

1

1 A robust approach to estimate relative phytoplankton cell 2 abundance from metagenomes

3

4 Running title: Relative phytoplankton abundance from metagenomes

5

6 Juan José Pierella Karlusich^{1,2 *}, Eric Pelletier^{2,3}, Lucie Zinger^{1,2}, Fabien Lombard^{2,4,5},

7 Adriana Zingone⁶, Sébastien Colin^{7,8,9}, Josep M. Gasol¹⁰, Richard G. Dorrell¹,

8 Eleonora Scalco⁶, Silvia G. Acinas¹⁰, Patrick Wincker^{2,3}, Colomban de Vargas^{2,8},

9 Chris Bowler^{1,2 *}

10 ¹ Institut de Biologie de l'ENS (IBENS), Département de biologie, École normale supérieure,
11 CNRS, INSERM, Université PSL, 75005 Paris, France

12 ² CNRS Research Federation for the study of Global Ocean Systems Ecology and Evolution,
13 FR2022/Tara Oceans GOSEE, 3 rue Michel-Ange, 75016 Paris, France

14 ³ Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry,
15 Université Paris-Saclay, 91057 Evry, France

16 ⁴ Sorbonne Universités, CNRS, Laboratoire d'Océanographie de Villefranche (LOV), 06230
17 Villefranche-sur-Mer, France

18 ⁵ Institut Universitaire de France (IUF), Paris, France

19 ⁶ Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy

20 ⁷ European Molecular Biology Laboratory, Heidelberg, Germany

21 ⁸ Sorbonne Université, CNRS, Station Biologique de Roscoff, UMR 7144, ECOMAP, 29680
22 Roscoff, France

23 ⁹ Max Planck Institute for Developmental Biology, Tübingen, Germany

24 ¹⁰ Department of Marine Biology and Oceanography, Institut de Ciències del Mar, CSIC,
25 Barcelona, Spain

26

27 *corresponding authors: Juan José Pierella Karlusich (pierella@biologie.ens.fr) and

28 Chris Bowler (cbowler@biologie.ens.fr)

29

30

31

Abstract

Phytoplankton account for >45% of global primary production, and have an enormous impact on aquatic food webs and on the entire Earth System. Their members are found among prokaryotes (cyanobacteria) and multiple eukaryotic lineages containing chloroplasts. Phytoplankton communities are generally studied by PCR amplification of bacterial (16S), nuclear (18S) or chloroplastic (16S) rRNA marker genes from DNA extracted from environmental samples. 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 genes do not provide insights into metabolic traits such as photosynthesis. In addition, rRNA marker genes fail to capture both cyanobacteria and photosynthetic eukaryotes simultaneously. Here, we targeted the photosynthetic gene *psbO* from metagenomes to circumvent these limitations: the method is PCR-free, and the gene is universally and exclusively present in photosynthetic prokaryotes and eukaryotes, mainly in one copy per genome. We applied and validated this new strategy with the *Tara* Oceans datasets, and showed 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 dinoflagellates. To facilitate the incorporation of *psbO* in molecular-based surveys, we compiled a curated database of >18,000 unique sequences. Overall, *psbO* appears to be a promising new gene marker for molecular-based evaluations of entire phytoplankton communities.

5

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

58

59 Introduction

60 Photosynthetic plankton, or phytoplankton, consist of unicellular organisms of
61 diverse evolutionary history and ecology. They are responsible for more than 45% of
62 Earth's primary production (Field, Behrenfeld, Randerson, & Falkowski, 1998), fueling
63 aquatic food webs, microbial decomposition, and the global ocean biological carbon
64 pump (Guidi et al., 2009). They include prokaryotes (cyanobacteria) and multiple
65 eukaryotic lineages that acquired photosynthesis either through the primary
66 endosymbiosis of cyanobacteria, or (and predominantly) through secondary and
67 higher endosymbioses of eukaryotic algae (Pierella Karlusich, Ibarbalz, & Bowler,
68 2020). They display a broad body size spectrum, from less than 1 micron (e.g.,
69 *Prochlorococcus*, *Ostreococcus*) to several millimetres (e.g., *Trichodesmium*
70 colonies, colonial green algae, and chain-forming diatoms), either due to cell size
71 variation, aggregation or symbioses (Beardall et al., 2009). This size variability partly
72 explains their different roles in the food web and in the biological carbon pump. For
73 example, cyanobacteria are generally thought to be recycled within the microbial
74 loop, whereas larger eukaryotic phytoplankton are usually considered more important
75 in energy transfer to higher trophic levels (through grazing by small protists,
76 zooplankton, and/or larvae (Ullah, Nagelkerken, Goldenberg, & Fordham, 2018)) and
77 in sequestering atmospheric CO₂ to the ocean interior through gravitational sinking of
78 particles (Guidi et al., 2009).

6

Surveys of the structure and composition of microbial communities are typically performed 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 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 (prokaryote and chloroplast) or 18S (eukaryotic nuclear) rRNA genes that are by far the most represented in reference databases (Guillou et al., 2013; Pawlowski et al., 2012; 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, Yu, Heidelberg, & Kirchman, 2011; Logares et al., 2012), but more recent studies have indicated severe limitations of this concept and only mRNA can be considered as an indicator of the metabolic state (Blazewicz, Barnard, Daly, & Firestone, 2013).

Although rRNA gene metabarcoding is widely used, it has multiple limitations (in addition to the error sources during DNA extraction or sequencing that also affect other molecular methods). 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, & Fuhrman, 2016; Polz & Cavanaugh, 1998; Wear, Wilbanks, Nelson, & Carlson, 2018) or 18S rRNA gene markers (Bradley, Pinto, & Guest, 2016). Shotgun sequencing is a PCR-free alternative and consists of the detection of these marker genes in metagenomes (Liu, Lozupone, Hamady, Bushman, & Knight, 2007; Logares et al., 2014; Obiol et al., 2020).

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 (Acinas, Marcelino, Klepac-Ceraj, & Polz, 2004; Kembel, Wu, Eisen, & Green, 2012; Větrovský & Baldrian, 2013), protists can differ by >5 orders of magnitude in their 18S rRNA gene copy numbers, from 33,000 in dinoflagellates to one in small chlorophytes (de Vargas et al., 2015; Godhe et al., 2008; Mäki, Salmi, Mikkonen, Kremp, & Tirola, 2017; Zhu, Massana, Not, Marie, & Vaulot, 2005). Due to a positive association between rRNA gene copy number and cell size, it was proposed that the rRNA gene metabarcoding reads reflect the relative biovolume proportion for a given taxon. Biovolume is a proxy of biomass, which is a relevant variable for studies of energy and matter fluxes such as food web structures and biogeochemical cycles. However, there is still little consensus for the use of rRNA gene as a biovolume estimator due to the poor correlations reported in many studies (Lamb et al., 2019; Lavrinienko, Jernfors, Koskimäki, Pirttilä, & Watts, 2021; Santoferrara, 2019; van der Loos & Nijland, 2021). Instead, there are attempts to infer relative cell abundances from rRNA gene metabarcoding by correcting the copy number variation. Although the copy number remains unknown for most microbial species, its assessment in different organisms could lead to the establishment of correction factors by assuming that the copy number is phylogenetically conserved. These approaches were applied for 16S rRNA gene in bacteria, but their accuracy is limited for taxa with no close representatives in reference phylogenies (Kembel et al., 2012; Louca, Doebeli, & Parfrey, 2018; Starke, Pylro, & Morais, 2020). In 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

126 of the haptophyte *Emiliana huxleyi* (Gong and Marchetti 2019). In addition, there are
127 major difficulties for generating a comprehensive database of 18S rRNA copy
128 numbers (Gong & Marchetti, 2019).

129 Finally, functional traits such as photosynthesis cannot be inferred solely from
130 rRNA genes or other housekeeping markers, whereby their knowledge is limited to a
131 restricted number of taxa known from experts and the literature. Indeed, while
132 photosynthesis occurs in almost all cyanobacteria (except a few symbiotic lineages
133 that have lost it (Nakayama et al., 2014; Thompson et al., 2012)), it is not necessarily
134 conserved within protist taxa, such as dinoflagellates - of which only around half of
135 known species are photosynthetic (Dorrell & Smith, 2011; Saldarriaga, Taylor,
136 Keeling, & Cavalier-Smith, 2001) - , chrysophytes (Dorrell et al., 2019; Dorrell &
137 Smith, 2011) and apicomplexa (Moore et al., 2008). This is an important issue
138 because we still do not know how extended among related lineages are the
139 independent events of chloroplast gains and losses or the extent of loss of
140 photosynthesis with retention of the plastids. Thus, it is not possible to annotate the
141 photosynthesis trait to those sequences whose taxonomic affiliation is, for example,
142 “unknown dinoflagellate”.

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

150 genes are extremely divergent (Green, 2011), and this approach can still capture
 151 non-photosynthetic plastids and kleptoplastids (functional plastids temporarily
 152 retained from ingested algal prey). Plastid-encoded markers directly involved in
 153 photosynthesis have also been used, such as *psbA* and *rbcL* (Man-Aharonovich et
 154 al., 2010a; Paul, Alfreider, & Wawrik, 2000; Zeidner, Preston, Delong, Massana,
 155 Post, Scanlan, & Beja, 2003). The *psbA* gene encodes the D1 protein of
 156 photosystem II and is also found in cyanophages (viruses) and the used primers
 157 target essentially the cyanobacterial and cyanophage sequences (Adriaenssens &
 158 Cowan, 2014). The *rbcL* gene encodes the large subunit of the ribulose-1,5-
 159 diphosphate carboxylase/oxygenase (RuBisCO). There are multiple *rbcL* types, even
 160 in non-photosynthetic organisms, and the gene location varies: form I is plastid
 161 encoded in plants and most photosynthetic protists (and is present in cyanobacteria)
 162 while form II is nuclear-encoded in peridinin dinoflagellates and chromerids (and is
 163 also present in proteobacteria) (Tabita, Hanson, Satagopan, Witte, & Kreel, 2008).
 164 The different *rbcL* variants thus prevent its use for covering the whole phytoplankton
 165 community.

166 Plastid-encoded genes (16S rRNA, *psbA*, *rbcL*) are affected by copy number
 167 variability among taxa not only at the level of gene copies (for example, four 16S
 168 rRNA gene copies in the plastid genome of the euglenophyte *Euglena gracilis* and six
 169 in the prasinophyte *Pedinomonas minor* (Decelle et al., 2015)), but also at the level of
 170 plastid genomes per plastid, and plastids per cell. The plastid number per cell varies
 171 from one or a few in most microalgal species to more than 100 in many centric
 172 diatoms (Decelle et al., 2015). In addition, this varies according to biotic interactions,
 173 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 during cell cycle progression (Armbrust & Virginia Armbrust, 1998; Coleman & Nerozzi, 1999; Hiramatsu, Nakamura, Misumi, Kuroiwa, & Nakamura, 2006; Koumandou & Howe, 2007; Oldenburg & Bendich, 2004). 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, gene metabarcoding 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 (Bachy, Dolan, López-García, Deschamps, & Moreira, 2013; Egge et al., 2013; Kembel et al., 2012; Mäki et al., 2017; Medinger et al., 2010; Pinto & 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 magnitude among protist taxa with regard to their relative abundances (Bachy et al., 2013; Egge et al., 2013; Mäki et al., 2017; Medinger et al., 2010; J. Pawlowski, Lejzerowicz, Apotheloz-Perret-Gentil, Visco, &

Esling, 2016). Most of these studies focused on the biases generated by primers and copy-number variations, but not on uncertainties in assigning photosynthetic potential (e.g., differentiating between functionally photosynthetic and secondarily non-photosynthetic species).

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 the manganese-stabilising polypeptide of the photosystem II oxygen evolving complex, and is essential for photosynthetic activity and has the additional advantage of lacking any non-photosynthetic homologs. We applied and validated this new strategy with the *Tara* Oceans datasets (Table I). 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

assignment of photosynthetic capacity based on the 18S rRNA gene. Finally, we show 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 (M. Kanehisa, 2000) database 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” program) 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

246 were considered for establishing the gene absence. This survey was also used for
247 determining gene copy number variation.

248 This survey resulted in a list of five genes that are core, nuclear-encoded and
249 present in low-copy numbers (Table II). For selecting a gene marker of phytoplankton
250 among them, we carried out a deeper sequence analysis to detect non-
251 photosynthetic homologs and to see if the phylogeny reflects the evolutionary history
252 of cyanobacteria and endosymbiosis. We first performed a sequence similarity
253 search using HMMer version 3.2.1 with gathering threshold option (<http://hmmer.org/>)
254 for the corresponding Pfam domain against the translated sequenced genomes and
255 transcriptomes from PhycoCosm and MMETSP as well as in the whole IMG
256 database (including viruses, archaea, bacteria and non-photosynthetic eukaryotes).
257 The Pfams used in the search were: MSP (PF01716) for PsbO, Rieske (PF00355) for
258 PetC, PRK (PF00485) for phosphoribulokinase, UbiA (PF01040) for chlorophyll-*a*
259 synthase, and NAD_binding_1 (PF00175) for ferredoxin:NADP⁺ reductase. CDHIT
260 version 4.6.4 (W. Li & Godzik, 2006) was used at a 80% identity cut-off to reduce
261 redundancy. These sequences were used for building a protein similarity network
262 using EFI-EST tool (Zallot, Oberg, & Gerlt, 2019) and Cytoscape visualization
263 (Shannon et al., 2003), and BlastKOALA with default parameters for functional
264 annotation (Minoru Kanehisa, Sato, & Morishima, 2016). These analyses led us to
265 focus on *psbO* as a gene marker for phytoplankton, for which we did a deeper
266 analysis by building its phylogeny in the following way. Protein sequences were
267 aligned with MAFFT version 6 using the G-INS-I strategy (Kato & Toh, 2008).
268 Phylogenetic trees were generated with PhyML version 3.0 using the LG substitution
269 model plus gamma-distributed rates and four substitution rate categories (Guindon et

23

270 al., 2010). The starting tree was a BIONJ tree and the type of tree improvement was
271 subtree pruning and regrafting. Branch support was calculated using the approximate
272 likelihood ratio test (aLRT) with a Shimodaira–Hasegawa-like (SH-like) procedure.

273

274 *Analysis of Tara Oceans datasets*

275 *Tara Oceans* expeditions between 2009 and 2013 performed a worldwide
276 sampling of plankton in the upper layers of the ocean (Sunagawa et al., 2020). To
277 capture the whole size spectrum of plankton, a combination of filter membranes with
278 different pore sizes (size-fractionation) was used to separate organisms by body size
279 (Pesant et al. 2015). There is an inverse logarithmic relationship between plankton
280 size and abundance (Belgrano, Allen, Enquist, & Gillooly, 2002; Pesant et al., 2015),
281 so small size fractions represent the numerically dominant organisms in terms of cell
282 abundance (albeit not necessarily in terms of total biovolume or biomass). Thus, the
283 protocols consisted in the filtering of higher seawater volumes for the larger size
284 fractions (Pesant et al., 2015). Five major organismal size fractions were collected:
285 picoplankton (0.2-3 μm size fraction), piconanoplankton (0.8-5 μm size fraction),
286 nanoplankton (5-20 μm size fraction), microplankton (20 to 180 μm), and
287 mesoplankton (180 to 2000 μm). These plankton samples were leveraged to
288 generate different molecular and optical datasets that were analysed in the current
289 work (Table I). We exclusively used the datasets corresponding to surface samples
290 (5 m depth).

291 *psbO*-based community data

292 To use the metagenomic and metatranscriptomic read abundances of *psbO* as
 293 a proxy of phytoplankton relative cell abundance and ‘activity’, respectively, we
 294 carried out a HMMer search as stated in the previous section against the two *Tara*
 295 Oceans gene catalogues: the Ocean Microbial Reference Gene Catalog version 2
 296 (OM-RGC.v2) covering prokaryotic and eukaryotic picoplankton (<3 µm), and the
 297 Marine Atlas of *Tara* Oceans Unigenes version 1 (MATOU.v1) covering eukaryotic
 298 plankton ranging from 0.8 to 2000 µm (Table I). The metagenomic and
 299 metatranscriptomic reads are already mapped onto both catalogs, thus we retrieved
 300 these values for those sequences obtained by our HMMer search. For the taxonomic
 301 assignment of *psbO* unigenes, we performed a phylogenetic placement of the
 302 translated sequences on the PsbO protein reference phylogenetic tree described in
 303 the previous section. For parallelization of the task, a set of 50 unigenes were
 304 translated and the PsbO specific Pfam PF01716 region was retrieved for the analysis
 305 in the following way. First, they were aligned against the reference alignment
 306 described in the previous section using the option --add of MAFFT version 6 with the
 307 G-INS-I strategy (Kato and Toh 2008). The resulting alignment was used for
 308 building a phylogeny with PhyML version 3.0 as described above (Guindon et al.,
 309 2010). The sequences were classified using the APE library in R (Paradis & Schliep,
 310 2019) according to their grouping in monophyletic branches of statistical support >0.7
 311 with reference sequences of the same taxonomic group.

312 Due to challenges of assembling eukaryotic genomes from complex
 313 metagenomes, the MATOU-v1 catalog only contains sequences assembled from
 314 poly-A-tailed RNA (Alberti et al., 2017; Carradec et al., 2018), which biases against

27

315 prokaryotic sequences. To determine the structure of the whole phytoplankton
316 community (including both cyanobacteria and eukaryotic phytoplankton), we aligned
317 all the metagenomic reads from *Tara* Oceans to a curated database of *psbO*
318 sequences (described below; see also Table I). The analysis was carried out using
319 the bwa tool version 0.7.4 (H. Li & Durbin, 2009) with the following parameters: -
320 minReadSize 70 -identity 80 -alignment 80 -complexityPercent 75 -
321 complexityNumber 30. Abundance values were expressed in rpkm (reads per
322 kilobase covered per million of mapped reads).

323 In general, the rpkm values for the different taxa under study were converted to
324 percentage of (either total or eukaryotic) phytoplankton. However, for a specific
325 analysis the *psbO* rpkm values were normalized by those values from single-copy
326 housekeeping genes: by bacterial *recA* (Sunagawa et al. 2013) to estimate the
327 contribution of cyanobacteria in the bacterioplankton, or by the average abundance
328 of 25 genes encoding ribosomal proteins (Ciccarelli et al. 2006; Carradec et al. 2018)
329 to estimate the contribution of phytoplankton among eukaryotes. The abundance
330 values for *recA* were retrieved from a previous work (Pierella Karlusich et al. 2021)
331 while the ribosomal proteins were recovered from the MATOU-v1 and OMRGC-v2
332 abundance tables.

333

334 *rRNA gene-based community data*

335 We used two different datasets generated by *Tara* Oceans for “traditional” DNA-
336 based methods: 16S rRNA gene miTags (size fraction 0.2-3 µm) and 18S rRNA gene
337 (V9 region) metabarcoding (sizes fractions 0.8-5, 5,20, 20-180, 180-2000 µm) (Table

338 I). We extracted the relative abundances for the 726 Operational Taxonomic Units
339 (OTUs) assigned to picophytoplankton (cyanobacteria and chloroplasts) from the 16S
340 miTags and the 31,930 OTUs assigned to eukaryotic phytoplankton from the V9-18S
341 metabarcoding data. The read abundances were expressed as relative abundance
342 (%) in relation to the picophytoplankton community for 16S miTags, and in relation to
343 eukaryotic phytoplankton for V9-18S metabarcoding.

344 The assignments of the 16S and V9-18S rRNA sequences to phytoplankton
345 were based on literature and expert information and included photosynthetic
346 dinoflagellates and chrysophytes when their taxonomic resolution was sufficient to
347 match known photosynthetic lineages. A full description of the 18S taxonomic
348 classification procedure is at <http://taraoceans.sb-roscoff.fr/EukDiv/> and the last
349 version of the trait reference database used in the current work is available at <https://zenodo.org/record/3768510#.YM4ny3UzbuE>. In the case of 16S miTags, the
350 taxonomic assignment was improved by building a phylogenetic tree with the 16S
351 miTags sequences and a curated set of references from NCBI and MMETSP.
352 Sequences were aligned using MAFFT v7.0 (Katoh & Standley, 2013) with --auto
353 setting option and then trimmed using trimal with the -gt 0.5 and -gt 0.8 settings, and
354 the resulted alignment was used for tree building using RAxML v8 (Stamatakis, 2014)
355 (100 bootstrap replicates, GTRCAT substitution model).

357

358 *Optical-based community data*

359 We also used quantitative optical data generated by *Tara Oceans* (Table I),
360 where cell abundance is assumed to be more accurate and less biased, and

31

361 additional features such as biovolume can be determined. The datasets cover: flow
362 cytometry for picoplankton, confocal microscopy for 5–20 μm size fraction, and light
363 microscopy for 20–180 μm size fraction.

364 Flow cytometry counts were determined on 1 ml replicated seawater samples
365 filtered through 200 μm that were fixed with cold 25% glutaraldehyde (final
366 concentration 0.125%) and stored at -80°C until analysis. Details about the
367 procedure can be found in (Gasol & Morán, 2015; Hingamp et al., 2013; Pierella
368 Karlusich et al., 2021). The cell biovolume was calculated using the equation of
369 (Calvo-Díaz & Morán, 2006) on the bead-standardized side scatter of the populations
370 and considering cells to be spherical.

371 Quantitative confocal microscopy was performed using environmental High
372 Content Fluorescence Microscopy (eHCFM) (Colin et al., 2017). Briefly, samples
373 were fixed with 10% monomeric formaldehyde (1 % final concentration) buffered at
374 pH 7.5 and 500 μl EM grade glutaraldehyde (0.25% final concentration) and kept at 4
375 $^{\circ}\text{C}$ until analysis. Sample collection and preparation as well imaging acquisition is
376 described in (Colin et al., 2017). The 5–20 μm size fraction has been classified at a
377 coarse taxonomic level (with an estimated accuracy of 93.8% at the phylum or class
378 level), into diatoms, dinoflagellates, haptophytes, and other/unclassified eukaryotic
379 phytoplankton (Colin et al., 2017). We used the major and minor axis of every image
380 to calculate their ellipsoidal equivalent biovolume. The 20–180 μm size fraction is also
381 available, but the curated taxonomic annotation is limited to symbiotic (*Richelia*,
382 *Calothrix*) and colony-forming (*Trichodesmium*) nitrogen-fixing cyanobacteria
383 (Pierella Karlusich et al. 2021), which were also used in the current work.

