environmental microbiology

Environmental Microbiology (2017) **00**(00), 00–00 doi:10.1111/1462-2920.13952

Comparison of coastal phytoplankton composition estimated from the V4 and V9 regions of the 18S rRNA gene with a focus on photosynthetic groups and especially Chlorophyta

Margot Tragin,[1] Adriana Zingone[2] and Daniel Vaulot[1]*

[1]*Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, CNRS, Station Biologique, Place Georges Teissier, 29680 Roscoff, France.*

[2]*Department of Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, Villa Comunale, Naples, Italy.*

Summary

We compared the composition of eukaryotic communities using two genetic markers (18S rRNA V4 and V9 regions) at 27 sites sampled during Ocean Sampling Day 2014, with a focus on photosynthetic groups and, more specifically green algae (Chlorophyta). Globally, the V4 and V9 regions of the 18S rRNA gene provided similar images of alpha diversity and ecological patterns. However, V9 provided 20% more OTUs built at 97% identity than V4. 34% of the genera were found with both markers and, of the remnant, 22% were found only with V4 and 44% only with V9. For photosynthetic groups, V4 and V9 performed equally well to describe global communities at different taxonomic levels from the division to the genus and provided similar Chlorophyta distribution patterns. However, at lower taxonomic level, the V9 dataset failed for example to describe the diversity of Dolichomastigales (Chlorophyta, Mamiellophyceae) emphasizing the lack of V9 sequences for this group and the importance of the reference database for metabarcode analysis. We conclude that in order to address questions regarding specific groups (e.g., a given genus), it is necessary to choose the marker based not only on the genetic divergence within this group but also on the existence of reference sequences in databases.

Received 10 April, 2017; accepted 30 September, 2017. *For correspondence. E-mail: vaulot@sb-roscoff.fr; Tel. (+33) 2 98 29 23 23; Fax (+33) 2 98 29 23 24

Introduction

Planktonic organisms are distributed throughout all branches of the tree of life (Baldauf, 2008) but share 'universal' genes presenting certain degrees of genetic variability, which allow them to be used as barcode markers to investigate biological diversity (Chenuil and Anne, 2006). The development of high-throughput sequencing (HTS) allows the acquisition of large metabarcoding datasets (i.e., one marker gene is amplified and sequenced for all organisms), which complement the time-consuming and expertise-demanding morphological inventories to explore the diversity and distribution of protist groups in the ocean. The 18S rRNA gene is commonly used to investigate eukaryotic diversity and community structures (López-García *et al.*, 2001; Moon-van der Staay *et al.*, 2001). The complete 18S rRNA gene (around 1,700 base pairs) from environmental clone libraries can only be sequenced by the Sanger method (Sanger and Coulson, 1975) using a combination of primers. In contrast, HTS provides a very large number of reads but allows only small fragments to be sequenced (van Dijk *et al.*, 2014). Small hypervariable regions of the 18S such as V9 (around 150 bp located near the end of the 18S rRNA gene) or V4 (around 450 bp in the first half of the gene) can be targeted depending on the sequence length allowed by the sequencing technology used. Initially, the Illumina technology only allowed to sequence the V9 region because of its relatively small size (Amaral-Zettler *et al.*, 2009). In recent years longer reads became possible (up to 2 $\times$ 300 bp with current Illumina technology, van Dijk *et al.*, 2014) allowing the sequencing of the V4 region. Both the V4 and V9 regions have been used recently to describe diversity and ecological patterns of protists in several large scale studies (Massana *et al.*, 2014; de Vargas *et al.*, 2015).

The performance of the 18S RNA hypervariable regions as barcodes and the interpretation of results produced remain a matter of debate. Hu and colleagues (2015) showed that the V4 region provides an image of diversity similar to that obtained from the entire 18S rRNA gene. The choice between V4 and V9 depends on the taxonomic

Table 1. Location of OSD 2014 stations, number of reads in initial datasets, percentage of reads subsampled and percentage of photosynthetic reads.

				V4			V9		
OSD	Station	Ocean	Region	Raw reads	% of reads subsampled	% of photo. reads	Raw reads	% of reads subsampled	% of photo. reads
2	Roscoff – SOMLIT	North Atlantic Ocean	English Channel	343 626	59.0	28.1	387 351	52.3	25.5
3	Helgoland	North Atlantic Ocean	North Sea	315 340	64.3	27.7	257 957	78.6	34.3
14	Banyuls	Mediterranean Sea	Western Basin	311 053	65.2	18.2	406 871	49.8	11.2
22	Marseille -Solemio SOMLIT	Mediterranean Sea	Western Basin	302 687	67.0	7.6	353 503	57.3	7.8
30	Tvärminne	North Atlantic Ocean	Gulf of Finland	296 892	68.3	8.8	346 294	58.5	4.3
37	Port Everglades	North Atlantic Ocean	East coast of USA	338 053	60.0	33.6	361 524	56.1	27.7
39	Charleston Harbor	North Atlantic Ocean	East coast of USA	332 841	60.9	73.1	296 868	68.3	49.0
43	SIO Pier	North Pacific Ocean	West coast of USA	320 295	63.3	10.9	388 996	52.1	21.0
49	Vida	Mediterranean Sea	Adriatic Sea	*202 710*	*100.0*	14.4	302 436	67.0	14.5
54	Maine Booth Bay	North Atlantic Ocean	East coast of USA	290 311	69.8	14.5	365 441	55.5	36.9
55	Maine Damariscotta River	North Atlantic Ocean	East coast of USA	237 919	85.2	30.7	276 076	73.4	43.3
60	South Carolina 2-North Inlet	North Atlantic Ocean	East coast of USA	268 351	75.5	50.3	353v390	57.4	33.9
72	Boknis Eck	North Atlantic Ocean	Kattegat	356 529	56.9	26.5	475 461	42.6	23.3
76	Foglia	Mediterranean Sea	Adriatic Sea	242 825	83.5	11.5	386 655	52.4	13.2
77	Metauro	Mediterranean Sea	Adriatic Sea	303 448	66.8	26.5	377 917	53.6	26.4
80	Young Sound	North Atlantic Ocean	Greenland Sea	349 267	58.0	17.4	436 165	46.5	23.0
99	C1	Mediterranean Sea	Adriatic Sea	339 739	59.7	14.2	449 242	45.1	12.7
123	Shikmona	Mediterranean Sea	Eastern Basin	286 203	70.8	8.9	416 420	48.7	8.3
124	Osaka Bay	North Pacific Ocean	Japan Sea	237 367	85.4	34.8	478 261	42.4	31.0
132	Sdot YAM	Mediterranean Sea	Eastern Basin	285 592	71.0	26.4	399 001	50.8	17.7
141	Raunefjorden	North Atlantic Ocean	Coast of Norway	308 267	65.8	0.8	402 413	50.4	1.6
143	Skidaway Institute of Oceanography	North Atlantic Ocean	East coast of USA	328 039	61.8	81.4	410 937	49.3	65.3
146	Fram Strait	North Atlantic Ocean	Greenland Sea	369 221	54.9	44.4	447 907	45.3	41.7
149	Laguna Rocha Norte	South Atlantic Ocean	Coast of Uruguay	324 063	62.6	52.2	323 981	62.6	44.0
150	Laguna Rocha Sur	South Atlantic Ocean	Coast of Uruguay	338 373	59.9	50.8	367 936	55.1	44.9
152	Compass Buoy Station	North Atlantic Ocean	Badford Basin	327 454	61.9	9.8	407 377	49.8	17.0
159	Brest – SOMLIT	North Atlantic Ocean	Celtic Sea	327 901	61.8	30.7	443 747	45.7	22.9

Table 2. Evolution of sequence number through the analysis pipeline.

Step	Step description	V4	V9
	Total number of sequences initially	8 844 871	11 393 040
1	Total number of sequence subsampled	5 473 170	5 473 170
	Total number of sequence subsampled (%)	61.9	48.0
	Total number of sequences per station	202 710	202 710
2	Unique sequences	1 430 038	916 411
3	Unique sequences after filtering (quality and size)	203 214	103 068
4	Unique sequences after chimera check and preclustering	57 383	28 134
5	Unique sequences after singleton removal	53 530	26 370
	Total number of sequences considered finally	3 796 476	4 651 851
6	OTUs (97% similarity)	13 169	16 383

levels as well as the specific groups targeted. It is necessary to make detailed comparisons of genetic distances for each targeted region between and within the groups of interest (Dunthorn *et al.*, 2012; Pernice *et al.*, 2013) and to determine whether reference sequences are available for the group of interest in the target region (Tragin *et al.*, 2016). The sequencing platform may also have some impact: using the 454 technology, Behnke and colleagues (2011) showed that the sequencing error rate was taxon dependent, but V4 error rates were higher than for V9. Analysis of mock communities has highlighted possible biases in molecular methods such as the generation of artificial diversity (Egge *et al.*, 2013). The primers used may also produce a bias against groups whose target fragments are not amplified. For example, some widely used V4 primers miss Haptophyta and Foraminifera, which are important groups of the marine plankton (Massana *et al.*, 2015). Finally bioinformatics steps such as raw sequence filtering based on sequence quality and length, clustering algorithm and threshold to regroup sequences into Operational Taxonomic Units (OTUs) may influence the final results (Majaneva *et al.*, 2015).

