



**HAL**  
open science

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

Margot Tragin, Adriana Zingone, Daniel Vaultot

► **To cite this version:**

Margot Tragin, Adriana Zingone, Daniel Vaultot. Comparison of coastal phytoplankton composition estimated from the V4 and V9 regions of 18S rRNA gene with a focus on photosynthetic groups and especially Chlorophyta. *Environmental Microbiology*, 2017, 20 (2), pp.506-520. 10.1111/1462-2920.13952 . hal-01628365

**HAL Id: hal-01628365**

**<https://hal.sorbonne-universite.fr/hal-01628365>**

Submitted on 3 Nov 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24

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

Margot Tragin<sup>1</sup>, Adriana Zingone<sup>2</sup>, Daniel Vaultot<sup>1\*</sup>

<sup>1</sup> Sorbonne Universités, UPMC Univ Paris 06, UMR 7144, CNRS, Station Biologique, Place Georges  
Teissier, 29680 Roscoff, France

<sup>2</sup> Department of Integrative Marine Ecology, Stazione Zoologica Anton Dohrn, Villa Comunale,  
Naples, Italy

Revised version – July 17, 2017

For *Environmental Microbiology*

**Keywords**

18S rRNA gene, V4 region, V9 region, diversity, ecology, marine systems

---

\* Corresponding author : vaultot@sb-roscoff.fr

## 25 Acknowledgments

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

## 32 Originality-Significance Statement

33 Environmental barcoding is now widely used to assess eukaryotic protist communities. The  
34 two most common markers correspond to the V4 and V9 regions of the 18S rRNA gene. This  
35 paper provides a detailed comparison of the V4 and V9 for the assessment of community  
36 composition on a large set of marine samples from the Ocean Sampling Day project using  
37 rigorously similar sequencing and analysis pipelines. We demonstrate that, by and large, these  
38 two markers provide similar community composition and point out some cases where this may  
39 not be true. These data should be very useful to anyone performing metabarcoding of marine  
40 protist communities.

41

42

43 [Abstract](#)

44           We compared the composition of eukaryotic communities using two genetic markers (18S  
45 rRNA V4 and V9 regions) at 27 sites sampled during Ocean Sampling Day 2014, with a focus on  
46 photosynthetic groups and, more specifically green algae (Chlorophyta). Globally, the V4 and V9  
47 regions of the 18S rRNA gene provided similar images of alpha diversity and ecological patterns.  
48 However, V9 provided 20% more OTUs built at 97% identity than V4 and 39% and 56% of the genera  
49 were found only in one dataset, respectively V4 and V9. For photosynthetic groups, V4 and V9  
50 performed equally well to describe global communities at different taxonomic levels from the division  
51 to the genus and provided similar Chlorophyta distribution patterns. However, at lower taxonomic level,  
52 the V9 dataset failed for example to describe the diversity of Dolichomastigales (Chlorophyta,  
53 Mamiellophyceae) emphasizing the lack of V9 sequences for this group and the importance of the  
54 reference database for metabarcode analysis. We conclude that in order to address specific questions  
55 regarding specific groups (e.g. a given genus), it is necessary to choose the marker based not only on  
56 the genetic divergence within this group but also on the existence of reference sequences in databases.

57

58

## 59 Introduction

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

79 The performance of the 18S RNA hypervariable regions as barcodes and the interpretation of  
80 results produced remains a matter of debate. Hu *et al.* (2015) showed that the V4 region provides an  
81 image of diversity similar to that obtained from the entire 18S rRNA gene. The choice between V4 and  
82 V9 depends on the taxonomic levels as well as the specific groups targeted. It is necessary to make  
83 detailed comparisons of genetic distances for each targeted region between and within the groups of  
84 interest (Dunthorn *et al.*, 2012; Pernice *et al.*, 2013) and to determine whether reference sequences are  
85 available for the group of interest in the target region (Tragin *et al.*, 2016). The sequencing platform  
86 may also have some impacts: using the 454 technology, Behnke *et al.* (2011) showed that the sequencing  
87 error rate was taxon dependent, but V4 error rates were higher than for V9. Analysis of mock  
88 communities have highlighted possible biases in molecular methods such as the generation of artificial  
89 diversity (Egge *et al.*, 2013). The primers used may also produce a bias against groups whose target  
90 fragments are not amplified. For example, some widely used V4 primers miss Haptophyta and  
91 Foraminifera, which are important groups of the marine plankton (Massana *et al.*, 2015). Finally  
92 bioinformatics steps such as raw sequence filtering based on sequence quality and length, clustering  
93 algorithm and threshold to regroup sequences into Operational Taxonomic Units (OTUs) may influence  
94 the final results (Majaneva *et al.*, 2015).

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

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

121

## 122 Material and Methods

123 Water samples were collected from 0-2 meter depth at 27 stations in the world ocean (Fig. 1 and  
124 Table 1). Metadata (Temperature, Salinity, Nitrates, Phosphates, Silicates and Chlorophyll *a*) are  
125 available at <https://github.com/MicroB3-IS/osd-analysis/wiki/Guide-to-OSD-2014-data>. Samples were  
126 filtered on 0.8 µm pore size polycarbonate membranes without prefiltration and flash frozen at -80°C or  
127 in liquid nitrogen. DNA was extracted using the Power Water isolation kit (MoBio, Carlsbad, CA, USA)  
128 following the manufacturer instructions. The V4 region was amplified using modified universal primer  
129 (Piredda *et al.*, 2017): V4\_18SNext.For primer (5' CCA GCA SCY GCG GTA ATT CC 3') and  
130 V4\_18SNext.Rev primer (5' ACT TTC GTT CTT GAT YRA TGA 3'). The V9 region was amplified  
131 using modified universal primer (Piredda *et al.*, 2017): V9\_18SNext.For (5' TTG TAC ACA CCG CCC  
132 GTC GC 3') and V9\_18SNexr.Rev (5' CC TTC YGC AGG TTC ACC TAC 3'). The library preparation  
133 was based on a modified version of the Illumina Nextera's protocol (Nextera DNA sample preparation  
134 guide, Illumina) and sequencing was done on an Illumina Miseq (NE08 Ocean Sampling Day protocols:  
135 [https://github.com/MicroB3-IS/osd-analysis/wiki/Guide-to-OSD-2014-data#analysis-of-workable-18s-](https://github.com/MicroB3-IS/osd-analysis/wiki/Guide-to-OSD-2014-data#analysis-of-workable-18s-rdna-datasets-sequenced-by-lifewatch-italy)  
136 [rdna-datasets-sequenced-by-lifewatch-italy](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: 2x250 paired  
137 end sequencing using MiSeq Reagent kit v3 and V9 region: 2x150 paired end sequencing using MiSeq  
138 Reagent kit v2) was done by the Laboratory of Molecular Biodiversity (MoBiLab) of LifeWatch-Italy.  
139 R1 and R2 were filtered based on quality and length and assembled by the OSD consortium which  
140 provided the so-called "workable" fasta files ([https://owncloud.mpi-](https://owncloud.mpi-bremen.de/index.php/s/RDB4Jo0PAayg3qx?path=/2014/silva-ngs/18s/lifewatch/)  
141 [bremen.de/index.php/s/RDB4Jo0PAayg3qx?path=/2014/silva-ngs/18s/lifewatch/](https://owncloud.mpi-bremen.de/index.php/s/RDB4Jo0PAayg3qx?path=/2014/silva-ngs/18s/lifewatch/)).  
142 All subsequent sequence analyses (Supplementary Data and Fig. S1, Table 2) were done with Mothur  
143 version 1.35.1 (Schloss *et al.*, 2009). To compare the two datasets (V4 and V9), twenty-seven OSD  
144 stations were selected and subsampled using the lowest number of reads (202,710) at a given station  
145 (station 49 for V4, Table 1). Sequences were then filtered by removing sequences shorter than 90 bases  
146 for the V9 region and shorter than 170 bp for the V4 region or containing ambiguities (N). Reads were  
147 aligned on SILVA release 119 seed alignment (Pruesse *et al.*, 2007) corrected by hand using the  
148 Geneious software version 7.1.7 (Kearse *et al.*, 2012). Gaps at the beginning and at the end of the  
149 alignment were deleted. Alignments were filtered by removing positions containing only insertions.  
150 Chimeras were removed using Uchime version 4.2.40 (Edgar *et al.*, 2011) as implemented in Mothur.  
151 The sequences were first pre-clustered and singletons were eliminated. After distance matrix calculation,  
152 reads were clustered using the Nearest Neighbor method and OTUs were built at 97% similarity  
153 (Supplementary Data). OTUs were assigned using Wang approach (Wang *et al.*, 2007) which is based  
154 on the calculation of Bayesian probabilities using kmer (8 bp by default) comparisons between dataset  
155 and database sequences. This method is complemented by a bootstrap step to confirm the taxonomical  
156 classification: assignation supported at a level lower than 80% were not taken into account.

