankton Diversity across Kingdoms of Life." *Cell* 179 (5): 1084–97.e21.
Jaffe, Alexander L., Cindy J. Castelle, Christopher L. Dupont, and Jillian F. Banfield. 2019. "Lateral Gene Transfer Shapes the Distribution of RuBisCO among Candidate Phyla Radiation Bacteria and DPANN Archaea." *Molecular Biology and Evolution* 36 (3): 435–46.
Kamikawa, Ryoma, Daniel Moog, Stefan Zauner, Goro Tanifuji, Ken-Ichiro Ishida, Hideaki Miyashita, Shigeki Mayama, et al. 2017. "A Non-Photosynthetic Diatom Reveals Early Steps of Reductive Evolution in Plastids." *Molecular Biology and Evolution* 34 (9): 2355–66.
Kanehisa, M. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." *Nucleic Acids Research*. https://doi.org/10.1093/nar/28.1.27.
Katoh, Kazutaka, and Hiroyuki Toh. 2008. "Recent Developments in the MAFFT Multiple Sequence Alignment Program." *Briefings in Bioinformatics* 9 (4): 286–98.
Keeling, Patrick J., Fabien Burki, Heather M. Wilcox, Bassem Allam, Eric E. Allen, Linda A. Amaral-Zettler, E. Virginia Armbrust, et al. 2014. "The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing." *PLoS Biology* 12 (6): e1001889.
Kembel, Steven W., Martin Wu, Jonathan A. Eisen, and Jessica L. Green. 2012. "Incorporating 16S Gene Copy Number Information Improves Estimates of Microbial Diversity and Abundance." *PLoS Computational Biology* 8 (10): e1002743.
Kirkham, Amy R., Ludwig E. Jardillier, Ana Tiganescu, John Pearman, Mikhail V. Zubkov, and David J. Scanlan. 2011. "Basin-Scale Distribution Patterns of Photosynthetic Picoeukaryotes along an Atlantic Meridional Transect." *Environmental Microbiology*. https://doi.org/10.1111/j.1462-2920.2010.02403.x.
Kirkham, Amy R., Cécile Lepère, Ludwig E. Jardillier, Fabrice Not, Heather Bouman, Andrew Mead, and David J. Scanlan. 2013. "A Global Perspective on Marine Photosynthetic Picoeukaryote Community Structure." *The ISME Journal*. https://doi.org/10.1038/ismej.2012.166.
Kono, Takunari, Sandhya Mehrotra, Chikako Endo, Natsuko Kizu, Mami Matusda, Hiroyuki Kimura, Eiichi Mizohata, et al. 2017. "A RuBisCO-Mediated Carbon Metabolic Pathway in Methanogenic Archaea." *Nature Communications*. https://doi.org/10.1038/ncomms14007.
Koumandou, V. L., and Christopher J. Howe. 2007. "The Copy Number of Chloroplast Gene Minicircles Changes Dramatically with Growth Phase in the Dinoflagellate Amphidinium Operculatum." *Protist*. https://doi.org/10.1016/j.protis.2006.08.003.
Lavrinienko, Anton, Toni Jernfors, Janne J. Koskimäki, Anna Maria Pirttilä, and Phillip C. Watts. 2021. "Does Intraspecific Variation in rDNA Copy Number Affect Analysis of Microbial Communities?" *Trends in Microbiology* 29 (1): 19–27.
Lebrun, Evelyne, Joanne M. Santini, Myriam Brugna, Anne-Lise Ducluzeau, Soufian Ouchane, Barbara Schoepp-Cothenet, Frauke Baymann, and Wolfgang Nitschke. 2006. "The Rieske Protein: A Case Study on the Pitfalls of Multiple Sequence Alignments and Phylogenetic Reconstruction." *Molecular Biology and Evolution* 23 (6): 1180–91.
Lepère, Cécile, Daniel Vaulot, and David J. Scanlan. 2009. "Photosynthetic Picoeukaryote