Several studies have compared the structure of microbial communities provided by the V4 versus V9 regions in specific environments such as an anoxic fjord in Norway (Stoeck *et al.*, 2010) or for specific planktonic group such as Radiolaria (Decelle *et al.*, 2014). Some of these studies pointed out that the relative number of V4 and V9 reads may be different depending on the taxonomic levels and groups considered (Stoeck *et al.*, 2010; Giner *et al.*, 2016). Stoeck and colleagues (2010) found that the V9 region recovered more diversity at higher taxonomic levels than the V4 region: the number of unique V4 reads was very low for ciliates and dinoflagellates in comparison to V9, while pelagophytes (Ochrophyta) were not detected at all when using V4. In contrast, both papers (Stoeck *et al.*, 2010; Giner *et al.*, 2016) found that V4 provided more Chlorophyta unique sequences than V9. However, these studies were relying on different technologies for V4 and V9 sequencing. Recently, Piredda and colleagues (2017) used the same sequencing technology to analyze both the V4 and the V9 regions of marine protist communities in different seasons in the Gulf of Naples. They showed that V4 and V9 performed equally well to describe temporal patterns of protist variations and recovered the same number of OTUs (at 95% similarity) with both markers. However, this study was limited to a single sampling site.

The Ocean Sampling Day project has sampled a large number (157) of mostly coastal stations at the summer solstice (June 21) of 2014 with the aim of determining the composition, structure and distribution of prokaryotic and eukaryotic microbial community in marine waters using metabarcode and metagenomic approaches (Kopf *et al.*, 2015). Within this project, the V4 and V9 regions of the 18S rRNA gene from 27 locations were sequenced using the Illumina technology. In the present study, we compare the V4 and V9 metabarcodes using identical sequence processing algorithms. We focus on different levels. First, we analyze the total protist community in terms of richness and diversity. Then, we look in detail at the community composition at the Class level for photosynthetic groups. We finally focus on the contribution at each station of Chlorophyta classes and of Mamiellophyceae genera, for which a high quality reference sequence database has been recently reannotated (Tragin *et al.*, 2016) and which have been the subject of recent ecological studies in oceanic waters (Monier *et al.*, 2016; Simmons *et al.*, 2016; Clayton *et al.*, 2017).

Results

Global eukaryotic community

Twenty-seven stations were selected for which both V4 and V9 metabarcodes were obtained (Table 1 and Fig. 1). The two datasets were subsampled in order to process the same number of reads per station. After subsampling, the V4 and V9 datasets were reduced to 62% and 48% of their original size respectively (Table 2). The number of unique sequences (Table 2) was higher for V4 (around 1 400 000) than for V9 (around 900 000). After filtering based on length and ambiguities, twice more reads were obtained

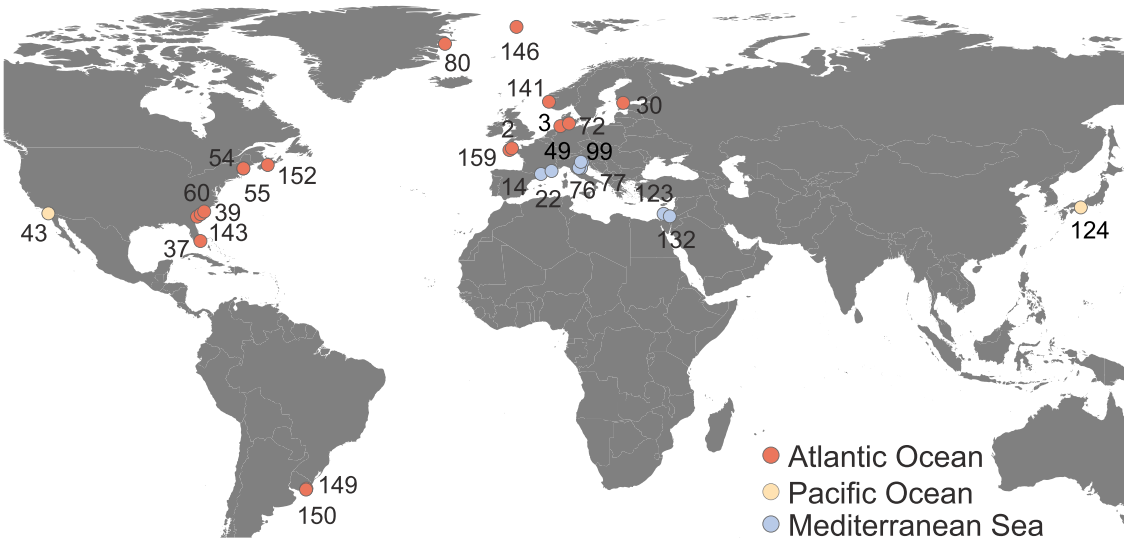


Fig. 1. Map of the 27 OSD stations sampled 2014 for which both V4 and V9 sequences were available. [Colour figure can be viewed at wileyonlinelibrary.com]

for V4 than for V9 (Table 2). Roughly 40 times more chimeras were found for V4 than for V9 (about 7500 against 170). Following taxonomic assignment, all eukaryotic groups were retained, not just protists.

Rarefaction curves computed for the global datasets as well as for each station (Supporting Information Figs S2A and S3) reached saturation, suggesting that the sequencing effort was sufficient. Global maximum richness varied between the datasets: 16 383 OTUs (4311 distinct assignations) were obtained for V9 against 13 169 OTUs (3412 distinct assignations) for V4. The two datasets yielded similar rank abundance curves (Supporting Information Fig. S2B), although V9 had larger OTUs as attested by the fact that the curve for V9 was above that for V4. The size of the largest OTU was equivalent (around 180 000 sequences).

The number of OTUs per station varied from 500 to about 3000 with respective averages of 1200 and 1600 for V4 and V9 respectively (Fig. 2A and Table 3). Although a positive correlation was found between the number of OTUs for V4 and V9 per station (R^2 = 0.99, Fig. 2A), the number of OTUs per stations was higher for V9 than for V4 (slope = 1.27, Fig. 2A) and this difference was confirmed by a Wilcoxon test (Table 3). The comparison of Simpson's diversity index per station for the two datasets (Fig. 2B and Table 3) showed that V4 and V9 diversity values were similar for large values between 0.9 and 1, irrespective of the OTU richness. For lower values (0.6–0.9), the Simpson's index was higher for V9 than V4 except at station OSD30 in the Gulf of Finland (Fig. 2B). At the latter station, one specific metazoan OTU (assigned to copepods and corresponding to 105 202 reads) was dominating the V9 reads, but this OTU did not dominate the V4 reads. If this copepod OTU is not taken into account (grey star in Fig. 2B), the V4 and V9 datasets have a similar alpha diversity (0.91 and 0.95 respectively). The number of genera (assignations without _X) found in the OSD datasets was equal to 3595, among which 34% were found with both markers and, of the remnant, 22% were found only with V4 and 44% only with V9. On average, four OTUs were assigned to the same genus and the maximum number of OTUs per genus reached 128 for V4 and 187 for V9. Ninety eight % of genera found only in one dataset were represented by less than 10 OTUs.

Non-parametric multidimensional scaling analysis (NMDS, Supporting Information Fig. S4A and B) and hierarchical clustering (Supporting Information Fig. S4C and D) were used to visualize the V4 and V9 communities based on OTUs (final stress values were, respectively, 0.187 and 0.195) using Bray–Curtis dissimilarity. Many stations grouped together in a similar way for both V4 and V9, some according to their geographic location (Supporting Information Fig. S4 and Fig. 1), as in the case of the Mediterranean Sea (OSD14, 22, 49, 76, 77 and 99) or of the subtropical Atlantic coast of the United States (OSD39, 60 and 143). Both V4 and V9 communities were structured by the same combination of environmental parameters with opposite gradients of nitrates, phosphates and chlorophyll on one side vs. silicates, temperature and salinity on the other side (Supporting Information Fig. S4A and B).

Photosynthetic groups

We next focused on photosynthetic groups for which taxonomic assignation relies on recently validated reference databases (Edvardsen *et al.*, 2016; Tragin *et al.*, 2016).

Table 3. General descriptive statistics: maximum, minimum, mean, standard deviation and results of the Wilcoxon test (*P* value) for V4 versus V9 OTU numbers, Simpson index (data from the Fig. 2) and photosynthetic groups relative contribution (see Supporting Information Figs S5, S6, S8 and S10).

	V4				V9				
Parameter	Max.	Min.	Mean	SD	Max.	Min.	Mean	SD	*P* value
OTU number	2906	522	1216	579	3216	911	1620	617.65	5.92E–06
Simpson Index	0.63	0.99	0.91	0.08	0.63	0.99	0.92	7.30E–02	4.90E–02
% of photosynthetic reads									
Ochrophyta	91.1	19.3	60.3	18.1	87.6	20.4	64.1	19.8	**0.4**
Chlorophyta	69.6	3.9	26.4	17.2	55.2	2.3	19.9	15.5	6.33E–05
Haptophyta	30.2	0.1	7.9	8.6	37.7	0.3	11.7	10.5	8.19E–07
Cryptophyta	13.6	0.0	4.2	3.6	19.7	0.3	4.9	4.2	8.00E–03
Bacillariophyta	89.5	8.6	49.1	21.6	86.2	8.1	51.8	23.4	9.50E–02
Dictyochophyceae	35.2	0.0	4.4	7.5	30.1	0.0	3.3	6.1	2.90E–03
Chryso-Synurophyceae	24.3	0.1	3.4	4.8	16.7	0.1	3.2	3.7	**0.5**
Pelagophyceae	7.0	0.0	0.7	1.6	8.0	0.0	0.7	1.6	***0.56***
Mamiellophyceae	49.6	0.0	12.1	14.2	45.7	0.2	10.9	12.9	**0.25**
Trebouxiophyceae	53.5	0.0	4.9	11.2	39.3	0.0	2.3	7.5	0.023
Chlorodendrophyceae	68.1	0.0	4.9	13.0	53.0	0.0	3.1	10.1	2.50E–05
Pyramimonadales	6.4	0.0	1.7	2.0	7.6	0.0	1.7	1.9	9.60E–03
Ulvophyceae	4.9	0.0	0.6	1.2	3.9	0.0	0.4	0.9	4.60E–02
Pseudoscourfieldiales	17.1	0.0	1.2	3.5	9.0	0.0	0.6	1.8	1.00E–03
% of Chlorophyta reads									
Micromonas	34.9	0.0	4.8	7.6	32.0	3.00E–02	4.3	6.8	**0.5**
Mamiella	2.5	0.0	0.2	0.5	1.8	0.0	0.2	0.3	***0.3***
Ostreococcus	35.8	0.0	4.0	9.6	40.0	0.0	4.4	10.6	***0.5***
Bathycoccus	4.0	0.0	0.7	1.1	3.7	0.0	0.6	1.0	***0.6***

P values in bold are above the 0.05 threshold indicating that V4 and V9 are not significantly different, while *P* values in italics were computed with datasets presenting ex aequo values.

