Structure and trophic diversity of picoeukaryotic primary producers across the global ocean

Laura Rubinat-Ripoll

Structure and trophic diversity of picoeukaryotic primary producers across the global ocean

Présentée pour obtenir le grade de

Docteur de Sorbonne Université

par

Laura Rubinat-Ripoll

Directeur: Dr. Colomban de Vargas

Co-directeurs: Dr. Ramiro Logares et Dr. Olivier Jaillon

Devant un jury composé de:

Présidente:	Dr. Catherine Boyen
Rapporteur I:	Dr. Raffaele Siano
Rapporteur II:	Dr. Susanne Wilken
Directeur de thèse:	Dr. Colomban de Vargas

To my parents:
you are in every spark of energy
lighting up these pages

Résumé

La production primaire est la synthèse de matière organique à partir de molécules inorganiques et, dans la plupart des écosystèmes, elle est réalisée à travers de la photosynthèse. Les microbiens eucaryotes phototrophes et mixotrophes sont les principaux contributeurs à la production primaire dans l'océan oligotrophe global, soutenant les processus de transfert trophique d'énergie et de biomasse à l'échelle planétaire. Malgré la valeur écologique de ces organismes, leur grande diversité taxonomique et fonctionnelle reste largement inconnue. Dans cette thèse, nous utilisons différents types de données moléculaires obtenues à partir de l'expédition circumglobale de *Tara* Oceans pour évaluer la composition et la diversité trophique des producteurs primaires picoeucaryotes dans l'océan oligotrophique. Dans la première partie de la thèse, nous comparons les données métagénomiques et de métabarcodage pour évaluer la diversité évolutive et l'abondance relative des organismes picophototrophes. Nous avons identifié les bactéries phototrophes comme trois fois plus abondantes et significativement moins phylogénétiquement diversifiées que les picoeukayrotes phototrophes. Les Prymnesiophyceae, Mamiellophyceae, Pelagophyceae et Dictiochophyceae sont apparus comme les groupes dominants de picoeucaryotes phototrophes en termes de richesse relative et d'abondance. Dans le quatrième chapitre, nous décrivons un modèle prédictif pour quantifier l'abondance des groupes trophiques dans les échantillons métagénomiques. Cette approche, qui ne dépend pas de la attribution taxonomique, a révélé la dominance des organismes phototrophes dans tous les bassins océaniques, tandis que la contribution des phagomixo- et phago- trophées oscillait autour de 25% de l'abondance relative dans la plupart des échantillons. Dans la dernière étude incluse dans cette thèse, nous évaluons la distribution d'une collection de single-cell amplified genomes dans les échantillons de *Tara* Oceans. Nos résultats soutiennent que la technique de séquençage unicellulaire a le potentiel de récupérer le génome des protistes dominants dans l'océan oligotrophique global avec un effort d'échantillonnage relativement modeste. Dans l'ensemble, ce travail décrit un certain nombre d'approches basées sur des données moléculaires pour évaluer la distribution et la diversité des producteurs primaires dans les environnements marins.

Mots-clés : producteurs primaires, eucaryotes microbiens, mixotrophie, méta-omiques, modélisation trophique, *Tara* Oceans.

Abstract

Primary production is the synthesis of organic matter out of inorganic molecules, and in most ecosystems is achieved through photosynthesis. Eukaryotic microbial phototrophs and mixotrophs are main contributors to primary production in the global oligotrophic ocean, supporting processes of energy and biomass trophic transfer at a planetary scale. Despite the ecological value of these organisms, their wide taxonomic and functional diversity remains largely unknown. In this thesis, we use different types of molecular data obtained from the *Tara* Oceans circumglobal expedition to assess the composition and trophic diversity of picoeukaryotic primary producers in the oligotrophic ocean. In the first part of the thesis, we compare metagenomic and metabarcoding data to assess the evolutionary diversity and relative abundance of picophototrophic organisms. We identified phototrophic bacteria as three-fold more abundant and significantly less phylogenetically diverse than phototrophic picoeukayrotes. Prymnesiophyceae, Mamiellophyceae, Pelagophyceae and Dictiochophyceae appeared as the dominant groups of phototrophic picoeukaryotes in terms of relative richness and abundance. In the fourth chapter, we describe a predictive model to quantify the abundance of trophic groups in metagenomic samples. This taxonomy-free approach revealed the dominance of photo-trophic organisms across all ocean basins, while the contribution of phagomixo- and phago-trophs oscillated around 25% of the relative abundance in most samples. In the last study included in this thesis we assess the distribution of a collection of single-cell amplified genomes across all *Tara* Oceans samples. Our results argue that single-cell sequencing technique has the potential to recover the genome of dominant protists in the global oligotrophic ocean with a relatively low sampling effort. Overall, this work describes a number of approaches based on molecular data for the assessment of primary producers distribution and diversity in marine environments.

Keywords: primary producers, microbial eukaryotes, mixotrophy, meta-omics, trophic modeling, *Tara* Oceans.

Acknowledgments

"And what in fluctuating appearance hovers,
you shall fix by lasting thoughts"

— Goethe

Writing this page and being able to thank the people who helped me during the journey is one of the most gratifying parts of wrapping up the thesis. Science wouldn't be possible without teamwork and I've been lucky enough to be surrounded with an amazing crew.

Colomban, thank you for giving me the opportunity to work with such privileged data. You're really moved by passion and I've learned lots from your endless enthusiasm. You value freedom above all and by giving me autonomy you've helped me so much to grow up as a scientist. Another big thank you to Ramiro and Olivier for your trust and for moving things forward when it was needed the most. This work wouldn't had been possible without your ideas and guidance.

Thank you Susanne, Raffaele and Catherine for going through these pages, complementing them with your experience and knowledge, and helping us writing the final dot.

Nico, Cédric, Ewen and Mary: I will miss you. Thank you for all the conversations that lighted up some bulb, for all the shared moments around good food, for teaching me your language, and for the respectful silences during some cloudy days. I could write a whole chapter about Nico's generosity, Cédric wiseness, the bouncy spirit of Ewen and Mary's vitality, but for getting a better insight I address the reader to pass by the office no. 329 from the Station. Coffee and cookies are served at 4:30pm.

To all the friends, colleagues and adoptive families from the Station, the ICM, Genoscope, HITS and the AMNH: thanks for your warm hosting. You've helped me in ways you probably don't imagine and you've made of this thesis a much richer experience.

John, I feel so fortunate to have had the chance to work with you. Thank you for your kindness and for teaching me about modelling. If I ever get to work on something meaningful to help the planet I'll try to keep in mind newyorkers' vibrant optimism.

To Ramon and Helena, thanks for caring about our projects and for always being ready to help. Thanks a lot also to the rest of colleagues from SINGEK and The Paper Mill; every trip with you has been lots of fun and lots of learning.

Als pares, al Josep M. i a la Dúnia, a la iaia, i a tota la superfamília de Cal Clavetaire i Cal Rubinat, els puntals més valuosos. Us toca fer paciència i encoratjar-me cada vegada que em complico la vida amb alguna originalitat i no falleu mai. No us dic prou com d'aprop us sento sempre que sóc lluny de casa. Gràcies per tot i per tant.

And a few last words for you, Pierre, the best thesis buddy. You're an expert of transforming fears into smiles and you've shown admirable patience skills. Walking some steps of this journey by your side has been the biggest present and I look forward to keep exploring life with you. Thanks a million, boy.

Contents

List of Figures

List of Tables

List of Acronyms

ADCL	Average Distance to the Closest Leaf
BLAST	Basic Local Alignment Search Tool
DCM	Deep Chlorophyll Maximum
HMM	Hidden Markov Model
metaB	metabarcoding
metaG	metagenomic
MMETSP	The Marine Microbial Eukaryote Transcriptome Sequencing Project
OTU	Operational Taxonomic Unit
PPE	phototrophic picoeukaryote
rDNA	Ribosomal DNA
SAG	Single-cell Amplified Genome
SUR	surface

1. General Introduction

1.1 Primary producers and their role in the biosphere

Life on Earth is sustained by primary producers: organisms capable to synthesise organic matter out of inorganic carbon-containing molecules. Virtually all primary production in our planet is achieved through oxygenic photosynthesis. This process is carried out in the so-called *reaction centres*, complexes of proteins and light-absorbing molecules stored within cells. Fuelled by carbon dioxide, water and light energy, oxygenic photosynthesis yields carbohydrates and oxygen, as described in the following equation:

$$CO_2 + H_2O + photons \rightarrow [CO_2] + O_2$$

Through this chemical reaction, photosynthetic organisms transform light energy from the sun into chemical energy that can later be used by the rest of organisms in the ecosystems. Besides transferring energy and biomass to upper trophic levels, phototrophs also rule two central processes for life on Earth: oxygen and carbon cycles. They synthesise the $\sim 8,390$ x10^{12} moles O2 yr^{-1} needed to maintain the current oxygen levels in the atmosphere (Walker, 1980), and absorb 104.9 Pg of CO2 per year (Field et al., 1998), contributing to carbon sequestration and balancing climate dynamics. Phototrophs are widespread around the terrestrial globe (Figure 1.1) and organisms from terrestrial and oceanic ecosystems contribute to the global net primary production in comparable sums (Field et al., 1998).

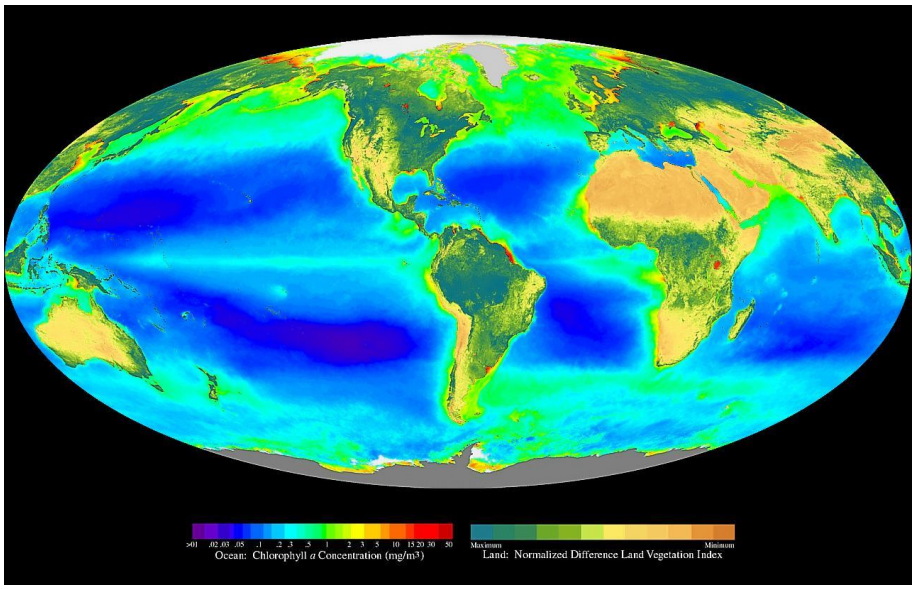


Figure 1.1 – Global distribution of phototrophs abundance across oceanic and terrestrial ecosystems from September 1997 to August 2000. The quantification of biomass was obtained through estimates of chlorophyll A in marine regions and through the normalised difference vegetation index (NDVI) in the terrestrial biosphere. The image is provided by the SeaWiFS Project, NASA/Goddard Space Flight Center and ORBIMAGE.

1.2 Evolution and diversity of marine primary producers

Photosynthesis evolved in cyanobacteria 2.7 billion years ago (Gould, Waller & McFadden, 2008). Cyanobacteria was the only group of organisms able to perform oxygenic photosynthesis up until the Proterozoic, when the capacity for photosynthesis was transferred to eukaryotes through an event of primary endosymbiosis (Reyes-Prieto, Weber & Bhattacharya, 2007). During that process, the ancestor of glaucophyta, red algae and green algae engulfed a cyanobacteria, giving rise to the origin of plastids (Figure 1.2).

Eukaryotic plastids radiated into further eukaryotic groups through events of secondary and tertiary endosymbiosis. Green algae were captured by euglenids, chlorarachnophytes and a group of dinoflagellates, while plastids from red algae were transferred to haptophytes, cryptophytes, stramenophyles and alveolates (Keeling et al. 2010). Some of the taxa descending from red algae lost their plastids and became non-photosynthetic, but the groups from this lineage that still maintain the capacity for photosynthesis (including prymnesiophyctes, dinoflagellates and ochrophytes) are the most abundant groups of eukaryotic phototrophs in today's oceans (Pierella et al., *in press*).

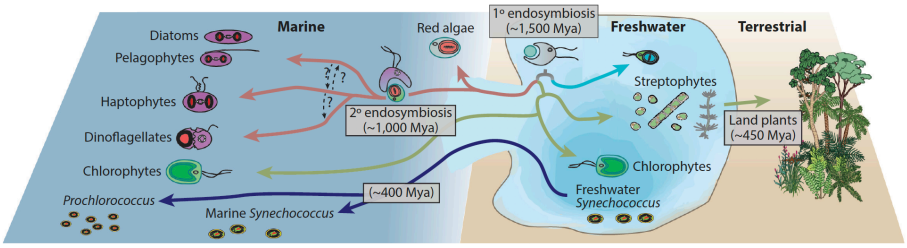


Figure 1.2 – Representation of the main events along the process of photosynthetic groups radiation and dispersion. Phototrophy was transferred to eukaryotes when a cyanobacteria was taken up by a eukaryotic heterotroph around 1.5 billion years ago. Glaucophyta, red algae and green algae diverged from that ancestor and eventually transitioned towards marine ecosystems. Dominant groups of eukaryotic phototrophs in nowadays oceans include chlorophytes, dinoflagellates, haptophytes, pelagophytes and diatoms; except for chlorophytes, all these groups are characterised by hosting plastids from red algae acquired through secondary or tertiary endosymbiosis. On the other hand, terrestrial ecosystems were colonised by land plants, which evolved from green algae ∼450 million years ago. Image reprinted from Pierella and Bowler (in press).

1.3 Mixotrophy: when photosynthesis is not enough

Mixotrophs are organisms capable to use multiple metabolic strategies to acquire nutrients or energy. Herein, we restrict the term "mixotrophy" (or phago-mixotrophy) to refer those microbes that feed through a combination of photosynthesis and phagocytosis. As described above, phototrophy is the capacity to fix inorganic carbon using the energy from the sun. On the other hand, phagocytosis involves the assimilation of prey through vacuole internalisation and its posterior digestion within a lysosome. Mixotrophs have been largely overlooked in planktonic communities due to the technical difficulties to study this trophic group. However, along the last decades it has become more and more clear that mixotrophy is not only a widely spread nutritional strategy among protists (Stoecker et al. 2017) but also a key piece for the functioning of planktonic systems (e.g. Ward and Follows, 2016).

Mixotrophy appeared through different evolutionary paths in groups that acquired the capacity for photosynthesis at the same time that they maintained the mechanisms for phagotrophy. Mixotrophs can show very different nutritional strategies, ranging from almost purely phototrophic metabolism to mostly relying on phagotrophy. Studies like Jones 1997, Stoecker 1998 and Mitra et al. 2016 attempted to delimit mixotrophs functional diversity by establishing classification frames based on physiological attributes. In the later study, Mitra et al. (2016) divided mixotrophs according to the following traits: inherent capacity for phototrophic and phagotrophic activity, retention of plastids from prey (kleptoplastidity), and achievement of autotrophy through phototrophic endosymbiosis. Organisms from each of these groups can respond very differently to prey and light availability (Liu et al., 2016), and because of the technical complexity to study

mixotrophs' physiology and to grow them in cultures, we still lack information about the exact nutritional strategy of many mixotrophs.

Recent studies on plankton modelling have shown that the presence of mixotrophs in microbial communities enhances primary production and energy transfer to higher trophic levels in ocean ecosystems (Mitra et al., 2014; Ward and Follows, 2016). The distribution and activity of mixotrophic organisms is predicted to be influenced by changes associated to climate change (Wilken et al., 2019). Discerning the response of the different mixotrophic species to an increase of temperature and stratification will be essential to predict ecosystem changes in future scenarios.

1.4 How do we study microbial primary producers in the ocean?

Back in the 1980s, the introduction of flow cytometers in microbiological research lead the first steps towards the automation of phytoplankton cell counting (Olson et al. 1983; Li 1989), pigment content evaluation (Olson et al., 1988) and biomass quantification (Li, Irwin, Dickie, 1993). Later on, high-performance liquid chromatography (Mackey et al., 1996) and the combination of flow cytometry with molecular probes (Simon et al., 1995; Not 2002 et al., 2002) allowed finer species identification based on pigments content. During the last decades, the improvement of DNA sequencing techniques has opened a new level of taxonomic resolution in the evaluation of microbial communities, unveiling a high diversity of protists (Massana et al., 2014; Liu et al. 2009; de Vargas et al. 2015).

DNA sequencing allows the identification of species present in environmental samples, as well as the characterisation of the genomic and functional traits of single species or whole microbial communities. In this thesis, we have used a combination of metabarcoding (metaB), metagenomic (metaG) and single-cell sequencing approaches to study phototrophic and mixotrophic microbes. MetaB technique consists on sequencing DNA regions with low intra-specific and high inter-specific variability for the taxonomic profiling of environmental samples. Typically, the DNA region (or "marker gene") chosen to identify bacterial species is 16S rDNA, while the barcoding of eukaryotic genome and plastids is based on 18S rDNA and 16S rDNA gene sequencing, respectively. On the other hand, metaG approach consists of sequencing genome fragments from all organisms in a sample. Once assembled, metaG reads allow us to characterise the functional potential of species present in a given community. Finally, single-cell sequencing technique involves isolating single cells from microbial samples and sequencing their genome or transcriptome individually. The obtention of single-cell amplified genomes (SAGs) or transcriptomes from environmental samples offers tremendous advantages for two main reasons: (i) it allows the obtention of genomes/transcriptomes from uncultivable taxa; and (ii) facilitates the study of genome intra-specific variability.

1.5 Assessing of the global map of marine primary producers with the data from *Tara* Oceans expedition

The exploration of the plankton systems in the global ocean started with the expedition Challenger (1872-1876). During the last decade, the Global Ocean Survey (2003-2010) and Malaspina-2010 campaigns surveyed the ocean with new molecular techniques, coinciding with the beginning of *Tara* Oceans expedition (2009-2013) (Karsenti, 2012). *Tara* Oceans (Figure 1.3) attempts to understand the spatio-temporal dynamics and evolutionary principles of plankton ecosystems at a planetary scale. For this purpose, the *Tara* Oceans oceanographic campaign generated eco-morpho-genetic data from 152 stations and 3 depths covering most of the oceanic regions. The golden repository of molecular data from *Tara* Oceans allows us to study planktonic communities in the global oceans in a standardised way.

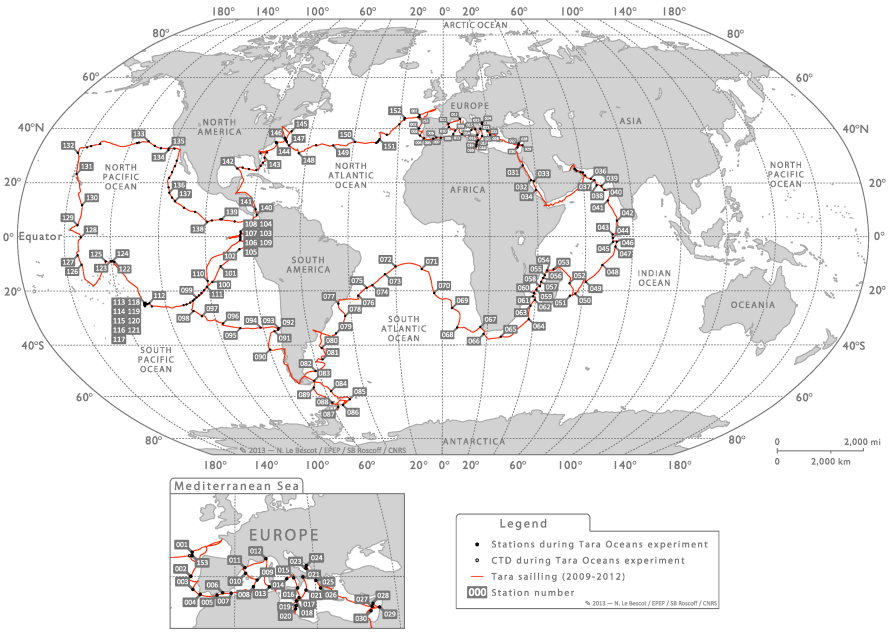


Figure 1.3 – Geographic location of the 152 scientific stations sampled during *Tara* Oceans expedition 2009-2012. Chart edited by Noan Le Bescot - *Tara* Oceans Foundation.

1.6 Objectives and content of this thesis

The aim of this thesis is to explore patterns of functional and phylogenetic diversity of microbial primary producers in the global ocean. The major accomplishments from this project include (i) the **characterisation of the global relative abundance and phylogenetic diversity of picophototrophs** in marine ecosystems;

(ii) the **assessment of trophic diversity in piconano- and nano- size fraction communities**, and (iii) the **evaluation of the potential and limitations of metabarcoding, metagenomic and single-cell sequencing data** for the study of microeukaryotic primary producers. The thesis is divided in the following chapters:

Chapter 2: describes the PhotoRefT 16S rDNA reference tree and a phylogeny-based pipeline for the assessment of picophototrophs richness and abundance.

Chapter 3: includes the evaluation of picophotorophs global distribution and evolutionary diversity through a combination of metagenomic and metabarcoding data.

Chapter 4: contains the description of a taxonomy-free model for the assessment of trophic diversity in metagenomic samples, together with the characterisation of photo-, phagomixo-, phago- and osmo-trophic groups distribution across *Tara* Oceans stations.

Chapter 5: includes the assessment of 903 single-cell sequenced genomes distribution across *Tara* Oceans stations to evaluate the potential of single-cell sequencing for obtaining the genomes of the most cosmopolitan and abundant protists in marine ecosystems.

Chapter 6: includes a discussion about future prospects of the study of marine primary producers in the context of the overall results described.

2. PhotoRefT: a general framework to assess picophototrophic communities through phylogenetic placement

The aim of this chapter is to introduce the PhotoRefT reference tree and its associated pipeline for the assessment of picophototrophic communities. PhotoRefT offers the foundation to study the abundance and diversity of phototrophs through phylogenetic placement out of 16S Ribosomal DNA (rDNA) metagenomic (metaG) and metabarcoding (metaB) data.

In this chapter, we use *Tara* Oceans 16S rDNA metaB and metaG data for the evaluation of the phylogenetic placement pipeline, and the description of the picophototrophic communities based on the two datasets is depicted in Chapter 3. Therefore, we refer the reader to Chapter 3 for details about the sequencing protocol and composition of 16S rDNA *Tara* Oceans metaB and metaG datasets. Here we describe the composition of PhotoRefT and we show it can work as a phylogenetic taxonomic annotation tool, and a promising alternative to annotation methods that depend on similarity thresholds -less adequate for marker genes like 16S rDNA which evolve at different rates among domains and lineages of life-.

2.1 INTRODUCTION

Phototrophs measuring $0.2-3\mu$m (namely picophototrophs) are major contributors to the primary production in the oligotrophic ocean (Li 1994; Pérez et al., 2006; Jardillier et al., 2010). After decades of studying these communities, which include Oxyphotobacteria and phototrophic picoeukaryotes (PPEs), we still lack a comprehensive picture of their taxonomic composition and phylogeographic distribution.

Past studies have evaluated picophototrophs through 16S rDNA (Shi et al., 2011; Kirkham et al., 2013), a marker gene that can identify at once both bacterial and eukaryotic phototrophs through their cellular and plastidial genomes, respectively. However, the lack of plastidial reference sequences has limited the potential of 16S rDNA in phototrophs diversity surveys. This situation changed recently, when Decelle et al. (2015) published PhytoREF, a database of plastidial 16S rDNA sequences. PhytoREF covers all major lineages of photosynthetic eukaryotes, including terrestrial plants, and provides a valuable resource for the annotation of environmental 16S rDNA metabarcodes obtained by high-throughput sequencing data.

In this study, we assembled reference sequences from eukaryotic plastids (including those published in PhytoREF), Oxyphotobacetria, and heterotrophic prokaryotes (as outgroups) to build a reference 16S rDNA phylogenetic tree for assessing phototrophic life. This reference tree, named PhotoRefT, covers the diversity of the main known picophototrophic lineages. The tree was further used to optimise a bioinformatics pipeline based on phylogenetic placement (Barbera et al., 2018) for the annotation of picophototrophs' metaB and metaG sequences. We show that by using this new phylogenetic framework we are able to identify a higher number of PPE environmental sequences than through the Basic Local Alignment Search Tool (BLAST) (i.e. a method based on sequences similarity). PhotoRefT provides a new tool for the characterisation of picophototrophs diversity which is designed to further evolve to represent additional phototrophic clades.

2.2 METHODS

2.2.1 Construction of PhotoRefT

PhotoRefT was built using sequences from the following databases: Genbank (Benson et al., 2017), RefSeq (O'Leary et al., 2016) and ENA (Harrison et al., 2019) for heterotrophic bacteria and archaea; PhytoREF (Decelle et al., 2015) for eukaryotic plastids, and SILVA SSURef_NR99_128 (Quast et al., 2013) for cyanobacteria.

From the PhytoREF database, we filtered out most of the sequences classified as Streptophyta, keeping only 6 representative sequences from that clade, and retaining all plastidial sequences from the other groups. From SILVA SSURef_NR99_128, we only selected cyanobacterial sequences. The taxonomy of Oxyphotobacteria in SILVA v128 was obtained using the Candidate Taxonomic Unit approach (Yarza et al. 2014) for establishing genus boundaries and its nomenclature relied on the Genome Taxonomy Database taxonomy (Parks et al. 2018). The annotation of the SILVA v128 cyanobacteria sequences was corrected with *SATIVA* (Kozlov et al., 2016), and sequences with incomplete classification were discarded. Additionally, we appended the sequences from 9 ecologically important *Prochlocococcus* and *Synechococcus* strains from an in-house database (labeled as PRJNA) that were not included in SILVA v128. To include representatives from the main groups of archaea and heterotrophic bacteria, we downloaded the 16S rDNA sequences from 85 bacteria and 10 archaea from Genbank, RefSeq and ENA. Finally, the IDs from the total

of 8,167 reference 16S rDNA sequences were standardised to include five taxonomic ranks.

In order to decrease the complexity of the PhotoRefT, these >8,000 reference sequences were clustered at 99% identity using *USEARCH* (sequences classified as Melainabacteria and Sericytochromatia, non-oxyphotogenic cyanobacteria, were clustered at 85% identity to keep only some representatives from these groups). The resulting 4,181 sequences were aligned with *MUSCLE* v3.8.31 (Edgar, 2004). We selected the V2-V9 region from the alignment in order to build the PhotoRefT with as much information as possible, and the references with incomplete V2-V9 fragment or looking clearly misaligned were filtered out. Exceptionally, reference sequences classified as Dinophyceae, Dictyochophyceae, Chrysophyceae, Cercozoa, Haptophyta, Chrysophyceae and Pelagophyceae were kept as long as they covered the V4-V5 region, because these are important groups of PPEs and their representation in PhytoREF is already limited. We realigned the output sequences using *MAFFT* v7.299b (Katoh & Standley, 2013) with the options -ep 0.2 and -op 2.5. Misaligned sequences were filtered out, leading to a final set of 3,042 reference sequences. Next, we trimmed non-informative positions from this reference alignment using *trimAL* v1.4.rev15 (Capella-Gutiérrez, Silla-Martínez & Gabaldón, 2009) with the options -gt 0.3 -st 0.001 -cons 70.

Finally, we build a Newick tree forcing well-known taxonomic paths of eukaryotes and Oxyphotobacteria to constrain the topology of our final inferred reference tree. In particular, we constrained the topology of Oxyphotobacteria using the five levels of their taxonomic annotation, while eukaryotic sequences were only clustered at domain level (letting sequences from the same plastidial lineage arrange accordingly). We inferred the final PhotoRefT using RAxML v8.2.9 (Stamatakis, 2014) by conducting 100 rapid bootstrap inferences and 20 independent maximum likelihood tree searches, using the GTRGAMMA model. Archaea reference sequences were used for arbitrary rooting. The obtained tree was visualised with *Archaeopteryx* (Han 2009). In order to remove redundant branches, we reduced the size of the tree to 802 leaves using the Average Distance to the Closest Leaf (ADCL) minimisation algorithm (Matsen, Gallagher & McCoy, 2013). The ADCL algorithm reduces the size of a tree by picking up the sequences that represent the local diversity (i.e. tries to find the central branches). The final version of the PhotoRefT and its alignment, together with the list of reference sequences' accession number and original classification, can be found online at https://github.com/lrubinat/PhotoRefT .

2.2.2 Meta-omics data for the evaluation of PhotoRefT

We used 16S rDNA metaG reads an V4-V5 amplicons from *Tara* Oceans to evaluate the performance of our pipeline. In total, we screened the data from 42 *Tara* Oceans stations covering a large extension of the temperate and oligotrophic global ocean. Details on the sequencing protocol and post-sequencing steps followed for the obtention of these datasets are specified in Chapter 3 (Section 3.2.1). The metaB dataset, rarefied at 201,404 reads per sample, contained a total of 8,257,564 reads

and 68,293 OTUs. MetaG data was rarefied at 40,590 reads per sample and contained 1,704,780 reads.

2.2.3 Framework for phylogenetic placement of environmental 16S rDNA sequences

The PhotoRefT tree only covers the diversity of heterotrophic bacteria superficially. In order to minimise possible misplacements of heterotrophs' 16S rDNA sequences into phototrophic branches of the PhotoRefT, we pre-filtered the metaB and metaG datasets. Sequences matching heterotrophic bacteria or archaea references in SILVA v123 (with identity cutoff $\geq$ 99% in BLAST search) were discarded. Following the pipeline described by Czech & Stamatakis (2019), we deduplicated the rest of the query sequences using the *GAPPA* command *chunkify,* and aligned to the PhotoRefT alignment with *PaPaRa* (Berger & Stamatakis, 2012). In the case of metaG, because sequences can cover any region of the PhotoRefT alignment, we selected the aligned reads that overlapped to the V4-V5 region with at least 125nt. Finally, we placed the aligned queries on the PhotoRefT with *EPA-ng* (Barbera et al., 2018) and subsequently reduplicated into per-sample result files using the *GAPPA* command *unchunkify*.

We observed long branch attraction in the placement that can lead to long pendant lengths (i.e. long lengths of the edges connecting queries with the reference tree). In order to discard possible misplacements due to this effect, we filtered the placement results with a custom program built with the *GENESIS* library (Czech, Barbera & Stamatakis, 2019) available at
https://github.com/Pbdas/genesis-apps/blob/master/jplace-filter.cpp. In this filtering step, we compared the pendant length of each placement with the average branch length of its 10 neighbour edges (or local average). The placements were cut away when the pendant length was four times greater than the local average.

We used the function *edit accumulate* from the *GAPPA program* to gather queries into the basal branches where they accumulated 0.95 of their likelihood weight mass. We visualized the results with the *GAPPA* subcommand *examine heat-tree*. Finally, we obtained the diversity and abundance values of our phototrophic groups of interest (see Table 2.1 for complete list of groups considered) with the *GAPPA* subcommand *prepare extract*. In this step, we selected those placements that observed an accumulated likelihood weight ratio (or confidence value) higher than 0.95 in any phototrophic clade. Queries placed in Apicomplexa, Streptophyta, archaea or heterotrophic bacteria were discarded.

2.2.4 Taxonomic assignment of metaB sequences through sequence similarity

In order to compare the placement results with the traditional annotation approach -based on sequences similarity-, we annotated the metaB dataset through BLAST and using PhytoREF as reference database. PhytoREF was clustered at 99% identity and metaB reads were blasted against it using an identity threshold of 97% and filtering by >80% coverage. Reads classified as Streptophyta were discarded.

Domain	Phylum	Class
Bacteria	Oxyphotobacteria	Limnotrichales
Bacteria	Oxyphotobacteria	Nostocales
Bacteria	Oxyphotobacteria	Pseudanabaenales
Bacteria	Oxyphotobacteria	Synechococcales
Eukaryota	Cercozoa	Chlorarachniophyceae
Eukaryota	Chlorophyta	Chlorodendrophyceae
Eukaryota	Chlorophyta	Chlorophyceae
Eukaryota	Chlorophyta	Mamiellophyceae
Eukaryota	Chlorophyta	Nephroselmidophyceae
Eukaryota	Chlorophyta	Pedinophyceae
Eukaryota	Chlorophyta	Other prasinophytes
Eukaryota	Chlorophyta	Chloropicophyceae
Eukaryota	Chlorophyta	Trebouxiophyceae
Eukaryota	Chlorophyta	Ulvophyceae
Eukaryota	Cryptophyta	Cryptophyceae
Eukaryota	Euglenozoa	Euglenophyceae
Eukaryota	Glaucophyta	Glaucocystophyceae
Eukaryota	Haptophyta	Pavlovophyceae
Eukaryota	Haptophyta	Prymnesiophyceae
Eukaryota	Haptophyta	Rappephyceae
Eukaryota	Ochrophyta	Bacillariophyceae
Eukaryota	Ochrophyta	Chrysophyceae
Eukaryota	Ochrophyta	Dictyochophyceae
Eukaryota	Ochrophyta	Eustigmatophyceae
Eukaryota	Ochrophyta	Pelagophyceae
Eukaryota	Ochrophyta	Phaeophyceae
Eukaryota	Ochrophyta	Raphidophyceae
Eukaryota	Rhodophyta	Bangiophyceae
Eukaryota	Rhodophyta	Florideophyceae
Eukaryota	Rhodophyta	Porphyridiophyceae
Eukaryota	Rhodophyta	Rhodellophyceae

Table 2.1 – List of phototrophic groups analysed classified by domain, phylum and class.

2.3 RESULTS

2.3.1 PhotoRefT composition

PhotoRefT contained a total of 802 reference sequences covering the main groups of picophototrophs (Figure 2.1). With 482 reference sequences (60% of the total), Oxyphotobacteria was the group with larger number of sequences. Eukaryotic plastids accounted for 239 reference sequences (∼30% of PhotoRefT), while archaea and heterotrophic bacteria included 83 reference sequences (∼10% of the total).

The branch length in PPE groups was visibly longer than in Oxyphotobacteria clades, illustrating the wide evolutionary radiation among plastids. The reference sequences from Chlorophyta, Cercozoa, Euglenida and the Dinoflagellata genus Lepidodinium clustered together naturally when inferring the tree, reflecting the shared evolutionary origin of their plastids. In a similar way, the plastid derived from the red algae lineage (Haptophyta, Ochrophyta and the rest of Dinoflagellates genera [Karlodinium, Gyrodinium and Karenia]) also grouped together despite the distinct evolutionary history encoded in their nuclear genomes.

2.3.2 Long branch attraction effect

The placement of metaB and metaG reads on PhotoRefT revealed comparable results of relative abundance in most of the groups, except for Dinoflagellata and Euglenozoa (Figure 2.2, upper treemaps). While Dinoflagellata accounted for $< 1\%$ in metaB, they contributed to ∼20% of the reads in metaG. On the other hand, Euglenozoa represented 0.07% of the reads in metaB and up to 2.6% in metaG.

The differences in Dinoflagellata and Euglenozoa between the two datasets was not only concerning their abundance but also the pendant length of their placements: reads classified as Dinoflagellata and Euglenozoa branched with longer pendant branches in metaG than in metaB. Most of the metaG reads classified as Dinoflagellata were placed in the inner branch of PhotoRefT that connects the group Lepidodinium with Chlorophyta; that branch is by far the longest node in the tree. In order to avoid possible misplacements in metaG caused by long branch attraction, we filtered placements by pendant length. As described in Section 1.2.3, the filtering step consisted of discarding those placements whose pendant length was four times greater than the local average branch length. After applying this filter, the relative abundance of Dinoflagellata and Euglenozoa in metaG decreased by more than 15-fold, while it barely affected their frequency in metaB (Figure 2.2, lower treemaps).

2.3.3 Comparison between annotation approaches

After feeding the classification algorithms with the collection of $2,544,771$ metaB sequences, we obtained $577,675$ reads classified as PPE through phylogenetic placement and $555,215$ reads through BLAST. Regardless that for most of the groups we

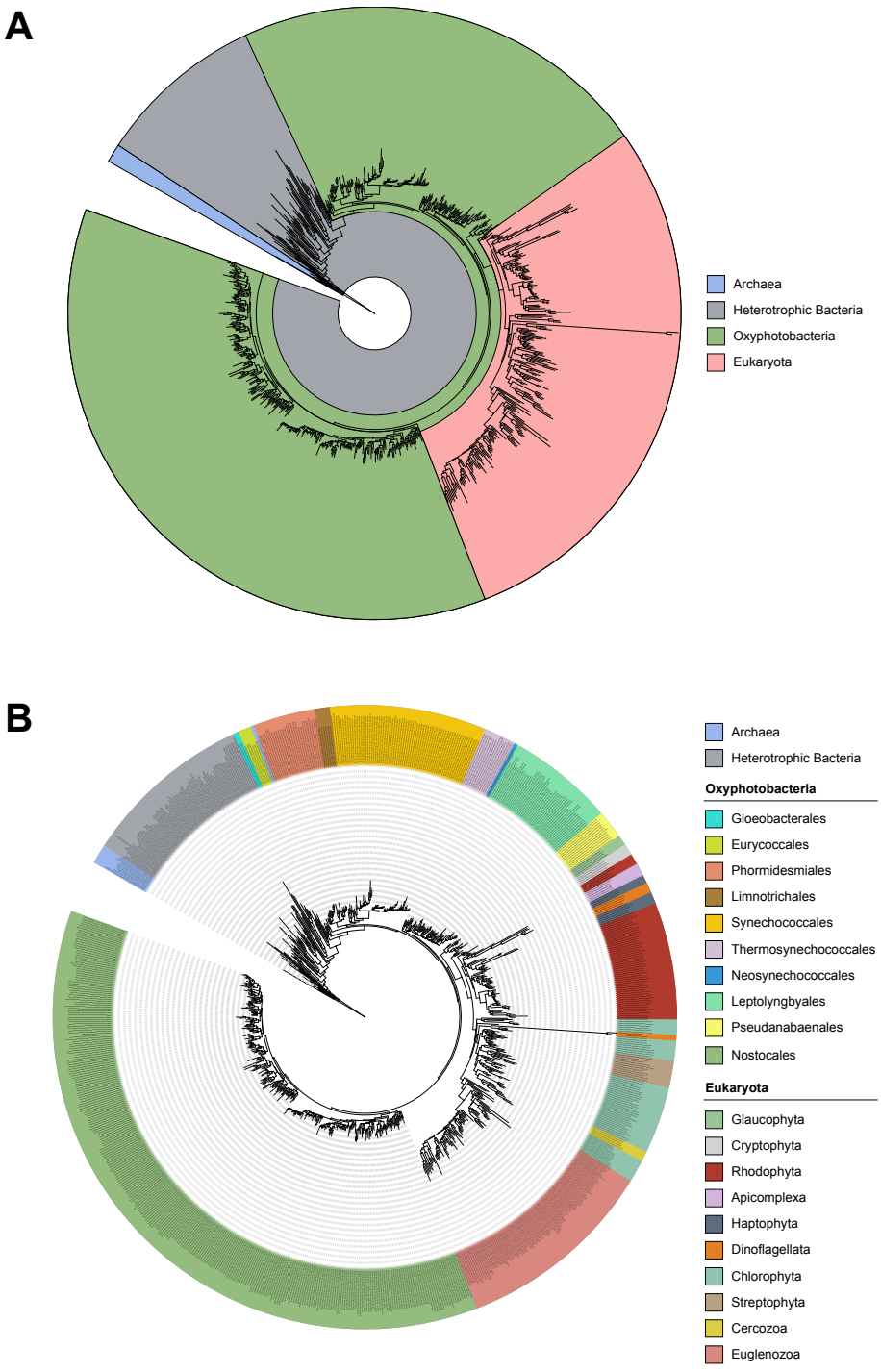


Figure 2.1 – Composition of PhotoRefT 16S rDNA sequences. (**A**) Distribution of PhotoRefT clades coloured by supergroups, include the following categories: archaea, heterotrophic bacteria, Oxyphotobacteria and Eukaryota. (**B**) Distribution of PhotoRefT clades coloured at group level (in the case of archaea and heterotrophic bacteria) and at phylum level (in the case of Oxyphotobacteria and Eukaryota). Oxyphotobacteria is represented by the following lineages: Eurycoccales, Geitlerinematales, Gloeobacterales, Leptolyngbyales, Limnotrichales, Neosynechococcales, Nostocales, Phormidesmiales, Pseudanabaenales, Synechococcales and Thermosynechococcales. Eukaryota include the following phyla: Apicomplexa, Cercozoa, Chlorophyta, Cryptophyta, Dinoflagellata, Euglenozoa, Glaucophyta, Haptophyta, Rhodophyta and Streptophyta.

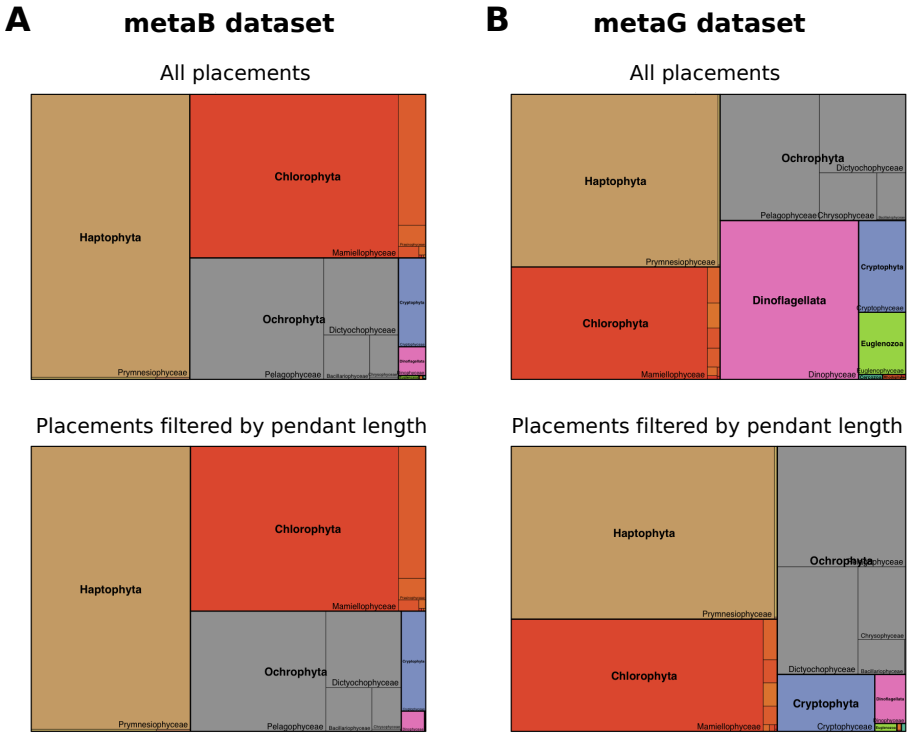


Figure 2.2 – Relative abundance of PPE groups according to (**A**) metaB data and (**B**) metaG data. The two upper treemaps show the contribution of each group before applying any filtering step. The two lower treemaps show the relative abundances after excluding reads placed with pendant branches 4 times longer than their local average.

recovered a higher amount of reads through placement, the composition of PPE community in terms of groups' relative abundance showed comparable results between the two annotation approaches (Figure 2.3). Haptophytes, Chlorophytes and Ochrophytes dominated PPE community according to both techniques. The only groups contributing to more than 1% of the relative abundance and showing significant differences between methods were Dictyochophyceae and Prasinophyceae Clade VII (or Chloropicophyceae). The relative abundance of Dictiochophyceae was two-fold higher in the assessment based on phylogenetic placement than in BLAST results. On the contrary, Prasinophyceae Clade VII accounted for 3% of the relative abundance according to placement and up to 7% according to BLAST. Pedinophyceae, Pavlovophyceae and Chrysophyceae accounted for less than 1% of the relative abundance according to BLAST approach while presented frequencies from 2 to 3 orders of magnitude higher through placement.

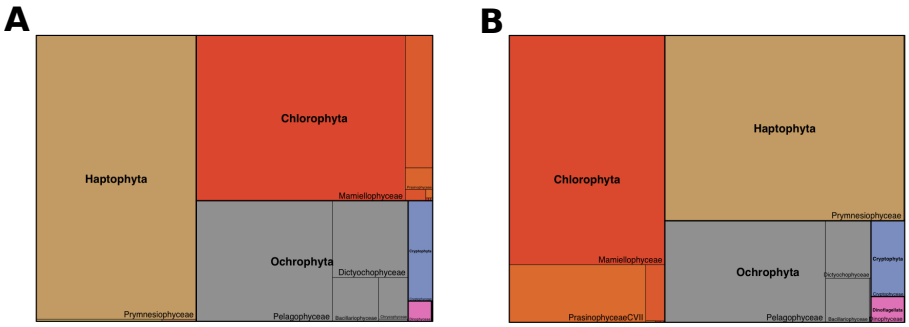


Figure 2.3 – Estimates of PPE relative abundance in metaB dataset obtained through (**A**) phylogenetic placement using PhotoRefT as reference tree and (**B**) BLAST screening against PhytoREF database.

2.4 DISCUSSION

PhotoRefT is, to our knowledge, the first reference phylogenetic tree covering the wide diversity of 16S rDNA from all known groups of photosynthetic organisms. Most of the reference sequences in the tree correspond to Oxyphotobacteria, as a result of the extensive representation of this group in reference databases. The higher phylogenetic diversity and longer branches observed in PPE clades, in comparison with the low radiation within Oxyphotobacteria, suggests that there is still a significant amount of diversity to uncover in PPE. The publication of PhytoREF database (Decelle et al., 2015) was an important contribution towards the identification of 16S rDNA environmental sequences, but a much broader diversity of reference plastidial sequences from eukaryotic microalgae need to be implemented into PhytoREF in the future. Dinoflagellata acquired plastids through the endosymbiosis of organelles from diatoms, haptophytes, cryptomonads and green algae (Keeling 2010). In PhotoRefT we have only included representatives from two of the complex evolutionary origins of Dinoflagellata plastids: the group of plastids acquired from haptophytes (with representative sequences from *Karenia*, *Karlodinium* and *Gyrodinium*) and

the plastids from the green algal lineage (with references from *Lepidodinium*). Until reference sequences representative from the other evolutionary origins of Dinoflagellata are not available, the evaluation of environmental sequences from this group through PhotoRefT will be incomplete.

Unlike metaB reads, metaG sequences seemed to be susceptible to long branch attraction in placement. Most likely, this bias was only due to the short length of metaG reads and could be corrected by excluding placements with long pendant length. The annotation of PPEs showed highly similar results between placement and BLAST approaches, with the main difference that phylogenetic placement allowed to annotate more than $20,000$ additional sequences.

Overall, we show that PhotoRefT provides a reliable and unifying support to evaluate the structure of phototrophic communities across any environment using metaomics data. This reference tree is meant to evolve in the future and include further 16S rDNA sequences from all phototrophic lineages.

3. Diversity and structure of photosynthetic picoplankton in the world surface ocean: a unifying, cross-domain perspective

Oceanic phytoplankton are composed of photosynthetic cyanobacteria (Oxyphotobacteria) and eukaryotes, and generates nearly half of planetary primary production. Despite their critical importance for the Earth system, ocean phytoplankton are still relatively poorly characterised in terms of taxonomic abundance, as compared to terrestrial plants. Traditionally, microbial phototrophs have been measured through flow-cytometry and chromatography analyses of pigment extracts on one hand, with very low taxonomic resolution, and PCR-based rDNA clone-libraries or metabarcoding one the other hand, with poor quantitative resolution. Here we re-assess the biodiversity of open ocean pico-phytoplankton using a phylogenetic placement approach of PCR-free 16S rDNA metagenomics data from two circum-global expeditions (*Tara* Oceans and Malaspina) onto a reference tree containing representative rDNA sequences from all known prokaryotic and eukaryotic phototrophs. This approach minimises the quantitative biases associated to amplicon-based metabarcoding, it uses a single marker gene to assess the diversity of both photosynthetic prokaryotes and eukaryotes, and allows taxonomic assignation across the entire phylogenetic diversity of photosynthetic life. Our results show that assessment of total phototrophic picoplankton through 16S rDNA extracted from metagenomes is highly congruent with measures based on 16S rDNA metabarcoding, as well as 18S rDNA recruitment from metagenomes. Oxyphotobacteria were three-fold more abundant than phototrophic pico-eukaryotes (PPEs) in surface waters of (sub)tropical to temperate oceans. However, PPEs were four-fold richer than Oxyphotobacteria, and

essentially composed of micro-algae derived from secondary or higher-level of plastid endosymbiosis, mostly Prymnesiophyceae, Pelagophyceae and Dictyochophyceae. PPEs also displayed a significantly higher evolutionary diversity than Oxyphotobacteria, presenting variances across groups and with Chlorophyta emerging as the eukaryotic microalgae with the highest plastidial phylogenetic diversity. Overall, our results support the use of metagenomic 16S rDNA phylogenetic placement as a unifying, trans-domain method to assess total phototrophic communities. Finally, the relatively good agreement between state-of-the-art meta-omics, old-style molecular ecology, and traditional organismal/biochemical approaches to assess picophytoplankton indicate that the knowledge of the biodiversity of the organisms that make up the bulk of primary production in the oceans is on its way to completion.

This Chapter corresponds to the draft manuscript submitted to *PeerJ* and is co-authored by my supervisors Colomban de Vargas and Ramiro Logares; our colleagues from the Heidelberg Institute for Theoretical Studies, Pierre Barbera and Lucas Czech; my group colleagues Nicolas Henry, Ewen Corre and Cédric Berney; Pelin Yilmaz, from the Max Planck Institute for Marine Microbiology, and members from Tara Oceans and Malaspina consortia Julie Poulain, Susana Agustí and Carlos M. Duarte. The body of the chapter is followed by the figures and tables corresponding to the Supplementary Material.

3.1 Introduction

Oceans compose the largest continuous ecosystem on Earth and are populated by an astronomical number of floating and drifting microbes. In the sunlit ocean, phytoplankton -all bacterial and eukaryotic microbes performing oxygenic photosynthesis- is generating nearly half of planetary primary production (Field et al. 1998). By doing so, phytoplankton sustains the whole ocean ecosystem, it also influences the concentration of CO_2 in the atmosphere (Westberry et al., 2008), and contributes to the export of organic matter into the deep ocean (biological carbon pump, Passow & Carlson, 2012). Understanding the compositional and functional dynamics of phytoplankton communities is critical to better predict the evolution of ocean eco- and geochemical-systems, especially in a context of climate change, when these mechanisms could be seriously perturbed (Ciais et al., 2013; Thornton, 2014; Ibarbalz et al., 2019).

Ocean primary production has traditionally been attributed mostly to Oxyphotobacteria, the cyanobacteria that perform oxygenic photosynthesis and whose origin goes back to around 2.7 billion years ago (Gould, Waller & McFadden, 2008). In the last two decades however, a few studies have reported eukaryotic plankton to be capable of fixing significantly higher amounts of carbon than Oxyphotobacteria despite being less abundant (Worden, Nolan & Palenik, 2004; Jardillier et al., 2010; Rii et al., 2016). Eukaryotic phytoplankton originated from the primary endosymbiosis between a heterotrophic protist and a cyanobacterium in the Proterozoic (Reyes-Prieto, Weber & Bhattacharya, 2007). The first major group of known photosynthetic eukaryotes, Chlorophyta, is thought to have diversified later in the

late Proterozoic / early Palaeozoic (Leliaert, Verbruggen & Zechman, 2011), and thrived in marine waters until the Permo-Triassic mass extinction event. In the early Mesozoic, numerous eukaryotic lineages with secondary plastids (e.g. haptophytes, dinoflagellates and ochrophytes), diversified and replaced chlorophytes as the dominant eukaryotic phytoplankton. In today' oceans, the phototrophic eukaryotes measuring <3µm (namely phototrophic picoeukaryotes or PPEs) have been found to be responsible for up to 50% of picoplanktonic biomass and more than 40% of CO_2 fixation in some oligotrophic waters (Not et al., 2008; Jardillier et al., 2010). Despite the obvious ecological importance of these communities, we still lack an accurate comparison on the overall abundance, richness and phylogenetic diversity of PPEs *vs.* Oxyphotobacteria in modern world oceans.