157 The reference database was a revised version (4.2  
 158 [https://figshare.com/articles/PR2\\_rRNA\\_gene\\_database/3803709/2](https://figshare.com/articles/PR2_rRNA_gene_database/3803709/2)) of the PR<sup>2</sup> database (Guillou *et al.*,  
 159 2013) for which the Chlorophyta sequences had been checked against the latest taxonomy (Tragin *et al.*,  
 160 2016). The PR<sup>2</sup> database considers 8 taxonomic levels (from Kingdom to Species). OTUs are considered  
 161 as assigned when their lowest taxonomic level (Level 8, "Species") differs from "unclassified". Note  
 162 that this level may not correspond to a single validly described species but may group several taxa (for  
 163 example Crustomastigaceae\_X\_sp., see details in Guillou *et al.*, 2013). Several OTUs can be assigned  
 164 to the same taxonomy if they, for example, correspond to the same "Species". OTUs assigned to  
 165 Chlorophyta were BLASTed against GenBank using 97% identity and 0.001 e-value cutoff thresholds  
 166 (Supplementary Data) and OTUs for which the best hit was not a Chlorophyta were removed from  
 167 further analysis.

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

$$171 \quad D = 1 - \sum_{i=1}^S p_i^2$$

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

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

182 The matrixes of the V4 and V9 relative contribution for photosynthetic groups, Chlorophyta and  
 183 Mamiellophyceae at each station were compared by the geometry-based procrustean method using the  
 184 *procrustes* and *protest* functions of Vegan. The distance matrix between stations based on the relative  
 185 contribution at the Class level for Chlorophyta and genus level for Mamiellophyceae were computed  
 186 using the Bray-Curtis distance and clustered using the hierarchical clustering "complete" method. Bray-  
 187 Curtis matrix distance was also computed for the global community (all OTUs considered) and the  
 188 communities were represented in a 2-dimensional space with the iterative ordination method  
 189 Nonparametric Multi-Dimensional Scaling (NMDS) plot using the *metaMDS* function of Vegan.



190 Hierarchical clustering was computed on the same Bray-Curtis distance matrix. The clustering  
191 dendrograms were cut with the *rect.hclust* function from the R stats package at a height  $h=0.9$ . Resulting  
192 groups were traced on the NMDS plot. Available OSD metadata were projected onto the NMDS plots  
193 using the *envfit* function from the Vegan with the *p.max* option set as 0.95.

## 194 Results

### 195 Global eukaryotic community

196 Twenty-seven stations were selected for which both V4 and V9 metabarcodes were obtained.  
197 The two datasets were subsampled in order to process the same number of reads per station. After  
198 subsampling, the V4 and V9 datasets were reduced to 62% and 48% of their original size, respectively  
199 (Table 2). The number of unique sequences (Table 2) was higher for V4 (around 1,400,000) than for V9  
200 (around 900,000). After filtering based on length and ambiguities, twice more reads were obtained for  
201 V4 than for V9 (Table 2). About forty times more chimeras were found for V4 than for V9 (about 7,500  
202 against 170). Following taxonomic assignment, all eukaryotic groups were retained, not just protists.

203 Rarefaction curves computed for the global datasets as well as for each station (Fig. S2A and  
204 S3) reached saturation, suggesting that the sequencing effort was sufficient. Global maximum richness  
205 varied between the datasets: 16,383 OTUs (4,311 distinct assignments) were obtained for V9 against  
206 13,169 OTUs (3,412 distinct assignments) for V4. The two datasets yielded similar rank abundance  
207 curves (Fig. S2 B) although V9 had larger OTUs as attested by the fact that the curve for V9 was above  
208 that for V4. The size of the largest OTU was equivalent (around 180,000 sequences).

209 The number of OTUs per station varied from 500 to about 3,000 with respective averages of  
210 1,200 and 1,600 for V4 and V9 (Fig. 2A and Table 3). Although a positive correlation was found  
211 between the number of OTUs for V4 and V9 per station ( $R^2=0.99$ , Fig. 2A), the number of OTUs per  
212 stations was higher for V9 than for V4 (slope=1.27, Fig. 2A) and this difference was confirmed by a  
213 Wilcoxon test (Table 3). The comparison of Simpson's diversity index per station for the two datasets  
214 (Fig. 2B and Table 3) showed that V4 and V9 diversity values were similar for large values between 0.9  
215 and 1, irrespective of the OTU richness. For lower values of the Simpson's index (0.6 to 0.9), it was  
216 higher for V9 than V4 except at station OSD30 in the Gulf of Finland (Fig. 2B). At the latter station,  
217 one specific metazoan OTU (assigned to copepods and corresponding to 105,202 reads) was dominating  
218 the V9 reads but this OTU did not dominate the V4 reads. If this copepod OTU is not taken into account  
219 (grey star on Fig. 2B), the V4 and V9 datasets have a similar alpha diversity (0.91 and 0.95 respectively).  
220 The number of genera (assignments without \_X) found in the OSD datasets was equal to 3,669, among  
221 which 39% (for V4) and 56% (for V9) were recovered only in one dataset. On average, four OTUs were  
222 assigned to the same genus and the maximum number of OTUs per genus reached 128 for V4 and 187  
223 for V9. 98 % of genera found only in one dataset were represented by less than 10 OTUs.

224 Non-parametric multidimensional scaling analysis (NMDS, Fig. S4A-B) and hierarchical  
225 clustering (Fig. S4C-D) were used to visualize the V4 and V9 communities based on OTUs (final stress  
226 values were respectively 0.187 and 0.195) using Bray-Curtis dissimilarity. Many stations grouped  
227 together in a similar way for both V4 and V9, some according to their geographic location (Fig. S4 and  
228 Fig. 1), as in the case of the Mediterranean Sea (OSD14, 22, 49, 76, 77, 99) or of the subtropical Atlantic

229 coast of the United States (OSD39, 60 and 143). Both V4 and V9 communities were structured by the  
 230 same combination of environmental parameters with opposite gradients of nitrates, phosphates and  
 231 chlorophyll on one side vs. silicates, temperature and salinity on the other side (Fig. S4A-B).