Dinophyceae were excluded from the analysis since about 50% of the species are not photosynthetic (Gómez, 2012). The percent of reads assigned to photosynthetic groups was quite similar between datasets: 28.6% versus 25.9% for V4 and V9, respectively. The four major photosynthetic groups were Ochrophyta (mostly diatoms), Chlorophyta (green algae), Haptophyta and Cryptophyta (Fig. 3A). The Rhodophyta, Cercozoa (Chlorarachniophyta) and Discoba (Euglenales) represented less than 1.5% of the photosynthetic groups in the two datasets (Fig. 3A). Procrustean analysis suggested that the relative contribution of photosynthetic groups per station was similar between V4 and V9 ($m^2 = 0.17$ and $r = 0.91$). The number of OTUs assigned to Ochrophyta was quite similar in the two datasets (1215 and 1250 in V4 and V9 respectively). In contrast, the number of V9 OTUs was almost twice that of V4 for Chlorophyta and Cryptophyta and three times for Haptophyta, but average OTUs size was similar (377, 64 and 91 and 573, 100 and 241 in V4 versus V9). For these three photosynthetic groups, average pairwise identity between the OTUs reference sequences was higher for V4 than V9 (76% vs. 72% for Chlorophyta, 84% vs. 76% for Cryptophyta and 86% vs. 77% for Haptophyta), indicating that V4 has lower genetic variability for these groups and, therefore, is less discriminating.

The relative contribution of photosynthetic groups was very different among the stations which ranged from estuarine to oligotrophic oceanic waters. Ochrophyta contribution was statistically similar for V4 and V9 (Table 3) and varied between 20% (OSD14, 146) and 90% (OSD159, 60) of the photosynthetic metabarcodes (Supporting Information Fig. S5A). Chlorophyta contribution varied between 5% (OSD76 and 159) and 70% (OSD14). Chlorophyta contribution was slightly higher in V4, and the difference was confirmed by the Wilcoxon test, except for stations OSD149 and 150 (Supporting Information Fig. S5B). Haptophyta contribution varied across stations from a few percent up to 40% (OSD22, 49 and 146) and was larger for V9 than for V4 (Table 3) except for OSD3 (Supporting Information Fig. S5C). Cryptophyta contribution was on average 4% (Table 3) in both datasets and varied between a few percent and 20% (OSD150). It was similar for V4 and V9 (Supporting Information Fig. S5D) except at OSD76, 149 and 150.

Among Ochrophyta, diatoms (Bacillariophyta) largely dominated, followed by Dictyochophyceae and Chrysophyceae–Synurophyceae (Fig. 3B). Diatom relative contribution to photosynthetic metabarcodes per stations was around 50% on average (Table 3) and varied between 15% (OSD30 and 146) to 90% (OSD159 and 60). Diatom

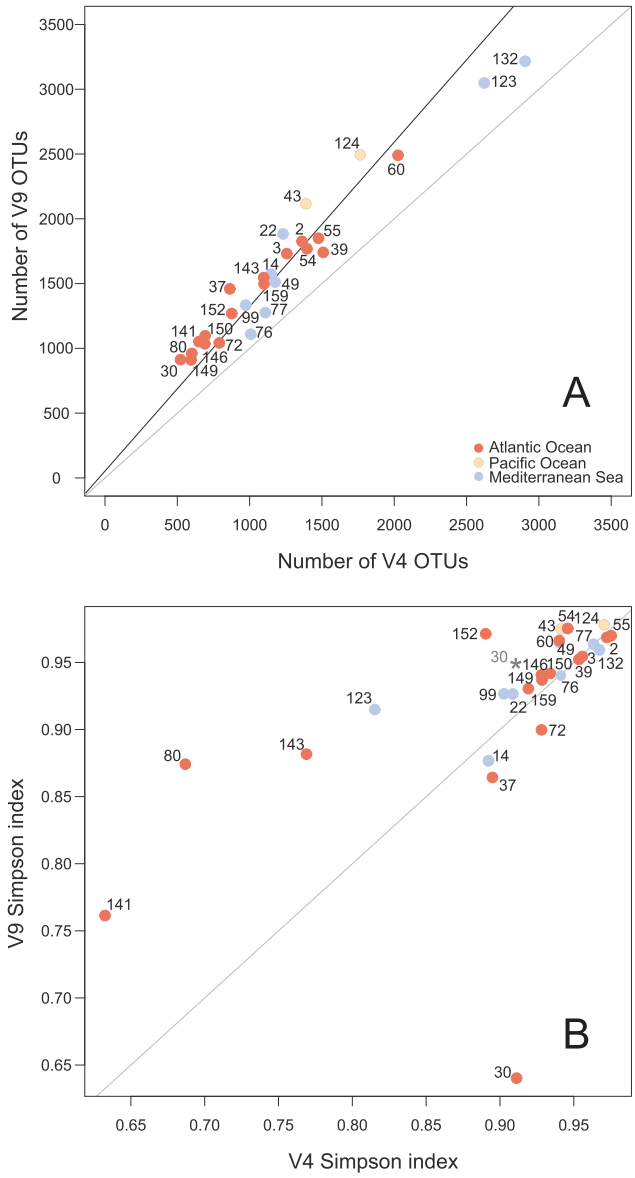


Fig. 2. A. Species richness: number of OTUs per stations for V4 versus V9. The grey line corresponds to $y = x$, and the black line corresponds to the regression $y = 1.27x + 53$ ($R^2 = 0.996$). B. Simpson's diversity index per stations for V4 versus V9. Grey star corresponds to the OSD30 Simpson's index after removal of the metazoan V9 OTU. [Colour figure can be viewed at wileyonlinelibrary.com]

contribution was statistically similar between V4 and V9 (Table 3). Dictyochophyceae relative contribution was below 10% except for five stations (OSD22, 149, 150, 152 and 72), where it reached 35% of photosynthetic reads (Supporting Information Fig. S6B). Dictyochophyceae contribution was slightly higher with V4 at these five stations. Chrysophyceae–Synurophyceae relative contribution was below 10% except for OSD76 (25%, Supporting Information Fig. S6C and Table 3), and V4 and V9 contributions were similar except at OSD30, where V9 was higher and OSD49 and 76 where V4 was higher (Supporting Information Fig. S6C). Pelagophyceae relative contribution was below 10% at individual stations, but V4 and V9 were similar (Supporting Information Fig. S6D and Table 3). Chlorophyta were dominated by Mamiellophyceae, followed by Trebouxiophyceae, Chlorodendrophyceae and Pyramimonadales (Fig. 3C). Trebouxiophyceae and Chlorodendrophyceae were more represented in V4, while Mamiellophyceae and Pyramimonadales were more represented in V9 (Fig. 3C). Other photosynthetic groups remained similar between the V4 and V9 datasets. Among Haptophyta, Prymnesiophyceae were largely dominating but two environmental clades HAP3 and HAP4 (Edvardsen *et al.*, 2016) were also recovered (Fig. 3D). For photosynthetic groups, the percentage