Estimates of abundance and/or diversity of Oxyphotobacteria and PPEs in marine waters have been performed primarily using flow cytometry and HPLC analyses of pigment extracts (i.e. Li, 1989; Mackey et al., 1996; Buitenhuis et al., 2012). While flow cytometry studies have demonstrated that Oxyphotobacteria account for ∼30% of the picophototrophs biomass in the global ocean (Buitenhuis et al., 2012), pigment data has shown the prominence of haptophytes, diatoms and chlorophytes amongst PPE lineages in marine waters (Latasa M & Bidigare, 1998; Not et al., 2005; Liu et al. 2009). Over the last three decades, the poor taxonomic resolution power of these methods was complemented by DNA metabarcoding survey of various marker genes, such as 16S rDNA (e.g. Huang et al., 2012) and petB (e.g. Farrant et al., 2016) for Oxyphotobacteria, or 18S rDNA (e.g. de Vargas et al. 2015) and plastidial 16S rDNA (e.g. Kirkham et al., 2013) for PPEs. However, these different molecular barcodes are hard to compare, and their amplification from environmental DNA extracts by PCR may lead to serious quantification biases, mainly for the 18S rDNA genes whose copy numbers vary extensively amongst eukaryotes (see supplementary figure W3 in de Vargas et al., 2015) and can display considerable genetic variation even in a single species (e.g. Pillet, Fontaine & Pawlowski, 2012). In addition, some of the PPEs, especially Dinophyceae and Chrysophyceae, contain a significant proportion of taxa that have lost photosynthesis and became heterotroph; studies based on nuclear 18S rDNA marker gene cannot distinguish them from their phototrophic siblings.

In Chapter 3, we use a single marker gene - the V4V5 fragment of the 16S rDNA- to assess and compare the entire bacterial and eukaryotic biodiversity of picophototrophs in the world sunlit tropical and temperate ocean. Oxyphotobacteria of the order of Synechococcales tend to have between one and four copies of 16S rDNA per genome (Engene et al. 2011), thus allowing theoretically unbiased quantification in this group. In eukaryotic chloroplasts, the number of plastidial genomes can vary from 50 to 100 copies per organelle (Decelle et al., 2015), and we examined how this could affect the quantification of PPE groups by comparing their abundance with complementary analyses of 18S rDNA marker gene. In addition to using metabarcoding datasets, we quantified the abundance of 16S rDNA reads directly from metagenomic libraries generated from picoplankton total DNA extracts. These are not subject to the exponential PCR amplification process used to generate amplicon

data and thus should provide more accurate quantification of abundance. Finally, in order to compare across domains and between sequencing methods (metabarcoding and metagenomics), we developed a unifying phylogenetic placement strategy allowing to quantify the number and diversity of DNA reads along a tree with representative sequences of all known picophototrophic life (except for dinoflagellates). Using this new framework, we reassessed the global relative abundance and diversity of PPEs and Oxyphotobacteria in 52 samples from the *Tara* Oceans (Karsenti, 2012)and Malaspina (Duarte, 2015)expeditions, providing and new global view of picoplanktonic phototrophic life in modern oceans.

3.2 Methods

Our goal was to reassess the diversity and distributions of marine pico-phytoplankton using a plastidial rDNA marker gene obtained via different DNA sequencing protocols and from two independent circumglobal oceanographic expeditions. We analysed sequence data collected from marine waters during the *Tara* Oceans and Malaspina-2010 expeditions. All samples analysed were collected from surface waters (3 to 5 meters) and corresponded to the pico-plankton size-fraction (0.2 - 3 µm). In the *Tara* Oceans dataset we selected the 42 stations available for combined metagenomics (metaG) and metabarcoding (metaB) data. The Malaspina-2010 dataset included 10 matching stations for metaG and metaB data. All together, our data covers large geographic areas of the tropical and temperate oceans in the Pacific, Atlantic, and to a smaller extent Indian basins (Figure S3.1). A detailed description of *Tara* Oceans and Malaspina-2010 sampling stations and biological material processing procedures can be found in Pesant et al., (2015)and Estrada et al. (2016), respectively.

3.2.1 DNA sequencing data

The abundance and phylogenetic diversity of Oxyphotobacteria and Photosynthetic Pico-Eukaryotes (PPEs) was compared by analysing 16S rDNA metabarcoding (metaB) and metagenomic (metaG) datasets generated from the same *Tara* Oceans and Malaspina-2010 samples (Section 3.2.2). The datasets from the two expeditions were analysed independently and used as replicates for 16S metaB *vs.* metaG comparison. Additionally, we used 18S rDNA metaG reads from *Tara* Oceans samples to cross-check the estimates obtained for the different groups of PPEs using both plastidial and nuclear marker genes (Section 3.2.4). Figure S3.2 describes the datasets used.

3.2.1.1 *Tara* Oceans metabarcoding (metaB) datasets

The *Tara* Oceans dataset included 16S V4-V5 rDNA metaB sequences from 41 picoplankton samples available at ENA repository XXXX. The primers used for amplifying 16S rDNA V4-V5 region were 515yF 5'-GTGYCAGCMGCCGCGGTAA-3' and 926R 5'-CCGYCAATTYMTTTRAGTTT-3' (Parada, Needham & Fuhrman,

2016). DNA amplification was followed by Illumina sequencing with the Hiseq2500 system. After cleaning and filtering the metaB as described in Mahé et al.(2017), the reads were clustered into operational taxonomic units (OTUs) using the *Swarm v2* clustering software (Mahé et al., 2015), resulting in 87,634 OTUs and 14,145,403 reads, each OTU containing at least 2 reads from 2 different samples. The resulting OTU table comprising 41 *Tara* Oceans samples represented an average of $\sim 345,010$ reads per sample and $\sim 7,408$ OTUs per sample. We rarefied the dataset to the lowest number of reads per sample (i.e. 201,404 reads) using the *rrarefy* function from the R package *vegan* (Oksanen et al., 2018). The rarified table had a total of 8,257,564 reads, 68,293 OTUs and an average of 5,652 OTUs per sample (Figure S3.2). We refer to the rarefied table as TO-16S-metaB (for *Tara* Oceans 16S rDNA metaB).

3.2.1.2 Malaspina-2010 metabarcoding datasets

For the Malaspina metaB dataset, we used the 16S rDNA data published in Logares et al., (2018). The 16S rDNA V4-V5 sequences of Malaspina samples were amplified by PCR using the same primers as in *Tara* Oceans metaB (515yF and 926R), and sequenced on an Illumina MiSeq sequencing platform. Amplicons were processed using the pipeline available at https://github.com/ramalok/amplicon_processing. In this pipeline, the obtained paired-end reads were merged using *PEAR* (Zhang et al., 2014), and only those assembled sequences measuring over 100 bp were selected. Next, *USEARCH* (Edgar, 2010) was used for read quality checking, dereplication, OTU clustering (UPARSE algorithm, similarity 99%) and chimera filtering using SILVA v.119 (Quast et al., 2013) as the reference database. The OTU table comprising the 10 Malaspina samples included 429,116 reads and 5,936 OTUs, with an average of $\sim 42,912$ reads per sample and $\sim 1,765$ OTUs per sample. After rarefying (same procedure as with TO-16S-metaB) to 5,557 reads per sample, the obtained dataset had 55,570 reads, 3,370 OTUs and an average of 750 OTUs per sample. The Malaspina rarefied OTU table is referred as MSP-16S-metaB (for Malaspina 16S rDNA metaB).

3.2.1.3 Metagenomic sequence datasets

For community analyses based on rDNA from metagenomic reads, we used datasets from 42 *Tara* Oceans samples (Sunagawa et al., 2015) and 10 Malaspina-2010 picoplankton samples. MiTag sequences are publicly available at ENA under the project YYYYY. In both cases, the environmental samples were sequenced with Illumina HiSeq2000. The total metaG reads were filtered with HMM profiles to extract the sequences belonging to 16S rDNA (in both *Tara* Oceans and Malaspina-2010 samples) and 18S rDNA (in *Tara* Oceans samples) and analysed as described in Logares et al. (2013).

We subsampled the 16S rDNA datasets from 42 *Tara* Oceans samples and 10 Malaspina samples to the lowest number of reads per sample in each dataset using *VSEARCH v2.4.3 (*Rognes et al., 2016). The subsampling value was 40,590 reads per sample for the *Tara* Oceans dataset and 14,808 reads per sample for the

Malaspina dataset. Hereafter, the 16S rDNA metaG datasets from *Tara* Oceans and Malaspina are referred as TO-16S-metaG and MSP-16S-metaG, respectively.

Contrary to the 16S rDNA datasets, we did not perform taxonomic assignment of *Tara* Oceans 18S rDNA metaG reads through phylogenetic placement (see Section 2.2.1), but rather through miTags clustering (Logares et al., 2013). We mapped the reads to reference OTUs deriving from an ad-hoc Sanger database that included SILVA v.119 for 18S rDNA reads. This database was pre-clustered at 97% using *USEARCH* and the cluster-representative sequences became reference OTUs. OTU delineation was performed using UCLUST with the following parameters: identity >99%, coverage 90%, --maxaccepts 5,000 and --maxrejects 5,000. The output table contained 183,939 reads and 7,663 miTags, with an average of $\sim 4,380$ reads and 1,348 miTags per sample. We refer to the table as TO-18S-metaG (for *Tara* Oceans 18S rDNA metaG).

3.2.2 Characterisation of phototrophic group richness and relative abundance

We assessed the community in picophototrophs in metaB and metaG data using the annotation pipeline described in Chapter 2. In short, the method consisted on placing 16S rDNA reads into the PhotoRefT reference tree. PhotoRefT contains 802 16S rDNA sequences representing the main lineages of Oxyphotobacteria and PPE, together with ∼90 outgroup sequences from Streptophyta, archaea and heterotrophic bacteria. This tree provides the support to identify high-throughput reads from picophototrophs through phylogenetic placement. In our study, MetaB and metaG reads were placed in PhotoRefT using EPA-ng (Barbera et al., 2017). Placements with pendant branches four-times longer than the local average were excluded. We obtained the estimates of abundance and diversity for the different groups under study (see Table 2.1) by filtering placements with a likelihood weight mass threshold of 0.95 using GAPPA (Czech & Stamatakis, 2019).

3.2.3 Phylogenetic diversity measurements

We compared the phylogenetic structure of the different phototrophic communities using Faith's phylogenetic diversity metric (PD) and the mean pairwise distance (MPD) divergence index. While PD measures the evolutionary diversity in a community, MPD tells us about the average phylogenetic diversity of species within samples. We limited our diversity measurements to the TO-16S-metaB dataset as its higher sequencing depth allows for the evaluation of the largest possible diversity of Oxyphotobacteria and PPEs in the analysed samples.

We selected the 12,709 OTU sequences from TO-16S-metaB which were classified as phototrophic after phylogenetic placement to build a phylogenetic tree. We then aligned these sequences using *MAFFT* v7.299b (method FFT-NS-2, parameters --ep 0.2 --op 2.5). The output was trimmed with *trimAL* using options -gt 0.2 -st 0.0005. The resulting alignment was passed to *RAxML-NG* (Kozlov et al., 2019)to infer 20

trees using the GTRGAMMA model. Tree searches were topologically constrained at phylum level and GloeobacterPCC7421_AF132791 was used as the outgroup. We selected the best-scoring maximum likelihood tree for PD and MPD analyses. PD was calculated using the *pd* function from the R package *Picante* (Kembel et al., 2010)and normalizing it by number of species in each sample. MPD was computed with the function *ses.mpd* from the same package, weighting species by abundance.

3.2.4 Cross-validation of meta-omics approaches

We compared the estimates of the PPE groups' relative abundances between meta-omics approaches (PCR-free metaG *vs.* PCR-based metaB) and marker genes (plastidial 16S rDNA *vs.* nuclear 18S rDNA) [SF1] using *Tara* Oceans datasets: TO-16S-metaG, TO-16S-metaB and TO-18S-metaG. The taxonomic groups that appeared in less than 50% of the samples were excluded from the analysis. We normalised the data using the centred log-ratio (clr) transformation to later test for variation between samples in each dataset through Principal Component Analysis (PCA). We compared the equivalence between dataset ordinations using the symmetric Procrustes rotation function *protest* from *vegan* with 999 permutations.

3.3 Results

3.3.1 Global contribution and distribution of PPEs and Oxyphotobacteria

A total of 1,970,069 reads and 12,709 OTUs from *T*O-16S-metaB were classified as phototrophs, representing 24% of reads and 19% of OTUs. While the Oxyphotobacteria accounted for >70% of picophototrophs' global relative abundance, they represented only ∼20% of their richness (Figure 3.1). Synechococcales was the dominant group in Oxyphotobacteria while the remaining phyla contributed to less than 3% of the phototrophic bacteria fraction abundance and richness. In the PPE community, Prymnesiophyceae, Mamiellophyceae, Pelagophyceae and Dictyochophyceae were the most abundant and diverse groups (Figure 3.1), jointly contributing to 90% and 83% of the total relative abundance and richness in PPEs. Remarkably, Prymnesiophyceae alone (with 5,917 OTUs) accounted for about 60% of the relative richness in the picoeukaryotic community and ∼47% of the overall richness of photosynthetic picoplankton in the analysed samples.

A total of 74,688 reads (4.4%) from TO-16S-metaG were placed in some phototrophic groups. The results of abundance and richness obtained in this metagenomic dataset (Figure S3.3) were remarkably comparable to the ones described for TO-16S-metaB. Most of the groups' abundances were correlated between metabarcoding and metagenomics sequencing techniques (Figure 3.2); only low abundance groups (relative abundance <1%) showed some differences. Groups like Cryptophyta, Euglenozoa, Chlorarachniophyceae, Rhodophyta, Ulvophyceae and Nephroselmidophyceae were

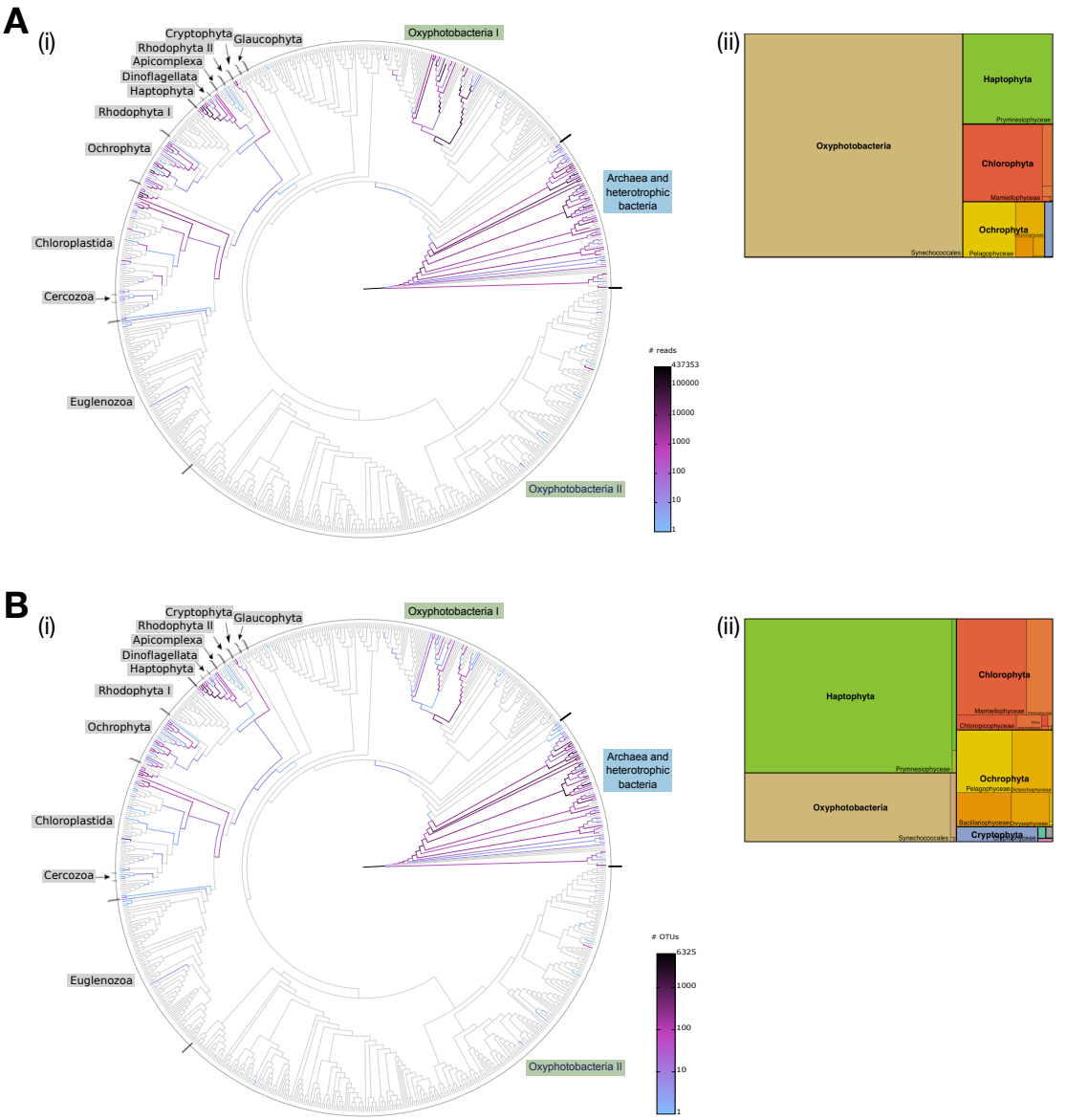


Figure 3.1 – Composition of Oxyphotobacteria and PPE communities in the 41 TO-16S-metaB samples according to (**A**) relative abundance and (**B**) richness. The relative contribution of each group is represented both as (**i**) heat trees -where each branch of the 16S rDNA reference tree is colored according to accumulated placement mass-, and (**ii**) tree maps. The heat trees include reference sequences from Oxyphotobacteria (green labels); archaea and heterotrophic bacteria (blue label), and PPEs (grey labels). PPE references include the following groups: Euglenozoa, Cercozoa, Chloroplastida, Ochrophyta, Rhodophyta, Haptophyta, Dinoflagellata, Apicomplexa, Cryptophyta and Glaucophyta. Oxyphotobacteria I encompasses Eurycoccales, Leptolyngbyales, Limnotrichales, Phormidesmiales, Pseudanabaenales, Synechococcales and Thermosynechococcales, while Oxyphotobacteria II corresponds to Nostococales.

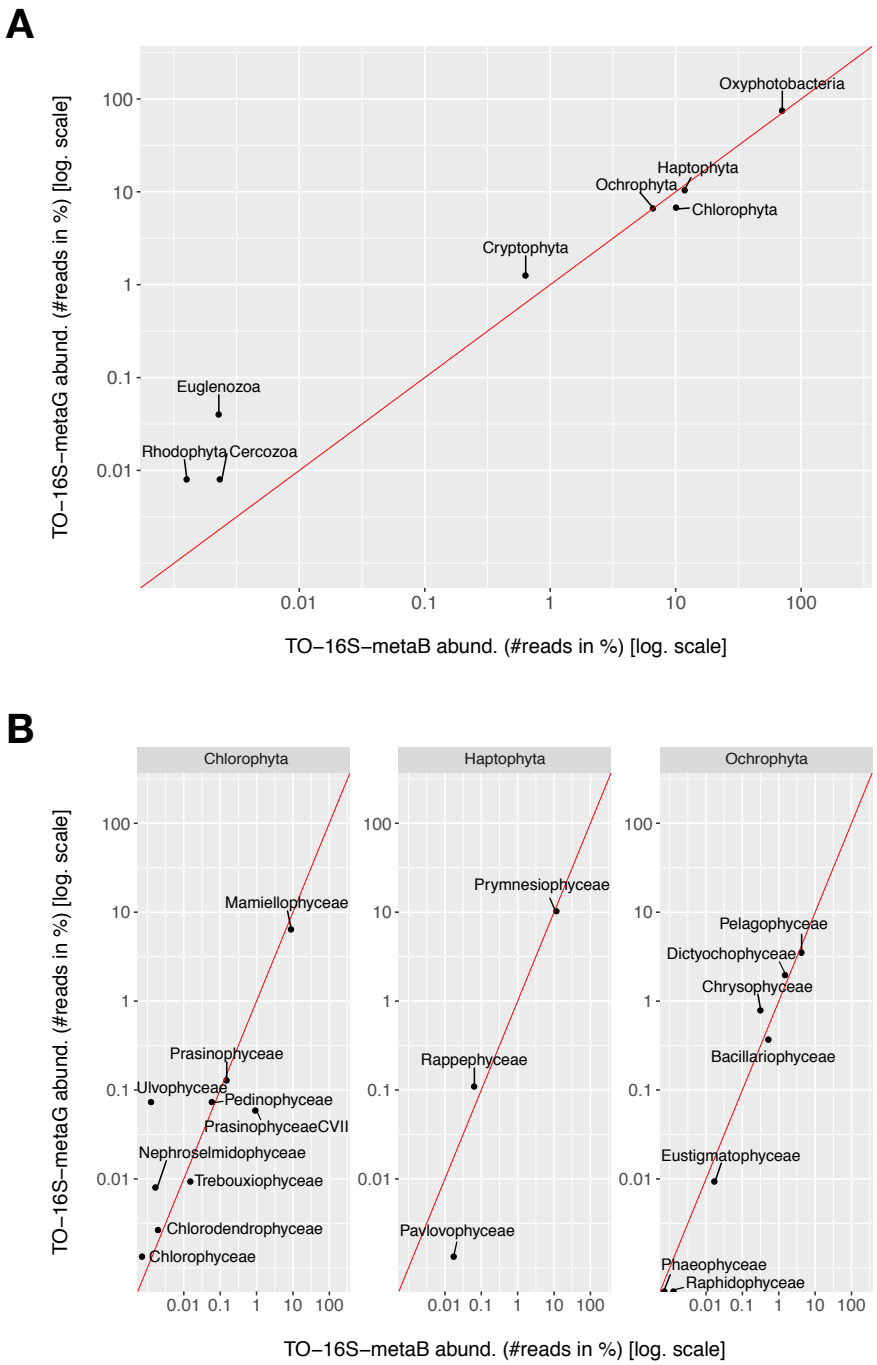


Figure 3.2 – Correlation of phototrophic groups' relative abundance obtained in metabarcoding (TO-16S-metaB) and metagenomic (TO-16S-metaG) datasets. (**A**) Correlation of Oxyphotobacteria and PPE phyla (Dinophyta excluded). (**B**) Correlation of the PPE groups within the three main eukaryotic phyla (Chlorophyta, Haptophyta and Ochrophyta).

better captured by metaG, while Chloropicophyceae, Pavlovophyceae and Raphidophyceae showed higher values in the metaB approaches (Figure 3.2).

Photosynthetic groups displayed biogeographic trends associated to oceanic basins (Figure S3.4, Figure S3.5 and Figure S3.6). For instance, Oxyphotobacteria showed significantly lower relative abundance in the South Atlantic Ocean, in comparison with the sampling sites in the South Pacific Ocean and Indian Ocean (Scheffe's Test $p < 0.05$). Contrarily, Mamiellophyceae constituted more than half of PPE abundance in some samples from the North Atlantic, Indian and first stations of the South Atlantic Oceans, while they were almost absent in most samples of the Pacific Ocean. Prymnesiophyceae fluctuated notably across regions, ranging from 8% to 78% in PPE relative abundance. Other dominant groups like Pelagophyceae and Dictyochophyceae showed their highest peaks in relative abundance (of up to 58% and 36%, respectively) around the stations of South Pacific and South Atlantic Oceans.

The relative richness of groups like Oxyphotobacteria, Haptophyta and Chlorophyta fluctuated proportionally to their contribution to the relative abundance (Figure S3.7), showing correlation values above +0.8 between the two indexes. In opposition, the distribution of Ochrophyta's relative richness was strikingly constant across sampling sites.

As for Malaspina results (Figure S3.8), in MSP-16S-metaB we classified 23,286 reads and 781 OTUs (42% and 23% of the total reads and OTUs, respectively) as phototrophic. In MSP-16S-metaG, we classified 2,761 reads (2% of the total) as Oxyphotobacteria or PPE. The main difference between the results from Malaspina and *Tara* Oceans data was that in the former dataset, Oxyphotobacteria accounted for a larger proportion of phototroph abundance (>98% in MSP-16S-metaB and >92% in MSP-16S-metaG). In terms of diversity, we found Oxyphotobacteria to be the dominant group in MSP-16S-metaB, with 74% of the OTUs. Cryptophyta, Rhodophyta, Euglenozoa and most of the groups under 1% of abundance in the *Tara* Oceans samples were not detected at all in Malaspina samples. Despite these discrepancies, the results from the Malaspina stations were also consistent between sequencing methods (metaB *vs.* metaG) (Figure S3.9).

3.3.2 Characterisation of phylogenetic diversity between phototrophic communities

The global phylogenetic diversity of phototrophic communities as measured by the PD index, was significantly higher ($p < 0.01$ Wilcoxon test) in PPEs ($PD\overline{x} = 0.092$, $PD_{SD} = 0.055$) when compared to that of Oxyphotobacteria ($PD\overline{x} = 0.032$, $PD_{SD} = 0.003$) (Figure 3.3A). Similarly, MPD (Figure 3.3B) was higher ($p < 0.01$ Wilcoxon test) in PPEs ($MPD\overline{x} = 0.984$, $MPD_{SD} = 0.325$) than in Oxyphotobacteria ($MPD\overline{x} = 0.047$, $MPD_{SD} = 0.018$).

Regarding PPE groups, the ones with highest PD values were Chlorophyta ($\overline{x} = 0.318$, $PD_{SD}= 0.25$), Rhodophyta ($PD\overline{x} = 0.374$, $PD_{SD} = 0.285$) and Euglenozoa

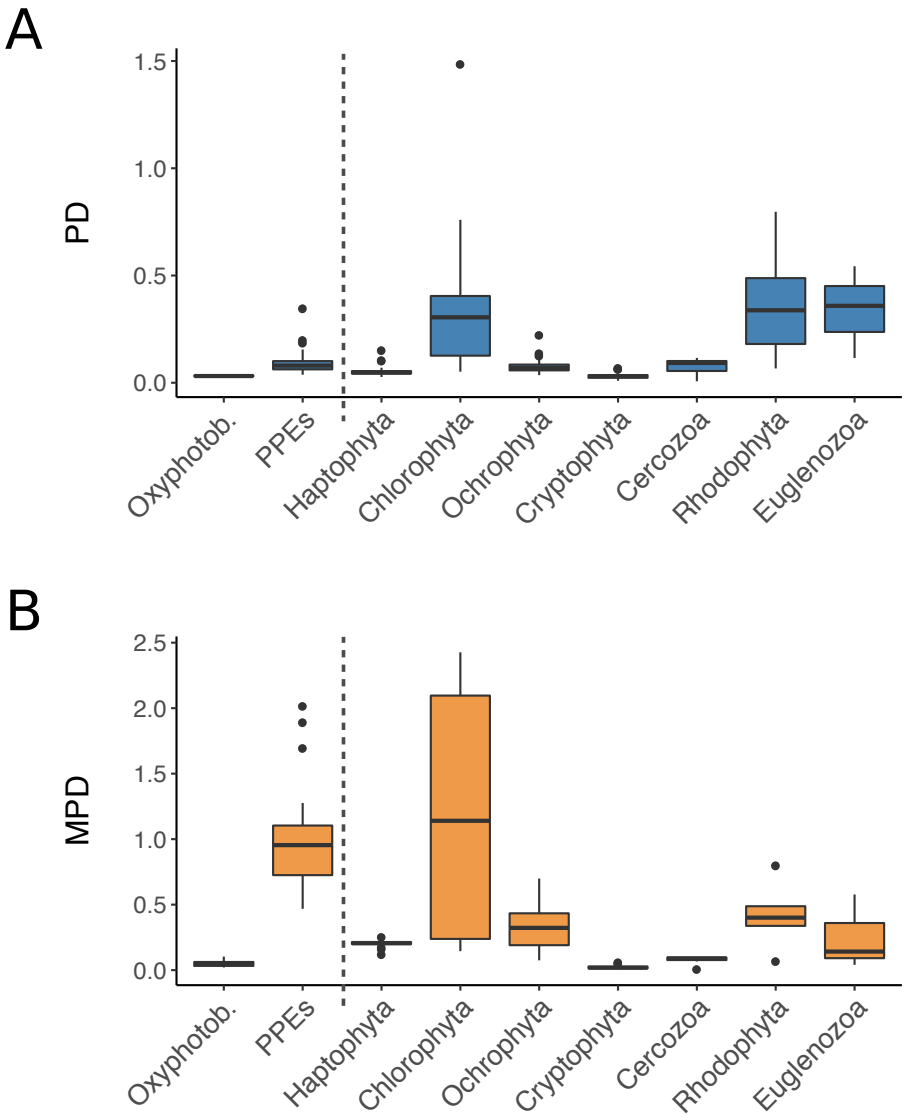


Figure 3.3 – Measures of phylogenetic diversity in Oxyphotobacteria and PPE within TO-16S-metaB. The analyses are computed per individual samples through (**A**) Faith's Phylogenetic Diversity (PD) and (**B**) Mean Pairwise Distances (MPD) indexes.

($PD\overline{x}$ = 0.34, PD_{SD} = 0.215). In contrast, the two other dominant PPE phyla, Haptophyta ($PD\overline{x}$ = 0.051, PD_{SD} = 0.023) and Ochrophyta ($PD\overline{x}$ = 0.076, PD_{SD} = 0.033), displayed lower PDs. Chlorophyta ($MPD\overline{x}$ = 1.13, MPD_{SD} = 0.863) was the group that contributed the most to PPEs' MPD and showed differences of up to 16-fold between the samples. Within Haptophyta, MPD was low and constant across samples ($MPD\overline{x}$ = 0.205, MPD_{SD} = 0.022). In comparison, Ochrophyta showed greater MPD average and variance ($MPD\overline{x}$ = 0.337, MPD_{SD} = 0.177).

3.3.3 Sequencing methods comparison

The quantification of PPE groups' relative abundances with different marker genes chloroplastic 16S *vs.* nuclear 18S rDNA) and meta-omics approaches (metaB *vs.* metaG) showed strongly correlated results (Table 3.1). The symmetric Procrustes analysis indicated that the correlations ($Procrustes_{corr}$) between datasets ordinations were above +0.8 in all cases, meaning that they provide a similar view of PPE community structure. The ordinations of TO-metaG-16S and TO-metaG-18S samples correlated slightly better between themselves ($Procrustes_{corr}$ = 0.896) than when they were compared against the results of TO-metaB-16S ($Procrustes_{corr}$ = 0.804, in both cases). The permutation p-values were 0.001 in all comparisons, indicating that the similarity between datasets ordinations was significant. The fact that TO-metaG-16S correlated better with TO-metaG-18S than with TO-metaB-16S argues that the quantification of PPEs was more susceptible to biases derived from the adopted meta-omics approach than from the differences that we could expect from amplifying two different marker genes. On the other hand, the strong correlation between metaG datasets points out that 16S and 18S rDNA provide close estimates of PPE when analyzed through metaG approach.

3.4 Discussion

3.4.1 MetaB and metaG sequencing approaches to assess the diversity of ocean picophytoplankton.

In general, our results show that the assessment of total phototrophic picoplankton (both prokaryotes and eukaryotes) through 16S rDNA metaB and metaG sequencing yield strikingly similar results. The strong correlation between both methods suggests that the plastidial sequence-based measure used herein is not significantly conditioned by the potential biases associated with each sequencing approach. In addition, this suggests that reads as short as 100bp (as found in metaB) hold sufficient information for taxonomic profiling of most of the PPE plastid groups, as it was observed in bacterial communities (Hao & Chen, 2012; Logares et al., 2013). In fact, low-abundance PPE groups such as Cryptophyta, Euglenozoa, Cercozoa, Rhodophyta, Ulvophyceae or Nephroselmidophyceae were detected in higher proportion through metagenomic data than in amplicons sequencing. Such difference could be explained by primer mismatch in these groups during metaB amplification, or PCR underestimation of low-abundance species (Gonzalez et al., 2012). On the

	TO-16S-metaB	TO-16S-metaG	TO-18S-metaG
TO-16S-metaB	1 *p*-value 0.001	0.8047 *p*-value 0.001	0.8049 *p*-value 0.001
TO-16S-metaG	0.8047 *p*-value 0.001	1 *p*-value 0.001	0.8960 *p*-value 0.001
TO-18S-metaG	0.8049 *p*-value 0.001	0.8960 *p*-value 0.001	1 *p*-value 0.001

Table 3.1 – Comparison of *Tara* Oceans datasets composition based on relative abundance of PPEs. Individual datasets variation was evaluated through PCA and later compared with the rest of datasets with Procrustes correlation. The correlation between datasets ordinations is above 0.8 in all comparisons and the significance of the Procrustes statistic obtained after 999 permutations is positive in all correlations (p-value = 0.001).

other hand, the low abundance or lack of detection of Chloropicophyceae, Pavlovophyceae and Raphidophyceae in 16S rDNA extracted from metagenomes could be explained by the significantly lower amount of data recovered by metaG as compared to deep-sequencing metaB. These groups could have been under-estimated or missed due to low sequencing coverage. Overall, low abundance PPE groups seem to be underrepresented in both sequencing approaches. While metaG sequencing might be more accurate when it comes to quantify the relative abundance of rare groups, metaB sequencing is more likely to detect their presence or absence without requiring the sequencing depth applied to metaG.

As in *Tara* Oceans datasets, the assessment of Oxyphotobacteria and PPE groups in Malaspina samples displayed comparable results between sequencing approaches. This suggests that the differences in the groups' relative abundances between the two expeditions may be related to variations in their sampling routes, temporal changes in communities or seasonality, or generated by the different DNA extraction protocols.

We did not identify significant differences between 16S rDNA and 18S rDNA on PPE screening, as reported in Shi et al. (2011). On the contrary, the high similarity between *T*O-16S-metaB and TO-16S-metaG with TO-18S-metaG validates the use of the bacterial/plastidial 515yF-926R primers and whole plankton metaG sequencing for assessment of PPE community. Furthermore, the positive correlation of PPE groups' relative abundances obtained through the plastidial and nuclear marker genes could indicate that the ratio between 16S and 18S copy number in PPEs remains proportional across groups.

3.4.2 Community structure of Oxyphotobacteria and PPEs

According to both metaG and metaB, Oxyphotobacteria were three-fold more abundant than PPEs in surface oligotrophic waters of temperate and tropical ocean regions. This is consistent with the higher biomass of Oxyphotobacteria in temperate surface waters reported previously by Bouman et al. (2011) and Buitenhuis et al. (2012). In agreement with preceding studies based on flow cytometry (e.g. Buitenhuis et al., 2012), molecular probes (Kirkham et al., 2013), and pigment markers (Bouman et al., 2011), we also observed an increase of PPE relative abundance as latitude increases. Eukaryotic plastidial genomes contain between 1 and 10 copies of 16S rDNA, although typically they hold two replicates of the gene (Decelle et al. 2015, Green 2011). Even if most PPEs tend to have one or very few plastids per cell, the number of plastidial genomes per chloroplast in eukaryotic cells ranges between 50 and 100 copies for what we know (Decelle et al. 2015). We still lack detailed data about the variability in the number of plastidial genomes per cell across the different eukaryotic groups, making it hard to estimate PPE cell number through 16S rDNA counts. In the case of Oxyphotobacteria though, we can infer cell number quite directly through ribosomal genes because the number of 16S rDNA copies per genome only oscillates between 1 and 4. The low number of 16S rDNA copies per cell in Oxyphotobacteria, in comparison with that in PPEs, reinforce the greater dominance of this group within picophototrophic communities in terms of cell number.

In terms of richness, PPEs accounted for ∼80% of picophototrophs total richness, thus largely dominating Oxyphotobacteria. The principal groups were Prymnesiophyceae, Mamiellophyceae, Pelagophyceae and Dictyochophyceae. Overall, eukaryotic microalgae from the 'red lineage' (containing 'red' plastids from secondary or higher-level endosymbiosis), essentially mixotrophic, are dominating open ocean PPE communities. The strong diversification and cosmopolitanism of Prymnesiophyceae in the open oceans, which account for ∼60% of PPE richness in our datasets, has been highlighted previously through molecular hybridisation and 16S rDNA (Jardillier et al., 2010; Kirkham et al., 2013) and 28S rDNA (Liu et al. 2009) Sanger-sequenced libraries, as well as via pigment data (Not et al. 2008, Liu et al. 2009), and using Illumina-sequenced 18S V9 amplicons data (De Vargas et al., 2015). Pelagophyceae displayed an average relative abundance of 14% within PPEs. Not et al. (2008) identified this group as dominant in the surface of the Indian Ocean, contributing 28% of picoeukaryotic pigments. Using 16S rDNA Sanger-sequenced libraries, Shi et al. (2011) also detected a high abundance of Pelagophyceae in the South Pacific, accounting for 60-17% of the sequences. Additionally, the abundance of Pelagophyceae in our results match fairly well the estimates from the *Tara* Oceans stations in common with our study reported by de Vargas et al. (2015) and based on massive 18S V9 rDNA metaB. Dictyochophyceae accounted for 5-8% of PPEs global relative abundance in metaB and metaG, respectively. Shi et al. (2011) and de Vargas et al. (2015), show relative contributions of Dictiochophyceae comparable to our observations.

Chlorophyta accounted for 26% of the global relative PPEs abundance and this group displayed high variability across basins: while they contributed to 12.4% of the reads in the South Pacific Ocean -consistently with the results reported by Shi et al. (2011)-, in the Indian Ocean they reached ∼46% of the reads in the community. In de Vargas et al. (2015), the contribution of Mamiellophyceae to PPE relative abundance is half of what we observed in the matching stations. The difference on the quantification of this group in comparison to our study could be due to the fact that piconano-size fraction in de Vargas et al. (2015) [0.8-5µm] is slightly bigger than the one used in this survey (0.2-3µm), but also because of primer mismatch in Mamiellophyceae 18S V9 rDNA or over-quantification of this group through 16S rDNA.

While Chrysophyceae accounted for <3% of the relative abundance in metaB and metaG, it is one of the dominant groups in the study of Kirkham et al. (2013), where they analyse over 90 stations covering the major ocean basins through 16S rDNA dot blot hybridisation and cloning-sequencing data. The frequency of Chrysophyceae is also comparatively higher in Shi et al. (2011), where the group accounts for ∼20% of the total relative abundance. In contrast, the contribution of Chrysophytes in de Vargas et al. (2015), as measured by massive Illumina sequencing of nuclear 18S-V9 metaB, is similar to that in our results. This suggests that the relatively high abundance detected previously for this group could be due to PCR amplification biases of the relatively larger DNA fragment used for Sanger-sequenced clone-libraries a decade ago.

Recent studies have reported Chloropicophyceae as a prevalent group in the oligotrophic ocean (Dos Santos et al., 2018; Tragin & Vaulot, 2018; de Vargas et al., 2015). Overall, we detected an abundance of less than 1% of Chloropicophyceae in both our metaB and metaG datasets. Since we do not observe differences between sequencing approaches, the underestimation of Chloropicophyceae in our results could derive from post-sequencing analyses like an incomplete representation of this group in the PhytoRefT database.

In summary, we show that the structure of picophototrophic communities obtained by metaB (using the 515yF-926R 16S rDNA primers) and metaG sequencing is broadly comparable to the estimates obtained through flow cytometry, molecular probes hybridisation, pigments data, 16S rDNA clone libraries or 18S rDNA amplicons. Most of the groups could be identified through phylogenetic placement and our approach provides phylogenetic information of picophototrophs diversity that we miss through methods that only rely on similarity to reference sequences for taxonomic identification.

3.4.3 Phylogenetic assessment of picophototrophic communities

Despite Oxyphotobacteria displayed higher richness than PPEs in some stations, we observed a significantly lower phylogenetic diversity in this group than in the PPE community. The diversity of Oxyphotobacteria, mainly concentrated in a single

taxonomic rank, contrasts with the pattern revealed in PPEs, where divergence is scattered through polyphyletic eukaryotic lineages. The higher evolutionary diversity observed in eukaryotic plastids reflects the wide variety of functional strategies known (and still unknown!) in photo/mixotrophic protists, and could also be explained by the fact that endosymbionts' genomes evolve faster than those of related free-living lineages (Moran, 1996).

The dominant PPE groups differed in PD and MPD: in comparison with Ochrophyta and Haptophyta, Chlorophyta plastids displayed a larger PD and MPD. The high evolutionary diversity in Chlorophyta could indicate faster evolutionary rates, longer evolutionary history, and/or lower purifying selection in the organelles of this group. However, plastid evolution is so complex that it cannot be inferred using 16S rDNA as the only marker gene (Keeling, 2004; Sanchez-Puerta & Delwiche, 2008). Analyses of further genes related to the photosynthetic machinery of PPE would help to better disentangle the causes for differential evolutions of PPEs' plastids. Besides, nuclear-encoded genes can show different evolutionary pathways compared to plastid genome-encoded markers (Cuvelier et al., 2010) and complementing these results with analyses of 18S rDNA PD and MPD will also be necessary to clarify if the diversity observed in PPE plastids is consistent with the evolutionary signal encoded in their nuclear genomes.

3.5 Supplementary Figures

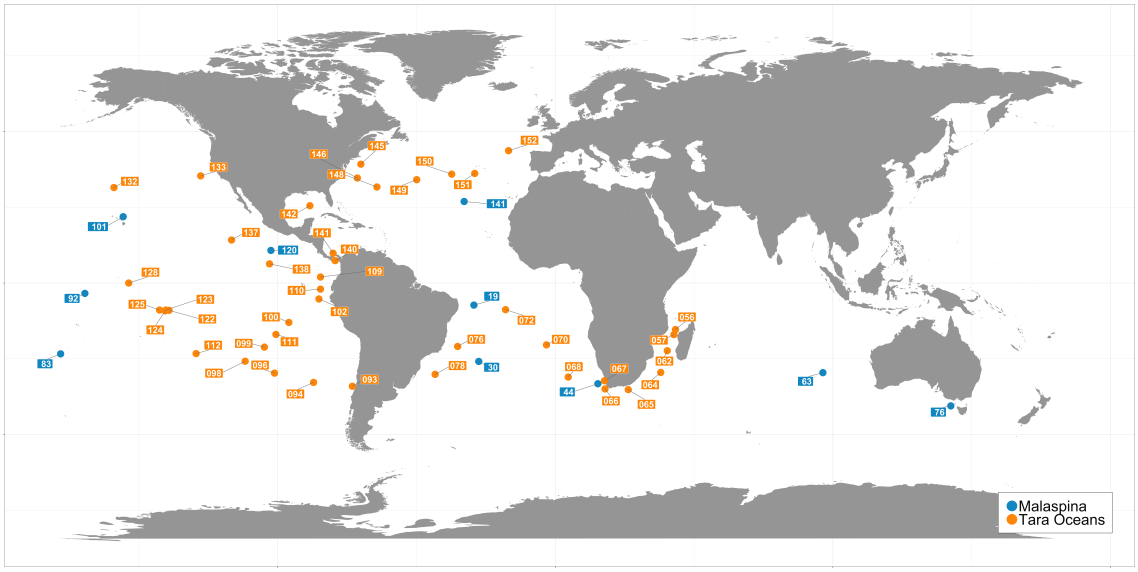


Figure S3.1 – World map with the distribution of the *Tara* Oceans and Malaspina samples analysed.

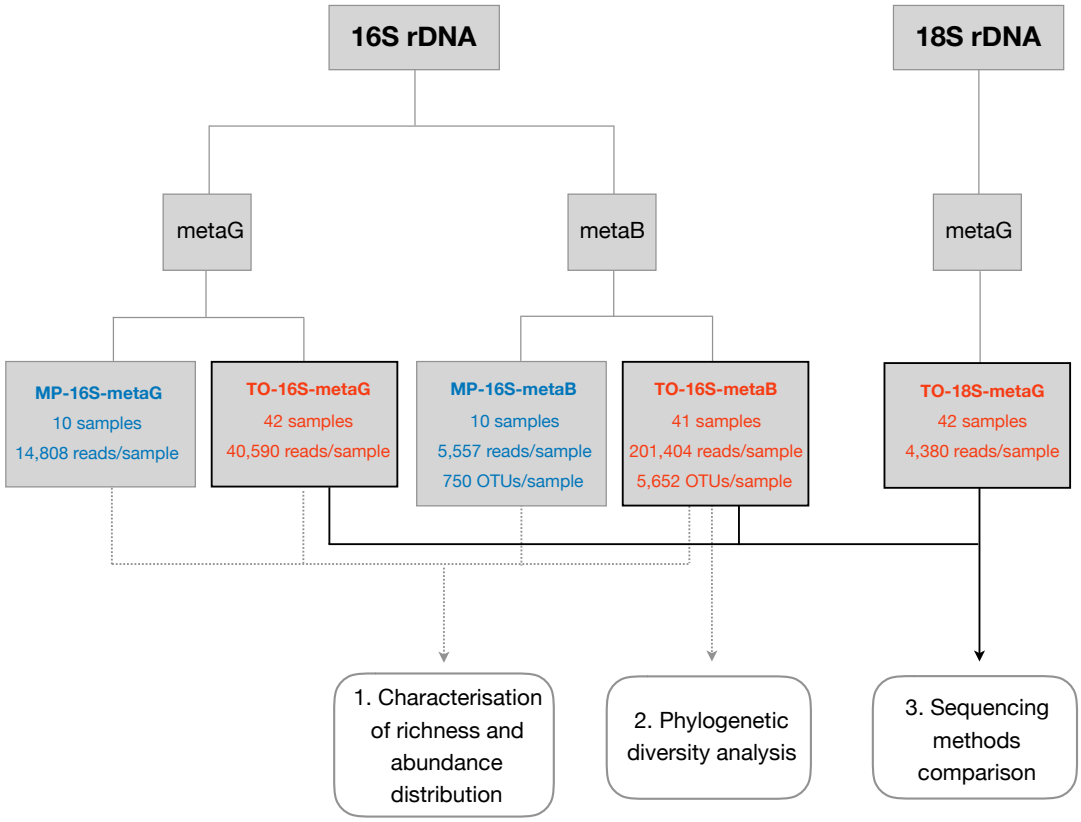


Figure S3.2 – Methods diagram describing the input datasets and their use in the different analyses.

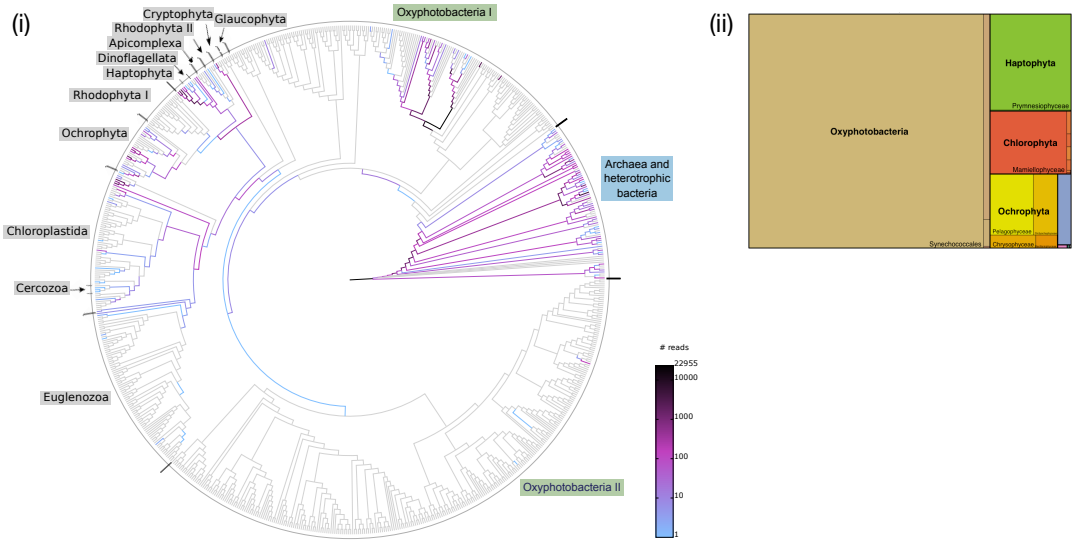


Figure S3.3 – Relative abundance of Oxyphotobacteria and PPEs communities in the 42 samples of TO-16S-metaG. The relative contribution of each group is represented both as (**i**) heat tree displaying the number of metagenomic reads (16S rDNA) placed in our summary tree of photosynthetic life, and (**ii**) tree map showing the relative proportion of metagenomic reads amongst photrophic phyla. The heat tree includes reference sequences from Oxyphotobacteria (green labels); archaea and heterotrophic bacteria (blue label), and PPEs (grey labels). List of groups as in Figure 3.1.

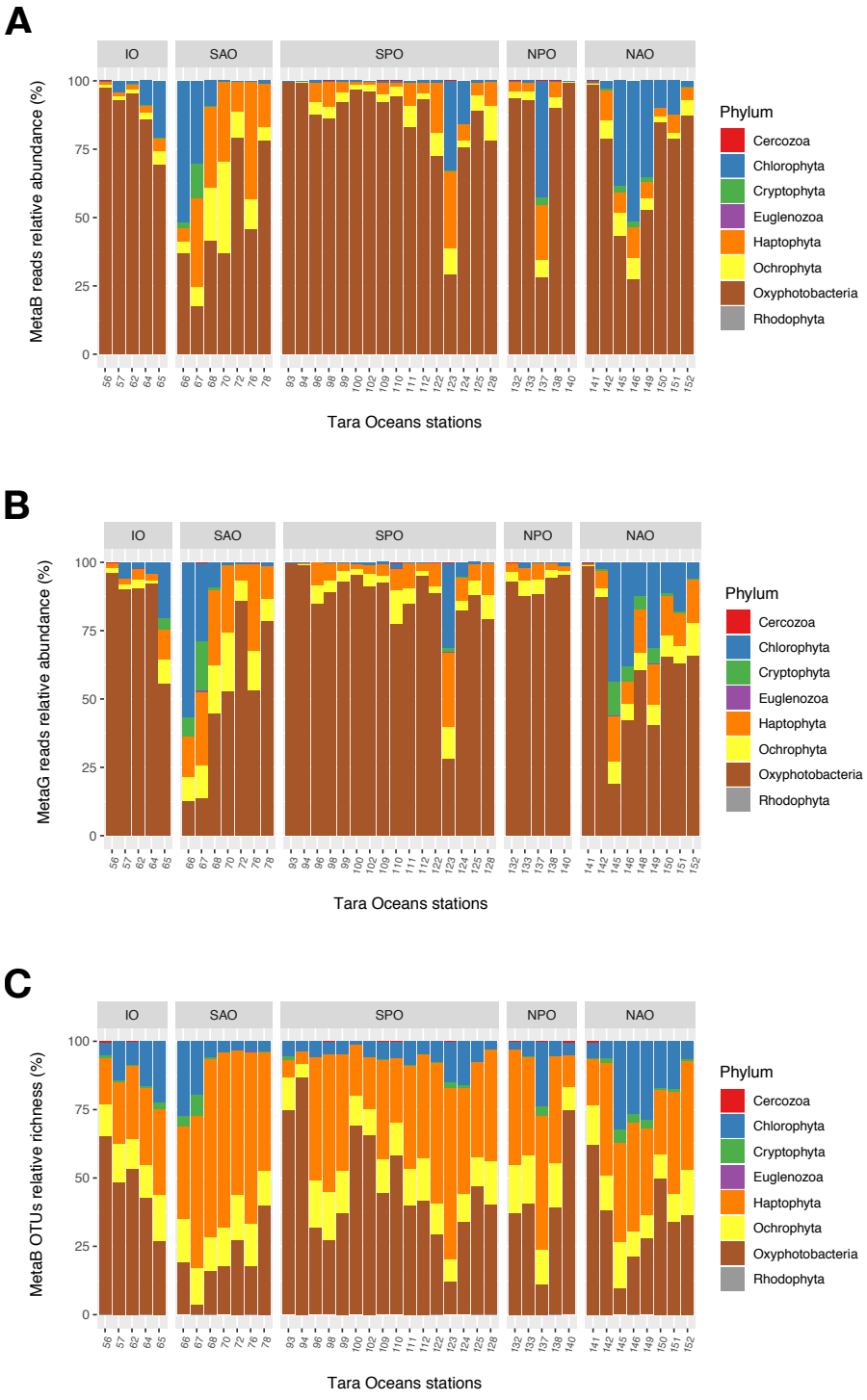


Figure S3.4 – Global relative abundance and richness of Oxyphotobacteria and PPE phyla across sampling stations. The ocean basins acronyms correspond to the following regions: IO, Indian Ocean; SAO, South Atlantic Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean. (**A**) Relative abundance of phototrophic groups in TO-16S-metaB stations. (**B**) Relative abundance of phototrophic groups in TO-16S-metaG stations. (**C**) Relative richness of phototrophic groups in TO-16S-metaB stations.

