ORIGINAL ARTICLE

Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters

Adriana Lopes dos Santos[1], Priscillia Gourvil[1], Margot Tragin[1], Mary-Hélène Noël[2], Johan Decelle[1,3], Sarah Romac[1] and Daniel Vaulot[1]

[1]Sorbonne Universités, UPMC Université Paris 06, CNRS, UMR 7144 Station Biologique de Roscoff, Roscoff, France; [2]National Institute for Environmental Studies, Tsukuba, Japan and [3]Department of Isotope Biogeochemistry, Helmholtz Centre for Environmental Research - UFZ, Leipzig, Germany

Prasinophytes clade VII is a group of pico/nano-planktonic green algae (division Chlorophyta) for which numerous ribosomal RNA (rRNA) sequences have been retrieved from the marine environment in the last 15 years. A large number of strains have also been isolated but have not yet received a formal taxonomic description. A phylogenetic analysis of available strains using both the nuclear 18S and plastidial 16S rRNA genes demonstrates that this group composes at least 10 different clades: A1–A7 and B1–B3. Analysis of sequences from the variable V9 region of the 18S rRNA gene collected during the Tara Oceans expedition and in the frame of the Ocean Sampling Day consortium reveal that clade VII is the dominant Chlorophyta group in oceanic waters, replacing Mamiellophyceae, which have this role in coastal waters. At some location, prasinophytes clade VII can even be the dominant photosynthetic eukaryote representing up to 80% of photosynthetic metabarcodes overall. B1 and A4 are the overall dominant clades and different clades seem to occupy distinct niches, for example, A6 is dominant in surface Mediterranean Sea waters, whereas A4 extend to high temperate latitudes. Our work demonstrates that prasinophytes clade VII constitute a highly diversified group, which is a key component of phytoplankton in open oceanic waters but has been neglected in the conceptualization of marine microbial diversity and carbon cycle.
The ISME Journal advance online publication, 25 October 2016; doi:10.1038/ismej.2016.120

Introduction

Ocean photosynthesis is dominated by phytoplankton, a functional group of single-cell organisms including prokaryotes and eukaryotes. Photosynthetic eukaryotes, although less abundant in numbers than cyanobacteria, can dominate carbon production and biomass in oceanic and coastal waters (Li, 1995; Worden et al., 2004; Rii et al., 2016). These unicellular organisms are taxonomically diverse with representatives in all of the supergroups of the eukaryotic tree of life (Not et al., 2012).

Four taxonomic groups have emerged as important factors within eukaryotic phytoplankton community: diatoms, dinoflagellates, haptophytes and chlorophyta (Not et al., 2012). The first three groups are mostly found in the micro-plankton (20–200 µm) and nano-plankton (2–20 µm) size ranges, whereas

Chlorophyta dominate the smaller size fractions (pico-plankton - 0.2–2 µm). Abundant groups from the micro- and nano-plankton, such as diatoms and dinoflagellates, have been broadly investigated, in contrast to members of the green lineage. The importance and diversity of Chlorophyta have not been extensively studied, in part because many are of small size and lack distinctive morphological features. Several early diverging clades of mainly marine unicellular algae gathered under the term prasinophytes located at the base of Chlorophyta evolution (Lemieux et al., 2014). The paraphyletic origin of prasinophytes is reflected in the wide range of cell shapes (Leliaert et al., 2012; Tragin et al., 2015) and photosynthetic pigments (Latasa et al., 2004) found among them. Within prasinophytes, nine distinct lineages (known as clades I–IX) have been distinguished on the base of 18S ribosomal RNA (rRNA) gene phylogeny (Guillou et al., 2004; Viprey et al., 2008). These lineages correspond to established taxonomic levels such as Class (for example, Mamiellophyceae) or Order (for example, Prasinococcales) or just regroup undescribed strains (clade VII) or even only environmental sequences (clade IX). The class Mamiellophyceae (Marin and

2

Melkonian, 2010), corresponding to clade II (Guillou *et al.*, 2004), contains three important genera of marine pico-phytoplankton: *Micromonas*, *Bathycoccus* and *Ostreococcus*. They are typical of coastal waters (Not *et al.*, 2004; Romari and Vaulot, 2004; Collado-Fabri *et al.*, 2011) where they can form sporadic blooms (O'Kelly *et al.*, 2003) but have also been observed in open oceanic waters (Foulon *et al.*, 2008; Monier *et al.*, 2012; Treusch *et al.*, 2012; Vaulot *et al.*, 2012). The genus *Micromonas* can be dominant in Arctic ecosystems (Lovejoy *et al.*, 2007; Balzano *et al.*, 2012). Members of these three genera are easily isolated into cultures and genome sequences from representative strains are available (Derelle *et al.*, 2006; Worden *et al.*, 2009; Moreau *et al.*, 2012), fostering their use as biological models for marine photosynthetic eukaryotes.

Another group of prasinophytes, clade VII (Guillou *et al.*, 2004), has been suggested to be important in oceanic waters (Moon-van der Staay *et al.*, 2000; Viprey *et al.*, 2008; Shi *et al.*, 2009). Prasinophytes clade VII are naked coccoid cells ranging in size from 2 to 5 µm with no specific morphological feature and remain without formal description, despite the fact that they have been in culture since 1965 (Potter *et al.*, 1997). Previous phylogenetic analysis of the 18S rRNA gene divided this group into three well-supported lineage, A, B and C (Guillou *et al.*, 2004), the latter formed by *Picocystis salinarum* only, a small species found in saline lakes (Lewin *et al.*, 2000; Roesler *et al.*, 2002; Krienitz *et al.*, 2012). Strains from lineages A and B harbor very similar set of photosynthetic pigments in contrast with those from lineage C (Lewin *et al.*, 2000; Lopes dos Santos *et al.*, 2016).

The first molecular data of prasinophytes clade VII were recovered from the environment by Moon-van der Staay *et al.* (2000). Since then, clade VII has been found in moderately oligotrophic areas from the Pacific Ocean, Mediterranean Sea (Viprey *et al.*, 2008; Shi *et al.*, 2009; Rii *et al.*, 2016) as well as in coastal waters (Romari and Vaulot, 2004; Wu *et al.*, 2015). Recently, Rii *et al.* (2016) confirmed through 18S rRNA clone libraries that prasinophytes clade VII are the dominant green algae in the mildly oligotrophic waters of the South East Pacific Ocean. Rates of ^{14}C-primary productivity estimated from radiolabeled cells sorted by flow cytometry revealed that the productivity of photosynthetic picoeukaryotes (< 3 µm) was nearly (44%) equivalent that of *Prochlorococcus* plus *Synechococcus* (56%), suggesting a key role for clade VII in these waters (Rii *et al.*, 2016)

In this study, we analyze the genetic diversity of clade VII based on a large number of strains hosted by the Roscoff Culture Collection (RCC) and National Institute of Environmental Sciences Collection (NIES) and determine the oceanic distribution of this group using V9 rRNA metabarcoding data sets obtained in the frame of the *Tara* Oceans expedition and Ocean Sampling Day (OSD) consortium.

We demonstrate that prasinophytes clade VII is a highly diversified group and one of the key component of phytoplankton in open ocean waters.

Materials and methods

Origin of the strains and culture conditions

Prasinophytes clade VII strains listed in Table 1 were used for genetic diversity analysis. These strains were selected from the RCC (http://www.roscoff-culture-collection.org) and from the NIES collection (National Institute for Environmental Studies, http://mcc.nies.go.jp). Strains were grown at 22 °C in K seawater medium (Keller *et al.*, 1987) under average 100 µmoles photons m^{-2} s^{-1} and 12:12-h light:dark regime.

Genomic DNA extraction and PCR amplification

Cells were harvested in exponential growth phase and concentrated by centrifugation. Total nucleic acids were extracted using the Nucleospin Plant II kit (Macherey-Nagel, Düren, Germany). The nearly full-length nuclear 18S rRNA and partial plastid 16S rRNA gene were PCR amplified using the eukaryotic primers Euk63f and Euk1818r (Lepère *et al.*, 2011), and PLA491f (Fuller *et al.*, 2006a, b) and OXY1313r (West *et al.*, 2001), respectively (Supplementary Table 1). Sequences were deposited to GenBank under accession numbers: KU843559–KU843599.

Environmental sequences

A reference data set of near complete sequences of gene coding for nuclear 18S rRNA and partial plastid 16S rRNA gene were compiled using sequences available for prasinophytes clade VII strains and the PhytoRef database for plastid 16S rRNA (Decelle *et al.*, 2015b). This data set was used to identify similar environmental sequences in GenBank (Release 208.0; June 2015) using BLAST (blastn algorithm, e-value e^{-10}). Only sequences longer than 500 and 400 bp were retrieved for plastid 18S and 16S rRNA, respectively (Supplementary Table 2). The nuclear 18S and plastid 16S rRNA environmental sequences were checked for chimeras using uchime as implemented in Mothur v.1.36.0 (Schloss *et al.*, 2009) and assigned manually based on a reference alignment as described below.