32

33

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

389 To be compared with the molecular data, the optical data were expressed as
390 relative abundance (%). In the case of flow cytometry, as % over the total number of
391 cells counted as picophytoplankton (*Prochlorococcus* + *Synechococcus* + eukaryotic
392 picophytoplankton). In the case of confocal and optical microscopy, as % over the
393 total number of cells counted as eukaryotic phytoplankton.

394

395 *psbO database generation*

396 We compiled, curated and annotated a database of >18,000 unique *psbO*
397 sequences covering cyanobacteria, photosynthetic protists, macroalgae and land
398 plants (Figure S1). It includes sequences retrieved from IMG, NCBI, MMETSP and
399 other sequenced genomes and transcriptomes from cultured isolates, as well as from
400 the environmental sequence catalogs from Global Ocean Sampling (Rusch et al.,
401 2007) and *Tara* Oceans (Carradec et al., 2018; Delmont et al., 2020, 2021; Salazar
402 et al., 2019). The database can be downloaded from the EMBL-EBI repository
403 BioStudies (www.ebi.ac.uk/biostudies) under accession S-BSST659. We expect to
404 maintain it updated to facilitate its incorporation in molecular-based surveys.

405

35

406 *Plotting and statistical analysis*

407 Graphical analyses were carried out in R language (<http://www.r-project.org/>)
 408 using *ggplot2* (Wickham, 2016) and treemaps were generated with *treemap*. Maps
 409 were generated with *borders* function in *ggplot2* and *geom_point* function for bubbles
 410 or *scatterpie* package for pie charts (Yu, 2018). Spearman's rho correlation
 411 coefficients and p-values were calculated using the *cor.test* function of the *stats*
 412 package. Shannon diversity indexes were calculated using the *vegan* package
 413 (Oksanen et al. 2020). Intra- and inter-specific genetic distances were calculated in
 414 MEGAX (Kumar, Stecher, Li, Knyaz, & Tamura, 2018) using the Maximum
 415 Composite Likelihood model.

416

417 **Results**

418

419 *Search for phytoplankton marker genes*

420 We first analysed transcriptomes and nuclear and plastid genomes derived from
 421 cultured strains to inventory photosynthetic genes in relation to their genome location
 422 (nuclear- vs plastid-encoded) and taxonomic prevalence (core vs non-core, i.e.,
 423 present in all phototrophs or not) (see Methods; Figure 1A). Among the plastid-
 424 encoded genes, we identified phytoplankton marker genes previously used in
 425 environmental surveys, such as *psbA* (Man-Aharonovich et al., 2010b; Zeidner,
 426 Preston, Delong, Massana, Post, Scanlan, & B  j  , 2003), *rbcL* (nuclear encoded in
 427 dinoflagellates containing peridinin) (Paul et al., 2000) and *petB* (Farrant et al., 2016)
 428 (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 chlorophyll biosynthesis (*chlG*) (Table II). The absence of non-photosynthetic homologs is a unique characteristic of *psbO* (Table II and Figures S2-S6), reflecting its essential role in photosynthetic oxygen evolution, and a clear advantage for its use as a marker gene for phytoplankton. Previous studies of secondarily non-photosynthetic eukaryotes have marked its presence or absence as being an effectively universal predictor of photosynthetic potential (Dorrell et al., 2019). Its phylogeny additionally reflects the evolutionary history of endosymbiosis (Figure 1B; Pierella Karlusich et al. 2015), with few or no post-endosymbiotic horizontal replacements known, so we focused on this gene for the analysis of environmental samples.

Although no global barcoding gap (i.e., a distance threshold set for all species) was detected when checking intra- vs inter-specific divergences for eukaryotic phytoplankton based on *psbO*, it was neither observed with the V9 region of the traditional marker 18S rRNA gene (Figure S7). This absence does not necessarily preclude specimen identification, which relies upon the presence of a 'local' barcoding gap (i.e., a query sequence being closer to a conspecific sequence than a different species), rather than the 'global' barcoding gap (i.e., a distance threshold set for all species) that is required for species discovery (Collins & Cruickshank, 2013).

We retrieved the *psbO* sequences from the two *Tara* Oceans gene catalogs (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 *Synechococcus* 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 and 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.

Marine phytoplankton community structure based on psbO shows 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 and plastids, 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 μ m), the picocyanobacteria *Prochlorococcus* and *Synechococcus* were prevalent, while ~60% of the average read abundance was 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

41

477 among eukaryotic phototrophs, especially in the 5–20 μm and 20–180 μm size
478 fractions (Figure 2B). In the 180–2000 μm fraction, diatoms and dinoflagellates were
479 still abundant, due to the presence of large diameter cells (*Tripos*, *Pyrocystis*), chain-
480 forming (e.g., *Chaetoceros*, *Fragilaria*) or epizoic (e.g., *Pseudohimantidium*) species,
481 without discarding that smaller species may be retained in samples of this size
482 fraction due to net clogging or within herbivorous guts and faecal pellets. Relative
483 abundance in the smaller 0.8–5 μm size fraction was much more homogeneously
484 distributed between the different groups.

485 For *psbO*-based methods, we found that metagenomic and metatranscriptomic
486 reads from *Synechococcus*, *Prochlorococcus*, pelagophytes, chlorophytes and
487 haptophytes to be dominant among picophytoplankton (0.2–3 μm), along with
488 dictyochophytes and chrysophytes (Figure 2A). In the larger size fractions,
489 haptophytes, chlorophytes and pelagophytes clearly dominated the eukaryotic
490 phytoplankton in the 0.8–5 μm size fraction, whereas diatoms and dinoflagellates
491 were more abundant in the three larger size ranges (5–20 μm , 20–180 μm , 180–2000
492 μm), although haptophytes, chlorophytes and pelagophytes were also detected in
493 large quantities (Figure 2B). The potential cyanobacteria present in these large size
494 fractions are presented later in another section due to the need to bypass the
495 sequences assembled from poly-A-tailed RNA for analysing prokaryotes (see
496 Methods and Table I).

497 We noted some differences in *psbO* read counts between metagenomic and
498 metatranscriptomic datasets. In the case of picophytoplankton, *Prochlorococcus* was
499 enriched in metagenomes in comparison to (total RNA) metatranscriptomes (likely
500 due to low transcript content needs for their low cell volume), the opposite occurred

42

for pelagophytes and haptophytes, whereas no major changes were observed for *Synechococcus* and chlorophytes (Figures 2A and S8A). In the case of larger photosynthetic protists, dinoflagellates were highly abundant at the (polyA) transcript level in comparison to gene abundance (probably they blanket overtranscribe genes as they predominantly perform post-transcriptional gene regulation (Roy, Jagus, and Morse 2018; Cohen et al. 2021)), 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 S8B).

The taxonomic abundance patterns based on *psbO* showed some differences with 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 S9). 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

45

525 size fractions (<7% average 18S read abundance). When we compared the
526 metatranscriptomic profile, it was more similar to the profile obtained with
527 metagenomes than to that obtained with V9-18S metabarcoding (Figure 2B).

528

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

531 To assess the accuracy of the *psbO* gene for determining phytoplankton cell
532 relative abundances, we carried out comparative analyses with imaging datasets. For
533 the 0.2-3 μm size fraction, we compared relative abundances based on 16S and
534 *psbO* counts with those inferred from flow cytometry (Figure 2C). Both genes were
535 found to correlate well with flow cytometry. Although the correlations for eukaryotic
536 picophytoplankton were strong (Spearman's $\rho=0.64-0.71$, $p\text{-value}\ll<0.001$), the
537 relationships were not linear and picoeukaryotes appeared at much higher relative
538 abundances in metagenomes than in flow cytometry. This is consistent with the fact
539 that flow cytometry can count cells of up to 10-20 μm diameter and was performed on
540 seawater aliquots pre-filtered through a 200- μm mesh (see Methods), whereas DNA
541 isolation of picoplankton was carried out on seawater volumes mainly filtered through
542 3- μm pore sizes. When we discarded eukaryotes to focus only on the ratio
543 *Synechococcus* / (*Synechococcus* + *Prochlorococcus*) (Figure S10), flow cytometry
544 data shows a linear relationship with *psbO* metagenomic reads, while 16S miTags
545 reads underestimated *Synechococcus* and the opposite occurred for *psbO*
546 metatranscriptomic reads. In addition, the highest correlation with flow cytometry data
547 occurred with the *psbO* metagenomic counts (Spearman's $\rho=0.92$, 0.90 and 0.75,

46

548 $p < < 0.001$, for *psbO* metagenomic reads, *psbO* metatranscriptomic reads and 16S
549 miTags, respectively).

550 For the 5–20 μm size fraction, the relative abundance of eukaryotic
551 photosynthetic organisms was determined by cell counts using high-throughput
552 confocal microscopy. We compared these results with the proportions based on V9-
553 18S metabarcoding and *psbO* metagenomic reads (Figure 2D). The metabarcoding
554 data for dinoflagellates and diatoms were in good agreement with the microscopy but
555 it clearly underestimated the relative abundance of haptophytes and other eukaryotic
556 phytoplankton. Regarding *psbO*, the metagenomic relative abundances were in
557 stronger agreement with the microscopy counts for the four defined phytoplankton
558 groups (Figure 2D). Therefore, in the 5–20 μm size fraction, diatoms and
559 dinoflagellates displayed robust patterns of relative abundance using either V9-18S
560 metabarcoding or *psbO* metagenomic counts, while haptophytes and the other
561 groups were better described by *psbO*.

562 In the 20-180 μm size fraction, the relative abundance of eukaryotic
563 phytoplankton was determined by light microscopy. Again, the metabarcoding data
564 for dinoflagellates and diatoms were in good agreement with the microscopy data but
565 clearly underestimated the relative abundance of haptophytes and other eukaryotic
566 phytoplankton groups (Figure 2D). The metagenomic read relative abundances of
567 *psbO* were in stronger agreement with the microscopy counts for the four defined
568 phytoplankton groups, although the correlation with haptophytes was weaker (Figure
569 2D). Therefore, in the 20-180 μm size fraction, diatoms and dinoflagellates displayed
570 robust patterns of relative abundance using either V9-18S metabarcoding or *psbO*

49

571 metagenomic counts, while haptophytes were weakly described by both methods
572 and the other groups were much better described by *psbO*.

573 We also compared the relative abundances based on optical methods against
574 those based on *psbO* metatranscriptomic reads, and in general we observed good
575 agreement (Figure S11). Some phytoplankton groups displayed stronger correlations
576 against optical methods using metatranscriptomic *psbO* counts than 16S miTAGs
577 (e.g., *Synechococcus*) or V9-18S metabarcoding (e.g., other eukaryotic
578 phytoplankton in the 5-20 μ m) (Figures 2D and S11). However, the consistency in
579 relative abundance of *psbO* reads with optical methods was always better for
580 metagenomes than for metatranscriptomes (Figures 2D and S11).

581

582 *Comparison with optical-based biovolume suggests that neither psbO nor rRNA*
583 *genes are good proxies for estimating relative proportion of biovolume*

584 We also compared the relative read abundances of the different marker genes
585 against the proportional biovolumes for each taxon (Figure 3). Although the copy
586 number of rRNA marker genes was previously proposed as a proxy of cell
587 biovolume, the correlation of biovolume against rRNA gene relative abundances was
588 not stronger than those against *psbO* (Figures 3 and S12). The relative read
589 abundances for *Prochlorococcus* and eukaryotic picophytoplankton based either on
590 16S rRNA gene or *psbO* were higher than their proportional biovolumes in the same
591 samples, while the opposite was the case for *Synechococcus*. In the 5-20 μ m size
592 fraction, the biovolume proportion for haptophytes was clearly described by their
593 *psbO* relative abundance, while their V9-18S rRNA gene reads were very low in

50

51

594 relation to their biovolume. Both V9-18S rRNA gene and *psbO* reads were correlated
595 with the relative biovolume for diatoms and dinoflagellates, but for V9-18S rRNA
596 gene the data points were somewhat scattered and for *psbO* the relative abundances
597 for the reads were higher in relation to their biovolume. As the biovolume of other
598 taxa was very low, their proportion of *psbO* reads was much higher than the
599 corresponding biovolume fraction, whereas there was no correlation between V9-18S
600 and biovolume.

601

602

603 *Diversity analysis: Shannon-index is robust to the biases introduced by the*
604 *traditional molecular methods*

605 We further analysed whether our method improved the widely used Shannon
606 index, a diversity index that relates monotonically to species richness but differs in
607 that it downweights rare species, whose numbers are highly sensitive to
608 undersampling and molecular artefacts (Calderón-Sanou, Münkemüller, Boyer,
609 Zinger, & Thuiller, 2020). We found a strong correlation between Shannon values for
610 eukaryotic phytoplankton defined either by 18S rRNA gene metabarcoding or by
611 *psbO* metagenomics or metatranscriptomics (Figure 4). This is in agreement with
612 previous reports showing no major effects of 16S rRNA gene copy number variation
613 on the Shannon index of bacterial communities (Ibarbalz et al., 2019; Milanese et al.,
614 2019). These results illustrate that not all subsequent analyses are affected by the
615 biases introduced by traditional molecular methods.

616

617

52

Combining housekeeping and photosynthetic marker genes improves 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 5A), especially for those OTUs whose taxonomic affiliation is “unknown dinoflagellate” (Figure S13). The uncertainty was especially significant in the 0.8-5 μm size fraction, where on average the ~80% of the total dinoflagellate read abundance remained unclassified (Figures 5A and S13).

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 (Sunagawa et al., 2013) and genes encoding ribosomal proteins for eukaryotes (Carradec et al., 2018; Ciccarelli et al., 2006)), 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 homologs (i.e., the nuclear-encoded genes for the Rieske subunits of the Cyt *bc*-type complexes from chloroplasts and mitochondria) (Table II and Figure S5). 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 S14), reflecting the relative paucity of described secondarily non-photosynthetic members of these groups. 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

55

642 ranges, which were also shown by the V9-18S rRNA gene metabarcoding method
643 (Figures 5B, S13 and S14). However, whereas the metabarcoding data showed a
644 dramatic increase in phototrophs towards the larger size classes of dinoflagellates,
645 the metagenomic analysis showed similar levels between the three larger size
646 fractions (5-20 μm , 20-180 μm , 180-2000 μm) (Figure 5B). These different patterns
647 between the two marker genes might be explained by differences in the unknown
648 trait assignment of the 18S rRNA gene barcodes and/or in the 18S rRNA gene copy
649 number (e.g., higher copy number in photosynthetic species in larger size fractions).

650 The approach suggested can be applied to unveil variation of phototrophs in
651 whole plankton communities, including both bacteria and eukaryotes. In order to do
652 so, we mapped the metagenomic reads against our comprehensive catalog of *psbO*
653 sequences (Figure S1). The highest proportion of phytoplankton among eukaryotes
654 was observed in the 0.8-5 μm size fraction, followed by the 5-20 μm size-fraction,
655 while the lowest value was found in the 180-2000 μm size range (Figure 5C), where
656 copepods are prevalent (considered one of the most abundant animals on the
657 planet). Surprisingly, the percentage of phototrophs among bacterioplankton did not
658 vary across size fractions (10-15 % on average; see next section). In the 0.2-3 μm
659 size fraction, very similar values were detected by 16S miTags, but when comparing
660 both molecular methods with flow cytometry, the *psbO/recA* ratio was better
661 correlated to flow cytometry (Spearman's rho of 0.82 vs 0.91, $p < 0.001$, and a closer
662 1:1 relationship) (Figure S15).

663

664

665 *Trans-domain comparison reveals unexpected abundance of picocyanobacteria*
666 *in large size fractions*

667 To further examine the distribution of both prokaryotic and eukaryotic
668 phytoplankton across the whole size spectrum, we continued the analysis of the
669 mapped metagenomic reads against our catalogue of *psbO* sequences (Figure S1).
670 We observed a high abundance of cyanobacteria in the large size fractions in relation
671 to the eukaryotic phytoplankton (Figure 6A). The nitrogen-fixers *Trichodesmium* and
672 *Richelia/Calothrix* were found principally in the 20-180 and 180-2000 μm size
673 fractions (Figure 6A), which is expected as the former forms filaments and colonies
674 while the second group are symbionts of certain diatoms (Figures 7B-D). These
675 genera were recently quantified in the high-throughput confocal microscopy dataset
676 from the 20-180 μm size fraction (Pierella Karlusich et al., 2021). We therefore
677 checked the correlations of these data with the *psbO* determinations and found them
678 to be very strongly related (Figure 6G).

679 To our surprise, we also detected a high abundance of both *Prochlorococcus*
680 and, in particular, *Synechococcus*, in the large size fractions (Figure 6A) across
681 multiple and geographically distinct basins of the tropical and subtropical regions of
682 the world's ocean (Figure 7). Picocyanobacteria have small cell diameters ($<1 \mu\text{m}$),
683 and therefore should readily pass through the filters with pore sizes of 5, 20 or 180
684 μm . Although smaller cells can get caught on larger filters, their abundance should be
685 limited and hence not responsible for the values observed. The reason why a
686 substantial fraction of picocyanobacteria were found in the largest size fractions may
687 be colony formation, symbiosis, attachment to particles, or their grazing by protists,
688 copepods and/or suspension feeders. We examined these possibilities by looking at

689 the *Tara* Oceans confocal microscopy dataset, and found many microscopy images
 690 evidencing colony formation and symbiosis in the 20-180 μm size fraction (Figure 6E-
 691 F). This is in agreement with the mapping of the *Tara* Oceans metagenomes against
 692 a recently sequenced single cell genome of a *Synechococcus* living as a
 693 dinoflagellate symbiont (Nakayama et al., 2019). In addition, there are reports of
 694 picocyanobacterial symbionts among isolates of planktonic foraminifers, radiolarians,
 695 tintinnids, and dinoflagellates (Bird et al., 2017; Foster, Collier, & Carpenter, 2006;
 696 Kim, Choi, & Park, 2021; Yuasa, Horiguchi, Mayama, Matsuoka, & Takahashi, 2012)
 697 and picocyanobacterial colonies were observed in a regional study based on optical
 698 methods (Masquelier & Vaulot, 2008) and in lab cultures (W. Deng, Cruz, & Neuer,
 699 2016; Wei Deng, Monks, & Neuer, 2015).

700 These results suggest that we should move from the traditional view of
 701 *Synechococcus/Prochlorococcus* as being exclusively part of picoplankton
 702 communities, and instead we should consider them as part of a broader range of the
 703 plankton size spectrum (in a similar way as occurs with other small-celled
 704 phytoplankton such as the haptophyte *Phaeocystis*; (Beardall et al., 2009; Decelle et
 705 al., 2019)). However, it should be borne in mind that these results correspond to
 706 estimates of relative cell abundance, and thus the picture is very different when
 707 translated to biovolume, due to the large differences in cell size (Figure S16). All in
 708 all, our approach allows us to make trans-domain comparisons, which can reveal
 709 photosymbiosis and cell aggregates (Figure 6), and allows us to examine the
 710 biogeography of the entire phytoplankton community simultaneously (Figures 7, S16
 711 and S17).

Discussion

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, *psbO* emerged as the most suitable due to its lack of non-photosynthetic homologs (but note that the other genes could be incorporated in future studies by discarding non-photosynthetic homologs by phylogenetic and/or sequence similarity methods). We applied this new approach by retrieving *psbO* sequences from the metagenomes generated by *Tara* Oceans, and successfully validated it using the optical determinations from the same expedition.

We also quantified the biases of “traditional” molecular approaches as compared to the optical methods. The 16S miTags approach avoids PCR biases and 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 was analysed in the current work, where most picoplankton cells only 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 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. It is important to take into account that not all analyses are affected by the biases introduced by traditional molecular methods, as we showed for the Shannon index.

736 While our work demonstrated that *psbO* reflects the relative cell abundance of
737 phytoplankton, some previous studies suggested that rRNA genes reflect the relative
738 biovolume of the corresponding taxa. However, there is still no clear consensus for
739 rRNA genes as proxies of biovolume. Here, we did not observe major differences
740 between rRNA gene or *psbO* when correlated against optical-based biovolumes.

741 In addition to correcting for the abovementioned biases, we revealed
742 unexpected ecological features missed by 18S metabarcoding. For example, our
743 trans-domain comparison detected picocyanobacteria in high numbers in large size
744 fractions, which was supported by the observation of numerous images of
745 picocyanobacterial aggregates and endosymbionts in the *Tara* Oceans imaging
746 dataset. Moreover, when the analysis of metagenomes includes housekeeping
747 marker genes (in addition to photosynthetic genes), we observed small
748 dinoflagellates (0.8-5 μ m) to be mainly heterotrophic, while those in the larger size
749 communities (>5 μ m) to be mainly photosynthetic.

750 In addition to metagenomes, we also analysed *psbO* in metatranscriptomes,
751 where dinoflagellates stood out from the rest due to their much higher *psbO*
752 abundance ratio of mRNA abundance to gene copy number. It will be of interest to
753 analyse if this reflects higher 'photosynthetic activity' or if it is an effect of their
754 predominant post-transcriptional regulation (Cohen et al., 2021; Roy, Jagus, &
755 Morse, 2018).)

756 The very deep sequencing of the *Tara* Oceans metagenomes (between $\sim 10^8$
757 and $\sim 10^9$ total reads per sample) allowed us to carry out taxonomic analysis based
758 on a unique gene, in spite of dilution of the signal. As reduced DNA sequencing costs
759 are leading to the replacement of amplicon-based methods by metagenome

65

760 sequencing, we expect the utility of our method to increase in future years. In the
761 short term, a barcode approach using *psbO* primers is a promising cheap alternative,
762 although it will be subject to PCR biases and affected by the presence of introns.