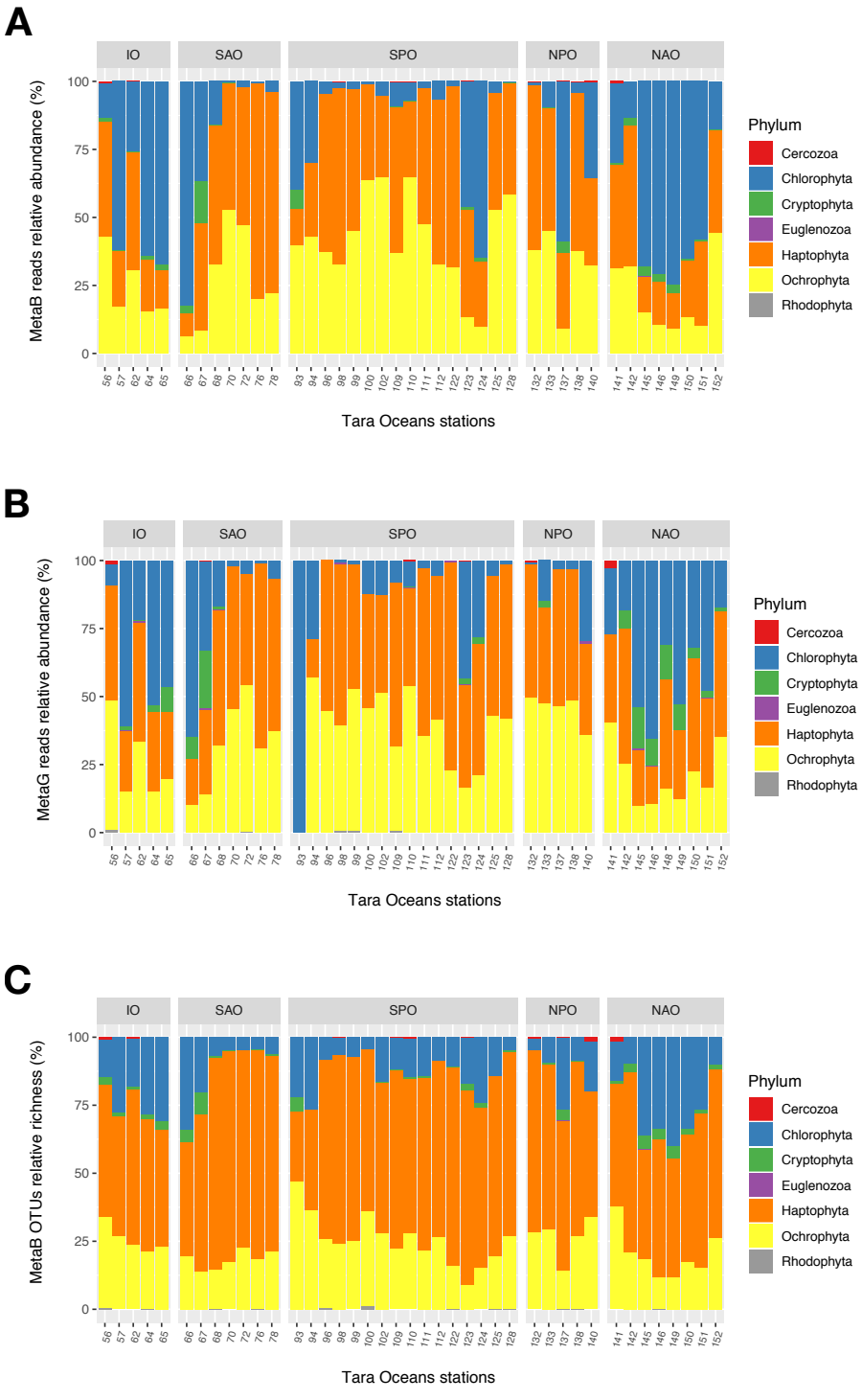


Figure S3.5 – Global relative abundance and richness of PPE phyla across sampling stations. Ocean basins acronyms as in Figure S3.4. (**A**) Relative abundance of PPE phyla in TO-16S-metaB stations. (**B**) Relative abundance of PPE phyla in TO-16S-metaG. (**C**) Relative richness of PPE phyla in TO-16S-metaB stations.

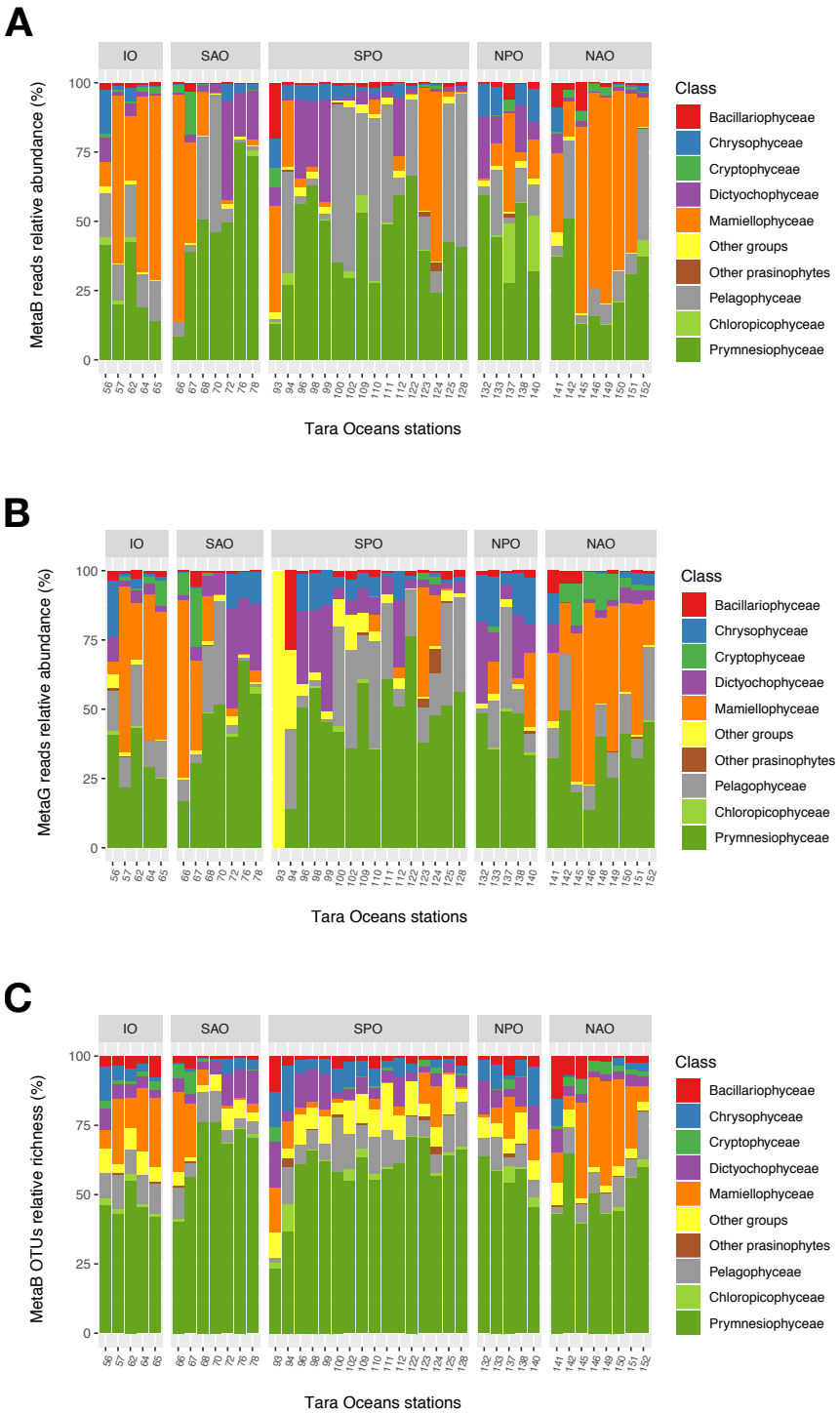


Figure S3.6 – Global abundance and diversity of PPE at class level across sampling stations. Ocean basins acronyms as in Figure S3.4. (**A**) Relative abundance of PPE groups in TO-16S-metaB stations. (**B**) Relative abundance of PPE groups in TO-16S-metaG. (**C**) Relative richness of PPE groups in TO-16S-metaB stations.

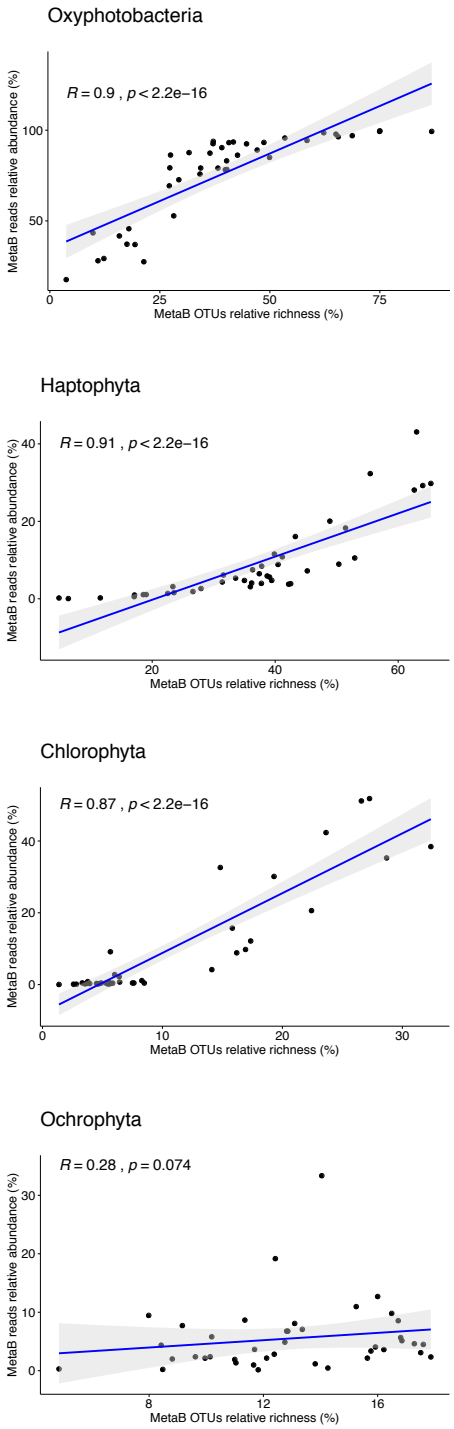


Figure S3.7 – Spearman correlation between group contribution to relative abundance and relative richness in TO-16S-metaB. Measures shown for Oxyphotobacteria, Haptophyta, Chlorophyta and Ochrophyta.

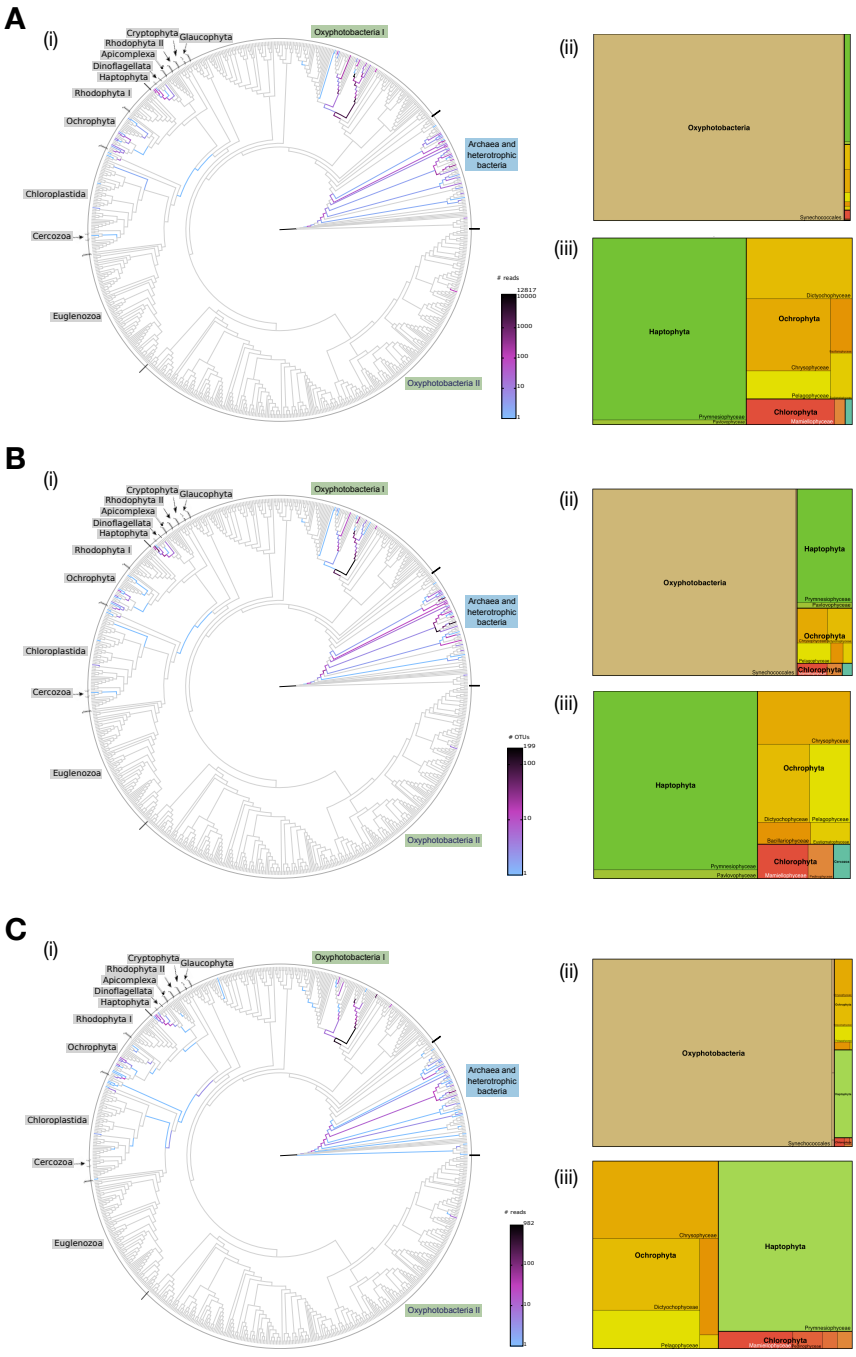


Figure S3.8 – Composition of Oxyphotobacteria and PPE communities in the 10 Malaspina samples. (**A**) Relative abundance of phototrophic groups in MSP-16S-metaB. (**B**) Relative abundance of phototrophic groups in MSP-16-metaG. (**C**) Relative richness of phototrophic groups in MSP-16S-metaB. The relative contribution of each group is represented both as (**i**) heat trees, (**ii**) tree maps combining Oxyphotobacteria and PPE phyla and (**iii**) tree maps with PPE phyla alone. The heat trees include reference sequences from Oxyphotobacteria (green labels); archaea and heterotrophic bacteria (blue label), and PPEs (grey labels). List of groups as in Figure 3.1.

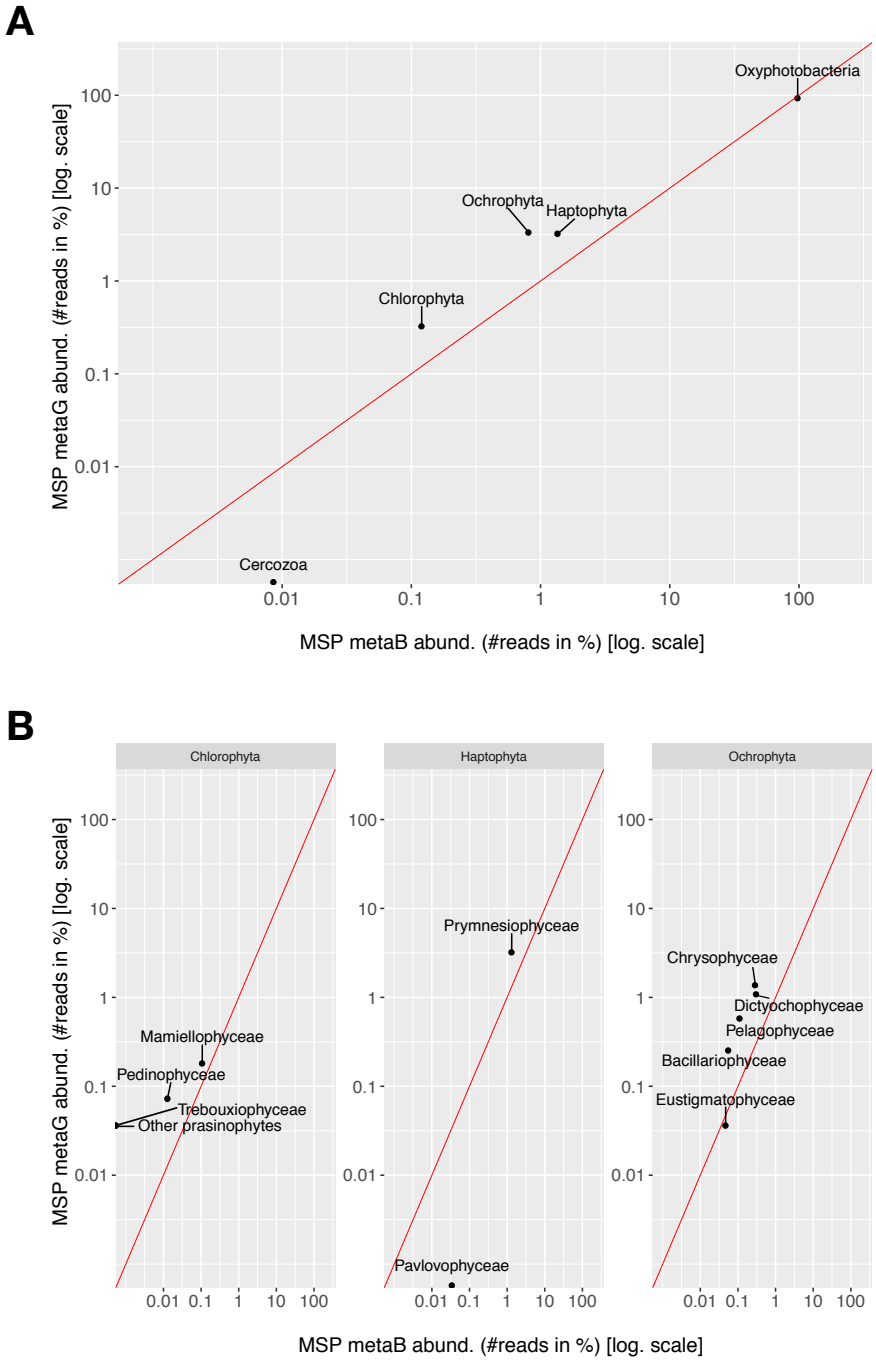


Figure S3.9 – Correlation of phototrophic groups' relative abundance obtained in MSP-16S-metaB and MSP-16S-metaG. (**A**) Correlation of Oxyphotobacteria and PPE phyla. (**B**) Correlation of the PPE groups within Chlorophyta, Haptophyta and Ochrophyta.

4. Modelling metagenomes to assess pico/nano-plankton trophic diversity across the world ocean

Marine microeukaryotes present a diversity of trophic strategies. While some of these organisms feed purely through photosynthesis or phagotrophy, phago-mixotrophy is increasingly being recognized as an inherent capability of many eukaryotic plankton groups that inhabit the photic-zone. The identification of a protist's trophic mode can be challenging because most groups are uncultured and their physiology can change across life-cycle, season and region. In this chapter, we built up and describe a new model that allows the quantification of different trophic groups (photo-, phagomixo-, phago- and osmo-trophs) in metaG samples. The model was fed with the collection of genes recruited from *Tara* Oceans metagenomic samples to evaluate the prevalence of the different trophic groups across the oligotrophic ocean and plankton size-fractions. Next, we evaluated the correlation of trophic groups' estimates with environmental parameters. Our results indicated that phototrophs account for more than half of the taxa in most piconano- and nano- size fraction samples. We also observed that phago- and phagomixo-trophs contribute to about one quarter of the planktonic communities each. The trophic composition of the eukaryotic communities showed comparable results across plankton size fractions and between surface and 'deep-chlorophyll maximum' depths, but we detected differences among size fractions and depths when compared with environmental factors.

The model described in this chapter was developed by our collaborators John Burns (Bigelow Laboratory for Ocean Sciences) and Giulio Trigila (Baruch College). The analysis of *Tara* Oceans metaG datasets was done in collaboration with John Burns, my supervisors Colomban de Vargas and Olivier Jaillon, and our colleagues Eric Pelletier and Nicolas Henry.

4.1 INTRODUCTION

Energy and mass flows in marine systems are still poorly understood. This is in part due to our limited knowledge about the exact nutritional strategy of many groups of microeukaryotes. The quantification of trophic groups in plankton ecosystems has been tackled through field surveys and, more recently, by means of modelling. Field studies based on taxonomic identification provide us with a broad picture of community composition (e.g. de Vargas et al., 2015), but the identification of trophic mode is only possible for those relatively few species or groups with known trophic physiology (Faure et al., 2019). On the other hand, numerical models have provided us with a holistic view of marine plankton ecosystems functioning (e.g. Mitra et al. 2014; Ward and Follows, 2016), but they lack physio-taxonomic precision and thus depend on highly idealized simulations. Despite the ecological importance of microbial food-webs in the ocean, we still miss a comprehensive measure of the global distribution and abundance of plankton trophic groups.

Recently, Burns et al. (2018) published a model that can predict the trophic mode of an organism based on its genotypic composition. In Burns' study, they screened polyphyletic reference genomes representative for phago-, photo- and proto-trophic organisms, and identified a set of proteins enriched for each trophic group. Based on the presence of such proteins in a query genome, their model predicts capacity for phagocytosis, phototrophy and prototrophy in individual species.

In the present study, we adapted the model described in Burns et al. (2018) to predict the proportion of photo-, phagomixo-, phago- and osmo-trophic groups in any planktonic community through its metaG signature. With this new approach, we screened the catalog of ∼116 million *unigenes* (or assembled transcripts) generated from *Tara* Oceans metaG samples (Carradec et al., 2018) to describe the trophic composition of eukaryotic plankton communities across the global oligotrophic ocean. Some phago-mixotrophic protists in large size fractions only have the intrinsic capacity to perform phagocytosis but are able to exchange nutrients with symbiotic phototrophs (Stoecker et al., 2017). Since this model is not able to distinguish phago-mixotrophic activity derived from symbiotic relationships (i.e. involving multiple genomes), we limited our analyses to piconano- (0.8-5μm) and nano- (5-20μm) size fractions. Lastly, we compared the frequencies of the different trophic modes with environmental parameters to try to identify which factors drive trophic communities' structure.

4.2 METHODS

4.2.1 Marine plankton unigenes selection

We screened the collection of ∼116 million transcripts from *Tara* Oceans to identify unigenes matching the proteins predictive for trophic mode. Such proteins were the ones used by the model published in Burns et al. (2018). The original version of Burns' model uses a set of ∼14,000 predictive proteins profiled into Hidden Markov

Models (HMMs). In the first round of unigene identification, we used a reduced version of the model based on 2,000 HMMs with the purpose of minimising the complexity of the unigene collection. We translated unigene sequences into the six possible protein coding frames using *transeq* (Madeira et al., 2019) and compared them against the 2,000 HMMs using *hmmsearch v3.1b2* (Eddy, 1998) to look for significant sequence matches. We normalised the alignment e-values based on the size of the unigene database and selected the hits with *hmm e-value* $\leq 10^{-5}$ and *domain e-value* $\leq 10^{-4}$. Next, we refined the identification of unigenes by feeding the obtained hits into the model from Burns et al. (2018) once more, this time using the version of the algorithm that includes 14,000 HMMs in order to obtain more accurate annotations. The results were normalised and filtered by e-value using the same parameters as in the first identification cycle. Among the selected unigenes, those classified as metazoa were excluded from downstream analyses. This yielded a total of 2,061,313 unigenes predictive for trophic mode.

We selected the unigenes appearing in samples from piconano- and/or nano- size fractions located in surface (SUR) and deep chlorophyll maximum (DCM) depths. Overall, the final unigenes dataset comprised 171 samples covering 63 sampling sites. We selected unigenes with ≥3 reads per sample and used unigene count data normalised per sample using the following formula:

$$A_{\mathrm{ij}} = \frac{\frac{R_j}{L_j}}{\sum_{j=1}^{N} \mathrm{Rij}} \tag{4.1}$$

where A_{ij} represents the relative abundance of unigene j in sample i; R_j is the number of reads mapped on unigene j; L_j is the length of unigene j covered by reads, and $\sum_{j=1}^{N} R_{ij}$ is the total amount of reads in sample i that mapped on all unigenes.

4.2.2 Model for trophic diversity prediction

The model described in this section was designed to determine the proportion of trophic groups (including photo-, phagomixo-, phago- and osmo-trophs) in a metaG sample based on its gene content. The approach used a set of ∼600 predictive genes from the collection of 14,000 proteins described in Burns et al. (2018), selected by virtue of their low copy number in the reference genomes (see the list of reference genomes in the Table 4.1).

To infer trophic group proportions from metaG data, the model of Burns et al. (2018) was extended to a likelihood mixture estimate approach whose mathematical procedures will be described in detail in a future publication by Trigila, Rubinat and Burns. In brief, the approach estimates the mixture of reference genomes that best represents the gene counts observed in metaG samples for proteins belonging to each predictive model. As represented in a simplified schema in (Figure 4.1, the likelihood mixture model is provided with (*i*) count data of the predictive genes in a metaG sample of unknown composition and (*ii*) the frequency data of the

PHOTOTROPHS	
Species	Group
Arabidopsis thaliana	Chloroplastida
Brachypodium distachyon	Chloroplastida
Chlamydomonas reinhardtii	Chloroplastida
Chlorella variabilis	Chloroplastida
Cyanidioschyzon merolae	Rhodophyta
Mimulus guttatus	Chloroplastida
Oryza sativa	Chloroplastida
Physcomitrella patens	Chloroplastida
Picea abies	Chloroplastida
Selaginella moellendorffii	Chloroplastida
Volvox carteri	Chloroplastida

PHAGOTROPHS	
Species	Group
Acanthamoeba castellani	Amoebozoa
Acytostelium subglobosum	Amoebozoa
Bodo saltans	Euglenozoa
Dictyostelium discoideum	Amoebozoa
Drosophila melanogaster	Insecta
Entamoeba histolytica	Amoebozoa
Fonticula alba	Holomycota
Mus musculus	Mammalia
Paramecium tetraurelia	Alveolata
Reticulomyxa filosa	Rhizaria
Rozella allomycis	Rozella
Tetrahymena thermophila	Alveolata
Thecamonas trahens	Apusozoa
Trichomonas vaginalis	Excavata

PHAGO-MIXOTROPHS	
Species	Group
Bigelliowiella natans	Rhizaria
Chrysochromulina tobin	Haptophyta
Cymbomonas tetramitiformis	Chloroplastida
Prymnesium parvum	Haptophyta

OSMOTROPHS	
Species	Group
Allomyces macrogynus	Fungi-Blastocladiomycota
Batrachochytrium dendrobatidis	Fungi-Chytridiomycota
Conidiobolus coronatus	Fungi-Zygomycota
Neocallimastix californiae	Fungi-Neocallimastigomycota
Puccinia sorghi	Fungi-Basidiomycota
Rhizophagus irregularis	Fungi-Glomeromycota
Saccharomyces cerevisiae	Fungi-Ascomycota
Schizosaccharomyces pombe	Fungi-Ascomycota

Table 4.1 – Taxa included in the collection of reference genomes for the selection of ∼600 genes predictive for trophic mode. Table adapted from of Burns et al. (2018).

predictive genes in each of the 37 reference genomes. The model tries to minimise the difference between the observed gene count distribution from metaG samples and the gene count distribution inferred from reference genome mixtures, as expressed in the following equation:

$$R = \sum_{k=1}^{N} (D_S^k - \sum_{g \in G} l_g D_g^k)^2 \tag{4.2}$$

where N is the number of genes considered; D_S^k is the distribution of gene k in sample S; G is the set of reference genomes g; D_g^k is the distribution of gene k in genome g; l_g is the weight of genome g, and R is the residue between the gene distributions in sample S and in the reference genomes mixture. The contribution of each reference genome in the mixture that best represented each metaG sample is summed according to its known trophic group to infer the relative proportions of trophic groups present in each metaG sample.

4.2.3 Community response to environmental variables

We analysed the relationship between *Tara* Oceans metadata (Pesant et al., 2015) variation and changes in trophic groups' abundances through Spearman correlations. We selected for results with $R > 0.4$ and $p < 0.05$. The parameters analysed included NO2, NO3, Si, PO4, chlorophyll A, latitude and temperature. The nutrients concentration in station 85 differed significantly from those in the rest of the samples, and we excluded this sampling site from the analyses to keep it from skewing the results. See Figures S4.1 and S4.2 for details on the variation range of the parameters along *Tara* Oceans' stations.

4.3 RESULTS

4.3.1 Distribution of trophic groups

From the original selection of 173 *Tara* Oceans metaG samples screened with the model, 137 of them contained enough predictive genes for reliable estimations and were selected for downstream ecological analyses. Piconano-plankton size fraction was represented with a total of 51 samples in surface (SUR) and 31 samples in Deep Chlorophyll Maximum (DCM), while from nano-plankton size fraction we analysed 35 samples from SUR and 17 samples from DCM.

The prediction of trophic groups abundance revealed that phototrophs dominated in all depths and size fractions, ranging from 30% to 80% in the overall samples (Figure 4.2). Phago-mixotrophs were the second most abundant group, accounting for 10%-40% of organisms in the community, while phagotrophs contributed to 5%-25% of the relative abundance. Osmotrophs were the least abundant organisms in all datasets, accounting for less than 10% of the relative abundance.

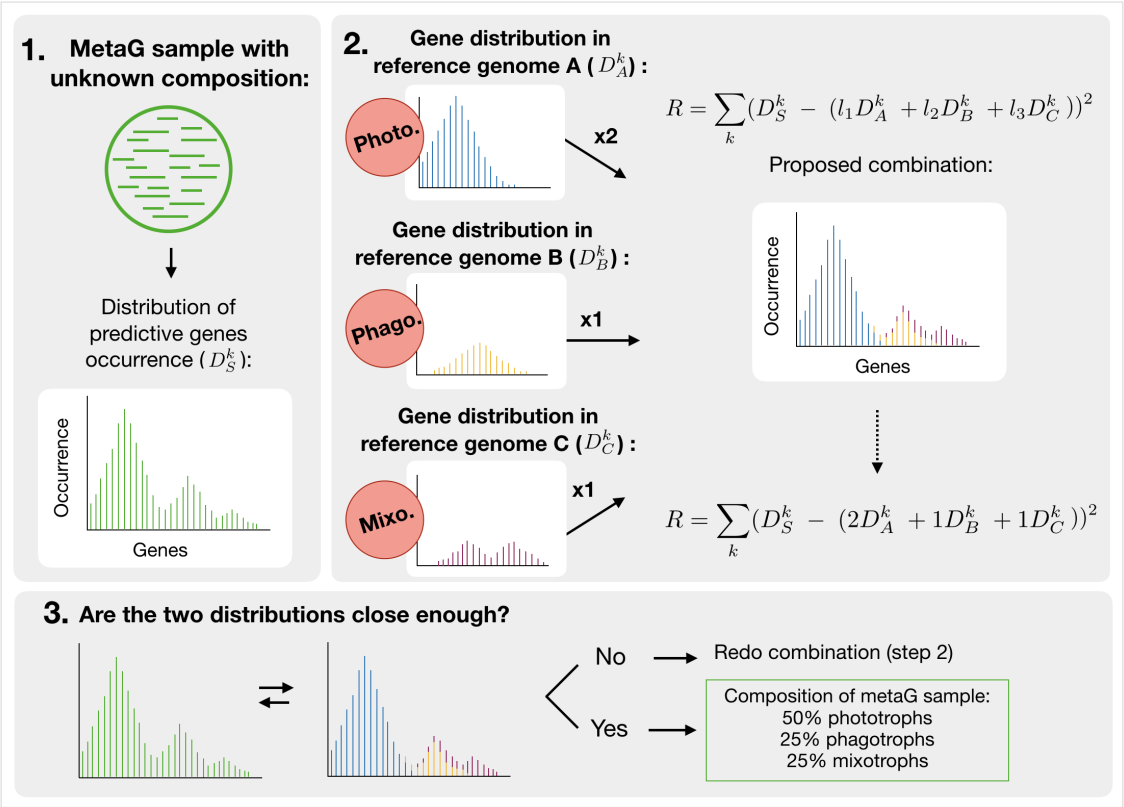


Figure 4.1 – Simplified representation of the steps followed by the likelihood mixture model. Firstly, the model is fed with the occurrence data of ∼600 predictive genes k in a query metaG sample S. Next, the model evaluates the occurrence of the same ∼600 genes in a number of reference genomes. Here, the reference genomes are named A, B and C. Following a maximum likelihood approximation, the model combines the distribution of predictive genes in the reference genomes (D_A^k, D_B^k and D_C^k) to find a configuration similar to the one observed in the metaG sample (D_S^k). In the third step, the model evaluates if the configuration of the gene occurrence in the metaG sample is close enough to the one proposed by combining the gene distribution of the reference genomes. If the two configurations are close enough, the model is able to tell the approximate proportion of the different trophic groups based on the classification of the reference genomes. In this particular example, the reference genomes A, B and C are classified as phototrophic, phagotrophic and mixotrophic, respectively. Phototrophs are predicted to account for 50% of the weight in the metaG sample, while phagotrophs and phagomixotrophs account for 25% of the weight each.

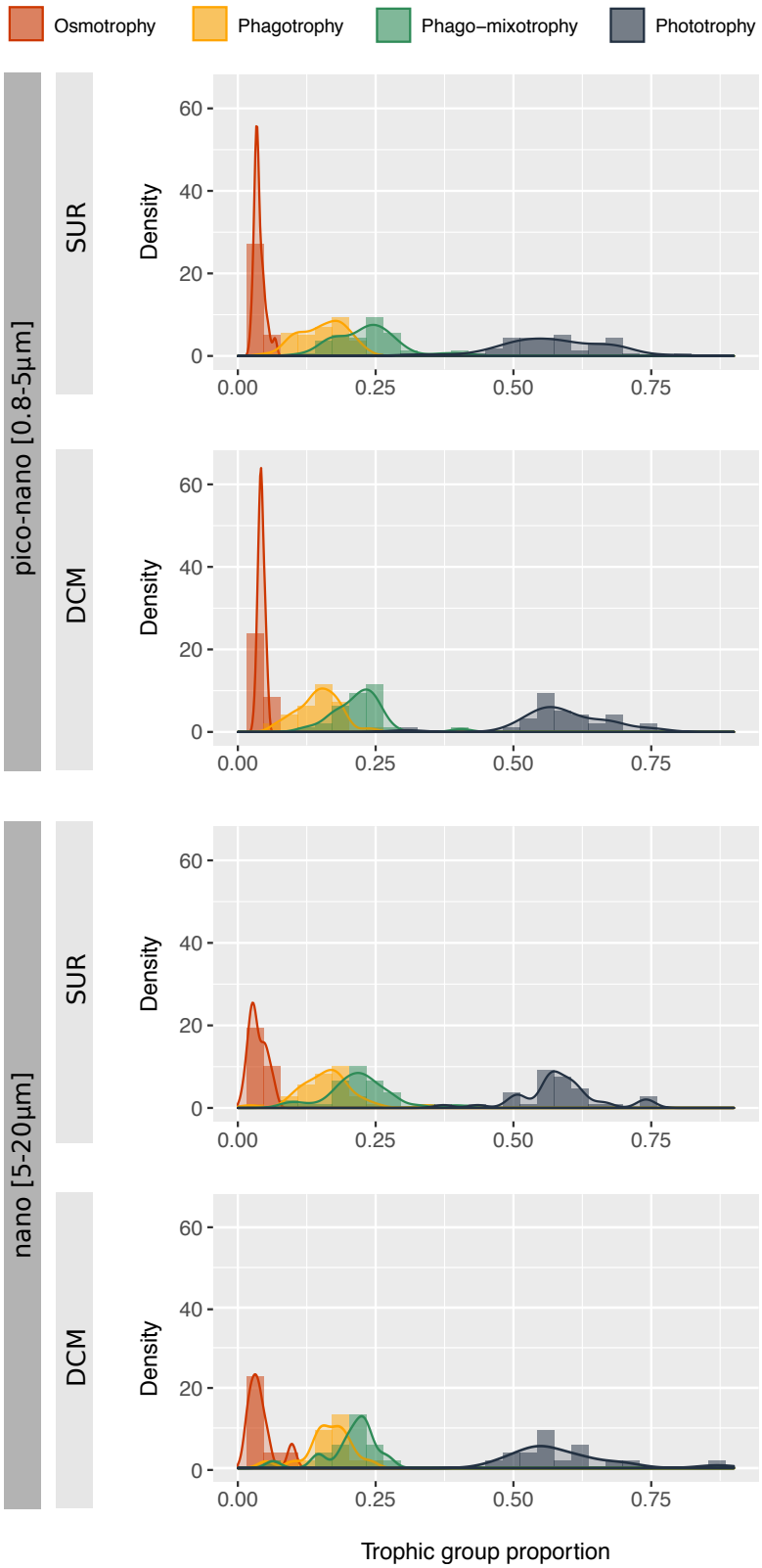


Figure 4.2 – Distribution of trophic groups' relative abundances. The histogram curves are computed through kernel density estimation, a variation of Histogram approach that uses kernel smoothing to visualize distributions over a continuous interval. Data shown by individual datasets, divided by size fraction and depth.

We observed no remarkable differences between group proportions in piconano- and nano- size fractions. Within piconano- samples, only osmotrophs were significantly lower in SUR waters (T-test, $p < 0.01$), in comparison with DCM samples. In the case of nano- size fraction, groups presented comparable abundances between depths.

The samples in the Southern Ocean showed a different composition than the rest of basins in several datasets (Figure 4.2). Piconanoeukaryotes in SUR showed lower relative abundance of phototrophs in the Southern Ocean than in the Mediterranean Sea and the Indian Ocean (Scheffe's Test, $p < 0.01$), while phago-mixotrophs were more abundant in the Southern Ocean than in the rest of basins (excluding the South Atlantic Ocean and the South Pacific Ocean) (Scheffe's Test, $p < 0.05$).

Similarly, in piconano- size fraction and DCM depth, phototrophs had a lower relative abundance in the Southern Ocean than in the rest of basins (Scheffe's Test, $p < 0.05$), while phago-mixotrophs had a higher relative abundance in the Southern Ocean than in the other regions (Scheffe's Test, $p < 0.01$). Additionally, we also observed a significantly lower proportion of phago-mixotrophs in the Mediterranean Sea than in the South Pacific Ocean (Scheffe's Test, $p < 0.05$). Phagotrophic picoeukaryotes in the DCM only showed a significantly higher relative abundance in the Southern Ocean in comparison with the Mediterranean Sea (Scheffe's Test, $p < 0.05$).

Most of the groups in nano-plankton size fraction presented no regional differences; only phagotrophs contributed in higher proportions in the North Atlantic Ocean than in the Indic Ocean (Scheffe's Test, $p < 0.05$).

4.3.2 Correlation between trophic groups distribution and environmental variables

In piconano- size fraction and SUR samples, phago-mixotrophs' abundances correlated positively with the concentrations of NO2, NO3 and PO4 (Figure 4.4), whilst phototrophs‘ relative abundances correlated negatively with these parameters. Phagotrophs and osmotrophs abundances showed weaker correlations with NO2, NO3 and PO4 ($R < 0.5$), which was positive in the case of phagotrophs and negative for osmotrophs. Contrarily, piconano- communities from DCM depth showed no correlation with NO2, NO3 and PO4 (Figure 4.5); only the proportions of phototrophs and phagotrophs were moderately correlated with Si concentrations.

Phototrophs' relative abundance within samples from nano- size fraction and SUR depth seemed to decrease with higher concentrations of NO2, while phago-mixotrophs proportions augmented (Figure 4.6). In the same samples, phagotrophic eukaryotes correlated positively with Latitude and decreased with Temperature. Temperature was the only environmental parameter correlated with changes in groups' abundance from nano- size fraction and DCM depth (Figure 4.7): while phototrophs tended to be more abundant with higher temperatures, phago-mixotrops and phagotrops decreased.

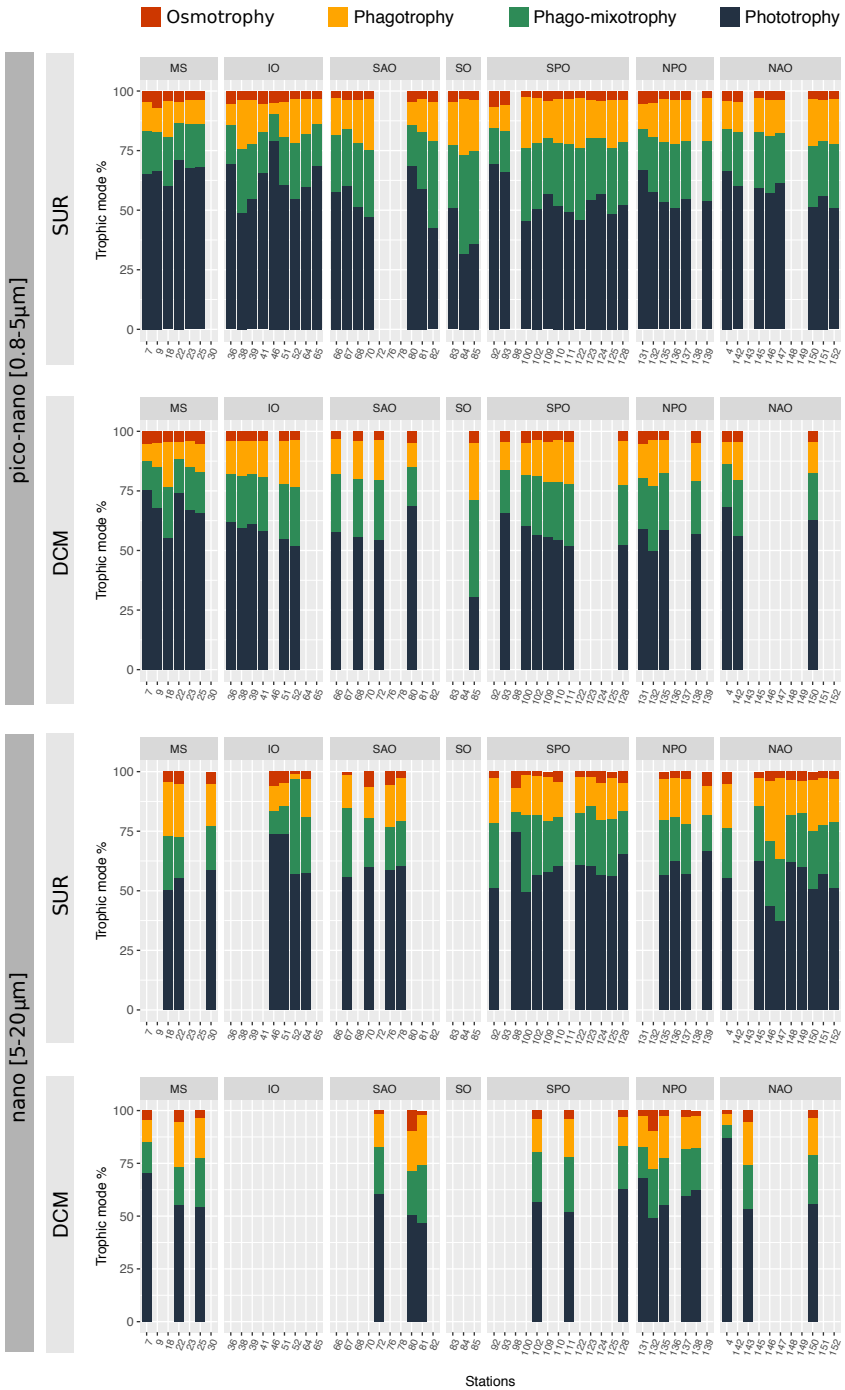


Figure 4.3 – Distribution of trophic groups' relative abundance across samples. Data shown by individual datasets, divided by size fraction and depth. The ocean basins acronyms correspond to the following regions: MS, Mediterranean Sea; IO, Indian Ocean; SAO, South Atlantic Ocean; SO, Southern Ocean; SPO, South Pacific Ocean; NPO, North Pacific Ocean; NAO, North Atlantic Ocean.

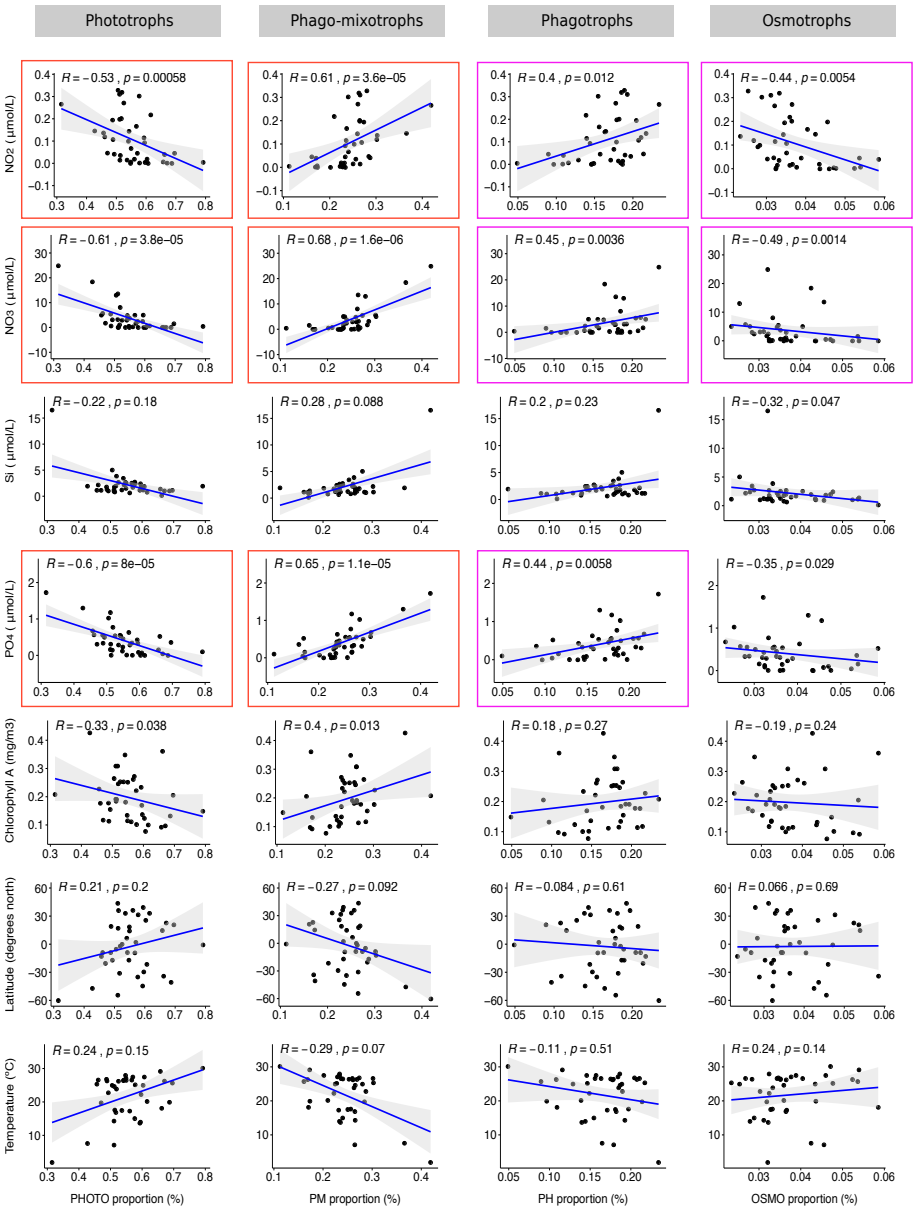


Figure 4.4 – Spearman correlation between trophic group's relative abundance and metadata in piconano- size fraction and SUR depth. Groups' abundances were compared against the following environmental parameters: NO2, NO3, Si, PO4, Chlorophyll A, Latitude and Temperature. Red squares highlight plots with $R \geq 0.5$;pink squares highlight plots with $R > 0.4$ and $p < 0.05$.

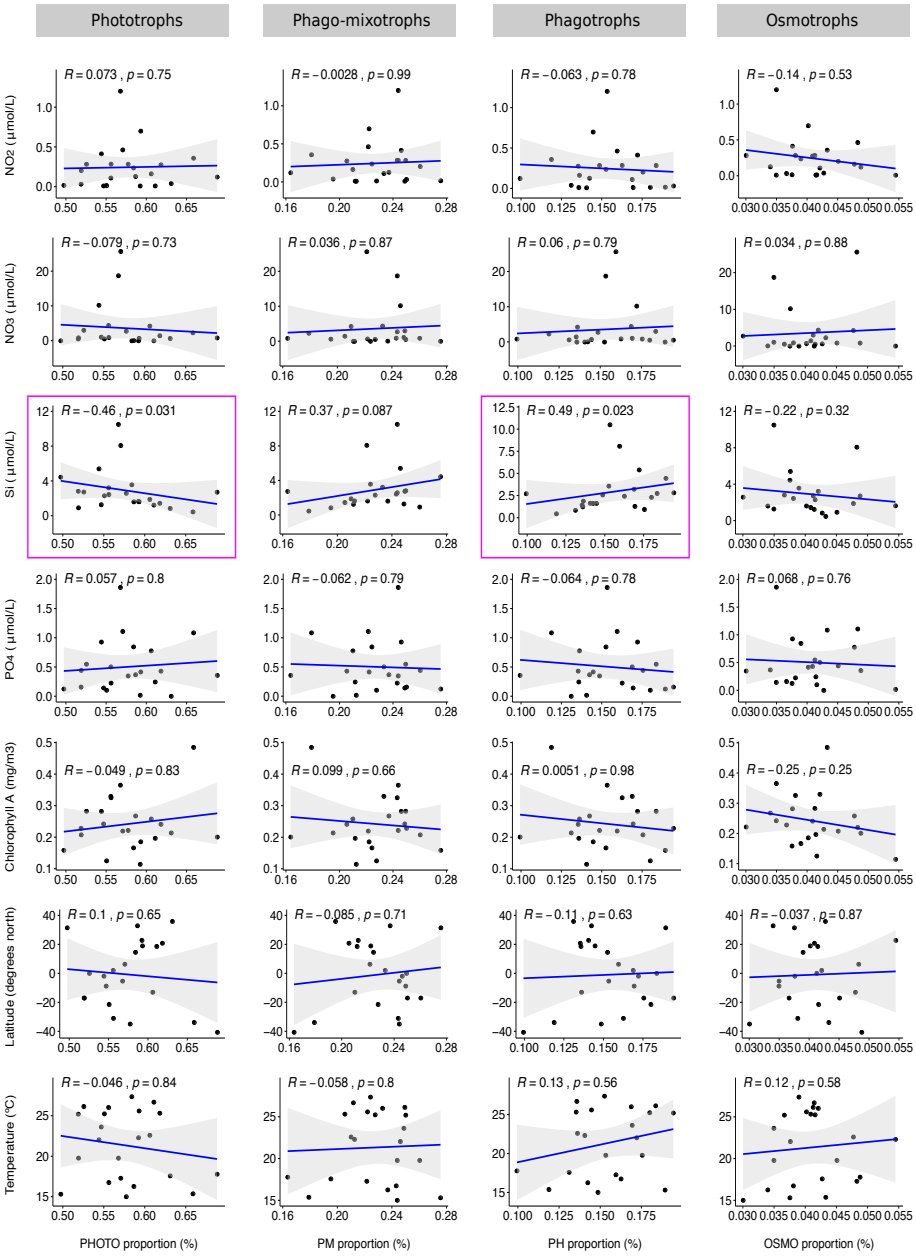


Figure 4.5 – Spearman correlation between trophic group's relative abundance and metadata in piconano- size fraction and DCM depth. Environmental parameters analysed as in Figure 4.4. Pink squares highlight plots with $R > 0.4$ and $p < 0.05$.