Community Structure in the South East Pacific Ocean Encompassing the Most Oligotrophic Waters on Earth." *Environmental Microbiology*. https://doi.org/10.1111/j.1462-2920.2009.02015.x.
Liu, Z., C. Lozupone, M. Hamady, F. D. Bushman, and R. Knight. 2007. "Short Pyrosequencing Reads Suffice for Accurate Microbial Community Analysis." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkm541.
Li, W., and A. Godzik. 2006. "Cd-Hit: A Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences." *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btl158.
Li, W. K. W. 2002. "Macroecological Patterns of Phytoplankton in the Northwestern North Atlantic Ocean." *Nature* 419 (6903): 154–57.
Logares, Ramiro, Stephane Audic, Sebastien Santini, Massimo C. Pernice, Colomban de Vargas, and Ramon Massana. 2012. "Diversity Patterns and Activity of Uncultured Marine Heterotrophic Flagellates Unveiled with Pyrosequencing." *The ISME Journal* 6 (10): 1823–33.
Logares, Ramiro, Shinichi Sunagawa, Guillem Salazar, Francisco M. Cornejo-Castillo, Isabel Ferrera, Hugo Sarmento, Pascal Hingamp, et al. 2014. "Metagenomic 16S rDNA Illumina Tags Are a Powerful Alternative to Amplicon Sequencing to Explore Diversity and Structure of Microbial Communities." *Environmental Microbiology* 16 (9): 2659–71.
Louca, Stilianos, Michael Doebeli, and Laura Wegener Parfrey. 2018. "Correcting for 16S rRNA Gene Copy Numbers in Microbiome Surveys Remains an Unsolved Problem." *Microbiome* 6 (1): 41.
Mäki, Anita, Pauliina Salmi, Anu Mikkonen, Anke Kremp, and Marja Tiirola. 2017. "Sample Preservation, DNA or RNA Extraction and Data Analysis for High-Throughput Phytoplankton Community Sequencing." *Frontiers in Microbiology* 8 (September): 1848.
Man-Aharonovich, Dikla, Alon Philosof, Benjamin C. Kirkup, Florence Le Gall, Tali Yogev, Ilana Berman-Frank, Martin F. Polz, Daniel Vaulot, and Oded Béjà. 2010a. "Diversity of Active Marine Picoeukaryotes in the Eastern Mediterranean Sea Unveiled Using Photosystem-II psbA Transcripts." *The ISME Journal*. https://doi.org/10.1038/ismej.2010.25.
———. 2010b. "Diversity of Active Marine Picoeukaryotes in the Eastern Mediterranean Sea Unveiled Using Photosystem-II psbA Transcripts." *The ISME Journal* 4 (8): 1044–52.
McDonald, S. M., D. Sarno, D. J. Scanlan, and A. Zingone. 2007. "Genetic Diversity of Eukaryotic Ultraphytoplankton in the Gulf of Naples during an Annual Cycle." *Aquatic Microbial Ecology*. https://doi.org/10.3354/ame01148.
Medinger, Ralph, Viola Nolte, Ram Vinay Pandey, Steffen Jost, Birgit Ottenwälder, Christian Schlötterer, and Jens Boenigk. 2010. "Diversity in a Hidden World: Potential and Limitation of next-Generation Sequencing for Surveys of Molecular Diversity of Eukaryotic Microorganisms." *Molecular Ecology*. https://doi.org/10.1111/j.1365-294x.2009.04478.x.
Milanese, Alessio, Daniel R. Mende, Lucas Paoli, Guillem Salazar, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Pascal Hingamp, et al. 2019. "Microbial Abundance, Activity and Population Genomic Profiling with mOTUs2." *Nature Communications* 10 (1): 1014.
Mohamed, M. E., A. Zaar, C. Ebenau-Jehle, and G. Fuchs. 2001. "Reinvestigation of a New Type of Aerobic Benzoate Metabolism in the Proteobacterium Azoarcus Evansii." *Journal of Bacteriology* 183 (6): 1899–1908.
Moore, Robert B., Miroslav Oborník, Jan Janouskovec, Tomás Chrudimský, Marie Vancová, David H. Green, Simon W. Wright, et al. 2008. "A Photosynthetic Alveolate Closely Related to Apicomplexan Parasites." *Nature* 451 (7181): 959–63.
Nakayama, Takuro, Mami Nomura, Yoshihito Takano, Goro Tanifuji, Kogiku Shiba, Kazuo Inaba, Yuji Inagaki, and Masakado Kawata. 2019. "Single-Cell Genomics Unveiled a Cryptic Cyanobacterial Lineage with a Worldwide Distribution Hidden by a Dinoflagellate Host." *Proceedings of the National Academy of Sciences of the United States of America* 116 (32): 15973–78.

Obiol, Aleix, Caterina R. Giner, Pablo Sánchez, Carlos M. Duarte, Silvia G. Acinas, and Ramon Massana. 2020. “A Metagenomic Assessment of Microbial Eukaryotic Diversity in the Global Ocean.” *Molecular Ecology Resources* 20 (3). https://doi.org/10.1111/1755-0998.13147.

Oldenburg, Delene J., and Arnold J. Bendich. 2004. “Changes in the Structure of DNA Molecules and the Amount of DNA per Plastid during Chloroplast Development in Maize.” *Journal of Molecular Biology* 344 (5): 1311–30.