### 232 Photosynthetic groups

233 We next focused on photosynthetic groups for which taxonomic assignment relies on recently  
 234 validated reference databases (Edwardsen *et al.*, 2016; Tragin *et al.*, 2016). Dinophyceae were excluded  
 235 from the analysis since about 50% of the species are not photosynthetic (Gómez, 2012). The percent of  
 236 reads assigned to photosynthetic groups was quite similar between datasets: 28.6% vs. 25.9 % for V4  
 237 and V9, respectively. The four major photosynthetic groups were Ochrophyta (mostly diatoms),  
 238 Chlorophyta (green algae), Haptophyta and Cryptophyta (Fig. 3A). The Rhodophyta, Cercozoa  
 239 (Chlorarachniophyta) and Discoba (Euglenales) represented less than 1.5% of the photosynthetic groups  
 240 in the two datasets (Fig. 3A). Procrustean analysis suggested that the relative contribution of  
 241 photosynthetic groups per station was similar between V4 and V9 ( $m^2=0.17$  and  $r=0.91$ ). The number  
 242 of OTUs assigned to Ochrophyta was quite similar in the two datasets (1215 and 1250 in V4 and V9,  
 243 respectively). In contrast, the number of V9 OTUs was almost twice that of V4 for Chlorophyta and  
 244 Cryptophyta and three times for Haptophyta, but average OTUs size was similar (377, 64, 91 and 573,  
 245 100, 241 in V4 versus V9). For these three photosynthetic groups, average pairwise identity between  
 246 the OTUs reference sequences was higher for V4 than V9 (76 vs. 72% for Chlorophyta, 84 vs. 76% for  
 247 Cryptophyta and 86 vs. 77% for Haptophyta), indicating that V4 has lower genetic variability for these  
 248 groups, and therefore is less discriminating.

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

259 Among Ochrophyta, diatoms (Bacillariophyta) largely dominated, followed by  
 260 Dictyochophyceae and Chrysophyceae-Synurophyceae (Fig. 3B). Diatom relative contribution to  
 261 photosynthetic metabarcodes per stations was around 50% on average (Table 3) and varied between  
 262 15% (OSD30, 146) to 90% (OSD159, 60). Diatom contribution was statistically similar between V4 and  
 263 V9 (Table 3). Dictyochophyceae relative contribution was below 10% except for five stations (OSD22,

264 149, 150, 152 and 72), where it reached 35% of photosynthetic reads (Fig. S6B). Dictyochophyceae  
 265 contribution was slightly higher with V4 at these five stations. Chrysophyceae-Synurophyceae relative  
 266 contribution was below 10% except for OSD76 (25%, Fig. S6C and Table 3) and V4 and V9 contribution  
 267 were similar except at OSD30, where V9 was higher and OSD49 and 76 where V4 was higher (Fig.  
 268 S6C). Pelagophyceae relative contribution was below 10% at individual stations but V4 and V9 were  
 269 similar (Fig. S6D and Table 3). Chlorophyta were dominated by Mamiellophyceae, followed by  
 270 Trebouxiophyceae, Chlorodendrophyceae and Pyramimonadales (Fig. 3C). Trebouxiophyceae and  
 271 Chlorodendrophyceae were more represented in V4 while Mamiellophyceae and Pyramimonadales  
 272 were more represented in V9 (Fig. 3C). Other photosynthetic groups remained similar between the V4  
 273 and V9 datasets. Among Haptophyta, Prymnesiophyceae were largely dominating but two  
 274 environmental clades, HAP3 and HAP4 (Edvardsen *et al.*, 2016), were also recovered (Fig. 3D). For  
 275 photosynthetic groups, the percentage of genera found either in only one dataset (V4 or V9) or in both  
 276 was class dependent, but globally 50% of the genera were recovered in both datasets (Fig. S7). Within  
 277 Ochrophyta, more genera were found using V9 in 5 out of 8 classes, but this was not the case for  
 278 Bacillariophyta and Xanthophyceae for which more genera were recovered with V4. Raphidophyceae  
 279 genera were almost all recovered in both V4 and V9 (Fig. S7). More red algae genera (Florideophyceae  
 280 and Bangiophyceae) were recovered with V9. For Haptophyta, Cryptophyta and Chlorarachniophyceae  
 281 most genera were found with both markers (Fig. S7).

## 282 Chlorophyta classes

283 The relative contributions of the 6 major Chlorophyta groups (Mamiellophyceae,  
 284 Trebouxiophyceae, Chlorodendrophyceae, Pyramimonadales, Ulvophyceae and Pseudoscourfieldiales)  
 285 in V4 and V9 were similar at most stations (Fig. 4A and Fig. S8) as supported by a procrustean  
 286 comparison ( $m^2=0.027$  and  $r=0.98$ ) but individual group contributions were not similar except for  
 287 Mamiellophyceae (Table 3). Mamiellophyceae were dominant at most stations, but the four stations  
 288 located in the Adriatic Sea (OSD49, 76, 77 and 99) shared a specific pattern with high contributions of  
 289 Pseudoscourfieldiales and Chlorodendrophyceae in both V4 and V9 datasets (Fig. 4A). Stations OSD30,  
 290 54, 55, 141, all located in North Atlantic coastal waters presented differences in Chlorophyta class  
 291 contribution recovered with V4 and V9 (Fig. 4A and Fig. S9A). For the first three, the Mamiellophyceae  
 292 contribution in V9 was partially replaced in V4 by classes from “core chlorophytes” such as  
 293 Chlorodendrophyceae and/or Trebouxiophyceae. At OSD141, prasinophytes clade VII were only  
 294 recovered with V9, while Chlorophyceae (*Chlamydomonas* sp.) were only recovered with V4 (Fig.  
 295 S9A). BLAST analysis and alignment of Chlorophyta OTUs (data not shown) revealed that the V9  
 296 region of some *Chlamydomonas* is very similar to that of prasinophytes clade VII A5 (Lopes dos Santos  
 297 *et al.*, 2016). Interestingly, the number of reads recovered in V4 and V9 for these 2 assignments (i.e.  
 298 *Chlamydomonas* sp. for V4 and prasinophytes clade VII A5 for V9) was similar (51 versus 47 reads,  
 299 respectively).

300 In general, Chlorophyta OTUs were well assigned by the Wang approach implemented in the  
 301 Mothur software (Wang et al., 2007) compared to the results of BLAST (Supplementary data 6 and 7).  
 302 However, some V9 reads initially assigned as Chlorophyta by the Wang approach hit bacterial sequences  
 303 and were not considered any further. Some V9 Chlorophyta OTUs shared 100% identity with several  
 304 different Chlorophyta genera with (mostly in the UTC clade, i.e. Ulvophyceae, Trebouxiophyceae and  
 305 Chlorophyceae) suggesting that the V9 region might not have the appropriate resolution to investigate  
 306 UTC clade diversity. A number of genera within the UTC clade were only recovered with one marker  
 307 in contrast to the Mamiellophyceae and Pyramimonadales for which almost all genera were recovered  
 308 in both datasets (Fig. S7).

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

#### 316 Mamiellophyceae genera

317 Mamiellophyceae that dominated at most OSD stations were further investigated at the genus  
 318 level. Nine genera of Mamiellophyceae were found in the OSD datasets, seven of which were found in  
 319 both datasets, one only in V4, assigned to RCC391, and one only in V9, assigned to *Monomastix*. The  
 320 latter is a freshwater genus and the OTUs assigned to it were badly assigned (BLAST analysis showed  
 321 100% identity with sequences of several land plants genera, see Supplementary data), while the RCC391  
 322 genus has eight references sequences for V4 against only one for V9. *Micromonas* and *Ostreococcus*  
 323 were the two dominant genera, except at OSD80 in the Greenland Sea where *Mantoniella* was dominant  
 324 and in the Adriatic Sea (OSD49, 76, 77 and 99) where Dolichomastigales and *Mamiella* were dominant  
 325 (Fig. 5A). Procrustean comparison showed that V4 and V9 provided similar Mamiellophyceae genus  
 326 distribution ( $m^2=0.075$  and  $r=0.96$ ). The relative contributions per station of the four major genera  
 327 *Micromonas*, *Mamiella*, *Ostreococcus* and *Bathycoccus* (Fig. S10) was overall statistically similar in  
 328 the two datasets (Table 3), although it could be different for *Mamiella* at specific stations (OSD22, 49,  
 329 132, 123). Stations located in the Adriatic Sea (OSD49, 76, 77, 99) showed a different pattern in the  
 330 heatmap (Fig. S9B) because V9 failed to discriminate the Dolichomastigales clades at the genus level.  
 331 V9 recorded only *Crustomastix* contribution while V4 found 4 to 6 different clades of Crustomastigaceae  
 332 and Dolichomastigaceae (Fig. 5A).. Hierarchical clustering based on Bray-Curtis distances always  
 333 grouped together V4 and V9 (Fig. 5B). Four groups of stations were observed depending on the genus  
 334 dominant at the station: *Micromonas*, *Ostreococcus*, Dolichomastigales or *Mantoniella* (Fig. 5B).