Phylogenetic analysis and sequence assignments

All culture and environmental sequences were aligned with MAFFT using the E-INS-i algorithm (Katoh *et al.*, 2002). Only 18S rRNA sequences longer than 1500 bp and environmental plastid 16S rRNA that completely covered the fragment obtained (approximately 850 bp) with the primers PLA491f/OXY1313r were considered in order to build a reference alignment for the phylogenetic analysis. The remaining shorter sequences were placed in the

Table 1 Origins and culture conditions of prasinophytes clade VII strains used in this study

Strain ID	Strain name	Other name	Clade	Ocean origin	Latitude	Depth of isolation (m)	18S rRNA	16S rRNA
RCC15	CCMP1205		A2	NA	NA	NA	U40921	AY702121, FN563080
RCC19	OLI 26 FG-B		A5	Pacific Ocean	−7.00	60	KT860855	**KU843563**
RCC138	CCMP1606		A2	Pacific Ocean	22.75	NA	KT860872	LN735241
RCC227	OLI 26 FB-A		A5	Pacific Ocean	−7.00	60	KT860875	**KU843564**
RCC287	NOUM15	NOUM97015	A3	Pacific Ocean	0.00	120	AY425302	AY702147
RCC297	Açores 3		A3	Atlantic Ocean	38.83	0	KT860659	LN735413
RCC696	IndianOcean_1–1		B2	Indian Ocean	−22.08	120	**KU843577**	LN735439
RCC700	IndianOcean_6–3		A5	Indian Ocean	−14.48	70	**KU843578**	LN735441
RCC701	IndianOcean_8–1		A5	Indian Ocean	−14.48	70	KF899839	**KU843565**
RCC712	IndianOcean_36–1		A1	Indian Ocean	−22.08	120	**KU843579**	LN735451
RCC713	IndianOcean_37–2		A1	Indian Ocean	−22.08	120	**KU843580**	LN735452
RCC717	IndianOcean_43–5		A2	Indian Ocean	−14.48	0	**KU843581**	LN735453
RCC719	IndianOcean_45–8		A1	Indian Ocean	−12.22	76	**KU843582**	LN735454
RCC722	IndianOcean_47–2		A4	Indian Ocean	−12.22	6	**KU843583**	**KU843566**
RCC726	IndianOcean_49–8		A4	Indian Ocean	−14.48	0	**KU843584**	LN735455
RCC856	Biosope_42 A2		A5	Pacific Ocean	−8.33	10	KF422631	LN735470
RCC857	Biosope_40 A2		A3	Pacific Ocean	−8.33	10	**KU843585**	LN735471
RCC917	Biosope_182_FL1–3		A4	Pacific Ocean	−33.35	5	FJ997211	LN735488
RCC996	Biosope_46 B4S		A	Pacific Ocean	−9.07	100	**KU843586**	LN735514
RCC997	Biosope_46 B5S	NIES-2675	A1	Pacific Ocean	−9.07	100	KT860935	LN735515
RCC998	Biosope_46 C3S	NIES-2676	A1	Pacific Ocean	−9.07	100	KF422632	LN735516
RCC999	Biosope_46 C4S		B2	Pacific Ocean	−9.07	100	**KU843587**	LN735517
RCC1019	Biosope_45 A2 478		A3	Pacific Ocean	−9.07	100	**KU843588**	LN735204
RCC1021	Biosope_46 B6		A5	Pacific Ocean	−9.07	100	**KU843589**	LN735205
RCC1032	Biosope_46 B7		A3	Pacific Ocean	−9.07	100	**KU843590**	LN735207
RCC1043	Biosope_47 B1		A3	Pacific Ocean	−9.07	30	**KU843591**	LN735210
RCC1124	PAP_AD		A4	Atlantic Ocean	48.83	10	**KU843592**	LN735219
RCC1871	RA090205–09		A4	English Channel	48.75	0	KF899840	LN735295
RCC2335	JST MH 317	NIES-2755	A4	Pacific Ocean	35.22	0	KF422627	LN735348
RCC2337	JST MH 335	NIES-2756	B2	Pacific Ocean	33.77	0	**KU843593**	LN735349
RCC2339	JST MH 340	NIES-2758	B	Pacific Ocean	33.77	0	KF422628	LN735350
RCC3368	CCMP2111		A	Indian Ocean	−5.47	NA	KF899838[a]	LN735423
RCC3373	CREST MH 504	NIES-3667	A4	Pacific Ocean	37.98	0	**KU843594**	**KU843566**
RCC3374	CCMP2152		A7	Pacific Ocean	22.75	NA	**KU843595**	**KU843568**
RCC3375	CCMP2175		A5	Pacific Ocean	22.75	NA	KF899843	**KU843569**
RCC3376	CCMP2113		A	Pacific Ocean	8.93	85	**KU843596**	**KU843570**
RCC4429	AMT 2013 - P180-A5		A4	Atlantic Ocean	44.11	2	**KU843597**	**KU843571**
RCC4430	AMT 2013 - P181-A1		A4	Atlantic Ocean	44.11	2	**KU843598**	**KU843572**
RCC4434	AMT 2013 - P182-H9		A6	Atlantic Ocean	17.06	2	**KU843599**	
RCC3402	*Picocystis salinarum* CCAP 158/2	CCMP1897	C	San Francisco salt pond	37.78	NA	FR865649	AB491631
NIES-3668	CREST MH 537		A4	Pacific Ocean	38.03	0	**KU843573**	**KU843559**
NIES-3670	CREST MH 533		A4	Pacific Ocean	42.16	0	**KU843575**	**KU843561**
NIES-3671	CREST MH 514		A3	Pacific Ocean	38.01	0	**KU843576**	**KU843562**
NIES-3669	CREST MH 509	RCC4656	B1	Pacific Ocean	42.16	0	**KU843574**	**KU843560**

Abbreviations: NA, not available; NIES, National Institute of Environmental Sciences Collection; RCC: Roscoff Culture Collection (www.roscoff-culture-collection.org); rRNA, ribosomal RNA.
[a]18S rRNA sequence from GenBank is a chimera between clade VII and *Micromonas*.
Sequences in bold were obtained during this study.

reference alignment using the online tool MAFFT with the –addfragments option (Katoh and Frith, 2012) and manually assigned.

Phylogenetic reconstructions were performed with three different methods: maximum likelihood (ML), distance (neighbor joining; NJ) and Bayesian analyses. The K2+G+I model was selected for the 18S rRNA and 16S rRNA sequence data based on the substitution model selected through the Akaike information criterion and the Bayesian information criterion options implemented in MEGA 6.06 (Tamura *et al.*, 2013). NJ and ML analysis was performed using MEGA 6.06 and PhyML 3.0 (Guindon *et al.*, 2010) with SPR (Subtree Pruning and Regrafting) tree topology search operations and approximate likelihood ratio test with Shimodaira–Hasegawa-like procedure, respectively. Markov chain Monte Carlo iterations were conducted for 1 000 000 generations sampling every 100 generations with a burning length 100 000 using MrBayes 3.2.2 (Ronquist and Huelsenbeck, 2003). MAFFT and MrBayes programs were run within Geneious 7.1.7 (Kearse *et al.*, 2012). For the definition of environmental clades, we adapted the criteria of Guillou *et al.* (2008): a clade must contain at least two environmental or strain sequences obtained from

4

different locations and/or samples at a given location and must have strong phylogenetic support for at least one of the gene markers used and specific V9 signatures in nuclear 18S gene (see details below). Clade nodes were considered as well supported when bootstrap, SH-like support values and Bayesian posterior probabilities were higher than 70, 0.7 and 0.95, respectively. The same criteria was used to represent the sequences on the phylogenetic trees.

18S rRNA metabarcodes
V9 rRNA metabarcoding data sets obtained in the frame of the *Tara* Oceans expedition and the OSD consortium were used to analyze the global environmental distribution of prasinophytes clade VII. The *Tara* Oceans samples were collected in oceanic waters over the course of a 2.5 year expedition (Pesant *et al.*, 2015) and OSD samples were collected during a simultaneous sampling campaign performed on 21 June 2014 at 191 different sites mostly in coastal areas (Kopf *et al.*, 2015).

For *Tara* Oceans, we used the publicly available eukaryotic metabarcoding data set, consisting of >530 million quality-checked eukaryotic V9 rRNA sequences obtained from 293 samples obtained at 47 stations (mostly concentrated in the Southern hemisphere), two depths (surface and deep chlorophyll maximum (DCM)) and four size fraction (0.8–5, 5–20, 20–180 and 180–2000 μm). V9 rRNA metabarcodes (129 bp on average) were sequenced with Illumina HiSeq (Illumina, San Diego, CA, USA) sequencing at a typical depth of 2 M sequences per sample (Supplementary Table 3). Metabarcode sequences from each sample set were clustered into OTUs, differing by one base at most, using the SWARM software (Mahé *et al.*, 2014). Detailed information on sampling, primers and sequence data processing are provided in Pesant *et al.* (2015), Amaral-Zettler *et al.* (2009) and de Vargas *et al.* (2015). We only considered OTUs represented by at least 10 sequences. OTUs were assigned to a taxonomic group using a BLAST-like strategy (see details at http://taraoceans.sb-roscoff.fr/EukDiv/#content2) against an annotated eukaryotic V9 database derived from the PR2 database (Guillou *et al.*, 2013), which provides eight levels of taxonomic hierarchy (Kingdom, Super-division, Division, Class, Order, Family, Genus and Species) (for detailed information see de Vargas *et al.*, 2015). Only OTUs that presented >80% similarity to at least one of the reference sequences were kept. Fifty-seven OTUs were classified as prasinophytes-and-relatives (Supplementary Table 4). Sequences representative of each OTU were aligned to reference sequences (see above) and assigned to a specific clade of prasinophytes clade VII (Supplementary Table 4 and Supplementary Figure 2).

Only OSD V9 rRNA sequences were considered in this study for comparison purpose. The metabarcode data set, sequenced with Illumina MiSeq sequencing by Life Watch (Italy), consisted of 11 million quality-checked eukaryotic V9 rRNA sequences obtained from 31 stations (surface) after filtration through 0.8 μm. Most of the OSD stations used in this study were concentrated in the Northern hemisphere. Detailed information on sampling and sequence data preprocessing are provided at: https://github.com/MicroB3-IS/osd-analysis/wiki/Guide-to-OSD-2014-data#data-deposited-in-public-archives-and-available-and-web-sites. Metabarcode sequences from each sample set were clustered into OTUs at 99% identity, using the nearest neighbor method implemented in Mothur (Schloss *et al.*, 2009). OTUs represented by at least 10 sequences were assigned to a taxonomic group using V9 regions extracted from the PR2 database (Guillou *et al.*, 2013) using the 'classify' Mothur command with the wang method (Wang *et al.*, 2007).

We compared the contribution of prasinophytes clade VII of Mamiellophyceae metabarcodes to the total number of 'photosynthetic' metabarcodes. Groups considered as photosynthetic included Chlorophyta, Rhodophyta, Cryptophyta, Haptophyta and Ochrophyta (that is, photosynthetic Stramenopiles such as diatoms) but not dinoflagellates for which the differentiation between photosynthetic and heterotrophic species is highly complex.

Analyses and plots were performed with the R software using the following libraries: ggplot2, mapproj, maps, mapdata, reshape2, scales, gridExtra (R Development Core Team, 2013).

The *Tara* Oceans and OSD data set are accessible at http://doi.pangaea.de/10.1594/PANGAEA.843017 and 843022 and https://owncloud.mpi-bremen.de/index.php/s/RDB4Jo0PAayg3qx?path=/2014/silva-ngs/18s/lifewatch/v9, respectively.