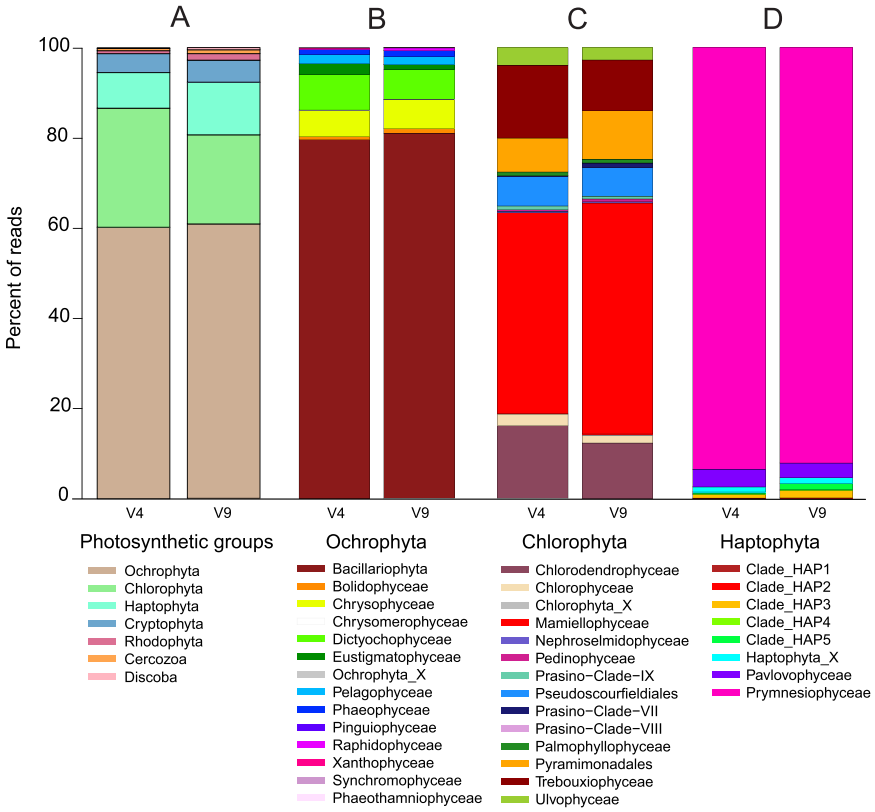


Fig. 3. A. Contribution of divisions to photosynthetic metabarcodes (Dinophyceae were excluded) for V4 and V9. B–D, Distribution of reads among classes for the three major photosynthetic divisions for V4 and V9: B, Ochrophyta; C, Chlorophyta; D, Haptophyta.

of genera found either in only one dataset (V4 or V9) or in both was class dependent, but globally 50% of the genera were recovered in both datasets (Supporting Information Fig. S7). Within Ochrophyta, more genera were found using V9 in five out of eight classes, but this was not the case for Bacillariophyta and Xanthophyceae for which more genera were recovered with V4. Raphidophyceae genera were almost all recovered in both V4 and V9 (Supporting Information Fig. S7). More red algae genera (Florideophyceae and Bangiophyceae) were recovered with V9. For Haptophyta, Cryptophyta and Chlorarachniophyceae most genera were found with both markers (Supporting Information Fig. S7).

Chlorophyta classes

The relative contributions of the six major Chlorophyta groups (Mamiellophyceae, Trebouxiophyceae, Chlorodendrophyceae, Pyramimonadales, Ulvophyceae and Pseudoscourfieldiales) in V4 and V9 were similar at most stations (Fig. 4A and Supporting Information Fig. S8) as supported by a procrustean comparison ($m^2 = 0.027$ and $r = 0.98$), but individual group contributions were not similar except for Mamiellophyceae (Table 3). Mamiellophyceae were dominant at most stations, but the four stations located in the Adriatic Sea (OSD49, 76, 77 and 99) shared a specific pattern with high contributions of Pseudoscourfieldiales and Chlorodendrophyceae in both V4 and V9 datasets (Fig. 4A). Stations OSD30, 54, 55 and 141, all located in North Atlantic coastal waters, presented differences in Chlorophyta class contribution recovered with V4 and V9 (Fig. 4A and Supporting Information Fig. S9A). For the first three, the Mamiellophyceae contribution in V9 was partially replaced in V4 by classes from 'core chlorophytes' such as Chlorodendrophyceae and/or Trebouxiophyceae. At OSD141, prasinophytes clade VII were only recovered with V9, while Chlorophyceae (*Chlamydomonas* sp.) were only recovered with V4 (Supporting Information Fig. S9A). BLAST analysis and alignment of Chlorophyta OTUs (data not shown) revealed that the V9 region of some *Chlamydomonas* is very similar to that of prasinophytes clade VII A5 (Lopes dos Santos *et al.*, 2016). Interestingly, the number of reads recovered in V4 and V9 for these 2 assignations (i.e., *Chlamydomonas* sp. for V4 and prasinophytes clade VII A5 for V9) was similar (51 vs. 47 reads respectively).

In general, Chlorophyta OTUs were well assigned by the Wang approach implemented in the Mothur software (Wang *et al.*, 2007) as validated by the results of BLAST (Supporting Information Data S6 and S7). However, some V9 reads initially assigned as Chlorophyta by the Wang approach hit bacterial sequences by BLAST and were not considered any further. Some V9 Chlorophyta OTUs shared 100% identity with several different Chlorophyta genera with (mostly in the UTC clade, i.e., Ulvophyceae, Trebouxiophyceae and Chlorophyceae) suggesting that

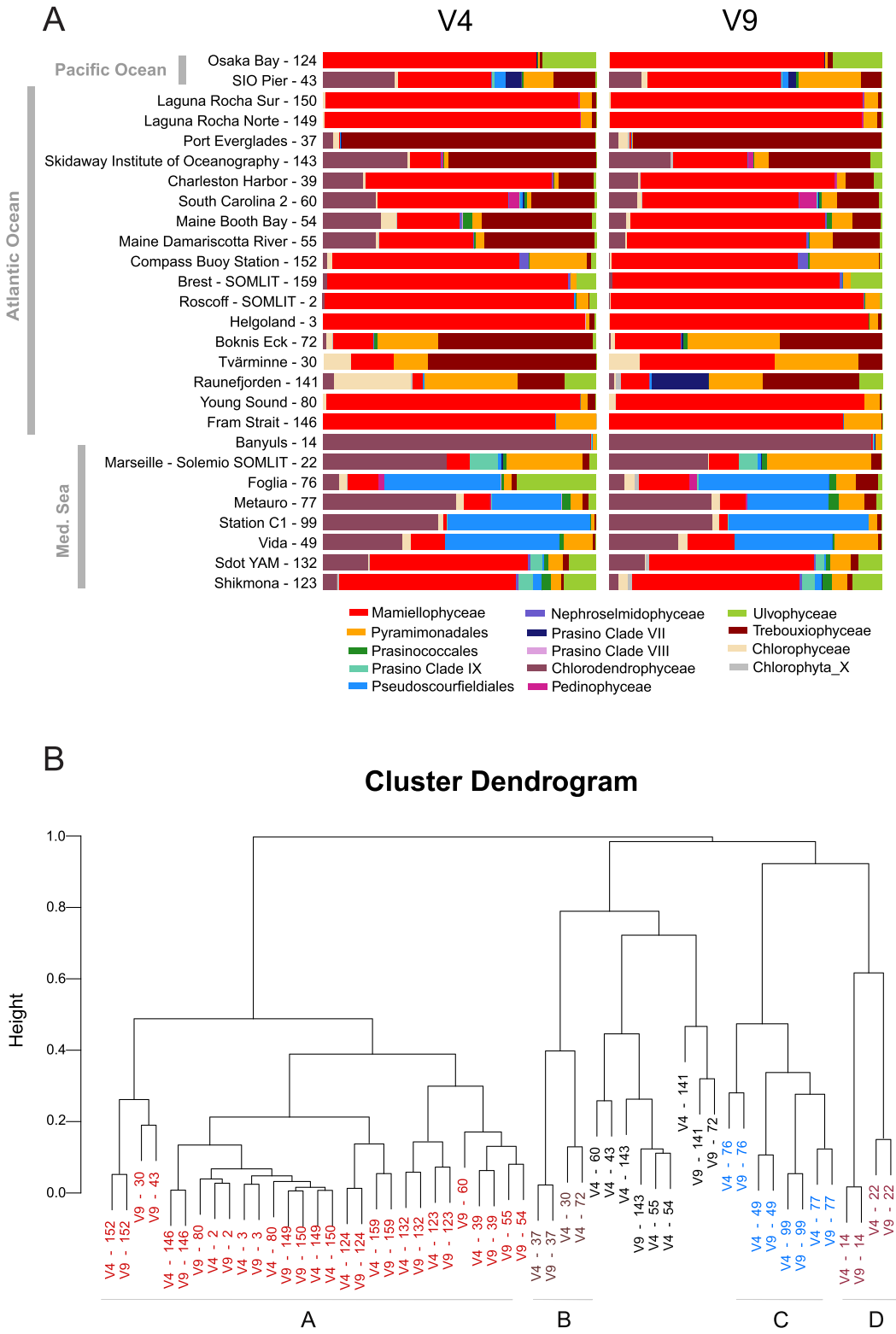


Fig. 4. A. Comparison of Chlorophyta read distribution (assigned at the class level) for 27 OSD stations. B. Comparison of Chlorophyta communities at the class level based hierarchical clustering for V9 and V4. The dissimilarity matrix was computed using Bray–Curtis distance. The stations were labelled by marker (V4 or V9). Stations where Mamiellophyceae represent more than 50% of the reads are colored in red (cluster A). Stations in blue are dominated by Pseudoscourfieldiales (cluster C), in brown by Trebouxiophyceae (cluster B) and in purple by Chlorodendrophyceae (cluster D).

the V9 region might not have the appropriate resolution to investigate UTC clade diversity. A number of genera within the UTC clade were only recovered with one marker in contrast to the Mamiellophyceae and Pyramimonadales for which almost all genera were recovered in both datasets (Supporting Information Fig. S7).

When Chlorophyta communities were clustered using the Bray–Curtis distance, V4 and V9 clustered together for individual stations except for OSD30, 43, 54, 55, 60, 72 and 143, (Fig. 4B). Clustering was strongly influenced by the contribution of Mamiellophyceae, because this class largely dominated in coastal waters and was present at almost all stations. A large group of stations where Mamiellophyceae were dominant formed a first cluster (Fig. 4B), whereas in four other groups of stations either another class was dominant (Trebouxiophyceae, Pseudoscourfieldiales or Chlorodendrophyceae) or none was really dominant (for example OSD141, Fig. 4B).