335

## 336 Discussion

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

348 The choice of an identity threshold to build OTUs affects the number of recovered OTUs and  
349 the final taxonomic resolution. An analysis of 2,200 full 18S sequences of protist (Caron *et al.*, 2009)  
350 showed that building OTUs at 95% identity provided a number of OTUs close to the expected number  
351 of species, but the authors remarked that a 98% identity threshold provides a better taxonomic resolution  
352 that allows to investigate interspecific diversity. In the present study, OTUs were built at 97% identity  
353 for both the V4 and the V9 regions of the 18S rRNA gene, in agreement with a number of recent studies  
354 that used these markers (e.g. Massana *et al.*, 2015; Ferrera *et al.*, 2016; Hu *et al.*, 2016). Clustering  
355 regions with different size (V4: 450 bp - V9: 150 bp) at the same identity level should produce more  
356 diverse OTUs for V4 than for V9, although regions where nucleotide changes are concentrated do not  
357 cover the whole amplicons and can be of different length in V4 and V9. For example in V4, most  
358 nucleotide diversity occurs within about 150 bp in the first half of the region (Monier *et al.*, 2016).

359 The V9 dataset provided 20% more OTUs than the V4. This difference between the number of  
360 OTUs for V4 and V9 is the same as the one unveiled in other environmental study such as the Naples  
361 times series results (Piredda *et al.*, 2017). Piredda *et al.* (2017) also found 20% more OTUs built at 97%  
362 identity for V9 than for V4. This could be linked to the size difference between V4 and V9 as discussed  
363 above. Interestingly, these authors showed that the number of OTUs built at 95% identity was similar  
364 for V4 and V9, suggesting that at lower identity thresholds, the size difference has a lower impact.

365 The number of OTUs for the main photosynthetic phyla Ochrophyta, Chlorophyta, Haptophyta  
366 and Cryptophyta falls in the range found in European coastal waters using the V4 and 97% identity  
367 OTUs (1905, 314, 221 and 77 respectively, Massana *et al.*, 2015) except for the Haptophyta for which  
368 three times less OTUs were found in the OSD V4 dataset. The number of OTUs of the main  
369 photosynthetic phyla in the OSD V9 dataset were considerably lower than the numbers of Tara Oceans  
370 V9 OTUs, 3900, 1420, 713 and 195 respectively (de Vargas *et al.*, 2015). However, the depth of

371 sequencing was much higher than in the OSD dataset (around one to two million reads per sample, i.e.  
372 20 to 40 more than for OSD) which increases the occurrence of the rare OTUs that had been filtered out  
373 in the OSD dataset because of the relatively low read number.

374 At six stations (OSD30, 80, 123, 141, 143, 152) the same species richness (OTU number) was  
375 observed but Simpson index was different between V4 and V9 (Fig. 2 A and B). This means that even  
376 if the same number of OTUs was found for V4 and V9, the proportion of each OTU was different. The  
377 V9 Simpson index of OSD80 and 123 (0.87 and 0.91 respectively) fall in the range of Simpson index  
378 calculated in similar environments: for example in Baffin Bay (0.88, Hamilton *et al.*, 2008) and off the  
379 Mediterranean Sea coast (0.92, Ferrera *et al.*, 2016), but the V4 Simpson index was lower (0.68 and  
380 0.81 respectively).

381 In the OSD dataset, photosynthetic groups (Dinophyceae excluded) varied widely from between  
382 0.8 and 81 and between 1.5 and 65 % at the different stations for V4 and V9, respectively, representing  
383 on average 29 and 26% of the sequences recovered, . These average numbers are comparable to those  
384 observed in other studies. For example, Massana and Pedrós-Alió (2008), synthesizing 35 picoplankton  
385 clone libraries of 18S gene from oceanic and coastal waters, found that photosynthetic sequences  
386 represented about 30% of eukaryotic sequences. The proportion of the main photosynthetic phyla  
387 Ochrophyta, Chlorophyta, Haptophyta and Cryptophyta, roughly 17%, 5-7%, 2-3% and 1.3%,  
388 respectively, in the OSD dataset are comparable to those found by Massana and Pedrós-Alió (2008)  
389 (15%, 7.7%, 2.4% and 2%, respectively).

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

397 Within Mamiellophyceae the same genus, most of the time either *Micromonas* or *Ostreococcus*,  
398 was dominant in both V4 and V9 datasets. Not *et al.* (2009) found *Micromonas* to be the most prevalent  
399 genus in the world ocean coastal waters and at more local scale *Micromonas* dominates coastal  
400 picoplankton in the Western English Channel (Not *et al.*, 2004). Rii *et al.* (2016) found that  
401 *Ostreococcus* was dominant in the upwelling-influenced coastal waters from Chile. OSD data also  
402 unveiled a high genetic diversity of the order Dolichomastigales especially in the Adriatic Sea. Viprey  
403 *et al.* (2008) made similar observations in oligotrophic Mediterranean surface waters and Monier *et al.*  
404 (2016) in the Tara *Oceans* survey.

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

#### 418 [Concluding remarks - What is the best choice: V4 or V9?](#)

419 The first element of choice between these two regions is based on the genetic divergence within  
 420 and between the groups of interests (Chenuil, 2006). For Chlorophyta, average similarity is in general  
 421 lower in V9 than V4 (Tragin *et al.*, 2016), which suggests that V9 will be more discriminating than V4  
 422 and will be the best choice. This is the case for example for prasinophytes clade VII, an important  
 423 oceanic group, for which the use of 99% threshold for V9 OTUs allows to discriminate all sub-clades  
 424 (e.g. A1 and A2) defined to date (Lopes dos Santos *et al.*, 2016), while in V4, several clades collapse  
 425 together, having identical sequences in that region. The V9 region of some *Chlamydomonas* is very  
 426 similar to that of prasinophytes clade VII A5, which could lead to misinterpret the distribution of this  
 427 specific sub-clade when using the V9 region. However this may not be the case for other groups such  
 428 as Nephroselmidophyceae for which the two markers are equally suitable (Tragin *et al.*, 2016). The  
 429 second element to take into account is the reference database that contains more representatives of each  
 430 of the taxa investigated. For example, in the present study, the V9 region of the 18S rRNA gene failed  
 431 to discriminate clades within Dolichomastigales (Fig. S9B) because there are only four  
 432 Dolichomastigales V9 reference sequences against 69 for V4 (Tragin *et al.*, 2016). In the same way,  
 433 obtaining accurate image of communities at stations which host rare or uncultured taxa is more difficult  
 434 with V9 than V4, because many sequences in public databases are short and do not extend to the end of  
 435 the 18S rRNA gene. For example, Viprey *et al.* (2008) discovered one novel prasinophyte group (clade  
 436 VIII) by using Chlorophyta specific primers that only amplified a short (around 910 base pairs) sequence  
 437 not extending to the V9 region, and therefore this group can only be studied using V4.

438 Metabarcoding analysis methods using assignment rely heavily on carefully curated public database  
 439 such as PR<sup>2</sup> (Guillou *et al.*, 2013) or, even better, on specifically tailored databases that include, besides  
 440 public sequences, reference sequences for the environment investigated, as for example Arctic specific



441 databases for polar environments (Comeau *et al.*, 2011; Marquardt *et al.*, 2016). Other approaches to  
442 analyze metabarcoding datasets do not rely on reference databases. For example, oligotyping relies on  
443 nucleotide signatures to cluster sequences and can reveal fine distribution patterns of specific taxonomic  
444 groups (Eren *et al.*, 2014; Berry *et al.*, 2017), but to our knowledge it has not been applied to eukaryotes  
445 yet. Phylogenetic placement methods such as pplacer (Matsen *et al.*, 2010) allows to investigate  
446 phylogenetic diversity without assignation against a reference database. Phylogenetic approach however  
447 may be impacted by the lack of reference sequences and have to be complemented by statistical testing  
448 of the consistency of phylogenetic signals (Kembel, 2009; Stegen *et al.*, 2012).