Results

Genetic diversity
The phylogenetic tree based on nearly full nuclear 18S rRNA sequences (>1500 bp) obtained from strains and retrieved from GenBank (Figure 1 and Supplementary Table 2) confirmed the existence of the three major lineages, A, B and C, defined by Guillou *et al.* (2004). However, each of these lineage can be further divided into finer clades. Lineage A is subdivided into seven clades, A1–A7, well supported by all three methods (NJ, ML and Bayes) used except for clades A2, A3 and A4. Clade A2 and A4 had no support from NJ (bootstrap 61% and 23%, respectively) and clade A3 neither by NJ (bootstrap 59%) nor Bayes analysis (0.56). A2 and A3 clades had identical sequences in the V4 region of the 18S rRNA gene and A4 was only distinguished by a unique position around 580 bp. Lineage B contains three clades, B1–B3, the latter composed only by environmental sequences. The three clades are well supported by all three methods used. Lineage C did not contain any subdivision based on the 18S rRNA
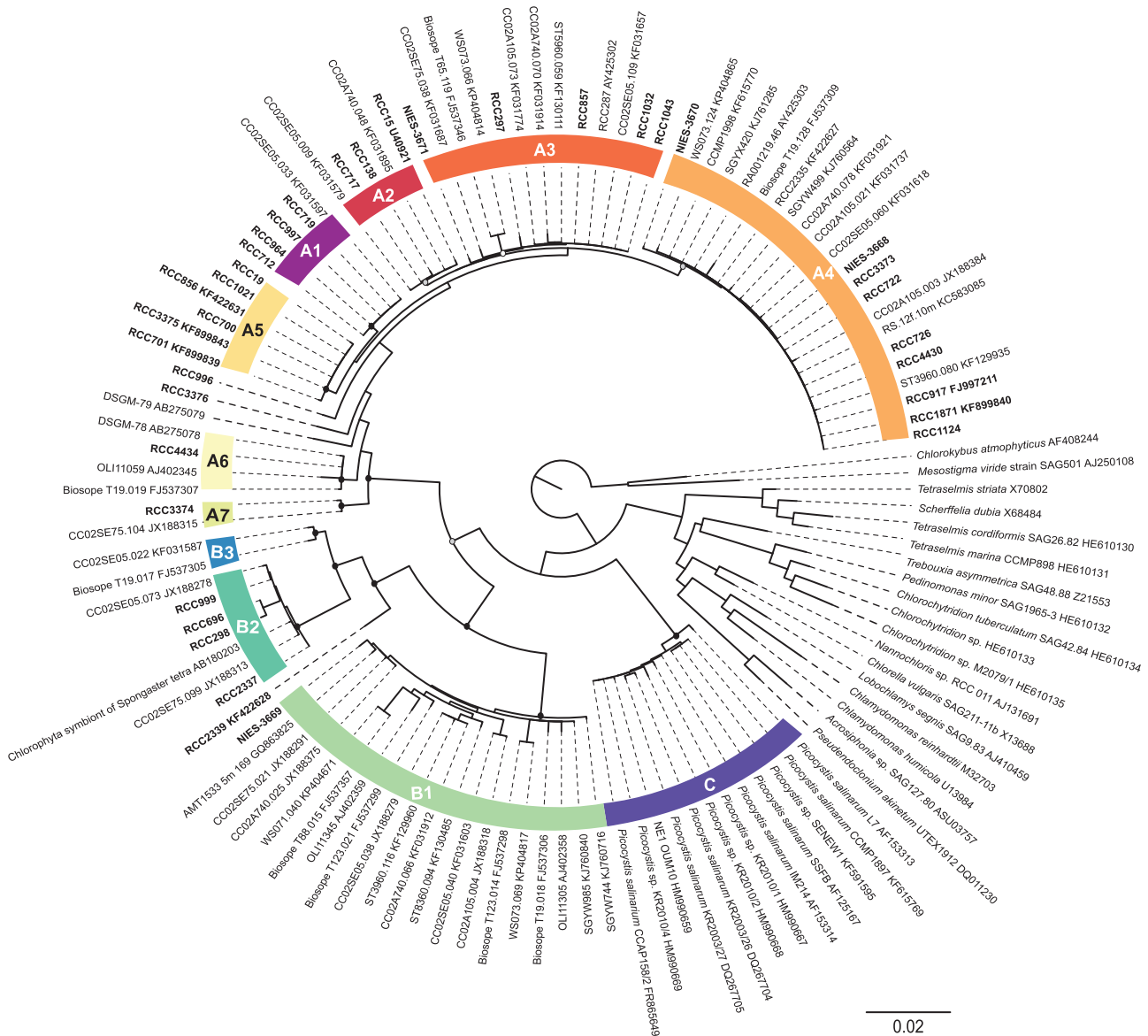
Figure 1 ML tree inferred from nuclear 18S rRNA sequences belonging to prasinophytes clade VII. Nodes supported by NJ, ML and Bayesian methods are shown by a solid dot (●). Nodes supported by ML and Bayes only are indicated by a gray dot (◐). Empty dot (○) represents a node supported by only one method, either ML or Bayes.

gene and is only composed of *Picocystis salinarum*, which is the only described species within prasinophytes clade VII so far. The average sequence identity within a clade calculated with the nuclear 18S rRNA data set ranged from 98.19% to 99.96%. Clades B1 and A6 showed lowest average sequence identity, 98.19 and 99.12, respectively (Supplementary Table 5).

Almost all clades within lineages A and B were well supported by the phylogenetic analysis of plastid 16S rRNA sequences longer than 850 bp by the three methods used, including A2, A3 and A4 (Figure 2). The only exceptions were B3 that was only represented by environmental nuclear 18S rRNA sequences and for which no equivalent was found with the plastid 16S rRNA gene, A7

represented by a single strain RCC3374 and A6 (RCC4434) for which we have no 16S rRNA sequence (Figure 2). The branching of RCC3374 and RCC3368 (for which we have no 18S sequence, see Table 2) was not supported by any of the methods used. Also, we observed that three plastid 16S rRNA clones previously assigned as environmental clade 16S-VIII by Lepère *et al.* (2009) corresponded to clades VII B1 and B2 as hypothesized by Shi *et al.* (2011). Plastid 16S rRNA phylogeny indicates that the *Picocystis* lineage is not a sister lineage of the two others lineages (A and B) of prasinophytes clade VII (Figure 2).

Strains RCC996, RCC3376 and RCC2339 as well as environmental sequence DSGM-79 (AB275079) did not belong to any clade in the nuclear 18S rRNA

Figure 2 ML tree inferred from nuclear 16S rRNA sequences belonging to prasinophytes clade VII. Nodes supported by NJ, ML and Bayesian methods are shown by a solid dot (●).

Table 2 Prasinophyte clade VII and Mamiellophyceae sequences as a function of size fraction and depth in the *Tara* Oceans and OSD metabarcode data sets

Data set	Fraction (μm)	Depth	# of samples	Chlorophyta/ photo (%)		% of samples with clade VII sequences	Clade VII/ photo (%)		Clade VII/ chloro (%)		Mamiellophyceae/ photo (%)		Mamiellophyceae/ chloro (%)	
				Mean	Max		Mean	Max	Mean	Max	Mean	Max	Mean	Max
Tara	0.8–5	SUR	40	15.7	54.5	100	10.8	53.1	43.4	97.5	2.7	19.2	24.8	93.7
Tara		DCM	33	20.4	58.6	100	11.5	58.2	45.4	99.3	7.2	34.1	36.6	92.1
Tara	5–20	SUR	41	9.5	80.3	95	6.6	80.2	54.7	99.9	0.5	4.5	13.4	90.0
Tara		DCM	33	10.1	91.6	97	5.4	24.8	61.8	98.9	0.5	3.2	13.7	100.0
Tara	20–180	SUR	42	7.9	39.9	98	5.8	29.0	55.9	98.4	0.2	3.0	6.3	63.4
Tara		DCM	28	9.0	48.7	100	5.6	23.0	61.6	98.9	0.8	19.6	9.1	63.3
Tara	180–2000	SUR	45	15.6	56.1	96	10.0	51.1	51.3	100.0	2.1	21.7	18.3	87.0
Tara		DCM	31	14.4	61.2	97	10.2	30.9	57.3	100.0	0.9	7.0	11.7	57.9
Tara	All	All	293	12.8	91.6	98	8.3	80.2	53.5	100.0	1.8	34.1	16.7	100.0
OSD	0.8–200	SUR	31	19.6	56.2	48	0.2	2.2	1.4	21.2	10.7	45.7	49.7	94.7

Abbreviations: DCM, deep chlorophyll maximum; OSD, ocean sampling day; SUR, surface.
Number of samples analyzed, percentage of samples with prasinophyte clade VII, mean and maximum contribution of prasinophyte clade VII and Mamiellophyceae to photosynthetic taxa and to Chlorophyta.

analysis. Strains RCC996 and RCC3376, as well as clone DSGM-79 belong to lineage A, whereas RCC2339 belongs to lineage B. These strains have unique signatures in both nuclear 18S rRNA and plastid 16S rRNA. Partial nuclear environmental sequences retrieved from GenBank as well as V9 metabarcodes were similar to these unique sequences (see below) demonstrating that these sequences are not chimera. However, as no other long sequences are available in public databases, we did not create clades for these sequences. It is likely that these unique branches will be attached to clades as more sequences and strains become available.

Prasinophytes clade VII in public databases
A total of 267 nuclear 18S and 109 plastid 16S rRNA clade VII sequences longer than 500 and 400 bp, respectively, were retrieved from GenBank (Figure 1 and Supplementary Table 2). These sequences were mostly from environmental clone libraries although 12 were from cultures, 11 of which were assigned to lineage C. The other strain sequence recovered from GenBank is labeled 'Prasinophyceae sp. CCMP1998' (KF615770), which had been wrongly attributed to the genus *Pycnococcus* (Duanmu *et al.*, 2014) but clearly belongs to prasinophyte clade VII.

Sequences from lineages A (43%) and B (48%) were all obtained from oceanic waters, including Atlantic, Indian and Pacific Oceans and China, Arabian, Red, Caribbean and Mediterranean Seas in contrast to sequences from lineage C (8%), which were not of oceanic origin. Geographical information about sampling origin was available for 85% of the sequences assigned to A and B. They were found between ~60° N and ~34° S with no sequence in cold temperate or polar regions (Supplementary Table 2).