763 The use of functional genes as taxonomic markers for phytoplankton has been
764 restricted to some surveys (using plastid-encoded genes) (Farrant et al., 2016; Man-
765 Aharonovich et al., 2010a; Paul et al., 2000; Zeidner, Preston, Delong, Massana,
766 Post, Scanlan, & Beja, 2003). This is not the case for other functional groups, such
767 as nitrogen-fixers, which are studied by targeting a gene encoding a subunit of the
768 nitrogenase enzymatic complex (Zehr & Paerl, 1998) and for which extensive
769 reference sequence databases are now available (<https://www.jzehrlab.com>; (Heller,
770 Tripp, Turk-Kubo, & Zehr, 2014)). To facilitate the incorporation of *psbO* into future
771 molecular-based surveys, we have generated a database of >18,000 annotated
772 *psbO* sequences (<https://www.ebi.ac.uk/biostudies/studies/S-BSST659>; Figure S1).
773 We hope that the release of this data, and the establishment of *psbO* as a new
774 biomarker for quantifying species abundances, opens new perspectives for
775 molecular-based evaluations of phytoplankton communities.

776 Based on the current analyses, we recommend the use of *psbO* as a proxy of
777 relative cell abundance of the whole phytoplankton community. However, when
778 focusing on either eukaryotes or prokaryotes, Shannon index is robust enough to be
779 based on rRNA genes. Finally, the use of molecular markers (either *psbO* or rRNA
780 genes) as proxies of relative phytoplankton biovolume is not established.

66

781 **Acknowledgments**

782 We would like to thank all colleagues from the *Tara* Oceans consortium as well as
 783 the Tara Ocean Foundation for their inspirational vision. We also acknowledge
 784 Quentin Carradec for his help with genes encoding ribosomal proteins. This project
 785 has received funding from the European Research Council (ERC) under the
 786 European Union's Horizon 2020 research and innovation programme (Diatomic;
 787 grant agreement No. 835067). Additional funding is acknowledged from the FFEM -
 788 French Facility for Global Environment (Fonds Français pour l'Environnement
 789 Mondial), and the French Government "Investissements d'Avenir" Programmes
 790 MEMO LIFE (Grant ANR-10-LABX-54), Université de Recherche Paris Sciences et
 791 Lettres (PSL) (Grant ANR-1253 11-IDEX-0001-02), France Genomique (ANR-10-
 792 INBS-09), and OCEANOMICS (Grant ANR-11-BTBR-0008). JJPK acknowledges
 793 postdoctoral funding from the Fonds Français pour l'Environnement Mondial. RGD
 794 acknowledges a CNRS Momentum Fellowship, awarded 2019-2021. This article is
 795 contribution number *** of *Tara* Oceans.

796 **Author contributions**

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

804 Data availability statement

805 All datasets analyzed for this study are of public access as described in Table I. The
806 curated *psbO* database was submitted to the EMBL-EBI repository BioStudies
807 (www.ebi.ac.uk/biostudies) under accession S-BSST659.

808

809 References

- 810 Acinas, S. G., Marcelino, L. A., Klepac-Ceraj, V., & Polz, M. F. (2004). Divergence and
811 Redundancy of 16S rRNA Sequences in Genomes with Multiple *rrn* Operons. *Journal of*
812 *Bacteriology*, Vol. 186, pp. 2629–2635. doi: 10.1128/jb.186.9.2629-2635.2004
- 813 Adriaenssens, E. M., & Cowan, D. A. (2014). Using signature genes as tools to assess
814 environmental viral ecology and diversity. *Applied and Environmental Microbiology*,
815 80(15), 4470–4480.
- 816 Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., ... Wincker, P.
817 (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans
818 expedition. *Scientific Data*, 4, 170093.
- 819 Armbrust, E. V., & Virginia Armbrust, E. (1998). Uniparental inheritance of chloroplast
820 genomes. *The Molecular Biology of Chloroplasts and Mitochondria in Chlamydomonas*,
821 pp. 93–113. doi: 10.1007/0-306-48204-5_6
- 822 Bachy, C., Dolan, J. R., López-García, P., Deschamps, P., & Moreira, D. (2013). Accuracy of
823 protist diversity assessments: morphology compared with cloning and direct
824 pyrosequencing of 18S rRNA genes and ITS regions using the conspicuous tintinnid
825 ciliates as a case study. *The ISME Journal*, 7(2), 244–255.
- 826 Beardall, J., Allen, D., Bragg, J., Finkel, Z. V., Flynn, K. J., Quigg, A., ... Raven, J. A. (2009).
827 Allometry and stoichiometry of unicellular, colonial and multicellular phytoplankton. *The*
828 *New Phytologist*, 181(2), 295–309.

- 829 Belgrano, A., Allen, A. P., Enquist, B. J., & Gillooly, J. F. (2002). Allometric scaling of
830 maximum population density: a common rule for marine phytoplankton and terrestrial
831 plants. *Ecology Letters*, Vol. 5, pp. 611–613. doi: 10.1046/j.1461-0248.2002.00364.x
- 832 Bird, C., Darling, K. F., Russell, A. D., Davis, C. V., Fehrenbacher, J., Free, A., ... Ngwenya,
833 B. T. (2017). Cyanobacterial endobionts within a major marine planktonic calcifier
834 (*Globigerina bulloides*, Foraminifera) revealed by 16S rRNA metabarcoding.
835 *Biogeosciences* , 14(4), 901–920.
- 836 Blazewicz, S. J., Barnard, R. L., Daly, R. A., & Firestone, M. K. (2013). Evaluating rRNA as
837 an indicator of microbial activity in environmental communities: limitations and uses. *The*
838 *ISME Journal*, 7(11), 2061–2068.
- 839 Bradley, I. M., Pinto, A. J., & Guest, J. S. (2016). Design and Evaluation of Illumina MiSeq-
840 Compatible, 18S rRNA Gene-Specific Primers for Improved Characterization of Mixed
841 Phototrophic Communities. *Applied and Environmental Microbiology*, Vol. 82, pp. 5878–
842 5891. doi: 10.1128/aem.01630-16
- 843 Calderón-Sanou, I., Münkemüller, T., Boyer, F., Zinger, L., & Thuiller, W. (2020). From
844 environmental DNA sequences to ecological conclusions: How strong is the influence of
845 methodological choices? *Journal of Biogeography*, 47(1), 193–206.
- 846 Calvo-Díaz, A., & Morán, X. A. G. (2006). Seasonal dynamics of picoplankton in shelf waters
847 of the southern Bay of Biscay. *Aquatic Microbial Ecology: International Journal*, 42, 159–
848 174.
- 849 Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden,
850 T. L. (2009). BLAST : architecture and applications. *BMC Bioinformatics*, Vol. 10, p. 421.
851 doi: 10.1186/1471-2105-10-421
- 852 Campbell, B. J., Yu, L., Heidelberg, J. F., & Kirchman, D. L. (2011). Activity of abundant and
853 rare bacteria in a coastal ocean. *Proceedings of the National Academy of Sciences*, Vol.
854 108, pp. 12776–12781. doi: 10.1073/pnas.1101405108
- 855 Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., ...

Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, 9(1), 373.

Chen, I.-M. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., ... Kyrpides, N. C. (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–D677.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., & Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311(5765), 1283–1287.

Cohen, N. R., McIlvin, M. R., Moran, D. M., Held, N. A., Saunders, J. K., Hawco, N. J., ... Saito, M. A. (2021). Dinoflagellates alter their carbon and nutrient metabolic strategies across environmental gradients in the central Pacific Ocean. *Nature Microbiology*, 6(2), 173–186.

Coleman, A. W., & Nerozzi, A. M. (1999). Temporal and Spatial Coordination of Cells with Their Plastid Component. *International Review of Cytology*, pp. 125–164. doi: 10.1016/s0074-7696(08)61780-5

Colin, S., Coelho, L. P., Sunagawa, S., Bowler, C., Karsenti, E., Bork, P., ... de Vargas, C. (2017). Quantitative 3D-imaging for cell biology and ecology of environmental microbial eukaryotes. *eLife*, 6. doi: 10.7554/eLife.26066

Collins, R. A., & Cruickshank, R. H. (2013). The seven deadly sins of DNA barcoding. *Molecular Ecology Resources*, 13(6), 969–975.

Decelle, J., Romac, S., Stern, R. F., Bendif, E. M., Zingone, A., Audic, S., ... Christen, R. (2015). PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Molecular Ecology Resources*, Vol. 15, pp. 1435–1445. doi: 10.1111/1755-0998.12401

Decelle, J., Stryhanyuk, H., Gallet, B., Veronesi, G., Schmidt, M., Balzano, S., ... Musat, N. (2019). Algal Remodeling in a Ubiquitous Planktonic Photosymbiosis. *Current Biology*, Vol. 29, pp. 968–978.e4. doi: 10.1016/j.cub.2019.01.073

- 883 Delmont, T. O., Gaia, M., Hinsinger, D. D., Fremont, P., Vanni, C., Fernandez Guerra, A., ...
884 Jaillon, O. (2020). Functional repertoire convergence of distantly related eukaryotic
885 plankton lineages revealed by genome-resolved metagenomics. *bioRxiv*. doi.
886 10.1101/2020.10.15.341214
- 887 Delmont, T. O., Pierella Karlusich, J. J., Veseli, I., Fuessel, J., Murat Eren, A., Foster, R. A.,
888 ... Pelletier, E. (2021). Heterotrophic bacterial diazotrophs are more abundant than their
889 cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. *bioRxiv*.
890 doi: 10.1101/2021.03.24.436778
- 891 Deng, W., Cruz, B. N., & Neuer, S. (2016). Effects of nutrient limitation on cell growth, TEP
892 production and aggregate formation of marine *Synechococcus*. *Aquatic Microbial*
893 *Ecology: International Journal*, 78(1), 39–49.
- 894 Deng, W., Monks, L., & Neuer, S. (2015). Effects of clay minerals on the aggregation and
895 subsequent settling of marine *Synechococcus*. *Limnology and Oceanography*, 60(3),
896 805–816.
- 897 de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., ... Karsenti, E.
898 (2015). Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. *Science*,
899 348(6237), 1261605.
- 900 Dorrell, R. G., Azuma, T., Nomura, M., de Kerdrel, G. A., Paoli, L., Yang, S., ... Kamikawa,
901 R. (2019). Principles of plastid reductive evolution illuminated by nonphotosynthetic
902 chrysophytes. *Proceedings of the National Academy of Sciences*, Vol. 116, pp. 6914–
903 6923. doi: 10.1073/pnas.1819976116
- 904 Dorrell, R. G., & Smith, A. G. (2011). Do red and green make brown?: perspectives on
905 plastid acquisitions within chromalveolates. *Eukaryotic Cell*, 10(7), 856–868.
- 906 Egge, E., Bittner, L., Andersen, T., Audic, S., de Vargas, C., & Edvardsen, B. (2013). 454
907 pyrosequencing to describe microbial eukaryotic community composition, diversity and
908 relative abundance: a test for marine haptophytes. *PloS One*, 8(9), e74371.
- 909 Farrant, G. K., Doré, H., Cornejo-Castillo, F. M., Partensky, F., Ratin, M., Ostrowski, M., ...

910 Garczarek, L. (2016). Delineating ecologically significant taxonomic units from global
911 patterns of marine picocyanobacteria. *Proceedings of the National Academy of Sciences*
912 *of the United States of America*, 113(24), E3365–E3374.

913 Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary production
914 of the biosphere: integrating terrestrial and oceanic components. *Science*, 281(5374),
915 237–240.

916 Foster, R. A., Collier, J. L., & Carpenter, E. J. (2006). Reverse transcription pcr amplification
917 of cyanobacterial symbiont 16s rRNA sequences from single non-photosynthetic
918 eukaryotic marine planktonic host cells¹. *Journal of Phycology*, 42(1), 243–250.

919 Fuller, N. J., Campbell, C., Allen, D. J., Pitt, F. D., Zwirgmaier, K., Le Gall, F., ... Scanlan, D.
920 J. (2006). Analysis of photosynthetic picoeukaryote diversity at open ocean sites in the
921 Arabian Sea using a PCR biased towards marine algal plastids. *Aquatic Microbial*
922 *Ecology*, Vol. 43, pp. 79–93. doi: 10.3354/ame043079

923 Fuller, N. J., Tarran, G. A., Cummings, D. G., Woodward, E. M. S., Orcutt, K. M., Yallop, M.,
924 ... Scanlan, D. J. (2006). Molecular analysis of photosynthetic picoeukaryote community
925 structure along an Arabian Sea transect. *Limnology and Oceanography*, Vol. 51, pp.
926 2502–2514. doi: 10.4319/lo.2006.51.6.2502

927 Gasol, J. M., & Morán, X. A. G. (2015). Flow cytometric determination of microbial
928 abundances and its use to obtain indices of community structure and relative activity. In
929 *Springer Protocols Handbooks* (pp. 159–187). Berlin, Heidelberg: Springer Berlin
930 Heidelberg.

931 Godhe, A., Asplund, M. E., Harnstrom, K., Saravanan, V., Tyagi, A., & Karunasagar, I.
932 (2008). Quantification of Diatom and Dinoflagellate Biomasses in Coastal Marine
933 Seawater Samples by Real-Time PCR. *Applied and Environmental Microbiology*, Vol.
934 74, pp. 7174–7182. doi: 10.1128/aem.01298-08

935 Gong, W., & Marchetti, A. (2019). Estimation of 18S Gene Copy Number in Marine
936 Eukaryotic Plankton Using a Next-Generation Sequencing Approach. *Frontiers in Marine*

- 937 *Science*, Vol. 6. doi: 10.3389/fmars.2019.00219
- 938 Green, B. R. (2011). Chloroplast genomes of photosynthetic eukaryotes. *The Plant Journal*,
- 939 Vol. 66, pp. 34–44. doi: 10.1111/j.1365-313x.2011.04541.x
- 940 Grigoriev, I. V., Hayes, R. D., Calhoun, S., Kamel, B., Wang, A., Ahrendt, S., ... Kuo, A.
- 941 (2021). PhycoCosm, a comparative algal genomics resource. *Nucleic Acids Research*,
- 942 49(D1), D1004–D1011.
- 943 Guidi, L., Stemmann, L., Jackson, G. A., Ibanez, F., Claustre, H., Legendre, L., ... Gorskya,
- 944 G. (2009). Effects of phytoplankton community on production, size, and export of large
- 945 aggregates: A world-ocean analysis. *Limnology and Oceanography*, 54(6), 1951–1963.
- 946 Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., ... Christen, R. (2013).
- 947 The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote
- 948 small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*,
- 949 41(Database issue), D597–D604.
- 950 Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010).
- 951 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing
- 952 the performance of PhyML 3.0. *Systematic Biology*, 59(3), 307–321.
- 953 Heller, P., Tripp, H. J., Turk-Kubo, K., & Zehr, J. P. (2014). ARBitrator: a software pipeline for
- 954 on-demand retrieval of auto-curated nifH sequences from GenBank. *Bioinformatics* ,
- 955 30(20), 2883–2890.
- 956 Hingamp, P., Grimsley, N., Acinas, S. G., Clerissi, C., Subirana, L., Poulain, J., ... Ogata, H.
- 957 (2013). Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial
- 958 metagenomes. *The ISME Journal*, 7(9), 1678–1695.
- 959 Hiramatsu, T., Nakamura, S., Misumi, O., Kuroiwa, T., & Nakamura, S. (2006). Morphological
- 960 changes in mitochondrial and chloroplast nucleoids and mitochondria during the
- 961 *Chlamydomonas reinhardtii* (Chlorophyceae) cell cycle. *Journal of Phycology*, Vol. 42,
- 962 pp. 1048–1058. doi: 10.1111/j.1529-8817.2006.00259.x
- 963 Ibarbalz, F. M., Henry, N., Brandão, M. C., Martini, S., Busseni, G., Byrne, H., ... Zinger, L.

81

(2019). Global Trends in Marine Plankton Diversity across Kingdoms of Life. *Cell*, 179(5), 1084–1097.e21.

Jaffe, A. L., Castelle, C. J., Dupont, C. L., & Banfield, J. F. (2019). Lateral Gene Transfer Shapes the Distribution of RuBisCO among Candidate Phyla Radiation Bacteria and DPANN Archaea. *Molecular Biology and Evolution*, 36(3), 435–446.

Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, Vol. 28, pp. 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Sato, Y., & Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *Journal of Molecular Biology*, 428(4), 726–731.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780.

Katoh, K., & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics*, 9(4), 286–298.

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., ... Worden, A. Z. (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, S. W., Wu, M., Eisen, J. A., & Green, J. L. (2012). Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Computational Biology*, 8(10), e1002743.

Kim, M., Choi, D. H., & Park, M. G. (2021). Cyanobiont genetic diversity and host specificity of cyanobiont-bearing dinoflagellate *Ornithocercus* in temperate coastal waters. *Scientific Reports*, 11(1), 9458.

Kirkham, A. R., Jardillier, L. E., Tiganescu, A., Pearman, J., Zubkov, M. V., & Scanlan, D. J. (2011). Basin-scale distribution patterns of photosynthetic picoeukaryotes along an

82

991 Atlantic Meridional Transect. *Environmental Microbiology*, Vol. 13, pp. 975–990. doi:
992 10.1111/j.1462-2920.2010.02403.x

993 Kirkham, A. R., Lepère, C., Jardillier, L. E., Not, F., Bouman, H., Mead, A., & Scanlan, D. J.
994 (2013). A global perspective on marine photosynthetic picoeukaryote community
995 structure. *The ISME Journal*, Vol. 7, pp. 922–936. doi: 10.1038/ismej.2012.166

996 Kono, T., Mehrotra, S., Endo, C., Kizu, N., Matusda, M., Kimura, H., ... Ashida, H. (2017). A
997 RuBisCO-mediated carbon metabolic pathway in methanogenic archaea. *Nature*
998 *Communications*, Vol. 8. doi: 10.1038/ncomms14007

999 Koumandou, V. L., & Howe, C. J. (2007). The Copy Number of Chloroplast Gene Minicircles
1000 Changes Dramatically with Growth Phase in the Dinoflagellate *Amphidinium*
1001 *operculatum*. *Protist*, Vol. 158, pp. 89–103. doi: 10.1016/j.protis.2006.08.003

1002 Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular
1003 Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and*
1004 *Evolution*, 35(6), 1547–1549.

1005 Lamb, P. D., Hunter, E., Pinnegar, J. K., Creer, S., Davies, R. G., & Taylor, M. I. (2019). How
1006 quantitative is metabarcoding: A meta-analytical approach. *Molecular Ecology*, 28(2),
1007 420–430.

1008 Lavrinienko, A., Jernfors, T., Koskimäki, J. J., Pirttilä, A. M., & Watts, P. C. (2021). Does
1009 Intraspecific Variation in rDNA Copy Number Affect Analysis of Microbial Communities?
1010 *Trends in Microbiology*, 29(1), 19–27.

1011 Lebrun, E., Santini, J. M., Brugna, M., Ducluzeau, A.-L., Ouchane, S., Schoepp-Cothenet, B.,
1012 ... Nitschke, W. (2006). The Rieske protein: a case study on the pitfalls of multiple
1013 sequence alignments and phylogenetic reconstruction. *Molecular Biology and Evolution*,
1014 23(6), 1180–1191.