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

452

453 [References](#)

- 454 Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., and Huse, S.M. (2009) A method for studying  
455 protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-  
456 subunit ribosomal RNA genes. *PLoS One* **4**: e6372.
- 457 Baldauf, S.L. (2008) An overview of the phylogeny and diversity of eukaryotes. *J. Syst. Evol.* **46**: 263–  
458 273.
- 459 Behnke, A., Engel, M., Christen, R., Nebel, M., Klein, R.R., and Stoeck, T. (2011) Depicting more  
460 accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU  
461 rRNA gene regions. *Environ. Microbiol.* **13**: 340–9.
- 462 Berry, M.A., White, J.D., Davis, T.W., Jain, S., Johengen, T.H., Dick, G.J., et al. (2017) Are oligotypes  
463 meaningful ecological and phylogenetic units? A case study of *Microcystis* in freshwater lakes.  
464 *Front. Microbiol.* **8**: 365.
- 465 Caron, D.A., Countway, P.D., Savai, P., Gast, R.J., Schnetzer, A., Moorthi, S.D., et al. (2009) Defining  
466 DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl. Environ.*  
467 *Microbiol.* **75**: 5797–808.
- 468 Cerino, F., Bernardi Aubry, F., Coppola, J., La Ferla, R., Maimone, G., Socal, G., and Totti, C. (2012)  
469 Spatial and temporal variability of pico-, nano- and microphytoplankton in the offshore waters of  
470 the southern Adriatic Sea (Mediterranean Sea). *Cont. Shelf Res.* **44**: 94–105.
- 471 Chenuil, A. (2006) Choosing the right molecular genetic markers for studying biodiversity: from  
472 molecular evolution to practical aspects. *Genetica* **127**: 101–20.
- 473 Clayton, S., Lin, Y.-C., Follows, M.J., and Worden, A.Z. (2017) Co-existence of distinct *Ostreococcus*  
474 ecotypes at an oceanic front. *Limnol. Oceanogr.* **62**: 75–88.
- 475 Comeau, A.M., Li, W.K.W., Tremblay, J.-É., Carmack, E.C., and Lovejoy, C. (2011) Arctic Ocean  
476 Microbial Community Structure before and after the 2007 Record Sea Ice Minimum. *PLoS One* **6**:  
477 e27492.
- 478 Decelle, J., Romac, S., Sasaki, E., Not, F., Mahé, F., Sogin, M., et al. (2014) Intracellular Diversity of  
479 the V4 and V9 Regions of the 18S rRNA in Marine Protists (Radiolarians) Assessed by High-  
480 Throughput Sequencing. *PLoS One* **9**: e104297.
- 481 van Dijk, E.L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014) Ten years of next-generation  
482 sequencing technology. *Trends Genet.* **30**: 418–426.
- 483 Dunthorn, M., Klier, J., Bunge, J., and Stoeck, T. (2012) Comparing the hyper-variable V4 and V9  
484 regions of the small subunit rDNA for assessment of ciliate environmental diversity. *J. Eukaryot.*  
485 *Microbiol.* **59**: 185–187.
- 486 Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011) UCHIME improves  
487 sensitivity and speed of chimera detection. *Bioinformatics* **27**: 2194–200.
- 488 Edvardsen, B., Egge, E.S., and Vaultot, D. (2016) Diversity and distribution of haptophytes revealed by  
489 environmental sequencing and metabarcoding – a review. *Perspect. Phycol.* **3**: 77–91.
- 490 Egge, E., Bittner, L., Andersen, T., Audic, S., de Vargas, C., and Edvardsen, B. (2013) 454  
491 pyrosequencing to describe microbial eukaryotic community composition, diversity and relative  
492 abundance: a test for marine Haptophytes. *PLoS One* **8**:
- 493 Eren, A.M., Morrison, H.G., Lescault, P.J., Reveillaud, J., Vineis, J.H., and Sogin, M.L. (2014)

- 494 Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning of high-  
495 throughputs marker gene sequences. *ISME J.* **9**: 968–979.
- 496 Ferrera, I., Giner, C.R., Reñé, A., Camp, J., Massana, R., Gasol, J.M., and Garcés, E. (2016) Evaluation  
497 of alternative high-throughput sequencing methodologies for the monitoring of marine  
498 picoplanktonic biodiversity based on rRNA gene amplicons. *Front. Mar. Sci.* **3**: 147.
- 499 Giner, C.R., Forn, I., Romac, S., Logares, R., de Vargas, C., and Massana, R. (2016) Environmental  
500 sequencing provides reasonable estimates of the relative abundance of specific picoeukaryotes.  
501 *Appl. Environ. Microbiol.* **82**: 4757–4766.
- 502 Gómez, F. (2012) A quantitative review of the lifestyle, habitat and trophic diversity of dinoflagellates  
503 (Dinoflagellata, Alveolata). *Syst. Biodivers.* **10**: 267–275.
- 504 Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., et al. (2013) The Protist Ribosomal  
505 Reference database (PR2): A catalog of unicellular eukaryote Small Sub-Unit rRNA sequences  
506 with curated taxonomy. *Nucleic Acids Res.* **41**: 597–604.
- 507 Hamilton, A.K., Lovejoy, C., Galand, P.E., and Ingram, R.G. (2008) Water masses and biogeography  
508 of picoeukaryote assemblages in a cold hydrographically complex system. *Limnol. Oceanogr.* **53**:  
509 922–935.
- 510 Hu, S., Campbell, V., Connell, P., Gellen, A.G., Liu, Z., Terrado, R., and Caron, D.A. (2016) Protistan  
511 diversity and activity inferred from RNA and DNA at a coastal ocean site in the eastern North  
512 Pacific. *FEMS Microb. Ecol.* 1–39.
- 513 Hu, S.K., Liu, Z., Lie, A.A.Y., Countway, P.D., Kim, D.Y., Jones, A.C., et al. (2015) Estimating  
514 Protistan Diversity Using High-Throughput Sequencing. *J. Eukaryot. Microbiol.* **62**: 688–693.
- 515 Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012) Geneious  
516 Basic: an integrated and extendable desktop software platform for the organization and analysis of  
517 sequence data. *Bioinformatics* **28**: 1647–9.
- 518 Kembel, S.W. (2009) Disentangling niche and neutral influences on community assembly: assessing the  
519 performance of community phylogenetic structure tests. *Ecol. Lett.* **12**: 949–960.
- 520 Kopf, A., Bicak, M., Kottmann, R., Schnetzer, J., Kostadinov, I., Lehmann, K., et al. (2015) The ocean  
521 sampling day consortium. *Gigascience* **4**: 27.
- 522 Lopes dos Santos, A., Gourvil, P., Tragin, M., Noël, M.-H., Decelle, J., Romac, S., and Vaulot, D. (2016)  
523 Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae  
524 in oceanic waters. *ISME J.* **11**: 512–528.
- 525 López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C., and Moreira, D. (2001) Unexpected diversity  
526 of small eukaryotes in deep-sea Antarctic plankton. *Nature* **409**: 603–607.
- 527 Majaneva, M., Hyytiäinen, K., Varvio, S.L., Nagai, S., and Blomster, J. (2015) Bioinformatic amplicon  
528 read processing strategies strongly affect eukaryotic diversity and the taxonomic composition of  
529 communities. *PLoS One* **10**: e0130035.
- 530 Marquardt, M., Vader, A., Stübner, E.I., Reigstad, M., and Gabrielsen, T.M. (2016) Strong Seasonality  
531 of Marine Microbial Eukaryotes in a High-Arctic Fjord (Isfjorden, in West Spitsbergen, Norway).  
532 *Appl. Environ. Microbiol.* **82**: 1868–80.
- 533 Massana, R., del Campo, J., Sieracki, M.E., Audic, S., and Logares, R. (2014) Exploring the uncultured  
534 microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J.*  
535 **8**: 854–66.