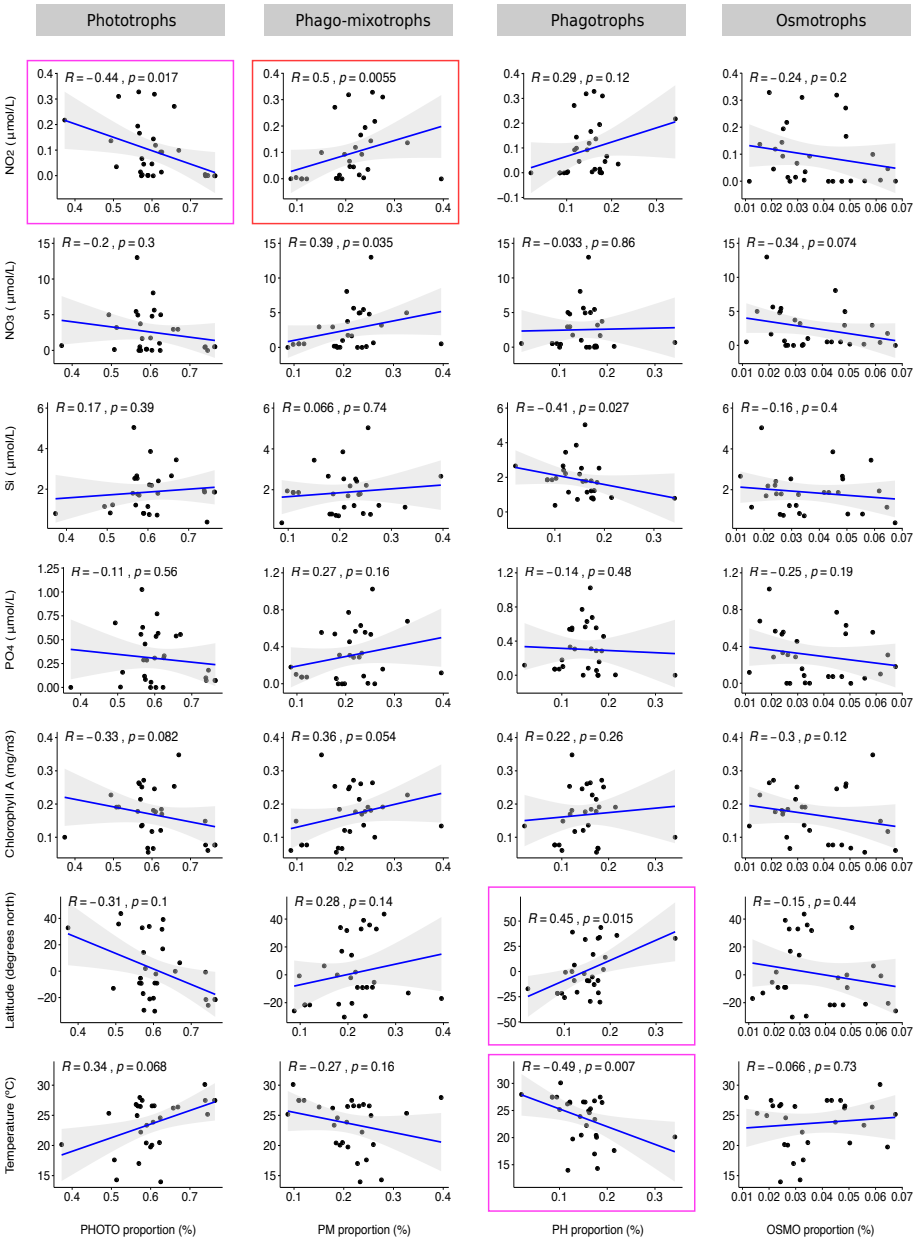


Figure 4.6 – Spearman correlation between trophic group's relative abundance and metadata in nano- size fraction and SUR depth. Environmental parameters analysed as in Figure 4.4. Red squares highlight plots with $R \geq 0.5$; pink squares highlight plots with $R > 0.4$ and $p < 0.05$.

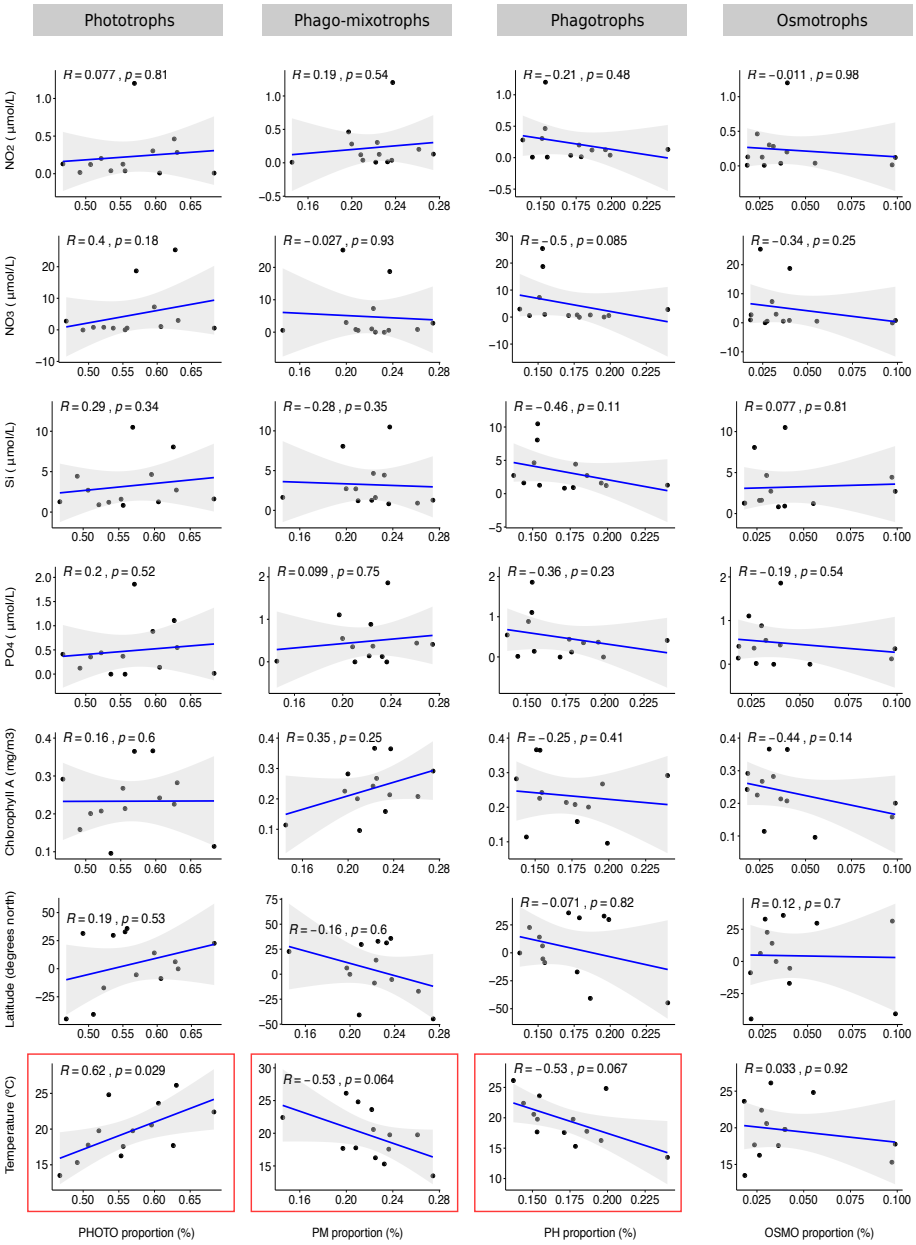


Figure 4.7 – Spearman correlation between trophic group's relative abundance and metadata in nano- size fraction and DCM depth. Environmental parameters analysed as in Figure 4.7. Red squares highlight plots with $R \geq 0.5$.

4.4 DISCUSSION

The model described in this study predicts dominance of phototrophs in piconano- and nano-plankton size fractions, at both SUR and DCM depths. Phago-mixotrophs and phagotrophs accounted for approximately half of the organisms in all eukaryotic communities. Osmotrophs, here understood as taxa that could not be classified as photo-/phagomixo-/phago-trophic, had a lower relative abundance in piconano- samples from SUR depth than in the rest of the datasets. This could indicate that piconano- samples from SUR depth contained a lower proportion of purely osmotrophic organisms like fungi. Yet, it could also mean that the model recognized a larger proportion of protists as photo-/phagomixo-/phago-trophs in these samples because their gene composition was closer to those of the photo-/phagomixo-/phago-trophic reference genomes from the training set.

Despite the overall contribution of the trophic groups being comparable between size fractions and depths, we observed different regional patterns among datasets. A higher relative abundance of picoeukaryotic phago-mixotrophs and a decrease of picophototrophs in the Southern Ocean appeared in both SUR and DCM layers. The dominance of phago-mixotrophs in these samples could be explained by community changes related with seasonal succession. Upwellings shape plankton ecosystems in Polar seas, and while they encourage the blooming of phototrophs and phagotrophs at the beginning of the community transition, the mature stages of the ecosystems are more advantageous for mixotrophs due to the decreasing availability of nutrients and abundance of prey (Mitra et al., 2014; Stoecker and Lavrentyev, 2018). *Tara* Oceans samples in the Southern Ocean were collected during early January, (i.e. towards the end of the Antarctica's summer), consistently with the higher relative abundance of mixotrophs expected when the ecosystem has transitioned to a mature state.

In the nano-plankton size fraction, the only group displaying regional differences were phagotrophs, showing lower relative abundances in the Indian Ocean than in the North Atlantic Ocean. Phagotrophs' relative abundance in this size fraction correlated negatively with temperature, but we do not expect this environmental parameter to explain the low frequencies of phagotrophs in the Indian Ocean, since high temperatures tend to favour growth rates of heterotrophs (Rose and Caron, 2007). The fact we only identified differences between two basins in nano- size fraction could be influenced by the low number of samples analyzed from this community.

Nutrients concentration affects biomass and productivity of all trophic groups either as direct food source in the case of phototrophs and phago-mixotrophs or by influencing prey availability in phago-mixotrophs and phagotrophs. Our results revealed differential relationships of environmental parameters in trophic groups' proportions among datasets. The increase of phagotrophs' and phago-mixotrophs' relative abundance in the piconano- size fraction correlated with concentrations of NO2, NO3 and PO4 in SUR depth. In the DCM layer, however, the proportion of phagotrophs' relative abundance was only moderately correlated with silica concentrations while

the proportion of phago-mixotrophs' relative abundance was not significantly associated with any environmental variable. In the nano- size fraction, temperature correlated with photo-, phagomixo-, and phago-trophs' relative abundances; phototrophs and phago-mixotrophs showed correlations with NO2 in SUR depth, and phagotrophs' relative abundance increased with latitude. Among all environmental parameters analysed, only NO2 and temperature correlated with trophic groups' frequencies in similar ways across different datasets. The disparity of correlations between trophic group's abundance and environmental parameters among datasets suggests that either some of these correlations are not strictly causative, and/or that the communities in the different size fractions and depths respond to environmental changes in different ways.

Based on theoretical trophic modelling, Ward and Follows (2016) observed that phago-mixotrophy could be an advantageous trophic strategy in environments with limiting nutrient concentration, particularly for nano-plankton. According to this study, phago-mixotrophs would be favoured in oligotrophic waters due to their ability to support and supplement photosynthesis through prey ingestion. Contrarily, our results pointed out that phago-mixotrophic and phagotrophic picoeukaryotes present a higher relative abundance than pure phototrophs in samples rich in inorganic nutrients. As commented above, the higher relative abundance of phago-mixotrophs and phagotrophs in these samples could be related with seasonal blooming, but enhanced grazing functionality could also be triggered by the presence of inorganic nutrients. Remarkably, the relationship between nutrients and trophic modes is absent or weaker in DCM and large size fractions.

Overall, the similar frequencies of trophic groups between size fractions and depths indicates that, despite the expected variation in taxonomic and functional diversity among datasets (de Vargas et al., 2015; Carradec et al., 2018), all communities maintain a remarkably stable trophic structure. However, the similarity in trophic composition does not necessarily reflect equal contribution of groups to the food network. For instance, mixotrophs present comparable frequencies across samples, but these organisms can follow very different nutritional strategies, favoring phototrophy or heterotrophy depending on their physiology and environmental conditions (Mitra et al., 2016).

In this study we show that trophic groups' frequencies in metaG samples can be quantified through gene frequencies. The obtained predictions indicate that the relative abundance of the different trophic groups is comparable between size fractions and depths, with few changes among oceanic regions. The environmental parameters analysed here did not show consistent correlations among the composition of communities from different sizes and depths. The predictions of our model could be inaccurate if the set of reference genomes does not represent the complete genomic signature of the different trophic groups; therefore, diversifying and complementing the training set with further reference genomes could help to reduce possible biases. Expanding the analyses with further environmental variables and additional samples could help explain the variability of trophic groups' frequencies among communities. Future work should also incorporate measurements of functional activity, so that we

can better couple variations in community composition with trophic contribution in the ecosystems.

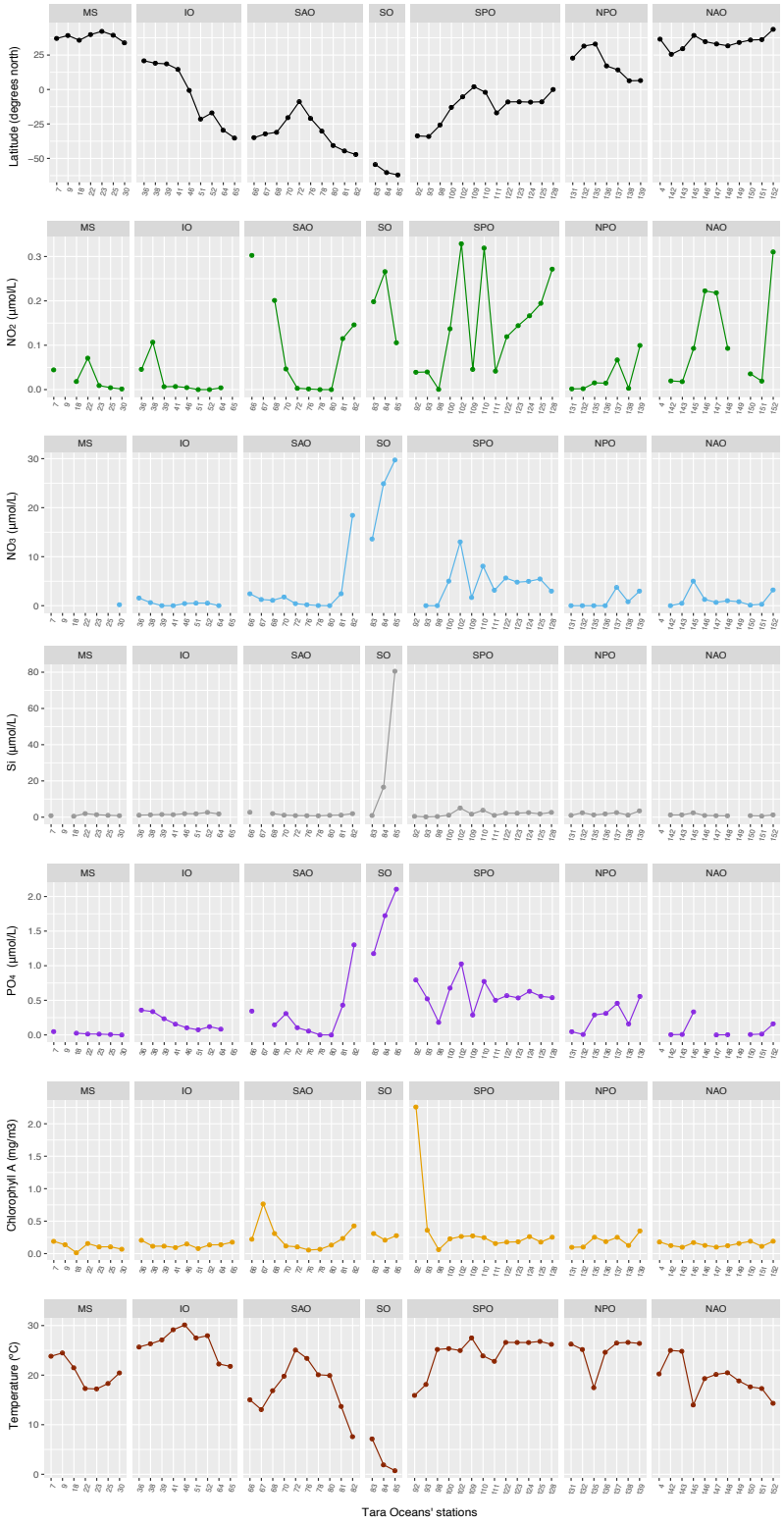


Figure S4.1 – Measurements of environmental parameters in SUR depth across *Tara* Oceans stations, including concentrations of NO2, NO3, Si, PO4 and Chlorophyll A; Temperature, and Latitude.

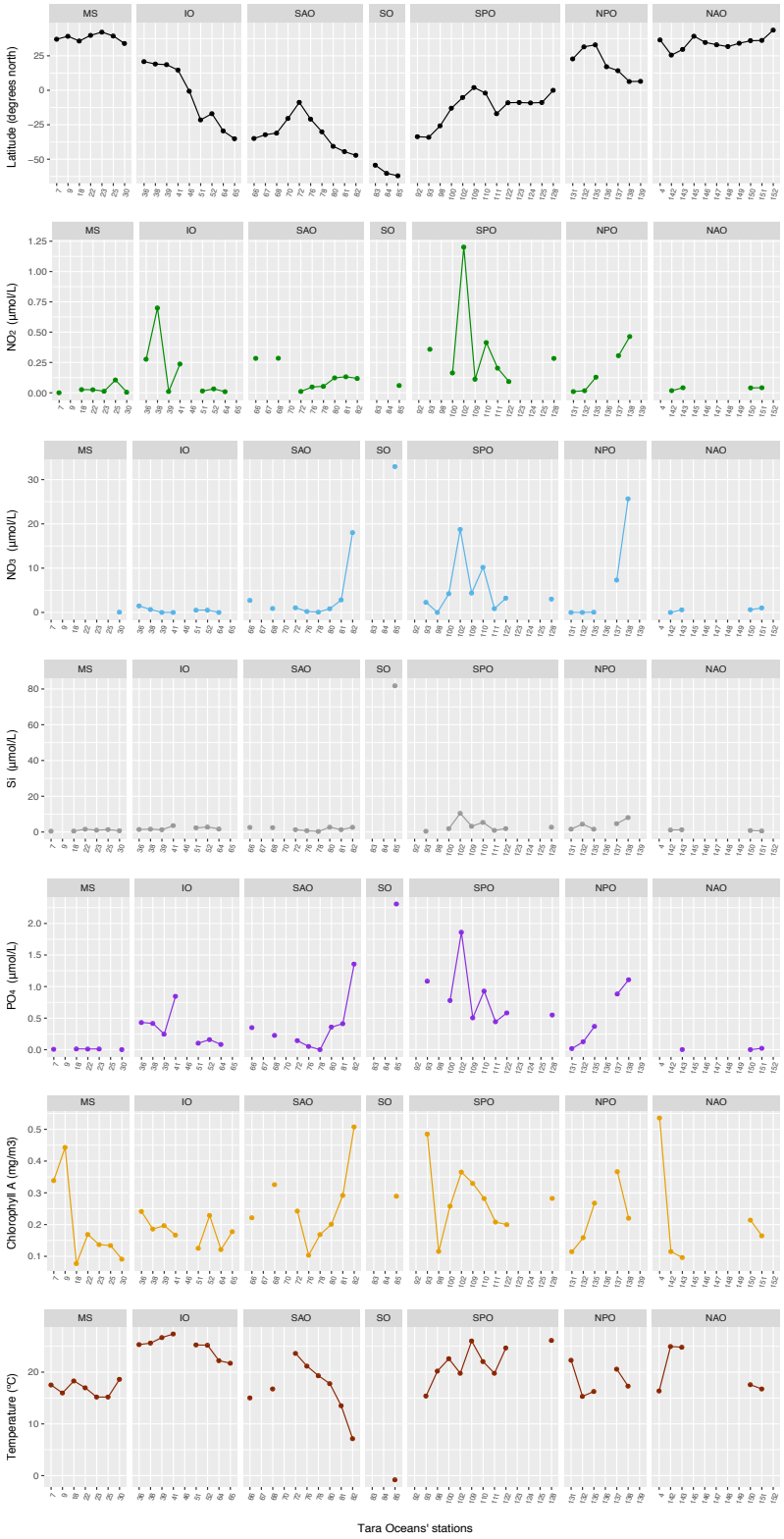


Figure S4.2 – Measurements of environmental parameters in DCM depth across *Tara* Oceans stations. Same variables as in Figure S4.1.

5. Assessment of single environmental cell sequencing for the obtention of genomes from the dominant protists in the ocean

Traditionally, the obtention of microbes' reference genomes has relied on the sequencing of cells grown *in vitro*. This is radically changing with the development of single-cell genomics, a technique that allows the sequencing of genomes or transcriptomes from individual environmental cells (e.g. Pachiadaki et al. 2019). Single-cell sequencing has an enormous potential in marine protistology, since the majority of taxonomic groups of microbial eukaryotes are unknown or cannot be grown in cultures. In this chapter, we explore the extent of the taxonomic diversity of marine protists covered by a collection of 903 Single-cell Amplified Genomes (SAGs) generated randomly from a few cryopreserved *Tara* Oceans samples. We show that by sequencing only a few hundred SAGs from 8 marine plankton samples it is possible to cover a large fraction of the most abundant and cosmopolitan taxa in the open ocean. Part of the results described in this chapter were integrated in the publication of Sieracki et al. (2019), which I co-authored and can be found in Appendix A.1.

5.1 Introduction

The use of molecular data in microbial ecology surveys is severely constrained by the lack of reference genes and genomes for taxonomic annotation. In particular in the case of microbial eukaryotes, the diversity of reference genomes available is still extremely limited (del Campo et al., 2014). The first large-scale effort to cover this gap started with the generation of reference eukaryotic transcriptomes in the frame of

the The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling et al., 2014). MMETSP assembled and annotated ∼650 transcriptomes from cultures of marine microbial eukaryotes. MMETSP cultures provide the first integrated, cross-eukaryotes information for gene annotation and activity analyses. Yet, a tiny minority of protists (mostly phototrophic) can be maintained in cultures and not all genes are equally expressed in culturing vs. natural conditions; thus, it is essential to complement these reference transcriptomes with genome sequencing from unculturable taxa.

Sieracki et al. (2019) did another step forward towards the expansion of protists' reference genomes databases by generating a collection of 903 single-cell amplified genomes (SAGs) from *Tara* Oceans cryopreserved samples . Sieracki et al. (2019) showed that the diversity of groups obtained through single-cell sequencing was comparable to that observed through metabarcoding, indicating that the technique is not biased against any phylogenetic group. Using the same SAG collection, here we assess the potential of single-cell technique to cover the diversity of protists in the global open ocean.

5.2 Methods

As described in Sieracki et al. (2019), the collection of 903 *Tara* Oceans eukaryotic SAGs was obtained from 8 stations located in the Adriatic Sea, the Arabian Sea and the Indian Ocean (Figure 5.1). Whole water subsamples of 4 mL were resuspended in glycine betaine, and cryopreserved on board *Tara* and in the laboratory before their molecular analysis. Individual cells measuring <5μm were then sorted in microplates by flow-cytometry, and their genomic DNA was amplified through multiple displacement amplification (Sieracki et al., 2019). The amplified genomic DNA from each SAG was used as a template for 18S rDNA amplicon sequencing to check their taxonomic identity. In our study, in order to compare the taxonomy and diversity of SAGs to that of MMETSP cultures and *Tara* Oceans eukaryotic metabarcodes, we retrieved the 18S rDNA V9 fragment from the 903 SAGs using the sequences from the primers 1389F 5'- TTGTACACACCGCCC -3' and 1510R 5'- CCTTCYGCAGGTTCACCTAC -3' (Amaral-Zettler et al., 2009), and we recovered a total of 868 sequences. Similarly, we fetched the V9 from 537 MMETSP transcriptomes by screening their 18S rDNA with the same primers, and retrieved a total of 385 V9 rDNA sequences from the MMETSP database.

We explored the distribution of the *Tara* Oceans SAGs and MMETSP transcriptomes across the world ocean by mapping their 18S V9 rDNA against the 18S rDNA V9 amplicons from *Tara* Oceans (see de Vargas et al., [2015] for details about data sampling and sequencing). Overall, we selected 337 *Tara* Oceans metaB samples corresponding to piconano- (0.8-5μm) plankton size fractions. These samples represented a total of 105 stations and were obtained from both SUR and DCM water depth. The metaB dataset contained 435,240,095 reads. We clustered the reads into OTUs using SWARM 2.1.1 (Mahé et al., 2015), generating a total of 271,787 OTUs. MMETSPs' and *Tara* Oceans SAGs' V9 rDNA sequences were then blasted

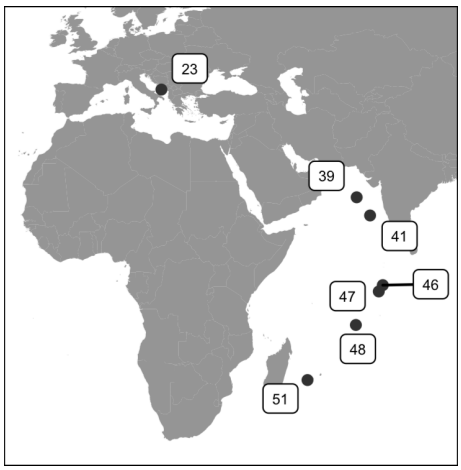


Figure 5.1 – Geographical location of the *Tara* Oceans stations at which the 903 SAGs were sampled.

on these OTUs ranked by their abundance and occurrence. We filtered the results by >97% similarity and >80% coverage and obtained a total of 671 SAGs and 212 MMETSP mapped on the collection of oceanic OTUs.

In order to assess the extent to which SAGs and MMETSP represent the taxa in the global oligotrophic ocean, we then mapped the 18S rDNA OTUs from *Tara* Oceans metaB on SAGs and MMETSP. The mapping was done using BLAST and the results were filtered by 80% coverage. For each *Tara* Oceans station, we aggregated the total amount of reads and OTUs mapped at 80-85%, <90%, <95%, <97%, <99%, <100%, 100% identity thresholds.

The taxonomic annotation from SAGs and MMETSP was obtained by mapping their V9 rDNA against the PR2 database (Guillou et al., 2013) at >97% similarity and ≥70% coverage. Additionally, we also used the published annotation of SAGs based on V4 rDNA mapping against MAS and NCBI databases (Sieracki et al., 2019), together with the original annotation from *Tara* Oceans metaB OTUs (de Vargas et al., 2015).

5.3 Results

The mapping of *Tara* Oceans SAGs against *Tara* Oceans 18S rDNA OTUs (Figure 5.2-A) showed that SAGs represent a considerable number of the most abundant and cosmopolitan taxa in the world oligotrophic ocean. At the same time, SAGs also matched a few OTUs identified as globally rare. In comparison with the *T*O SAGs snapshot, the number of MMETSP cultures represented in the 18S rDNA dataset was much lower (Figure 5.2-B). MMETSP matched few of the dominant OTUs and were spread across the gradient of OTUs' occurrence and abundance.

The frequency of SAGs and MMETSP in the individual samples from *Tara* Ocean was explored in detail in Figure 5.3 (available at `https://figshare.com/articles/`

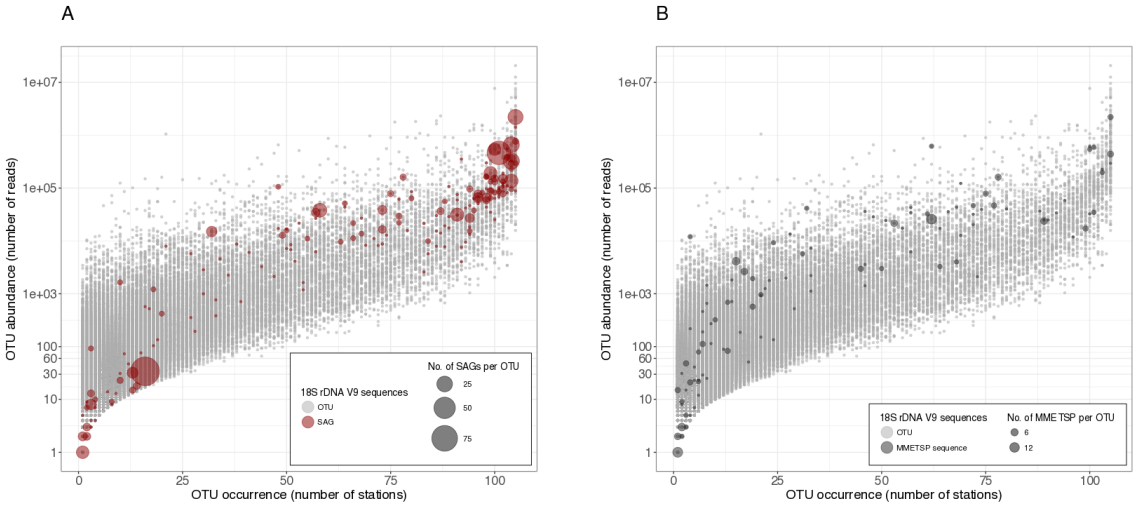


Figure 5.2 – Abundance and occurrence of (A) SAGs and (B) MMETSP across Tara Oceans stations. SAGsand MMETSP (coloured bubbles) are mapped on Tara Oceans 18S rDNA V9 OTUs (grey dots). The size of the bubbles represents the number of mapped SAGs or MMETSP.

`Figure_5_3_Inference_of_SAGs_and_MMETSP_relative_abundance_across_Tara_Oceans_stations_based_on_the_distribution_of_the_OTUs_to_which_they_have_been_assigned_/11342576` due to its large size). As expected, the stations that were better represented in the SAGs collection were those around the Indian Ocean, while the sampling sites from the Southern Ocean were the ones presenting lower relative abundance of V9 rDNA amplicons matching SAGs and MMETSP. The SAGs and MMETSP showing higher abundance and widespread distribution belonged essentially to the following groups: MAST (MArine STramenopiles), MALV (MArine ALveolates) Pelagophyceae, Prymnesiophyceae, Telonemida, Chrysophyceae, Mamiellophyceae and Dictiochophyceae.

Furthermore, the amount of 18S V9 rDNA amplicons matching SAGs at 100% identity was 7.4% per sample, on average (Figure 5.4-A). In the case of MMETSP (Figure 5.4-B), the number of reads mapped with the same similarity threshold was significantly lower (4.3% on average). When decreasing the identity threshold of the mapping, we obtained similar relative abundances of reads mapped to SAGs and MMETSPs. In most of the stations, ∼25% of the 18S rDNA reads matched SAGs and/or MMETSPs at ≥97% identity, and around 50% mapped at ≥90% identity.

Regarding richness, the amount of OTUs mapped at 100% against SAGs accounted for ∼1.5% of the OTUs in each sample, on average, and was significantly higher than the amount of OTUs matching MMETSP at 100% identity (∼0.6% of reads per sample) [Figure 5.5]. The amount of OTUs with matches at lower similarity thresholds was slightly higher when mapped against SAGs than against MMETSP, but overall the comparison against the two datasets showed similar trends of OTU aggregation across stations. Most of the stations contained a minimum of 50% of OTUs mapped at ≥85% similarity on some SAGor MMETSP.

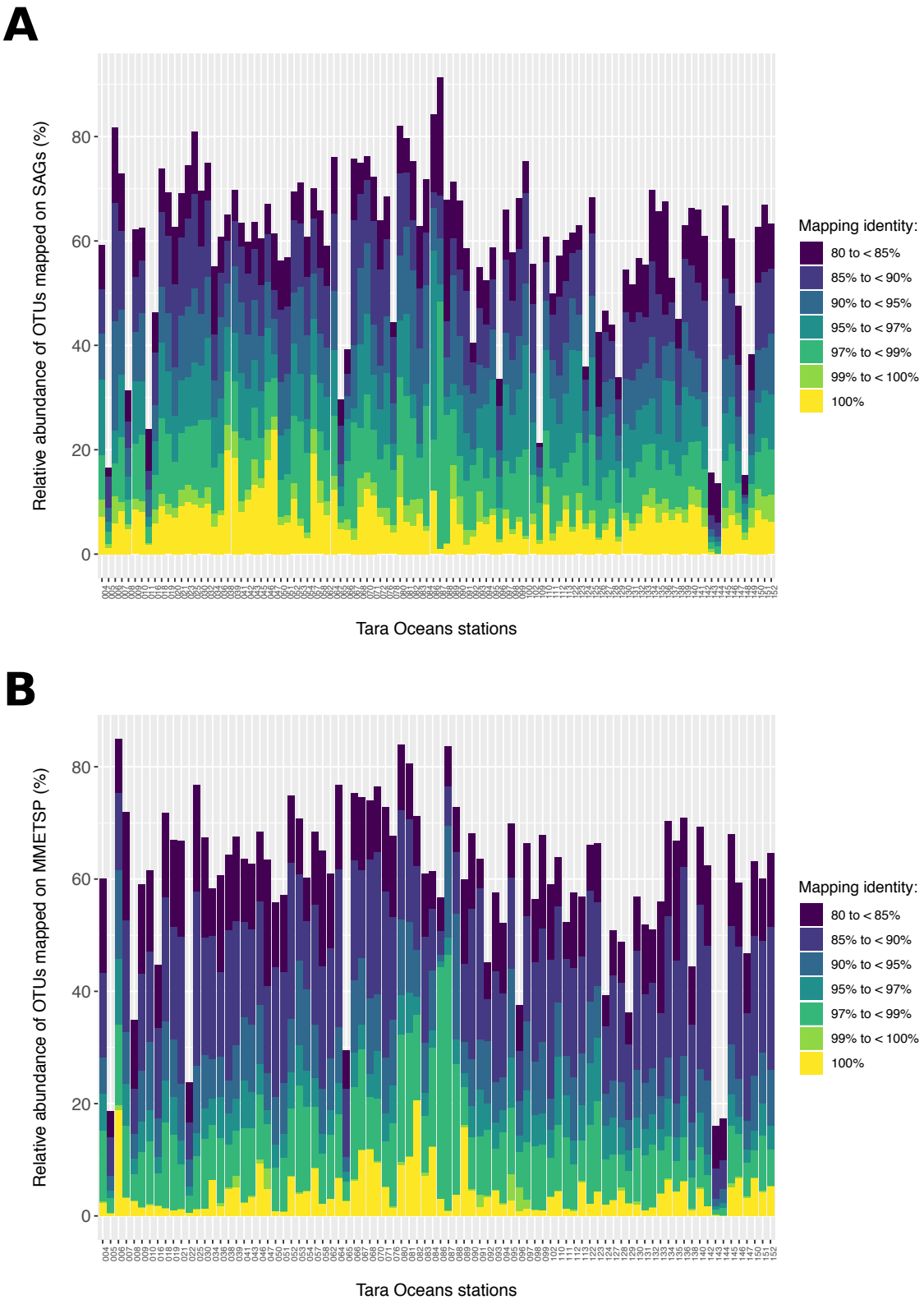


Figure 5.4 – Accumulated relative abundance of OTUs from *Tara* Oceans stations with a match in (A) SAGs and (B) MMETSP collections. For each sample, we represent the accumulated proportion of OTUs with matches at different identity thresholds (80-85%, <90%, <95%, <97%, <99%, <100%, 100%).

5.4 Discussion

The relatively few *Tara* Oceans SAGs analysed from a handful of Tara Oceans stations represented many of the most abundant and cosmopolitan OTUs in *Tara* Oceans metaB dataset, indicating that single-cell sequencing is potentially an extremely powerful approach for recovering the genomes of dominant taxa in the ocean. In contrast, MMETSP did not seem to represent as widely the dominant contributors of protists in open ocean communities.

In the publication of MMETSP dataset (Keeling et al., 2014), the authors remark that the collection is mainly composed of phototrophic organisms due to the difficulty of maintaining heterotrophic organisms in cultures. They also predict that the advent of single-cell sequencing will help correcting this bias, diversifying the trophic diversity of protists' reference genomes available. Indeed, some of the most abundant SAGs in our collection corresponded to taxa classified as heterotrophic (e.g. MAST and Telonemida), mixotrophic (e.g. Prymnesiophyceae, Chrysophyceae and Dictiochophyceae) or parasitic organisms (e.g. MALV), confirming that single-cell sequencing is a valid and critical alternative for accessing the genomes of uncultured micro-eukaryotes that make up the large majority of oceanic protists (de Vargas et al. 2015).

In our results we identified that a number of SAGs as locally abundant and globally rare, manifesting the influence of geographic structuring in the composition of SAGs collection. Future efforts should focus on diversifying the collection of *Tara* Oceans SAGs by sorting cells from different oceanic regions and environments.

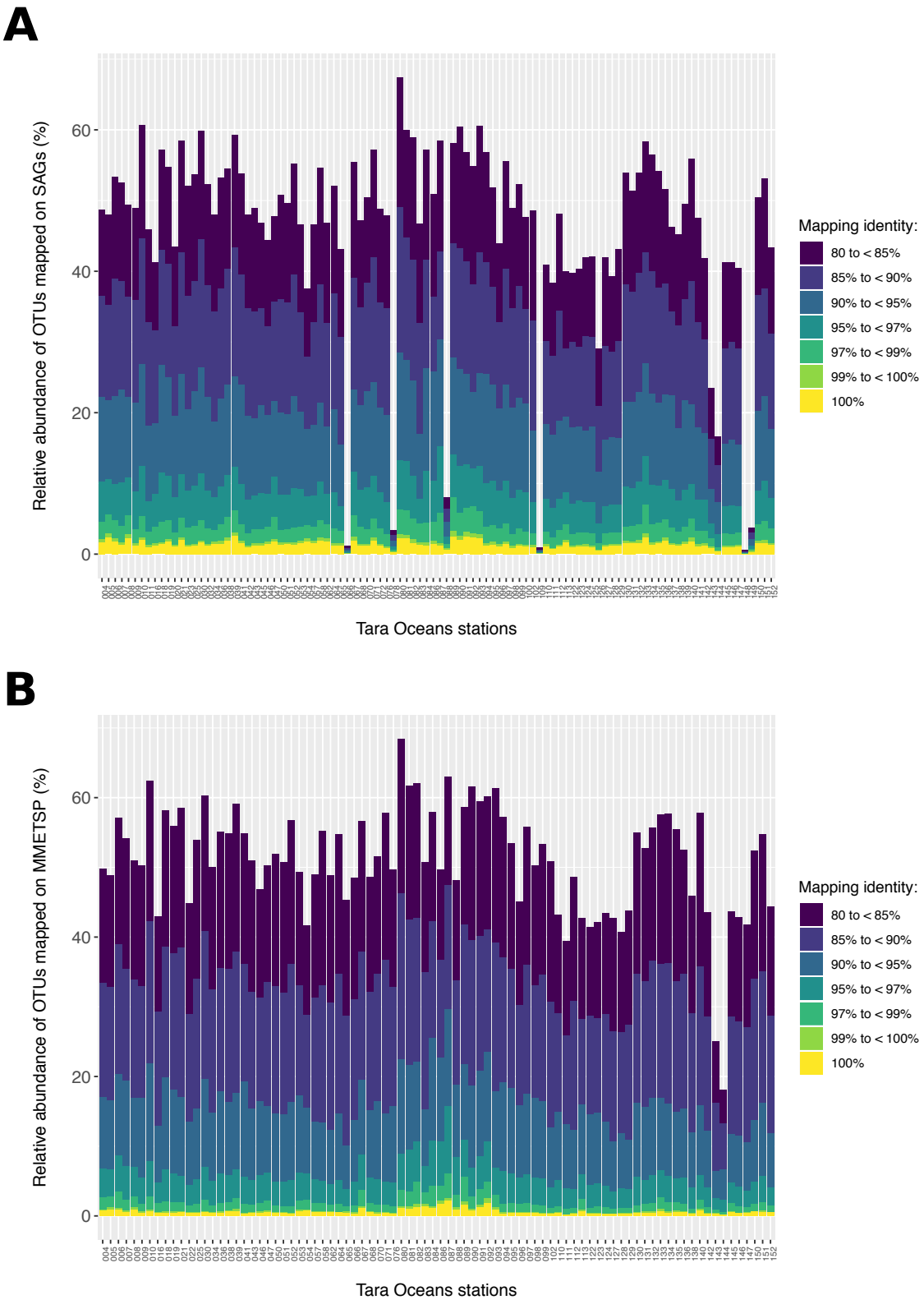


Figure 5.5 – Accumulated relative richness of OTUs from *Tara* Oceans stations with a match in (A) SAGs and (B) MMETSP collections. Similarly as in Figure 5.4, f or each sample, we represent the accumulated richness of OTUs with matches at different identity thresholds.

6. General Discussion

6.1 Summary of the main results

In this thesis we have explored the diversity and distribution of microbial eukaryotic primary producers inhabiting the global oligotrophic ocean. For this purpose, we have analysed different types of molecular data, including metabarcodes, metagenomes and single-cell genomes. Our work contributes to better understand the composition of marine pico/nano phototrophic and mixotrophic communities by means of taxonomic mapping, phylogenetics, as well as trophic modeling .

In Chapter 2, we described PhotoRefT and the pipeline used for picophototrophs phylogeny-based assessment. PhotoRefT is a 16S rDNA reference tree containing sequences from the main known groups of picophototrophic eukaryotes and Oxyphotobacteria. This tree can be used as a support for the taxonomic annotation of 16S rDNA metaB and metaG reads through phylogenetic placement. The annotation pipeline was implemented in Chapter 3 for the identification of Oxyphotobacteria and eukaryotic plastid sequences from marine communities.

In Chapter 3, we analysed 16S rDNA metaB and metaG data from over 50 stations and observed that Oxyphotobacteria are three-fold more abundant than PPEs plastids in surface waters of the tropical and temperate ocean. Within the PPE plastidial community, Prymnesiophyceae, Mamiellophyceae, Pelagophyceae and Dictyochophyceae were the dominant groups in terms of abundance and richness. PPE plastids accumulated higher evolutionary diversity than Oxyphotobacteria and showed differences in PD and MPD magnitude between groups. The predictions obtained for picophototrophs taxonomic profiling using 16S rDNA 515yF-926R were comparable to those yielded by 16S rDNA and 18S rDNA metaG. Overall, our results supported the use of 16S rDNA metaG and metaB sequencing as trans-domain, unifying methods for the assessment of picophototrophs.

In Chapter 4, we described the results of a new mathematical model to predict the trophic diversity of pico- and nano-eukaryotic communities out of metagenomic

samples. The approach is based on the quantification of genes identified as enriched in photo-, phagomixo-, phago- and osmo-trophs in individual samples and further comparison against gene frequencies in reference genomes. This model was used to screen *Tara* Oceans metaG datasets. The results showed remarkably comparable and stable proportions of trophic groups across size fractions and depths. We observed regional variations in trophic groups distribution and differences in communities' response to environmental variables. This work provides the first taxonomy-free assessment of pico- and nano-eukaryotes' trophic structuring.

In Chapter 5, we tested the potential of single-cell sequencing to obtain the genomes from abundant marine microeukaryotes. We compared a set of 903 SAGs with a collection of 18S rDNA oceanic OTUs and observed that a fraction of the SAGs mapped highly abundant and cosmopolitan taxa. In comparison, the genomes obtained from MMETSP transcriptomes represented less extensively the community of piconano- and nano- microeukaryotes in the open ocean.

6.2 Potential and limitations of phylogeny-based community assessments

The increasing generation of sequencing data from environmental samples has allowed the inference of phylogenies for many microbial groups. The use of phylogenies for biodiversity studies provides information about species' evolutionary and phenotypic relationships that cannot be predicted through sole taxonomic description of microbial communities.

The need for building larger reference trees faces scalability challenges: as the number of taxa in trees increases, the running time for their construction grows exponentially. The development of **faster and more scalable tools for tree inference** will be key to meet the demand for larger phylogenies. On the other hand, building larger trees will require **longer alignments to keep up tree inference accuracy**. In this regard, multigene trees and long read sequencing will become valuable tools to obtain information enough for inferring larger phylogenies.

Methods for environmental sequences' annotation based on phylogeny can produce more accurate results than traditional approaches relying on sequence similarity (Berger et al. 2010). This idea motivated the construction of PhotoRefT and the use of phylogenetic placement to explore picophototrophs' communities. Although PhotoRefT has proven useful, further efforts to improve the tree should focus on **expanding the diversity of references covering picoeukaryotic phototrophs**. The databases of plastidial 16S rDNA sequences are still limited (Decelle et al., 2015) and do not encompass the diversity of important groups like Dinophyceae. This gap of information could be quickly cut down using cultivation-independent tools like long read and single-cell sequencing (e.g. Jami et al., 2019; Pachiadaki et al., 2019) for the obtention of complete ribosomal genes from environmental samples.

6.3 Next steps in the exploration of trophic diversity through modelling

The characterization of plankton food-webs is an essential step towards the understanding of ocean ecosystems and biogeochemical cycles. Until now, several studies have attempted to model plankton contribution to energy transfer across trophic levels (e.g. Stickney, Hood, Stoecker, 2000; Mitra et al., 2014; Ward and Follows, 2016). These approaches rely on simplified, theoretical pictures of the system's biocomplexity and would greatly benefit from more accurate information about the actual, measured trophic composition of planktonic communities.

The model described in Chapter 4 allows a **taxonomy-free assessment** of trophic structuring. This model will improve accuracy as more reference genomes representing well-defined trophic groups become available. The collection of reference genomes used by the model is fairly complete for phototrophs and phagotrophs, but **the quantification of phago-mixotrophs would strongly benefit from adding further genomes from this group in the training set, and eventually sub-divide them by effective trophic strategy.**

In a recent study, Ward and Follows (2016) compared different numerical simulations of marine plankton and predicted a shift from autotrophic to heterotrophic nutrition with increasing size fraction. In contrast, our survey did not reflect compositional differences between size fractions. Future work should focus on **complementing our predictions with phenotypic information**. Including information about prey ingestion rates and phototrophic activity would allow to better compare trophic groups' abundance with their contribution into energy and mass flows.

Finally, modeling approaches require validations in which their predictions are compared to real-world data. In our work, we did a first attempt to validate the model described in Chapter 4 by comparing its output with that obtained from binned genomes (also named "metagenomic based transcriptomes" or MGTs). The procedure consisted on classifying the MGTs obtained from *Tara* Oceans samples (Vorobev et al., 2019) as photo-, phago- and mixo-trophic using the model from Burns et al. (2018). Later, we compared their contribution (as photo-, phago- and mixo-trophs) in each sample with the relative abundance of each trophic group predicted by our model. However, only 80 out of 924 MGTs could be classified into some trophic group due to low genome coverage, representing less than 10% of the relative abundance in most of the samples. Since the coverage and abundance of the MGTs is still not high enough for a proper validation, we are planning to test the model using synthetic mock-communities as an alternative. This approach would consist of mixing known proportions of cultures from strains representing the different trophic modes and sequencing the metagenome of the resultant populations. Next, we would feed the model with the obtained metaG data to test if its predictions agree with the trophic composition of the culture mixtures.

6.4 Advent of single-cell sequencing

Single-cell genomics and transcriptomics have great potential to generate reference genomes from unicellular eukaryotes, to study cell-to-cell interactions, and to explore genome and transcriptome heterogeneity within species (e.g. Mangot et al., 2017; Yoon et al., 2011; Liu et al., 2017). Because of the apparent homogeneity of microbial communities across the temperate and tropical oceans, **this approach seems to be able to recover the most abundant taxa in these ecosystems by sorting a relatively low number of samples** (Sieracki et al., 2019; Pachiadaki et al., 2019).

However, the use of single-cell sequencing still suffers from important technical limitations when applied to eukaryotic cells. For instance, there is no universal method to cell lysis due to the wide diversity of membranes and covers in eukaryotic cells (Lynn DH, Pinheiro M, 2009; Woyke T, Doud DFR, Schulz F, 2017), and the low coverage in genome completion requires the co-assembly of multiple genomes and lost of infraspecific genetic variability (Liu et al., 2017; Mangot et al., 2017). While **cell lysis issues might require approaches individually adapted to the different cell types**, **improvements in whole-genome amplification techniques are on track** (Stepanauskas et al., 2017) and will probably help expanding protists SAGs collections very fast.

6.5 Future perspectives for the assessment of primary producers in the ocean

Amplicon sequencing of 16S rDNA enables a universal and low-priced approach for the taxonomic profiling of primary producers in environmental samples. However, this technique does not allow the inference of functional traits nor the distinction of phototrophs from phago-mixotrophs. These limitations can be tackled through metagenomes assembly or modeling. Yet, similarly as in metaB sequencing, metaG-based approaches also depend on annotation steps against reference databases that are still limited. In order to improve precision in the analysis of picoeukaryotic phototrophs and mixotrophs through molecular methods, we have to **guide sequencing efforts towards the expansion of reference genomes and marker gene databases**. As discussed along these pages, single-cell sequencing and long read sequencing will become valuable resources for the recovery of environmental references.

Molecular techniques allow us to obtain a rather complete view of the diversity and taxonomic composition of the phototrophic and mixotrophic communities in the oceans. What is largely missing in this picture is detailed information about the metabolic activity of these communities. In this regard, the **data obtained from genomic and transcriptomic data should be complemented with physiological analyses of protists' feeding behaviour under different environmental scenarios**. This information will be essential to improve the indirect measurements of microbes trophic activity through modeling, so we can better predict the evolution of Earth's climate and ecosystems dynamics.

A. Co-autored papers

A.1 Single cell genomics yields a wide diversity of small planktonic protists across major ecosystems (Sieracki et al., 2019)

In this study, Sieracki et al. generated a collection of 903 single-cell amplified genomes from picoeukaryotes and showed that single-cell sequencing approach is not biassed against or towards any taxonomic group. My contribution to this manuscript consisted of analysing the distribution of the SAGs collection across Tara Oceans stations and comparing it with the one of MMETSP transcriptomes (Figure 6). These results are included in the Chapter 5 of the thesis.

www.nature.com/scientificreports

SCIENTIFIC REPORTS

OPEN

Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems

Received: 1 June 2018
Accepted: 28 March 2019
Published online: 15 April 2019

M. E. Sieracki[1], N. J. Poulton[2], O. Jaillon[3], P. Wincker[3], C. de Vargas[4], L. Rubinat-Ripoll[4], R. Stepanauskas[2], R. Logares[5] & R. Massana[5]

Marine planktonic protists are critical components of ocean ecosystems and are highly diverse. Molecular sequencing methods are being used to describe this diversity and reveal new associations and metabolisms that are important to how these ecosystems function. We describe here the use of the single cell genomics approach to sample and interrogate the diversity of the smaller (pico- and nano-sized) protists from a range of oceanic samples. We created over 900 single amplified genomes (SAGs) from 8 *Tara* Ocean samples across the Indian Ocean and the Mediterranean Sea. We show that flow cytometric sorting of single cells effectively distinguishes plastidic and aplastidic cell types that agree with our understanding of protist phylogeny. Yields of genomic DNA with PCR-identifiable 18S rRNA gene sequence from single cells was low (15% of aplastidic cell sorts, and 7% of plastidic sorts) and tests with alternate primers and comparisons to metabarcoding did not reveal phylogenetic bias in the major protist groups. There was little evidence of significant bias against or in favor of any phylogenetic group expected or known to be present. The four open ocean stations in the Indian Ocean had similar communities, despite ranging from 14°N to 20°S latitude, and they differed from the Mediterranean station. Single cell genomics of protists suggests that the taxonomic diversity of the dominant taxa found in only several hundreds of microliters of surface seawater is similar to that found in molecular surveys where liters of sample are filtered.

Planktonic protists in the surface ocean are ubiquitous, abundant and highly diverse. They range in size from the smallest known eukaryote, *Ostreococcus* spp. (0.8 µm)[1], to large ciliates, radiolarians, and protist colonies 100's of micrometers across and visible to the naked eye. They function as primary producers, grazers and parasites, and influence the packaging and recycling of carbon and nutrients in marine ecosystems. Protists associate with prokaryotes, sometimes endosymbiotically, to conduct fundamental biogeochemical transformations such as nitrogen fixation[2]. Despite this ecological importance in the structure and function of marine ecosystems the smaller planktonic eukaryotes are not as well characterized as the larger microplankton due to their small size, lack of distinctive morphological features, and the lack of cultures of many dominant forms, especially of the aplastidic bacterivorous protists[3].