1015 Lepère, C., Vaultot, D., & Scanlan, D. J. (2009). Photosynthetic picoeukaryote community
1016 structure in the South East Pacific Ocean encompassing the most oligotrophic waters on
1017 Earth. *Environmental Microbiology*, Vol. 11, pp. 3105–3117. doi: 10.1111/j.1462-

- 1018 2920.2009.02015.x
- 1019 Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler
1020 transform. *Bioinformatics* , 25(14), 1754–1760.
- 1021 Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D., & Knight, R. (2007). Short
1022 pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids*
1023 *Research*, Vol. 35, pp. e120–e120. doi: 10.1093/nar/gkm541
- 1024 Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of
1025 protein or nucleotide sequences. *Bioinformatics*, Vol. 22, pp. 1658–1659. doi:
1026 10.1093/bioinformatics/btl158
- 1027 Logares, R., Audic, S., Santini, S., Pernice, M. C., de Vargas, C., & Massana, R. (2012).
1028 Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled
1029 with pyrosequencing. *The ISME Journal*, 6(10), 1823–1833.
- 1030 Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F. M., Ferrera, I., Sarmiento, H., ...
1031 Acinas, S. G. (2014). Metagenomic 16S rDNA Illumina tags are a powerful alternative to
1032 amplicon sequencing to explore diversity and structure of microbial communities.
1033 *Environmental Microbiology*, 16(9), 2659–2671.
- 1034 Louca, S., Doebeli, M., & Parfrey, L. W. (2018). Correcting for 16S rRNA gene copy numbers
1035 in microbiome surveys remains an unsolved problem. *Microbiome*, 6(1), 41.
- 1036 Mäki, A., Salmi, P., Mikkonen, A., Kremp, A., & Tirola, M. (2017). Sample preservation, DNA
1037 or RNA extraction and data analysis for high-throughput phytoplankton community
1038 Sequencing. *Frontiers in Microbiology*, 8, 1848.
- 1039 Malviya, S., Scalco, E., Audic, S., Vincent, F., Veluchamy, A., Poulain, J., ... & Bowler, C.
1040 (2016). Insights into global diatom distribution and diversity in the world's ocean.
1041 *Proceedings of the National Academy of Sciences*, 113(11), E1516-E1525.
- 1042 Man-Aharonovich, D., Philosof, A., Kirkup, B. C., Le Gall, F., Yogev, T., Berman-Frank, I., ...
1043 Béjà, O. (2010a). Diversity of active marine picoeukaryotes in the Eastern
1044 Mediterranean Sea unveiled using photosystem-II psbA transcripts. *The ISME Journal*,

- 1045 Vol. 4, pp. 1044–1052. doi: 10.1038/ismej.2010.25
- 1046 Man-Aharonovich, D., Philosof, A., Kirkup, B. C., Le Gall, F., Yogev, T., Berman-Frank, I., ...
- 1047 Béjà, O. (2010b). Diversity of active marine picoeukaryotes in the Eastern
- 1048 Mediterranean Sea unveiled using photosystem-II psbA transcripts. *The ISME Journal*,
- 1049 4(8), 1044–1052.
- 1050 Masquelier, S., & Vault, D. (2008). Distribution of micro-organisms along a transect in the
- 1051 South-East Pacific Ocean (BIOSCOPE cruise) using epifluorescence microscopy.
- 1052 *Biogeosciences* , 5(2), 311–321.
- 1053 McDonald, S. M., Sarno, D., Scanlan, D. J., & Zingone, A. (2007). Genetic diversity of
- 1054 eukaryotic ultraphytoplankton in the Gulf of Naples during an annual cycle. *Aquatic*
- 1055 *Microbial Ecology*, Vol. 50, pp. 75–89. doi: 10.3354/ame01148
- 1056 Medinger, R., Nolte, V., Pandey, R. V., Jost, S., Ottenwälder, B., Schlötterer, C., & Boenigk,
- 1057 J. (2010). Diversity in a hidden world: potential and limitation of next-generation
- 1058 sequencing for surveys of molecular diversity of eukaryotic microorganisms. *Molecular*
- 1059 *Ecology*, Vol. 19, pp. 32–40. doi: 10.1111/j.1365-294x.2009.04478.x
- 1060 Milanese, A., Mende, D. R., Paoli, L., Salazar, G., Ruscheweyh, H.-J., Cuenca, M., ...
- 1061 Sunagawa, S. (2019). Microbial abundance, activity and population genomic profiling
- 1062 with mOTUs2. *Nature Communications*, 10(1), 1014.
- 1063 Mohamed, M. E., Zaar, A., Ebenau-Jehle, C., & Fuchs, G. (2001). Reinvestigation of a new
- 1064 type of aerobic benzoate metabolism in the proteobacterium *Azoarcus evansii*. *Journal*
- 1065 *of Bacteriology*, 183(6), 1899–1908.
- 1066 Moore, R. B., Oborník, M., Janouskovec, J., Chrudimský, T., Vancová, M., Green, D. H., ...
- 1067 Carter, D. A. (2008). A photosynthetic alveolate closely related to apicomplexan
- 1068 parasites. *Nature*, 451(7181), 959–963.
- 1069 Nakayama, T., Kamikawa, R., Tanifuji, G., Kashiyama, Y., Ohkouchi, N., Archibald, J. M., &
- 1070 Inagaki, Y. (2014). Complete genome of a nonphotosynthetic cyanobacterium in a
- 1071 diatom reveals recent adaptations to an intracellular lifestyle. *Proceedings of the*

- 1072 *National Academy of Sciences of the United States of America*, 111(31), 11407–11412.
- 1073 Nakayama, T., Nomura, M., Takano, Y., Tanifuji, G., Shiba, K., Inaba, K., ... Kawata, M.
- 1074 (2019). Single-cell genomics unveiled a cryptic cyanobacterial lineage with a worldwide
- 1075 distribution hidden by a dinoflagellate host. *Proceedings of the National Academy of*
- 1076 *Sciences of the United States of America*, 116(32), 15973–15978.
- 1077 Obiol, A., Giner, C. R., Sánchez, P., Duarte, C. M., Acinas, S. G., & Massana, R. (2020). A
- 1078 metagenomic assessment of microbial eukaryotic diversity in the global ocean.
- 1079 *Molecular Ecology Resources*, 20(3). doi: 10.1111/1755-0998.13147
- 1080 Oksanen, J., Blanchet, G., Friendly, M., Kindt, R., Legendre, P., McGlinn, ... Wagner, H
- 1081 (2020). *vegan*: Community Ecology Package. R package version 2.5-7. [https://CRAN.R-](https://CRAN.R-project.org/package=vegan)
- 1082 [project.org/package=vegan](https://CRAN.R-project.org/package=vegan)
- 1083 Oldenburg, D. J., & Bendich, A. J. (2004). Changes in the structure of DNA molecules and
- 1084 the amount of DNA per plastid during chloroplast development in maize. *Journal of*
- 1085 *Molecular Biology*, 344(5), 1311–1330.
- 1086 Parada, A. E., Needham, D. M., & Fuhrman, J. A. (2016). Every base matters: assessing
- 1087 small subunit rRNA primers for marine microbiomes with mock communities, time series
- 1088 and global field samples. *Environmental Microbiology*, 18(5), 1403–1414.
- 1089 Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and
- 1090 evolutionary analyses in R. *Bioinformatics*, Vol. 35, pp. 526–528. doi:
- 1091 10.1093/bioinformatics/bty633
- 1092 Paul, J. H., Alfreider, A., & Wawrik, B. (2000). Micro- and macrodiversity in rbcL sequences
- 1093 in ambient phytoplankton populations from the southeastern Gulf of Mexico. *Marine*
- 1094 *Ecology Progress Series*, Vol. 198, pp. 9–18. doi: 10.3354/meps198009
- 1095 Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., ... de Vargas, C. (2012).
- 1096 CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant,
- 1097 and Fungal Kingdoms. *PLoS Biology*, Vol. 10, p. e1001419. doi:
- 1098 10.1371/journal.pbio.1001419

- 1099 Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J., & Esling, P. (2016).
1100 Protist metabarcoding and environmental biomonitoring: Time for change. *European*
1101 *Journal of Protistology*, 55(Pt A), 12–25.
- 1102 Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., ... Tara
1103 Oceans Consortium Coordinators. (2015). Open science resources for the discovery and
1104 analysis of Tara Oceans data. *Scientific Data*, 2, 150023.
- 1105 Pierella Karlusich, J. J., & Carrillo, N. (2017). Evolution of the acceptor side of
1106 photosystem I: ferredoxin, flavodoxin, and ferredoxin-NADP oxidoreductase.
1107 *Photosynthesis Research*, Vol. 134, pp. 235–250. doi: 10.1007/s11120-017-0338-2
- 1108 Pierella Karlusich, J. J., Ceccoli, R. D., Graña, M., Romero, H., & Carrillo, N. (2015).
1109 Environmental selection pressures related to iron utilization are involved in the loss of
1110 the flavodoxin gene from the plant genome. *Genome Biology and Evolution*, 7(3), 750–
1111 767.
- 1112 Pierella Karlusich, J. J., Ibarbalz, F. M., & Bowler, C. (2020). Phytoplankton in the Ocean.
1113 *Annual Review of Marine Science*, 12, 233–265.
- 1114 Pierella Karlusich, J. J., Pelletier, E., Lombard, F., Carsique, M., Dvorak, E., Colin, S., ...
1115 Foster, R. A. (2021). Global distribution patterns of marine nitrogen-fixers by imaging
1116 and molecular methods. *Nature Communications*, 12(1), 4160.
- 1117 Pinto, A. J., & Raskin, L. (2012). PCR biases distort bacterial and archaeal community
1118 structure in pyrosequencing datasets. *PloS One*, 7(8), e43093.
- 1119 Polz, M. F., & Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate
1120 PCR. *Applied and Environmental Microbiology*, 64(10), 3724–3730.
- 1121 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... Glöckner, F. O.
1122 (2013). The SILVA ribosomal RNA gene database project: improved data processing
1123 and web-based tools. *Nucleic Acids Research*, 41(Database issue), D590–D596.
- 1124 Roy, S., Jagus, R., & Morse, D. (2018). Translation and Translational Control in
1125 Dinoflagellates. *Microorganisms*, 6(2). doi: 10.3390/microorganisms6020030

- 1126 Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., ...
1127 Craig Venter, J. (2007). The Sorcerer II Global Ocean Sampling Expedition: Northwest
1128 Atlantic through Eastern Tropical Pacific. *PLoS Biology*, Vol. 5, p. e77. doi:
1129 10.1371/journal.pbio.0050077
- 1130 Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M., ...
1131 Wincker, P. (2019). Gene Expression Changes and Community Turnover Differentially
1132 Shape the Global Ocean Metatranscriptome. *Cell*, Vol. 179, pp. 1068–1083.e21. doi:
1133 10.1016/j.cell.2019.10.014
- 1134 Saldarriaga, J. F., Taylor, F. J., Keeling, P. J., & Cavalier-Smith, T. (2001). Dinoflagellate
1135 nuclear SSU rRNA phylogeny suggests multiple plastid losses and replacements.
1136 *Journal of Molecular Evolution*, 53(3), 204–213.
- 1137 Santoferrara, L. F. (2019). Current practice in plankton metabarcoding: optimization and error
1138 management. *Journal of Plankton Research*, 41(5), 571–582.
- 1139 Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., ... Ideker, T.
1140 (2003). Cytoscape: a software environment for integrated models of biomolecular
1141 interaction networks. *Genome Research*, 13(11), 2498–2504.
- 1142 Shi, X. L., Lepère, C., Scanlan, D. J., & Vault, D. (2011). Plastid 16S rRNA gene diversity
1143 among eukaryotic picophytoplankton sorted by flow cytometry from the South Pacific
1144 Ocean. *PloS One*, 6(4), e18979.
- 1145 Singer, A., Poschmann, G., Mühlich, C., Valadez-Cano, C., Hänsch, S., Hüren, V., ...
1146 Nowack, E. C. M. (2017). Massive Protein Import into the Early-Evolutionary-Stage
1147 Photosynthetic Organelle of the Amoeba Paulinella chromatophora. *Current Biology:*
1148 *CB*, 27(18), 2763–2773.e5.
- 1149 Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of
1150 large phylogenies. *Bioinformatics* , 30(9), 1312–1313.
- 1151 Starke, R., Pylro, V. S., & Morais, D. K. (2020). 16S rRNA Gene Copy Number Normalization
1152 Does Not Provide More Reliable Conclusions in Metataxonomic Surveys. *Microbial*

- 1153 *Ecology*. doi: 10.1007/s00248-020-01586-7
- 1154 Sunagawa, S., Coordinators, T. O., Acinas, S. G., Bork, P., Bowler, C., Eveillard, D., ... de
- 1155 Vargas, C. (2020). Tara Oceans: towards global ocean ecosystems biology. *Nature*
- 1156 *Reviews Microbiology*. doi: 10.1038/s41579-020-0364-5
- 1157 Sunagawa, S., Mende, D. R., Zeller, G., Izquierdo-Carrasco, F., Berger, S. A., Kultima, J. R.,
- 1158 ... Bork, P. (2013). Metagenomic species profiling using universal phylogenetic marker
- 1159 genes. *Nature Methods*, 10(12), 1196–1199.
- 1160 Tabita, F. R., Hanson, T. E., Satagopan, S., Witte, B. H., & Kreel, N. E. (2008). Phylogenetic
- 1161 and evolutionary relationships of RubisCO and the RubisCO-like proteins and the
- 1162 functional lessons provided by diverse molecular forms. *Philosophical Transactions of*
- 1163 *the Royal Society of London. Series B, Biological Sciences*, 363(1504), 2629–2640.
- 1164 Thompson, A. W., Foster, R. A., Krupke, A., Carter, B. J., Musat, N., Vaultot, D., ... Zehr, J.
- 1165 P. (2012). Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga.
- 1166 *Science*, 337(6101), 1546–1550.
- 1167 Ullah, H., Nagelkerken, I., Goldenberg, S. U., & Fordham, D. A. (2018). Climate change
- 1168 could drive marine food web collapse through altered trophic flows and cyanobacterial
- 1169 proliferation. *PLoS Biology*, 16(1), e2003446.
- 1170 van der Loos, L. M., & Nijland, R. (2021). Biases in bulk: DNA metabarcoding of marine
- 1171 communities and the methodology involved. *Molecular Ecology*, 30(13), 3270–3288.
- 1172 Veit, S., Takeda, K., Tsunoyama, Y., Baymann, F., Nevo, R., Reich, Z., ... Rexroth, S.
- 1173 (2016). Structural and functional characterisation of the cyanobacterial PetC3 Rieske
- 1174 protein family. *Biochimica et Biophysica Acta*, 1857(12), 1879–1891.
- 1175 Větrovský, T., & Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial
- 1176 genomes and its consequences for bacterial community analyses. *PloS One*, 8(2),
- 1177 e57923.
- 1178 Wang, J., Chu, S., Zhu, Y., Cheng, H., & Yu, D. (2015). Positive selection drives
- 1179 neofunctionalization of the UbiA prenyltransferase gene family. *Plant Molecular Biology*,

- 1180 87(4-5), 383–394.
- 1181 Wear, E. K., Wilbanks, E. G., Nelson, C. E., & Carlson, C. A. (2018). Primer selection
- 1182 impacts specific population abundances but not community dynamics in a monthly time-
- 1183 series 16S rRNA gene amplicon analysis of coastal marine bacterioplankton.
- 1184 *Environmental Microbiology*, Vol. 20, pp. 2709–2726. doi: 10.1111/1462-2920.14091
- 1185 Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer.
- 1186 Yu, G. (2018) *scatterpie*: Scatter pie plot. <https://CRAN.R-project.org/package=scatterpie>
- 1187 Yoon, H. S., Reyes-Prieto, A., Melkonian, M., & Bhattacharya, D. (2006). Minimal plastid
- 1188 genome evolution in the Paulinella endosymbiont. *Current Biology: CB*, 16(17), R670–
- 1189 R672.
- 1190 Yuasa, T., Horiguchi, T., Mayama, S., Matsuoka, A., & Takahashi, O. (2012). Ultrastructural
- 1191 and molecular characterization of cyanobacterial symbionts in Dictyocoryne profunda
- 1192 (polycystine radiolaria). *Symbiosis* , 57(1), 51–55.
- 1193 Zallot, R., Oberg, N., & Gerlt, J. A. (2019). The EFI web resource for genomic enzymology
- 1194 tools: leveraging protein, genome, and metagenome databases to discover novel
- 1195 enzymes and metabolic pathways. *Biochemistry*, 58(41), 4169–4182.
- 1196 Zehr, J. P., & Paerl, H. (1998). Nitrogen fixation in the marine environment: genetic potential
- 1197 and nitrogenase expression. *Molecular Approaches to the Study of the Ocean*, pp. 285–
- 1198 301. doi: 10.1007/978-94-011-4928-0_13
- 1199 Zeidner, G., Preston, C. M., Delong, E. F., Massana, R., Post, A. F., Scanlan, D. J., & Beja,
- 1200 O. (2003). Molecular diversity among marine picophytoplankton as revealed by psbA
- 1201 analyses. *Environmental Microbiology*, Vol. 5, pp. 212–216. doi: 10.1046/j.1462-
- 1202 2920.2003.00403.x
- 1203 Zeidner, G., Preston, C. M., Delong, E. F., Massana, R., Post, A. F., Scanlan, D. J., & Béjà,
- 1204 O. (2003). Molecular diversity among marine picophytoplankton as revealed by psbA
- 1205 analyses. *Environmental Microbiology*, 5(3), 212–216.
- 1206 Zhu, F., Massana, R., Not, F., Marie, D., & Vaulot, D. (2005). Mapping of picoeucaryotes in

99

1207 marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiology*
1208 *Ecology*, 52(1), 79–92.

1209

1210 **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 work
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 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)*	18S rRNA gene (V9 region) metabarcoding	PCR amplification of the V9 region (~130 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 <i>psbO</i> 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 <i>psbO</i> 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 <i>psbO</i> sequences	EBI accessions: PRJEB1787 PRJEB1788 PRJEB4352 PRJEB4419 PRJEB9691 PRJEB9740 PRJEB9742
Prokaryotes and eukaryotes	<200 μm	Flow cytometry		Abundances and biovolume of picocyanobacteria and	Hingamp et al. 2013; Gasol and Morán 2015 https://

				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
Prokaryotes	20-180 µm	Confocal microscopy		Abundance of the symbiotic cyanobacteria <i>Richelia/Calothrix</i> and the colony-forming <i>Trichodesmium</i>	Pierella Karlusich et al., 2021 https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-021-24299-y/MediaObjects/41467_2021_24299_MOESM11_ESM.xlsx

1211 *0.8-5 µm, 5-20 µm, 20-180 µm, 180-2000 µm

1212 **0.2-3 µm, 0.8-5 µm, 5-20 µm, 20-180 µm, 180-2000 µm

1213 **Table II: List of nuclear-encoded photosynthetic genes present in all cyanobacteria and eukaryotic phytoplankton.** These genes are always nuclear-
 1214 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
 1215 the origin of all other known plastids (Singer et al., 2017; Yoon, Reyes-Prieto, Melkonian, & Bhattacharya, 2006).

Gene	Pathway	Function	Copies	Non-photosynthetic homologs	References
<i>prk</i> (phosphoribulokinase)	Calvin-Benson-Bassham cycle	phosphorylation of ribulose-5-phosphate to ribulose-1,5-bisphosphate, the RuBisCO substrate	1	PRKs from archaea and bacteria	(Jaffe, Castelle, Dupont, & Banfield, 2019; Kono et al., 2017)
<i>chlG</i> (chlorophyll-a synthase)	chlorophyll-a biosynthesis	last step of chlorophyll-a biosynthesis	1	Prenyltransferases with UbiA domain	(Wang, Chu, Zhu, Cheng, & Yu, 2015)
<i>petH</i> (ferredoxin-NADP ⁺ oxidoreductase)	Photosynthetic electron transport chain	last step of the linear electron flow (NADP ⁺ reduction by 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	(Pierella Karlusich & Carrillo, 2017; Mohamed, Zaar, Ebenau-Jehle, & Fuchs, 2001)
<i>petC</i>	Photosynthetic electron transport chain	Rieske subunit of the chloroplast Cyt <i>b₆f</i> complex	2-3	Rieske proteins from mitochondria, bacteria and archaea	(Lebrun et al., 2006; Veit et al., 2016)
<i>psbO</i>	Photosynthetic electron transport chain	Manganese-stabilizing protein of photosystem II	1-2	No	(Pierella Karlusich, Ceccoli, Graña, Romero, & Carrillo, 2015)

1216

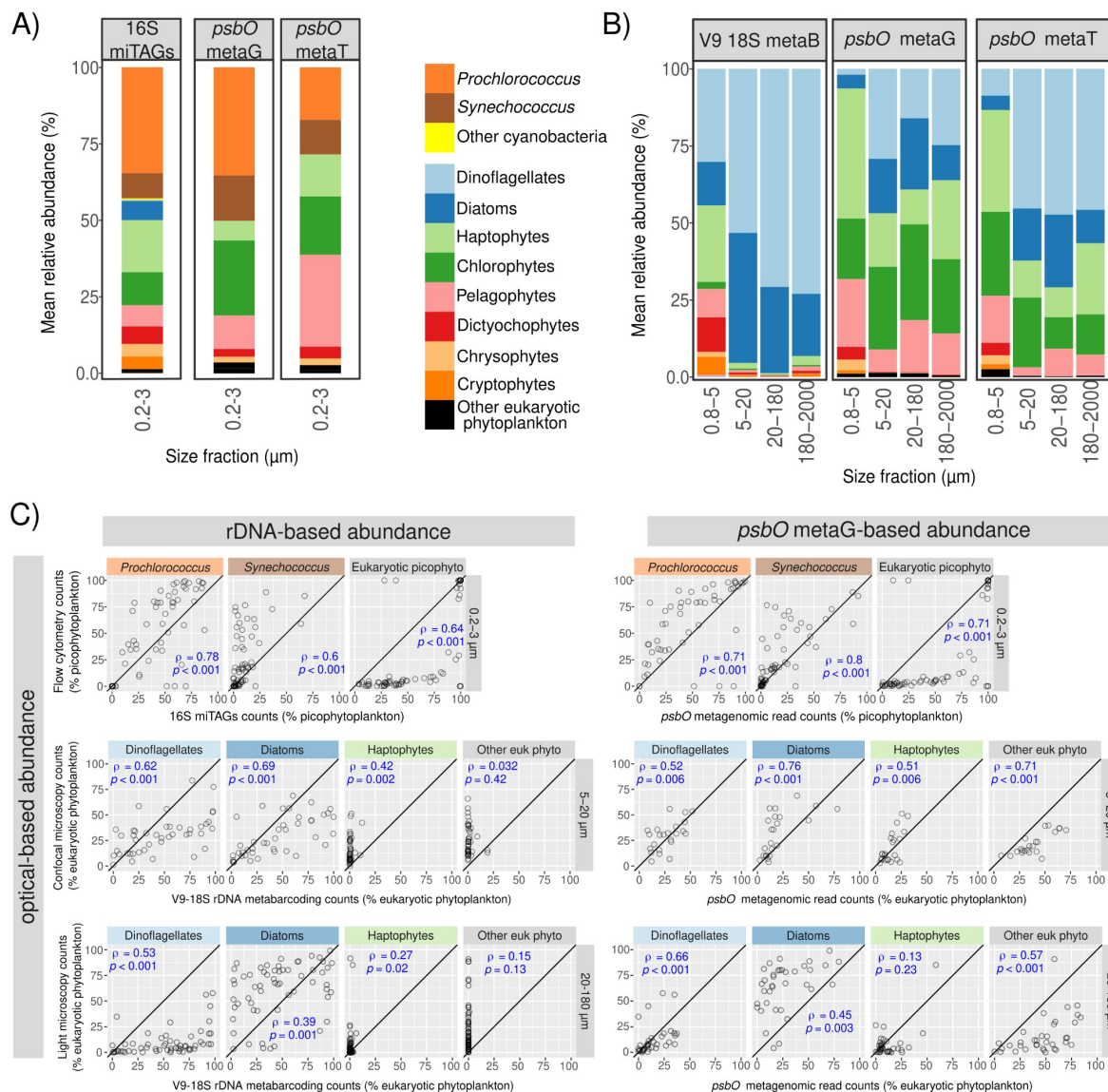


Figure 2: Congruence in relative abundances of the main phytoplankton groups based on different gene markers. A-B) Average relative abundances for all surface samples of each size fraction using different marker genes. In A, picocyanobacteria and eukaryotic picophytoplankton (0.2-3 μm) were analysed 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 S11.