- 536 Massana, R., Gobet, A., Audic, S., Bass, D., Bittner, L., Boutte, C., et al. (2015) Marine protist diversity  
537 in European coastal waters and sediments as revealed by high-throughput sequencing. *Environ.*  
538 *Microbiol.* **17**: 4035–4049.
- 539 Massana, R. and Pedrós-Alió, C. (2008) Unveiling new microbial eukaryotes in the surface ocean. *Curr.*  
540 *Opin. Microbiol.* **11**: 213–218.
- 541 Matsen, F.A., Kodner, R.B., and Armbrust, E.V. (2010) pplacer: linear time maximum-likelihood and  
542 Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*  
543 **11**: 538.
- 544 Monier, A., Worden, A.Z., and Richards, T.A. (2016) Phylogenetic diversity and biogeography of the  
545 Mamiellophyceae lineage of eukaryotic phytoplankton across the oceans. *Environ. Microbiol. Rep.*  
546 **8**: 461–469.
- 547 Moon-van der Staay, S.Y., De Wachter, R., and Vaultot, D. (2001) Oceanic 18S rDNA sequences from  
548 picoplankton reveal unsuspected eukaryotic diversity. *Nature* **409**: 607–10.
- 549 Not, F., del Campo, J., Balagué, V., de Vargas, C., and Massana, R. (2009) New insights into the  
550 diversity of marine picoeukaryotes. *PLoS One* **4**..
- 551 Not, F., Latasa, M., Marie, D., Cariou, T., Vaultot, D., and Simon, N. (2004) A single species,  
552 *Micromonas pusilla* (Prasinophyceae), dominates the eukaryotic picoplankton in the Western  
553 English Channel. *Appl. Environ. Microbiol.* **70**: 4064–72.
- 554 Pernice, M.C., Logares, R., Guillou, L., Massana, R., and Franz, M. (2013) General Patterns of Diversity  
555 in Major Marine Microeukaryote Lineages. *PLoS One* **8**: e57170.
- 556 Piredda, R., Tomasino, M.P., D’Erchia, A.M., Manzari, C., Pesole, G., Montresor, M., et al. (2017)  
557 Diversity and temporal patterns of planktonic protist assemblages at a Mediterranean Long Term  
558 Ecological Research site. *FEMS Microbiol. Ecol.* **93**..
- 559 Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glockner, F.O. (2007)  
560 SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA  
561 sequence data compatible with ARB. *Nucleic Acids Res.* **35**: 7188–7196.
- 562 Revelante, N. and Gilmartin, M. (1976) Temporal succession of phytoplankton in the northern adriatic.  
563 *Netherlands J. Sea Res.* **10**: 377–396.
- 564 Rii, Y.M., Duhamel, S., Bidigare, R.R., Karl, D.M., Repeta, D.J., and Church, M.J. (2016) Diversity  
565 and productivity of photosynthetic picoeukaryotes in biogeochemically distinct regions of the  
566 South East Pacific Ocean. *Limnol. Oceanogr.* **61**: 806–824.
- 567 Sanger, F. and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed  
568 synthesis with DNA polymerase. *J. Mol. Biol.* **94**: 441–448.
- 569 Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., et al. (2009)  
570 Introducing mothur: open-source, platform-independent, community-supported software for  
571 describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**: 7537–41.
- 572 Simmons, M.P., Sudek, S., Monier, A., Limardo, A.J., Jimenez, V., Perle, C.R., et al. (2016) Abundance  
573 and biogeography of picoprasinophyte ecotypes and other phytoplankton in the Eastern North  
574 Pacific Ocean. *Appl. Environ. Microbiol.* **82**: 1693–705.
- 575 Simpson, E.H. (1949) Measurement of diversity. *Nature* **163**: 688–688.
- 576 Stegen, J.C., Lin, X., Konopka, A.E., and Fredrickson, J.K. (2012) Stochastic and deterministic  
577 assembly processes in subsurface microbial communities. *ISME J.* **6**: 1653–64.

- 578 Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D.M., Breininger, H.-W., and Richards, T.A.  
579 (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex  
580 eukaryotic community in marine anoxic water. *Mol. Ecol.* **19**: 21–31.
- 581 Tragin, M., Lopes dos Santos, A., Christen, R., and Vaultot, D. (2016) Diversity and ecology of green  
582 microalgae in marine systems: an overview based on 18S rRNA gene sequences. *Perspect. Phycol.*  
583 **3**: 141–154.
- 584 de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., et al. (2015) Eukaryotic plankton  
585 diversity in the sunlit ocean. *Science* **348**: 1261605–1261605.
- 586 Viprey, M., Guillou, L., Ferréol, M., and Vaultot, D. (2008) Wide genetic diversity of picoplanktonic  
587 green algae (Chloroplastida) in the Mediterranean Sea uncovered by a phylum-biased PCR  
588 approach. *Environ. Microbiol.* **10**: 1804–1822.
- 589 Wang, Q., Garrity, G.M., Tiedje, J.M., and Cole, J.R. (2007) Naive Bayesian classifier for rapid  
590 assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol.* **73**:  
591 5261–5267.
- 592 Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bull.* **1**: 80.
- 593
- 594

595 [Table legends](#)

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

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

599 Table 3: General descriptive statistics: maximum, minimum, mean, standard deviation and results of the  
600 Wilcoxon test (P value) for V4 versus V9 OTU numbers, Simpson index (data from the Fig. 2) and  
601 photosynthetic groups relative contribution (see Fig. S5, S6, S8 and S10). P values in bold are above the  
602 0.05 threshold indicating that V4 and V9 are not significantly different while P values in italics were  
603 computed with datasets presenting ex aequo values.

604

605 **Figure legends**

606 Fig. 1: Map of the 27 OSD stations sampled 2014 for which both V4 and V9 sequences were available.

607 Fig. 2: A. Species richness: number of OTUs per stations for V4 versus V9. The grey line corresponds  
608 to  $y=x$ , and the black line corresponds to the regression  $y=1.27x+53$  ( $R^2 = 0.996$ ). B. Simpson's diversity  
609 index per stations for V4 vs. V9. Grey star corresponds to the OSD30 Simpson's index after removal of  
610 the metazoan V9 OTU.

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

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

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

626

627 [Supplementary Figures](#)

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

632 Supplementary Fig. S2: A. Rarefaction curves; B. Rank abundance distribution. x-axis represents OTUs  
633 by decreasing number of sequences.

634 Supplementary Fig. S3: Rarefaction curves per station A. V4; B. V9.

635 Supplementary Fig. S4: A and B. Non-metric Multi Dimensional Scaling (NMDS) representation of  
636 communities based on lowest taxonomic level (OTUs) for V4 (A) and V9 (B). The dissimilarity matrix  
637 was computed using Bray Curtis distance. C and D. Hierarchical cluster analysis based on the Bray  
638 Curtis matrix for V4 (C) and V9 (D). Stations in panels A and B were grouped together based on clusters  
639 from panels C and D using a fixed threshold (0.9).

640 Supplementary Fig. S5: Correlation between V4 and V9 relative contribution to photosynthetic  
641 metabarcodes in major photosynthetic phyla. A. Ochrophyta. B. Chlorophyta. C. Haptophyta. D.  
642 Cryptophyta.

643 Supplementary Fig. S6: Correlation between V4 and V9 relative contribution of the four major  
644 Ochrophyta Classes. A. Bacillariophyta. B. Dictyochophyceae., C. Chrysophyceae-Synurophyceae. D.  
645 Pelagophyceae.

646 Supplementary Fig. S7: Percentage of genera from photosynthetic groups found either only in V4 (blue),  
647 or only in V9 (red), or in both datasets (grey). Only taxonomically valid genera and only Classes with  
648 at least 5 genera were taken into account. Numbers below each group indicate the total number of genera  
649 recorded.