Mamiellophyceae genera

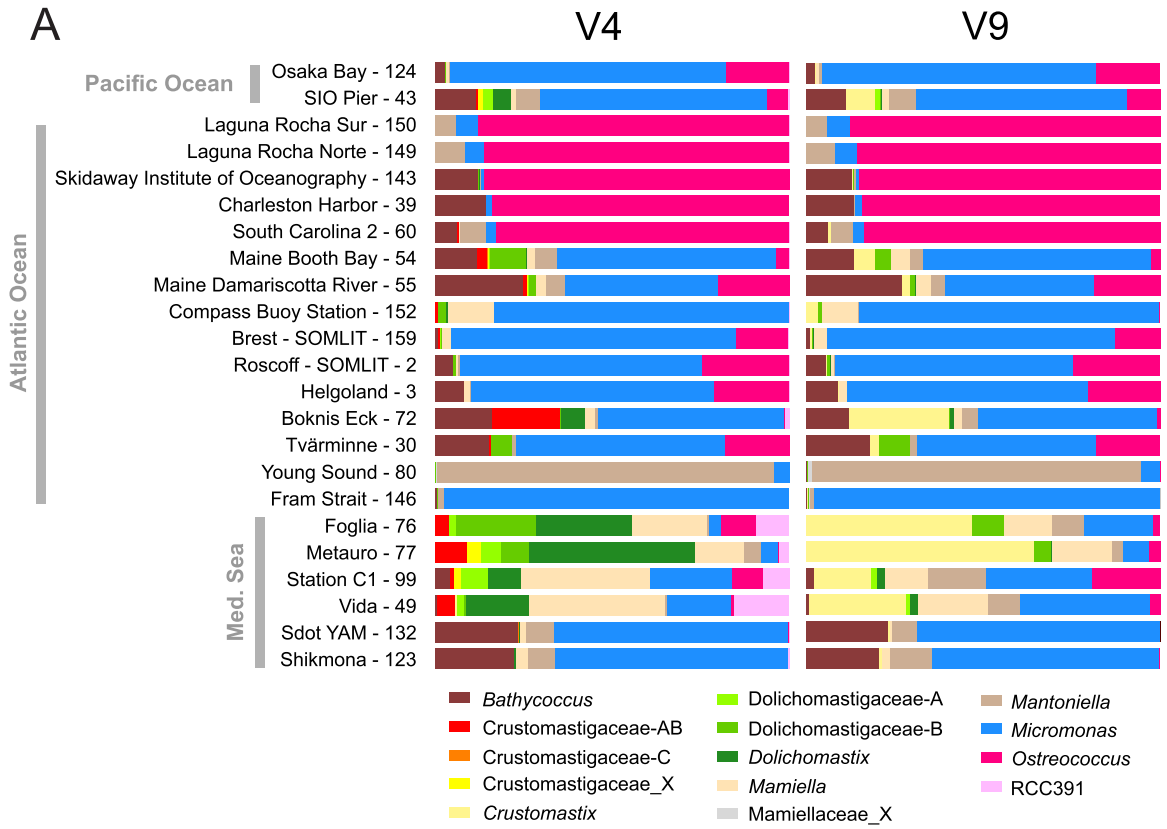
Mamiellophyceae that dominated at most OSD stations were further investigated at the genus level. Nine genera of Mamiellophyceae were found in the OSD datasets, seven of which were found in both datasets, one only in V4, assigned to RCC391, and one only in V9, assigned to *Monomastix*. The latter is a freshwater genus, and the OTUs assigned to it were badly assigned (BLAST analysis showed 100% identity with sequences of several land plants genera, see Supporting Information), while the RCC391 genus has eight references sequences for V4 against only one for V9. *Micromonas* and *Ostreococcus* were the two dominant genera, except at OSD80 in the Greenland Sea where *Mantoniella* was dominant and in the Adriatic Sea (OSD49, 76, 77 and 99) where Dolichomastigales and *Mamiella* were dominant (Fig. 5A). Procrustean comparison showed that V4 and V9 provided similar Mamiellophyceae genus distribution ($m^2 = 0.075$ and $r = 0.96$). The relative contributions per station of the four major genera *Micromonas*, *Mamiella*, *Ostreococcus* and *Bathycoccus* (Supporting Information Fig. S10) were overall statistically similar in the two datasets (Table 3), although it could be different for *Mamiella* at specific stations (OSD22, 49, 132 and 123). Stations located in the Adriatic Sea (OSD49, 76, 77 and 99) showed a different pattern in the heatmap (Supporting Information Fig. S9B) because V9 failed to discriminate the Dolichomastigales clades at the genus level. V9 recorded only *Crustomastix* contribution while V4 found 4–6 different clades of Crustomastigaceae and Dolichomastigaceae (Fig. 5A). Hierarchical clustering based on Bray–Curtis distances always grouped together V4 and V9 (Fig. 5B). Four groups of stations were observed depending on the genus dominant at the station: *Micromonas*, *Ostreococcus*, Dolichomastigales or *Mantoniella* (Fig. 5B).

Discussion

The OSD LifeWatch dataset, with its uniform sampling protocol, provides a unique opportunity to compare protist communities from a wide range of stations based on the two most widely used 18S rRNA markers, the V4 and V9 regions. In contrast to previous studies (e.g., Giner *et al.*, 2016), sequencing was performed on the same platform (Illumina), the same number of reads was analyzed at all stations for both V4 and V9. Bioinformatics analyses were conducted using exactly the same pipeline with the widespread software Mothur (Schloss *et al.*, 2009). A marked difference between the V4 and V9 datasets was the much larger number of chimeras found in V4. This could be due to the fact that the longer the amplified sequence is, the higher the chance is to have them recombining. Moreover, in contrast to the V9 region, the V4 region is composed of hypervariable regions as well as conserved regions (Monier *et al.*, 2016), which facilitates recombination. Finally, bioinformatics programs better detect chimeras on longer amplicons (Edgar *et al.*, 2011).

The choice of an identity threshold to build OTUs affects the number of recovered OTUs and the final taxonomic resolution. An analysis of 2200 full 18S sequences of protists (Caron *et al.*, 2009) showed that building OTUs at 95% identity provided a number of OTUs close to the expected number of species, but the authors remarked that a 98% identity threshold provides a better taxonomic resolution that allows to investigate interspecific diversity. In the present study, OTUs were built at 97% identity for both the V4 and the V9 regions of the 18S rRNA gene, in agreement with a number of recent studies that used these markers (e.g., Massana *et al.*, 2015; Ferrera *et al.*, 2016; Hu *et al.*, 2016). Clustering regions with different sizes (V4: 450 bp and V9: 150 bp) at the same identity level should produce more diverse OTUs for V4 than for V9, although regions where nucleotide changes are concentrated do not cover the whole amplicons and can be of different length in V4 and V9. For example in V4, most nucleotide diversity occurs within about 150 bp in the first half of the region (Monier *et al.*, 2016).

The V9 dataset provided 20% more OTUs than the V4. This difference between the number of OTUs for V4 and V9 is the same as the one unveiled in other environmental study. In the Naples times series, Piredda and colleagues (2017) also found 20% more OTUs built at 97% identity for V9 than for V4. This is in contrast to what would be expected based on the size difference between V4 and V9 discussed above. Interestingly, these authors showed that the number of OTUs built at 95% identity was similar for



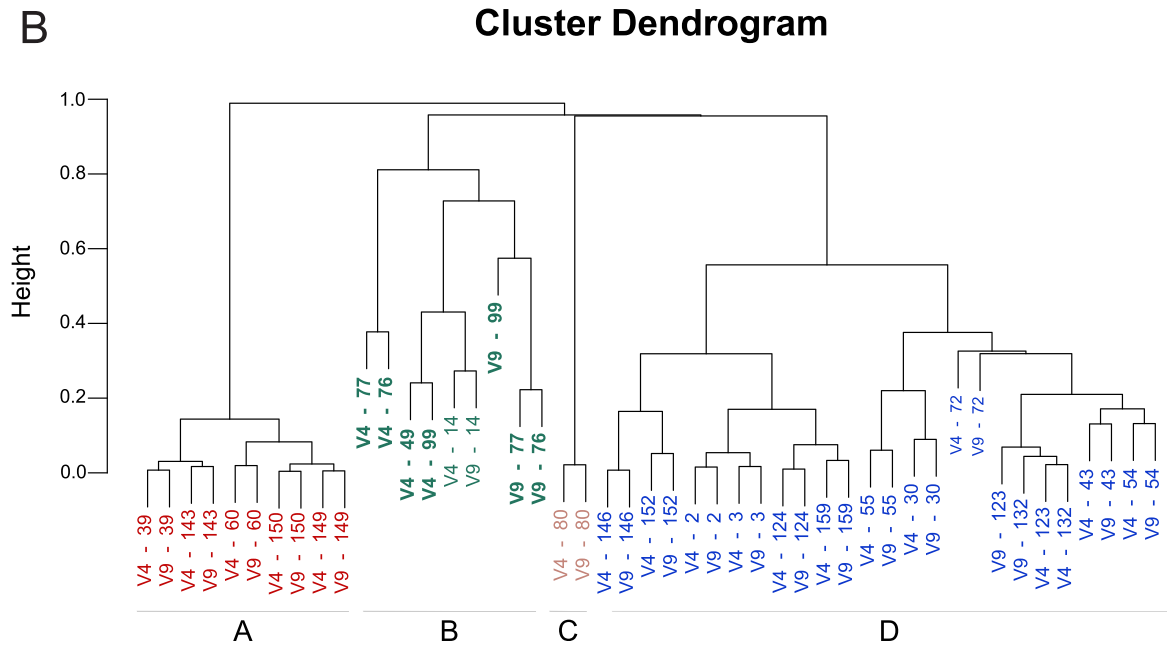


Fig. 5. A. Comparison of Mamiellophyceae read distribution (assigned at the genus level) for 23 OSD stations. Stations, where the number of reads assigned to Mamiellophyceae was lower than 100 were removed (OSD14, 22, 37 and 141). B. Comparison of Mamiellophyceae communities at the genus level by hierarchical clustering using V9 and V4. The stations were labelled by marker (V4 or V9) and station name. Stations in blue are dominated by *Micromonas* (cluster D), in red by *Ostreococcus* (cluster A), in green by Dolichomastigales (cluster B) and in grey by *Mantoniella* (cluster C).