111

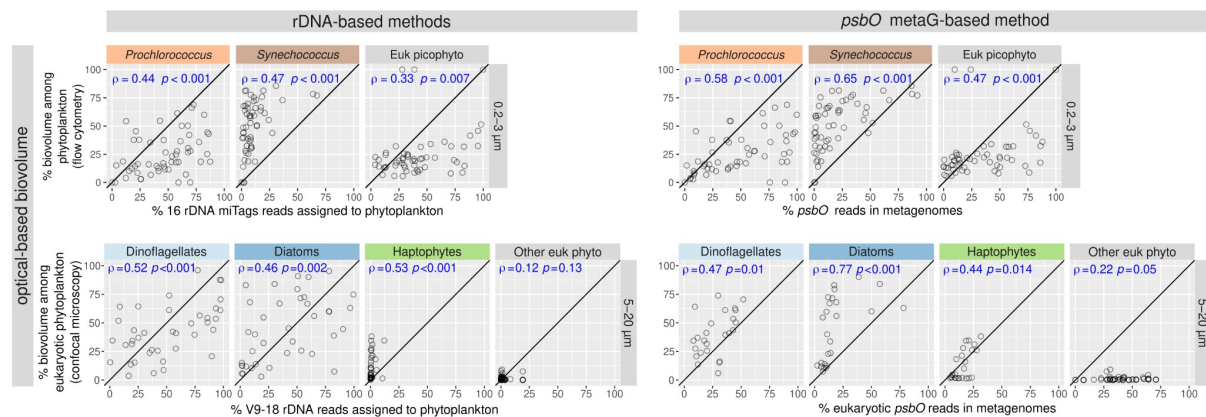


Figure 3: Correlation between relative biovolume (based on optical methods) and relative abundances based on different molecular methodologies. The upper panels show the correlations for picophytoplankton (size fraction 0.2-3 µm). The vertical axis corresponds to the relative biovolume based on flow 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 rRNA gene 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.

112

113

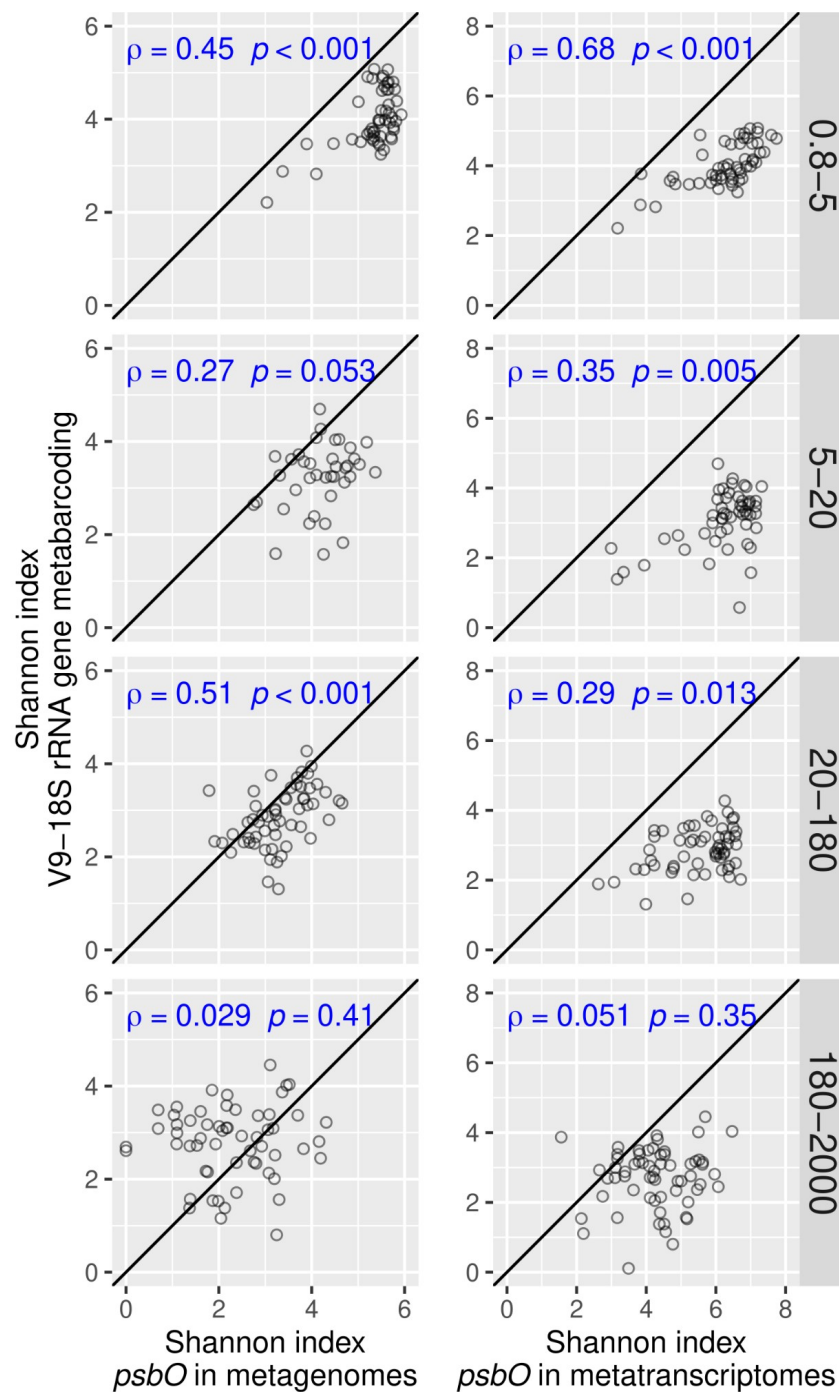


Figure 4: Correlation between the Shannon values derived from different molecular methods for eukaryotic phytoplankton communities. The values derived from *psbO* metagenomics (left) and metatranscriptomics (right) were compared with those derived from V9-18S rRNA gene metabarcoding. Spearman correlation coefficients and p-values are displayed in blue. Axis are in the same scale and the diagonal line corresponds to a 1:1 slope.

114

115

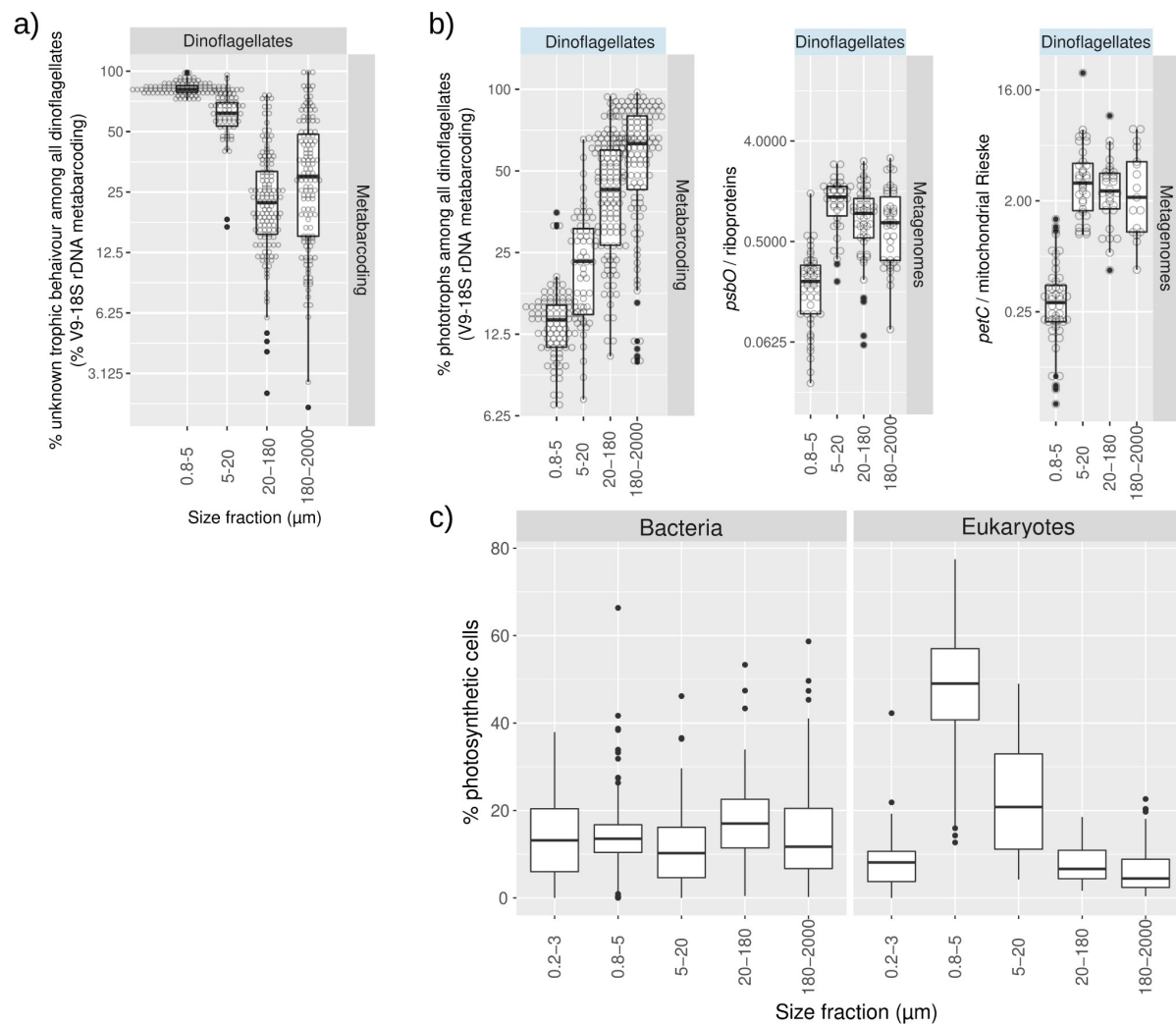


Figure 5: Variations in the abundance of phototrophs vs heterotrophs across size fractions. A) Read abundance of V9-18S rRNA gene metabarcoding assigned to dinoflagellates of unknown capacity for photosynthesis. B) Relative abundance of phototrophs among dinoflagellates based on different molecular methods. The first panel corresponds to the trait classification of V9-18S rRNA gene metabarcodes based on the literature (a description of the trait classification can be found at <http://taraoceans.sb-roscoff.fr/EukDiv/> and the trait reference database is available at <https://zenodo.org/record/3768951#.YM4odnUzbuE>). 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 homolog). 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).

116

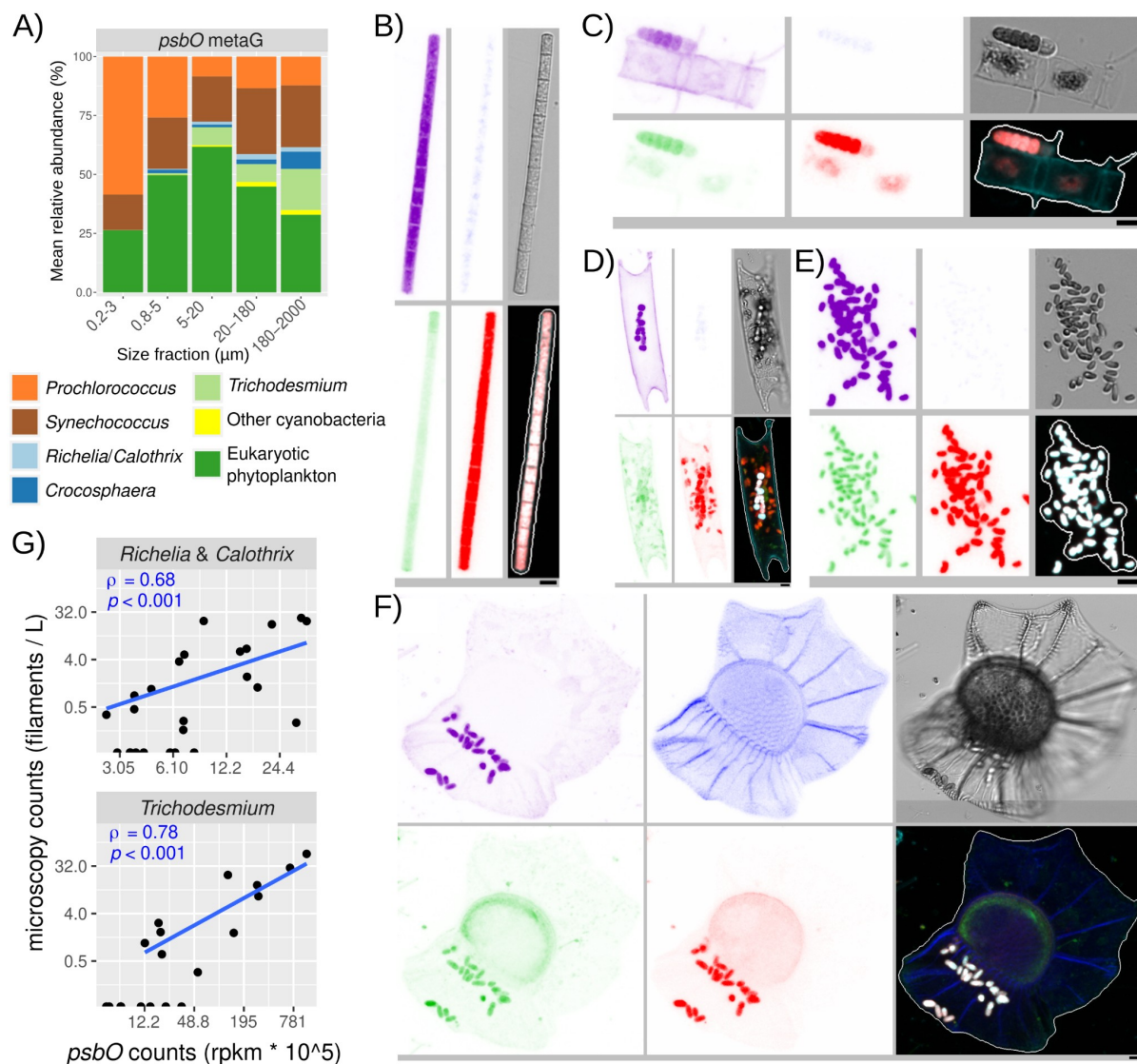


Figure 6: Prokaryotic and eukaryotic phytoplankton community structure across the entire plankton size spectrum. A) Average relative cell abundance of phototrophs across all metagenomes based on *psbO* metagenomic reads. (B-F) Examples of confocal microscopy detection of cyanobacteria in the 20-180 μm 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 in size fraction 20-180 μm. Spearman rho's correlations coefficients and p-values are indicated. rpkm: reads per kilobase per million mapped reads.

119

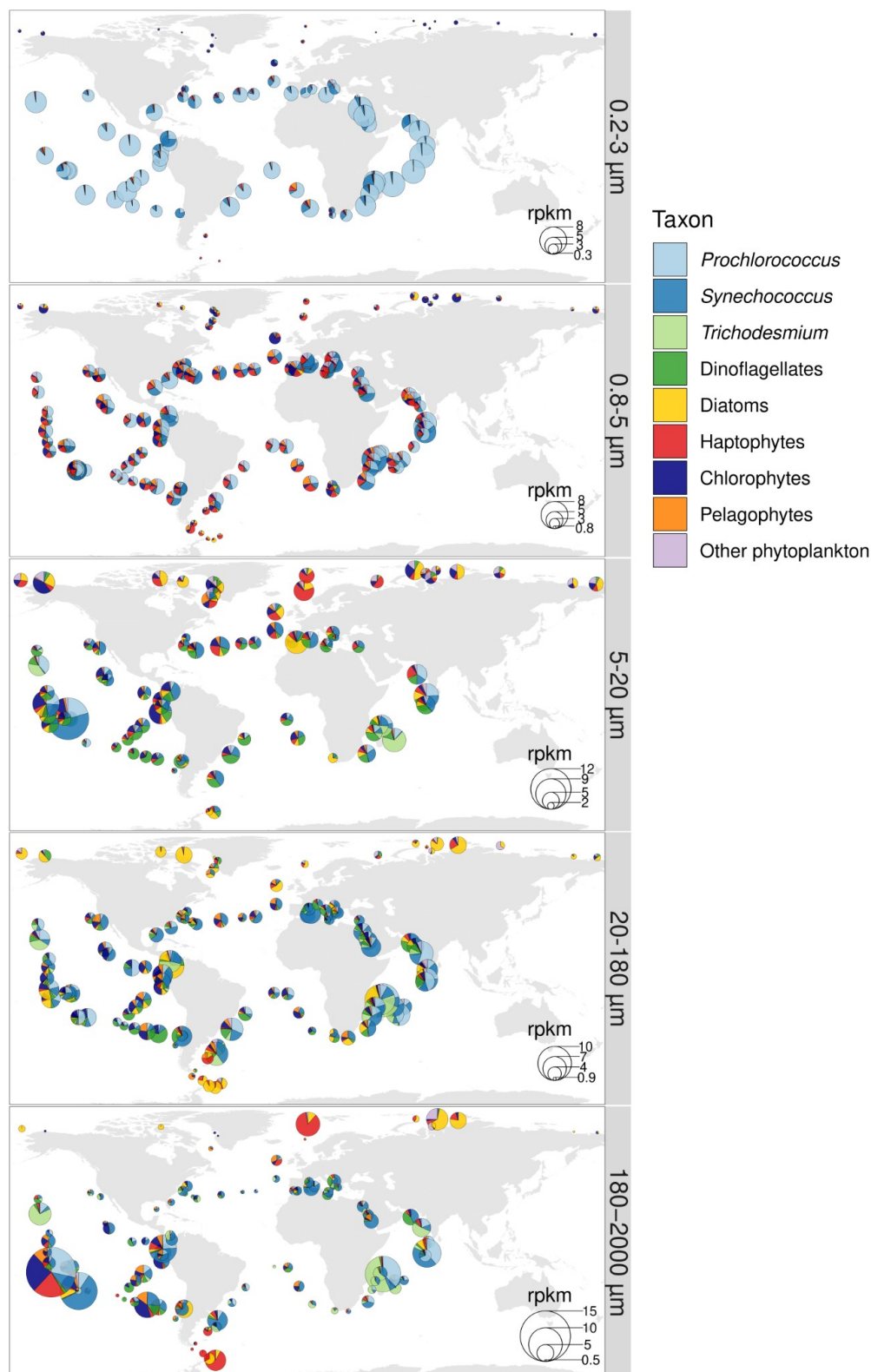


Figure 7: Global biogeographical patterns of marine phytoplankton in surface waters. The pie charts show the *psbO* relative abundance of the main cyanobacteria and eukaryotic phytoplankton in metagenomes derived from different size-fractionated samples. Values are displayed as rpkms (reads per kilobase per million mapped reads). The comparison between the *psbO*-based relative cell abundances versus the patterns corrected by biovolume are displayed in Figure S16. The distribution of the main phytoplankton groups in the size fraction in which they were most prevalent is shown in Figure S17.

120

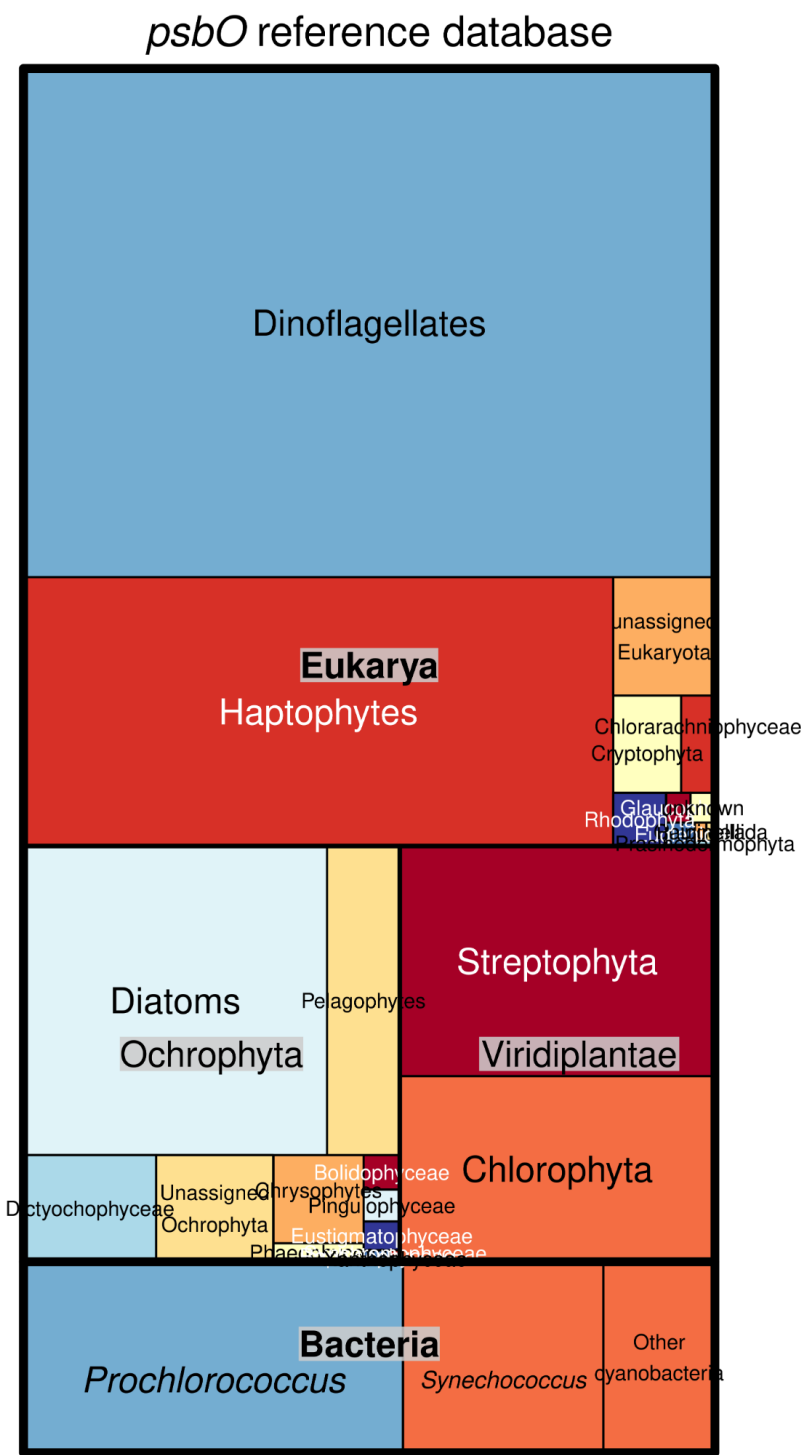


Figure S1: Taxonomic distribution of the curated *psbO* database generated in the current work. It consists of 18,378 unique sequences covering cyanobacteria, photosynthetic protists, macroalgae and land plants. The sequences were retrieved from sequenced genomes and transcriptomes from cultured isolates as well as from the environmental sequence catalogs of Global Ocean Sampling (Rusch et al. 2007) and *Tara* Oceans (Salazar et al. 2019; Carradec et al. 2018; Delmont et al. 2020; 2021). The database can be downloaded from the EMBL-EBI repository BioStudies (www.ebi.ac.uk/biostudies) under accession S-BSST659.