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

654 Supplementary Fig. S9: Heatmap of differences between V9 and V4 (V9-V4) relative contribution: A.  
655 Chlorophyta classes B. Mamiellophyceae genera. The colors correspond to the difference from - 50% (-  
656 0.5) to + 50 % (0.5).



657 Supplementary Fig. S10: Correlation between V4 and V9 relative contribution to Chlorophyta  
658 metabarcodes for major Mamiellophyceae genera. A. *Micromonas*. B. *Mamiella*. C. *Ostreococcus*. D.  
659 *Bathycoccus*.

660

661 [Supplementary data](#)

662 The data are deposited on Figshare at <https://figshare.com/s/ff26853527ced961326a>

663 Supplementary Data 1. Mothur script for sequence analysis

664 Supplementary Data 2. Fasta file of Chlorophyta OTUs for V4

665 Supplementary Data 3. Fasta file of Chlorophyta OTUs for V9

666 Supplementary Data 4. Chlorophyta OTUs for V4 with assignation and read abundance at the different  
667 stations (Excel file).

668 Supplementary Data 5. Chlorophyta OTUs for V9 with assignation and read abundance at the different  
669 stations (Excel file).

670 Supplementary Data 6. Top 10 BLAST hits against Genbank nr database for Chlorophyta V4 OTUs.  
671 Red lines correspond to OTUs badly assigned to non-Chlorophyta and green corresponds to OTUs badly  
672 assigned to another Chlorophyta representative.

673 Supplementary Data 7. Top 10 BLAST hits against Genbank nr database for Chlorophyta V9 OTUs.  
674 Red lines correspond to OTUs badly assigned to non-Chlorophyta and green lines corresponds to OTUs  
675 badly assigned to another Chlorophyta representative.