As with prokaryotes, genetic methods have revealed remarkably diverse ocean planktonic protist communities[4]. These methods include direct cloning of environmental DNA, fingerprinting methods, tag sequencing, and metagenomics of filtered or sorted fractions of the community. These methods have various advantages and disadvantages depending upon the science question being addressed[5]. For assessing the diversity of the dominant forms present in seawater, clone libraries and tag sequencing have been the favored approaches. These methods have the disadvantage of being biased in favor of particular, often larger, cell types, which can have 10's to 100's

[1]National Science Foundation, 2415 Eisenhower Ave., Alexandria, VA, 22314, USA. [2]Bigelow Laboratory for Ocean Sciences, 60 Bigelow Drive, East Boothbay, ME, 04544, USA. [3]Génomique Métabolique, Genoscope, Institut de biologie François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Université Evry, Université Paris-Saclay, Evry, France. [4]Sorbonne Universités, UPMC Université Paris 06, CNRS, UMR7144, Station Biologique de Roscoff, 29680, Roscoff, France. [5]Department of Marine Biology and Oceanography, Institute of Marine Sciences (ICM)-CSIC, Pg. Maritim de la Barceloneta, 37-49, Barcelona, E-08003, Catalonia, Spain. Correspondence and requests for materials should be addressed to M.E.S. (email: mike.sieracki@gmail.com)

of copies of target genes per cell (in particular the 18S rDNA[6]) and thus are artificially overrepresented in these surveys[7]. This seems to be especially true of the alveolates, including marine alveolate groups I and II, dinoflagellates, and ciliates. Fine plankton nets and filter fractionation is often used to characterize plankton communities, but these can break up fragile animals, colonies, and individual cells, sending their DNA into small size fractions[8].

It has been known for some time that many marine protists are mixotrophic and are not easily assigned to photo- or heterotrophic categories[9,10]. More recent results confirm that many of the small planktonic chlorophyll-containing cells are mixotrophs, ingesting bacteria[11,12]. Flow cytometry easily distinguishes cells containing chlorophyll from those that do not by the presence of chlorophyll autofluorescence. We, therefore, use the terms "plastidic" and "aplastidic" here to distinguish the presence or absence of chloroplasts, without assigning a trophic category to them. In this nomenclature plastidic cells are most likely phototrophic or mixotrophic, although some could also be heterotrophs with a phototroph in their food vacuole. Aplastidic cells are most likely to be heterotrophic, but could be phototrophs or mixotrophs with reduced chloroplasts or faint chlorophyll fluorescence.

The single cell approach has proven its power in the discovery of new potential metabolisms in uncultured prokaryotes[13], and has the advantage of yielding large amounts of genomic DNA from individual microorganisms for further sequencing and investigation. Early results from one coastal Maine sample revealed significantly higher protist diversity in whole water samples using the single cell approach compared to clone libraries[7]. The bias due to gene copy number in the clone libraries was the likely cause of the underestimation. Sequencing of three "picobiliphyte" (now Picozoa) SAGs from that sample showed how this approach can reveal microbial interactions between eukaryotes, prokaryotes and viruses[14].

Here we report on a larger set of protist SAGs collected across a wider range of ocean samples for single cell genomics on the *Tara* Oceans expedition[15]. We analyzed the 18S rDNA sequences of over 900 SAGs from the Adriatic Sea, Arabian Sea and the Indian Ocean across a wide latitudinal gradient. A subset of these SAGs has recently been more fully sequenced and assembled[16–19]. These studies demonstrate that, although individual SAGs may represent only a portion of the cellular genome, the co-assembly of multiple SAGs can improve genome recovery significantly. For MAST-4 clade, the co-assembly of 14 SAGs yielded ~74% of genome recovery[16], and for *Monosiga brevicollis* four co-assembled SAGs yielded 46% of the genome[17]. Single cell genomes were used as reference genomes to match with metagenomic data and reveal biogeographic patterns of *Bathycoccus*[18], and unexpected functional diversity of the dominant MAST-4 heterotroph clade[19]. The work reported here shows that our sampling and cell handling approach appears to accurately sample the diversity of the dominant oceanic eukaryotes in the pico- to nanoplankton size range (<5 µm).

Methods

Cryopreservation and flow cytometric detection. Tests were done to confirm and optimize the cryopreservation of marine protists for single cell genomics using marine samples from 1 m depth at the dock in Booth Bay, Maine, a coastal Atlantic site. The cryoprotectant glycine betaine[20] had previously proven to preserve prokaryotic cells, allow identification of nucleic acid stained cells by flow cytometry, and not interfere with single cell amplification, PCR screening, and sequencing reactions[21]. Live (aplastidic) protists had been sorted and successfully amplified and PCR screened using light scatter properties and Lysotracker staining[7,22]. Lysotracker, however, only stains live cells with active vacuoles and does not stain cryopreserved cells. As an alternative, we adopted the method of Zubkov, *et al.*[23] based on SYBR Green I staining for detecting aplastidic cells. Plastidic cells are easily distinguished by red autofluorescence of chlorophyll emitted by chloroplasts using flow cytometry. An experiment was conducted to compare the cell numbers of aplastidic cells obtained by flow cytometry using both fresh and cryopreserved samples with the cryoprotectants glycine betaine (GBe, 7% v/v, Sigma) and glycerol-TE (Gly-TE, 5% glycerol + 1x TE buffer, Sigma). Counts of the cryopreserved samples were determined after staining with SYBR Green I (1:5,000 dilution; ThermoFisher Scientific, USA), while counts of fresh samples were determined after samples stained with both SYBR Green I (SYBR, 1:5,000 dilution; ThermoFisher Scientific, USA) and Lysotracker (LT, 75 nmol; ThermoFisher, Scientific, USA).

Ocean sampling. Whole water samples were taken from surface ocean water, or from the deep chlorophyll maximum (DCM), by a submerged impeller pump. Sample sites included the Adriatic Sea, Arabian Sea and the Equatorial Indian Ocean. Subsamples were dispensed into replicate 4 mL cryovials containing GBe as a cryoprotectant (7% w/v, final conc.). The cryovials were flash frozen and stored in liquid nitrogen (LN) until SV Tara reached a shipping port.

Hydrographic data, including salinity and temperature, was determined at each station using a CTD with a bottle rosette onboard SV Tara. Bottle samples were analyzed for chlorophyll by HPLC, and for counts of the small cells using standard flow cytometry methods[24].

Single cells. Samples were express-shipped on dry ice to Bigelow Laboratory for Ocean Sciences where they were stored in LN until sorting. SAG generation and identification were performed at the Single Cell Genomics Center at Bigelow (scgc.bigelow.org). On the sorting day tubes were thawed at room temperature and a subsample was stained with SYBR Green I. Sorting was conducted on a Beckman-Coulter MoFlo sorter outfitted with a Cyclone™ robotic arm for sorting into plates. Single plastidic cells were sorted using the natural chlorophyll autofluorescence within an unstained subsample and single aplastidic cells were sorted using a SYBR Green I (1:5000 dilution) stained subsample[23]. All single cells were sorted into 384 well plates containing 0.6 µL TE buffer per well. Multiple plates were prepared for unstained plastidic and stained aplastidic cells from each sample. After sorting, all plates were stored frozen at −80 °C.

								Chl			Plastidic euks		No. of SAGs	
Stn	Date mm/dd/yy	Site[a]	Lat (deg. N)	Lon (deg. E)	Depth (m)	Temp. (°C)	Salinity (psu)	µg L^{-1} *(s.e.)*	*Syn* # mL^{-1}	*Pro* # mL^{-1}	Small # mL^{-1}	Large # mL^{-1}	Plast.	Aplast.
23	11/18/09	Ad	42.18986	17.71670	55	17.32	38.201	0.139 *(0.04)*	10,448	19,390	392	699	24	118
39	03/18/10	Ar	18.57138	66.53050	S	26.82	36.285	0.099 *(0.02)*	146,758	125,121	3,275	2,165	38	52
41S	03/30/10	IO	14.59540	69.98100	S	29.09	36.025	0.020 *(0.02)*	13,703	119,176	3,507	1,393	57	88
41D	"	IO	"	"	59	27.21	36.499	0.373 *(0.11)*	3,809	245,757	888	nd	64	141
46	04/15/10	M	−0.66245	73.16097	S	30.13	35.111	0.122 *(0.01)*	178,299	157,827	10,607	596	72	78
47	04/16/10	IO	−2.04653	72.15680	S	30.20	34.912	0.007 *(0.01)*	210,569	1,398	677	677	26	26
48	04/19/10	IO	−9.40295	66.36804	S	29.83	34.175	nd	497	119,104	498	nd	37	21
51	05/11/10	IO	−21.50212	54.35328	S	27.26	34.901	0.040 *(nd)*	1,336	221,147	653	nd	35	26

Table 1. Samples station locations, dates, water characteristics, pico- and nanoplankton cell abundances, and numbers of plastidic and aplastidic SAGs obtained. The small and large plastidic eukaryotes were identified and counted by flow cytometry triggered on chlorophyll autofluorescence, so aplastidic protists are not counted here. [a]Ad = Adriatic Sea, Ar = Arabian Sea, IO-Indian Ocean, M = Addu Atoll, Maldives, s.e. = standard error, S = surface (<3 m), sample depths greater than 50 m were targeted at the subsurface chlorophyll maximum, nd = not detected.

Lysis and MDA. Attempts to improve the amplification yield of single cells were made by increasing the number of freeze-thaw cycles and incubating with KOH at 20 °C. Incubating with KOH at higher temperatures resulted in lower yields, probably due to DNA degradation. We settled on 5 cycles of freeze-thaw as optimal.

Genomic DNA from single cells was amplified using the phi-29 polymerase (real-time multiple displacement amplification, rtMDA) method in 384-well format[13]. Amplification reactions were run overnight (ca. 18 h) and monitored in real time based on DNA fluorescence. Critical point (Cp) values for each well were determined as the reaction time when well fluorescence reached half the maximum value. Based on these Cp values we selected those SAGs having Cp values below 14 h for further analysis.

PCR screening. The genomic DNA produced by MDA served as template for screening using universal 18S rRNA gene eukaryotic PCR primers. All wells were screened regardless of their MDA Cp values. Primers used were Euk528 (forward)[6] and Euk B (reverse)[25] which amplify two thirds of the gene (ca. 1200 bp)[26]. PCR amplicons were sequenced using Sanger technology using the same two primers. Sequences were curated manually and compared to sequences in GenBank using BLAST to determine similarity to known sequences. Closest matches and closest cultured matches were recorded. Sequences were aligned using MAFFT and compared to each other, and to reference sequences for some groups, using maximum likelihood trees (RAxML) to achieve a final phylogenetic assignment. To assess primer bias in sampling protistan diversity we additionally screened one plate each of plastidic and aplastidic protists from one sample, Stn 41 surface, with two additional primer sets targeting the variable V4 and V9 regions of the rRNA gene (see PCR protocols and primer sequences in refs.[27,28], for V4 and V9 regions, respectively).

Comparison of SAG sequences with metabarcoding data. We compared the relative community composition at three *Tara* Oceans stations using available V9 metabarcodes at the group level with the SAG samples. Detailed information on sampling and metabarcoding (iTAG) sequencing can be found in Pesant *et al.*[29] and de Vargas *et al.*[28], respectively. We separated the iTAG sequences into plastidic and aplastidic types by assigning them to class-level groups, removing groups not targeted in the SAGs such as ciliates, diatoms, dinoflagellates, MALVs, radiolarians and unassigned. Then the proportions were calculated on the remaining 33 groups for comparison with SAGs.

We also used the V9 metabarcodes obtained from Tara Oceans samples to explore the occurrence and abundance of SAGs in the global ocean. For this analysis, we only considered samples obtained from the photic zone (surface and DCM) and the smaller size fractions, piconano- (0.8–5 µm) and nano- (5–20 µm) sized cells. We ended up with a dataset containing barcodes from 337 samples deriving from 105 stations. The resulting metabarcode table had 435,240,095 V9 sequence reads grouped into 4,298,066 valid barcodes. The barcodes were clustered into OTUs using SWARM 2.1.129 with default options (local clustering threshold d = 1), generating a total of 271,787 OTUs[28]. We mapped the V9 sequences of 868 SAGs on these OTUs using BLAST 2.6.0 and selected the 671 hits that were retrieved with similarity >97% and coverage >80%.

For comparison, we also mapped the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP)[30] V9 sequences onto these oceanic OTUs. Among the 385 MMETSP transcriptomes for which we could recover sequences of the V9 region, 212 were mapped on some OTU at similarity >97% and coverage >80%.

Results

Sample locations and water properties. Sample locations, water properties, and number of recovered SAGs are shown in Table 1. All stations are open water except for station 46, which was in the middle of a tropical lagoon (Supplementary Fig. S1).

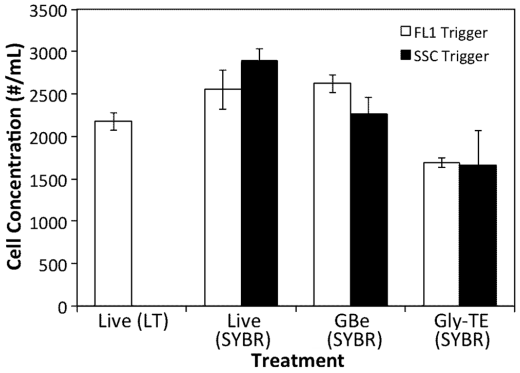


Figure 1. Comparison of flow cytometric cell counts of aplastidic protists counted live and with two cryoprotectants. Triplicate samples of coastal Maine water were run live and stained with either Lysotracker (LT) or SYBR green. Replicate samples were stored with the cryoprotectants glycine betaine (GBe) or glycerol – TE (Gly-TE) at −80 °C and then stained with SYBR green before enumeration. Each replicate was run with the flow cytometer triggered by green fluorescence (FL1) or by side scatter (SSC). Errors bars show standard deviations of triplicate samples.

Cryopreservation. Preliminary tests using the cryoprotectant glycine betaine (Gbe) indicated that this method worked well for sorting and single cell genomics of protists. Chlorophyll fluorescence was preserved in the plastidic cells for discrimination by flow cytometry (Supplementary Fig. S2). Aplastidic cells preserved this way could be stained using SYBR Green I and a sort region was created similar to that in Zubkov *et al.*[23] (Supplementary Fig. S2). We then compared cryopreservation with the live staining methods for cell counts of aplastidic protists, and found that the GBe method showed the lowest cell loss compared to live samples, or the other cryoprotectant glycerol-TE (glyTE) (Fig. 1). In fact, live cell counts determined by SYBR Green I staining were higher than by Lysotracker staining, and the counts in the GBe cryotreatment were not significantly different from those using SYBR Green I stained cells.

MDA and PCR performance. Initial rtMDA results for the protists yielded fewer positive wells (<20%) than what we usually observe for oceanic prokaryotes (average 27%)[31]. We tried different lysis protocols including multiple freeze-thaw cycles and increasing the temperature of the 10 minute KOH incubation, but these simple modifications did not improve yield significantly (data not shown).

Generally we found that plastidic cells had a lower yield of good quality 18S rRNA gene sequences than aplastidic cells (Supplementary Table S1). Twenty-nine plates were processed and analyzed, comprising 9,135 one cell wells (non-controls). We recovered good quality 18S rRNA gene sequence from 7.2% of the plastidic cell wells, and from 14.7% wells of aplastidic cells. These were statistically different at a confidence level of 95% (Student's t-test, $p = 0.046$). There are 3,408 one-cell wells (37%) with good MDA, but no 18S rRNA gene identity. These wells are likely to contain amplified eukaryote genomic DNA, but where 18S rRNA genes could not be recovered due to uneven MDA, PCR primer mismatches, long inserts in 18S rRNA genes, or other interferences[32]. Conversely, there were some wells (124, 1.4%) with good 18S rRNA gene identity but with poor MDA (Cp > 14 h). These could contain a limited amount of genomic DNA. The list of SAGs with good 18S rRNA gene identity is given in Supplementary Table S2. Sequences have been submitted to the European Nucleotide Archive (ENA, accession numbers PRJEB31452).

V4–V9 primer screens. The numbers of SAGs identified by each of the three primer sets used showed that the addition of the V9 primer screen identified 11 additional SAGs from each plate tested, one plastidic and one aplastidic sort (Supplementary Fig. S2). The addition of V4 only identified six more SAGs from the plate of plastidic cells, and three more from the aplastidic plate. Conversely, the V4 primers missed 23 SAGs (13 plastidic, 10 aplastidic) that were identified by the Euk528/B primer set, and the V9 primers missed 16 SAGs (8 from each plate). There were no major new groups that were picked up by the new primers in these plates.

Distribution of sorted cells across phylogenetic groups. The distributions of plastidic and aplastidic cells, as determined by flow cytometry (i.e. presence or absence of chlorophyll fluorescence), is shown in Fig. 2 for the defined taxonomic groups. Twenty-two of these groups were represented essentially by aplastidic cells and twelve of them by plastidic cells. Interestingly, some of the groups represented by cells from both sorts (i.e. Chrysophyceae, Chlorarachniophyta, Dictyochophyceae) are also well known for containing both plastidic and colorless species. More intriguing was the presence of MALV-II among the two sorts. Generally, the distribution of chloroplasts across these groups is as expected based on what we know about their phylogeny and evolution[33].

Protist communities. The protist communities recovered by the SAG approach were quite diverse in most samples (Fig. 3). Richness, calculated at the level of the groups defined here, was highest for station 41 surface,

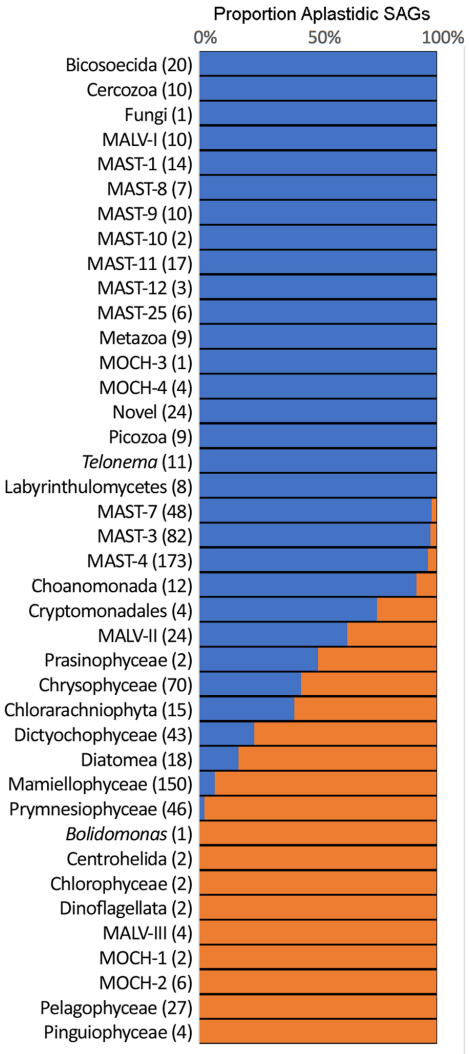


Figure 2. Distribution of plastidic and aplastidic SAGs within the major taxonomic groups found. The bars show the proportion of SAGs in each group that were sorted as aplastidic (blue bars) and plastidic (orange bars) cells. The groups are ranked by proportion and the number of identified SAGs for each is given in parentheses.

and lowest at stations 23, 39, and 46, while diversity (Shannon H) was highest at station 51, and evenness was highest at station 48 (Supplementary Table 3).

The plastidic cells for three stations, 39, 41 deep chlorophyll maximum, and 46, were dominated by Mamiellophyceae (Fig. 3a), specifically mixed blooms of *Micromonas* and *Ostreococcus spp*. with identical 18S rRNA sequences. Station 23D, in the Adriatic, was distinctive in that Pelagophyceae dominated the plastidic cells. The remaining stations (48, 51, 47 and 41S) had more diverse and similar community compositions with a mix of Prymnesiophyceae, Chrysophyceae, and Dictyochophyceae dominating.

The aplastidic protist communities were more similar to each other than the plastidic communities across stations at the taxonomic levels chosen (Fig. 3b). Overall three marine stramenopiles types, MAST-4, 3 and 7, made up about 50% of the aplastidic cells. Other dominant types of the aplastidic community across the other stations were Chrysophyceae, Bicosoecida, MALV-I, MALV-II, Telonema, Dichtyochophyceae and Picozoa. There are 9 cells that are from metazoans, mostly ctenophores and salps. One metazoan SAG with a novel 18S rRNA gene was

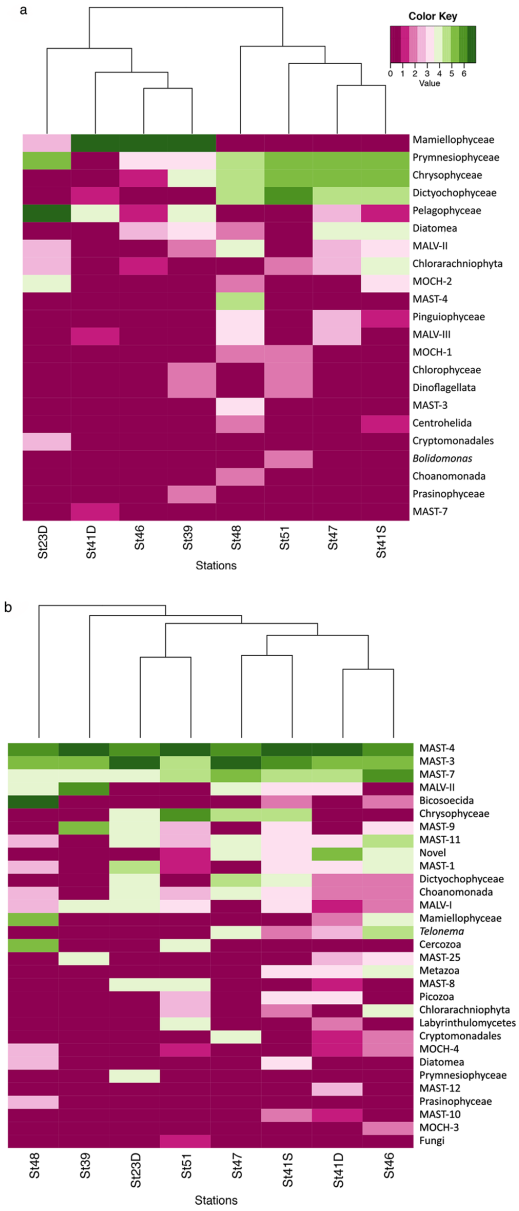


Figure 3. Heat maps showing the SAG composition of the (**a**) plastidic community and (**b**) aplastidic community at each station. The values on the color scale are the number of SAGs of each type transformed by $\log_2(x)+1$ (with zeros left as zeros)[40].

found and appears to be from an acorn worm. Metazoan SAGs could have come from single cells from damaged animal tissue, fecal material, or as free-swimming gametes.

The analysis of iTag sequences[28] from the 3 samples where we could directly compare, revealed a general positive trend with the SAG groups with the exception of some groups found by the iTag method, but not in the SAGs (Fig. 4). The metabarcode (iTag) approach revealed groups that were not detected in the SAGs (points on the left of Fig. 4). This is likely due to differences in sampling between the two methods, most importantly the fewer cells

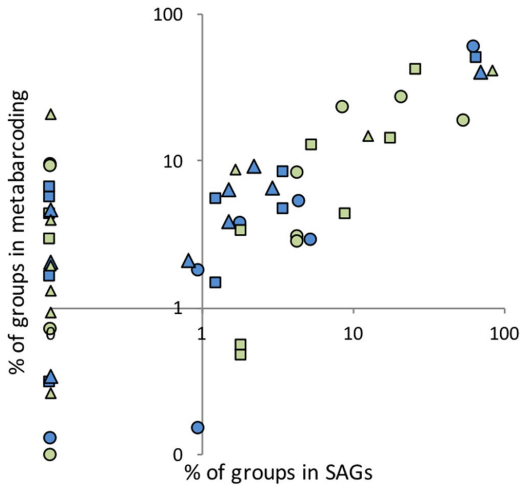


Figure 4. Comparison of the relative abundance of taxonomic groups found in the SAG collections and in metabarcoding dataset (iTAGs) at three stations for plastidic (green symbols) and aplastidic (blue) protists. Each station has a different plot symbol: circle (23-D), square (41-S), and triangle (41-D).

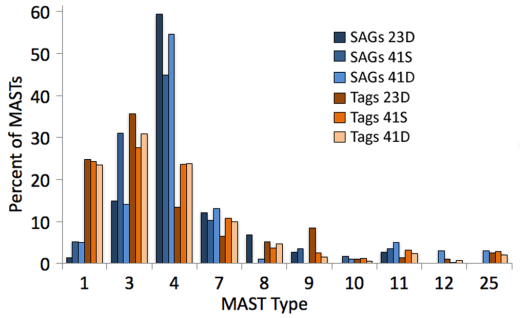


Figure 5. Comparison of the proportion of types found in SAGs (blue columns) and metabarcoding (iTAGs, orange columns) at three stations (23, 41S and 41D, represented by the different color shades) for the various MAST clades. Data calculated as percent of total MASTs.

identified per sample through the SAG approach. The SAGs were derived from small, whole seawater samples (ca. 300 microliters subsampled from several milliliters), whereas the metabarcoding samples were comprised of many liters of seawater size fractionated and collected on filters. For the SAG samples only the small, and most dominant protists were chosen for sorting. Analysis of the outliers - groups that were relatively underrepresented in the SAG data - was not particularly enlightening (data not shown). In the comparison of only marine stramenopile groups (Fig. 5) showed a similar distribution between the methods.

The comparison of the SAG sequences against oceanic metabarcoding V9-swarms (Fig. 6a) shows that the SAGs represent the most abundant types (bubbles in the upper right) presumably corresponding to ecologically important groups. Some matches were also found to types rarer in the metabarcodes (Fig. 6b, lower left). In contrast to the SAGs, the sequences in the Marine Microbial Eukaryote Transcriptome Sequencing Project database (Fig. 6b) were more representative of rare types in the oceanic metabarcodes, with fewer matches in the upper right compared to the SAGs (Fig. 6a).

Discussion

We have found that single cell methods developed to preserve samples and amplify genomic DNA for planktonic prokaryotes were generally transferrable to eukaryotes. The major difference is that Gly-TE, although a preferable cryoprotectant for prokaryotes, appears to be less effective for protists than GBe. We noticed that the yields of SAGs from protist plates, especially for the plastidic types, were lower than what we usually observed for

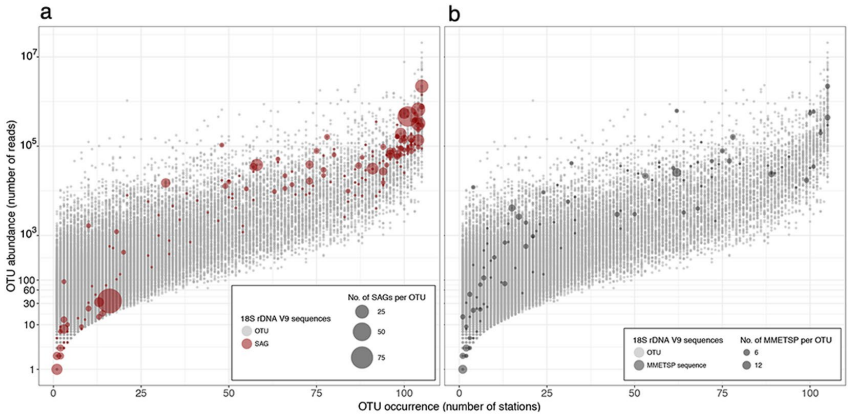


Figure 6. Mapping of SAG (**a**) and MMETSP (**b**) V9 sequences onto the Tara Oceans metabarcode V9 OTUs. The size of the colored bubbles represents the number of matching SAG or MMETPS sequences in each OTU. SAGs dataset represents a number cosmopolitan and abundant taxa while the MMETSP sequences have fewer matches and are spread across the range of dominance distribution.

planktonic prokaryotes. Our attempts to make simple modifications to our lysis methods failed to significantly increase yields of SAGs. The test of additional PCR primers to identify positive cases did yield some more identified SAGs, but did not reveal whole new cell types not seen with the Euk528/B primer set. We conclude from this test that the use of these additional primer screens slightly increased yield, but not diversity.

We observed that more of the cells sorted as aplastidic yielded successful MDA product than those sorted as plastidic. This could be due to lower lysis success with these groups, or interference with the MDA reaction by constituents such as polysaccharides, either within the cells or on the cell wall of plastidic cells. While the single cell approach avoids some biases, there may be others that affect our results. These include a possible lysis bias with some cell types being less likely to be opened and their genomes available for amplification than other types.

The sorting strategy using plastid autofluorescence was very effective, as the majority of the groups were sorted either in aplastidic or in plastidic sorts (not both, Fig. 1). Groups containing a mix of plastidic and aplastidic cells (Fig. 1) could be explained by several factors. Some groups (e.g. Choanomonada, MAST-3, and -4) are predominantly aplastidic bacterivores with only a few instances of plastidic types. These could be herbivores with a recently ingested plastidic cell[19]. For instance, it has been seen that MAST-4 is able to graze both on bacteria and plastidic picoeukaryotes[33]. In this case the fluorescence in their food vacuoles would cause it to be classified as plastidic. Other groups (e.g. Diatomea, Mamiellophyceae, and Prymnesiophyceae) are predominantly plastidic phototrophs where some cells might have very weak autofluorescence (little chlorophyll) and were not detected as plastidic. More interesting were the groups with similar numbers of aplastidic and plastidic cells (e.g. MALV-II, Chrysophyceae, Chlorarachniophyta and Dictyochophyceae) as most of these groups are known to contain plastidic and aplastidic species. They can also include species with weak fluorescence plastids or that change the pigment content depending upon circumstances. We may also have sorted infected autotrophs with degraded host nuclear DNA. The presence of the putative MALV-II parasite within this category is intriguing and deserves further analyses.

In this set of protist SAGs we found 9 Picozoa cells[34] (formerly Picobiliphytes[35]), all in aplastidic sorts. This continues to confirm our observations of these organisms from Booth Bay, Maine[7], where these types only appeared in sorts of cells without chlorophyll fluorescence. Yoon *et al.*[14] found no genetic evidence of plastids in the partial genomes of three SAGs, and Seenivasan *et al.*[34] obtained the first picozoan culture and found no evidence of plastids in serial thin sections.

There are a variety of factors that can bias our determination by the single cell approach of the community composition of marine microeukaryotes[7]. Koid *et al.*[5] found that diatoms appear to be underrepresented in clone libraries, likely due to difficulties in lysing the cells and releasing the genomic DNA. Amacher *et al.*[36] noted biases in clone libraries related to abundances of both target and co-occurring species. In our results we obtained 18 diatoms: 15 out of 353 plastidic SAGs, and an additional 3 from the 550 aplastidic SAGs. This might seem a small number, but we targeted a flow cytometric region that only contained small cells (about 2–5 µm in size) and diatoms are generally larger than this. Therefore, it is not clear if we missed diatoms due to inefficient lysis or because they were not included in the sorting gates. At any rate, the data shown here reveals that the community composition derived from the analysis of dozens of SAGs per sample is comparable with the more common metabarcoding molecular surveys.

The community composition of protists we observed in the Indian Ocean (Fig. 3) has some similarities and differences from previous observations based on clone libraries[37]. On a cruise from the southwest to the northeast Indian Ocean, adjacent to the Tara Oceans Expedition, reaching similar latitudes (25°–12°S) east of our transect,

clone libraries of the cells passing a 3 µm filter revealed a somewhat different pattern of community structure than what we observed at our comparable stations (41, 47, 48 and 51)[37]. They found higher proportions of dinoflagellates, marine alveolates (esp. MALV-I), and radiolarians than we observed. They also found lower proportions of MAST, Chrysophytes, Dichtyochophytes, and Prymnesiophytes. Prasinophytes were common in both studies. The major differences between these studies seem to relate to the biases we have seen in clone libraries due to gene copy number. The types overrepresented in the Not *et al.*[37] study relative to this study are types known to contain many copies of the targeted rDNA operon.

Assembling whole genomes from SAGs is generally difficult, and gets particularly challenging for eukaryotic genomes, which can be complicated by heterozygosity, and putatively massive repeated regions. New information can be obtained from partially assembled genomes, however, especially from uncultivated cell types[14,16–19,26]. Due to the nature of MDA, some sections of the genomic DNA are over amplified while other sections may not be amplified at all[32], preventing sequencing and assembly of complete genomes from single cells. This appears to involve stochastic processes when amplifying a single DNA molecule as a starting template, as well as a bias against sequences with high G + C content[31]. Recent work has shown that co-assembly of eukaryote SAGs from several cells of the same population significantly increases the proportion of the genome that can be recovered[16,17].

As with prokaryotes, the high diversity of single celled eukaryotes in marine ecosystems is problematic for metagenomic or metatranscriptomic approaches alone. Without assembled genomes it is difficult to assign functional genes to species[38]. In addition, most marine eukaryotes have not been cultivated, especially the heterotrophic types[39], so traditional genome sequencing is not a viable option. The Marine Microbial Eukaryote Transcriptome Sequencing Project sought to sequence the transcriptomes of about 650 important marine microbial eukaryotes, and has produced a powerful sequence dataset[30]. It is limited, however, in only including cultured types with an emphasis on phototrophs. Keeling *et al.*[30] acknowledge that single cell genomics will play an important complementary approach to gain understanding of these diverse protists. The approach is a powerful complement to environmental metatranscriptomics[38].

For prokaryotes the high potential metabolic diversity of communities is well established, and advances in understanding the relationships between genetic diversity and ecosystem function is currently an area of active ecological research not only in environmental systems, but microbiomes of metazoans including humans. The high diversity of eukaryotic protists in the ocean is less well appreciated, but similarly enigmatic. Conventional knowledge has limited their metabolic or ecological function to phototrophy and heterotrophy. New evidence, however, reveals complexities of mixotrophy, endosymbiosis, and parasitism that could dominate the functions of marine protists. More intricate relationships based on small scale physical structuring, resource sharing, and chemical communication could be the basis of niche separation allowing the high diversity observed. Sequencing single eukaryotic cells sampled directly from the ocean as described here offers a way forward in deciphering who is doing what and how in the ocean.

References

1. Chrétiennot-Dinet, M. J. *et al*. A new marine picoeucaryote: *Ostreococcus tauri* gen. et sp. nov. (chlorophyta, prasinophyceae). *Phycologia* **34**, 285–292 (1995).
2. Thompson, A. W. *et al*. Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* **337**, 1546–50 (2012).
3. Massana, R., Pernice, M., Bunge, J. A. & Del Campo, J. Sequence diversity and novelty of natural assemblages of picoeukaryotes from the Indian Ocean. *ISME J.* **5**, 184–95 (2011).
4. Moon-van der Staay, S. Y., De Wachter, R. & Vaulot, D. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**, 607–610 (2001).
5. Koid, A., Nelson, W. C., Mraz, A. & Heidelberg, K. B. Comparative analysis of eukaryotic marine microbial assemblages from 18S rRNA gene and gene transcript clone libraries by using different methods of extraction. *Appl. Environ. Microbiol.* **78**, 3958–65 (2012).
6. Zhu, F., Massana, R., Not, F., Marie, D. & Vaulot, D. Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol. Ecol.* **52**, 79–92 (2005).
7. Heywood, J. L., Sieracki, M. E., Bellows, W., Poulton, N. J. & Stepanauskas, R. Capturing diversity of marine heterotrophic protists: one cell at a time. *ISME J*, https://doi.org/10.1038/ismej.2010.155 (2010).
8. Sørensen, N., Daugbjerg, N. & Richardson, K. Choice of pore size can introduce artefacts when filtering Picoeukaryotes for molecular biodiversity studies. *Microb. Ecol.* **65**, 964–968 (2013).
9. Stoecker, D. K., Michaels, A. E. & Davis., K. H. Large proportion of marine planktonic ciliates found to contain functional chloroplasts. *Nature* **326**, 790–792 (1987).
10. Sanders, R. W., Berninger, U.-G., Lim, E. L., Kemp, P. F. & Caron, D. A. Heterotrophic and mixotrophic nanoplankton predation on picoplankton in the Sargasso Sea and on Georges Bank. *Mar. Ecol. Progr. Ser.* **192**, 103–118 (2000).
11. Zubkov, M. V. & Tarran, G. A. High bacterivory by the smallest phytoplankton in the North Atlantic Ocean. *Nature* **455**, 224–6 (2008).
12. Frias-Lopez, J. *et al*. Microbial community gene expression in ocean surface waters. *Proc. Natl. Acad. Sci.* **105**, 3805–3810 (2008).
13. Swan, B. K. *et al*. Potential for chemolithoautotrophy among ubiquitous bacteria lineages in the dark ocean. *Science* **333**, 1296–300 (2011).
14. Yoon, H. S. *et al*. Single cell genomics reveals organismal interactions in uncultivated marine protists. *Sciences.* **332**, 714–717 (2011).
15. Karsenti, E. *et al*. A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biol* **9**, e1001177 (2011).
16. Mangot, J.-F. *et al*. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Sci. Rep.* **7**, 41498 (2017).
17. López-Escardó, D. *et al*. Evaluation of single-cell genomics to address evolutionary questions using three SAGs of the choanoflagellate *Monosiga brevicollis*. *Sci. Rep*. **7** (2017).
18. Vannier, T. *et al*. Survey of the green picoalga *Bathycoccus* genomes in the global ocean. *Sci. Rep*. **6** (2016).
19. Seeleuthner, Y. *et al*. Single-cell genomics of multiple uncultured stramenopiles reveals underestimated functional diversity across oceans. *Nat. Commun*. **9**, 310 (2018).
20. Cleland, D., Krader, P., McCree, C., Tang, J. & Emerson, D. Glycine betaine as a cryoprotectant for prokaryotes. *J. Microbiol. Methods* **58**, 31–38 (2004).
21. Stepanauskas, R. & Sieracki, M. E. Matching phylogeny and metabolism in the uncultured marine bacteria: one cell at a time. *Proc. Natl. Acad. Sci.* **104**, 9052–9057 (2007).

22. Rose, J. M., Caron, D. A., Sieracki, M. E. & Poulton, N. Counting heterotrophic nanoplanktonic protists in cultures and aquatic communities by flow cytometry. *Aquat. Microb. Ecol.* **34**, 263–277 (2004).
23. Zubkov, M. V., Burkill, P. H. & Topping, J. N. Flow cytometric enumeration of DNA-stained oceanic planktonic protists. *J. Plankton Res.* **29**, 79–86 (2006).
24. Vaulot, D., Simon, N. & Marie, D. Phytoplankton cell counting by flow cytometry. in *Algal Culturing Techniques* (ed. Anderson, R.) pp 253–268 (Academic Press, 2005).
25. Díez, B., Pedrós-Alió, C. & Massana, R. Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing. *Appl. Environ. Microbiol.* **67**, 2932–2941 (2001).
26. Martinez-Garcia, M. *et al*. Unveiling *in situ* interactions between marine protists and bacteria through single cell sequencing. *ISME J*, https://doi.org/10.1038/ismej.2011.126 (2011).
27. Massana, R. *et al*. Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environmental Microbiology* **17** (2015).
28. de Vargas, C. *et al*. Eukaryotic plankton diversity in the sunlit ocean. *Sci.* **348**, https://doi.org/10.1126/science.1261605 (2015).
29. Pesant, S. *et al*. Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data* **2**, 150023, https://doi.org/10.1038/sdata.2015.23 (2015).
30. Keeling, P. J. *et al*. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol.* **12** (2014).
31. Stepanauskas, R. *et al*. Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nat. Commun.* **8**, 84 (2017).
32. Zhang, K. *et al*. Sequencing genomes from single cells by polymerase cloning. *Nat. Biotech.* **24**, 680–686 (2006).
33. Adl, S. M. *et al*. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol* **52**, 399–451 (2005).
34. Seenivasan, R., Sausen, N., Medlin, L. K. & Melkonian, M. *Picomonas judraskeda* gen. et sp. nov.: the first identified member of the Picozoa phylum nov., a widespread group of picoeukaryotes, formerly known as 'picobiliphytes'. *PLoS One* **8**, e59565 (2013).
35. Not, F. *et al*. Picobiliphytes: A Marine Picoplanktonic Algal Group with Unknown Affinities to Other Eukaryotes. *Science* **315**, 253–255 (2007).
36. Amacher, J. A., Baysinger, C. W. & Neuer, S. The importance of organism density and co-occurring organisms in biases associated with molecular studies of marine protist diversity. *J. Plankton Res.* **33**, 1762–1766 (2011).
37. Not, F. *et al*. Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. Deep Sea Res. Part I Oceanogr. *Res. Pap.* **55**, 1456–1473 (2008).
38. Carradec, Q. *et al*. A global ocean atlas of eukaryotic genes. *Nat. Commun.* **9**, 373 (2018).
39. Del Campo, J., Not, F., Forn, I., Sieracki, M. E. & Massana, R. Taming the smallest predators of the oceans. *ISME J.* 1–8, https://doi.org/10.1038/ismej.2012.85 (2012).
40. Anderson, M. J., Ellingsen, K. E. & McArdle, B. H. Multivariate dispersion as a measure of beta diversity. *Ecol. Lett.* **9**, 683–693 (2006).

Acknowledgements

We thank the staff of the Single Cell Genomics Center in Bigelow for generating and screening the single amplified genomes. We thank the analysis by J.M. Gasol, S.G. Acinas and the ICM group for flow cytometry cell counts. Funding was provided by the following sponsors: U.S. NSF grant DEB-1031049; CNRS (in particular Groupement de Recherche GDR3280); European Molecular Biology Laboratory (EMBL), Genoscope/CEA; the French Government 'Investissements d'Avenir' programmes OCEANOMICS (ANR-11-BTBR-0008) and FRANCE GENOMIQUE (ANR-10-INBS-09-08; Agence Nationale de la Recherche; European Union FP7 (MicroB3/No.287589); and EU project SINGEK (H2020-MSCA-ITN-2015-675752). RL was supported by a RyC fellowship (RYC-2013-12554, MINECO, Spain). We also thank the support and commitment of agnès b. and Etienne Bourgois, the Veolia Environment Foundation, Region Bretagne, Lorient Agglomeration, World Courier, Illumina, the Eléctricité de France (EDF) Foundation, Fondation pour la recherche sur la biodiversité (FRB), the Foundation Prince Albert II de Monaco, the Tara Foundation, its schooner and teams. We are also grateful to the French Ministry of Foreign Affairs for supporting the expedition and to the countries who graciously granted sampling permissions. Tara Oceans would not exist without continuous support from 23 institutes (http://oceans.taraexpeditions.org/en/m/science/labs-involved/). The authors further declare that all data reported herein are fully and freely available from the date of publication, with no restrictions, and that all of the samples, analyses, publications, and ownership of data are free from legal entanglement or restriction of any sort by the various nations whose waters the Tara Oceans expedition sampled in. Specific bioinformatics analyses were performed at the MARBITS platform of the Institut de Ciències del Mar (ICM; http://marbits.icm.csic.es). This article is Tara Oceans contribution number 87. The views expressed in this paper do not necessarily reflect the views of the National Science Foundation.

Author Contributions

M.S. and R.M. led the writing of the manuscript. M.S., N.P. and R.S. developed the sample methodology and produced the SAGs. Data interpretation and analysis was by M.S., O.J., P.W. D.dV. L.R.-R., R.L. and R.M. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-42487-1.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

A.2 Disentangling the mechanisms shaping the surface ocean microbiota (Logares et al., (2019)

This study describes the influence of natural selection, ecological drift and dispersal as mechanisms shaping the composition of eukaryotic and prokaryotic picoplanktonic communities in marine ecosystems. I contributed to this manuscript by curating part of the Malaspina-2010 metadata, used for comparing the effects of environmental variables in community dynamics.

A.2.1 Main Paper

Disentangling the mechanisms shaping the surface ocean microbiota

Ramiro Logares[1,2*] (ramiro.logares@icm.csic.es), Ina M. Deutschmann[1] (ina@icm.csic.es), Pedro C. Junger[3] (pedro.junger@gmail.com)[3], Caterina. R. Giner[1] (c.giner@oceans.ubc.ca), Anders K. Krabberød[2] (a.k.krabberod@ibv.uio.no), Thomas S. B. Schmidt[4] (sebastian.schmidt@embl.de), Laura Rubinat-Ripoll[5] (laura.rubinat@sb-roscoff.fr), Mireia Mestre[1,6,7] (mireia.mestre.martin@gmail.com), Guillem Salazar[1,8] (guillems@ethz.ch), Clara Ruiz-González[1] (clararg@icm.csic.es), Marta Sebastián[1,9] (msebastian@icm.csic.es), Colomban de Vargas[5] (vargas@sb-roscoff.fr), Silvia G. Acinas[1] (sacinas@icm.csic.es), Carlos M. Duarte[10] (carlos.duarte@kaust.edu.sa), Josep M. Gasol[1,11] (pepgasol@icm.csic.es), Ramon Massana[1] (ramonm@icm.csic.es)

[1] Institute of Marine Sciences (ICM), CSIC, 08003, Barcelona, Catalonia, Spain.
[2] Section for Genetics and Evolutionary Biology, Department of Biosciences, University of Oslo, N-0316 Oslo, Norway.
[3] Laboratory of Microbial Processes & Biodiversity. Department of Hydrobiology – DHB, Universidade Federal de São Carlos – UFSCar, São Carlos, 13565-905, SP, Brazil.
[4] European Molecular Biology Laboratory, Meyerhofstr. 1, D-69117, Heidelberg, Germany.
[5] Sorbonne Universités, UPMC Univ Paris 06, CNRS UMR 7144, Adaptation et Diversité en Milieu Marin, Equipe EPEP, Station Biologique de Roscoff, 29680 Roscoff, France.
[6] Instituto de Ciencias Marinas y Limnológicas, Universidad Austral de Chile, Valdivia, Chile.
[7] Centro FONDAP de Investigación en Dinámica de Ecosistemas Marinos de Altas Latitudes (IDEAL), Valdivia, Chile.
[8] ETH Zurich, Institute of Microbiology, 8093 Zurich, Switzerland.
[9] Oceanography and Global Change Institute, IOCAG, University of Las Palmas de Gran Canaria, ULPGC, Gran Canaria 35214, Spain.
[10] King Abdullah University of Science and Technology (KAUST), Red Sea Research Center (RSRC), Thuwal, Saudi Arabia.
[11] Centre for Marine Ecosystems Research, School of Science, Edith Cowan University, Joondalup, WA, Australia.

*** Corresponding author:**
Ramiro Logares
Institute of Marine Sciences (ICM), CSIC, Passeig Marítim de la Barceloneta, 37-49, 08003, Barcelona, Catalonia, Spain.
e-mail: ramiro.logares@icm.csic.es. Tel: +34 93 2309500; Fax: +34 93 2309555

Manuscript for: *Microbiome*

ABSTRACT

Background

The ocean microbiota modulates global biogeochemical cycles and changes in its configuration may have largescale consequences. Yet, the underlaying ecological mechanisms structuring it are unclear. Here we investigate how fundamental ecological mechanisms (*selection*, *dispersal* and *ecological drift*) shape the smallest members of the tropical and subtropical surface-ocean microbiota: prokaryotes and minute eukaryotes (picoeukaryotes). Furthermore, we investigate the agents exerting abiotic selection on this assemblage as well as the spatial patterns emerging from the action of ecological mechanisms. To explore the previous, we analysed the composition of surface-ocean prokaryotic and picoeukaryotic communities using DNA-sequence data (16S- and 18S-rRNA genes) collected during the circumglobal expeditions *Malaspina-2010* and *TARA-Oceans*.

Results

We found that the two main components of the tropical and subtropical surface-ocean microbiota, prokaryotes and picoeukaryotes, appear to be structured by different ecological mechanisms. Picoeukaryotic communities were predominantly structured by dispersal-limitation, while prokaryotic counterparts appeared to be shaped by the combined action of dispersal-limitation, selection and drift. Temperature-driven selection appeared as a major factor influencing species co-occurrence networks in prokaryotes but not in picoeukaryotes, indicating that association patterns may contribute to understand ocean microbiota structure and response to selection. Other measured abiotic variables seemed to have limited selective effects on community

structure in the tropical and subtropical ocean. Picoeukaryotes displayed a higher spatial differentiation between communities and a higher distance decay when compared to prokaryotes, consistent with a scenario of higher dispersal limitation in the former after considering environmental heterogeneity. Lastly, random dynamics or *drift* seemed to have a more important role in structuring prokaryotic communities than picoeukaryotic counterparts.

Conclusions

The differential action of ecological mechanisms seems to cause contrasting biogeography, in the tropical and subtropical ocean, among the smallest surface plankton, prokaryotes and picoeukaryotes. This suggests that the idiosyncrasy of the main constituents of the ocean microbiota should be considered in order to understand its current and future configuration, which is especially relevant in a context of global change, where the reaction of surface ocean plankton to temperature increase is still unclear.

Keywords: microbiota, ocean, picoeukaryotes, prokaryotes, ecological processes, plankton, selection, dispersal, drift, structure

BACKGROUND

The surface ocean microbiota is a pivotal underpinning of global biogeochemical cycles [1, 2]. The smallest ocean microbes, the picoplankton, have a key role in the global carbon cycle, being responsible for an important fraction of the total atmospheric carbon and nitrogen fixation in the ocean [3-5], which supports ~46% of

the global primary productivity [6]. Oceanic picoplankton plays a fundamental role in processing organic matter by recycling nutrients and carbon to support additional production as well as by channelling organic carbon to upper trophic levels through food webs [5, 7, 8]. The ocean picoplankton includes prokaryotes (both bacteria and archaea) and tiny unicellular eukaryotes (hereafter picoeukaryotes), which feature fundamental differences in terms of cellular structure, feeding habits, metabolic diversity, growth rates and behaviour [9]. Even though marine picoeukaryotes and prokaryotes are usually investigated separately, they are intimately connected through biogeochemical and food web networks [10-12].

The underlying ecological mechanisms determining the biogeography of prokaryotes and picoeukaryotes in the global ocean are unclear. In particular, we do not know whether these crucial components of the ocean microbiota are structured by the action of the same or different ecological processes. Comprehending such processes is fundamental, as their differential action can produce changes in the ocean microbiota composition that could impact global ecosystem function [13-15]. A recent ecological synthesis explains the structure of communities and the emergence of biogeography as a consequence of the action of four main processes: *selection*, *dispersal*, *ecological drift* and *speciation* [16]. Selection involves deterministic reproductive differences among individuals from different or the same species as a response to biotic or abiotic conditions. Selection can act in two opposite directions, it can constrain (*homogeneous selection*) or promote (*heterogeneous selection*) the divergence of communities [17]. Dispersal is the movement of organisms across space, and rates can be high (*homogenising dispersal*), moderate, or low (*dispersal limitation*) [17]. Dispersal limitation occurs when species are absent from suitable

habitats because potential colonizers are too far away [18], and the significance of dispersal limitation increases as geographic scale increases [19]. Ecological drift (hereafter *drift*) in a local community refers to random changes in species' relative abundances derived from stochastic birth, death, offspring production, immigration and emigration [16]. The action of drift in a *metacommunity*, that is, local communities that are connected via dispersal of multiple species [20], may lead to neutral dynamics [19], where random dispersal is the main mechanism of community assembly. Finally, speciation is the evolution of new species [16], and it will not be considered hereafter as it is expected to have a small impact in the turnover of communities that are connected via dispersal [21], being also difficult to measure this ecological process in the wild.

The action of the previous ecological processes is typically manifested as different taxonomic or phylogenetic patterns of community turnover, that is, -diversity. At the moment, there are several estimators of -diversity which capture different aspects of community turnover [22]. Most of these indices consider taxonomic or phylogenetic aspects of communities, but not species-association patterns, which can also manifest the action of ecological processes. For example, selection exerted by an environmental variable can drive species co-occurrences generating groups of highly associated species or modules in association networks that correspond with specific environmental conditions [23]. Different members of these modules may be more abundant in specific regions of the ocean, contributing to increase -diversity estimates between these regions when based on standard compositional or phylogenetic -diversity metrics. Yet, -diversity estimates based on association-aware metrics may point to higher similarity between these regions, as

taxa belong to the same modules. Furthermore, modules may display correlations with environmental heterogeneity. Thus, association aware metrics of -diversity may allow unveiling community patterns and their relationships with environmental variables (i.e. selection), which would be missed by standard approaches [24]. So far, most studies investigating the structure of the ocean microbiota have not considered species associations in their analyses of -diversity.

The differential action of selection, dispersal and drift may generate different microbial assemblages that could feature diverse metabolisms and ecologies [14, 15]. Moderate or high selection together with moderate dispersal rates may couple environmental heterogeneity with combinations of species, leading to a spatial pattern known as *species sorting* [25]. In contrast, high or low levels of dispersal may decouple environmental heterogeneity (i.e. selection) from the composition of species assemblages. High dispersal rates may maintain populations in habitats to which they are maladapted [14, 20]. Inversely, low dispersal rates may promote microbial assemblages that become more different as the geographic distance between them increases (*distance decay*). If environmental heterogeneity and geographic distance covary, then distance decay could reflect both selection and dispersal limitation [26]. Drift is expected to cause important random effects in local community composition in cases where selection is weak and populations are small [13, 27].