123

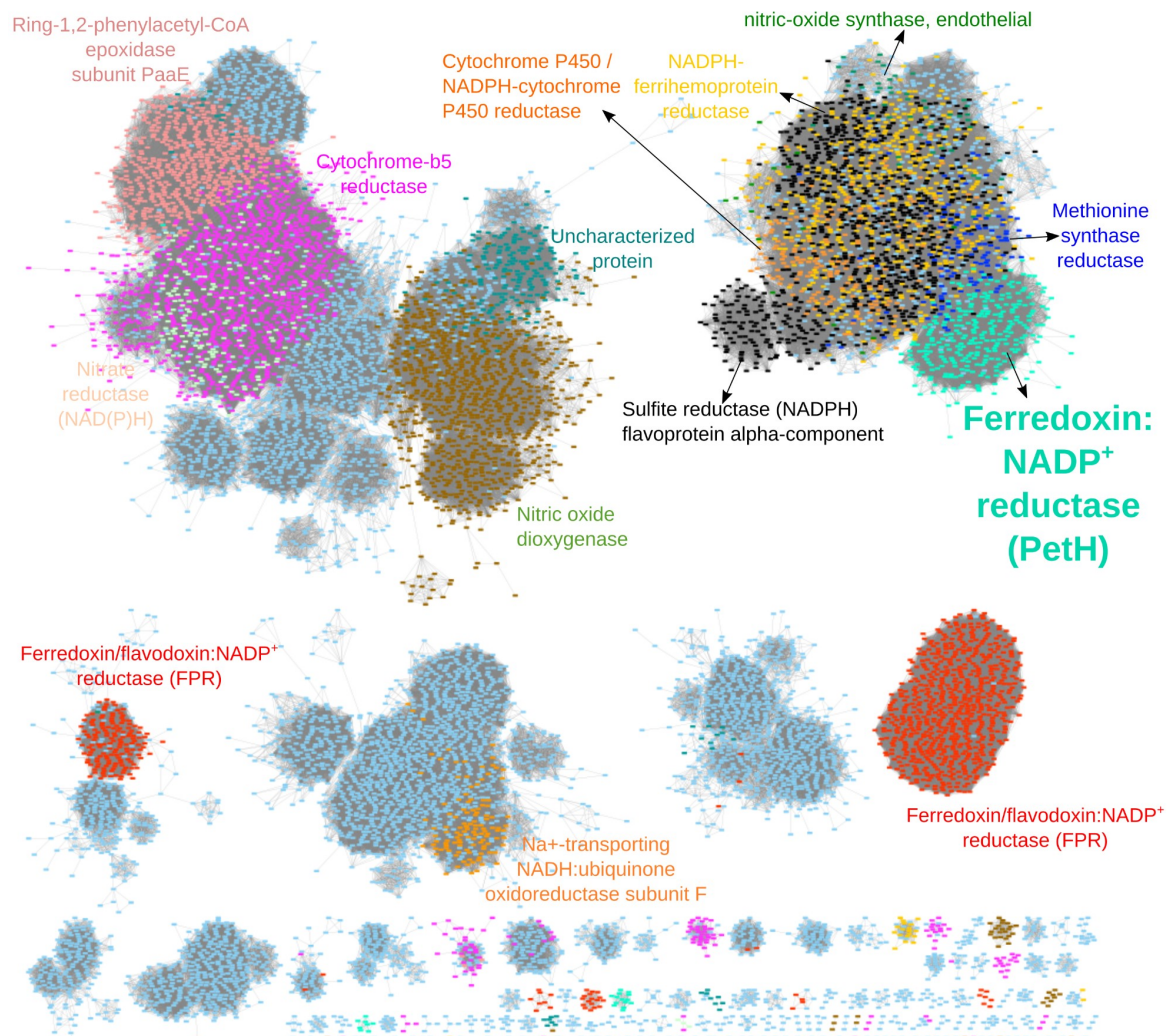


Figure S2: Sequence analysis of ferredoxin:NADP⁺ reductase (PetH) and homologs. Protein similarity network for the Pfam domain NAD_binding_1 (PF00175). Each node corresponds to a representative sequence (clustered at 80% identity by CDHIT) and those sequences with similarity higher than a score cutoff are linked (score cut-off of 22 in blastp alignment). The network was built with sequences retrieved from the literature and from reference genomes and transcriptomes. Nodes are coloured according to their functional assignment based on BlastKOALA. The nodes for FNR are in light green, and includes photosynthetic FNRs as well as FNRs involved in nitrogen metabolism and in non-photosynthetic plastids (Pierella Karlusich & Carrillo 2017 *Phot Res* 134:235–250) and FNR from heterotrophic bacteria acquired by horizontal gene transfer (Catalano Dupuy et al. 2011 *PLoS One* 6:e26736)..

124

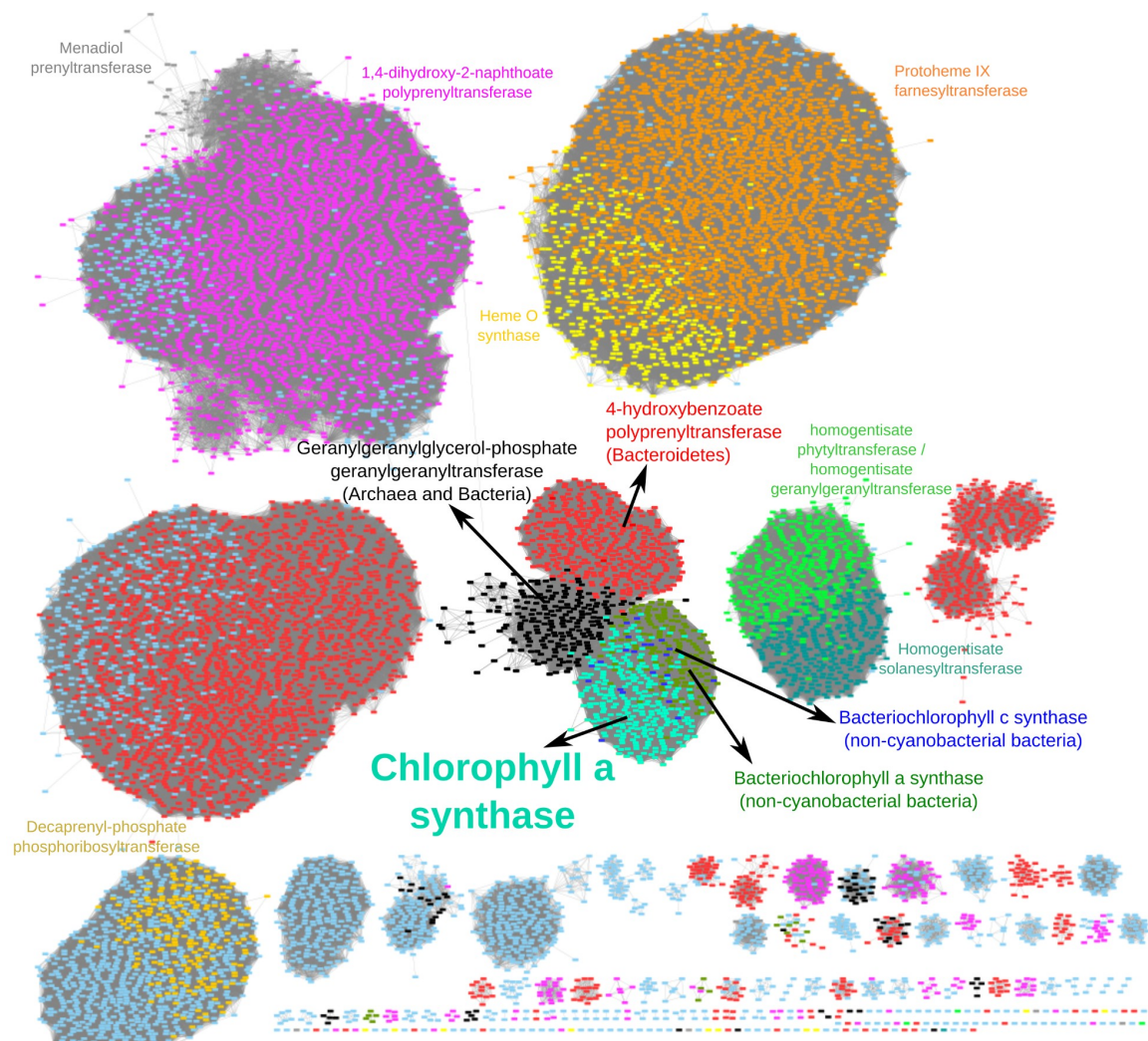


Figure S3: Sequence analysis of chlorophyll *a* synthase (ChlG) and homologs. Protein similarity network for the Pfam domain UbiA (PF01040). Each node corresponds to a representative sequence (clustered at 80% identity by CDHIT) and those sequences with similarity higher than a score cutoff are linked (score cut-off of 25 in blastp alignment). The network was built with sequences retrieved from the literature and from reference genomes and transcriptomes. Nodes are coloured according to their functional assignment based on BlastKOALA.

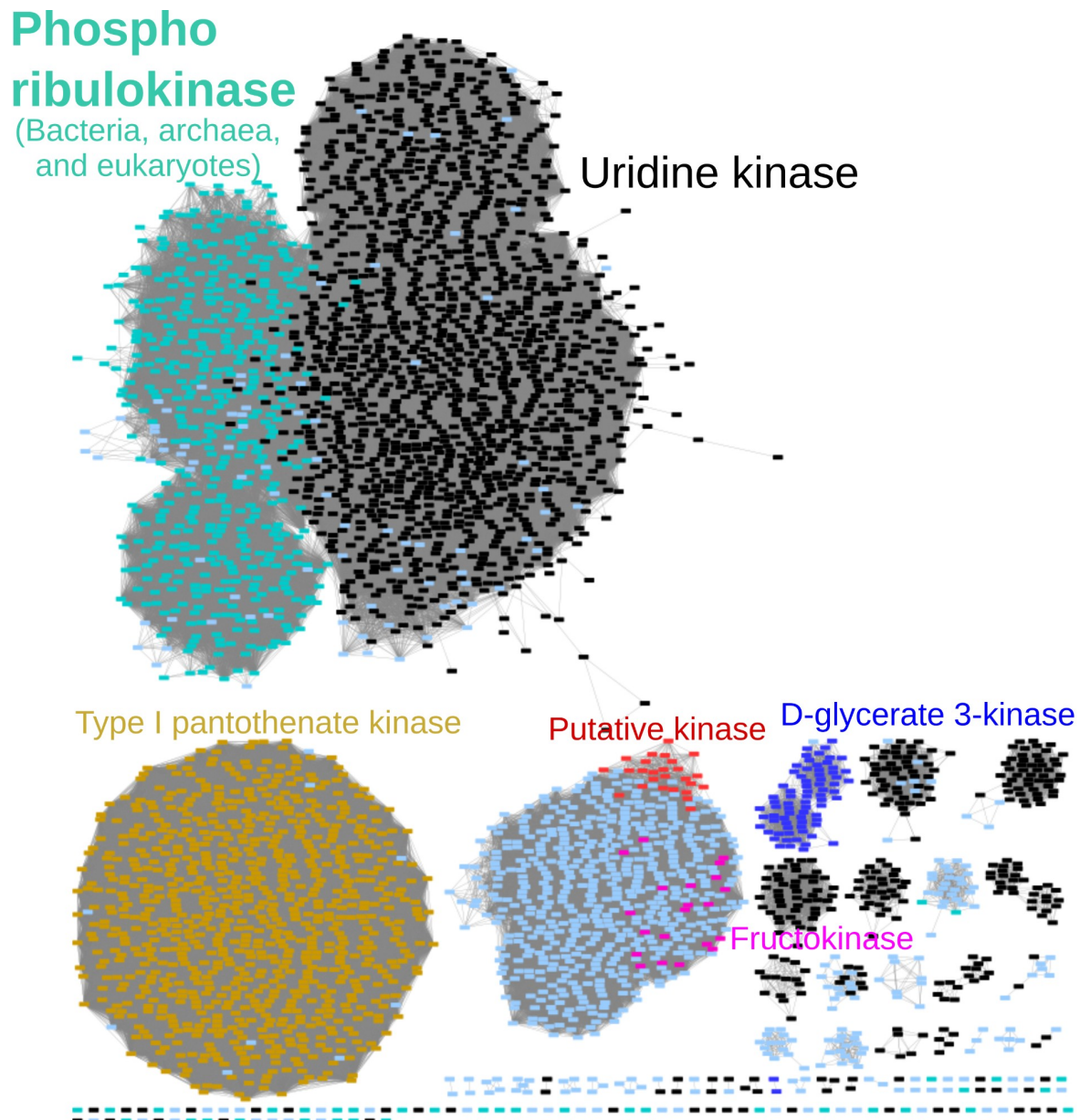


Figure S4: Sequence analysis of phosphoribulokinase (PRK) and homologs. Protein similarity network for the Pfam domain PRK (PF00485). Each node corresponds to a representative sequence (clustered at 80% identity by CDHIT) and those sequences with similarity higher than a score cutoff are linked (score cut-off of 25 in blastp alignment). The network was built with sequences retrieved from the literature and from reference genomes and transcriptomes. Nodes are coloured according to their functional assignment based on BlastKOALA. The nodes for PRK are in light green, and include photosynthetic PRKs as well as those from archaea and non-cyanobacterial bacteria.

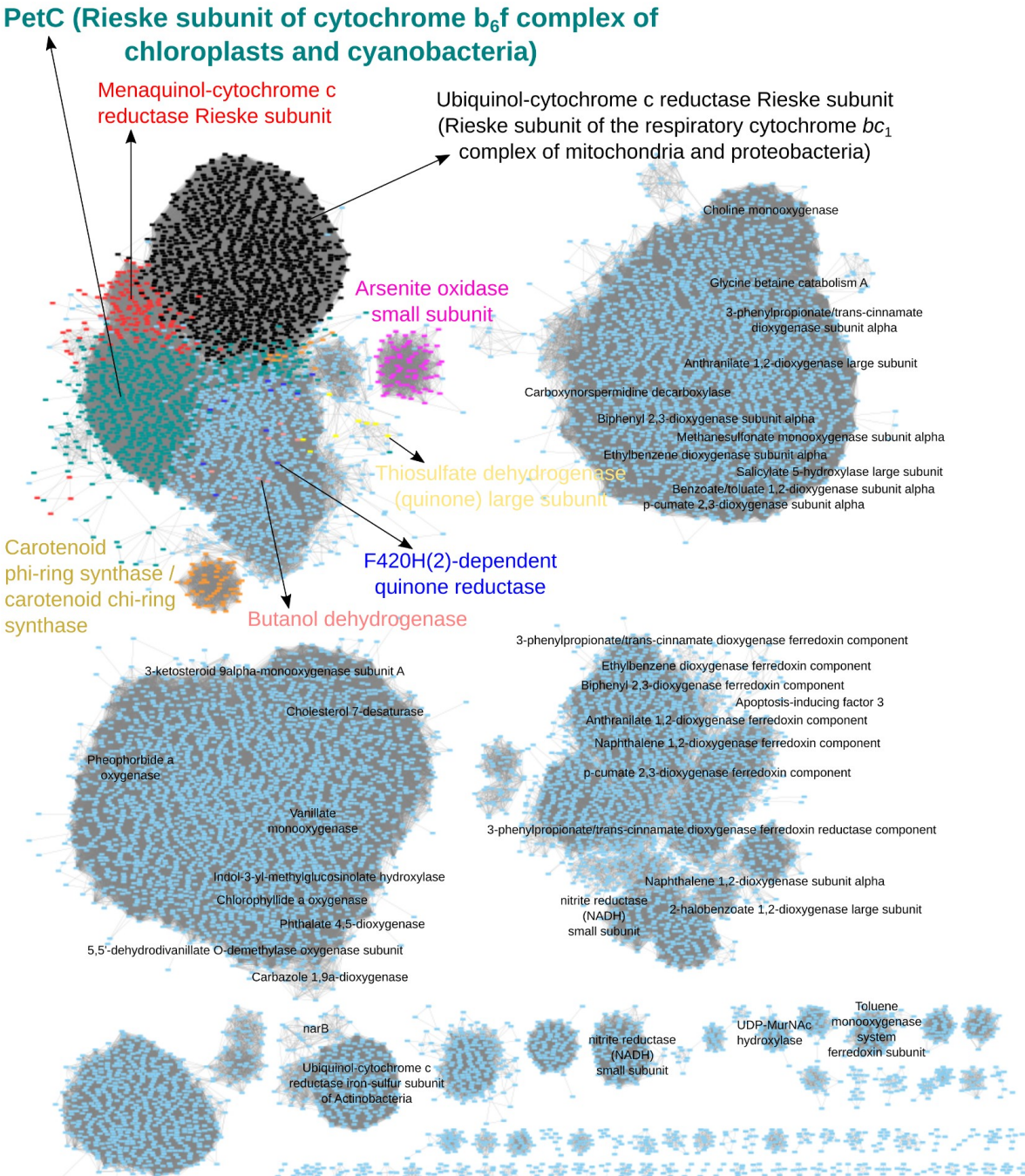


Figure S5: Sequence analysis of PetC (Rieske subunit of the Cytochrome *b₆f* complex) and homologs. Protein similarity network for the Pfam domain Rieske (PF00355). Each node corresponds to a representative sequence (clustered at 80% identity by CDHIT) and those sequences with similarity higher than a score cutoff are linked (score cut-off of 18 in blastp alignment). The network was built with sequences retrieved from the literature and from reference genomes and transcriptomes. Labels correspond to the functional assignment based on BlastKOALA, and for the cluster of interest nodes are coloured according to the functional assignment of their sequences. The nodes for PetC are in green and those for the Rieske subunit of the respiratory Cytochrome *bc₁* complex are in black.

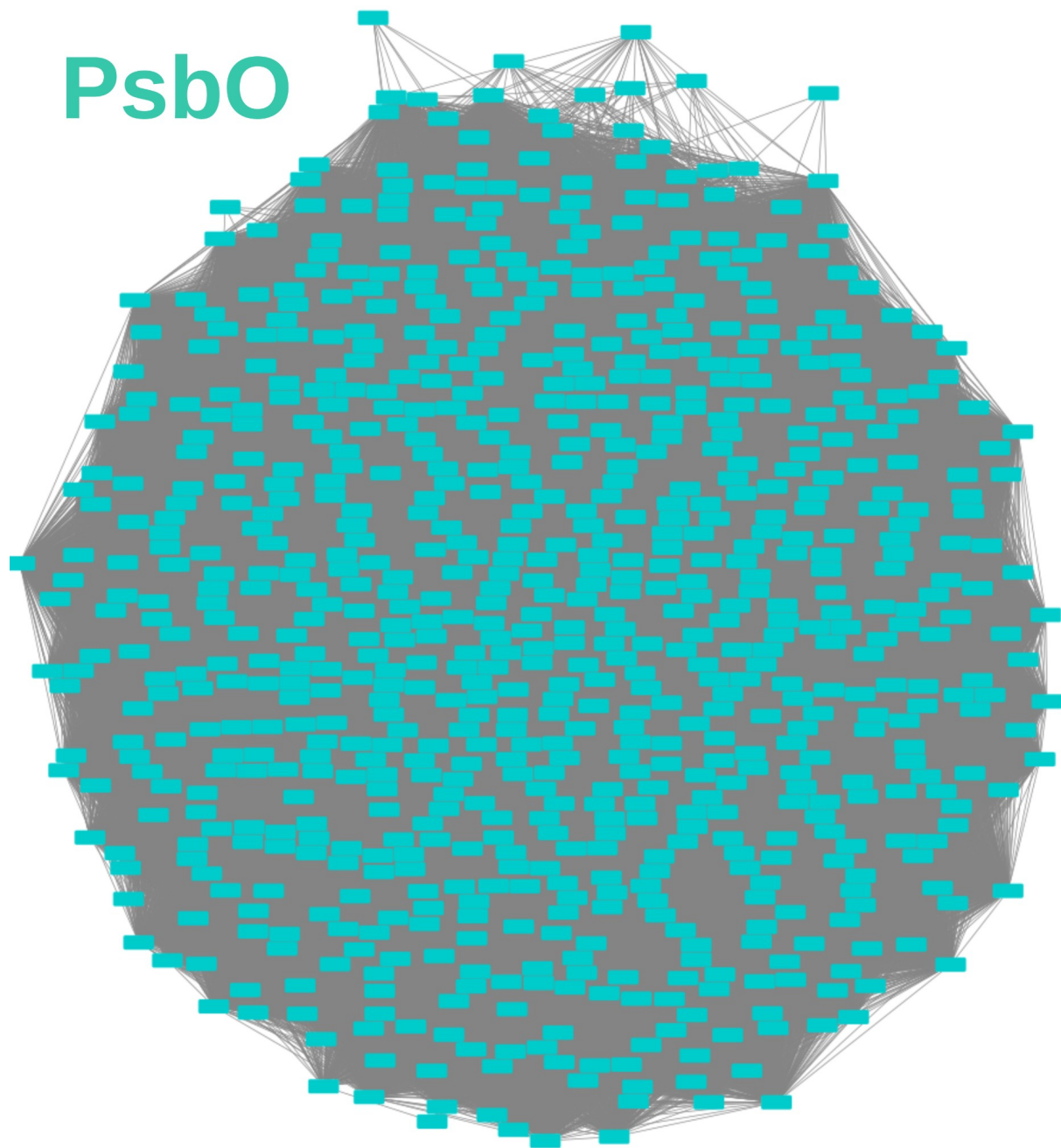


Figure S6: Sequence analysis of PsbO protein. Protein similarity network for the Pfam domain MSP (PF01716). Each node corresponds to a representative sequence (clustered at 80% identity by CDHIT) and those sequences with similarity higher than a score cutoff are linked (score cut-off of 30 in blastp alignment). The network was built with sequences retrieved from the literature and from reference genomes and transcriptomes. Nodes are coloured according to their functional assignment based on BlastKOALA.