Among the nuclear 18S nuclear sequences, clades A4 and B1 were the most abundant (Supplementary Figure 3A). About 7% of the sequences were only assigned at the lineage level because the aligned region did not have signatures strong enough to identify them at the clade level. For example, most of the 18S rRNA sequence not assigned corresponded to the V4 region for which clades A2 and A3 are identical. Also most plastid 16S rRNA sequences from GenBank were obtained using bacteria 16S rRNA primers: despite being long sequences, the alignment of these sequences with the fragments generated by our primers (PLA491F/OXY1313R) was short. About 2% of 18S rRNA sequences were assigned to the solitary branches RCC996, RCC2339 and clone DSGM-79 (Supplementary Figure 4).

Oceanic distribution
In order to access the oceanic distribution of prasinophytes clade VII, we searched for 18S rRNA signatures of this group in V9 rRNA metabarcoding data sets obtained in the frame of the *Tara* Oceans expedition (de Vargas *et al.*, 2015) and OSD (Kopf *et al.*, 2015). The *Tara* Oceans stations were distributed over the major oceans with a bias towards the Southern hemisphere, whereas OSD stations were mostly coastal and in the Northern hemisphere (Supplementary Figure 1). OSD sampling was synchronized on the northern summer solstice while *Tara* Oceans expedition extended over all four seasons. The sampling strategy of *Tara* Oceans included four different size fractions (0.8–5, 5–20, 20–180 and 180–2000 μm) in contrast to only one size fraction for OSD (0.8–200 μm). Also two water column depths corresponding to subsurface and the DCM were sampled in *Tara* Oceans, whereas OSD samples were only recovered from surface. In this work, we only considered the 31 OSD and 47 *Tara* Oceans stations for which V9 18S rRNA data sets are publicly available.

The contribution of Chlorophyta to *Tara* Oceans and OSD metabarcodes affiliated to photosynthetic groups (excluding dinoflagellates) was 12.8% and 19.6%, respectively (Table 2). For the *Tara* Oceans metabarcodes, Chlorophyta contribution was highest, on average, for the smallest size fraction and higher at the DCM compared with the surface (Table 2).

In coastal waters, Mamiellophyceae was the dominant Chlorophyta group. In the OSD data set, they contributed on average 10.7% to the total photosynthetic sequences (49.7% of the Chlorophyta) against 0.2% (1.4% of Chlorophyta) for prasinophytes clade VII (Figure 3a and Table 2). Mamiellophyceae were dominant at most coastal station from OSD but also at some *Tara* Oceans stations, including high latitudes stations both in the northern and southern hemispheres, and off South Africa in the Benguela current (Figure 4).

In contrast, clade VII appeared as the dominant group in oceanic waters. In the *Tara* data set, their average contribution to the total photosynthetic sequences was about 8%, while that of Mamiellophyceae was only 1.8% (Figure 3a and Table 2). The overall contribution of prasinophytes clade VII to total *Tara* Oceans photosynthetic sequences varied between 5% and 11% on average for the different depths and size fractions but could reach up to 80% in specific samples (Table 2). The dominance of clade VII within Chlorophyta held for all size classes and for both surface and DCM samples (Figure 3b). At most *Tara* oceanic stations (Figure 4), the 0.8–5 μm fraction was dominated by prasinophytes clade VII but other Chlorophyta groups could prevail at specific locations, for example, at several Mediterranean Sea stations where Pyramimonadales were the major group. Their highest contribution was in the pico-eukaryote fraction (0.8–5 μm) and unexpectedly in the mesoplankton fraction (180–200 μm). In fact, in surface, for these two size fraction, the overall contribution of prasinophytes clade VII to photosynthetic sequences was very similar with an average of 10% and a maximum over 50% (Table 2). Clade VII contribution to the total photosynthetic pico-eukaryote size fraction was highest in equatorial
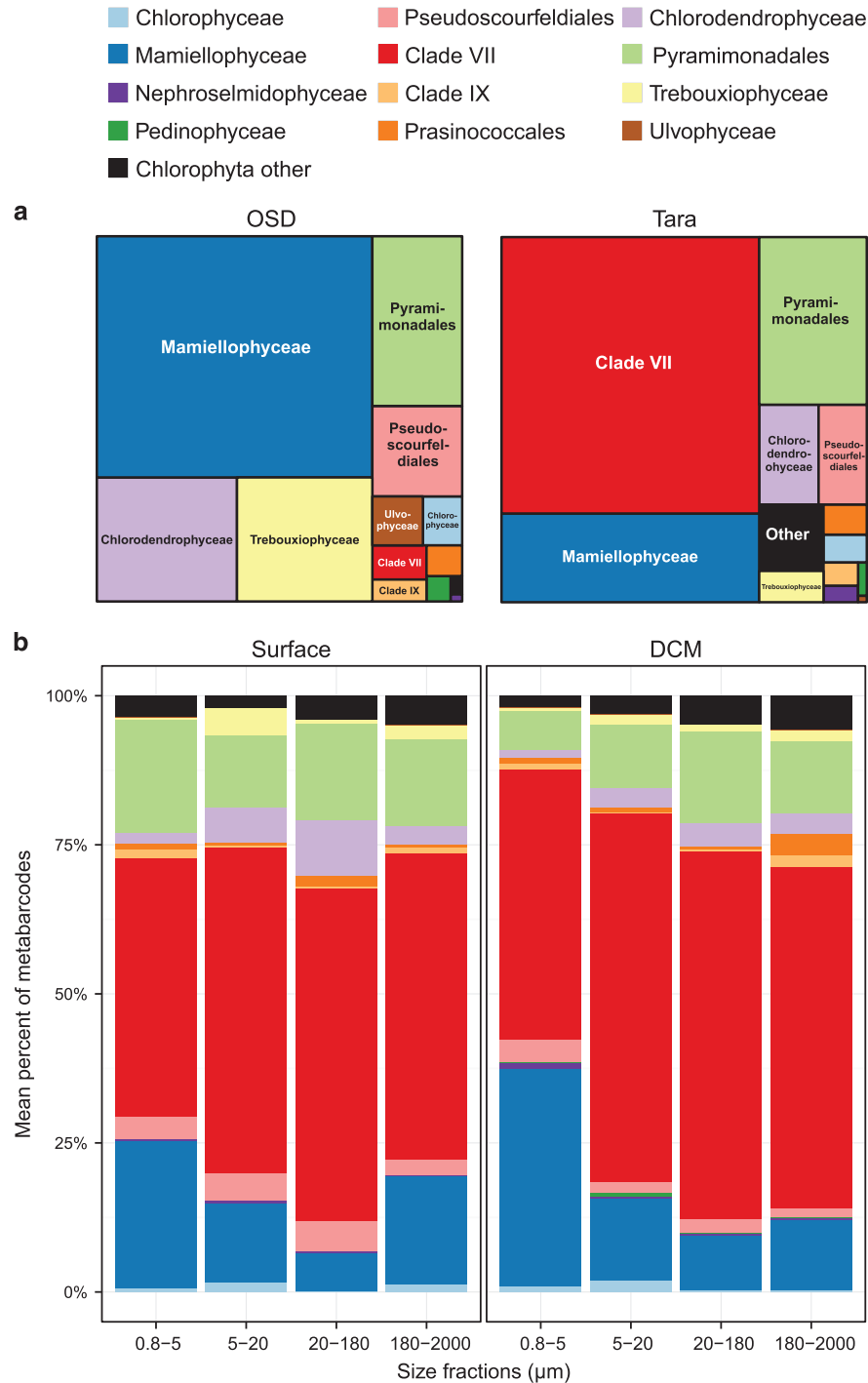
Figure 3 (**a**) Average contribution of the different groups of Chlorophyta in *Tara* Oceans (average of the three size fractions 0.8–5, 5–20 and 20–180 μm) and in OSD (0.8–200 μm fraction). (**b**) Average contribution of the different groups of Chlorophyta in the Tara Oceans data set as a function of size fraction in surface (left) and at the DCM (right).

and subtropical waters (Figures 5a and b, bottom graphs).

Fifty-seven *Tara* Oceans and eight OSD OTUs were classified as prasinophytes clade VII (Supplementary Table 4). *Tara* Oceans OTUs represented a total of over 2 million sequences and clade VII was present at all stations. In contrast, OSD had just over 1000 sequences of clade VII, which was only present at 16