Here we investigate the mechanisms that shape the surface-ocean microbiota by using DNA-sequence data collected in two of the largest circumglobal oceanographic expeditions to date, *Malaspina 2010* [28] and *TARA Oceans* [29]. Specifically, we ask: What is the relative importance of selection, dispersal and drift in structuring the surface ocean microbiota? Do these processes act similarly on main

components of this microbiota (prokaryotes and picoeukaryotes)? What are the main agents that exert abiotic selection? Do species association networks reflect the action of selection in the upper ocean microbiota? What are the main spatial-structure patterns that emerge due to the action of selection, dispersal and drift?

RESULTS

Quantifying the mechanisms that structure the surface ocean picoplankton

We analysed 16S and 18S rRNA-genes from prokaryotes and picoeukaryotes in 120 globally-distributed tropical and subtropical stations sampled during the *Malaspina 2010* expedition [28] **(Figure 1A)**. Note that the nature of the *TARA Oceans* data precluded using them in these analyses (see Methods). Operational Taxonomic Units were delineated at 99% similarity ($OTUs_{-99\%}$) and as unique sequence variants ($OTUs_{-ASVs}$, the maximum resolution for the 18S and 16S rRNA-gene). Analyses using both, $OTUs_{-99\%}$ and $OTUs_{-ASVs}$ indicated that dispersal limitation was the dominant factor structuring picoeukaryotic communities, explaining 76-67% of community turnover, while this process had a lower importance in prokaryotes (~35-25%) **[Figure 1B]**. Note that percentage refers to the percentage of pairs of communities that appear to be driven by dispersal limitation. In contrast, homogenizing dispersal had a very limited role in the structuring of the tropical and subtropical upper-ocean microbiota (<3% for both picoeukaryotes and prokaryotes). Drift had a limited role in the structuring of

picoeukaryotic communities as indicated by both $OTUs_{-99\%}$ and $OTUs_{-ASVs}$, representing ~21-6% of community turnover **(Figure 1B)**. In contrast, drift appeared as a relevant factor structuring prokaryotic communities, explaining ~44-31% of the community turnover according to $OTUs_{-99\%}$ and $OTUs_{-ASVs}$ **(Figure 1B)**. The role of selection was higher in prokaryotes compared to picoeukaryotes according to both $OTUs_{-99\%}$ and $OTUs_{-ASVs}$, explaining ~34-27% of the turnover of prokaryotic communities, and ~17-11% of that in picoeukaryotes **(Figure 1B)**. Heterogeneous selection had a relatively higher importance in structuring picoeukaryotes as compared to prokaryotes (~16-7% vs. ~9-4%, respectively). Instead, homogeneous selection appeared more important in structuring prokaryotic (~24-23%) than picoeukaryotic (~ 1-4%) communities **(Figure 1B)**.

Our quantifications indicated different roles of ecological processes in structuring communities of marine prokaryotes and picoeukaryotes populating the tropical and subtropical surface-ocean **(Figure 1B)**. We then aimed at confirming these results using other more traditional approaches. In these analyses, considering *Malaspina* data, we used $OTUs_{-99\%}$, given that these likely correspond to well-defined lineages, while $OTUs_{-ASVs}$ may reflect, in some cases, intraspecific variation [30]. We found moderate correlations between picoeukaryotic and prokaryotic -diversity (Bray Curtis: =0.58, gUniFrac: =0.61, p=0.01, Mantel tests; **Figure S2, Additional file 2**). Given that rare species tend to occupy less sites than more abundant ones [31], communities featuring different proportions of abundant or rare species may display different spatial turnover. We found that picoeukaryotes had proportionally more regionally rare (i.e. mean abundances across all samples <0.001%) species than prokaryotes (71% vs. 48% respectively) **[Table S1, Additional file 3; Figure S3,**

Additional file 4]. This is consistent with the observation that picoeukaryotes had more restricted species distributions (i.e., occurring in <20% of the stations) than prokaryotes (95% vs. 88% of the species respectively) **[Table S2, Additional file 5; Figure S3, Additional file 4]**.

Selection acting on the microbiota

We investigated the agents exerting abiotic selection on the tropical and subtropical surface-ocean microbiota by analysing -diversity together with the environmental variables included in the *Meta-119* dataset (Temperature (ºC), Conductivity (S m^{-1}), Fluorescence, Salinity and Dissolved Oxygen (ml L^{-1})). We used different indices that capture distinct facets of -diversity (Bray-Curtis, $TINA_w$, $PINA_w$, gUniFrac; see Methods). Water temperature was the most important driver of selection on prokaryotes **(Figure 2)**, ranging between 15.7 - 29.3 C, with a mean of 24.5 C and a standard deviation of 3.2 C across the whole *Meta-119 Malaspina* dataset **(Figure 1A)**. Furthermore, water temperature appeared to affect prokaryotic association networks, given that $TINA_w$ [24] explained ~50% of community variance (ADONIS R^2) **[Figure 2]**, while other used -diversity indices that do not consider species associations explained considerably lower proportions **(Figure 2)**. In contrast, temperature had limited effects on picoeukaryotic community turnover **(Figure 2)**. Analyses using both the *Malaspina* and *TARA Oceans* datasets indicated stronger positive correlations between $TINA_w$ and water-temperature differences in prokaryotes (Mantel r =0.8-0.5, $p<0.01$) than in picoeukaryotes [Mantel r =0.3, $p<0.05$] **(Figure 3)**. *TARA Oceans* samples displayed a higher correlation with water temperature than *Malaspina samples* **(Figure 3)**, which likely reflects the wider

temperature range covered by *TARA Oceans* samples (range ~0-30 C, mean ~21 C, standard deviation ~7 C) compared to *Malaspina* (range ~15-30 C, mean ~24 C, standard deviation ~3 C). Overall, $TINA_w$ results indicate that locations with similar temperatures include prokaryotic species that tend to co-occur, with this pattern disappearing as the temperature difference between stations increases. The previous pattern was either weak or non-existent in microbial eukaryotes **(Figure 3)**.

We expanded the exploration of the role of abiotic selection on microbiota structuring by analysing a larger number of environmental variables (total 17) that were available for only 57 globally distributed *Malaspina* stations (see details in **Supplementary Methods, Additional file 6**; **Figure S4, Additional file 7**). Results supported the importance of temperature-driven selection for prokaryotic community structuring **(Figure S5, Additional file 8)** and indicated that fluorescence (a proxy for Chlorophyll *a* concentration) explained 31% of $PINA_w$-based prokaryotic community variance (ADONIS R^2), being non-significant for picoeukaryotes **(Figure S5, Additional file 8)**. The remaining tested abiotic variables explained a minor fraction of community variance, suggesting that abiotic selection, at the whole ocean-microbiota level, operates via few agents, mainly temperature, although we cannot rule out that other unmeasured abiotic variables may also be exerting selection.

The different correlations between temperature and -diversity as measured by $TINA_w$ in prokaryotes and picoeukaryotes suggest that they may feature different species association networks. We found that prokaryotes sampled in both *Malaspina* and *TARA Oceans* were more associated between themselves than protists **(Figure S6, Additional file 9; Table S3, Additional file 10; Table S4, Additional file 11; Table S5, Additional file 12)**. Furthermore, the prokaryotic networks were more modular

(in terms of cliques) than the picoeukaryotic counterparts **(Table S3, Additional file 10)**, which may reflect to certain extent, temperature-driven selection [23].

Given that selection exerted by variables that lack phylogenetic signal, typically biotic variables, could inflate estimates of dispersal limitation, we have checked whether the high dispersal limitation we estimated for picoeukaryotes could reflect zooplankton grazing. For that, we have analysed globally-distributed surface *TARA Oceans* stations for which we could estimate both the community composition of picoeukaryotes (here defined as the 0.8-5 µm size-fraction; 36 or 38 stations) as well as that of microzooplankton (20-180 µm size-fraction; 36 stations) or mesozooplankton (180-2,000 µm size-fraction; 38 stations) based on 18S-rRNA genes [32]. Analyses considering abiotic (total 6, see **Supplementary Methods, Additional file 6**) and biotic (estimated zooplankton abundance) variables indicated that micro- and mesozooplankton had a minor influence on picoeukaryotic community structure (~5% of the variance explained, ADONIS R^2). In addition, the correlation between picoeukaryotic and zooplankton -diversity was either weak (microzooplankton, =0.34) or absent (mesozooplankton) [$p<0.01$, Mantel tests]. Thus, zooplankton grazing does not appear to influence -diversity in picoeukaryotes.

Selection acting on single species

The previous analyses investigated how selection may operate on the entire assemblage of species, without considering the different responses to selection that are expected in individual species. We therefore evaluated the potential action of selection on single species by determining their individual correlations with multiple abiotic environmental variables using the Maximum Information Coefficient (MIC). In the

Malaspina dataset **(Figure 1A)**, temperature was the variable with the highest number of associated prokaryotic species (1.7%), representing ~17% of the 16S rRNA gene-sequence abundance, while picoeukaryotic species displayed limited associations with temperature (~0.3% of the species representing ~5% of the 18S rRNA gene-sequence abundance) **[Figure S7, Additional file 13]**. Picoeukaryotic and prokaryotic species were also associated with oxygen, conductivity and salinity **(Figure S7, Additional file 13)**, which co-vary with temperature. The remaining variables displayed limited associations with individual prokaryotic or picoeukaryotic species **(Figure S7, Additional file 13)**, thus agreeing with our previous results suggesting that abiotic selection on the tropical and subtropical surface-ocean microbiota operates via few variables, with a dominant role for temperature among prokaryotes. Overall, prokaryotes featured proportionally more individual-species associations with environmental parameters than picoeukaryotes **(Figure S7, Additional file 13**), suggesting that environmental heterogeneity in the tropical and subtropical surface-ocean has a stronger effect on prokaryotic assemblages than on picoeukaryotic counterparts. Analyses of *TARA Oceans* data supported the previous results by indicating that prokaryotic species were associated predominantly with temperature and oxygen in the upper global ocean, while unicellular eukaryotes had weak associations to multiple variables (**Table S6, Additional file 14**).

Dispersal

Abiotic environmental conditions in adjacent stations over the trajectory of the *Malaspina* cruise, typically separated by 250-500 km, in the tropical and sub-tropical ocean **(Figure 1A)** are generally comparable [33]. Therefore, compositional

differences between pairs of neighbouring communities could manifest the differential capability of distinct microbial assemblages to disperse. Following these premises, we analysed the change in picoeukaryotic and prokaryotic community composition along the trajectory of the *Malaspina* cruise by comparing each community to the one sampled immediately before in a sequential manner (i.e. sequential -diversity) **[Figure 4]**. Both picoeukaryotic and prokaryotic communities displayed variable amounts of sequential -diversity **(Figure 4, Panels A and B)**, although picoeukaryotes featured, on average, a higher sequential -diversity than prokaryotes **(Figure 4**, **Panel C)**. This agrees with the overall mean -diversity, which was significantly higher for picoeukaryotes than for prokaryotes **(Figure S8, Additional file 15)**. Tests by subsampling the number of picoeukaryotic $OTUs_{-99\%}$ to the same number of prokaryotic ones (7,025) indicated that different numbers of $OTUs_{-99\%}$ in these groups did not affect mean Bray-Curtis estimates of -diversity displayed in **Figure S8, Additional file 15** [34].

When geographic distance covary with environmental heterogeneity, spatial community variance may be the manifestation of both selection and/or dispersal limitation. -diversity in picoeukaryotes and prokaryotes displayed positive correlations with geographic distance (i.e. distance decay) predominantly within 1,000 km **(Figure 4, Panel D)**. Yet, correlations were weaker in prokaryotes than in picoeukaryotes, pointing to stronger dispersal limitation or selection in the latter. Variance partitioning analyses considering both environmental [Temperature (ºC), Conductivity (S m^{-1}), Fluorescence, Salinity and Dissolved Oxygen (ml L^{-1})] and geographic variables (ocean basin and subdivisions, as well as Longhurst biogeographic provinces [35] **, Figure S1, Additional file 1**) indicated that in prokaryotes, geographic variables

explained most of the variance (24%), while environmental variables explained 10%, and 13% was explained by both variables; 53% of the variance remained unexplained. In contrast, picoeukaryotes displayed non-significant results in the same analyses. Still, after controlling for the effects of the most important environmental variables, Longhurst provinces (but not ocean basins nor subdivisions) accounted for ~20-25% of community variance in both picoeukaryotes and prokaryotes (ADONIS R^2) **[Figure 2]**. All in all, the previous analyses seem coherent with our quantifications of ecological processes **(Figure 1B)**, in the sense that they indicate that both selection and dispersal limitation (represented by geographic variables such as distance or ocean provinces), do seem to have a role in the structuring of the surface ocean picoplankton.

Selection and dispersal limitation may operate more strongly in geographic areas that constitute ecological boundaries, leading to abrupt changes in microbiota composition. We identified 14 communities where sequential -diversity displayed abrupt changes, with 11 of them coinciding for both picoeukaryotes and prokaryotes **(Figure 4, Panels A** & **B)**. The Local Contributions to Beta Diversity (LCBD) index [36] **(Figure S9, Additional file16)** indicated that ~22% of both picoeukaryotic and prokaryotic communities (26 stations each, totaling 36 different stations) contributed the most to the -diversity, with 16 communities coinciding for both prokaryotes and picoeukaryotes (**Figure S9, Additional file16; Table S7, Additional file 17**). In addition, 8 of the 36 stations featuring a significant LCBD were also identified as zones of abrupt community change in sequential -diversity analyses **(Table S7, Additional file 17)**. These zones point to selection or dispersal operating simultaneously and strongly upon both prokaryotic and picoeukaryotic communities in

the surface ocean.

DISCUSSION

Applying an innovative ecological framework [21] allowed us to quantify the mechanisms that shape the tropical and subtropical upper-ocean microbiota. Yet, this approach has limitations (summarised by Zhou & Ning [17]) that need to be considered in the context of our results. *First*, our results represent the overall action of ecological processes at the whole microbiota level, and not their operation on every taxonomic group or lineage (for example, different taxonomic Classes may be structured by different processes). In addition, our results reflect the action of ecological mechanisms at the global ocean level, and we expect that other spatial scales (ocean basin for example) may lead to other results. Furthermore, our results provide a snapshot of the importance of ecological processes at the global-ocean scale, and future studies should investigate how the relative importance of these mechanisms change over time [37]. *Second*, the measured ecological mechanisms are associated with the evolutionary diversification that is reflected by the variation in the chosen molecular markers. $OTUs_{-99\%}$ and $OTUs_{-ASVs}$ based on the 16S and 18S rRNA genes likely reflect defined species (or gene flow units [38]) or in some cases population variation [30], and therefore, the measured ecological mechanisms in the tropical and subtropical ocean apply to those evolutionary levels. Hence, our results do not reflect the mechanisms shaping intra-population variation or those shaping taxonomic ranks above the species level. Furthermore, our results indicate that delineating OTUs based on sequence clustering ($OTUs_{-99\%}$) or sequence variants ($OTUs_{-ASVs}$) can affect measurements of ecological mechanisms, although in our study, main trends were

maintained. It could be hypothesized that $OTUs_{-99\%}$ and $OTUs_{-ASVs}$ may represent different taxonomic units in prokaryotes or picoeukaryotes, especially if one group was evolving faster than the other. Yet, both prokaryotes and picoeukaryotes show a wide range of evolutionary rates [39, 40], including lineages evolving slow or fast, therefore potential differences in unit definitions associated to different evolutionary rates will likely compensate when analysing complex assemblages of species. *Third*, failure to detect selection could inflate estimates of dispersal limitation. We consider that our estimates indicating substantial dispersal limitation in picoeukaryotes were not inflated, as picoeukaryotes displayed more restricted spatial distributions than prokaryotes and important biotic variables, such as potential zooplankton grazing, did not seem to affect the structure of picoeukaryotic assemblages. Furthermore, another study also suggests that dispersal limitation influences protist distributions in the global ocean [32]. Altogether, the used framework [21] can be considered as a guide that can provide important insights on the ecological mechanisms structuring the global ocean microbiota, while more data (e.g. single nucleotide variants in genes or genomes) and experiments are necessary to understand such mechanisms in further detail.

Our results indicated that the differential action of ecological processes may promote different biogeographic patterns in prokaryotic and picoeukaryotic assemblages in the upper global-ocean. This is consistent with other works using similar approaches to ours indicating that protistan and bacterial assemblages are shaped by different ecological processes [37, 41-43]. In particular, selection, which is known to have an important role in structuring prokaryotic communities [25, 26], explained a higher proportion of community turnover in surface-ocean prokaryotes (~

34-27% of the turnover) than in picoeukaryotes (~17-11%). This modest role of selection in structuring the tropical and subtropical sunlit-ocean microbiota is consistent with the moderate environmental gradients characterizing this habitat. In other habitats featuring a higher selective pressure, the role of selection in structuring prokaryotes is expected to be higher [41]. The quantifications of the importance of selection are also associated to the global scale of our survey. Thus, for example, at smaller geographic scales, where dispersal limitation is expected to have a lower impact than at global scales [18], the relative importance of selection could increase. Congruently, in surface waters of the East China Sea it was found that selection was 40% more important than dispersal limitation in structuring bacterial communities [42], while in our global study, selection and dispersal limitation had a similar importance in structuring prokaryotes. Furthermore, the previous study [42] found that selection was considerably more important than dispersal limitation in structuring communities of microbial eukaryotes. In contrast, our global assessment yields dispersal limitation to be ~5 times higher than selection in structuring picoeukaryotic communities.

We found that heterogeneous selection was more important in structuring picoeukaryotic than prokaryotic communities, while homogeneous selection was more important in structuring prokaryotic than picoeukaryotic communities. This suggest that prokaryotes and picoeukaryotes respond differently to the same environmental heterogeneity, which in the tropical and subtropical surface-ocean would be preventing community divergence in prokaryotes while promoting it in picoeukaryotes. Different adaptations in prokaryotes and picoeukaryotes [9] may determine such contrasting responses to the same environmental heterogeneity. For

example, a given environmental heterogeneity could select for a few species featuring wide environmental tolerance or several species that are adapted to narrow environmental conditions.

Diverse large-scale studies have indicated that temperature is one of the main variables structuring the ocean microbiota [44-49]. Consistent with our results, Sunagawa et al. [50] reported strong correlations between prokaryotic global-ocean microbiota composition and temperature, and weak correlations with nutrients. The previous agrees with our results indicating that temperature is one of the most important agents exerting abiotic selection on the surface-ocean microbiota, although we cannot rule out the selective action of other unmeasured abiotic factors. Furthermore, temperature is one of the main abiotic variables structuring microbial assemblages in seasonal time-series, pointing also to the importance of this variable at local scales over yearly cycles [51-53]. Our analyses also unveiled an additional layer of information by indicating that temperature-driven selection affects prokaryotic taxa co-occurrences, a pattern not observed in picoeukaryotes. Such -diversity related to species associations is typically not captured by classic compositional indices like Bray Curtis, possibly due to variations in the relative abundance of the co-occurring species [54]. In contrast to prokaryotes, less is known about the effects of temperature on the community structure of ocean picoeukaryotes, which according to our results are modest. Yet, specific picoeukaryotic lineages, such as MAST-4, do seem to be affected by temperature [55], pointing to taxonomic-group specific responses to selection. One of the possible reasons why picoeukaryotes do not show co-occurrence patterns comparable to those observed in prokaryotes is dispersal limitation, which precludes picoeukaryotic species with similar niches to share the same geographic

zone. Overall, our work indicates that species association patterns are informative on the -diversity of marine prokaryotes, therefore taxa association networks should be contemplated in future analyses of the ocean microbiota.

To what extent dispersal limitation affects the distribution of ocean microbes is a matter of debate. The impact of dispersal limitation is expected to increase with increasing body size [56], therefore, larger protists are expected to be more limited by dispersal than smaller prokaryotes. Ocean protists seem to follow the previous tenet, as it has been observed that dispersal limitation appears to increase with increasing cell size [32]. Furthermore, in surface open-ocean waters, prokaryotes typically display abundances of 10^6 cells/mL, while picoeukaryotes normally have abundances of 10^3 cells/mL [57]. Due to random dispersal alone, the more abundant prokaryotes are expected to be distributed more thoroughly than the less abundant picoeukaryotes [31]. Thus, both cell size and abundance could partially explain our results indicating a higher dispersal limitation in picoeukaryotes than in prokaryotes. Yet, multiple studies of aquatic unicellular eukaryotes point to restricted dispersal [32, 58, 59], while other studies indicate the opposite [55, 60, 61]. This could reflect different dispersal capabilities among unicellular eukaryotes [58, 62] and the generation of dormant cysts in some species [63, 64], which may increase dispersal. Yet, cyst formation has not been reported for picoeukaryotes [9] and this may partially explain their limited dispersal. Regarding prokaryotes, previous studies indicate that dispersal limitation has a modest influence in the structure of marine communities [50, 65, 66] which is coherent with our results. In particular, Louca et al. [67] indicate that there is virtually no dispersal limitation in surface ocean prokaryotes within specific ocean regions, suggesting that the importance of dispersal limitation may increase across

large oceanic regions or basins. Nevertheless, dormancy in prokaryotes seems to be more common than in picoeukaryotes [9, 68], and this may allow the former to disperse more thoroughly by reducing their metabolisms when moving through unfavorable habitats [69].

The importance of drift in structuring microbial communities is unclear [25, 70]. Our results, considering both $OTUs_{-99\%}$ and $OTUs_{-ASVs}$ indicated that drift has a modest role in structuring picoeukaryotic communities in the tropical and subtropical surface ocean. In contrast, both $OTUs_{-99\%}$ and $OTUs_{-ASVs}$ indicated that drift has a more important role in structuring prokaryotic counterparts. Another study also found a larger importance of drift in determining the community structure of bacteria when compared with phytoplankton populating freshwater and brackish habitats [71]. In contrast, drift was the prevalent community-structuring mechanism in unicellular eukaryotes populating lakes in a relatively small geographic area that features a strong salinity gradient, having a low importance for the structuring of prokaryotic counterparts [41]. Drift is expected to have a more important role in small populations, which is normally not the case for microbes in the global ocean. Yet, other random processes could have effects similar to drift in large microbial populations. For example, the evolution or arrival of a new bacteriophage may attack abundant bacteria in a local community, randomly reshuffling species abundances. Likely, the importance of drift in structuring microbiotas is dependent on taxa adaptations (e.g. the presence of habitat generalists or specialists), selection strength, ecological redundancy, as well as on the dimensions of the analysed habitats.

A decrease in community similarity with geographic distance (distance decay) can be the manifestation of selection and/or dispersal limitation [26]. Distance decay

has been evidenced in diverse studies focusing on the surface and deep ocean microbiotas [65, 72, 73]. In our study, we have used different analyses (variance partitioning and ADONIS based on measured environmental variables and geographic features; see Methods) to interpret the measured distance decay. Variance partitioning suggested that both geography (i.e. dispersal limitation) and environmental variation (selection) likely explain distance decay in prokaryotes, with geography having potentially a more important role, which agrees with our ADONIS analyses based on Bray Curtis and gUnifrac distances **(Figure 2)**. Interestingly, variance partitioning was not significant in picoeukaryotes, although ADONIS analyses based on Bray Curtis and gUnifrac distances indicated that geography, and to a lesser extent temperature, would partially explain picoeukaryotic distance decay **(Figure 2)**. Overall, provincialism, as measured by Longhurst provinces **(Figure S1, Additional file 1)** was the most relevant spatial feature for the community structuring of both prokaryotes and picoeukaryotes **(Figure 2).** Possibly, this reflects dispersal limitation, as the selective effects of main environmental variables that covary with these provinces were considered during ADONIS analyses. Longhurst provinces may also reflect, to certain extent, different water masses or currents that impose restrictions to microbial dispersal. Yet, it may be possible that dispersal limitation in picoeukaryotes is only partially reflected by provincialism, thus explaining the lack of significance in variance partitioning analyses as well as the differences between the magnitudes of dispersal limitation suggested by provincialism analyses using ADONIS **(Figure 2)** and those estimated by analyses of ecological processes **(Figure 1B)**. Alternatively, dispersal limitation in picoeukaryotes may be better reflected by geographic distances between communities, as suggested by sequential Bray-Curtis analyses **(Figure 4C)** as

well as their stronger distance decay when compared to prokaryotes **(Figure 4D)**. Interestingly, a study investigating surface marine bacteria along ~12,000 km in the Atlantic Ocean found that provincialism explained an amount of community variance comparable to our results [65]. Furthermore, and consistent with our results, a study of the sunlit global-ocean eukaryotic microbiota indicated that basin, which may be associated to provincialism and dispersal limitation, was one of the most important variables explaining community turnover [32]. In sum, Longhurst provinces, as a spatial feature, seem to partially reflect the dispersal limitation measured for upper-ocean picoplankton.

In the surface ocean, drastic changes in microbial species composition across space may point to strong changes in abiotic selection (as expected to occur across oceanographic fronts [74, 75]), or high immigration. We identified 14 stations featuring abrupt changes in prokaryotic or picoeukaryotic community composition as well as 36 stations with a "unique" species composition. Some of these areas correspond to nutrient-rich (selection) coastal zones (the South African Atlantic coast and the South Australia Bight) or potential upwelling (dispersal) zones, such as the Equatorial Pacific and Atlantic as well as the Costa Rica Dome. These findings were coherent with Spatial Abundance Distributions (SpAD) of bacterioplankton in the tropical and subtropical surface-ocean [33]. Altogether, the previous suggests strong selective changes or immigration from deep water layers into the surface associated to upwellings, affecting both prokaryotic and picoeukaryotic community structure. Such immigration events into the surface, when random, may partially explain the measured drift.

CONCLUSION

Our results indicate that selection, dispersal and drift have different roles in shaping the main components of the picoplankton (prokaryotes and picoeukaryotes) in the tropical and subtropical surface ocean. This highlights the importance of comprehending the characteristics of the different constituents of microbiotas in order to understand their structure. Our results also suggest that the surface ocean picoplankton may not show a single response to global change, and that perhaps prokaryotes will display more pronounced changes in their community structure as a response to temperature increase than picoeukaryotes, considering that temperature seems to affect more prokaryotic than picoeukaryotic assemblages. Future studies on the ocean microbiota should investigate the change in the role of selection, dispersal and drift with ocean scale (from meters to kilometers), depth, latitude and longitude as well as with time, taxonomic ranks (e.g. Class, Family, etc.) and molecular markers that evolve at different rates. Such studies will likely provide a more comprehensive understanding of the underlaying mechanisms shaping the ocean microbiota at different evolutionary levels (from lineages to populations) and will also provide insights on the environmental variables that could modify its current configuration.

METHODS

Sample collection

Surface waters (3 m depth) from a total of 120 globally-distributed stations located in the tropical and sub-tropical ocean **(Figure 1A)** were sampled as part of the *Malaspina 2010* expedition [28]. Sampling took place between December 2010 and July 2011 and the cruise was organized in a way so that most regions were sampled

during similar meteorological seasons. Samples were obtained with a 20 L Niskin bottle deployed simultaneously to a CTD profiler that measured conductivity, temperature, oxygen, fluorescence and turbidity for each sample. About 12 L of seawater were sequentially filtered through a 20 µm nylon mesh, followed by a 3 µm and 0.2 µm polycarbonate filters of 47 mm diameter (Isopore, Millipore, Burlington, MA, USA). Only the smallest size-fraction (0.2 -3 µm, here called "picoplankton" [8]) was used in downstream analyses. Samples for inorganic nutrients (NO_3^-, NO_2^-, PO_4^{3-}, SiO_2) were collected from the Niskin bottles and measured spectrophotometrically using an Alliance Evolution II autoanalyzer (Frépillon, France) [76]. Chlorophyll measurements were obtained from Estrada et al. [77]. In specific samples nutrient concentrations were estimated using the World Ocean Database [78] due to issues with the measurements. Since not all environmental parameters were available for all stations, two contextual datasets were generated: *Meta-119*, including 119 stations, 5 environmental parameters and 5 spatial features **(**all except one station in **Figure 1A)** and *Meta-57* **(Figure S4, Additional file 7)**, including 57 stations and 17 environmental parameters (the 5 environmental parameters included in *Meta-119* were considered here as well). See **Supplementary Methods, Additional file 6**.

DNA extraction, sequencing and bioinformatics

DNA was extracted using a standard phenol-chloroform protocol [79]. Both the 18S and 16S rRNA-genes were amplified from the same DNA extracts. The hypervariable V4 region of the 18S rRNA gene (~380 bp) was amplified with the primers TAReukFWD1 and TAReukREV3 [80], while the hypervariable V4-V5 (~400bp) region of the 16S rRNA gene was amplified with the primers 515F-Y - 926R [81],

which target both Bacteria and Archaea. Amplicon libraries were then paired-end sequenced on an *Illumina* (San Diego, CA, USA) MiSeq platform (2x250bp) at the Research and Testing Laboratory facility (http://www.researchandtesting.com/).

Reads were processed following and in-house protocol [82]. Operational Taxonomic Units (OTUs) were delineated at 99% similarity using UPARSE V8.1.1756 [83], producing 42,505 picoeukaryotic and 10,158 prokaryotic $OTUs_{-99\%}$. Taxonomic assignment of $OTUs_{-99\%}$ was generated by BLASTing OTU-representative sequences against different reference databases. BLAST hits were filtered prior to taxonomy assignment using an in-house python script, considering a percentage of identity >90%, a coverage >70%, a minimum alignment length of 200 bp and an e-value > 0.00001. Metazoan, Streptophyta, nucleomorphs, Chloroplast and mitochondrial OTUs were removed from the $OTUs_{-99\%}$ tables. See **Supplementary Methods, Additional file 6** and **Table S8, Additional file 18**.

Additionally, to investigate the effects of clustering on the estimation of ecological mechanisms **(Fig. 1B)**, we determined OTUs as Amplicon Sequence Variants (ASVs) using DADA2 [84]. For the 18S, we trimmed the forward reads at 240 bp and the reverse reads at 180 bp, while for the 16S, forward reads were trimmed at 220 bp and reverse reads at 200 bp. Then, for the 18S, the maximum number of expected errors (maxEE) was set to 12 and 20 for the forward and reverse reads respectively, while for the 16S, the maxEE was set to 2 for the forward reads and to 4 for the reverse reads. $OTUs_{-ASVs}$ were assigned taxonomy using the naïve Bayesian classifier method [85] together with the SILVA version 132 [86] database as implemented in DADA2. Eukaryotic $OTUs_{-ASVs}$ were also BLASTed [87] against the Protist Ribosomal Reference database (PR^2, version 4.11.1; [88]). Streptophyta,

Metazoa, nucleomorphs, chloroplasts and mitochondria were removed from OTUs$_{-ASVs}$ tables. Tables of OTUs$_{-ASVs}$ were rarefied to 20,000 reads per sample with the function *rrarefy* in Vegan. Only OTUs$_{-ASVs}$ with abundances >100 reads were used for the calculation of ecological mechanisms **(Fig. 1B)**.

We used publicly-available data from the *TARA Oceans* global expedition [29] in multiple analyses. This expedition took place between September 2009 - March 2012, and includes samples from the same hemisphere during different meteorological seasons. Due to the nature of the *TARA Oceans* dataset, we did not perform all the analyses that were run for the *Malaspina* dataset. Specifically, short V9 18S rRNA-gene reads or 16S rRNA-gene miTags [89] from *TARA Oceans* precluded robust phylogenetic reconstructions, which instead were possible with the longer reads produced for *Malaspina.* We used data from *TARA Oceans* surface (~5 m depth) stations only, including 41 samples (40 stations) for pico-nano eukaryotes (0.22-3 m [1 sample] and 0.8-5 m [40 samples]; 18S-V9 rRNA gene amplicon data) [32] as well as 63 stations for prokaryotes (picoplankton, 0.22-3 m [45 samples] and 0.22-1.6 m [18 samples]; 16S rRNA genes, miTags) [50].

General analyses and phylogenetic inferences

Tables including OTUs$_{-99\%}$ were sub-sampled to 4,060 reads per sample using *rrarefy* in *Vegan* [90], resulting in sub-sampled tables containing 18,775 picoeukaryotic and 7,025 prokaryotic OTUs. OTUs$_{-99\%}$ with mean relative abundances >0.1% or <0.001% were defined as regionally abundant or rare respectively [91]. Phylogenetic trees were constructed by aligning 16S or 18S OTUs$_{-99\%}$ representative sequences against an aligned SILVA [86] template using mothur [92]. Afterwards, poorly aligned regions

or sequences were removed using trimAl [93]. A phylogenetic tree was inferred using FastTree v2.1.9 [94]. Most analyses were performed in the R statistical environment [95] using *APE* [96], *ggplot2* [97], *gUniFrac* [98], *Maps*, *Mapplots*, *Picante* [99] and *Vegan*. The *Vegan* function *adonis* and *adonis2* were used to investigate the amount of variance in community composition explained by environmental or geographic variables. Variance partitioning analyses were run with *varpart* in *Vegan* and tested for significance with ANOVA. Distance decay, which refers to the decrease in microbial community similarity as geographic distance between communities increases was investigated in R using Mantel correlograms between geographic distance and -diversity, considering distance classes of 1,000 km. Local Contributions to Beta Diversity (LCBD) [36], which indicates the degree of uniqueness of each community in terms of its species composition, was measured with *adespatial* [100]. See **Supplementary Methods, Additional file 6**.

Quantification of selection, dispersal and drift

These processes were quantified using an approach that relies on null models, consisting of two main sequential steps: the first uses OTU phylogenetic turnover to infer the action of selection and the second uses OTU compositional turnover to infer the action of dispersal and drift [21]. The action of selection, dispersal and drift was quantified using both $OTUs_{-99\%}$ and $OTUs_{-ASVs}$. In order to determine the action of selection using phylogenetic turnover, we first checked whether habitat preferences of phylogenetically closely related taxa (according to the 16S and 18S rRNA-genes) were more similar to each other than to those of more distantly related taxa, what is known as *phylogenetic signal* [101, 102]. We tested for phylogenetic signal using

temperature and fluorescence, which were the two variables that explained the highest fraction of community variance. We detected phylogenetic signal at relatively short phylogenetic distances **(Figure S10, Additional file 19; Figure S11, Additional file 20)**, which is coherent with previous work [21, 103, 104]. We measured phylogenetic turnover using the abundance-weighted -Mean Nearest Taxon Distance (MNTD) metric [17, 21], which quantifies the mean phylogenetic distances between the evolutionary-closest OTUs in two communities. MNTD values can be larger, smaller or equal to the values expected when selection is not affecting community turnover (that is, expected by chance). MNTD values higher than expected by chance indicate that communities experience heterogeneous selection [17]. In contrast, MNTD values which are lower than expected by chance indicate that communities experience homogeneous selection. Null models included 999 randomizations [21]. Differences between the observed MNTD and the mean of the null distribution are denoted as -Nearest Taxon Index (NTI), with |NTI|>2 being considered as significant departures from random phylogenetic turnover, pointing to the action of selection.

The second step uses OTU turnover to calculate whether the -diversity of communities not structured by selection could be generated by drift (i.e. chance) or dispersal. We calculated the Raup-Crick metric [105] using Bray-Curtis dissimilarities (hereafter RC_{bray}) [21]. RC_{bray} compares the measured -diversity against the -diversity that would be obtained under random community assembly (drift); randomizations were run 9,999 times. RC_{bray} values between -0.95 and +0.95 point to a community assembly governed by drift. On the contrary, RC_{bray} values >+0.95 or <-0.95 indicate that community turnover is driven by dispersal limitation or homogenizing dispersal respectively [105]. See **Supplementary Methods, Additional file 6**.

Estimation of interaction-adjusted indices

Taxa INteraction-Adjusted (TINA) and Phylogenetic INteraction Adjusted (PINA) indices were estimated following Schmidt et al. [24]. TINA is based on taxa co-occurrences while PINA considers phylogenetic similarities. TINA quantifies -diversity as the average association strength between all taxa in different samples. Thus, communities which are identical or include taxa that are perfectly associated will give a TINA value of 1. TINA values will approach 0.5 in communities sharing no taxa or having neutral associations, and approach 0 if taxa display high avoidance. Dissimilarity matrices were generated as 1-TINA and used in downstream analyses **(**e.g. **Figure 3)**. Full picoeukaryotic and prokaryotic subsampled $OTU_{-99\%}$ tables were used to calculate the abundance-weighted $TINA_w$ and $PINA_w$. $TINA_w$ was calculated using picoeukaryotic and prokaryotic data from 119 *Malaspina* surface stations (most stations in **Figure 1A**). In addition, $TINA_w$ was calculated using data from *TARA Oceans*, including 63 surface stations for prokaryotes and 40 surface station for small unicellular eukaryotes **(Figure 3)**.

Associations between taxa and environmental parameters

We analysed whether $OTUs_{-99\%}$ displayed associations with environmental variables and between themselves. Firstly, we used the Maximum Information Coefficient (MIC) which captures diverse relationships between two pairs of variables [106]. The *Malaspina* dataset consisted of 119 stations and 17 environmental variables. In the *TARA Oceans* dataset, prokaryotes were analysed across 63 surface stations (including 8 environmental variables), while microbial eukaryotes were analysed across 40 surface stations (including 6 environmental variables) [see **Supplementary Methods,**

Additional file 6]. In both datasets, MIC analyses were run using CV=0.5, B=0.6, and statistically significant relationships with MIC ≥0.4 (*Malaspina*) or MIC ≥0.5 (*TARA Oceans*) were considered (MIC thresholds were adjusted to the characteristics of the datasets). MIC significance was assessed using precomputed p-values [106]. Secondly, we constructed association networks with the *Malaspina* dataset considering $OTUs_{-99\%}$ with >100 reads using SparCC [107] as implemented in FastSpar [108]. To determine correlations, FastSpar was run with 1,000 iterations, including 1,000 bootstraps to infer p-values. We used $OTUs_{-99\%}$ associations with absolute correlation scores >0.3 and $p<0.01$. Networks were visualized and analysed with Cytoscape [109] and igraph [110].

DECLARATIONS

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

DNA sequences and metadata from the *Malaspina* expedition are publicly available at the European Nucleotide Archive (http://www.ebi.ac.uk/ena; accession numbers PRJEB23913 [18S rRNA genes] & PRJEB25224 [16S rRNA genes]). The data used from *TARA Oceans* is publicly available through *Pangaea*

(https://doi.org/10.1594/PANGAEA.873275) as well as in (http://ocean-microbiome.embl.de/companion.html) [32, 50]. The code for generating $OTUs_{-99\%}$ tables is available in: https://doi.org/10.5281/zenodo.259579. R-Scripts for calculating the -Nearest Taxon Index and the Raup-Crick metric are available in https://github.com/stegen/Stegen_etal_ISME_2013. The code for calculating the TINA and PINA indices is available in https://github.com/defleury/Schmidt_et_al_2016_community_similarity, while the code for calculating MIC is available at http://www.exploredata.net. All used *R* packages as well as other software are cited in Methods.

Competing interests

The authors declare that they have no competing interests

Funding

RL was supported by a Ramón y Cajal fellowship (RYC-2013-12554, MINECO, Spain). IMD was supported by an ITN-SINGEK fellowship (ESR2-EU-H2020-MSCA-ITN-2015, Grant Agreement 675752 [ESR2] to RL), PCJ by Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP (PhD grant 2017/26786-1) and CR-González by a Juan de la Cierva (IJCI-2015-23505, MINECO, Spain) fellowship. MM was partially supported by CONICYT (FONDAP-IDEAL 15150003), Chile. This work was supported by the projects *Malaspina 2010* Expedition (CSD2008-00077, MINECO, Spain to CMD), INTERACTOMICS (CTM2015-69936-P, MINECO, Spain to RL), REMEI (CTM2015-70340-P, MINECO, Spain to JMG) and MicroEcoSystems (240904, RCN, Norway to RL).

Authors' contributions

RL designed the study. SGA, CMD, JMG and RM organized the sampling expedition. CR-Giner, GS, and MS collected samples, extracted the DNA and organized its sequencing. RL, IMD, AKK, PCJ analysed the DNA data, while CR-Giner, TSBS, LRR, MM, GS, CR-González, MS, CdV, RM, and JMG provided contextual ecological or environmental pre-processed data. RL, IMD, AKK, PCJ, TSBS, GS, CdV and RM interpreted the results. RL wrote the manuscript. All authors contributed substantially to manuscript revisions. All authors read and approved the final manuscript.

Acknowledgements

We thank all scientists from the *Malaspina 2010* expedition and crews from the R/V Hespérides. Bioinformatics analyses were performed at the MARBITS platform of the Institut de Ciències del Mar (ICM; http://marbits.icm.csic.es) as well as in MareNostrum (Barcelona Supercomputing Center) via grants obtained from the Spanish Network of Supercomputing (RES) to RL. We thank the CSIC Open Access Publication Support Initiative through the Unit of Information Resources for Research (URICI) for helping to cover publication fees.

REFERENCES

1. Falkowski P. The power of plankton. Nature. 2012; 483(7387):S17-20.
2. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's

biogeochemical cycles. Science. 2008; 320(5879):1034-1039.

3. Jardillier L, Zubkov MV, Pearman J, Scanlan DJ. Significant CO_2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. ISME J. 2010; 4(9):1180-1192.
4. Li WKW. Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. Limnology and Oceanography. 1994; 39(1):169-175.
5. Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ. Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes. Science. 2015; 347(6223):1257594.
6. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. Primary production of the biosphere: integrating terrestrial and oceanic components. Science. 1998; 281(5374):237-240.
7. del Giorgio PA, Duarte CM. Respiration in the open ocean. Nature. 2002; 420(6914):379-384.
8. Massana R. Eukaryotic picoplankton in surface oceans. Annual review of microbiology. 2011; 65:91-110.
9. Massana R, Logares R. Eukaryotic versus prokaryotic marine picoplankton ecology. Environ Microbiol. 2013; 15(5):1254-1261.
10. Massana R. Picoeukaryotes. In: *Encyclopedia of Microbiology.* Edited by Schaechter M. Oxford: Elsevier; 2009: 674-688.
11. Seymour JR, Amin SA, Raina JB, Stocker R. Zooming in on the phycosphere: the ecological interface for phytoplankton-bacteria relationships. Nat Microbiol. 2017; 2:17065.

12. Jürgens K, Massana R. Protistan grazing on marine bacterioplankton, 2nd edn. Hoboken, New Jersey: Wiley-Blackwell; 2008.
13. Nemergut DR, Schmidt SK, Fukami T, O'Neill SP, Bilinski TM, Stanish LF, Knelman JE, Darcy JL, Lynch RC, Wickey P *et al*. Patterns and processes of microbial community assembly. Microbiol Mol Biol Rev. 2013; 77(3):342-356.
14. Leibold MA, Chase JM, Ernest SK. Community assembly and the functioning of ecosystems: how metacommunity processes alter ecosystems attributes. Ecology. 2017; 98(4):909-919.
15. Mori AS, Isbell F, Seidl R. beta-Diversity, Community Assembly, and Ecosystem Functioning. Trends Ecol Evol. 2018; 33(7):549-564.
16. Vellend M. The theory of ecological communities. Princeton: Princeton University Press; 2016.
17. Zhou J, Ning D. Stochastic Community Assembly: Does It Matter in Microbial Ecology? Microbiol Mol Biol Rev. 2017; 81(4):e00002-00017.
18. Heino J, Melo AS, Siqueira T, Soininen J, Valanko S, Bini LM. Metacommunity organisation, spatial extent and dispersal in aquatic systems: patterns, processes and prospects. Freshwater Biology. 2015; 60(5):845-869.
19. Hubbell SP. A unified neutral theory of biodiversity and biogeography. Princeton, NJ: Princeton University Press; 2001.
20. Holyoak M, Leibold MA, Holt RD. Metacommunities: Spatial Dynamics and Ecological Communities. Chicago: The University of Chicago Press; 2005.
21. Stegen JC, Lin X, Fredrickson JK, Chen X, Kennedy DW, Murray CJ, Rockhold ML, Konopka A. Quantifying community assembly processes and

identifying features that impose them. ISME J. 2013; 7(11):2069-2079.

22. Magurran AE, McGill BJ. Biological Diversity: Frontiers in measurements and assessment: Oxford University Press; 2011.

23. Röttjers L, Faust K. From hairballs to hypotheses - biological insights from microbial networks. FEMS Microbiol Rev. 2018.

24. Schmidt TS, Matias Rodrigues JF, von Mering C. A family of interaction-adjusted indices of community similarity. ISME J. 2017; 11(3):791-807.

25. Lindström ES, Langenheder S. Local and regional factors influencing bacterial community assembly. Environ Microbiol Rep. 2012; 4:1-9.

26. Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JB. Beyond biogeographic patterns: processes shaping the microbial landscape. Nature reviews Microbiology. 2012; 10(7):497-506.

27. Fisher CK, Mehta P. The transition between the niche and neutral regimes in ecology. Proc Natl Acad Sci U S A. 2014; 111(36):13111-13116.

28. Duarte CM. Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. Limnology and Oceanography Bulletin. 2015; 24(1):11-14.

29. Karsenti E, Acinas SG, Bork P, Bowler C, De Vargas C, Raes J, Sullivan M, Arendt D, Benzoni F, Claverie JM *et al*. A holistic approach to marine eco-systems biology. PLoS biology. 2011; 9(10):e1001177.

30. Caron DA, Hu SK. Are We Overestimating Protistan Diversity in Nature? Trends Microbiol. 2019; 27(3):197-205.

31. Gaston KJ, Blackburn TM, Greenwood JJD, Gregory RD, Quinn RM, Lawton JH. Abundance–occupancy relationships. Journal of Applied Ecology. 2000;

37(s1):39-59.

32. de Vargas C, Audic S, Henry N, Decelle J, Mahe F, Logares R, Lara E, Berney C, Le Bescot N, Probert I *et al*. Eukaryotic plankton diversity in the sunlit ocean. Science. 2015; 348(6237):1261605.
33. Ruiz-Gonzalez C, Logares R, Sebastian M, Mestre M, Rodriguez-Martinez R, Gali M, Sala MM, Acinas SG, Duarte CM, Gasol JM. Higher contribution of globally rare bacterial taxa reflects environmental transitions across the surface ocean. Mol Ecol. 2019; 28(8):1930-1945.
34. Kraft NJ, Comita LS, Chase JM, Sanders NJ, Swenson NG, Crist TO, Stegen JC, Vellend M, Boyle B, Anderson MJ *et al*. Disentangling the drivers of beta diversity along latitudinal and elevational gradients. Science. 2011; 333(6050):1755-1758.
35. Longhurst AR. Ecological Geography of the Sea: Academic Press; 2007.
36. Legendre P, De Caceres M. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. Ecol Lett. 2013; 16(8):951-963.
37. Vass M, Székely AJ, Lindström ES, Langenheder S. Using null models to compare bacterial and microeukaryotic metacommunity assembly under shifting environmental conditions. bioRxiv. 2019.
38. Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF. A Reverse Ecology Approach Based on a Biological Definition of Microbial Populations. Cell. 2019; 178(4):820-834.e814.
39. Pernice MC, Logares R, Guillou L, Massana R. General patterns of diversity in major marine microeukaryote lineages. PLoS One. 2013; 8(2):e57170.
40. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ,

Fourment M, Holmes EC. Genome-scale rates of evolutionary change in bacteria. Microb Genom. 2016; 2(11):e000094-e000094.

41. Logares R, Tesson SVM, Canback B, Pontarp M, Hedlund K, Rengefors K. Contrasting prevalence of selection and drift in the community structuring of bacteria and microbial eukaryotes. Environ Microbiol. 2018; 20(6):2231-2240.

42. Wu W, Lu HP, Sastri A, Yeh YC, Gong GC, Chou WC, Hsieh CH. Contrasting the relative importance of species sorting and dispersal limitation in shaping marine bacterial versus protist communities. ISME J. 2018; 12(2):485-494.

43. Brislawn CJ, Graham EB, Dana K, Ihardt P, Fansler SJ, Chrisler WB, Cliff JB, Stegen JC, Moran JJ, Bernstein HC. Forfeiting the priority effect: turnover defines biofilm community succession. ISME J. 2019; 13(7):1865-1877.

44. Sul WJ, Oliver TA, Ducklow HW, Amaral-Zettler LA, Sogin ML. Marine bacteria exhibit a bipolar distribution. Proc Natl Acad Sci U S A. 2013; 110(6):2342-2347.

45. Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH. A latitudinal diversity gradient in planktonic marine bacteria. Proc Natl Acad Sci U S A. 2008; 105(22):7774-7778.

46. Pommier T, Canbäck B, Riemann L, Boström H, Simu K, Lundberg P, Tunlid A, Hagström Å. Global patterns of diversity and community structure in marine bacterioplankton. Molecular ecology 2007; 16:867-880.

47. Rutherford S, D'Hondt S, Prell W. Environmental controls on the geographic distribution of zooplankton diversity. Nature. 1999; 400(6746):749-753.

48. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW.

Niche partitioning among Prochlorococcus ecotypes along ocean-scale environmental gradients. Science. 2006; 311(5768):1737-1740.

49. Raes J, Letunic I, Yamada T, Jensen LJ, Bork P. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. Molecular Systems Biology. 2011; 7(1):473.
50. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A *et al*. Structure and function of the global ocean microbiome. Science. 2015; 348(6237):1261359.
51. Giner CR, Balague V, Krabberod AK, Ferrera I, Rene A, Garces E, Gasol JM, Logares R, Massana R. Quantifying long-term recurrence in planktonic microbial eukaryotes. Mol Ecol. 2019; 28(5):923-935.
52. Lambert S, Tragin M, Lozano J-C, Ghiglione J-F, Vaulot D, Bouget F-Y, Galand PE. Rhythmicity of coastal marine picoeukaryotes, bacteria and archaea despite irregular environmental perturbations. ISME J. 2019; 13(2):388-401.
53. Bunse C, Pinhassi J. Marine Bacterioplankton Seasonal Succession Dynamics. Trends Microbiol. 2017; 25(6):494-505.
54. Chase JM. Community assembly: when should history matter? Oecologia. 2003; 136(4):489-498.
55. Rodriguez-Martinez R, Rocap G, Salazar G, Massana R. Biogeography of the uncultured marine picoeukaryote MAST-4: temperature-driven distribution patterns. ISME J. 2013; 7(8):1531-1543.
56. De Bie T, De Meester L, Brendonck L, Martens K, Goddeeris B, Ercken D, Hampel H, Denys L, Vanhecke L, Van der Gucht K *et al*. Body size and

dispersal mode as key traits determining metacommunity structure of aquatic organisms. Ecology Letters. 2012; 15(7):740-747.

57. Kirchman DL. Microbial Ecology of the Oceans. Hoboken, New Jersey: John Wiley & Sons; 2008.

58. Foissner W. Biogeography and Dispersal of Micro-organisms: A Review Emphasizing Protists. Acta Protozoologica. 2006; 45:111-136.

59. Casteleyn G, Leliaert F, Backeljau T, Debeer AE, Kotaki Y, Rhodes L, Lundholm N, Sabbe K, Vyverman W. Limits to gene flow in a cosmopolitan marine planktonic diatom. Proc Natl Acad Sci U S A. 2010; 107(29):12952-12957.

60. Cermeno P, Falkowski PG. Controls on diatom biogeography in the ocean. Science. 2009; 325(5947):1539-1541.

61. Whittaker KA, Rynearson TA. Evidence for environmental and ecological selection in a microbe with no geographic limits to gene flow. Proc Natl Acad Sci U S A. 2017; 114(10):2651-2656.

62. Bass D, Richards TA, Matthai L, Marsh V, Cavalier-Smith T. DNA evidence for global dispersal and probable endemicity of protozoa. BMC Evol Biol. 2007; 7(1):162.

63. Lewis J, Harris ASD, Jones KJ, Edmonds RL. Long-term survival of marine planktonic diatoms and dinoflagellates in stored sediment samples. J Plankton Res. 1999; 21(2):343-354.

64. Billard C, Inouye I. What is new in coccolithophore biology? In: *Coccolithophores: From Molecular Processes to Global Impact.* Edited by Thierstein HR, Young JR. Berlin, Heidelberg: Springer Berlin Heidelberg;

2004: 1-29.

65. Milici M, Tomasch J, Wos-Oxley ML, Decelle J, Jauregui R, Wang H, Deng ZL, Plumeier I, Giebel HA, Badewien TH *et al*. Bacterioplankton Biogeography of the Atlantic Ocean: A Case Study of the Distance-Decay Relationship. Front Microbiol. 2016; 7:590.