V4 and V9, suggesting that at lower identity thresholds, the size difference has a lower impact.

The number of OTUs for the main photosynthetic phyla Ochrophyta, Chlorophyta, Haptophyta and Cryptophyta falls in the range found in European coastal waters using the V4 and 97% identity OTUs (1905, 314, 221 and 77, respectively, Massana *et al.*, 2015) except for the Haptophyta for which three times less OTUs were found in the OSD V4 dataset. The number of OTUs of the main photosynthetic phyla in the OSD V9 dataset was considerably lower than the numbers of Tara Oceans V9 OTUs, 3900, 1420, 713 and 195 respectively (de Vargas *et al.*, 2015). However, the depth of sequencing was much higher than in the OSD dataset (around one to two million reads per sample, i.e., 20–40 more than for OSD), which increases the occurrence of the rare OTUs that had been filtered out in the OSD dataset because of the relatively low read number.

At six stations (OSD30, 80, 123, 141, 143 and 152), the same species richness (OTU number) was observed but Simpson index was different between V4 and V9 (Fig. 2A and B). This means that even if the same number of OTUs was found for V4 and V9, the proportion of each OTU was different. The V9 Simpson index of OSD80 and 123 (0.87 and 0.91, respectively) fall in the range of Simpson index calculated in similar environments: for example in Baffin Bay (0.88, Hamilton *et al.*, 2008) and off the Mediterranean Sea coast (0.92, Ferrera *et al.*, 2016), but the V4 Simpson index was lower (0.68 and 0.81 respectively).

In the OSD dataset, photosynthetic groups (Dinophyceae excluded) varied widely between 0.8% and 81% and between 1.5% and 65% at the different stations for V4 and V9, respectively, representing on average 29% and 26% of the sequences recovered. These average numbers are comparable to those observed in other studies. For example, Massana and Pedrós-Alió (2008), synthesizing 35 picoplankton clone libraries of 18S gene from oceanic and coastal waters, found that photosynthetic sequences represented about 30% of eukaryotic sequences. The proportion of the main photosynthetic phyla Ochrophyta, Chlorophyta, Haptophyta and Cryptophyta, roughly 17%, 5–7%, 2–3% and 1.3%, respectively, in the OSD dataset are comparable to those found by Massana and Pedrós-Alió (2008) (15%, 7.7%, 2.4% and 2% respectively).

Mamiellophyceae dominated Chlorophyta in nutrient rich coastal waters, which is consistent with studies in European coastal waters (Massana *et al.*, 2015), in particular in the English Channel (Not *et al.*, 2004), and in the South East Pacific Ocean (Rii *et al.*, 2016). The stations located in the Adriatic Sea (OSD49, 76, 77 and 99) showed a specific pattern with a high contribution of Pseudoscourfieldiales and Chlorodendrophyceae. Several studies using optical microscopy found in the Adriatic Sea a high contribution of phytoflagellates, most of which could not be identified (Revelante and Gilmartin, 1976; Cerino *et al.*, 2012).

Within Mamiellophyceae the same genus, most of the time either *Micromonas* or *Ostreococcus*, was dominant in both V4 and V9 datasets. Not and colleagues (2009) found *Micromonas* to be the most prevalent genus in the world ocean coastal waters and at a more local scale *Micromonas* dominates coastal picoplankton in the Western English Channel (Not *et al.*, 2004). Rii and colleagues (2016) found that *Ostreococcus* was dominant in the upwelling-influenced coastal waters from Chile. OSD data also unveiled a high genetic diversity of the order Dolichomastigales especially in the Adriatic Sea. Viprey and colleagues (2008) made similar observations in oligotrophic Mediterranean surface waters and Monier and colleagues (2016) in the Tara *Oceans* survey.

Clustering based on taxonomic assignment, either Chlorophyta classes or Mamiellophyceae genera, confirmed that for most stations, the V4 and V9 communities clustered together as observed previously for Illumina vs 454 data obtained on picoplankton (Ferrera *et al.*, 2016). However, for Chlorophyta, V4 and V9 of five stations (OSD30, 43, 54, 55 and 60) did not cluster together (Fig. 4B). OSD43 and 60 were not close in the cluster dendrogram but no clear differences are seen either in the barplot (Fig. 4A) or in the heatmap (Supporting Information Fig. S9A). In contrast, OSD141 V4 and V9 communities clustered together in spite of obvious differences in the barplot (Fig. 4A and B) and in the heatmap (Supporting Information Fig. S9A). At OSD30, 54 and 55, the latter two being spatially close on the Eastern US coast, more Trebouxiophyceae and Chlorodendrophyceae were found with V4 which were replaced by Mamiellophyceae for V9. This could be explained by the fact that the reference sequences of the Trebouxiophyceae and Chlorodendrophyceae found at these stations do not cover the V9 region and that the corresponding V9 OTUs were classified as Mamiellophyceae, because of their similarity to the V9 regions of the latter class.

Concluding remarks – what is the best choice: V4 or V9?

The first element of choice between these two regions is based on the genetic divergence within and between the groups of interests (Chenuil and Anne, 2006). For Chlorophyta, average similarity is in general lower in V9 than V4 (Tragin *et al.*, 2016), which suggests that V9 will be more discriminating than V4 and will be the best choice. This is the case for example for prasinophytes clade VII, an important oceanic group, for which the use of 99% threshold for V9 OTUs allows to discriminate all sub-clades defined to date (Lopes dos Santos *et al.*, 2017), while in V4, several clades collapse together, having identical sequences in

that region. The V9 region of some *Chlamydomonas* is very similar to that of prasinophytes clade VII A5, which could lead to misinterpret the distribution of this specific sub-clade when using the V9 region. However this may not be the case for other groups such as Nephroselmidophyceae for which the two markers are equally suitable (Tragin *et al.*, 2016). The second element to take into account is the reference database that contains more representatives of each of the taxa investigated. For example, in the present study, the V9 region of the 18S rRNA gene failed to discriminate clades within Dolichomastigales (Supporting Information Fig. S9B), because there are only four Dolichomastigales V9 reference sequences against 69 for V4 (Tragin *et al.*, 2016). In the same way, obtaining accurate image of communities at stations which host rare or uncultured taxa is more difficult with V9 than with V4, because many sequences in public databases are short and do not extend to the end of the 18S rRNA gene. For example, Viprey and colleagues (2008) discovered one novel prasinophyte group (clade VIII) by using Chlorophyta-specific primers that only amplified a short (around 910 base pairs) sequence not extending to the V9 region and, therefore, this group can only be studied using V4.

Metabarcode analysis methods using assignation rely heavily on carefully curated public database such as PR^2 (Guillou *et al.*, 2013) or, even better, on specifically tailored databases that include, besides public sequences, reference sequences for the environment investigated, as for example Arctic specific databases for polar environments (Comeau *et al.*, 2011; Marquardt *et al.*, 2016). Other approaches to analyze metabarcoding datasets do not rely on reference databases. For example, oligotyping relies on nucleotide signatures to cluster sequences and can reveal fine distribution patterns of specific taxonomic groups (Eren *et al.*, 2014; Berry *et al.*, 2017), but to our knowledge it has not been applied to eukaryotes yet. Phylogenetic placement methods such as pplacer (Matsen *et al.*, 2010) allow to investigate phylogenetic diversity without assignation against a reference database. Phylogenetic approach however may be impacted by the lack of reference sequences and have to be complemented by statistical testing of the consistency of phylogenetic signals (Kembel, 2009; Stegen *et al.*, 2012).

Despite all these caveats, our analyses demonstrate overall that in most cases V4 and V9 provide similar images of the distribution specific photosynthetic groups such as the Chlorophyta and therefore that global studies using either of these markers are comparable.