133

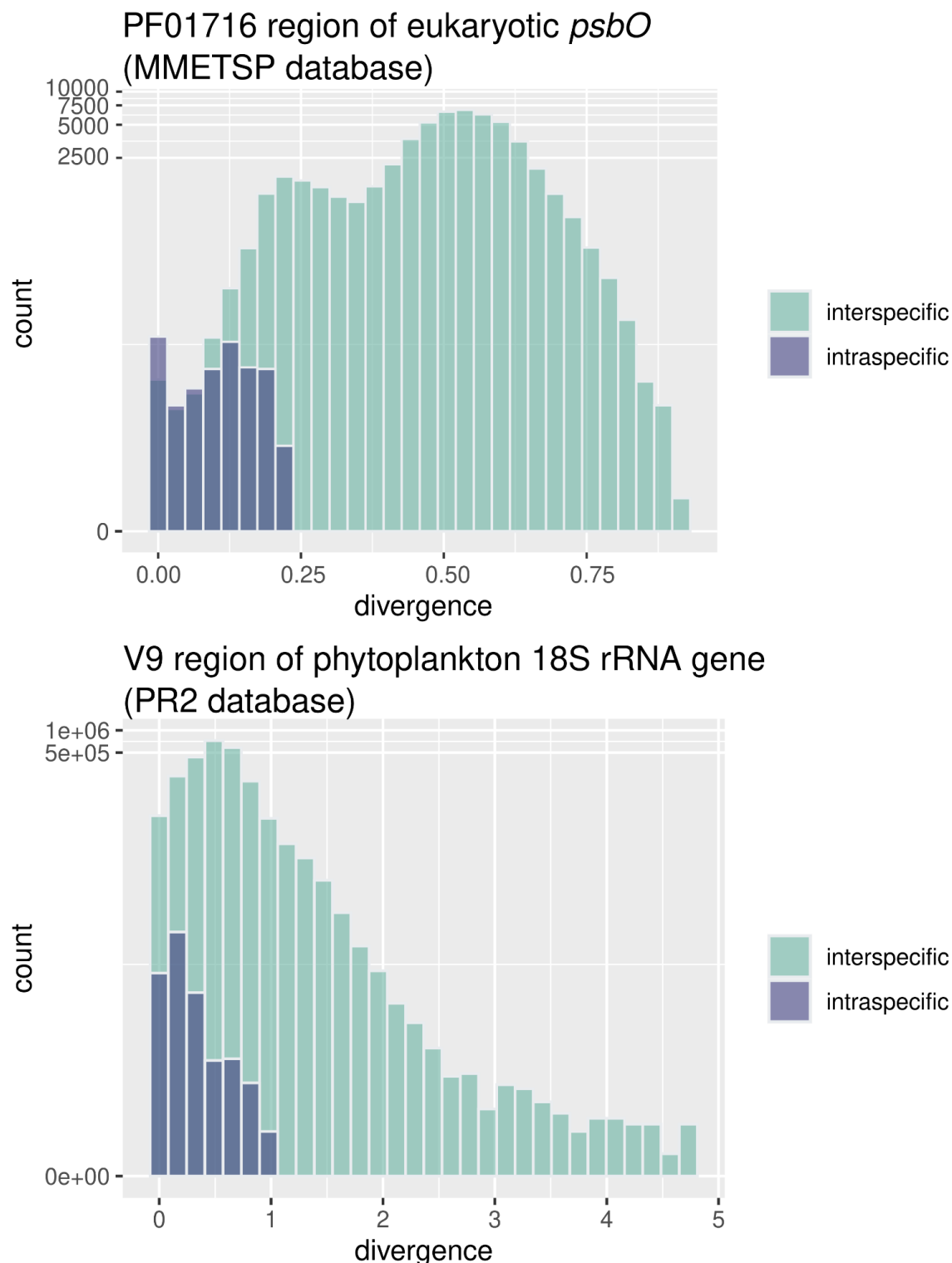


Figure S7: Intraspecific (blue) and interspecific (green) variation in genetic distances of eukaryotic phytoplankton sequences for the region of *psbO* coding for the Pfam domain PF01716 (upper panel) and the V9 region of 18S rRNA gene (lower panel). Sequences were retrieved from MMETSP project (Keeling et al. 2014) for *psbO* and from PR2 database (Guillou et al. 2012) for V9-18S maker. In this later case, the sequence assigned to phytoplankton were selected based on a public functional database available at <https://zenodo.org/record/3768951#.YM4odnUzbuE>.

134

135

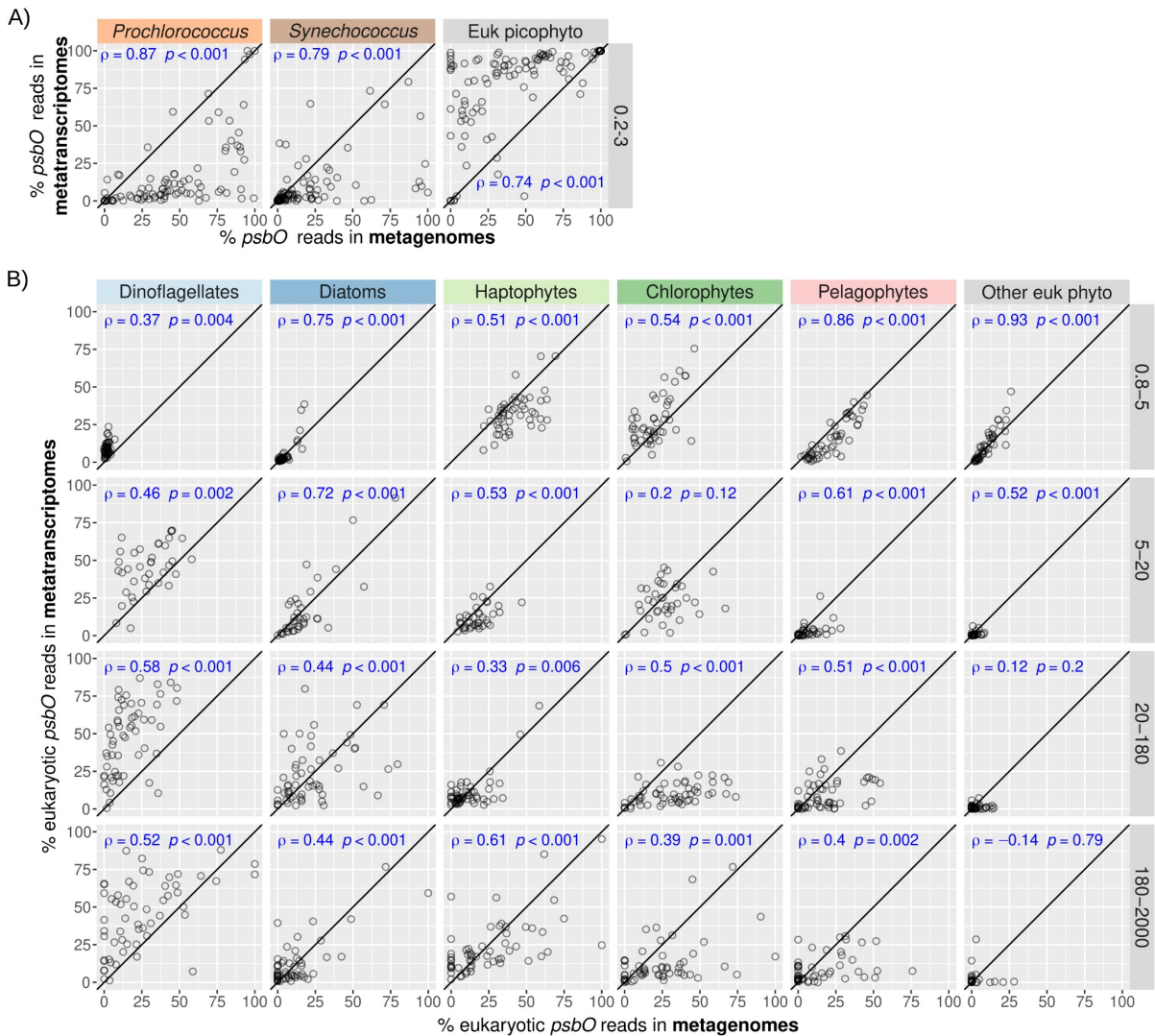


Figure S8: Comparison of relative abundances of *psbO* reads between metagenomes and metatranscriptomes of size fractionated samples. A) Picocyanobacteria and eukaryotic picophytoplankton (size fraction 0.2-3 μ m). B) Eukaryotic phytoplankton in the large size fractions (0.8-5 μ m, 5-20 μ m, 20-180 μ m, 180-2000 μ m), 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.

137

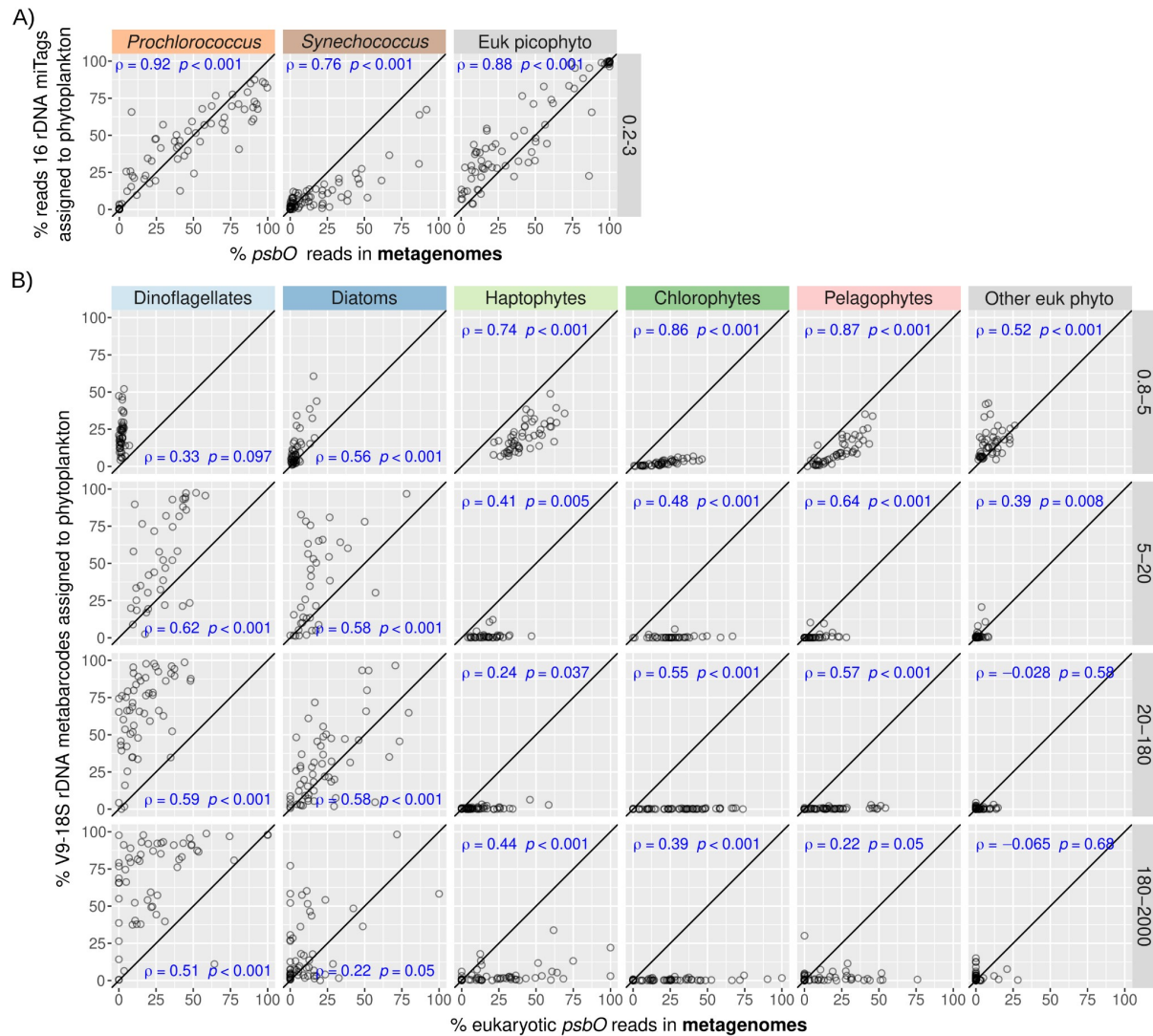


Figure S9: Comparison of relative read abundances of *psbO* and rRNA genes of size fractionated samples. A) Picocyanobacteria and eukaryotic picophytoplankton (0.2-3 μ m) were analysed using the relative abundances for 16S rRNA gene miTags and for *psbO* metagenomic reads. B) Eukaryotic phytoplankton were analysed in the large size fractions (0.8-5 μ m, 5-20 μ m, 20-180 μ m, 180-2000 μ m) 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.

138

139

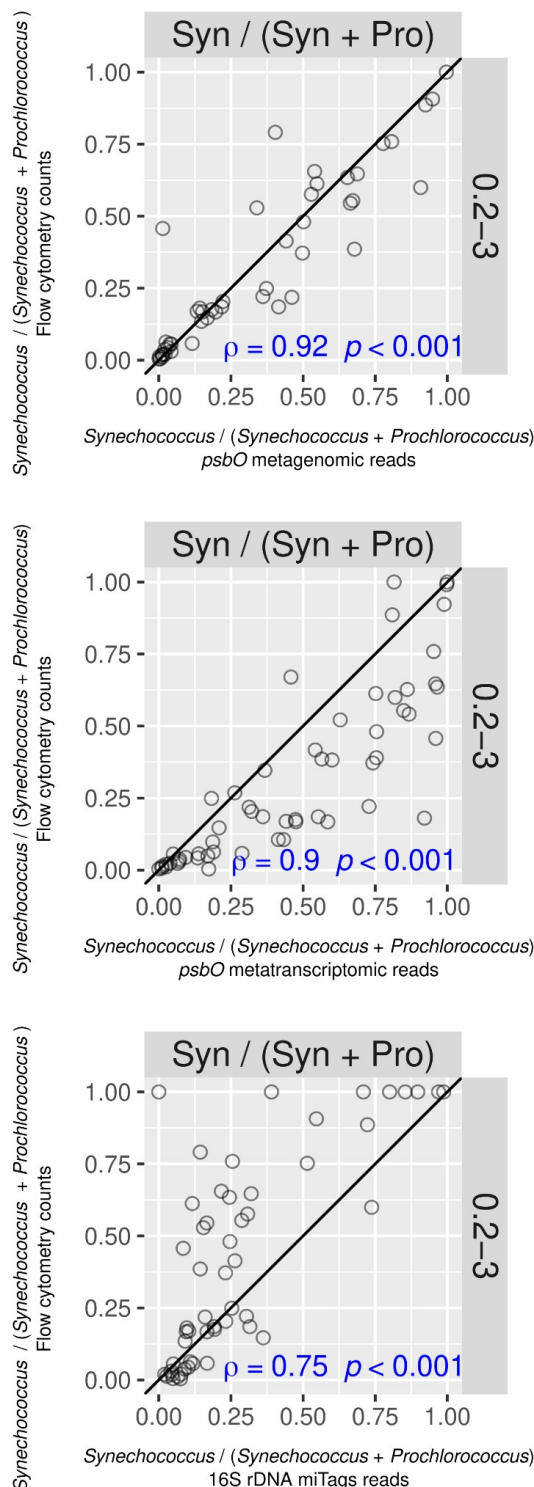


Figure S10: Correlation between the abundance ratio of the picocyanobacteria *Synechococcus* and *Prochlorococcus* obtained with different methodologies. The vertical axis corresponds to the ratio based on flow cytometry while the horizontal axis corresponds to the ratio based on *psbO* metagenomic reads (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.

140

141

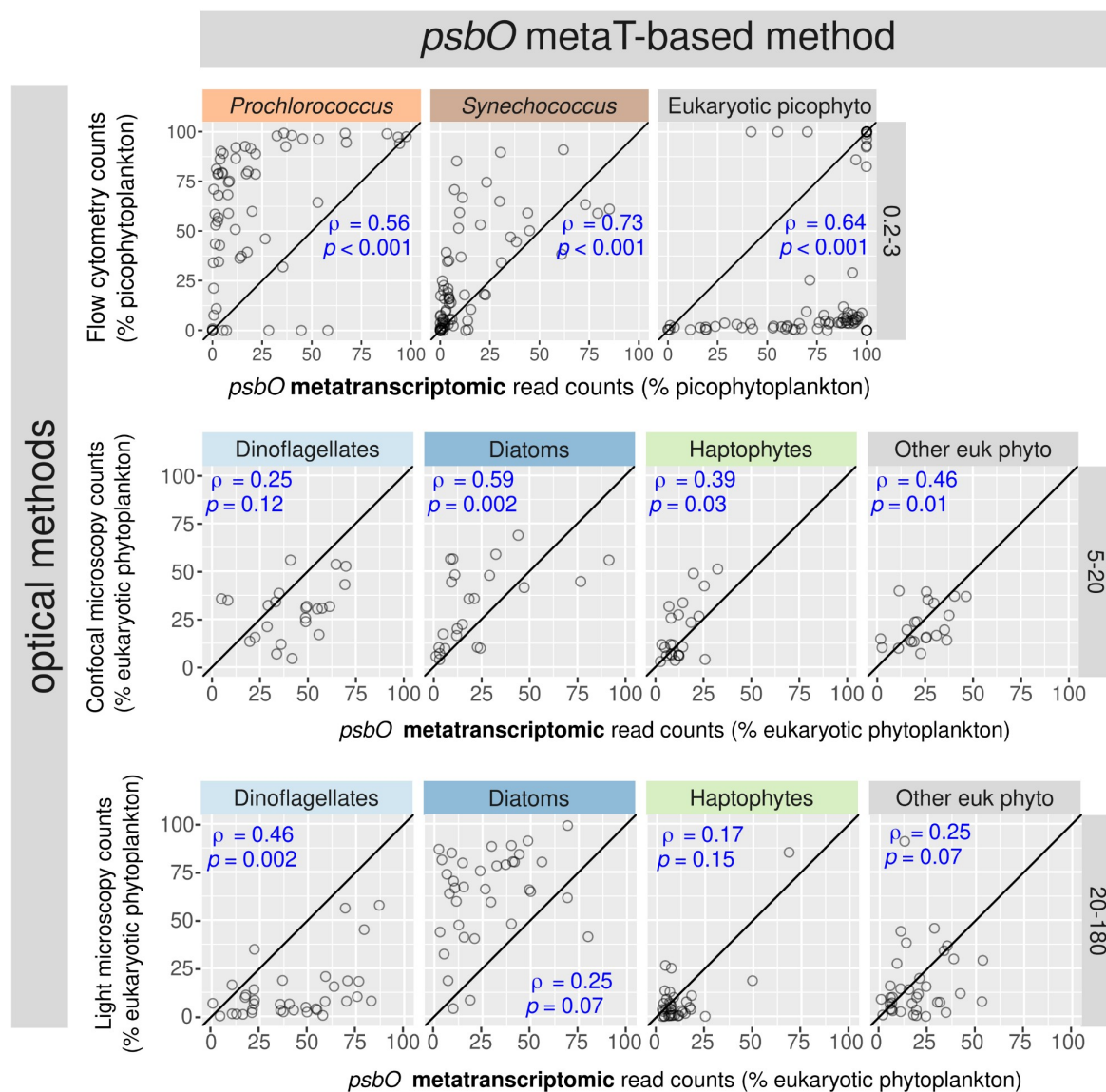


Figure S11: Correlation between relative abundances based on *psbO* metatranscriptomic reads against those based on optical quantifications of different phytoplankton groups. Upper panel: metatranscriptomic *psbO* relative abundances from picophytoplankton (size fraction 0.2-3 μm) were compared with 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.