out of 32 stations. The eight most abundant clade VII *Tara* Oceans OTUs regrouped > 99% of the sequences. By careful alignment of the sequence representative of each OTU with references sequences from cultures or the environment, we were able to assign the major OTUs to the clades defined above (Supplementary Figure 2 and Supplementary Table 4). Interestingly, no sequence
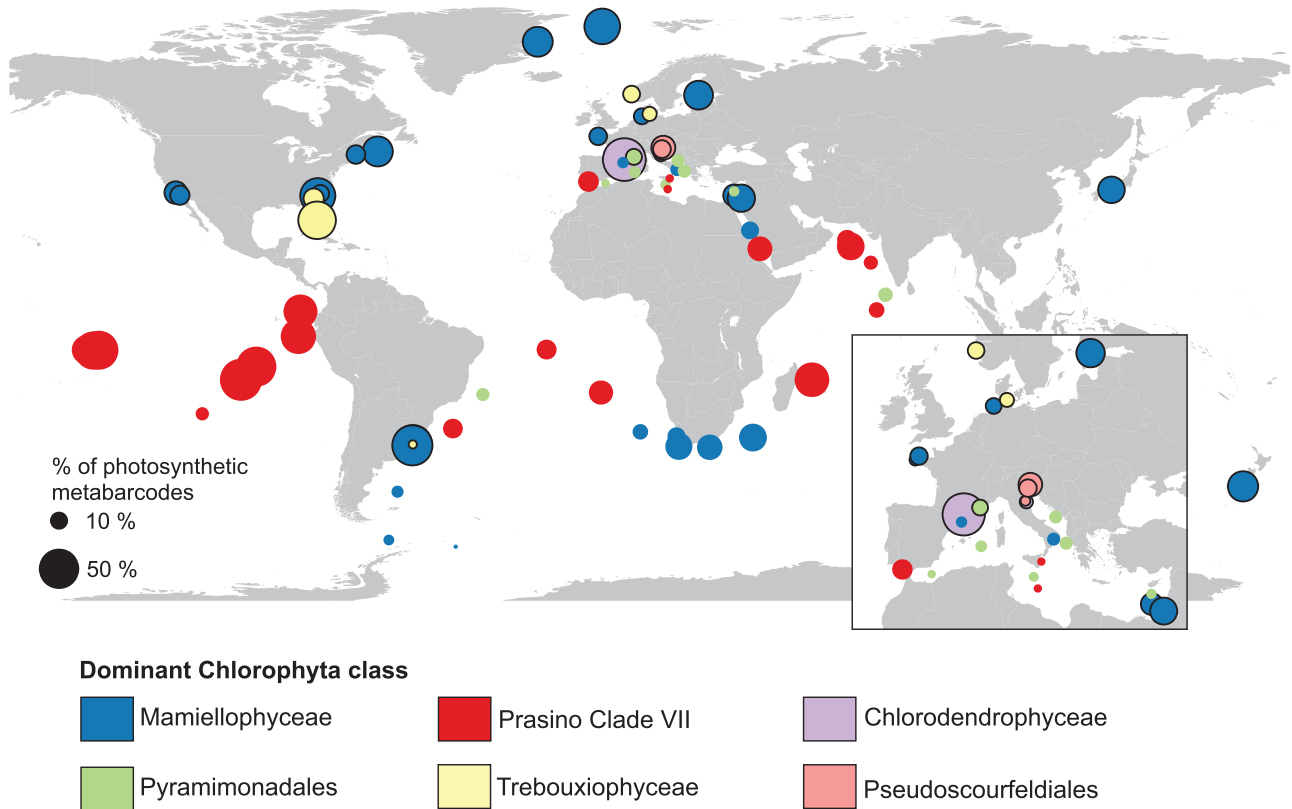
Figure 4 Map of the contribution of Chlorophyta to photosynthetic sequences (size of circles) and dominant Chlorophyta group (color of circles) for *Tara* Oceans (0.8–5 μm) and OSD samples in surface samples. Circles with black border correspond to OSD stations.

from either *Tara* Oceans or OSD could be assigned to the C clade.

Overall, the most abundant clades were B1, A4, A6 and A3, respectively, in the *Tara* data set (Table 3 and Supplementary Figure 3A). Remarkably, these clades were also the most abundant in OSD (A4, A6, B1 and A3) and in GenBank 18S (A4, B1, A3 and A6) (Supplementary Figure 3A). Two of these clades (B1 and A6) have only a single representative in culture, whereas A4 and A3 have strains from the Pacific, Atlantic and Indian Oceans (Table 1). The relative contribution of the different clades did not seem to vary much with the size fraction nor with the depth level (DCM vs surface) in the *Tara* Oceans samples (Supplementary Figure 3B).

The clade composition varied widely between the different regions sampled (Figure 5). Mediterranean Sea surface waters were dominated by clade A6, which was also present at depth in the Eastern, more oligotrophic basin (stations 22–30, Figure 5). Despite the low contribution of clade VII to the total photosynthetic pico-eukaryote fraction in the Mediterranean Sea, A6 notably contributed to about 10% of the total photosynthetic sequences at the lowest DCM recorded (Figure 6a).

The North Indian and tropical Pacific Oceans were dominated by clade B1. This clade was always present when clade VII significantly contributed to the total photosynthetic pico-eukaryote fraction, either at the surface in the Pacific Ocean or in a more pronounced way at the DCM in the Indian Ocean (Figure 5). In the Pacific Ocean, B1 was complemented by A1 both in surface and at depth, A1 being relatively more important at the DCM, which fits with the fact that all A1 strains have been isolated from deep euphotic waters (Figure 5 and Table 1). Deep waters seem to be the preferential niche for A1 where their contribution to the total photosynthetic was higher than in surface (Figure 6a). Clade A4 also had a large contribution to total photosynthetic sequences at the bottom of the euphotic zone (Figure 6a), although all A4 strains have been isolated from surface waters (Table 1). In the Indian Ocean, B2, A4 and A3 seemed to complement B1 and in the Atlantic Ocean B1, A3 and A4 are the most important clades (Figure 5).

Prasinophytes clade VII seem to have a preference for warm waters being mostly found between 20 °C and 30 °C (Figure 6b). Among the different clades, A1 seems to have a narrower temperature range and A4 two distinct temperature niches around 20 °C and 27 °C, where they had the larger contribution to the total photosynthetic sequences (Figure 6b). In contrast, B1 seems to have a wide temperature range (Figure 6b).

Discussion

Our phylogenetic analyses based on nuclear 18S rRNA sequences recovered the three main lineages

Diversity and distribution of oceanic prasinophytes
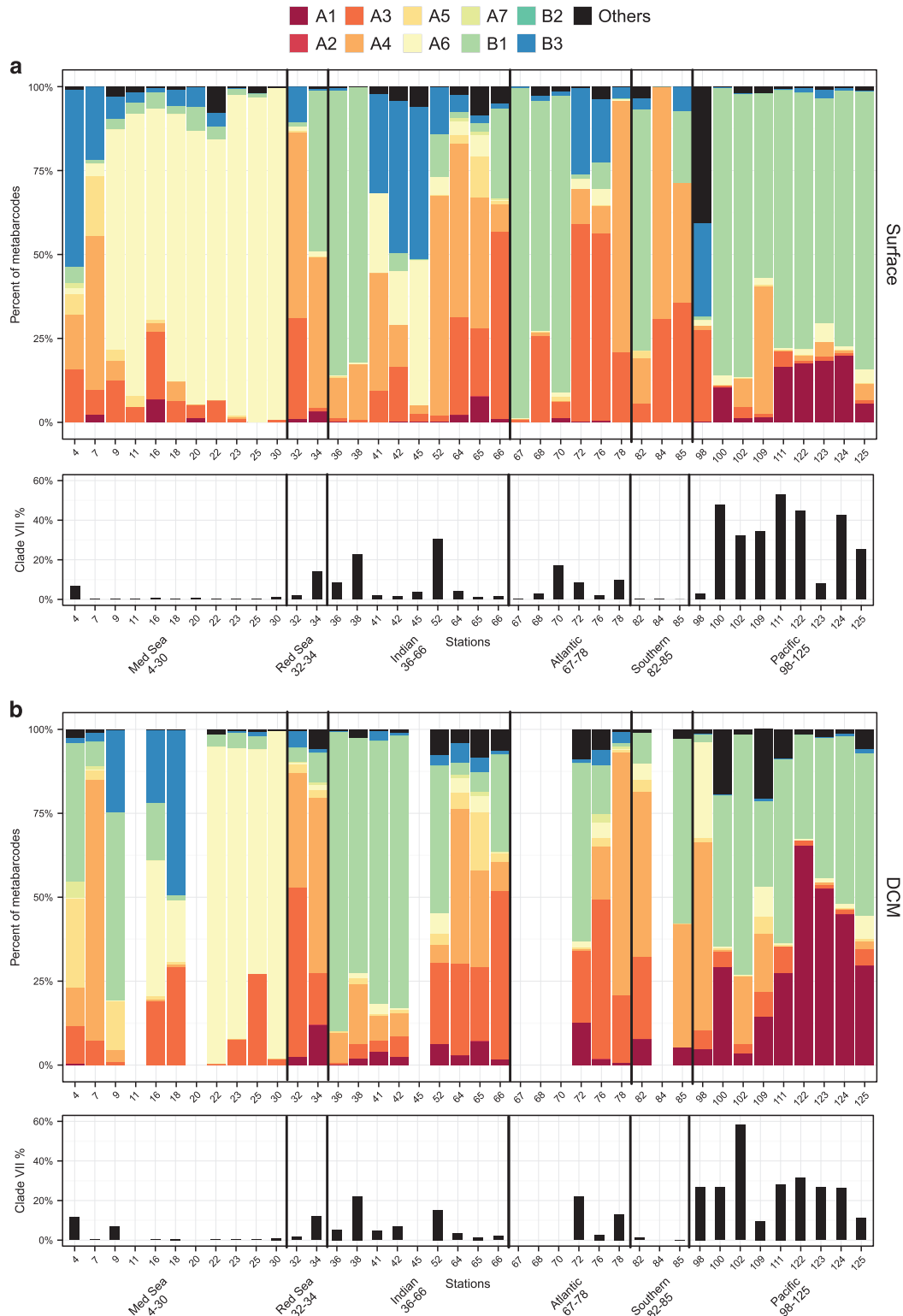A Lopes dos Santos *et al*

Figure 5 Top graph: contribution of the different clades of prasinophytes clade VII to the smallest size fraction (0.8–5 µm) at each *Tara* stations. Bottom graph: relative contribution of clade VII sequences to the total photosynthetic sequences. (**a**) Surface. (**b**) DCM.

(A, B, and C) of prasinophytes from clade VII as proposed by Guillou *et al.* (2004). A and B are clearly marine lineages with sequences and strains recovered only from marine environments. On the other hand, lineage C (encompassing the only formally

described species *Picocystis salinarum)* has been found in inland saline lakes (Hollibaugh *et al.*, 2001; Roesler *et al.*, 2002; Krienitz *et al.*, 2012) or ponds (Lewin *et al.*, 2000) where it was originally isolated. Thus, *Picocystis* seems to be a prasinophyte typical

Table 3 Contribution of the different subclades of prasinophytes

Group	Tara Oceans—293 samples			OSD—31 samples		
	Mean % of clade VII	Max % of clade VII	# of samples where present	Mean % of clade VII	Max % of clade VII	# of samples where present
Clade VII A1	8.5	70.0	184	—	—	—
Clade VII A2	0.02	2.8	19	—	—	—
Clade VII A3	14.6	100.0	247	12.4	40.9	8
Clade VII A4	21.0	100.0	240	28.0	100.0	9
Clade VII A5	2.4	90.1	169	—	—	—
Clade VII A6	19.0	100.0	235	19.6	80.4	4
Clade VII A7	0.2	5.8	49	—	—	—
Clade VII B1	25.2	100.0	245	17.7	97.1	5
Clade VII B2	—	—	—	—	—	—
Clade VII B3	4.5	56.5	192	2.3	22.2	4
Clade VII others	4.4	92.5	199	20.0	100.0	3

Abbreviation: OSD, ocean sampling day.
Number of samples analyzed, mean and maximum contribution, number of samples where present.

of inland saline waters rather than marine with a very restricted ecological range of habitats, as evidenced by the absence of similar sequences in the *Tara* and OSD data sets. Also, all our environmental sequences recovered and assigned to *Picocystis* were from hypersaline lakes in USA, China and Kenya with one exception. The environmental sequence Ola1.E10. invm13r (AB691200) was recovered from Terrebone Bay, a complex collection of small bayous of the Mississippi River Delta for which an increase of salinity has been reported (Steyer *et al.*, 2008).