Experimental procedures

Water samples were collected from 0 to 2 m depth at 27 stations in the world ocean (Fig. 1 and Table 1). Metadata (temperature, salinity, nitrates, phosphates, silicates and chlorophyll *a*) are available at https://github.com/MicroB3-IS/osd-analysis/wiki/Guide-to-OSD-2014-data. Samples were filtered on 0.8 μm pore size polycarbonate membranes without prefiltration and flash frozen at −80°C or in liquid nitrogen. DNA was extracted using the Power Water isolation kit (MoBio, Carlsbad, CA) following the manufacturer instructions. The V4 region was amplified using modified universal primer (Piredda *et al.*, 2017): V4_18SNext.For primer (5′ CCA GCA SCY GCG GTA ATT CC 3′) and V4_18SNext.Rev primer (5′ ACT TTC GTT CTT GAT YRA TGA 3′). The V9 region was amplified using modified universal primer (Piredda *et al.*, 2017): V9_18SNext.For (5′ TTG TAC ACA CCG CCC GTC GC 3′) and V9_18SNexr.Rev (5′ CC TTC YGC AGG TTC ACC TAC 3′). The library preparation was based on a modified version of the Illumina Nextera's protocol (Nextera DNA sample preparation guide, Illumina) and sequencing was done on an Illumina Miseq (NE08 Ocean Sampling Day protocols: https://github.com/MicroB3-IS/osd-analysis/wiki/Guide-to-OSD-2014-data#analysis-of-workable-18s-rdna-datasets-sequenced-by-lifewatch-italy). Amplicon PCR and sequencing (V4 region: 2 × 250 paired end sequencing using MiSeq Reagent kit v3 and V9 region: 2x150 paired end sequencing using MiSeq Reagent kit v2) was done by the Laboratory of Molecular Biodiversity (MoBiLab) of LifeWatch-Italy. R1 and R2 were filtered based on quality and length and assembled by the OSD consortium which provided the so-called 'workable hx2019; fasta files (https://owncloud.mpi-bremen.de/index.php/s/RDB4Jo0PAayg3qx?path=/2014/silva-ngs/18s/lifewatch/).

All subsequent sequence analyses (Supporting Information Fig. S1 and Table 2) were done with Mothur version 1.35.1 (Schloss *et al.*, 2009). To compare the two datasets (V4 and V9), twenty-seven OSD stations were selected and subsampled using the lowest number of reads (202,710) at a given station (station 49 for V4, Table 1). Sequences were then filtered by removing sequences shorter than 90 bases for the V9 region and shorter than 170 bp for the V4 region or containing ambiguities (N). Reads were aligned on SILVA release 119 seed alignment (Pruesse *et al.*, 2007) corrected by hand using the Geneious software version 7.1.7 (Kearse *et al.*, 2012). Gaps at the beginning and at the end of the alignment were deleted. Alignments were filtered by removing positions containing only insertions. Chimeras were removed using Uchime version 4.2.40 (Edgar *et al.*, 2011) as implemented in Mothur. The sequences were first pre-clustered and singletons were eliminated. After distance matrix calculation, reads were clustered using the Nearest Neighbor method and OTUs were built at 97% similarity (Supporting Information). OTUs were assigned using Wang approach (Wang *et al.*, 2007) which is based on the calculation of Bayesian probabilities using kmer (8 bp by default) comparisons between dataset and database sequences. This method is complemented by a bootstrap step to confirm the taxonomical classification: assignation supported at a level lower than 80% were not taken into account.

The reference database was a revised version (4.2 https://figshare.com/articles/PR2_rRNA_gene_database/3803709/2) of the PR^2 database (Guillou *et al.*, 2013) for which the Chlorophyta sequences had been checked against the latest taxonomy (Tragin *et al.*, 2016). The PR^2 database considers eight taxonomic levels (from Kingdom to Species). OTUs are considered as assigned when their lowest taxonomic level

(Level 8, 'Species') differs from 'unclassified'. Note that this level may not correspond to a single validly described species but may group several taxa (e.g., Crustomastigaceae_X_sp., see details in Guillou *et al.*, 2013). Several OTUs can be assigned to the same taxonomy if they, for example, correspond to the same 'Species'. OTUs assigned to Chlorophyta were BLASTed against GenBank using 97% identity and 0.001 e-value cutoff thresholds (Supporting Information) and OTUs for which the best hit was not a Chlorophyta were removed from further analysis.

Diversity analyses were conducted using the R software version 3.0.2 (http://www.R-project.org/). The Vegan package (https://cran.r-project.org/web/packages/vegan/) was used to compute rarefaction curves and Simpson diversity indexes (*D*, Simpson, 1949) at each station.

$$D = 1 - \sum_{i=1}^{S} p_i^2$$

S is the number of species in the sample and p_i the proportion of species *i*. D is relatively little influenced by sample size and does not require any hypothesis on the species distribution. D depends on the number of OTUs recorded as well as the distribution of the sequences within the OTUs. For example, in a sample with two species recorded ($S = 2$), D will be larger if the two species are equally distributed ($p_1 = p_2 = 0.5$) than if one is dominant ($p_1 = 0.9$ and $p_2 = 0.1$).

Descriptive statistics for V4 versus V9 were computed using the R functions *summary* and *sd* (Table 3). A non-parametric rank Wilcoxon test (Wilcoxon, 1945) was performed to compare both results using the *wilcoxon.test* function from the R package stats. Since the V4 and V9 regions were sequenced from the same DNA sample, the paired option was set as true. This test did not return exact *P* values for sample in which null or ex-aequo values occurred.

The matrixes of the V4 and V9 relative contribution for photosynthetic groups, Chlorophyta and Mamiellophyceae at each station were compared by the geometry-based procrustean method using the *procrustes* and *protest* functions of Vegan. The distance matrix between stations based on the relative contribution at the Class level for Chlorophyta and genus level for Mamiellophyceae were computed using the Bray–Curtis distance and clustered using the hierarchical clustering 'complete' method. Bray–Curtis matrix distance was also computed for the global community (all OTUs considered) and the communities were represented in a two-dimensional space with the iterative ordination method nonparametric multi-dimensional scaling (NMDS) plot using the *metaMDS* function of Vegan. Hierarchical clustering was computed on the same Bray–Curtis distance matrix. The clustering dendrograms were cut with the *rect.hclust* function from the R stats package at a height $h = 0.9$. Resulting groups were traced on the NMDS plot. Available OSD metadata was projected onto the NMDS plots using the *envfit* function from the Vegan with the p.max option set as 0.95. All supplementary data and scripts are available at https://figshare.com/articles/Comparison_of_coastal_phytoplankton_composition_estimated_from_the_V4_and_V9_regions_of_18S_rRNA_gene_with_a_focus_on_Chlorophyta/4252646.

Acknowledgements

MT was supported by a PhD fellowship from the Université Pierre et Marie Curie and the Région Bretagne. We would like also to thank the Ocean Sampling Day consortium (supported by EU project MicroB3/FP7–287589) for the sample collection and DNA extraction and the Biomolecular Thematic Centre (MoBiLab – Molecular Biodiversity Laboratory) of the ESFRI LifeWatch-Italia, which carried out the Illumina sequencing. We extend our warm thanks to Fabrice Not for his critical reading of the article. The authors declare no conflict of interest.

References

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., Huse, S.M., and Langsley, G. (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.

Baldauf, S.L. (2008) An overview of the phylogeny and diversity of eukaryotes. *J Syst Evol* **46:** 263–273.

Behnke, A., Engel, M., Christen, R., Nebel, M., Klein, R.R., and Stoeck, T. (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. *Environ Microbiol* **13:** 340–349.

Berry, M.A., White, J.D., Davis, T.W., Jain, S., Johengen, T.H., and Dick, G.J. (2017) Are oligotypes meaningful ecological and phylogenetic units? A case study of *Microcystis* in freshwater lakes. *Front Microbiol* **8:** 365.

Caron, D.A., Countway, P.D., Savai, P., Gast, R.J., Schnetzer, A., and Moorthi, S.D. (2009) Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl Environ Microbiol* **75:** 5797–5808.

Cerino, F., Bernardi Aubry, F., Coppola, J., La Ferla, R., Maimone, G., Socal, G., and Totti, C. (2012) Spatial and temporal variability of pico-, nano- and microphytoplankton in the offshore waters of the southern Adriatic Sea (Mediterranean Sea). *Cont Shelf Res* **44:** 94–105.

Chenuil, A., and Anne, C. (2006) Choosing the right molecular genetic markers for studying biodiversity: from molecular evolution to practical aspects. *Genetica* **127:** 101–120.

Clayton, S., Lin, Y.-C., Follows, M.J., and Worden, A.Z. (2017) Co-existence of distinct *Ostreococcus* ecotypes at an oceanic front. *Limnol Oceanogr* **62:** 75–88.

Comeau, A.M., Li, W.K.W., Tremblay, J.-É., Carmack, E.C., Lovejoy, C., and Gilbert, J.A. (2011) Arctic Ocean microbial community structure before and after the 2007 Record Sea Ice Minimum. *PLoS One* **6:** e27492.

Decelle, J., Romac, S., Sasaki, E., Not, F., Mahé, F., and Lovejoy, C. (2014) Intracellular diversity of the V4 and V9 Regions of the 18S rRNA in marine protists (Radiolarians) assessed by high-throughput sequencing. *PLoS One* **9:** e104297.

van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014) Ten years of next-generation sequencing technology. *Trends Genet* **30:** 418–426.

Dunthorn, M., Klier, J., Bunge, J., and Stoeck, T. (2012) Comparing the hyper-variable V4 and V9 regions of the small subunit rDNA for assessment of ciliate environmental diversity. *J Eukaryot Microbiol* **59:** 185–187.

Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27:** 2194–2200.

Edvardsen, B., Egge, E.S., and Vaulot, D. (2016) Diversity and distribution of haptophytes revealed by environmental sequencing and metabarcoding – a review. *Perspect Phycol* **3:** 77–91.

Egge, E., Bittner, L., Andersen, T., Audic, S., de Vargas, C., Edvardsen, B., and Lin, S. (2013) 454 pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: a test for marine Haptophytes. *PLoS One* **8:** e74371.

Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., and Sogin, M.L. (2014) Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *ISME J* **9:** 968–979.

Ferrera, I., Giner, C.R., Reñé, A., Camp, J., Massana, R., Gasol, J.M., and Garcés, E. (2016) Evaluation of alternative high-throughput sequencing methodologies for the monitoring of marine picoplanktonic biodiversity based on rRNA gene amplicons. *Front Mar Sci* **3:** 147.