142

143

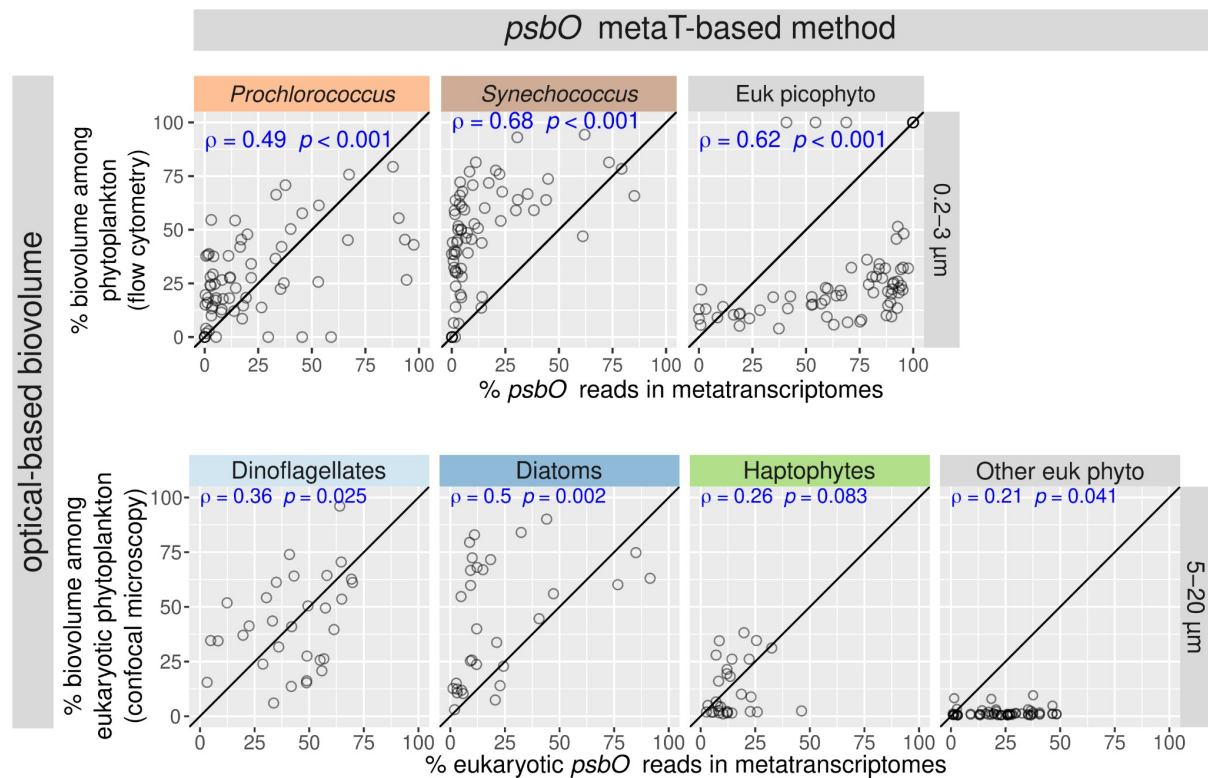


Figure S12: Correlation between relative biovolume (based on optical methods) and relative abundances based on *psbO* metatranscriptomic reads. The upper panel shows the correlations for picophytoplankton (size fraction 0.2-3 μm). The vertical axis corresponds to the relative biovolume based on flow 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.

144

145

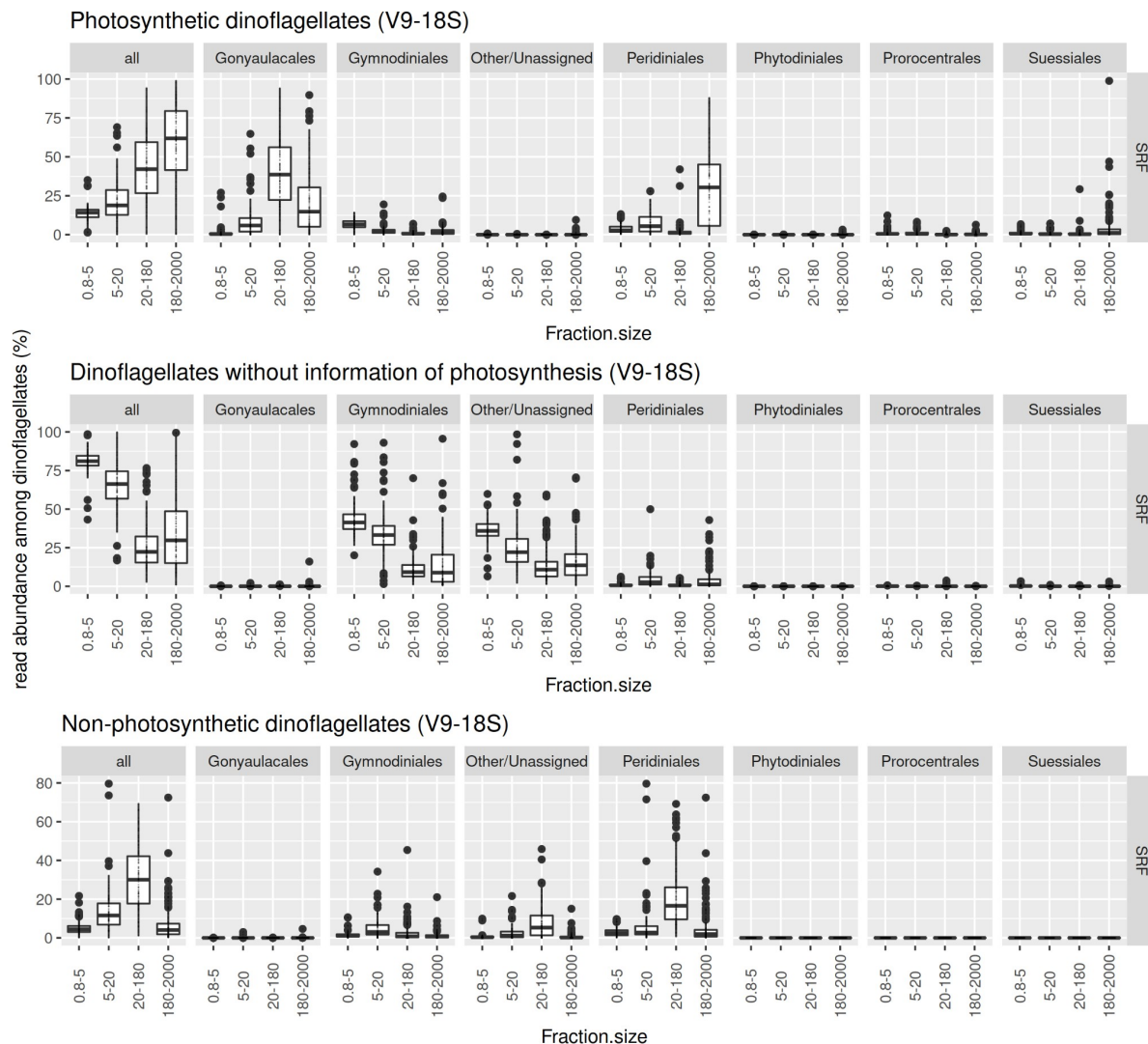


Figure S13: Variations in the abundance of phototrophs vs heterotrophs in the dinoflagellate community of different size fractions based on V9-18S rRNA gene metabarcoding. The plots show the relative abundance of dinoflagellates containing chloroplasts (upper panel) or not (lower panel) as well as those that cannot be classified (middle panel). Note that most of the reads that cannot be reliably classified as a chloroplast-containing taxon correspond to those reads mapping OTUs assigned either as “unknown dinoflagellate” or Gymnodinales order. A description of the trait classification can be found at <http://taraoceans.sb-roscoff.fr/EukDiv/> and the trait reference database is available at <https://zenodo.org/record/3768951#.YM4odnUzbuE>.

146

147

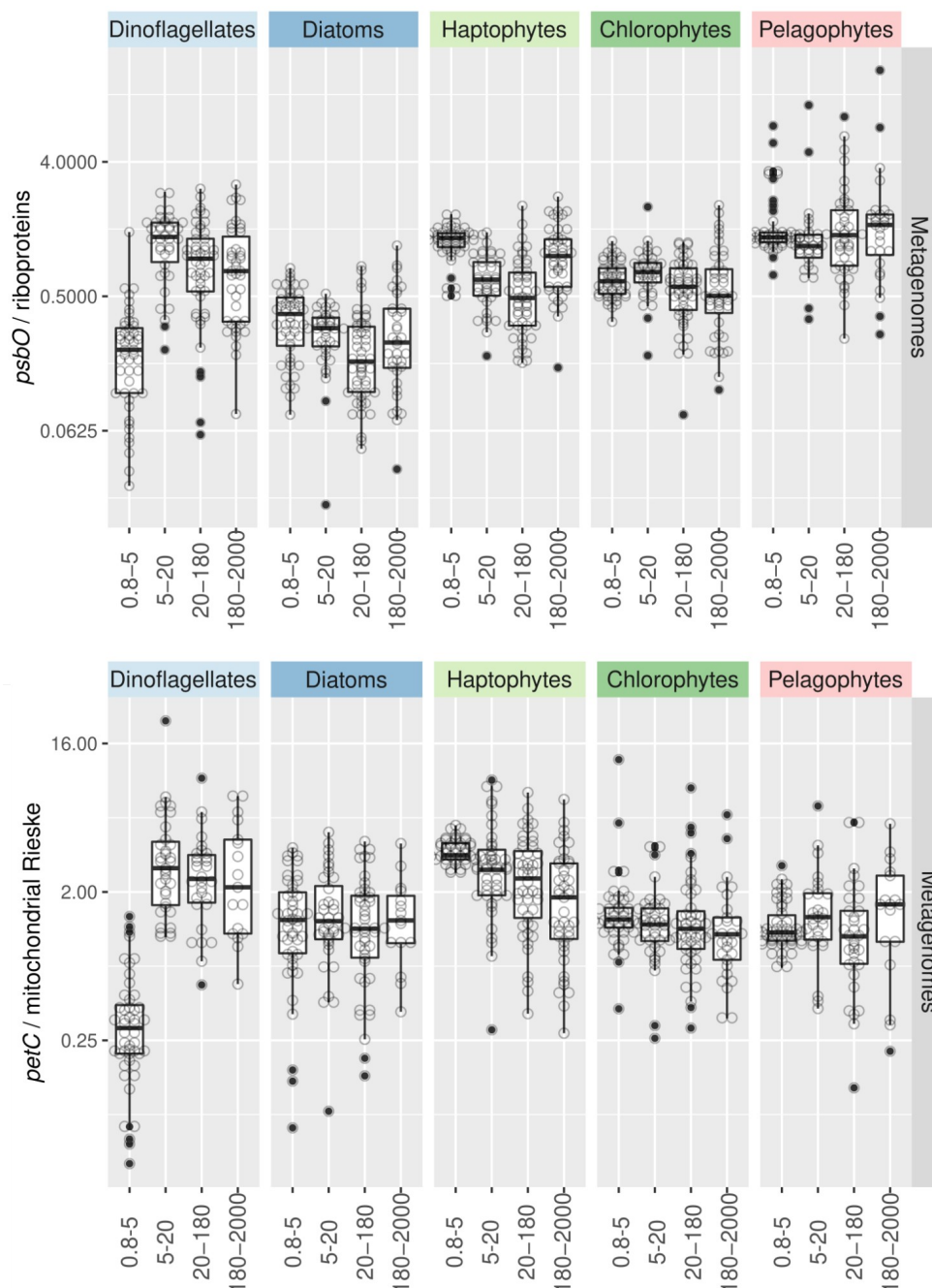


Figure S14: Variations in the abundance of phototrophs vs heterotrophs across size fractions based on different marker genes in metagenomes. The estimation were based on the ratio of metagenomic reads of photosynthetic vs housekeeping single-copy nuclear-encoded genes: *psbO* vs genes coding for ribosomal 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).

1468

148

149

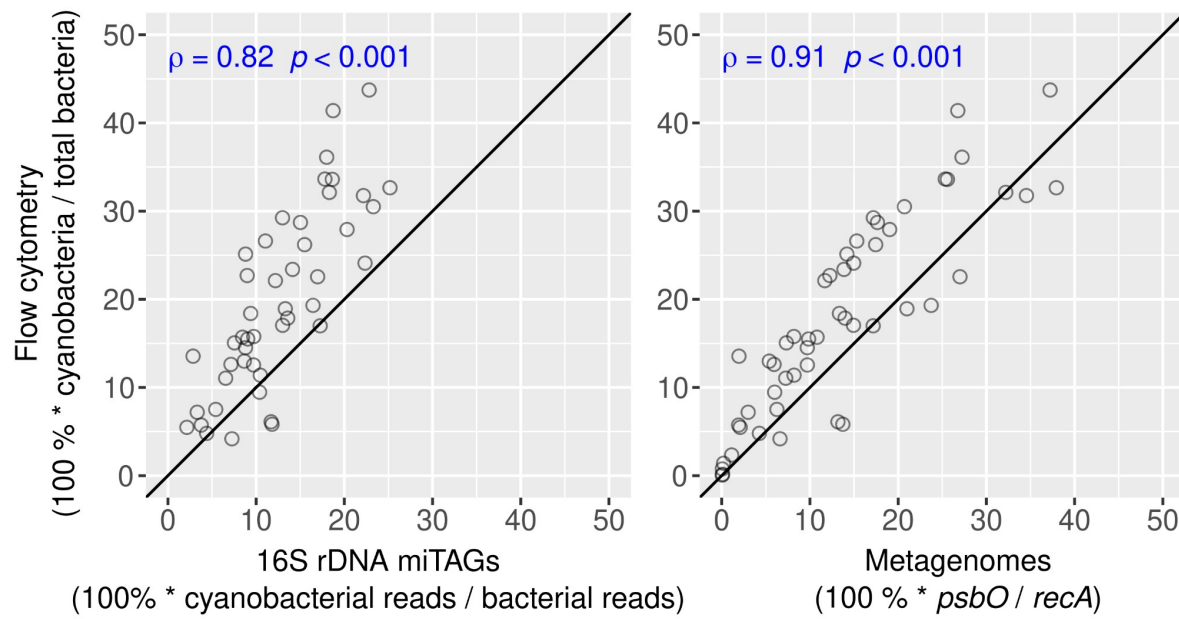


Figure S15: Congruence in the relative abundance of phototrophs among bacterioplankton in 0.2-3 size fraction based on different methods. The vertical axis corresponds to the relative cell abundance based on flow cytometry while the horizontal axis corresponds to the relative read abundance based on 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.

150

151

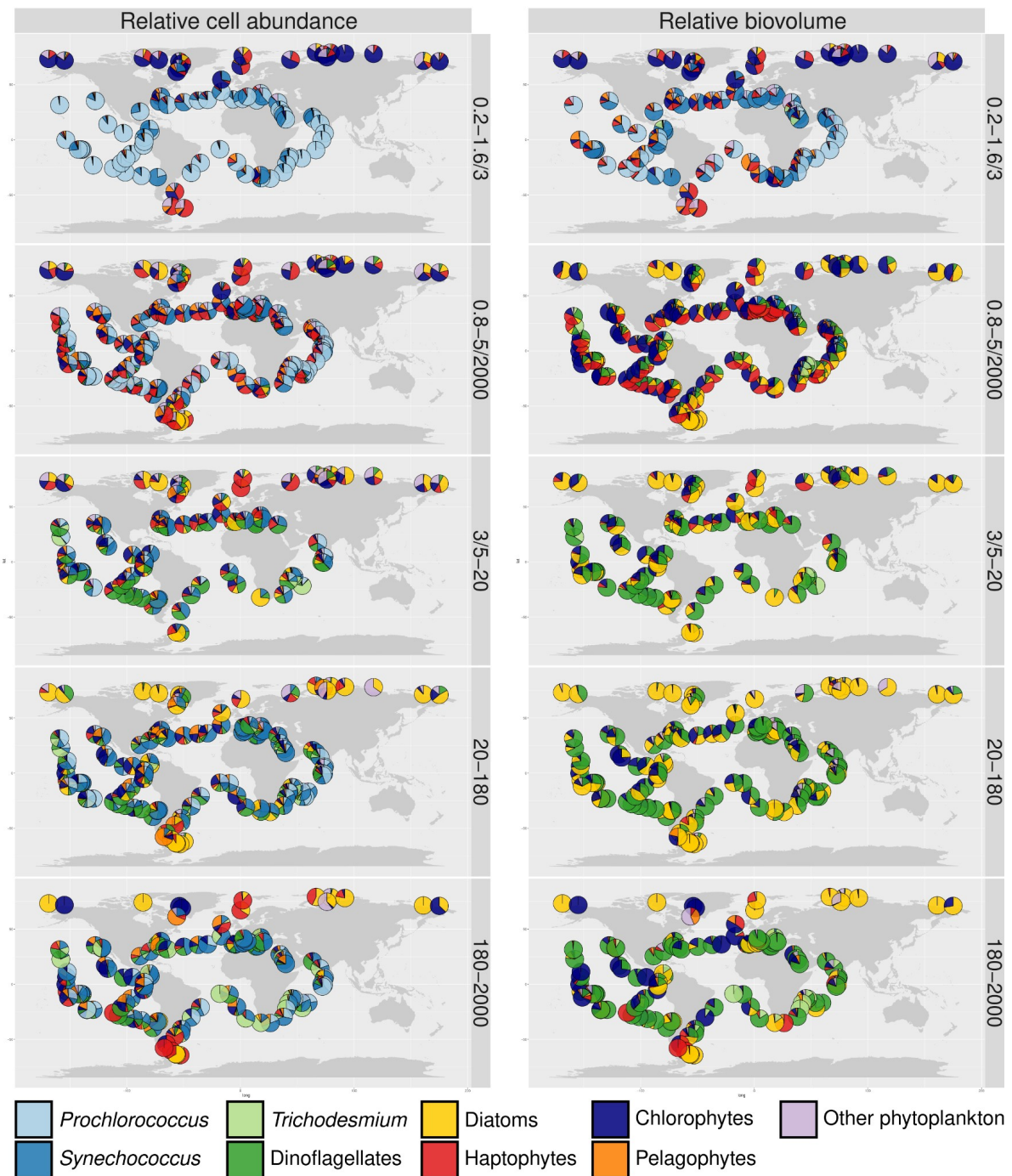


Figure S16: Global biogeographical patterns of relative cell abundance and biovolume for marine phytoplankton in surface waters. A) Relative cell abundances of the main cyanobacteria and eukaryotic phytoplankton based on *psbO* counts in metagenomes derived from different size-fractionated samples. B) Relative biovolume of the main cyanobacteria and eukaryotic phytoplankton based on *psbO* counts corrected by the mean cell biovolume for each taxon (based on optical measurements in *Tara* Oceans samples).

152

153

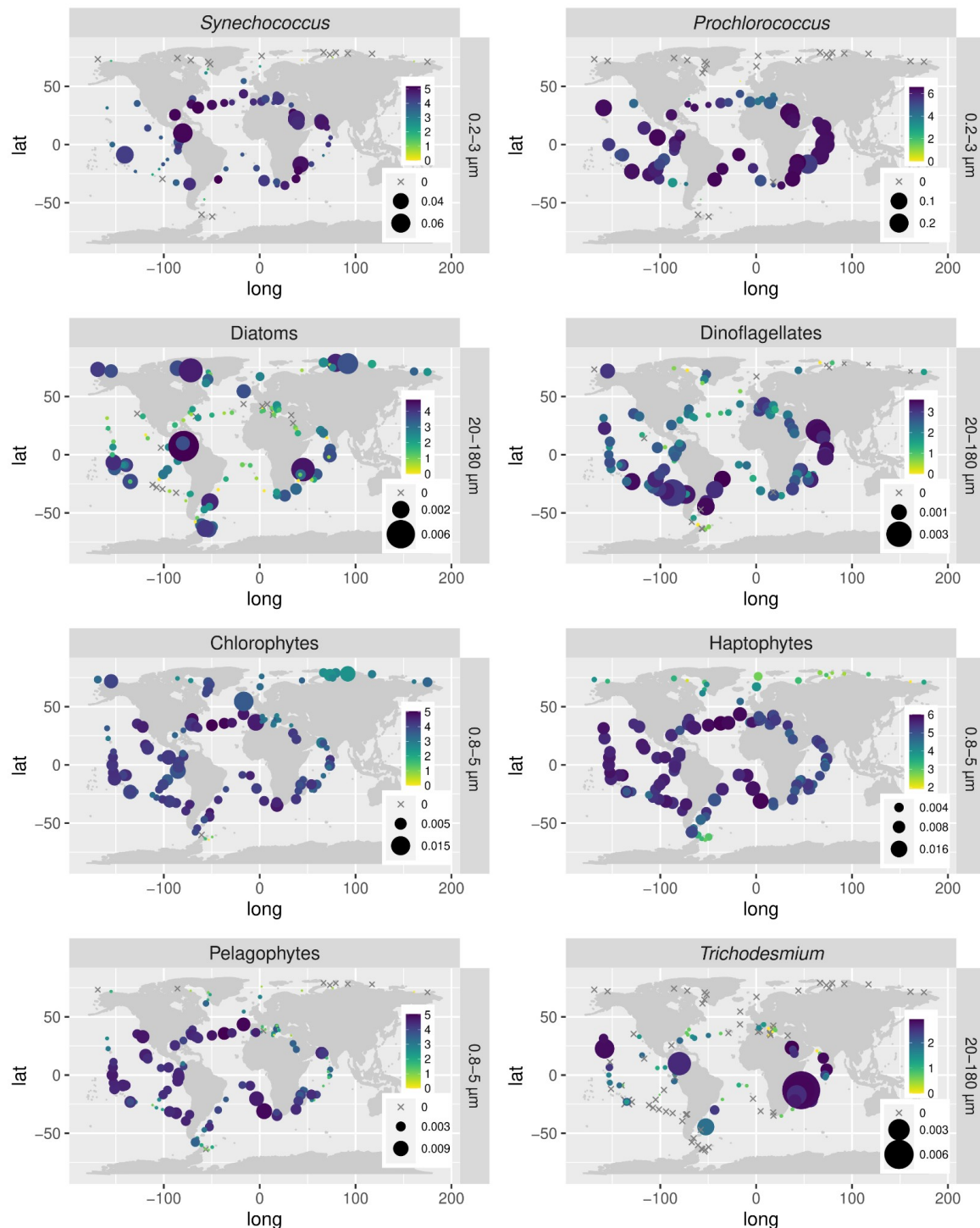


Figure S17: Global biogeographical patterns for main groups of marine phytoplankton in surface waters. The bubbles sizes vary according to the *psbO* relative abundance of the main cyanobacteria and eukaryotic phytoplankton in metagenomes, while color code corresponds to the Shannon index values. Relative abundance values are displayed as rpk (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 7.

154