In contrast to 18S analysis, plastid 16S rRNA phylogeny suggests that *Picocystis salinarum* forms a lineage separate from prasinophytes clade VII A and B. Several evidences have accumulated supporting this hypothesis, including phylogenetic analyses using the complete nuclear (Marin, 2012) and plastid encoded rRNA operons (Marin and Melkonian, 2010; Marin, 2012) and chloroplast genomes (Lemieux *et al.*, 2014). The chloroplast genomes of *Picocystis salinarum* and one strain of clade VII (RCC15 or CCMP1205, clade A2) differ in size (clade VII harbor the smallest chloroplast genome yet reported for a photosynthetic green algae), structure (clade VII chloroplast contains 100 genes unequally distributed between the two DNA strands) and gene content (Lemieux *et al.*, 2014). Very recently, pigment composition analysis confirmed that *Picocystis* genus harbor monadoxanthin and diatoxanthin as accessory pigments (Lopes dos Santos *et al.*, 2016). These pigments are usually found in cryptophytes or diatoms (Takaichi, 2011), algae that belong to the Red lineage, but have also been reported in *Coccomyxa*, a green alga belonging to the Chlorophyceae (Crespo *et al.*, 2009). Therefore, the *Picocystis* lineage should be considered as an independent lineage of prasinophytes as already suggested by Lemieux *et al.* (2014).

Nuclear and plastid SSU rRNA gene failed to resolve confidently the relationships between the different clades (low bootstrap and different tree topologies). For ancient groups of eukaryotes, such as prasinophytes, multi-gene analysis or secondary structure of the internal transcribed spacer 2 from several taxa are usually required to resolve relationships at high taxonomic levels.

A small number of oceanic studies, focusing on photosynthetic picoeukaryotes, have already pointed prasinophytes clade VII as an important oceanic group, especially in the mesotrophic and moderately oligotrophic waters of the Pacific Ocean (Moon-van der Staay *et al.*, 2000; Viprey *et al.*, 2008; Shi *et al.*, 2009, 2011; Rii *et al.*, 2016). Our analysis showed a clear dominance of clade VII over Chlorophyta in oceanic waters where this group contributed on average to 54% of the total Chlorophyta sequences and 8% of the total photosynthetic sequences of the oceanic *Tara* data set. Mamiellophyceae, usually pointed as important factors of phytoplankton communities in coastal waters (Romari and Vaulot, 2004; Collado-Fabbri *et al.*, 2011), was the dominant Chlorophyta group in coastal OSD data set where contributed with 10.7% to the total photosynthetic sequences (49.7% to the Chlorophyta) against 0.2% for clade VII. One could argue that clade VII dominance among 18S metabarcodes is a result of large number of copies of the rRNA operon in their genomes. Although, no genome sequence is available for any member of this group, prasinophytes clade VII cells have a size between 2 and 5 µm in the same range than Mamiellophyceae, and therefore their 18S rRNA gene copy number is probably similar (Zhu *et al.*, 2005).

An unexpected result was the high contribution of prasinophytes clade VII to the mesoplankton fraction (180–2000 µm). In some stations, they were the only green algal group present in this fraction (for example, in the Red Sea or the Indian Ocean). Many studies have reported that small cells can be incorporated into large aggregates allowing them to sink through the water column at a rate faster than their cell size would allow (Jackson, 2001; Richardson and Jackson, 2007). Sequences from
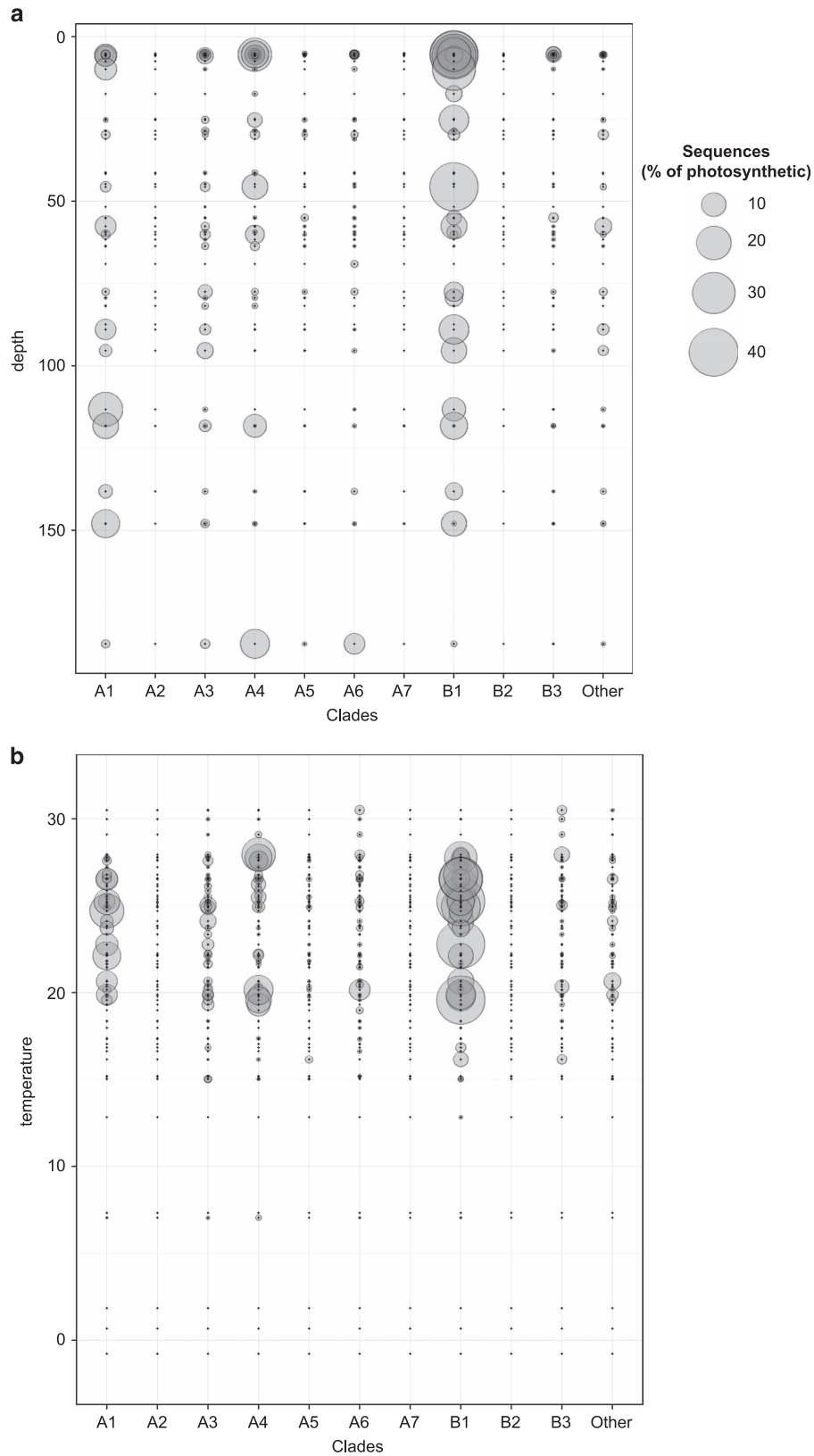
Figure 6 Abundance of the prasinophytes clade VII clades as a function of depth (**a**) and temperature (**b**). The radius of the circles is proportional to the contribution of clade to the total number of photosynthetic sequences.

Micromonas, *Bathycoccus* and *Ostreococcus* have been recovered in traps deployed at 200 and 500 m eastern subtropical North Atlantic (Amacher *et al.*, 2009). Among the clade VII sequences from Gen-Bank, a fraction was from deep sea waters, sediments, traps or marine snow (Supplementary Table 2). Alternatively, symbiotic associations could exist between some members of prasinophyte clade VII and larger plankton organisms. Symbioses are poorly characterized across open ocean plankton community and recent work has shown the importance and diversity of interactions among plankton members, especially the Rhizaria group (Decelle *et al.*, 2015a) for which a wide diversity of symbionts has been described including the green algae *Pedinomonas symbiotica* (Cachon and Caram, 1979). One GenBank clade VII sequence, AB180203, has been indeed associated to a symbiont of the radiolarian *Spongaster tetras*. However, all the strains present in the RCC or NIES collections have been isolated as free living cells.

Tara Oceans, OSD and GenBank data sets pointed clades A3, A4, A6 and B1 as the most abundant among the 10 clades of prasinophytes clade VII. The ecological importance of clades and ecotypes have been extensively address for the marine genera of cyanobacteria *Prochlorococcus* and *Synechococcus* (Scanlan *et al.*, 2009). The success of these cyanobacteria in the modern oceans has been attributed to their wide genetic diversity. For the three marine important genera of pico-phytoplankton, *Micromonas*, *Bathycoccus* and *Ostreococcus*, clades or ecotypes have also been described and ecological hypotheses about the distribution have been made. For example, *Ostreococcus* clade A seems to be typical from surface waters, whereas clade B appears to be associated with deeper layers of the euphotic zone (Rodríguez *et al.*, 2005; Six *et al.*, 2005). However, the physiological and environment parameter influencing clade distribution are more complex than irradiance alone, as showed for *Ostreococcus* (Simmons *et al.*, 2016).