66. Sintes E, De Corte D, Ouillon N, Herndl GJ. Macroecological patterns of archaeal ammonia oxidizers in the Atlantic Ocean. Mol Ecol. 2015; 24(19):4931-4942.

67. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome. Science. 2016; 353(6305):1272-1277.

68. Jones SE, Lennon JT. Dormancy contributes to the maintenance of microbial diversity. Proc Natl Acad Sci U S A. 2010; 107(13):5881-5886.

69. Locey KJ. Synthesizing traditional biogeography with microbial ecology: the importance of dormancy. Journal of Biogeography. 2010; 37(10):1835-1841.

70. Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, O'Connor MI, Ackermann M, Hahn AS, Srivastava DS, Crowe SA *et al*. Function and functional redundancy in microbial systems. Nature Ecology & Evolution. 2018; 2:936-943.

71. Östman Ö, Drakare S, Kritzberg ES, Langenheder S, Logue JB, Lindström ES. Regional invariance among microbial communities. Ecology letters. 2010; 13(1):118-127.

72. Salazar G, Cornejo-Castillo FM, Benitez-Barrios V, Fraile-Nuez E, Alvarez-Salgado XA, Duarte CM, Gasol JM, Acinas SG. Global diversity and biogeography of deep-sea pelagic prokaryotes. ISME J. 2016; 10(3):596-608.

73. Zinger L, Boetius A, Ramette A. Bacterial taxa–area and distance–decay relationships in marine environments. Mol Ecol. 2014; 23(4):954-964.

74. Díez B, Massana R, Estrada M, Pedrós-Alió C. Distribution of eukaryotic picoplankton assemblages across hydrographic fronts in the Southern Ocean, studied by denaturing gradient gel electrophoresis. Limnology and Oceanography. 2004; 49(4):1022-1034.

75. Flaviani F, Schroeder D, Lebret K, Balestreri C, Schroeder J, Moore K, Paszkiewicz K, Pfaff M, Rybicki E. Distinct oceanic microbiomes (from viruses to protists) found either side of the Antarctic Polar Front. Front Microbiol. 2018; 9.

76. Grasshoff K, Ehrhardt M, Kremling K. Methods on seawater analysis; 1983.

77. Estrada M, Delgado M, Blasco D, Latasa M, Cabello AM, Benitez-Barrios V, Fraile-Nuez E, Mozetic P, Vidal M. Phytoplankton across Tropical and Subtropical Regions of the Atlantic, Indian and Pacific Oceans. PLoS One. 2016; 11(3):e0151699.

78. Boyer TP, Antonov JI, Baranova OK, Coleman C, Garcia HE, Grodsky A, Johnson DR, Locarnini RA, Mishonov AV, O'Brien TD *et al*. World Ocean Database 2013. In: *NOAA Atlas NESDIS 72*. Edited by Levitus S, Mishonov A. Silver Spring, MD: NOAA; 2013.

79. Massana R, Murray AE, Preston CM, DeLong EF. Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. Appl Environ Microbiol. 1997; 63(1):50-56.

80. Stoeck T, Bass D, Nebel M, Christen R, Jones MD, Breiner HW, Richards TA. Multiple marker parallel tag environmental DNA sequencing reveals a

highly complex eukaryotic community in marine anoxic water. Mol Ecol. 2010; 19 Suppl 1:21-31.

81. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol. 2016; 18(5):1403-1414.
82. Logares R. Workflow for Analysing MiSeq Amplicons based on Uparse v1.5. In.: https://doi.org/10.5281/zenodo.259579; 2017.
83. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods. 2013; 10(10):996-998.
84. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJ, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016; 13(7):581-583.
85. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007; 73(16):5261-5267.
86. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013; 41(Database issue):D590-596.
87. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990; 215(3):403-410.
88. Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, de Vargas C, Decelle J *et al*. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA

sequences with curated taxonomy. Nucleic Acids Res. 2013; 41(Database issue):D597-604.

89. Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, Hingamp P, Ogata H, de Vargas C, Lima-Mendez G *et al*. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. Environ Microbiol. 2014; 16(9):2659-2671.

90. Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Stevens MHH, Wagner H. vegan: Community Ecology Package. R package version 1.15-0. In.; 2008.

91. Logares R, Audic S, Bass D, Bittner L, Boutte C, Christen R, Claverie JM, Decelle J, Dolan JR, Dunthorn M *et al*. Patterns of rare and abundant marine microbial eukaryotes. Current Biology. 2014; 24(8):813-821.

92. Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ *et al*. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol. 2009; 75(23):7537-7541.

93. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009; 25(15):1972-1973.

94. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009; 26(7):1641-1650.

95. R-Development-Core-Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2008.

96. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004; 20(2):289-290.

97. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag; 2009.

98. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics. 2012; 28(16):2106-2113.

99. Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. Picante: R tools for integrating phylogenies and ecology. Bioinformatics. 2010; 26(11):1463-1464.

100. Dray S, Blanchet G, Borcard D, Clappe S, Guenard G, Jombart T, Larocque G, Legendre P, Madi N, Wagner HH. adespatial: Multivariate Multiscale Spatial Analysis. In.; 2017.

101. Cavender-Bares J, Kozak KH, Fine PV, Kembel SW. The merging of community ecology and phylogenetic biology. Ecology letters. 2009; 12(7):693-715.

102. Losos JB. Phylogenetic niche conservatism, phylogenetic signal and the relationship between phylogenetic relatedness and ecological similarity among species. Ecol Lett. 2008; 11(10):995-1003.

103. Stegen JC, Lin X, Konopka AE, Fredrickson JK. Stochastic and deterministic assembly processes in subsurface microbial communities. ISME J. 2012;

6(9):1653-1664.

104. Andersson AF, Riemann L, Bertilsson S. Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. ISME J. 2010; 4(2):171-181.

105. Chase JM, Kraft NJB, Smith KG, Vellend M, Inouye BD. Using null models to disentangle variation in community dissimilarity from variation in α-diversity. Ecosphere. 2011; 2(2):1-11.

106. Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Lander ES, Mitzenmacher M, Sabeti PC. Detecting novel associations in large data sets. Science. 2011; 334(6062):1518-1524.

107. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 2012; 8(9):e1002687.

108. Watts SC, Ritchie SC, Inouye M, Holt KE. FastSpar: Rapid and scalable correlation estimation for compositional data. bioRxiv. 2018.

109. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13(11):2498-2504.

110. Csardi G, Nepusz T. The igraph software package for complex network research. InterJournal. 2006; Complex Systems:1695.

FIGURE LEGENDS

Figure 1. Ecological mechanisms shaping the tropical and subtropical surface-ocean picoplankton. Panel A) Position of the 120 stations included in this work that were sampled as part of the *Malaspina-2010* expedition (green dots) in the tropical and subtropical ocean. A snapshot of the global sea surface temperature, a main environmental driver affecting microbial distributions, is shown as a general representation of the temperature gradients in the surface ocean (as inferred using the 'optimum interpolation sea surface temperature' dataset from the NOAA corresponding to the 17 of March of 2018). Note that temperatures measured *in situ* were used in all analyses, not the ones displayed here. Panel B) Percentage of the community turnover associated to different ecological processes in prokaryotes and picoeukaryotes in the tropical and subtropical ocean as calculated using $OTUs_{-99\%}$ and $OTUs_{-ASVs}$. Note that percentage refers to the percentage of pairs of communities that appear to be driven by a given process.

Figure 2. Main variables influencing the structure of the surface-ocean microbiota as captured by different -diversity metrics. Percentage of variance in picoeukaryotic and prokaryotic community composition (ADONIS R^2) explained by

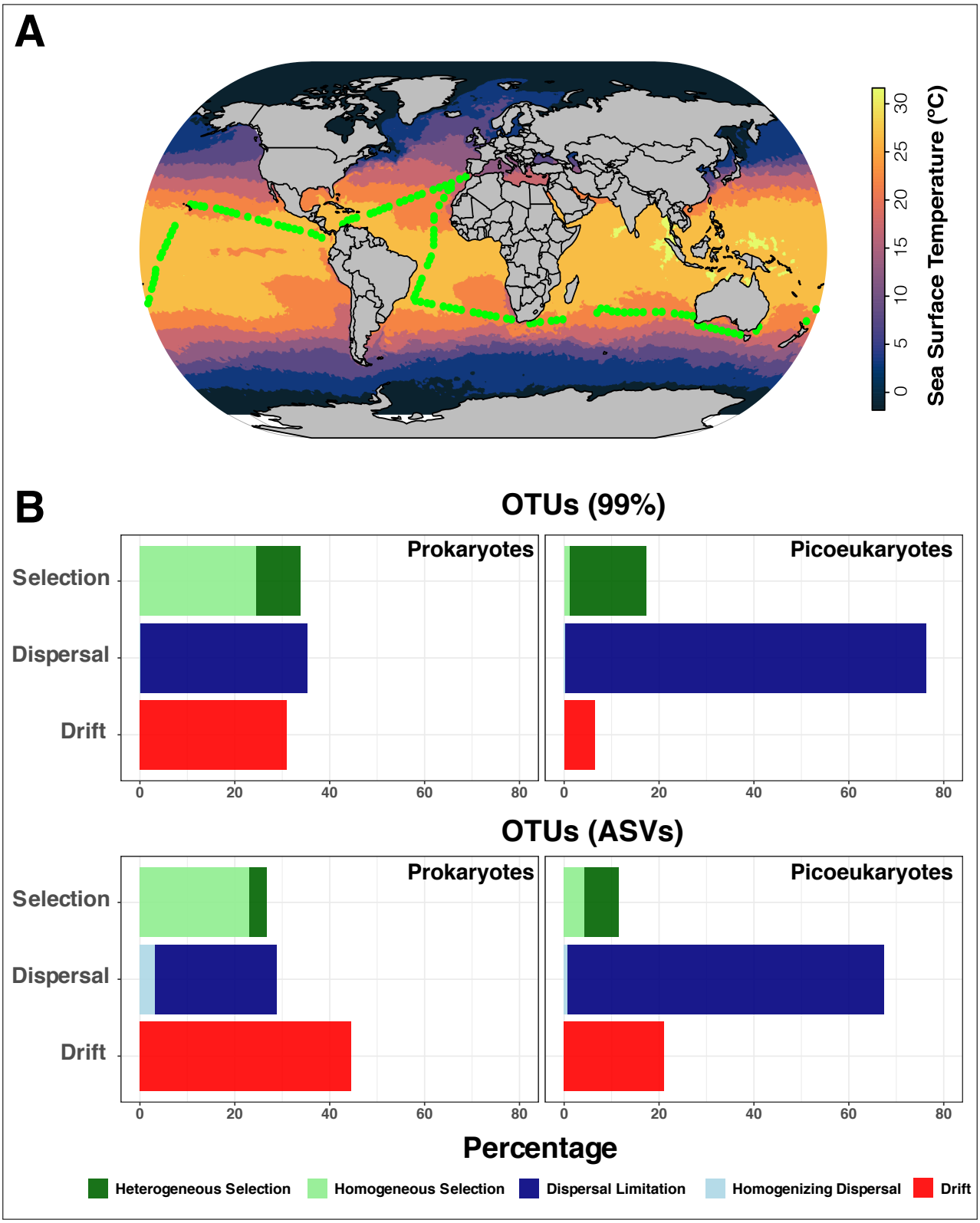
Water Temperature and Longhurst Provinces when using different -diversity metrics. Figure based on the *Malaspina Meta-119* dataset (see Methods). $TINA_w$: TINA weighted, gUniFrac: Generalized Unifrac, PINAw: PINA weighted. N.S. = Non-Significant. Note that $TINA_w$, which considers species association networks, captures a significantly higher proportion of community variance associated to temperature than Bray-Curtis, a compositional index, in prokaryotes.

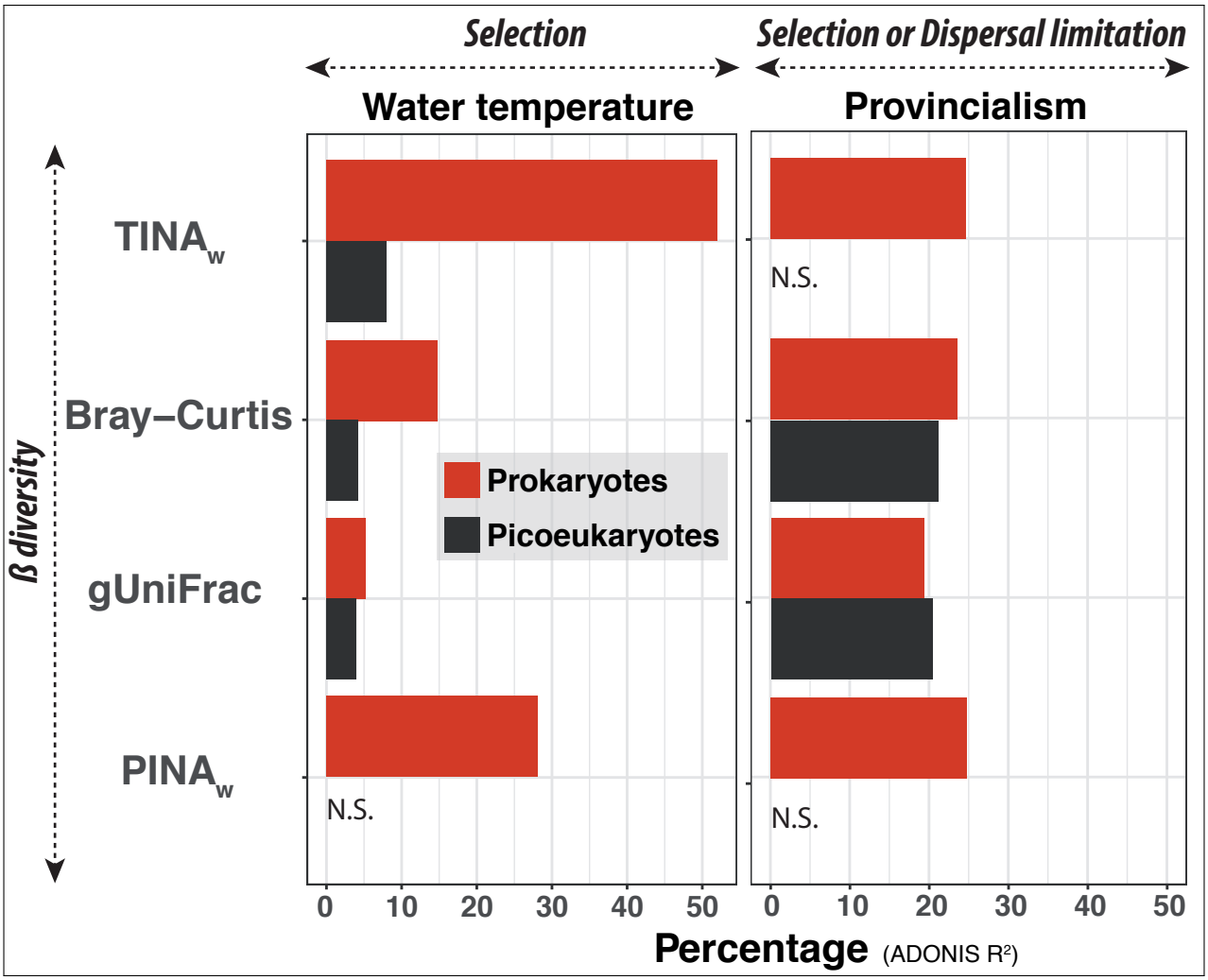
Figure 3. Temperature-driven selection seems to affect species association networks in prokaryotes but not in pico-/nano-eukaryotes. Differences in community composition (as 1-[TINA-weighted] = $TINA_w$ dissimilarities) vs. temperature differences (as Euclidean distances based on dimensionless z-scores) for both small unicellular eukaryotes and prokaryotes sampled during the *Malaspina* and *TARA Oceans* expeditions. Note that, in contrast to other indices, $TINA_w$ considers species-association patterns (i.e. co-occurrences and co-exclusions) when estimating -diversity [24]. NB: While only picoeukaryotes were contemplated in *Malaspina* (cell sizes <3 μm), *TARA Oceans* data included pico- and nano-eukaryotes (cell sizes <5 μm). Pico- and nanoeukaryotes from both expeditions (left panels) displayed low or no correlations between $TINA_w$ distances and temperature differences (Mantel test results included in the panels). On the contrary, prokaryotes (right panels) displayed high to moderate correlations between $TINA_w$ distances and temperature differences. These differences in the correlations are likely due to the wider temperature ranges covered by *TARA Oceans* compared to *Malaspina* (see Results). The regression line is shown in red (*Malaspina* microbial eukaryotes N.S., *Malaspina* Prokaryotes R^2=0.3, *TARA Oceans* microbial eukaryotes R^2=0.1, *TARA Oceans* Prokaryotes R^2=0.7; $p<0.05$). The maps at the bottom indicate the surface stations from the expeditions

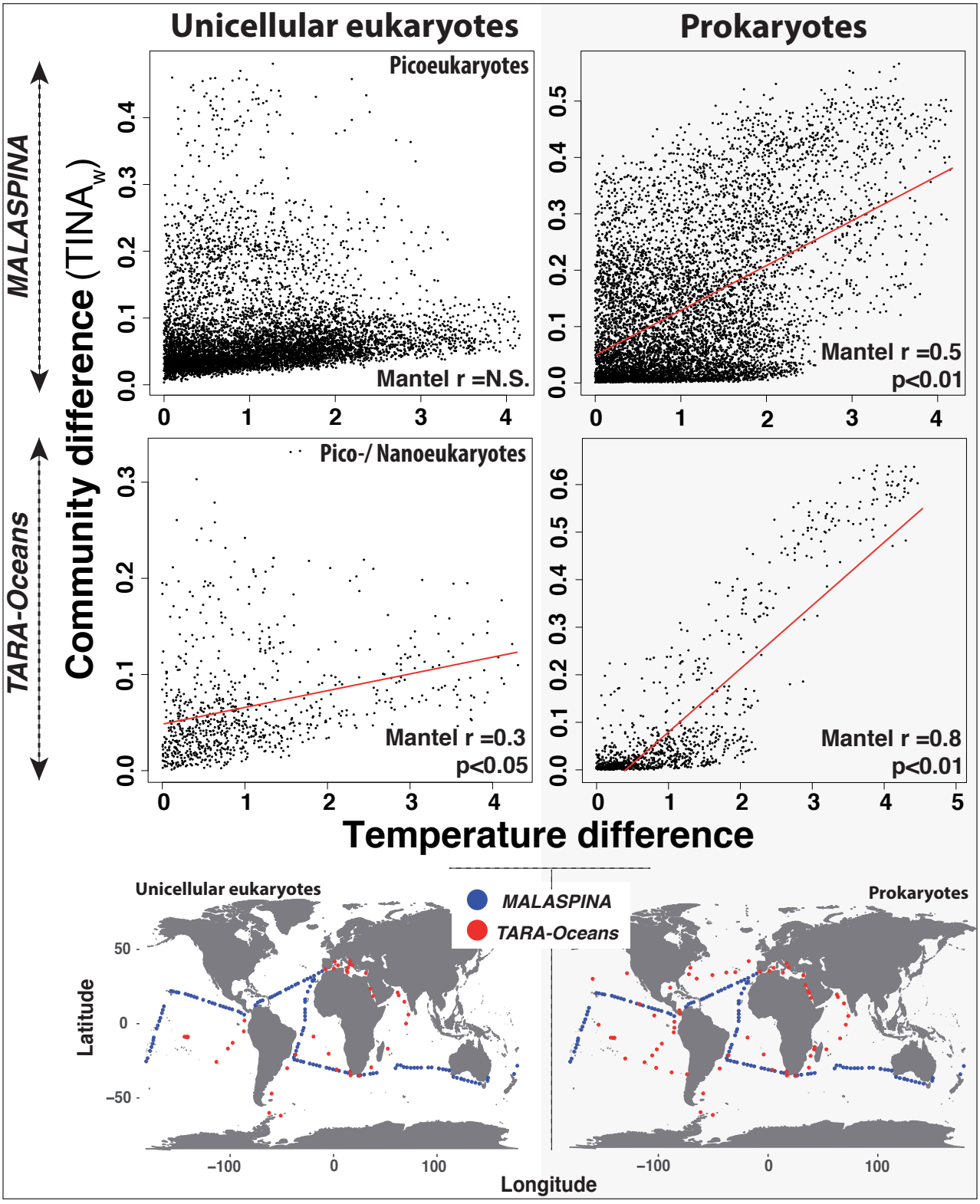
Malaspina (119 stations for both prokaryotes and picoeukaryotes) and *TARA Oceans* (63 stations for prokaryotes and 40 stations for small unicellular eukaryotes) that were used to calculate $TINA_w$.

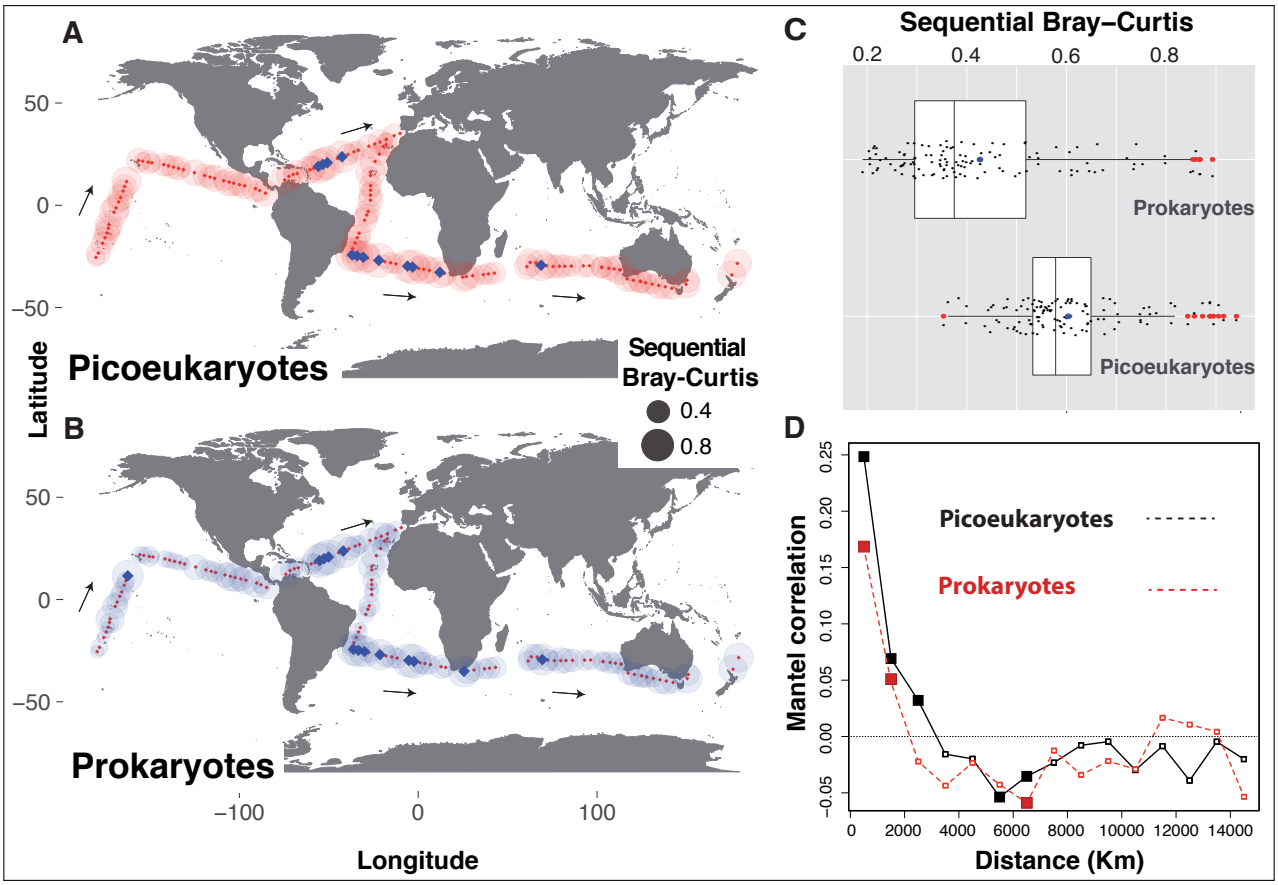
Figure 4. Picoeukaryotic communities display a higher spatial differentiation than prokaryotic counterparts in the tropical-subtropical surface-ocean. Panels A-C: Sequential change in community composition across space (sequential -diversity). Communities were sampled along the *Malaspina* expedition (Panels A and B, black arrows), and the composition of each community was compared against its immediate predecessor. In Panels A and B, the size of each bubble represents the Bray-Curtis dissimilarity between a given community and the community sampled previously. Blue squares in Panels A and B represent the stations where -diversity displayed abrupt changes (Bray Curtis values >0.8 for picoeukaryotes and >0.7 for prokaryotes). Abrupt changes coincided in a total of 11 out of 14 stations for both picoeukaryotes and prokaryotes, while one station displayed marked changes only for picoeukaryotes and two only for prokaryotes. Panel C summarizes the sequential Bray-Curtis values for prokaryotes and picoeukaryotes (Means were significantly different between domains [Wilcoxon text, $p<0.05$]). Panel D indicates the differences in distance-decay between prokaryotes and picoeukaryotes in the tropical and subtropical surface-ocean. Mantel correlograms between geographic distance and -diversity featuring distance classes of 1,000 km for both picoeukaryotes and

prokaryotes are shown. Coloured squares indicate statistically significant correlations ($p<0.05$). Note that -diversity in picoeukaryotes displayed positive correlations with increasing distances up to ~3,000 km, while prokaryotes had positive correlations with distances up to ~2,000 km. Correlations tended to be smaller in prokaryotes than in picoeukaryotes, indicating smaller distance decay in the former compared to the latter.


A
Sea Surface Temperature (°C)
0
5
10
15
20
25
30
B
OTUs (99%)
Prokaryotes
Picoeukaryotes
Selection
Dispersal
Drift
0
20
40
60
80
OTUs (ASVs)
Prokaryotes
Picoeukaryotes
Selection
Dispersal
Drift
0
20
40
60
80
Percentage
Heterogeneous Selection
Homogeneous Selection
Dispersal Limitation
Homogenizing Dispersal
Drift


Selection
Selection or Dispersal limitation
Water temperature
Provincialism
ß diversity
TINA$_w$
Bray–Curtis
gUniFrac
PINA$_w$
Prokaryotes
Picoeukaryotes
N.S.
N.S.
N.S.
0 10 20 30 40 50
0 10 20 30 40 50
Percentage (ADONIS R²)


Unicellular eukaryotes
Prokaryotes
MALASPINA
TARA-Oceans
Community difference (TINA$_w$)
Picoeukaryotes
Mantel r =N.S.
Mantel r =0.5
p<0.01
Pico-/ Nanoeukaryotes
Mantel r =0.3
p<0.05
Mantel r =0.8
p<0.01
Temperature difference
Unicellular eukaryotes
Prokaryotes
MALASPINA
TARA-Oceans
Latitude
Longitude


A
50
0
−50
Picoeukaryotes
Sequential
Bray-Curtis
0.4
0.8
B
Prokaryotes
Latitude
−100
0
100
Longitude
C
Sequential Bray−Curtis
0.2
0.4
0.6
0.8
Prokaryotes
Picoeukaryotes
D
Mantel correlation
−0.05
0.00
0.05
0.10
0.15
0.20
0.25
Picoeukaryotes
Prokaryotes
0
2000
4000
6000
8000
10000
12000
14000
Distance (Km)

A.2.2 Supplement

Supplementary Information for

Different processes shape prokaryotic and picoeukaryotic assemblages in the sunlit ocean microbiome

Ramiro Logares, Ina M. Deutschmann, Caterina. R. Giner, Anders K. Krabberød, Thomas S. B. Schmidt, Laura Rubinat-Ripoll, Mireia Mestre, Guillem Salazar, Clara Ruiz-González, Marta Sebastián, Colomban de Vargas, Silvia G. Acinas, Carlos M. Duarte, Josep M. Gasol & Ramon Massana

Ramiro Logares
Email: ramiro.logares@gmail.com

This PDF file includes:

Supplementary text
Figs. S1 to S8
Tables S1 to S8
References

Supplementary Methods

Sample collection

Surface waters (3m depth) from a total of 120 globally-distributed stations located in the tropical and sub-tropical global ocean **(Fig. S1)** were sampled from December 2010 to July 2011 as a part of the MALASPINA-2010 expedition (1) conducted on the R/V Hespérides. Water samples were obtained with a large (20L) Niskin bottle deployed simultaneously to a CTD profiler that included sensors for conductivity, temperature, oxygen, fluorescence and turbidity. After collection, ~12L of seawater were subsequently pre-filtered through a 200µm nylon mesh to remove large plankton, and then sequentially filtered, using a peristaltic pump, through a 20µm nylon mesh (at the entrance of the tubing) and 3µm and 0.2µm polycarbonate filters of 47mm diameter (Isopore, Millipore). Filtration time was ~15 minutes. After filtration, filters were flash-frozen in liquid N_2 and stored at -80ºC until downstream analyses. Samples for inorganic nutrients (NO_3^-, NO_2^-, PO_4^{3-}, SiO_2) were collected from the Niskin bottle, kept frozen, and measured spectrophotometrically using an Alliance Evolution II autoanalyzer (2). In specific samples, where the previous method failed or was not applied, we estimated nutrient concentration using the World Ocean Database (3). Given that not all environmental parameters were available for all stations, two contextual datasets were generated: *Meta-119*, including 119 stations, 5 environmental parameters and 5 spatial features and *Meta-57*, considering 57 stations and 17 environmental parameters (See below; **Fig. S4**). In statistical analyses, continuous environmental variables were standardized as z-scores, that is, deviations of the values from the global mean in standard deviation units.

DNA extraction, amplicon sequencing and bioinformatic analyses

DNA was extracted using a standard phenol-chloroform protocol (4). Fragments from both the 18S and 16S rRNA-gene were amplified from the same DNA extracts. The hypervariable V4 region of the 18S (~380 bp) was amplified with the primers TAReuk454FWD1 and TAReukREV3 (5), while the hypervariable V4-V5 (~400bp) region of the 16S was amplified with the primers 515F-Y - 926R (6), which targeted both Bacteria and Archaea. Samples were amplified for sequencing in a two-step process. In the first step, the forward primer was constructed with the Illumina i5 sequencing primer (5'-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG-3') and the TAReukFWD1 (18S) or 515F-Y (16S) primers. The reverse primer was constructed with the Illumina i7 sequencing primer (5'-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG-3') and the TAReukREV3 (18S) or 926R (16S) primers. Amplifications were performed in 25 µl reactions with Qiagen HotStar Taq master mix (Qiagen Inc, Valencia, California), 1µl of each 5µM primer, and 1µl of template. Reactions were performed on ABI Veriti thermocyclers (Applied Biosytems, Carlsbad, California) under the following thermal "touchdown" profile: 95°C for 5 min, then 10 cycles of 94°C for 30 sec, 50°C for 40 sec (+0.5°C per cycle), 72°C for 1 min, followed by 25 cycles of 94°C for 30 sec, 54°C for 40 sec, 72°C for 1 min, and finally, one cycle of 72°C for 10 min. Products from the first amplification step were added to a second PCR based on qualitatively determined concentrations. Primers for the second PCR were designed based on the Illumina Nextera PCR primers as follows: Forward - 5'-AATGATACGGCGACCACCGAGATCTACAC[i5index]TCGTCGGCAGCGTC-3' and Reverse - 5'-

CAAGCAGAAGACGGCATACGAGAT[i7index]GTCTCGTGGGCTCGG-3'. The second stage amplification consisted in 95°C for 5 min, then 10 cycles of 94°C for 30 sec, 54°C for 40 sec, 72°C for 1 min, followed by one cycle of 72°C for 10 min.

Amplification products were visualized with eGels (Life Technologies, Grand Island, New York). Products were then pooled equimolar and each pool was size selected in two rounds using Agencourt AMPure XP (BeckmanCoulter, Indianapolis, Indiana) in a 0.75 ratio for both rounds. Size selected pools were then quantified using the Qubit 2.0 fluorometer (Life Technologies) and loaded on an *Illumina* MiSeq (Illumina, Inc. San Diego, California) flow cell at 10 pM. Sequencing was performed using 2x250 bp. Amplicon library construction and sequencing was performed at the Research and Testing Laboratory facility (Lubbock, TX, USA; http://www.researchandtesting.com/).

A total of 71,391,060 (2 x 35,695,530) reads were produced for picoeukaryotes, while 17,129,672 (2 x 8,564,836) reads were produced for prokaryotes. Reads were processed following an in-house pipeline (7). Briefly, raw reads were corrected using BayesHammer (8) following Schirmer et al. (9) Corrected paired-end reads were subsequently merged with *PEAR* (10) and sequences longer than 200 bp were quality-checked (maximum expected errors 0.5) and de-replicated using *USEARCH* (11). OTUs were delineated at 99% similarity using *UPARSE* V8.1.1756 (12). To obtain OTU abundances, reads were mapped back to OTUs at 99% similarity using an exhaustive search *(-maxaccepts 20 -maxrejects 50,000-100,000*). Chimera check and removal was performed both *de novo* and using the SILVA reference database (13). After our stringent quality control, a total of 16,460,248 18S and 5,697,779 16S reads were considered, which were associated to 42,505 18S and 10,158 16S OTUs. See more details on

sequence processing in **Table S8**. Taxonomic assignment of 18S OTUs was generated by BLASTing (14) OTU-representative sequences against three reference databases, PR^2 (15) and two in-house marine protist databases (available at https://github.com/ramalok), one based on a collection of Sanger sequences from molecular surveys (16) and the other based on 454 reads from the BioMarKs project (17). Metazoan, Charophyta and nucleomorph OTUs were removed. Similarly, for 16S OTUs, taxonomic assignment was based on BLASTing OTU-representative sequences against SILVA v123. Chloroplast and mitochondrial sequences were removed. BLAST hits were filtered prior to taxonomy assignment using an in-house python script, considering a percentage of identity >90%, a coverage >70%, a minimum alignment length of 200 bp and an e-value > 0.00001. Computing analyses were performed at the MARBITS bioinformatics platform of the Institut de Ciències del Mar (ICM; http://marbits.icm.csic.es) as well as in MareNostrum (Barcelona Supercomputing Center). Sequences are publicly available at the European Nucleotide Archive (http://www.ebi.ac.uk/ena; accession numbers PRJEB23913 [18S] & PRJEB25224 [16S])

Statistical analyses and phylogenetic inferences

In order to allow for comparisons, both picoeukaryotic and prokaryotic datasets were sub-sampled to 4,060 reads per sample using rrarefy in Vegan (18). The sub-sampled picoeukaryotic and prokaryotic OTU tables contained 18,881 and 7,025 OTUs respectively. All OTUs with mean relative abundances above 0.1% and below 0.001% were defined as regionally abundant or rare respectively (19).

Most analyses and graphs were performed in the R statistical environment (20) using *ggplot2* (21), *Maps* (22), *Mapplots* (23) and *Vegan*. Local Contributions to Beta Diversity (24) was measured with *adespatial* (25). Phylogenetic trees were constructed for both the 16S and 18S datasets using OTU-representative sequences. Reads were aligned against an aligned SILVA template using *mothur* (26). Afterwards, poorly aligned regions or sequences were removed using *Trimal* [parameters: -gt 0.3 -st 0.001] (27). A phylogenetic tree was inferred with *FastTree* (28) using the Generalized Time Reversible (GTR) model of nucleotide substitution considering a CAT/Gamma-distributed rate of variation across sites (including 20 rate categories). Other phylogenetic analyses were performed with the *R*-packages *picante* (29), *APE* (30) and *gUniFrac* (31). *gUniFrac* was run with an alpha value of 0.5.

Quantification of Selection, Drift and Dispersal

These processes were quantified using the approach proposed by Stegen et al. (32) This methodology consists of two main sequential steps: the first step uses phylogenetic turnover and the second step uses OTU turnover. Phylogenetic turnover is measured by calculating the abundance-weighted β-mean nearest taxon distance (βMNTD), which quantifies the phylogenetic distances between the evolutionary closest OTUs in two communities. Short phylogenetic distances are considered in this approach, as it has been found that closely related taxa have habitat preferences which are more similar than the habitat preferences of distantly related taxa (32). βMNTD values can be larger, smaller or equal to the values expected when selection is not affecting community turnover (that is, expected under a random distribution). βMNTD values higher than

expected indicate that communities are under heterogeneous selection (33). In contrast, βMNTD values which are lower than expected indicate that communities are experiencing homogeneous selection. Null models were constructed using 999 randomizations as in Stegen et al. (32) Differences between the observed βMNTD values and the mean of the null distribution are denoted as β-Nearest Taxon Index (βNTI), with | βNTI | > 2 being considered as significant departures from random phylogenetic turnover, pointing to the action of selection.

The second step of this process calculates whether the observed β-diversity, based in OTU turnover, could be generated by drift or other processes. For this, we calculated the Raup-Crick metric (34) using Bray-Curtis dissimilarities [hereafter RC_{bray}], following Stegen et al. (32). RC_{bray} compares the measured β-diversity against the β-diversity that would be obtained if drift was driving community turnover (that is, under random community assembly). Randomizations were run 9,999 times and only OTUs with >1,000 reads over the entire dataset were considered. These should provide the best evidence on whether dispersal vs. drift structure communities, as such abundant OTUs should be less prone to the potential effects of subsampling or sampling biases. RC_{bray} values between -0.95 and +0.95 point to a community assembly governed by drift. On the contrary, RC_{bray} values > +0.95 or < -0.95 indicate that community turnover is driven by low/high dispersal respectively (34). According to Stegen et al.(32), dispersal limitation is only expected to produce significant RC_{bray} values when coupled to drift, which introduces stochastic changes in community composition that magnify their differentiation leading eventually to RC_{bray} values > +0.95. In contrast, homogenizing dispersal (similar to mass effects) could generate RC_{bray} values < -0.95, reflecting a

process in which the composition of two communities is more similar than expected by chance due to high immigration rates.

The previous framework was applied as proposed by Stegen et al. (32): First, we determined the fraction of total pairwise comparisons with a $|\beta NTI| > 2$. This proportion was interpreted as the overall action of selection in our picoplankton dataset. As a consequence, the turnover of communities featuring $|\beta NTI| < 2$ should be driven by dispersal limitation, homogenizing dispersal or drift. Thus, the second step in this procedure was to calculate the RC_{bray} for all those community pairs whose turnover was not governed by selection (that is, those with $|\beta NTI| < 2$). Here, values of $RC_{bray} > +0.95$ are interpreted as dispersal limitation, values of $RC_{bray} < -0.95$ are interpreted as homogenizing dispersal, while values of $| RC_{bray} | < +0.95$ are associated to drift. Subsequently, for the pairwise comparisons that did not indicate the action of selection, we calculated the proportion of total comparisons that could be assigned to dispersal limitation, homogenizing dispersal or drift according to their RC_{bray} values.

Environmental datasets

The *Meta-119* dataset, included 119 stations, 5 environmental parameters, and 5 spatial features for most stations. The 5 environmental parameters were: Temperature (ºC), Conductivity (S m^{-1}), Fluorescence, Salinity and Dissolved Oxygen (ml L^{-1}). *Meta-119* also considered the following spatial features: Longhurst Province (35), Ocean, Ocean Subdivision, Distance to the coast <370km and Terrestrial influence.

The *Meta-57* dataset considered 57 stations **(Fig. S4)** and 17 environmental parameters for most stations. The 17 environmental parameters were: Temperature (ºC),

Conductivity (S m^{-1}), Fluorescence, PAR (Photosynthetically Active Radiation; measured with a sensor attached to the CTD), Turbidity, Salinity, Dissolved Oxygen (ml L^{-1}), Chlorophyll concentration (µg L^{-1}) (36), Fluorescent Dissolved Organic Matter (FDOM; four peaks associated to humic and amino-acid substances were measured, indicated as Fmax1, Fmax2, Fmax3, Fmax4; see (37)), TEP [Transparent Exopolymer Particles] (38), POC (Particulate Organic Carbon) (38), NO_3_Mala_WOA13 (µmol L^{-1}) [Nitrate, values from Malaspina and WOA13], PO_4_Mala_WOA13 (µmol L^{-1}) [Phosphate, values from Malaspina and WOA13], SiO_4_Mala_WOA13 (µmol L^{-1}) [Silicate, values from Malaspina and WOA13] (37, 39, 40).

Maximal Information Coefficient (MIC) analyses

In MIC analyses (41), the same 17 environmental parameters used in the *Meta-57* dataset were considered (see above ***Environmental datasets***). In analyses of picoeukaryotic or prokaryotic OTUs vs. environmental parameters, all OTUs were considered, while in analyses including comparisons of all OTUs against each other plus environmental parameters, only OTUs with ≥100 reads were included, due to computational limitations.

MIC analyses using the TARA Oceans datasets included 8 environmental parameters for prokaryotes (63 stations): Temperature (ºC), Salinity, Oxygen (µmol/kg), NO_3 (µmol/L), NO_2 (µmol/L), PO_4 (µmol/L), NO_2NO_3 (µmol/L) and SI (µmol/L). These data are publicly available in: http://ocean-microbiome.embl.de/companion.html. MIC analyses of microbial eukaryotes from TARA Oceans considered 6 environmental parameters (40 stations / 41 samples): Temperature (ºC), Salinity, Oxygen (µmol/kg),

NO_3 (µmol/L), PAR, Chlorophyll *a* (mg/m^3). These data are available in: http://taraoceans.sb-roscoff.fr/EukDiv/. In MIC analyses of OTUs against environmental parameters, only OTU with ≥ 30 reads were used for both microbial eukaryotes (10,115 OTUs, 61,407,151 reads) and prokaryotes (5,029 OTUs, 6,402,539 reads). Given the large number of possible pairwise comparisons in analyses considering all OTUs and environmental parameters against each other, only OTUs with ≥ 500 reads were used for prokaryotes (1,656 OTUs, 5,930,665 reads) while OTUs with ≥ 1,000 reads were used for microbial eukaryotes (2,026 OTUs, 59,897,456 reads).

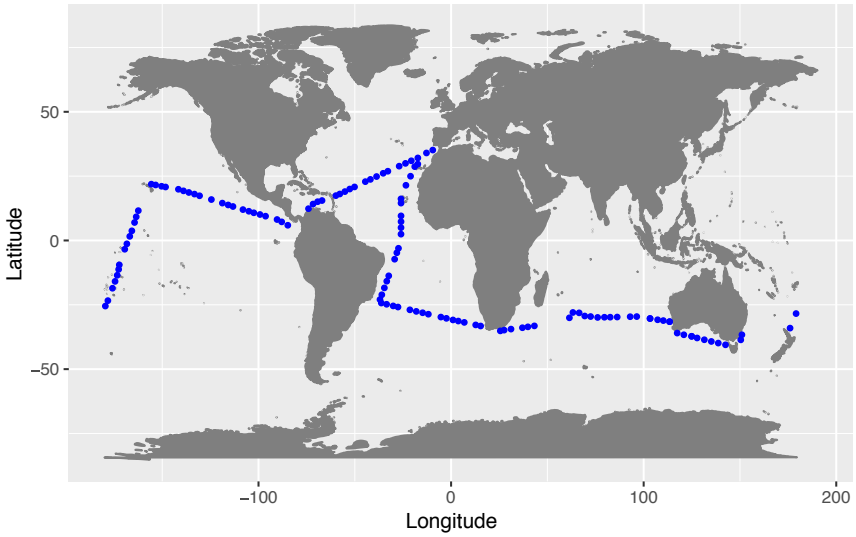


Fig. S1. Position of the 120 stations included in this work that were sampled as part of the Malaspina-2010 expedition.

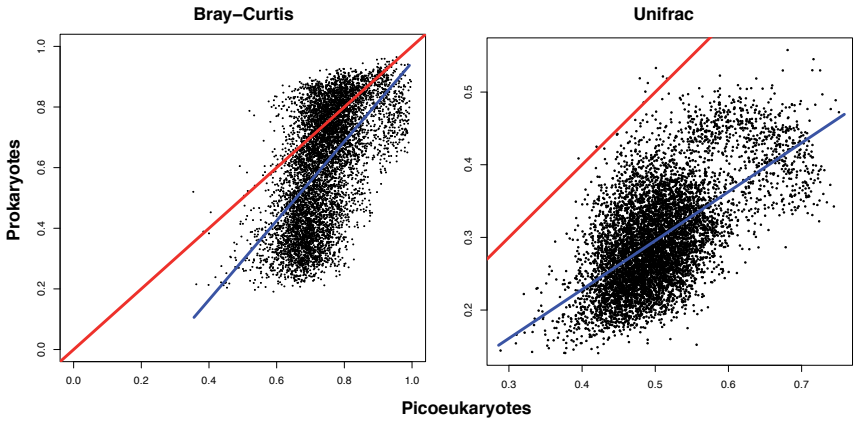


Fig. S2. Bray Curtis and gUniFrac distances between picoeukaryotes and prokaryotes. Regression (blue) and 0:1 (red) lines are indicated.

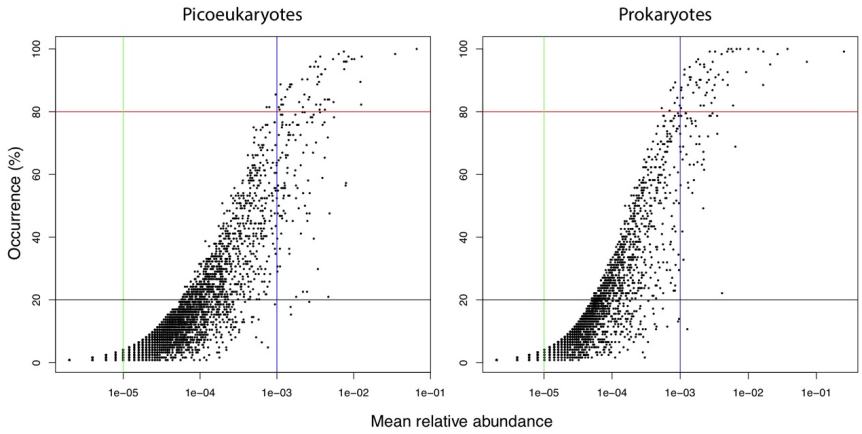


Fig. S3. OTU mean relative abundance (i.e. regional abundance) vs. occurrence (i.e. number of samples in which each OTU is present). The red and black horizontal lines indicate percentages of occurrences of 80% and 20% respectively. Cosmopolitan OTUs were considered as those with a percentage of occurrence >80%, while restricted OTUs were those with a percentage of occurrence <20% (see **Table S2**). Blue and green vertical lines indicate regional abundances above and below which OTUs are considered regionally abundant (>0.1%) or rare (<0.001%) respectively.

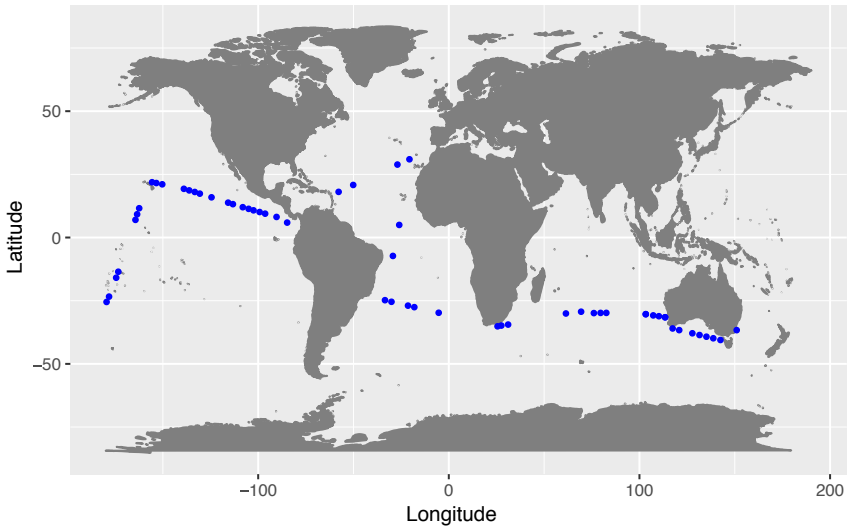


Fig. S4. The 57 Malaspina stations for which 17 environmental parameters were available (*Meta-57* dataset).

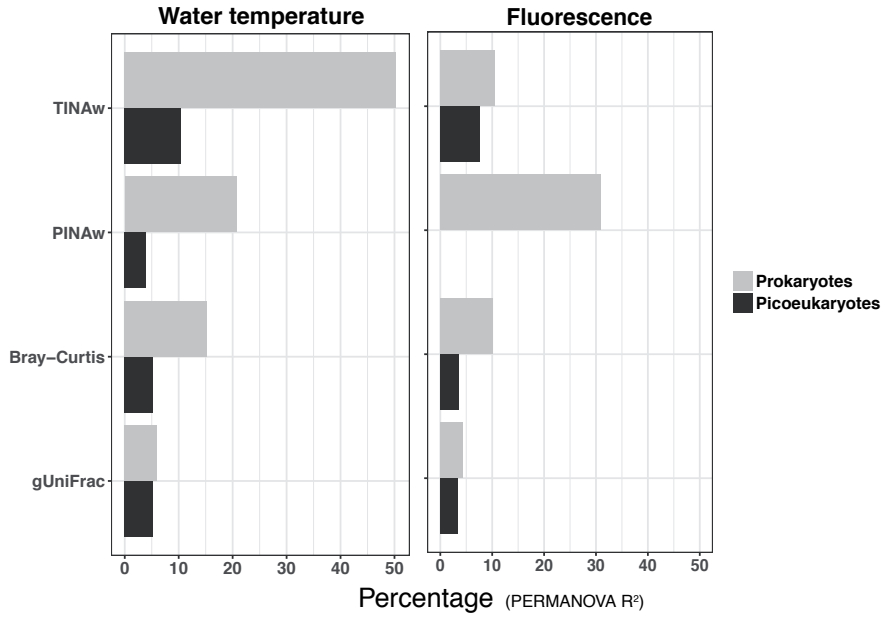


Fig. S5. Percentage of variance in Picoeukaryotic and Prokaryotic community composition (PERMANOVA R^2) explained by water temperature and fluorescence when using different ß-diversity metrics. Figure based on the *Meta-57* dataset.

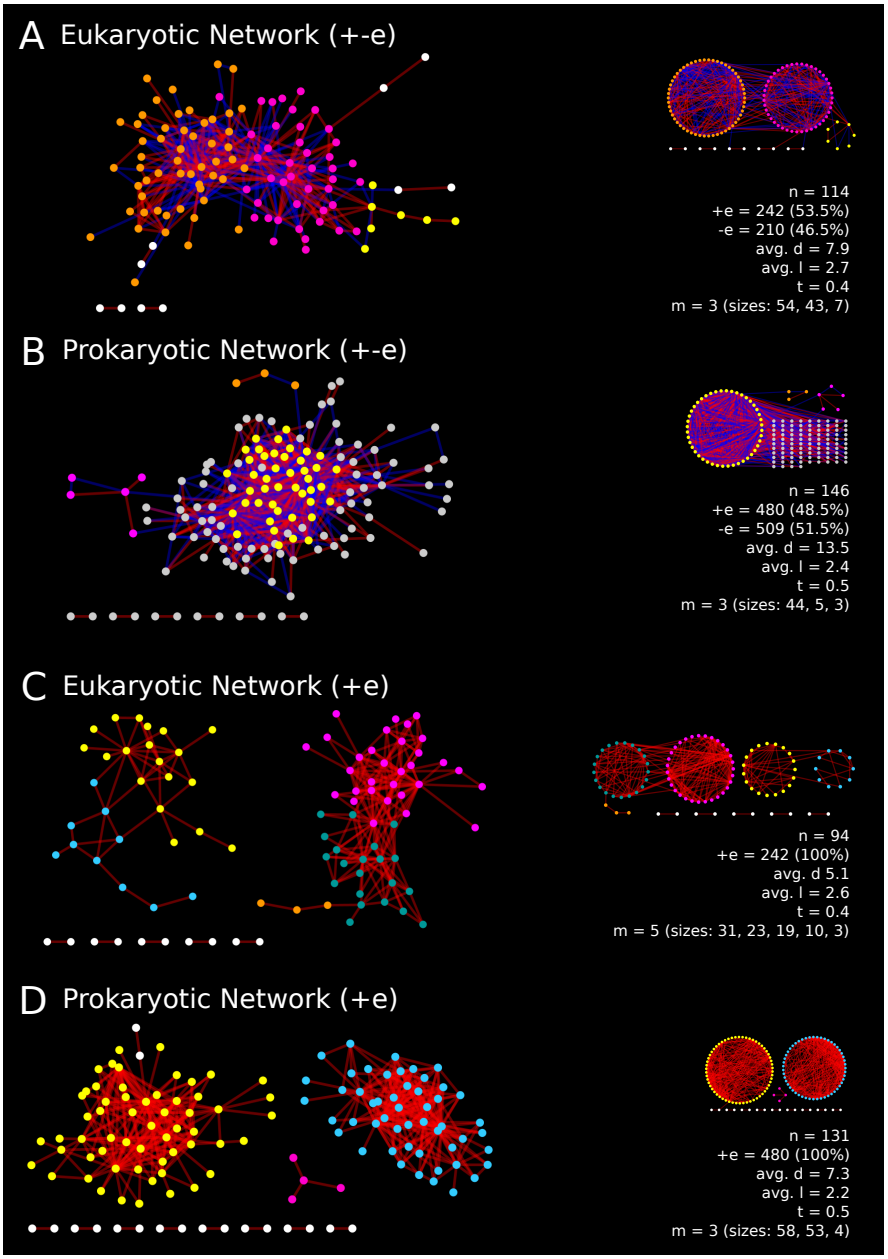


Fig. S6. Left-hand side: Association networks of picoeukaryotes and prokaryotes considering positive (red) and negative (blue) correlations in panels A) [Eukaryotic Network (+-e)] and B) [Prokaryotic Network (+-e)], and only positive correlations in C) [Eukaryotic Network (+e)] and

D) [Prokaryotic Network (+e)]. On the right-hand side, we present an alternative visualization of the network as well as the following network characteristics: number of nodes (n), number of edges with positive correlation (+e) and negative correlation (-e), average degree (avg. d), average path length (avg. l), global transitivity (t), number of modules with a least 3 nodes (m) and the number of nodes in each of those modules (sizes:). The smaller network visualization on the right-hand side groups the nodes according to the modules. The color of nodes in Left- and Right-hand side networks indicate the modules to which they belong.