Parada, Alma E., David M. Needham, and Jed A. Fuhrman. 2016. “Every Base Matters: Assessing Small Subunit rRNA Primers for Marine Microbiomes with Mock Communities, Time Series and Global Field Samples.” *Environmental Microbiology* 18 (5): 1403–14.

Paradis, Emmanuel, and Klaus Schliep. 2019. “Ape 5.0: An Environment for Modern Phylogenetics and Evolutionary Analyses in R.” *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty633.

Paul, J. H., A. Alfreider, and B. Wawrik. 2000. “Micro- and Macrodiversity in rbcL Sequences in Ambient Phytoplankton Populations from the Southeastern Gulf of Mexico.” *Marine Ecology Progress Series*. https://doi.org/10.3354/meps198009.

Pawlowski, Jan, Stéphane Audic, Sina Adl, David Bass, Lassaâd Belbahri, Cédric Berney, Samuel S. Bowser, et al. 2012. “CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms.” *PLoS Biology*. https://doi.org/10.1371/journal.pbio.1001419.

Pawlowski, J., F. Lejzerowicz, L. Apotheloz-Perret-Gentil, J. Visco, and P. Esling. 2016. “Protist Metabarcoding and Environmental Biomonitoring: Time for Change.” *European Journal of Protistology* 55 (Pt A): 12–25.

Pesant, Stéphane, Fabrice Not, Marc Picheral, Stefanie Kandels-Lewis, Noan Le Bescot, Gabriel Gorsky, Daniele Iudicone, et al. 2015. “Open Science Resources for the Discovery and Analysis of Tara Oceans Data.” *Scientific Data* 2 (May): 150023.

Pierella Karlusich, Juan J., Romina D. Ceccoli, Martín Graña, Héctor Romero, and Néstor Carrillo. 2015. “Environmental Selection Pressures Related to Iron Utilization Are Involved in the Loss of the Flavodoxin Gene from the Plant Genome.” *Genome Biology and Evolution* 7 (3): 750–67.

Pierella Karlusich, Juan José, and Néstor Carrillo. 2017. “Evolution of the Acceptor Side of Photosystem I: Ferredoxin, Flavodoxin, and Ferredoxin-NADP Oxidoreductase.” *Photosynthesis Research*. https://doi.org/10.1007/s11120-017-0338-2.

Pierella Karlusich, Juan José, Federico M. Ibarbalz, and Chris Bowler. 2020a. Phytoplankton in the *Tara* Ocean. *Annual Review of Marine Science* 12:233-265. https://doi.org/10.1146/annurev-marine-010419-010706

Pierella Karlusich, Juan José, Eric Pelletier, Madeline Carsique, Etienne Dvorak, Sébastien Colin, Marc Picheral, Rainer Pepperkok, et al. 2020b. “Global Distribution Patterns of Marine Nitrogen-Fixers by Imaging and Molecular Methods.” https://doi.org/10.1101/2020.10.17.343731.

Pinto, Ameet J., and Lutgarde Raskin. 2012. “PCR Biases Distort Bacterial and Archaeal Community Structure in Pyrosequencing Datasets.” *PloS One* 7 (8): e43093.

Polz, M. F., and C. M. Cavanaugh. 1998. “Bias in Template-to-Product Ratios in Multitemplate PCR.” *Applied and Environmental Microbiology* 64 (10): 3724–30.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. “The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools.” *Nucleic Acids Research* 41 (Database issue): D590–96.

Rusch, Douglas B., Aaron L. Halpern, Granger Sutton, Karla B. Heidelberg, Shannon Williamson, Shibu Yooseph, Dongying Wu, et al. 2007. “The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific.” *PLoS Biology*. https://doi.org/10.1371/journal.pbio.0050077.

Salazar, Guillem, Lucas Paoli, Adriana Alberti, Jaime Huerta-Cepas, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Christopher M. Field, et al. 2019. “Gene Expression

Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome." *Cell*. https://doi.org/10.1016/j.cell.2019.10.014.
Saldarriaga, J. F., F. J. Taylor, P. J. Keeling, and T. Cavalier-Smith. 2001. "Dinoflagellate Nuclear SSU rRNA Phylogeny Suggests Multiple Plastid Losses and Replacements." *Journal of Molecular Evolution* 53 (3): 204–13.
Shi, Xiao Li, Cécile Lepère, David J. Scanlan, and Daniel Vaulot. 2011. "Plastid 16S rRNA Gene Diversity among Eukaryotic Picophytoplankton Sorted by Flow Cytometry from the South Pacific Ocean." *PloS One* 6 (4): e18979.
Starke, Robert, Victor Satler Pylro, and Daniel Kumazawa Morais. 2020. "16S rRNA Gene Copy Number Normalization Does Not Provide More Reliable Conclusions in Metataxonomic Surveys." *Microbial Ecology*. https://doi.org/10.1007/s00248-020-01586-7.
Sunagawa, Shinichi, Tara Oceans Coordinators, Silvia G. Acinas, Peer Bork, Chris Bowler, Damien Eveillard, Gabriel Gorsky, et al. 2020. "Tara Oceans: Towards Global Ocean Ecosystems Biology." *Nature Reviews Microbiology*. https://doi.org/10.1038/s41579-020-0364-5.
Sunagawa, Shinichi, Daniel R. Mende, Georg Zeller, Fernando Izquierdo-Carrasco, Simon A. Berger, Jens Roat Kultima, Luis Pedro Coelho, et al. 2013. "Metagenomic Species Profiling Using Universal Phylogenetic Marker Genes." *Nature Methods* 10 (12): 1196–99.
Tabita, F. Robert, Thomas E. Hanson, Sriram Satagopan, Brian H. Witte, and Nathan E. Kreel. 2008. "Phylogenetic and Evolutionary Relationships of RubisCO and the RubisCO-like Proteins and the Functional Lessons Provided by Diverse Molecular Forms." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 363 (1504): 2629–40.
Thompson, Anne W., Rachel A. Foster, Andreas Krupke, Brandon J. Carter, Niculina Musat, Daniel Vaulot, Marcel M. M. Kuypers, and Jonathan P. Zehr. 2012. "Unicellular Cyanobacterium Symbiotic with a Single-Celled Eukaryotic Alga." *Science* 337 (6101): 1546–50.
Ullah, Hadayet, Ivan Nagelkerken, Silvan U. Goldenberg, and Damien A. Fordham. 2018. "Climate Change Could Drive Marine Food Web Collapse through Altered Trophic Flows and Cyanobacterial Proliferation." *PLoS Biology* 16 (1): e2003446.
Vargas, Colomban de, Stéphane Audic, Nicolas Henry, Johan Decelle, Frédéric Mahé, Ramiro Logares, Enrique Lara, et al. 2015. "Ocean Plankton. Eukaryotic Plankton Diversity in the Sunlit Ocean." *Science* 348 (6237): 1261605.
Veit, Sebastian, Kazuki Takeda, Yuichi Tsunoyama, Frauke Baymann, Reinat Nevo, Ziv Reich, Matthias Rögner, Kunio Miki, and Sascha Rexroth. 2016. "Structural and Functional Characterisation of the Cyanobacterial PetC3 Rieske Protein Family." *Biochimica et Biophysica Acta* 1857 (12): 1879–91.
Větrovský, Tomáš, and Petr Baldrian. 2013. "The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses." *PloS One* 8 (2): e57923.
Wang, Jiao, Shanshan Chu, Ying Zhu, Hao Cheng, and Deyue Yu. 2015. "Positive Selection Drives Neofunctionalization of the UbiA Prenyltransferase Gene Family." *Plant Molecular Biology* 87 (4-5): 383–94.
Wear, Emma K., Elizabeth G. Wilbanks, Craig E. Nelson, and Craig A. Carlson. 2018. "Primer Selection Impacts Specific Population Abundances but Not Community Dynamics in a Monthly Time-Series 16S rRNA Gene Amplicon Analysis of Coastal Marine Bacterioplankton." *Environmental Microbiology*. https://doi.org/10.1111/1462-2920.14091.
Zehr, Jonathan P., and Hans Paerl. 1998. "Nitrogen Fixation in the Marine Environment: Genetic Potential and Nitrogenase Expression." *Molecular Approaches to the Study of the Ocean*. https://doi.org/10.1007/978-94-011-4928-0_13.
Zeidner, Gil, Christina M. Preston, Edward F. Delong, Ramon Massana, Anton F. Post, David J. Scanlan, and Oded Beja. 2003. "Molecular Diversity among Marine

Picophytoplankton as Revealed by psbA Analyses." *Environmental Microbiology*. https://doi.org/10.1046/j.1462-2920.2003.00403.x.

Zeidner, Gil, Christina M. Preston, Edward F. Delong, Ramon Massana, Anton F. Post, David J. Scanlan, and Oded Béjà. 2003. "Molecular Diversity among Marine Picophytoplankton as Revealed by psbA Analyses." *Environmental Microbiology* 5 (3): 212–16.

Zhu, Fei, Ramon Massana, Fabrice Not, Dominique Marie, and Daniel Vaulot. 2005. "Mapping of Picoeucaryotes in Marine Ecosystems with Quantitative PCR of the 18S rRNA Gene." *FEMS Microbiology Ecology* 52 (1): 79–92.