Clade B1 was always associated with the highest contribution of clade VII to the photosynthetic sequences, especially in the Pacific Ocean. Despite being the most represented clade, B1 was the last one for which a strain has been isolated. This strain (NIES-3669) obtained in 2015 from western North Pacific waters off the south coast of Hokkaido, Japan, was difficult to isolate and grows fastidiously compared with others prasinophytes clade VII (MHN, personal observation). In contrast, three strains have been isolated from clade B2, which occurrence is much lower in environmental sequences and seven strains have been obtained for each clade A3 and A5, which are always sporadic. It is noteworthy that abundant groups of marine microbes have defied cultivation efforts. For example, the most abundant marine heterotrophic bacteria, *Pelagibacter*, was only isolated (Rappé *et al.*, 2002) 12 years after being discovered by Giovannoni

et al. (1990). The heterotrophic flagellates group MAST-4 (Massana *et al.*, 2004) is widespread in surface marine waters (Massana *et al.*, 2006; Rodríguez-Martínez *et al.*, 2009) but remains uncultured to date.

Interestingly, clade B1 arrived only in third position in the OSD data set, which suggests that it could be a more oceanic clade. Two Indian Ocean *Tara* stations, 41 and 42, with significant contribution of clade VII to the total photosynthetic sequences, showed dominance of B1 at the DCM and not in surface at the other stations where this clade is dominant. This could indicate an ecotype (or another clade) within the B1 clade adapted to deeper water. B1 showed the lowest average of sequence identity within prasinophytes clade VII clades (Supplementary Table 5) and the largest number of *Tara* OTUs (Supplementary Table 4), which could indicate the presence of different ecotypes within this clade.

Clade A1 was not as abundant as B1 but had a large contribution to the total photosynthetic sequences at the DCM and seemed to complement B1 at the Pacific stations. All A1 strains have been isolated from deep samples (Table 1) and possess an intron 500 bp long around position 1000 in the 18S rRNA gene.

Prasinophyte clade A4 seems to be a more coastal clade as it was more abundant in OSD and at some *Tara* stations closer to shore, like 32 in the Red Sea and 78 and 84 in the Atlantic and Southern Ocean (Figure 5 and Supplementary Figure 4). This clade is the one for which the largest number of strains has been isolated across wide oceanic and latitudinal ranges (Table 1). Clade A4 strains were isolated from temperate latitudes, as far North as ~ 49° N or as far South as ~ 33° S but also from tropical latitudes. The *Tara* Oceans metadata analysis suggest two temperature niches. The analysis of accessory pigment composition of two temperate A4 strains (RCC1124 and RCC1871, Table 1) showed that, in contrast with the others strains from lineage A and B, they do not contain loroxanthin (Lopes dos Santos *et al.*, 2016). It will be interesting to analyze the pigment composition of tropical A4 strains to know whether loroxanthin is associated with lower temperature niches.

Among *Tara* stations where B1 was not abundant, we did not observe a prevailing clade except in the Mediterranean Sea where A6 seemed to be dominant except at stations 4 and 7, which presented an assemblage of clades (Figure 5, Supplementary Figure 4). These latter stations are close to the Strait of Gibraltar, a well-known region of water mass exchange between Atlantic Ocean and Mediterranean Sea marked by salinity and temperature gradients (Gascard and Richez, 1985; Béranger *et al.*, 2005). The Mediterranean Sea presents a range of trophic conditions and it is known to harbor endemic phytoplankton species (for example, Gómez, 2006). However, no strain of prasinophytes clade VII has been isolated from the Mediterranean Sea and the only strain representing

14

A6 was isolated in 2013 from the Atlantic Ocean. Also only one environmental sequence among the nine available from the Mediterranean Sea was assigned to A6. The dominance of A6 in the Mediterranean Sea could be seasonal because all the stations at this oceanic region were sampled during late fall–early winter.

The distribution of the other clades, A2, A3, A5, A7 and B2, is too sporadic to draw major conclusions. Sensitive and quantitative techniques to assess the abundance and distribution of these clades would help to understand the role of these low abundance clades.

Conclusion

Amplicon-based analysis such as metabarcoding are influenced by several factors such as DNA extraction, amplicon size, primer preference, target gene copy number, sequencing analysis methods like OTU generation (Zinger *et al.*, 2012). Most of these bias will affect the interpretation of the results in the light of quantitative methods such as estimation of richness. Yet, metabarcoding is powerful to determine overall trends of dominants groups (de Vargas *et al.*, 2015). Here we analyzed the distribution of prasinophytes clade VII in two marine large 18S rRNA data sets from different environments: the coastal OSD and the oceanic *Tara* Oceans. Despite the bias and discrepancies in number of sequences between these two data sets, we could establish that prasinophytes clade VII is the dominant group of the green algae in oceanic waters and probably important primary producers as recently demonstrated (Rii *et al.*, 2016). Our phylogenetic analyses point out, however, wide heterogeneity within lineages A and B and the precise habitat of each clade is still unclear although the *Tara* data helped us to formulate some hypotheses such as the occurrence of A4 in temperate waters or dominance of A6 in the Mediterranean Sea. The power of defining clades for abundant groups such as prasinophytes clade VII is to be able to refine their distribution and correlation with environmental factors in order to understand their ecology. Each clade described here has a precise signature in the V9 region (Supplementary Figure 2) that allowed us to map their distribution through different oceans regions. These signatures could be further used to design fluorescence *in situ* hybridization or quantitative PCR probes in order to assess more precisely the abundance of the most prevalent clades and their contribution to the total photosynthetic community in marine waters.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

References

Amacher J, Neuer S, Anderson I, Massana R. (2009). Molecular approach to determine contributions of the protist community to particle flux. *Deep Sea Res Part I Oceanogr Res Pap* **56**: 2206–2215.

Amaral-Zettler LA, McCliment EA, Ducklow HW, Huse SM. (2009). A method for studying protistan diversity doing massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA Genes. *PLoS One* **4**: e6372.

Balzano S, Marie D, Gourvil P, Vaulot D. (2012). Composition of the summer photosynthetic pico and nanoplankton communities in the Beaufort Sea assessed by T-RFLP and sequences of the 18S rRNA gene from flow cytometry sorted samples. *ISME J* **6**: 1480–1498.

Béranger K, Mortier L, Crépon M. (2005). Seasonal variability of water transport through the Straits of Gibraltar, Sicily and Corsica, derived from a high-resolution model of the Mediterranean circulation. *Prog Oceanogr* **66**: 341–364.

Cachon M, Caram B. (1979). A symbiotic green alga, Pedinomonas symbiotica sp. nov. (Prasinophyceae), in the radiolarian *Thalassolampe margarodes*. *Phycologia* **18**: 177–184.

Collado-Fabri S, Vaulot D, Ulloa O. (2011). Structure and seasonal dynamics of the eukaryotic picophytoplankton community in a wind-driven coastal upwelling ecosystem. *Limnol Oceanogr* **56**: 2334–2346.

Crespo C, Rodríguez H, Segade P, Iglesias R, García-Estévez JM. (2009). *Coccomyxa* sp. (Chlorophyta: Chlorococcales), a new pathogen in mussels (*Mytilus galloprovincialis*) of Vigo estuary (Galicia, NW Spain). *J Invertebr Pathol* **102**: 214–219.

Decelle J, Colin S, Foster RA. (2015a). Photosymbiosis in marine planktonic protistsIn *Marine Protists: Diversity and Dynamics*. Springer: Japan, pp 465–500.

Decelle J, Romac S, Stern R, Bendif EM, Zingone A, Guiry M *et al.* (2015b). PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Mol Ecol Resour* **15**: 1435–1445.

Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, Robbens S *et al.* (2006). Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc Natl Acad Sci USA* **103**: 11647–11652.

de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R *et al.* (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**: 1261605.

Duanmu D, Bachy C, Sudek S, Wong C-H, Jiménez V, Rockwell NC *et al.* (2014). Marine algae and land plants share conserved phytochrome signaling systems. *Proc Natl Acad Sci USA* **111**: 15827–15832.

Foulon E, Not F, Jalabert F, Cariou T, Massana R, Simon N. (2008). Ecological niche partitioning in the picoplanktonic green alga *Micromonas pusilla*: evidence from environmental surveys using phylogenetic probes. *Environ Microbiol* **10**: 2433–2443.

Fuller NJ, Campbell C, Allen DJ, Pitt FD, Le Gall F, Vaulot D *et al.* (2006a). Analysis of photosynthetic picoeukaryote diversity at open ocean sites in the Arabian Sea using a PCR biased towards marine algal plastids. *Aquat Microb Ecol* **43**: 79–93.

Fuller NJ, Tarran G, Cummings D, Woodward M, Orcutt DM, Yallop M *et al.* (2006b). Molecular analysis of photosynthetic picoeukaryote community structure along an Arabian Sea transect. *Limnol Oceanogr* **51**: 2502–2514.

Gascard JC, Richez C. (1985). Water masses and circulation in the Western Alboran sea and in the Straits of Gibraltar. *Prog Oceanogr* **15**: 157–216.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**: 60–63.

Gómez F. (2006). Endemic and Indo-Pacific plankton in the Mediterranean Sea: a study based on dinoflagellate records. *J Biogeogr* **33**: 261–270.

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L *et al.* (2013). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small SubUnit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41**: D597–D604.

Guillou L, Eikrem W, Chrétiennot-Dinet M-J, Le Gall F, Massana R, Romari K *et al.* (2004). Diversity of picoplanktonic prasinophytes assessed by direct nuclear SSU rDNA sequencing of environmental samples and novel isolates retrieved from oceanic and coastal marine ecosystems. *Protist* **155**: 193–214.

Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R *et al.* (2008). Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). *Environ Microbiol* **10**: 3349–3365.

Guindon S, Dufayard J-F, Lefort V, Anisimova M. (2010). New algorithms and methods to estimate maximum-likelihoods phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

Hollibaugh JT, Wong PS, Bano N, Pak SK, Prager EM, Orrego C. (2001). Stratification of microbial assemblages in Mono Lake, California, and response to a mixing event. *Hydrobiologia* **466**: 45–60.

Jackson GA. (2001). Effect of coagulation on a model planktonic food web. *Deep Res Part I-Oceanographic Res Pap* **48**: 95–123.

Katoh K, Frith MC. (2012). Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics* **28**: 3144–3146.

Katoh K, Misawa K, Kuma K, Miyata T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30**: 3059–3066.

Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S *et al.* (2012). Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–1649.

Keller MD, Selvin RC, Claus W, Guillard RRL. (1987). Media for the culture of oceanic ultraphytoplankton. *J Phycol* **23**: 633–638.

Kopf A, Bicak M, Kottmann R, Schnetzer J, Kostadinov I, Lehmann K *et al.* (2015). The ocean sampling day consortium. *Gigascience* **4**: 27.