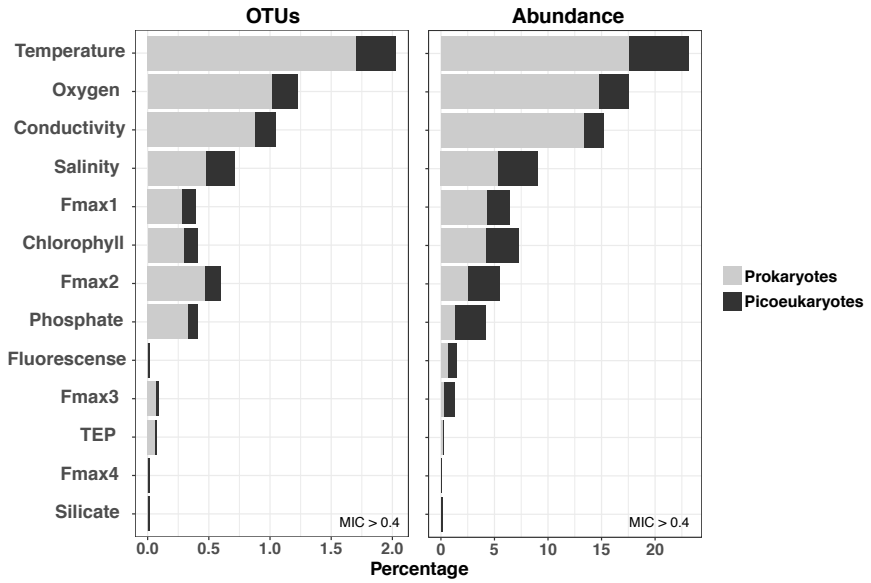


Fig. S7. Percentage of OTUs significantly associated to different environmental variables (MIC > 0.4) and their corresponding contribution to total abundance. NB: Temperature, Oxygen, Conductivity and Salinity are correlated. OTUs can be associated to more than one variable.

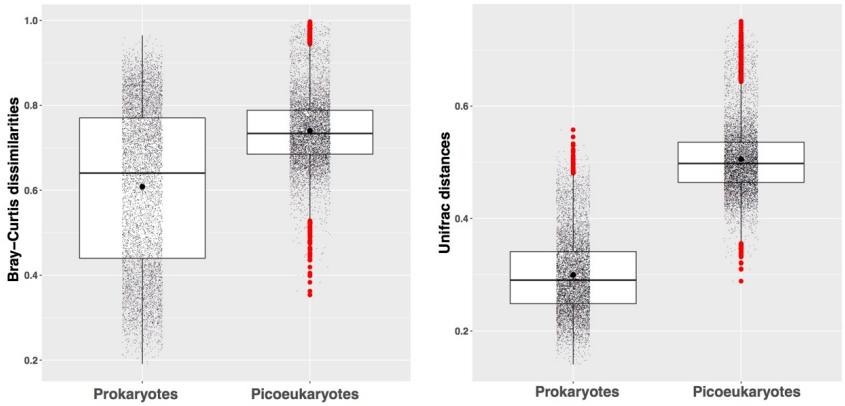


Fig. S8. Bray-Curtis dissimilarities and gUniFrac distances in Prokaryotes and Picoeukaryotes. In both cases, mean differences were significant (Wilcoxon text, $p<0.05$). Prokaryotes (Bray Curtis mean=0.61, SD=0.19; gUniFrac mean=0.30, SD=0.07); Picoeukaryotes (Bray Curtis mean=0.74, SD=0.08; gUniFrac mean=0.50, SD=0.06).

Table S1. Regionally abundant or rare prokaryotic and picoeukaryotic OTUs.

	Prokaryotes (%)	Picoeukaryotes (%)
Regionally abundant OTUs (>0.1%)[1]	1.46 (103)	0.84 (158)
Regionally rare OTUs (<0.001%)[2]	47.6 (3,343)	71.5 (13,499)

[1] OTUs featuring a mean relative abundance >0.1%. [2] OTUs featuring a mean relative abundance <0.001%. Percentages as well as corresponding number of OTUs (within parenthesis) are indicated.

Table S2. OTUs displaying Cosmopolitan, Intermediate and Restricted distributions.

	Cosmopolitan (>80%)[1]	Intermediate (80-20%)[2]	Restricted (<20%)[3]
Picoeukaryotes	0.3 (57)	5.1 (954)	94.6 (17,870)
Prokaryotes	1.0 (68)	11.1 (779)	87.9 (6,178)

[1] Percentage of OTUs occurring in >80% of the samples (Cosmopolitan). [2] Percentage of OTUs occurring in 80-20% of the samples (Intermediate distribution). [3] Percentage of OTUs occurring in <20% of the samples (Restricted distribution). The number of OTUs is indicated within parentheses.

Table S3. Summary of association networks based on SparCC (42).

	Connected nodes[1]	Positive edges[2]	Negative edges[3]	Average degree[4]	Transitivity (global)[5]	Average path length[6]	Number of cliques (≥3 nodes)[7]	Mean clique size[8]	Number of modules (≥3 nodes)[9]
Eukaryotes (+-e)	114 (17.5%)	242 (53.5%)	210 (46.5%)	7.9	0.4	2.7	120	5.4	3 (54,43,7)
Prokaryotes (+-e)	146 (32.6%)	480 (48.5%)	509 (51.5%)	13.5	0.5	2.4	262	11.4	3 (44,5,3)
Eukaryotes (+e)	94 (14.4%)	242 (100%)	-	5.1	0.4	2.6	58	4.3	5 (31,23,19,10,3)
Prokaryotes (+e)	131 (29.2%)	480 (100%)	-	7.3	0.5	2.2	103	6.2	3 (58,53,4)

The eukaryotic and prokaryotic networks included 651 and 448 nodes (OTUs) respectively, featuring absolute correlation scores >0.3 with $p<0.01$. Networks considering associations (edges) with both positive and negative scores (+-e) as well as networks including positive scores only (+e) are indicated. [1] Number of nodes with at least one edge as well as the percentage they represent of all the analysed nodes. [2] Number of edges with a positive correlation score and their percentage. [3] Number of edges with a negative correlation score and their percentage. [4] The degree indicates the number of edges connected to a node; the average degree refers to the sum of all degrees divided by the number of connected nodes. [5]Transitivity measures the probability that two nodes that are connected to a third node are also connected (also known as Clustering Coefficient). [6]The Average Path Length is the mean shortest distance between any pair of nodes in the network. [7] Cliques represent fully connected subnetworks of a network; the number of cliques with at least 3 nodes is indicated. [8] Average number of nodes in the network cliques of at least 3 nodes. [9] Number of modules (highly connected areas of a network) with at least 3 nodes calculated with the method considering edge betweenness (43); the size of each module, in terms of number of nodes, is indicated.

Table S4. Summary of significant OTU-associations using MIC (41).

	Associations[1]	% OTUs (#)[2]	Abundance (%)[3]	MIC-ρ^2 > 0.2 (%)[4]
Eukaryote – Eukaryote (MIC>0.4)[5]	121	15 (97)	16	51
Eukaryote – Eukaryote (MIC>0.3)[5]	914	60 (388)	61	71
Prokaryote – Prokaryote (MIC>0.4)[5]	618	51 (229)	71	54
Prokaryote – Prokaryote (MIC>0.3)[5]	3,163	87 (389)	83	53
Eukaryote – Prokaryote (MIC>0.4)[5]	143	-	-	75
Eukaryote – Prokaryote (MIC>0.3)[5]	1,507	-	-	73
Eukaryotes in Eukaryote – Prokaryote associations (MIC>0.4) [5]	-	7 (49)	8	-
Eukaryotes in Eukaryote – Prokaryote associations (MIC>0.3) [5]	-	47 (302)	51	-
Prokaryotes in Eukaryote – Prokaryote associations (MIC>0.4) [5]	-	12 (52)	46	-
Prokaryotes in Eukaryote – Prokaryote associations (MIC>0.3) [5]	-	54 (244)	73	-
Eukaryotes – Environment (MIC>0.4)[6]	269	1 (159)	15	87
Eukaryotes – Environment (MIC>0.3)[6]	3,165	8 (1430)	74	92
Prokaryotes – Environment (MIC>0.4)[6]	403	3 (212)	30	78
Prokaryotes – Environment (MIC>0.3)[6]	3,186	17 (1192)	88	85

All MIC values have a $p<0.05$. [1] Number of associations (NB: OTUs may feature more than one association). [2] Percentage of OTUs involved in associations; corresponding OTU numbers are given within parentheses. [3] Percentage of total abundance of OTUs involved in associations. [4] Percentage of non-linear associations (MIC-ρ^2 >0.2) out of all associations[1]. [5] Analyses included OTUs with >100 reads from sub-sampled OTU tables (Picoeukaryotes: 648 OTUs; Prokaryotes:448 OTUs). [6] Analyses were done with an OTU table including all OTUs. Prokaryotes: Total OTUs = 7,025, Total reads: 489,039; Eukaryotes: Total OTUs = 18,775, Total reads: 491,260. Included environmental parameters are indicated in *Environmental datasets* above.

Table S5. Summary of significant OTU-associations with MIC > 0.5 for the TARA-Oceans dataset based on MIC.

	Associations[1]	% OTUs (#)[2]	Abundance (%)[3]	MIC-ρ^2 > 0.2 (%)[4]
Eukaryote – Eukaryote (MIC>0.8)	747	21.5 (437)	22.7	91.3
Eukaryote – Eukaryote (MIC>0.7)	3,238	50.7 (1,028)	66.4	89.6
Eukaryote – Eukaryote (MIC>0.5)	49,839	96.5 (1,956)	99.6	88.1
Prokaryote – Prokaryote (MIC>0.8)	2,270	46.6 (771)	69.0	23.6
Prokaryote – Prokaryote (MIC>0.7)	6,698	72.6 (1,203)	84.9	37.2
Prokaryote – Prokaryote (MIC>0.5)	73,705	97.2 (1,611)	99.2	63.4
Eukaryote – Environment (MIC>0.7)	248	1.96 (199)	4.0	99.1
Eukaryote – Environment (MIC>0.5)	2,811	19.3 (1957)	28.5	94.8
Prokaryotes – Environment (MIC>0.7)	66	1.0 (50)	0.99	92.4
Prokaryotes – Environment (MIC>0.5)	1,099	14.7 (740)	17.9	72.6

[1] Number of associations (NB: OTUs may feature more than one association). [2] Percentage of OTUs involved in associations; corresponding OTU numbers are given within parentheses. [3] Percentage of total abundance of OTUs involved in associations. [4] Percentage of non-linear associations (MIC-ρ^2 >0.2) out of all associations[1]

Table S6. Significant MIC associations (MIC > 0.5) between OTUs and environmental parameters for the TARA-Oceans dataset.

Environmental variable	Number of OTUs[1]	%[2]	OTU abundance (# reads)[3]	%[4]
PROKARYOTES				
Temperature (ºC)	441	40.1	1,111,758	37.4
Salinity	4	0.4	6,365	0.2
Oxygen (μmol/kg)	555	50.5	1,473,130	49.5
NO_3 (μmol/L)	2	0.2	7,189	0.2
NO_2 (μmol/L)	37	3.4	176,421	5.9
PO_4 (μmol/L)	25	2.3	66,969	2.2
NO_2NO_3 (μmol/L)	31	2.8	125,608	4.2
SI (μmol/L)	4	0.4	7,209	0.2
MICROBIAL EUKARYOTES				
Temperature (ºC)	836	29.7	7,166,962	27.7
Salinity	464	16.5	4,172,647	16.1
Oxygen (μmol/kg)	662	23.5	4,024,073	15.5
NO_3 (μmol/L)	471	16.7	4,050,804	15.6
PAR	106	3.8	717,541	2.8
Chlorophyll a (mg/m^3)	272	9.7	5,747,883	22.2

[1] Number of OTUs associated to each environmental parameter. [2] Percentage of total OTUs. [3] Number of reads represented by the OTUs. [4] Percentage of total reads. NB: percentage columns do not sum 100% as OTUs can be involved in associations with more than one environmental variable or they can present no associations with them.

Table S7. The 36 stations (out of 120) featuring a significant ($p<0.05$) Local Contribution to Beta Diversity [LCBD (24)] in prokaryotes and/or picoeukaryotes.

#	Station ID	Sample	Prokaryotes	Picoeukaryotes
1	1	ST_1_MD28	*	N.S.
2	7	ST_7_MD98	N.S.	*
3	**27** [a]	ST_27_MD458	*	*
4	**29** [a]	ST_29_MD506	*	*
5	30	ST_30_MD528	N.S.	*
6	**37** [a]	ST_37_MD646	*	*
7	**38** [a]	ST_38_MD664	*	*
8	39	ST_39_MD684	*	N.S.
9	**40**	ST_40_MD712	*	*
10	**41**	ST_41_MD734	*	*
11	**43** [a]	ST_43_MD753	*	*
12	**44**	ST_44_MD778	*	*
13	45 [a]	ST_45_MD806	N.S.	*
14	53	ST_53_MD962	N.S.	*
15	**54**	ST_54_MD985	*	*
16	58	ST_58_MD1080	N.S.	*
17	67	ST_67_MD1246	N.S.	*
18 [b]	**71**	ST_71_MD1318 [b]	*	*
18 [b]	71	ST_71_MD1324 [b]	*	N.S.
19	72	ST_72_MD1331	*	N.S.
20	73	ST_73_MD1354	*	N.S.
21	**74**	ST_74_MD1368	*	*
22	75	ST_75_MD1398	*	N.S.
23	76	ST_76_MD1421	*	N.S.
24	77	ST_77_MD1425	N.S.	*
25	89	ST_89_MD1629	*	N.S.
26	92	ST_92_MD1672	*	N.S.
27	94	ST_94_MD1724	*	N.S.
28	**95**	ST_95_MD1744	*	*
29	96	ST_96_MD1772	*	N.S.
30	97	ST_97_MD1798	N.S.	*
31	**124**	ST_124_MD2332	*	*
32	**125**	ST_125_MD2340	*	*
33	130	ST_130_MD2474	N.S.	*
34	132	ST_132_MD2562	N.S.	*
35	**133** [a]	ST_133_MD2594	*	*
36	**135** [a]	ST_135_MD2662	*	*

* LCBD $p<0.05$. N.S.: Non-significant. Both picoeukaryotes and prokaryotes featured 26 stations each with LCBD $p<0.05$, totaling 36 stations. A total of 16 stations displayed samples with LCBD $p<0.05$ for both prokaryotes and picoeukaryotes (shown in **bold**). [a] Stations identified also in sequential β diversity analyses (8 stations) as points of abrupt β diversity change. [b] Two samples from the same station.

Table S8. Reads and OTUs processed during different steps of our in-house workflow (7) for picoeukaryotes and prokaryotes.

Processing step	Picoeukaryotes	Prokaryotes
Total reads	35,695,530 (100%)	8,564,836 (100%)
Merged reads (PEAR)	34,768,276 (97.4%)	8,516,553 (99.4%)
Reads passing quality control (max_ee =0.5)	19,230,661 (55.3%)	7,234,052 (84.9%)
Reads after HMM[1] (rDNA validation)[2]	**19,230,187 (99.9%)**	**7,234,049 (99.9%)**
De-replicated reads (incl. singletons)	5,644,422	1,748,584
Singletons	4,526,205	1,450,578
De-replicated reads (without singletons)	1,118,217	298,006
Chimeric reads (de-novo detection during Uparse)	192,186 (17.2%)	105,470 (35.4%)
Total OTUs (99% clustering UPARSE)	51,571 (100%)	13,300 (100%)
Chimeric OTUs (reference-based: positive)	6,605 (12.8%)	2,413 (18.1%)
Chimeric OTUs (reference-based: uncertain)	2,461 (4.8%)	729 (5.5%)
OTUs 99% - Non-Chimeric	**42,505 (82.4%)**	**10,158 (76.4%)**
Reads[2] mapped back to OTUs (99% similarity)	**16,460,248 (85.6%)**	**5,697,779 (78.8%)**
Reads[2] not mapping back to OTUs (99% similarity)	2,769,939 (14.4%)	1,536,270 (21.2%)

[1] Hidden Markov Models. [2]Indicate the reads that were mapped back to OTUs.

REFERENCES

1. Duarte CM (2015) Seafaring in the 21St Century: The Malaspina 2010 Circumnavigation Expedition. *Limnology and Oceanography Bulletin* 24(1):11-14.
2. Grasshoff K, Ehrhardt M, & Kremling K (1983) *Methods on seawater analysis*.
3. Boyer TP*, et al.* (2013) World Ocean Database 2013. in *NOAA Atlas NESDIS 72*, eds Levitus S & Mishonov A (NOAA, Silver Spring, MD).
4. Massana R, Murray AE, Preston CM, & DeLong EF (1997) Vertical distribution and phylogenetic characterization of marine planktonic Archaea in the Santa Barbara Channel. *Appl Environ Microbiol* 63(1):50-56.
5. Stoeck T*, et al.* (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* 19 Suppl 1:21-31.
6. Parada AE, Needham DM, & Fuhrman JA (2016) Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol* 18(5):1403-1414.
7. Logares R (2017) Workflow for Analysing MiSeq Amplicons based on Uparse v1.5.
8. Nikolenko SI, Korobeynikov AI, & Alekseyev MA (2013) BayesHammer: Bayesian clustering for error correction in single-cell sequencing. *BMC Genomics* 14 Suppl 1:S7.
9. Schirmer M*, et al.* (2015) Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. *Nucleic Acids Res* 43(6):e37.
10. Zhang J, Kobert K, Flouri T, & Stamatakis A (2014) PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30(5):614-620.
11. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460-2461.
12. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* 10(10):996-998.
13. Quast C*, et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41(Database issue):D590-596.
14. Altschul SF, Gish W, Miller W, Myers EW, & Lipman DJ (1990) Basic local alignment search tool. *Journal of molecular biology* 215(3):403-410.
15. Guillou L*, et al.* (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* 41(Database issue):D597-604.
16. Pernice MC, Logares R, Guillou L, & Massana R (2013) General patterns of diversity in major marine microeukaryote lineages. *PLoS One* 8(2):e57170.
17. Massana R*, et al.* (2015) Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ Microbiol* 17(10):4035-4049.
18. Oksanen J*, et al.* (2008) vegan: Community Ecology Package. R package version 1.15-0.

19. Logares R, *et al.* (2014) Patterns of rare and abundant marine microbial eukaryotes. *Curr Biol* 24(8):813-821.
20. R-Development-Core-Team (2008) *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, Vienna, Austria).
21. Wickham H (2009) *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag).
22. Becker RA, Wilks AR, Brownrigg R, Minka TP, & Deckmyn A (2017) maps: Draw Geographical Maps.
23. Gerritsen H (2014) mapplots: Data Visualisation on Maps.
24. Legendre P & De Caceres M (2013) Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecol Lett* 16(8):951-963.
25. Dray S, *et al.* (2017) adespatial: Multivariate Multiscale Spatial Analysis.
26. Schloss PD, *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75(23):7537-7541.
27. Capella-Gutierrez S, Silla-Martinez JM, & Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15):1972-1973.
28. Price MN, Dehal PS, & Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* 26(7):1641-1650.
29. Kembel SW, *et al.* (2010) Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26(11):1463-1464.
30. Paradis E, Claude J, & Strimmer K (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* 20(2):289-290.
31. Chen J, *et al.* (2012) Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* 28(16):2106-2113.
32. Stegen JC, *et al.* (2013) Quantifying community assembly processes and identifying features that impose them. *ISME J* 7(11):2069-2079.
33. Zhou J & Ning D (2017) Stochastic Community Assembly: Does It Matter in Microbial Ecology? *Microbiol Mol Biol Rev* 81(4):e00002-00017.
34. Chase JM, Kraft NJB, Smith KG, Vellend M, & Inouye BD (2011) Using null models to disentangle variation in community dissimilarity from variation in α-diversity. *Ecosphere* 2(2):1-11.
35. Longhurst AR (2007) *Ecological Geography of the Sea* (Academic Press).
36. Estrada M, *et al.* (2016) Phytoplankton across Tropical and Subtropical Regions of the Atlantic, Indian and Pacific Oceans. *PLoS One* 11(3):e0151699.
37. Catalá TS, *et al.* (2016) Drivers of fluorescent dissolved organic matter in the global epipelagic ocean. *Limnology and Oceanography* 61(3):1101-1119.
38. Pérez-Mazuecos I (2015) Exopolymer particles in the ocean: production by microorganisms, carbon export and mesopelagic respiration. (University of Granada).
39. Catalá TS, *et al.* (2016) Chromophoric signatures of microbial by-products in the dark ocean. *Geophysical Research Letters* 43(14):7639-7648.
40. Fernández-Castro B, *et al.* (2015) Importance of salt fingering for new nitrogen supply in the oligotrophic ocean. *Nature communications* 6:8002.

41. Reshef DN, *et al.* (2011) Detecting novel associations in large data sets. *Science* 334(6062):1518-1524.
42. Friedman J & Alm EJ (2012) Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 8(9):e1002687.
43. Girvan M & Newman ME (2002) Community structure in social and biological networks. *Proc Natl Acad Sci U S A* 99(12):7821-7826.

Bibliography

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA Genes. *PLoS ONE* 4:e6372. DOI: 10.1371/journal.pone.0006372.

Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, Stamatakis A. 2018. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Systematic Biology* 68:365–369. DOI: 10.1093/sysbio/syy054.

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2017. GenBank. *Nucleic Acids Research* 46:D41–D47. DOI: 10.1093/nar/gkw1070.

Berger SA, Krompass D, Stamatakis A. 2011. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Systematic Biology* 60:291–302. DOI: 10.1093/sysbio/syr010.

Berger S, Stamatakis A. 2012. PaPaRa 2.0: A Vectorized Algorithm for Probabilistic Phylogeny-Aware Alignment Extension. *Heidelberg Institute for Theoretical Studies*.

Bouman HA, Ulloa O, Barlow R, Li WKW, Platt T, Zwirglmaier K, Scanlan DJ, Sathyendranath S. 2011. Water-column stratification gover11ns the community structure of subtropical marine picophytoplankton. *Environmental Microbiology Reports* 3:473–482. DOI: 10.1111/j.1758-2229.2011.00241.x.

Buitenhuis ET, Li WKW, Vaulot D, Lomas MW, Landry MR, Partensky F, Karl DM, Ulloa O, Campbell L, Jacquet S, Lantoine F, Chavez F, MacIas D, Gosselin M, McManus GB. 2012. Picophytoplankton biomass distribution in the global ocean. *Earth System Science Data* 4:37–46. DOI: 10.5194/essd-4-37-2012.

Burns JA, Pittis AA, Kim E. 2018. Gene-based predictive models of trophic modes suggests Asgard archaea are not phagocytotic. *Nature Ecology and Evolution* 2:697-704. DOI: 10.1038/s41559-018-0477-7.

Del Campo J, Sieracki ME, Molestina R, Keeling P, Massana R, Ruiz-Trillo I. 2014. The others: Our biased perspective of eukaryotic genomes. *Trends in Ecology and Evolution* 29:252–259. DOI: 10.1016/j.tree.2014.03.006.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973. DOI: 10.1093/bioinformatics/btp348.

Carradec Q, Pelletier E, Da Silva C, Alberti A, Seeleuthner Y, Blanc-Mathieu R, Lima-Mendez G, Rocha F, Tirichine L, Labadie K, Kirilovsky A, Bertrand A, Engelen S, Madoui MA, Méheust R, Poulain J, Romac S, Richter DJ, Yoshikawa G, Dimier C, Kandels-Lewis S, Picheral M, Searson S, Acinas SG, Boss E, Follows M, Gorsky G, Grimsley N, Karp-Boss L, Krzic U, Pesant S, Reynaud EG, Sardet C, Sieracki M, Speich S, Stemmann L, Velayoudon D, Weissenbach J, Jaillon O, Aury JM, Karsenti E, Sullivan MB, Sunagawa S, Bork P, Not F, Hingamp P, Raes J, Guidi L, Ogata H, De Vargas C, Iudicone D, Bowler C, Wincker P. 2018. A global ocean atlas of eukaryotic genes. *Nature Communications* 9:373. DOI: 10.1038/s41467-017-02342-1.

Ciais P, Sabine C, Bala G, Bopp L, Brovkin V, Canadell J, Chhabra A, DeFries R, Galloway J, Heimann M, Jones C, Quéré C Le, Myneni RB, Piao S, Thornton P. 2013. Carbon and Other Biogeochemical Cycles. In: Stocker TF, Qin D, Plattner G-K, Tignor M, Allen SK, Boschung J, Nauels A, Xia Y, Bex V, Midgle PM eds. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 465–570. DOI: 10.1017/CBO9781107415324.015.

Cuvelier ML, Allen AE, Monier A, McCrow JP, Messié M, Tringe SG, Woyke T, Welsh RM, Ishoey T, Lee JH, Binder BJ, DuPont CL, Latasa M, Guigand C, Buck KR, Hilton J, Thiagarajan M, Caler E, Read B, Lasken RS, Chavez FP, Worden AZ. 2010. Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proceedings of the National Academy of Sciences of the United States of America* 107:14679–14684. DOI: 10.1073/pnas.1001665107.

Czech L, Barbera P, Stamatakis A. 2019. Genesis and Gappa: Processing, Analyzing and Visualizing Phylogenetic (Placement) Data. *bioRxiv*:647958. DOI: 10.1101/647958.

Czech L, Stamatakis A. 2019. Scalable methods for analyzing and visualizing phylogenetic placement of metagenomic samples. *PLoS ONE* 14:e0219925. DOI: 10.1371/journal.pone.0217050.

Decelle J, Romac S, Stern RF, Bendif EM, Zingone A, Audic S, Guiry MD, Guillou L, Tessier D, Le Gall F, Gourvil P, Dos Santos AL, Probert I, Vaulot D, de Vargas C, Christen R. 2015. PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Molecular Ecology Resources* 15:1435–1445. DOI: 10.1111/1755-0998.12401.

Duarte CM. 2015. Seafaring in the 21st century: The Malaspina 2010 circumnavigation expedition. *Limnology and Oceanography Bulletin* 24:11–14. DOI: 10.1002/lob.10008.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* 14:755–763. DOI: 10.1093/bioinformatics/14.9.755.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32:1792–1797. DOI: 10.1093/nar/gkh340.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26:2460–2461. DOI: 10.1093/bioinformatics/btq461.

Engene N, Gerwick WH. 2011. Intra-genomic 16S rRNA gene heterogeneity in cyanobacterial genomes. *Fottea* 1:17–24. DOI: 10.5507/fot.2011.003.

Estrada M, Delgado M, Blasco D, Latasa M, Cabello AM, Benítez-Barrios V, Fraile-Nuez E, Mozetič P, Vidal M. 2016. Phytoplankton across tropical and subtropical regions of the Atlantic, Indian and Pacific oceans. *PLoS ONE* 11:e0151699. DOI: 10.1371/journal.pone.0151699.

Farrant GK, Doré H, Cornejo-Castillo FM, Partensky F, Ratin M, Ostrowski M, Pitt FD, Wincker P, Scanlan DJ, Iudicone D, Acinas SG, Garczarek L. 2016. Delineating ecologically significant taxonomic units from global patterns of marine picocyanobacteria. *Proceedings of the National Academy of Sciences of the United States of America* 113:E3365–E3374. DOI: 10.1073/pnas.1524865113.

Faure E, Not F, Benoiston AS, Labadie K, Bittner L, Ayata SD. 2019. Mixotrophic protists display contrasted biogeographies in the global ocean. *ISME Journal* 13:1072–1083. DOI: 10.1038/s41396-018-0340-5.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. 1998. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* 281:237–240. DOI: 10.1126/science.281.5374.237.

Giner CR, Balagué V, Krabberød AK, Ferrera I, Reñé A, Garcés E, Gasol JM, Logares R, Massana R. 2019. Quantifying long-term recurrence in planktonic microbial eukaryotes. *Molecular Ecology* 28:923–935. DOI: 10.1111/mec.14929.

Gonzalez JM, Portillo MC, Belda-Ferre P, Mira A. 2012. Amplification by PCR artificially reduces the proportion of the rare biosphere in microbial communities. *PLoS ONE* 7:e29973. DOI: 10.1371/journal.pone.0029973.

Gould SB, Waller RF, McFadden GI. 2008. Plastid Evolution. *Annual Review of Plant Biology* 59:491–517. DOI: 10.1146/annurev.arplant.59.032607.092915.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. *Plant Journal* 66:34–44. DOI: 10.1111/j.1365-313X.2011.04541.x.

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, De Vargas C, Decelle J, Del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Logares R, Mahé F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet AL, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R. 2013. The Protist Ribosomal Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research* 41:D597–604. DOI: 10.1093/nar/gks1160.

Han M V., Zmasek CM. 2009. PhyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics* 10:356. DOI: 10.1186/1471-2105-10-356.

Hao X, Chen T. 2012. OTU Analysis Using Metagenomic Shotgun Sequencing Data. *PLoS ONE* 7:e49785. DOI: 10.1371/journal.pone.0049785.

Harrison PW, Alako B, Amid C, Cerdeño-Tárraga A, Cleland I, Holt S, Hussein A, Jayathilaka S, Kay S, Keane T, Leinonen R, Liu X, Martínez-Villacorta J, Milano A, Pakseresht N, Rajan J, Reddy K, Richards E, Rosello M, Silvester N, Smirnov D, Toribio AL, Vijayaraja S, Cochrane G. 2019. The European Nucleotide Archive in 2018. *Nucleic Acids Research* 47:D84–D88. DOI: 10.1093/nar/gky1078.

Huang S, Wilhelm SW, Harvey HR, Taylor K, Jiao N, Chen F. 2012. Novel lineages of prochlorococcus and synechococcus in the global oceans. *ISME Journal* 6:285–297. DOI: 10.1038/ismej.2011.106.

Ibarbalz F, Henry N, Brandão MC, Martini S, Busseni G, Byrne H, Coelho LP, Endo H, Gasol JM, Gregory AC, Mahé F, Rigonato J, Royo-Llonch M, Salazar G, Sanz-Sáez I, Scalco E, Soviadan D, Zayed AA, Zingone A, Labadie K, Ferland J, Marec C, Kandels S, Picheral M, Dimier C, Poulain J, Pisarev S, Carmichael M, Pesant S, Tara Oceans Coordinators, Babin M, Boss E, Iudicone D, Jaillon O, Acinas SG, Ogata H, Pelletier E, Stemmann L, Sullivan MB, Sunagawa S, Bopp L, de Vargas C, Karp-Boss L, Wincker P, Lombard F, Bowler C, Zinger L. 2019. Global trends in marine plankton diversity across kingdoms of life. *Cell* 179:1068–1083.e21. DOI: 10.1016/j.cell.2019.10.008

Jamy M, Foster R, Barbera P, Czech L, Kozlov AM, Stamatakis AM, Bass D, Burki F. 2019. Long metabarcoding of the eukaryotic rDNA operon to phylogenetically and taxonomically resolve environmental diversity. *Molecular ecology resources* 00:1–15. DOI: 10.1101/627828.

Jardillier L, Zubkov M V., Pearman J, Scanlan DJ. 2010. Significant CO2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *ISME Journal* 4:1180–1192. DOI: 10.1038/ismej.2010.36.

Jones H. 1997. A classification of mixotrophic protists based on their behaviour. *Freshwater Biology* 37.

Karsenti E. 2012. A journey from reductionist to systemic cell biology aboard the schooner Tara. *Molecular biology of the cell* 23:2403–2406. DOI: 10.1091/mbc.E11-06-0571.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30:772–780. DOI: 10.1093/molbev/mst010.

Keeling P. 2004. A brief history of plastids and their hosts. *Protist* 155:3–7. DOI: 10.1078/1434461000156.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, Beszteri B, Bidle KD, Cameron CT, Campbell L, Caron DA, Cattolico RA, Collier JL, Coyne K, Davy SK, Deschamps P, Dyhrman ST, Edvardsen B, Gates RD, Gobler CJ, Greenwood SJ, Guida SM,

Jacobi JL, Jakobsen KS, James ER, Jenkins B, John U, Johnson MD, Juhl AR, Kamp A, Katz LA, Kiene R, Kudryavtsev A, Leander BS, Lin S, Lovejoy C, Lynn D, Marchetti A, McManus G, Nedelcu AM, Menden-Deuer S, Miceli C, Mock T, Montresor M, Moran MA, Murray S, Nadathur G, Nagai S, Ngam PB, Palenik B, Pawlowski J, Petroni G, Piganeau G, Posewitz MC, Rengefors K, Romano G, Rumpho ME, Rynearson T, Schilling KB, Schroeder DC, Simpson AGB, Slamovits CH, Smith DR, Smith GJ, Smith SR, Sosik HM, Stief P, Theriot E, Twary SN, Umale PE, Vaulot D, Wawrik B, Wheeler GL, Wilson WH, Xu Y, Zingone A, Worden AZ. 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:e1001889. DOI: 10.1371/journal.pbio.1001889.

Kembel SW, Cowan PD, Helmus MR, Cornwell WK, Morlon H, Ackerly DD, Blomberg SP, Webb CO. 2010. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* 26:1463–1464. DOI: 10.1093/bioinformatics/btq166.

Kirkham AR, Lepère C, Jardillier LE, Not F, Bouman H, Mead A, Scanlan DJ. 2013. A global perspective on marine photosynthetic picoeukaryote community structure. *ISME Journal* 7:922–936. DOI: 10.1038/ismej.2012.166.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* btz305:1–3. DOI: 10.1093/bioinformatics/btz305.

Kozlov AM, Zhang J, Yilmaz P, Glöckner FO, Stamatakis A. 2016. Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research* 44:5022–5033. DOI: 10.1093/nar/gkw396.

Latasa M, Bidigare RR. 1998. A comparison of phytoplankton populations of the Arabian Sea during the Spring Intermonsoon and Southwest Monsoon of 1995 as described by HPLC-analyzed pigments. *Deep-Sea Research Part II: Topical Studies in Oceanography* 2:2133–2170. DOI: 10.1016/S0967-0645(98)00066-6.

Leliaert F, Verbruggen H, Zechman FW. 2011. Into the deep: New discoveries at the base of the green plant phylogeny. *BioEssays* 33:683–692. DOI: 10.1002/bies.201100035.

Li WKW. 1989. Shipboard analytical flow cytometry of oceanic ultraphytoplankton. *Cytometry* 10:564–579. DOI: 10.1002/cyto.990100512.

Li WKW, Irwin BD, Dickie PM. 1993. Dark fixation of 14C: Variations related to biomass and productivity of phytoplankton and bacteria. *Limnology and Oceanography* 38:483–494. DOI: 10.4319/lo.1993.38.3.0483.

Li WKW. 1994. Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. *Limnology and Oceanography* 39:169–175. DOI: 10.4319/lo.1994.39.1.0169.

Liu H, Probert I, Uitz J, Claustre H, Aris-Brosou S, Frada M, Not F, De Vargas C. 2009. Extreme diversity in noncalcifying haptophytes explains a major pigment

paradox in open oceans. *Proceedings of the National Academy of Sciences of the United States of America* 106:12803–12808. DOI: 10.1073/pnas.0905841106.

Liu Z, Campbell V, Heidelberg KB, Caron DA. 2016. Gene expression characterizes different nutritional strategies among three mixotrophic protists. *FEMS Microbiology Ecology* 92:fiw106. DOI: 10.1093/femsec/fiw106.

Liu Z, Hu SK, Campbell V, Tatters AO, Heidelberg KB, Caron DA. 2017. Single-cell transcriptomics of small microbial eukaryotes: Limitations and potential. *ISME Journal* 11 :1282-1285. DOI: 10.1038/ismej.2016.190.

Logares R, Deutschmann IM, Giner CR, Krabberød AK, Schmidt TSB, Rubinat-Ripoll L, Mestre M, Salazar G, Ruiz-González C, Sebastián M, de Vargas C, Acinas SG, Duarte CM, Gasol JM, Massana R. 2018. Different processes shape prokaryotic and picoeukaryotic assemblages in the sunlit ocean microbiome. *bioRxiv*:374298. DOI: 10.1101/374298.

Logares R, Sunagawa S, Salazar G, Cornejo-Castillo FM, Ferrera I, Sarmento H, Hingamp P, Ogata H, de Vargas C, Lima-Mendez G, Raes J, Poulain J, Jaillon O, Wincker P, Kandels-Lewis S, Karsenti E, Bork P, Acinas SG. 2013. Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental microbiology* 16:2659–2671. DOI: 10.1111/1462-2920.12250.

Lynn DH, Pinheiro M. 2009. A survey of polymerase chain reaction (pcr) amplification studies of unicellular protists using single-cell pcr. *Journal of Eukaryotic Microbiology* 56:406–412. DOI: 10.1111/j.1550-7408.2009.00439.x.

Mackey MD, Mackey DJ, Higgins HW, Wright SW. 1996. CHEMTAX - A program for estimating class abundances from chemical markers: Application to HPLC measurements of phytoplankton. *Marine Ecology Progress Series* 144:265–283. DOI: 10.3354/meps144265.

Madeira F, Park Y mi, Lee J, Buso N, Gur T, Madhusoodanan N, Basutkar P, Tivey ARN, Potter SC, Finn RD, Lopez R. 2019. The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Research* 47:W636–W641. DOI: 10.1093/nar/gkz268.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. 2015. Swarm v2: highly-scalable and high-resolution amplicon clustering. *PeerJ* 3:e1420. DOI: 10.7717/peerj.1420.

Mahé F, De Vargas C, Bass D, Czech L, Stamatakis A, Lara E, Singer D, Mayor J, Bunge J, Sernaker S, Siemensmeyer T, Trautmann I, Romac S, Berney C, Kozlov A, Mitchell EAD, Seppey CVW, Egge E, Lentendu G, Wirth R, Trueba G, Dunthorn M. 2017. Parasites dominate hyperdiverse soil protist communities in Neotropical rainforests. *Nature Ecology and Evolution* 1. DOI: 10.1038/s41559-017-0091.

Mangot JF, Logares R, Sánchez P, Latorre F, Seeleuthner Y, Mondy S, Sieracki ME, Jaillon O, Wincker P, Vargas C De, Massana R. 2017. Accessing the genomic information of unculturable oceanic picoeukaryotes by combining multiple single cells. *Scientific Reports* 7:41498. DOI: 10.1038/srep41498.

Massana R, Del Campo J, Sieracki ME, Audic S, Logares R. 2014. Exploring the uncultured microeukaryote majority in the oceans: Reevaluation of ribogroups within stramenopiles. *ISME Journal* 8:854–866. DOI: 10.1038/ismej.2013.204.

Matsen FA, Gallagher A, McCoy CO. 2013. Minimizing the average distance to a closest leaf in a phylogenetic tree. *Systematic Biology* 62:824–836. DOI: 10.1093/sysbio/syt044.

Mitra A, Flynn KJ, Burkholder JM, Berge T, Calbet A, Raven JA, Granéli E, Glibert PM, Hansen PJ, Stoecker DK, Thingstad F, Tillmann U, Väge S, Wilken S, Zubkov M V. 2014. The role of mixotrophic protists in the biological carbon pump. *Biogeosciences* 11:995–1005. DOI: 10.5194/bg-11-995-2014.

Mitra A, Flynn KJ, Tillmann U, Raven JA, Caron D, Stoecker DK, Not F, Hansen PJ, Hallegraeff G, Sanders R, Wilken S, McManus G, Johnson M, Pitta P, Våge S, Berge T, Calbet A, Thingstad F, Jeong HJ, Burkholder JA, Glibert PM, Granéli E, Lundgren V. 2016. Defining Planktonic Protist Functional Groups on Mechanisms for Energy and Nutrient Acquisition: Incorporation of Diverse Mixotrophic Strategies. *Protist* 167:106–120. DOI: 10.1016/j.protis.2016.01.003.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America* 93:2873–2878. DOI: 10.1073/pnas.93.7.2873.

Not F, Simon N, Biegala IC, Vaulot D. 2002. Application of fluorescent in situ hybridization coupled with tyramide signal amplification (FISH-TSA) to assess eukaryotic picoplankton composition. *Aquatic Microbial Ecology* 28:157–166. DOI: 10.3354/ame028157.

Not F, Massana R, Latasa M, Marie D, Colson C, Eikrem W, Pedrós-Alió C, Vaulot D, Simon N. 2005. Late summer community composition and abundance of photosynthetic picoeukaryotes in Norwegian and Barents Seas. *Limnology and Oceanography* 50:1677–1686. DOI: 10.4319/lo.2005.50.5.1677.

Not F, Latasa M, Scharek R, Viprey M, Karleskind P, Balagué V, Ontoria-Oviedo I, Cumino A, Goetze E, Vaulot D, Massana R. 2008. Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. *Deep-Sea Research Part I: Oceanographic Research Papers* 55:1456–1473. DOI: 10.1016/j.dsr.2008.06.007.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD. 2016. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* 44:D733–D745. DOI: 10.1093/nar/gkv1189.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Henry MH, Szoecs E, Wagner H. 2018. vegan: Community Ecology Package. *R package*.

Olson RJ, Frankel SL, Chisholm SW, Shapiro HM. 1983. An inexpensive flow cytometer for the analysis of fluorescence signals in phytoplankton: Chlorophyll and DNA distributions. *Journal of Experimental Marine Biology and Ecology* 68:129–144. DOI: 10.1016/0022-0981(83)90155-7.

Olson RJ, Chisholm SW, Zettler ER, Armbrust E V. 1988. Analysis of Synechococcus pigment types in the sea using single and dual beam flow cytometry. *Deep Sea Research Part A, Oceanographic Research Papers* 35:425–440. DOI: 10.1016/0198-0149(88)90019-2.

Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, Poulton NJ, Burkart MD, La Clair JJ, Chisholm SW, Stepanauskas R. 2019. Charting the Complexity of the Marine Microbiomethrough Single-Cell Genomics. *Cell* 179:1623–1635. DOI: 10.1016/j.cell.2019.11.017

Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: Assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environmental Microbiology* 18:1403–1414. DOI: 10.1111/1462-2920.13023.

Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology* 36:996–1004. DOI: 10.1038/nbt.4229.

Passow U, Carlson CA. 2012. The biological pump in a high CO2 world. *Marine Ecology Progress Series* 470:249–272. DOI: 10.3354/meps09985.

Pérez V, Fernández E, Marañón E, Morán XAG, Zubkov M V. 2006. Vertical distribution of phytoplankton biomass, production and growth in the Atlantic subtropical gyres. *Deep-Sea Research Part I: Oceanographic Research Papers* 1:1616–1634. DOI: 10.1016/j.dsr.2006.07.008.

Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, Iudicone D, Karsenti E, Speich S, Trouble R, Dimier C, Searson S. 2015. Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data* 2:150023. DOI: 10.1038/sdata.2015.23

Pierella Karlusich JJ, Ibarbalz FM, Bowler C. 2020. Phytoplankton in Tara Oceans. *Annual Review of Marine Science* 12:18.1–18.33. DOI: 10.1146/annurev-marine-010419-010706.

Pillet L, Fontaine D, Pawlowski J. 2012. Intra-genomic ribosomal RNA polymorphism and morphological variation in elphidium macellum suggests inter-specific hybridization in foraminifera. *PLoS ONE* 7:e32373. DOI: 10.1371/journal.pone.0032373.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: Improved data

processing and web-based tools. *Nucleic Acids Research* 41:D590–D596. DOI: 10.1093/nar/gks1219.

Rii YM, Duhamel S, Bidigare RR, Karl DM, Repeta DJ, Church MJ. 2016. Diversity and productivity of photosynthetic picoeukaryotes in biogeochemically distinct regions of the South East Pacific Ocean. *Limnology and Oceanography* 61:806–824. DOI: 10.1002/lno.10255.

Reyes-Prieto A, Weber APM, Bhattacharya D. 2007. The Origin and Establishment of the Plastid in Algae and Plants. *Annual Review of Genetics* 41:147–68. DOI: 10.1146/annurev.genet.41.110306.130134.

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* 4:e2584. DOI: 10.7717/peerj.2584.

Rose JM, Caron DA. 2007. Does low temperature constrain the growth rates of heterotrophic protists? Evidence and implications for algal blooms in cold waters. *Limnology and Oceanography* 52:886–895. DOI: 10.4319/lo.2007.52.2.0886.

Sanchez-Puerta MV, Delwiche CF. 2008. A hypothesis for plastid evolution in chromalveolates. *Journal of Phycology* 44:1097–1107. DOI: 10.1111/j.1529-8817.2008.00559.x.

Dos Santos AL, Gourvil P, Tragin M, Noël MH, Decelle J, Romac S, Vaulot D. 2017. Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. *ISME Journal* 11:512–528. DOI: 10.1038/ismej.2016.120.

Shi XL, Lepère C, Scanlan DJ, Vaulot D. 2011. Plastid 16S rRNA gene diversity among eukaryotic picophytoplankton sorted by flow cytometry from the South Pacific Ocean. *PLoS ONE* 6:e18979. DOI: 10.1371/journal.pone.0018979.

Sieracki ME, Poulton NJ, Jaillon O, Wincker P, de Vargas C, Rubinat-Ripoll L, Stepanauskas R, Logares R, Massana R. 2019. Single cell genomics yields a wide diversity of small planktonic protists across major ocean ecosystems. *Scientific reports* 9:6025. DOI: 10.1038/s41598-019-42487-1.

Simon N, Lebot N, Marie D, Partensky F, Vaulot D. 1995. Fluorescent in situ hybridization with rRNA-targeted oligonucleotide probes to identify small phytoplankton by flow cytometry. *Applied and Environmental Microbiology* 61:2506–2513.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313. DOI: 10.1093/bioinformatics/btu033.

Stepanauskas R, Fergusson EA, Brown J, Poulton NJ, Tupper B, Labonté JM, Becraft ED, Brown JM, Pachiadaki MG, Povilaitis T, Thompson BP, Mascena CJ, Bellows WK, Lubys A. 2017. Improved genome recovery and integrated cell-size analyses of individual uncultured microbial cells and viral particles. *Nature Communications* 8:1–10. DOI: 10.1038/s41467-017-00128-z.

Stickney HL, Hood RR, Stoecker DK. 2000. The impact of mixotrophy on planktonic marine ecosystems. *Ecological Modelling* 125:203–230. DOI: 10.1016/S0304-3800(99)00181-7.

Stoecker DK. 1998. Conceptual models of mixotrophy in planktonic protists and some ecological and evolutionary implications. *European Journal of Protistology* 34:281–290. DOI: 10.1016/S0932-4739(98)80055-2.

Stoecker DK, Hansen PJ, Caron DA, Mitra A. 2017. Mixotrophy in the Marine Plankton. *Annual Review of Marine Science* 9:311–335. DOI: 10.1146/annurev-marine-010816-060617.

Stoecker DK, Lavrentyev PJ. 2018. Mixotrophic plankton in the polar seas: A pan-Arctic review. *Frontiers in Marine Science* 5:292. DOI: 10.3389/fmars.2018.00292.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, D'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Boss E, Follows M, Karp-Boss L, Krzic U, Reynaud EG, Sardet C, Sieracki M, Velayoudon D, Bowler C, De Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P. 2015. Structure and function of the global ocean microbiome. *Science* 348:1261359. DOI: 10.1126/science.1261359.

Thornton DCO. 2014. Dissolved organic matter (DOM) release by phytoplankton in the contemporary and future ocean. *European Journal of Phycology* 49:20–46. DOI: 10.1080/09670262.2013.875596.

Tragin M, Vaulot D. 2018. Green microalgae in marine coastal waters: The Ocean Sampling Day (OSD) dataset. *Scientific Reports* 8:14020. DOI: 10.1038/s41598-018-32338-w.

De Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney Ć, Le Bescot N, Probert I, Carmichael M, Poulain J, Romac S, Colin S, Aury JM, Bittner L, Chaffron S, Dunthorn M, Engelen S, Flegontova O, Guidi L, Horák A, Jaillon O, Lima-Mendez G, Lukeš J, Malviya S, Morard R, Mulot M, Scalco E, Siano R, Vincent F, Zingone A, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Acinas SG, Bork P, Bowler C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Not F, Ogata H, Pesant S, Raes J, Sieracki ME, Speich S, Stemmann L, Sunagawa S, Weissenbach J, Wincker P, Karsenti E, Boss E, Follows M, Karp-Boss L, Krzic U, Reynaud EG, Sardet C, Sullivan MB, Velayoudon D. 2015. Eukaryotic plankton diversity in the sunlit ocean. *Science* 348:1261605. DOI: 10.1126/science.1261605.

Vorobev A, Dupouy M, Carradec Q, Delmont TO, Annamalé A, Wincker P, Pelletier E. 2019. Transcriptome reconstruction and functional analysis of eukaryotic marine plankton communities via high-throughput metagenomics and metatranscriptomics. *bioRxiv*. DOI: 10.1101/812974.

Walker J.C.G. 1980. The Oxygen Cycle. In: The Natural Environment and the Biogeochemical Cycles. *The Handbook of Environmental Chemistry*. Springer, Berlin, Heidelberg. DOI 10.1007/978-3-662-24940-6_5.

Ward BA, Follows MJ. 2016. Marine mixotrophy increases trophic transfer efficiency, mean organism size, and vertical carbon flux. *Proceedings of the National Academy of Sciences of the United States of America* 113:2958–2963. DOI: 10.1073/pnas.1517118113.

Westberry T, Behrenfeld MJ, Siegel DA, Boss E. 2008. Carbon-based primary productivity modeling with vertically resolved photoacclimation. *Global Biogeochemical Cycles* 22:GB2024. DOI: 10.1029/2007GB003078.

Wilken S, Yung CCM, Hamilton M, Hoadley K, Nzongo J, Eckmann C, Corrochano-Luque M, Poirier C, Worden AZ. 2019. The need to account for cell biology in characterizing predatory mixotrophs in aquatic environments. *Phil. Trans. R. Soc. B* 374:20190090. DOI: 10.1098/rstb.2019.0090.

Worden AZ, Nolan JK, Palenik B. 2004. Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnology and Oceanography* 49:168–179. DOI: 10.4319/lo.2004.49.1.0168.

Woyke T, Doud DFR, Schulz F. 2017. The trajectory of microbial single-cell sequencing. *Nature Methods* 14:1045–1054. DOI: 10.1038/nmeth.4469.

Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, Whitman WB, Euzéby J, Amann R, Rosselló-Móra R. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology* 12:635–645. DOI: 10.1038/nrmicro3330.

Yoon HS, Price DC, Stepanauskas R, Rajah VD, Sieracki ME, Wilson WH, Yang EC, Duffy S, Bhattacharya D. 2011. Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* 332:714–717. DOI: 10.1126/science.1203163.

Zhang J, Kobert K, Flouri T, Stamatakis A. 2014. PEAR: A fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30:614–620. DOI: 10.1093/bioinformatics/btt593.