Giner, C.R., Forn, I., Romac, S., Logares, R., de Vargas, C., and Massana, R. (2016) Environmental sequencing provides reasonable estimates of the relative abundance of specific picoeukaryotes. *Appl Environ Microbiol* **82:** 4757–4766.

Gómez, F. (2012) A quantitative review of the lifestyle, habitat and trophic diversity of dinoflagellates (Dinoflagellata, Alveolata). *Syst Biodivers* **10:** 267–275.

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 Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Res* **41:** 597–604.

Hamilton, A.K., Lovejoy, C., Galand, P.E., and Ingram, R.G. (2008) Water masses and biogeography of picoeukaryote assemblages in a cold hydrographically complex system. *Limnol Oceanogr* **53:** 922–935.

Hu, S., Campbell, V., Connell, P., Gellen, A.G., Liu, Z., Terrado, R., and Caron, D.A. (2016) Protistan diversity and activity inferred from RNA and DNA at a coastal ocean site in the eastern North Pacific. *FEMS Microb Ecol* **92:** 1–39.

Hu, S.K., Liu, Z., Lie, A.A.Y., Countway, P.D., Kim, D.Y., Jones, A.C., *et al.* (2015) Estimating Protistan diversity using high-throughput sequencing. *J Eukaryot Microbiol* **62:** 688–693.

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.

Kembel, S.W. (2009) Disentangling niche and neutral influences on community assembly: assessing the performance of community phylogenetic structure tests. *Ecol Lett* **12:** 949–960.

Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., and Lehmann, K. (2015) The Ocean Sampling Day consortium. *Gigascience* **4:** 27.

Lopes dos Santos, A., Gourvil, P., Tragin, M., Noël, M.-H., Decelle, J., Romac, S., and Vaulot, D. (2017) Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. *ISME J* **11:** 512–528.

López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C., and Moreira, D. (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409:** 603–607.

Majaneva, M., Hyytiäinen, K., Varvio, S.L., Nagai, S., Blomster, J., and Berg, G. (2015) Bioinformatic amplicon read processing strategies strongly affect eukaryotic diversity and the taxonomic composition of communities. *PLoS One* **10:** e0130035.

Marquardt, M., Vader, A., Stübner, E.I., Reigstad, M., and Gabrielsen, T.M. (2016) Strong seasonality of marine microbial eukaryotes in a High-Arctic Fjord (Isfjorden, in West Spitsbergen, Norway). *Appl Environ Microbiol* **82:** 1868–1880.

Massana, R., del Campo, J., Sieracki, M.E., Audic, S., and Logares, R. (2014) Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J* **8:** 854–866.

Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., *et al.* (2015) Marine protist diversity in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ Microbiol* **17:** 4035–4049.

Massana, R., and Pedrós-Alió, C. (2008) Unveiling new microbial eukaryotes in the surface ocean. *Curr Opin Microbiol* **11:** 213–218.

Matsen, F.A., Kodner, R.B., and Armbrust, E.V. (2010) pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics* **11:** 538.

Monier, A., Worden, A.Z., and Richards, T.A. (2016) Phylogenetic diversity and biogeography of the Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environ Microbiol Rep* **8:** 461–469.

Moon-van der Staay, S.Y., De Wachter, R., and Vaulot, D. (2001) Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409:** 607–610.

Not, F., Latasa, M., Marie, D., Cariou, T., Vaulot, D., and 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., del Campo, J., Balagué, V., de Vargas, C., Massana, R., and Earley, R.L. (2009) New insights into the diversity of marine picoeukaryotes. *PLoS One* **4:** e7143.

Pernice, M.C., Logares, R., Guillou, L., Massana, R., and Badger, J.H. (2013) General patterns of diversity in major marine microeukaryote lineages. *PLoS One* **8:** e57170.

Piredda, R., Tomasino, M.P., D'erchia, A.M., Manzari, C., Pesole, G., Montresor, M., *et al.* (2017) Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term Ecological Research site. *FEMS Microbiol Ecol* **93:** fiw200.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glockner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35:** 7188–7196.

Revelante, N., and Gilmartin, M. (1976) Temporal succession of phytoplankton in the northern Adriatic. *Netherlands J Sea Res* **10:** 377–396.

Rii, Y.M., Duhamel, S., Bidigare, R.R., Karl, D.M., Repeta, D.J., and Church, M.J. (2016) Diversity and productivity of photosynthetic picoeukaryotes in biogeochemically distinct regions of the South East Pacific Ocean. *Limnol Oceanogr* **61:** 806–824.

Sanger, F., and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* **94:** 441–448.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* **75:** 7537–7541.

Simmons, M.P., Sudek, S., Monier, A., Limardo, A.J., Jimenez, V., and Perle, C.R. (2016) Abundance and biogeography of picoprasinophyte ecotypes and other phytoplankton in the Eastern North Pacific Ocean. *Appl Environ Microbiol* **82:** 1693–1705.

Simpson, E.H. (1949) Measurement of diversity. *Nature* **163:** 688–688.

Stegen, J.C., Lin, X., Konopka, A.E., and Fredrickson, J.K. (2012) Stochastic and deterministic assembly processes in subsurface microbial communities. *ISME J* **6:** 1653–1664.

Stoeck, T., Bass, DD., Nebel, M., Christen, R., Jones, M., Breiner, H.-W., and Richards, T. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol Ecol* **19:** 21–31.

Tragin, M., Lopes dos Santos, A., Christen, R., and Vaulot, D. (2016) Diversity and ecology of green microalgae in marine systems: an overview based on 18S rRNA gene sequences. *Perspect Phycol* **3:** 141–154.

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:** 1261605–1261605.

Viprey, M., Guillou, L., Ferréol, M., and 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, G.M., Tiedje, J.M., and Cole, J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73:** 5261–5267.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bull* **1:** 80.

Supporting information

Additional supporting Information may be found in the online version of this article at the publisher's web-site. All supplementary data are available at https://figshare.com/articles/Comparison_of_coastal_phytoplankton_composition_estimated_from_the_V4_and_V9_regions_of_18S_rRNA_gene_with_a_focus_on_Chlorophyta/4252646:

Fig. S1. A. Bioinformatics pipeline use to build and assigned OTUs from V4 and V9 datasets. Reference alignment was SILVA seed release 119. The Chlorophyta curated PR^2 database (Tragin *et al.*, 2016) was used as taxonomic reference. The number of sequences at each step appears in Table 2.

Fig. S2. A. Rarefaction curves. B. Rank abundance distribution. *x*-axis represents OTUs by decreasing number of sequences.

Fig. S3. Rarefaction curves per station A. V4 and B. V9.

Fig. S4. A and B. Non-metric multi-dimensional scaling (NMDS) representation of communities based on lowest taxonomic level (OTUs) for V4 (A) and V9 (B). The dissimilarity matrix was computed using Bray–Curtis distance. C and D. Hierarchical cluster analysis based on the Bray–Curtis matrix for V4 (C) and V9 (D). Stations in panels A and B were grouped together based on clusters from panels C and D using a fixed threshold (0.9).

Fig. S5. Correlation between V4 and V9 relative contribution to photosynthetic metabarcodes in major photosynthetic phyla. A, Ochrophyta; B, Chlorophyta; C, Haptophyta; D, Cryptophyta.

Fig. S6. Correlation between V4 and V9 relative contribution of the four major Ochrophyta Classes. A, Bacillariophyta; B, Dictyochophyceae; C, Chrysophyceae–Synurophyceae; D, Pelagophyceae.

Fig. S7. Percentage of genera from photosynthetic groups found either only in V4 (blue), or only in V9 (red), or in both datasets (grey). Only taxonomically valid genera and only Classes with at least five genera were taken into account. Numbers below each group indicate the total number of genera recorded.

Fig. S8. Correlation between V4 and V9 relative contribution to photosynthetic metabarcodes for major Chlorophyta classes. A, Mamiellophyceae; B, Trebouxiophyceae; C, Chlorodendrophyceae (OSD14 is not represented on the scatter plot with 65% and 60% for V4 and V9, respectively); D, Pyramimonadales; E, Ulvophyceae; F, Pseudoscourfieldiales.

Fig. S9. Heatmap of differences between V9 and V4 (V9–V4) relative contribution: A, Chlorophyta classes; B, Mamiellophyceae genera. The colors correspond to the difference from −50% (–0.5) to +50% (0.5).

Fig. S10. Correlation between V4 and V9 relative contribution to Chlorophyta metabarcodes for major Mamiellophyceae genera. A, *Micromonas*; B, *Mamiella*; C, *Ostreococcus*; D, *Bathycoccus*.

Data S1. Mothur script for sequence analysis.

Data S2. Fasta file of Chlorophyta OTUs for V4.

Data S3. Fasta file of Chlorophyta OTUs for V9.

Data S4. Chlorophyta OTUs for V4 with assignation and read abundance at the different stations (Excel file).

Data S5. Chlorophyta OTUs for V9 with assignation and read abundance at the different stations (Excel file).

Data S6. Top 10 BLAST hits against Genbank nr database for Chlorophyta V4 OTUs. Red lines correspond to OTUs badly assigned to non-Chlorophyta and green corresponds to OTUs badly assigned to another Chlorophyta representative.

Data S7. Top 10 BLAST hits against Genbank nr database for Chlorophyta V9 OTUs. Red lines correspond to OTUs badly assigned to non-Chlorophyta and green lines corresponds to OTUs badly assigned to another Chlorophyta representative.