Krienitz L, Bock C, Kotut K, Luo W. (2012). *Picocystis salinarum* (Chlorophyta) in saline lakes and hot springs of East Africa. *Phycologia* **51**: 22–32.

Latasa M, Scharek R, Le Gall F, Guillou L. (2004). Pigment suites and taxonomic groups in Prasinophyceae. *J Phycol* **40**: 1149–1155.

Leliaert F, Smith DR, Moreau H, Herron MD, Verbruggen H, Delwiche CF *et al.* (2012). Phylogeny and molecular evolution of the green algae. *CRC Crit Rev Plant Sci* **31**: 1–46.

Lemieux C, Otis C, Turmel M. (2014). Six newly sequenced chloroplast genomes from prasinophyte green algae provide insights into the relationships among prasinophyte lineages and the diversity of streamlined genome architecture in picoplanktonic species. *BMC Genomics* **15**: 857.

Lepère C, Demura M, Kawachi M, Romac S, Probert I, Vaulot D. (2011). Whole genome amplification (WGA) of marine photosynthetic eukaryote populations. *FEMS Microbiol Ecol* **76**: 516–523.

Lepère C, Vaulot D, Scanlan DJ. (2009). Photosynthetic picoeukaryote community structure in the South East Pacific Ocean encompassing the most oligotrophic waters on Earth. *Environ Microbiol* **11**: 3105–3117.

Lewin RA, Krienltz L, Oerickei RG, Takeda H, Hepperle D. (2000). *Picocystis salinarum* gen. et sp. nov. (Chlorophyta) - a new picoplanktonic green alga. *Phycologia* **39**: 560–565.

Li WK. (1995). Composition of ultraphytoplankton in the central north Atlantic. *Mar Ecol Prog Ser* **122**: 1–8.

Lopes dos Santos A, Gourvil P, Rodriguez-Hernandez F, Garrido JL, Vaulot D. (2016). Photosynthetic pigments of oceanic Chlorophyta belonging to prasinophytes clade VII. *J Phycol* **52**: 148–155.

Lovejoy C, Vincent WF, Bonilla S, Roy S, Martineau MJ, Terrado R *et al.* (2007). Distribution, phylogeny, and growth of cold-adapted picoprasinophytes in Arctic seas. *J Phycol* **43**: 78–89.

Mahé F, Rognes T, Quince C, de Vargas C, Dunthorn M. (2014). Swarm: robust and fast clustering method for amplicon-based studies. *PeerJ* **2**: e593.

Marin B. (2012). Nested in the Chlorellales or independent class? Phylogeny and classification of the Pedinophyceae (Viridiplantae) revealed by molecular phylogenetic analyses of complete nuclear and plastid-encoded rRNA operons. *Protist* **163**: 778–805.

Marin B, Melkonian M. (2010). Molecular phylogeny and classification of the Mamiellophyceae class. nov. (Chlorophyta) based on sequence comparisons of the nuclear- and plastid-encoded rRNA Operons. *Protist* **161**: 304–336.

Massana R, Castresana J, Balagué V, Romari K, Groisillier A, Valentin K *et al.* (2004). Phylogenetic and ecological analysis of novel marine Stramenopiles. *Appl Environ Microbiol* **70**: 3528–3534.

Massana R, Terrado R, Forn I, Lovejoy C, Pedrós-Alió C. (2006). Distribution and abundance of uncultured heterotrophic flagellates in the world oceans. *Environ Microbiol* **8**: 1515–1522.

Monier A, Welsh RM, Gentemann C, Weinstock G, Sodergren E, Armbrust EV *et al.* (2012). Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. *Environ Microbiol* **14**: 162–176.

16

Moon-van der Staay S, Van der Staay GWM, Guillou L, Vaulot D, Claustre H, Medlin L. (2000). Abundance and diversity of prymnesiophytes in the picoplankton community from the equatorial Pacific Ocean inferred from 18S rDNA sequences. *Limnol Oceanogr* **45**: 98–109.

Moreau H, Verhelst B, Couloux A, Derelle E, Rombauts S, Grimsley N *et al.* (2012). Gene functionalities and genome structure in *Bathycoccus prasinos* reflect cellular specializations at the base of the green lineage. *Genome Biol* **13**: R74.

Not F, Latasa M, Marie D, Cariou T, Vaulot D, Simon N. (2004). A single species *Micromonas pusilla* (Prasinophyceae) dominates the eukaryotic picoplankton in the western English Channel. *Appl Environ Microbiol* **70**: 4064–4072.

Not F, Siano R, WHCF Kooistra, Simon N, Vaulot D, Probert I. (2012). Diversity and ecology of eukaryotic marine phytoplankton. *Adv Bot Res* **64**: 1–53.

O'Kelly CJ, Sieracki ME, Thier EC, Hobson IC. (2003). A transient bloom of *Ostreococcus* (Chlorophyta, Prasinophyceae) in West Neck Bay, Long Island, New York. *J Phycol* **39**: 850–854.

Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G *et al.* (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Sci Data* **2**: 150023.

Potter D, Lajeunesse TC, Saunders GW, Andersen RA. (1997). Convergent evolution masks extensive biodiversity among marine coccoid picoplankton. *Biodivers Conserv* **6**: 99–107.

R Development Core Team. (2013). R: a language and environment for statistical computing. *R Found Stat Comput* **1**: 409.

Rappé MS, Connon SA, Vergin KL, Giovannoni SJ. (2002). Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**: 630–633.

Richardson TL, Jackson GA. (2007). Small phytoplankton and carbon export from the surface ocean. *Science* **315**: 838–840.

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. *Limnol Oceanogr* **61**: 806–824.

Rodríguez F, Derelle E, Guillou L, Le Gall F, Vaulot D, Moreau H. (2005). Ecotype diversity in the marine picoeukaryote *Ostreococcus* (Chlorophyta, Prasinophyceae). *Environ Microbiol* **7**: 853–859.

Rodríguez-Martínez R, Labrenz M, Del Campo J, Forn I, Jürgens K, Massana R. (2009). Distribution of the uncultured protist MAST-4 in the Indian Ocean, Drake Passage and Mediterranean Sea assessed by real-time quantitative PCR. *Environ Microbiol* **11**: 397–408.

Roesler CS, Culbertson CW, Etheridge SM, Goericke R, Kiene RP, Miller LG *et al.* (2002). Distribution, production, and ecophysiology of *Picocystis* strain ML in Mono Lake, California. *Limnol Oceanogr* **47**: 440–452.

Romari K, Vaulot D. (2004). Composition and temporal variability of picoeukaryote communities at a coastal site of the English Channel from 18S rDNA sequences. *Limnol Oceanogr* **49**: 784–798.

Ronquist F, Huelsenbeck JP. (2003). MrBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Scanlan DJ, Ostrowski M, Mazard S, Dufresne A, Garczarek L, Hess WR *et al.* (2009). Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* **73**: 249–299.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB *et al.* (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537–7541.

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.

Shi XL, Marie D, Jardillier L, Scanlan DJ, Vaulot D. (2009). Groups without cultured representatives dominate eukaryotic picophytoplankton in the oligotrophic South East Pacific Ocean. *PLoS One* **4**: e7657.

Simmons MP, Sudek S, Monier A, Limardo AJ, Jimenez V, Perle CR *et al.* (2016). Abundance and biogeography of picoprasinophyte ecotypes and other phytoplankton in the Eastern North Pacific Ocean. *Appl Environ Microbiol* **82**: 1693–1705.

Six C, Worden AZ, Rodríguez F, Moreau H, Partensky F. (2005). New insights into the nature and phylogeny of prasinophyte antenna proteins: *Ostreococcus tauri*, a case study. *Mol Biol Evol* **22**: 2217–2230.

Steyer G, Sasser C, Evers E, Swenson E, Suir GM, Sapkota S. (2008). Influence of the Houma navigation canal on salinity patterns and landscape configuration in coastal Louisiana. *US Geological Survey Open-File Report*, 2008–1127.

Takaichi S. (2011). Carotenoids in algae: distributions, biosyntheses and functions of carotenoids in algae. *Mar Drugs* **9**: 1101–1118.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol Biol Evol* **30**: 2725–2729.

Tragin M, Lopes dos Santos A, Christen R, Vaulot D. (2016). Diversity and ecology of green microalgae in marine systems: an overview based on 18S rRNA gene sequences. *Perspect Phycol*; e-pub ahead of print 22 June 2016; doi:10.1127/pip/2016/0059.

Treusch AH, Demir-Hilton E, Vergin KL, Worden AZ, Carlson C a, Donatz MG *et al.* (2012). Phytoplankton distribution patterns in the northwestern Sargasso Sea revealed by small subunit rRNA genes from plastids. *ISME J* **6**: 481–492.

Vaulot D, Lepère C, Toulza E, de la Iglesia R, Poulain J, Gaboyer F *et al.* (2012). Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS One* **7**: e39648.

Viprey M, Guillou L, Ferréol M, Vaulot D. (2008). Wide genetic diversity of picoplanktonic green algae (Chloroplastida) in the Mediterranean Sea uncovered by a phylum-biased PCR approach. *Environ Microbiol* **10**: 1804–1822.

Wang Q, Garrity GM, Tiedje JM, Cole JR. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**: 5261–5267.

West NJ, Schönhuber W a, Fuller NJ, Amann RI, Rippka R, Post A F *et al.* (2001). Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by *in situ* hybridization using 16S rRNA-targeted oligonucleotides. *Microbiology* **147**: 1731–1744.

Worden AZ, Lee J-H, Mock T, Rouzé P, Simmons MP, Aerts AL *et al.* (2009). Green evolution and

dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**: 268–272.

Worden AZ, Nolan JK, Palenik B. (2004). Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnol Oceanogr* **49**: 168–179.

Wu W, Wang L, Liao Y, Huang B. (2015). Microbial eukaryotic diversity and distribution in a river plume and cyclonic eddy-influenced ecosystem in the South China Sea. *Microbiologyopen* **4**: 826–840.

Zhu F, Massana R, Not F, Marie D, Vaulot D. (2005). Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* **52**: 79–92.

Zinger L, Gobet A, Pommier T. (2012). Two decades of describing the unseen majority of aquatic microbial diversity. *Mol Ecol* **21**: 1878–1896.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)