676

677

Fig.S1

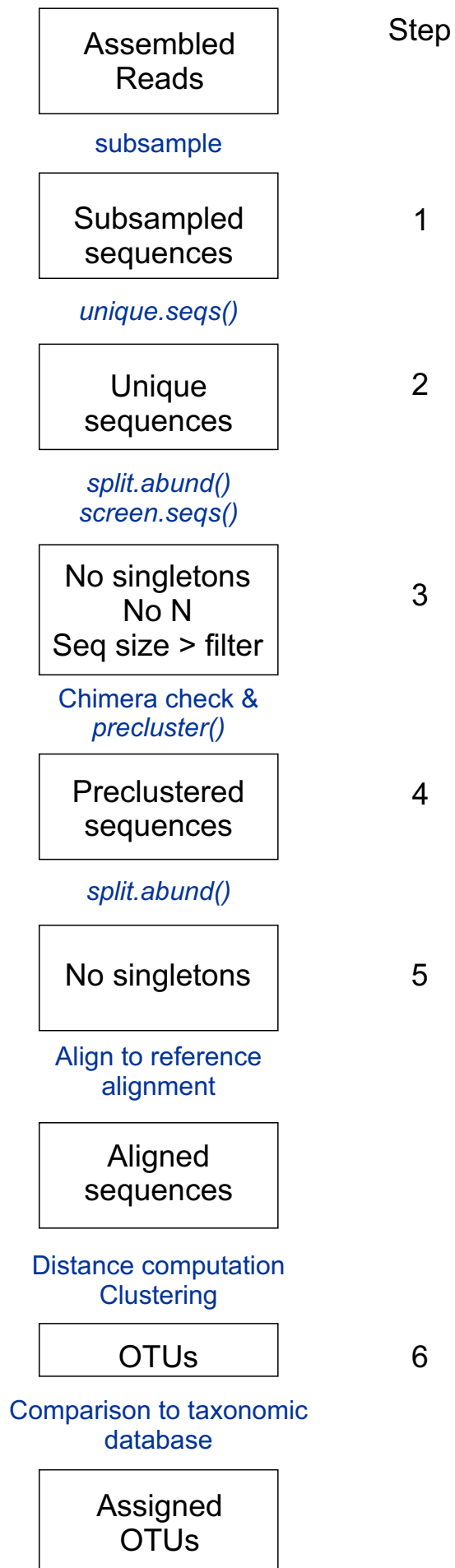


Fig.S2

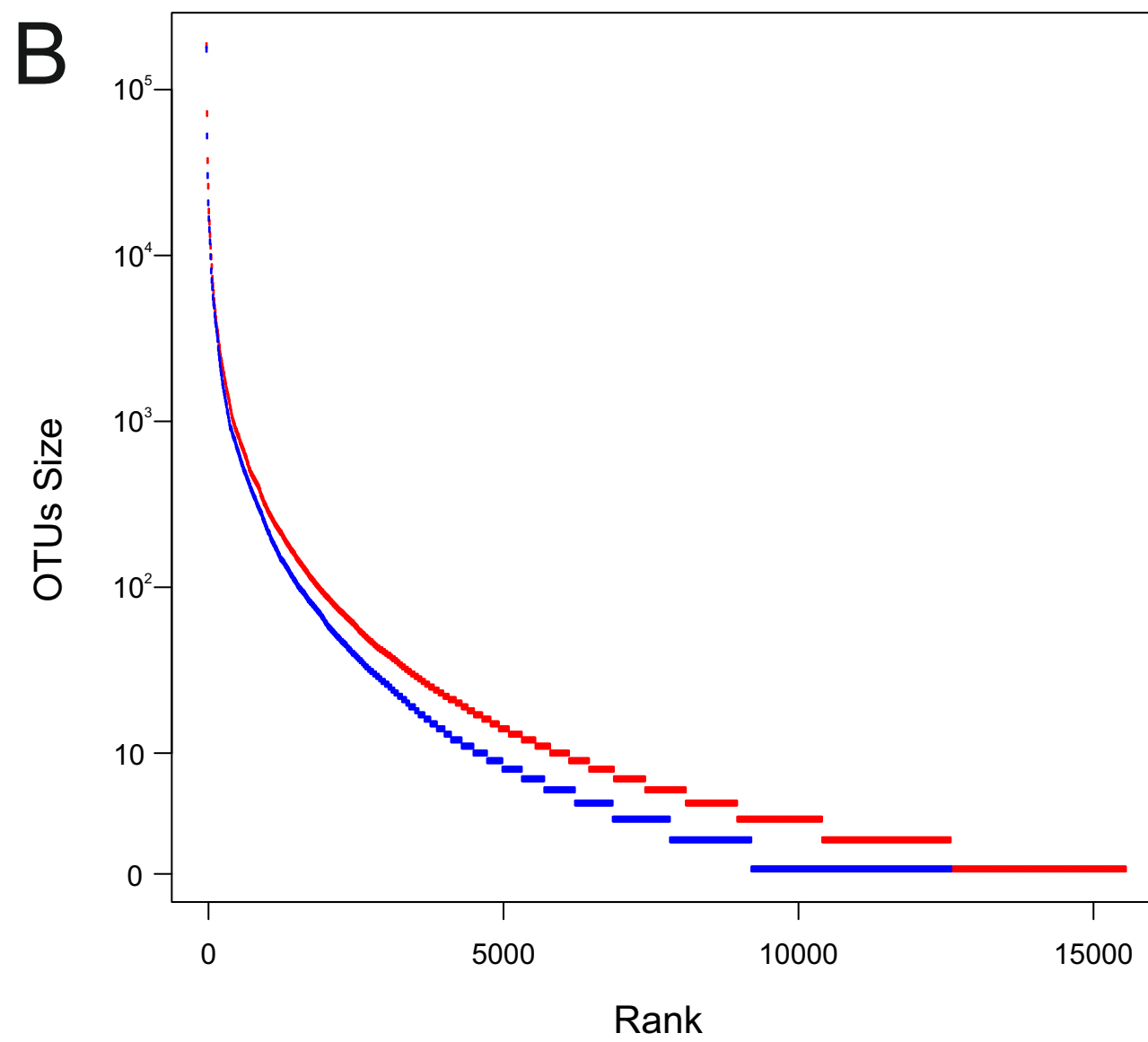
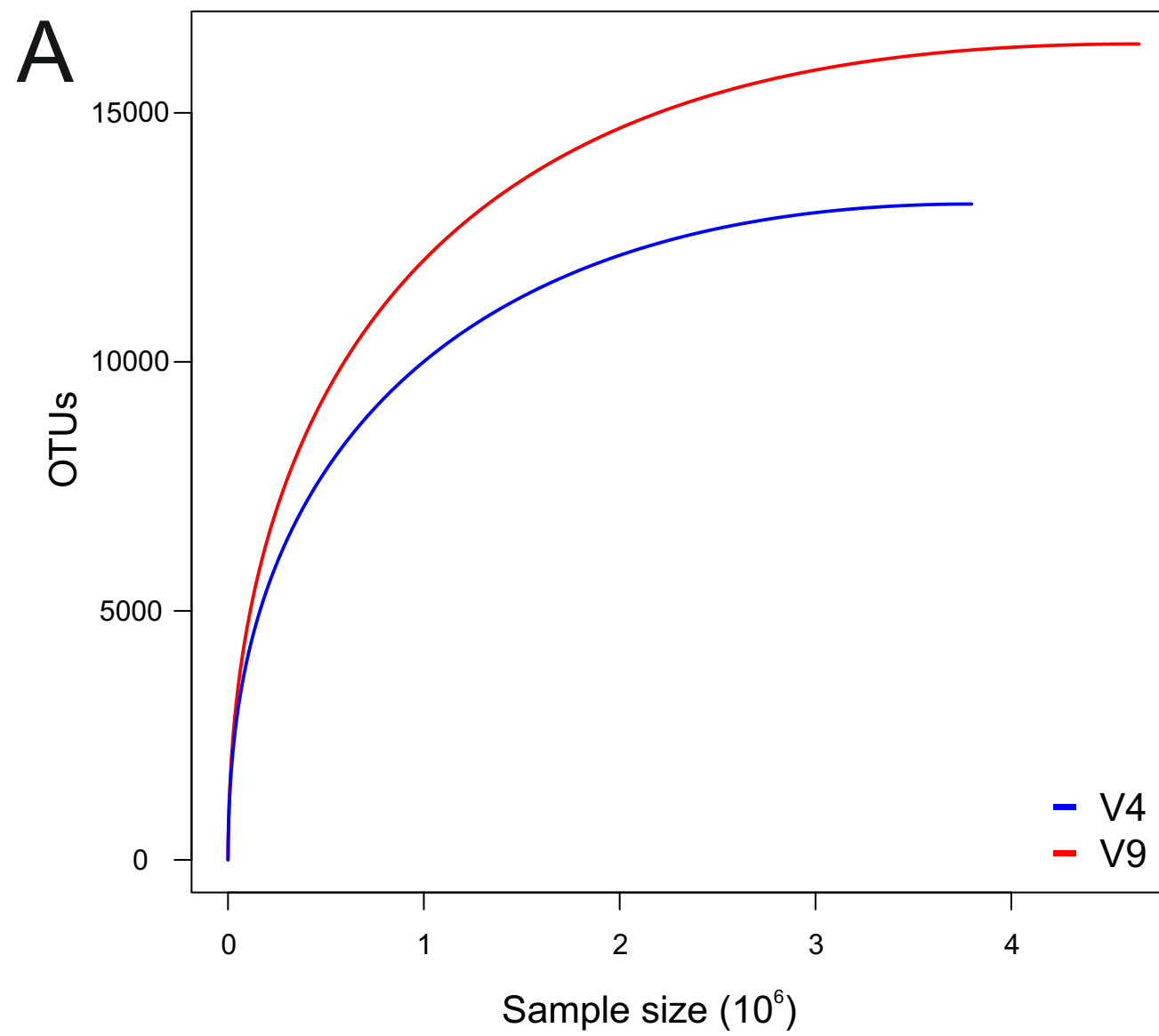


Fig.S3

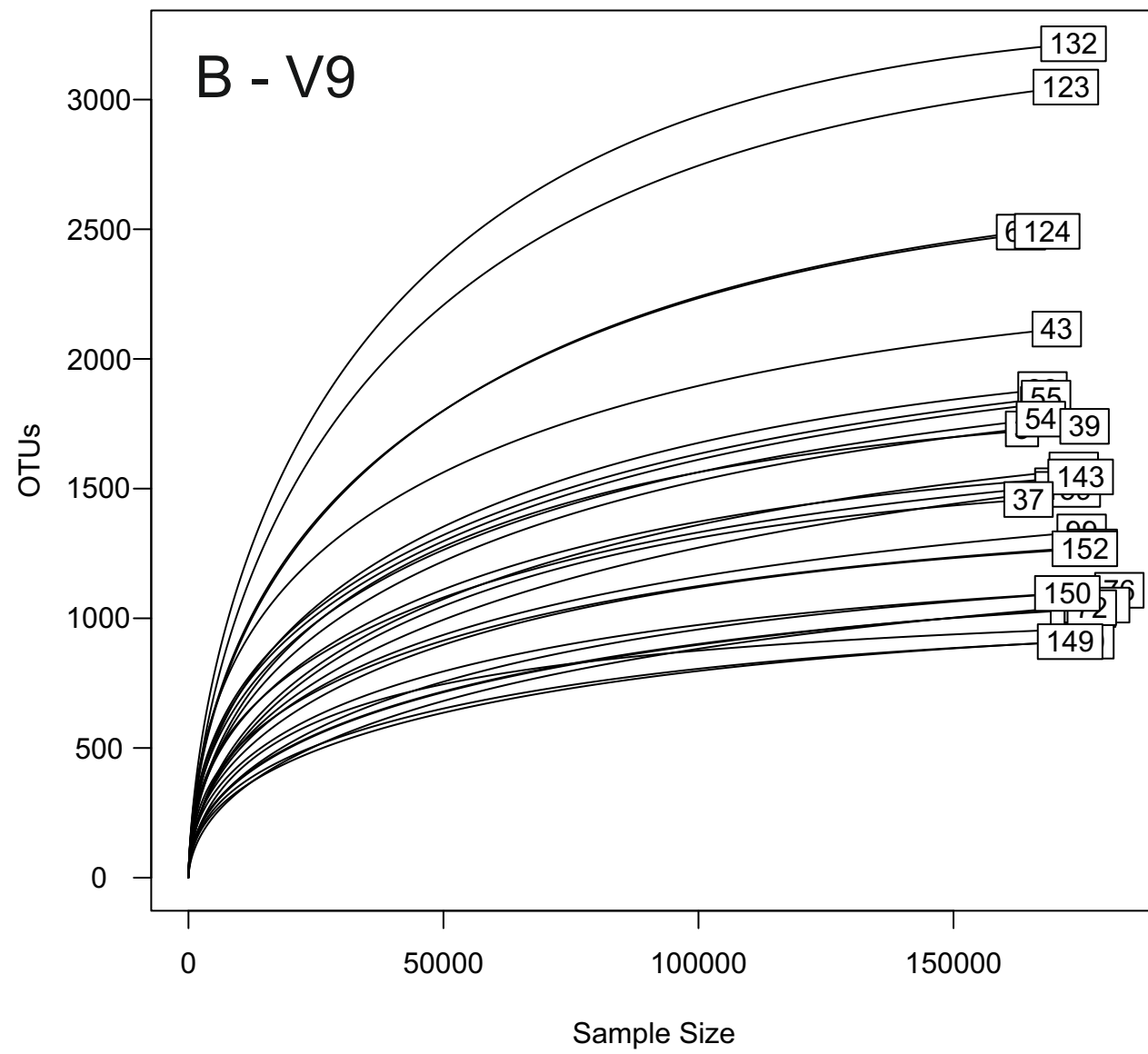
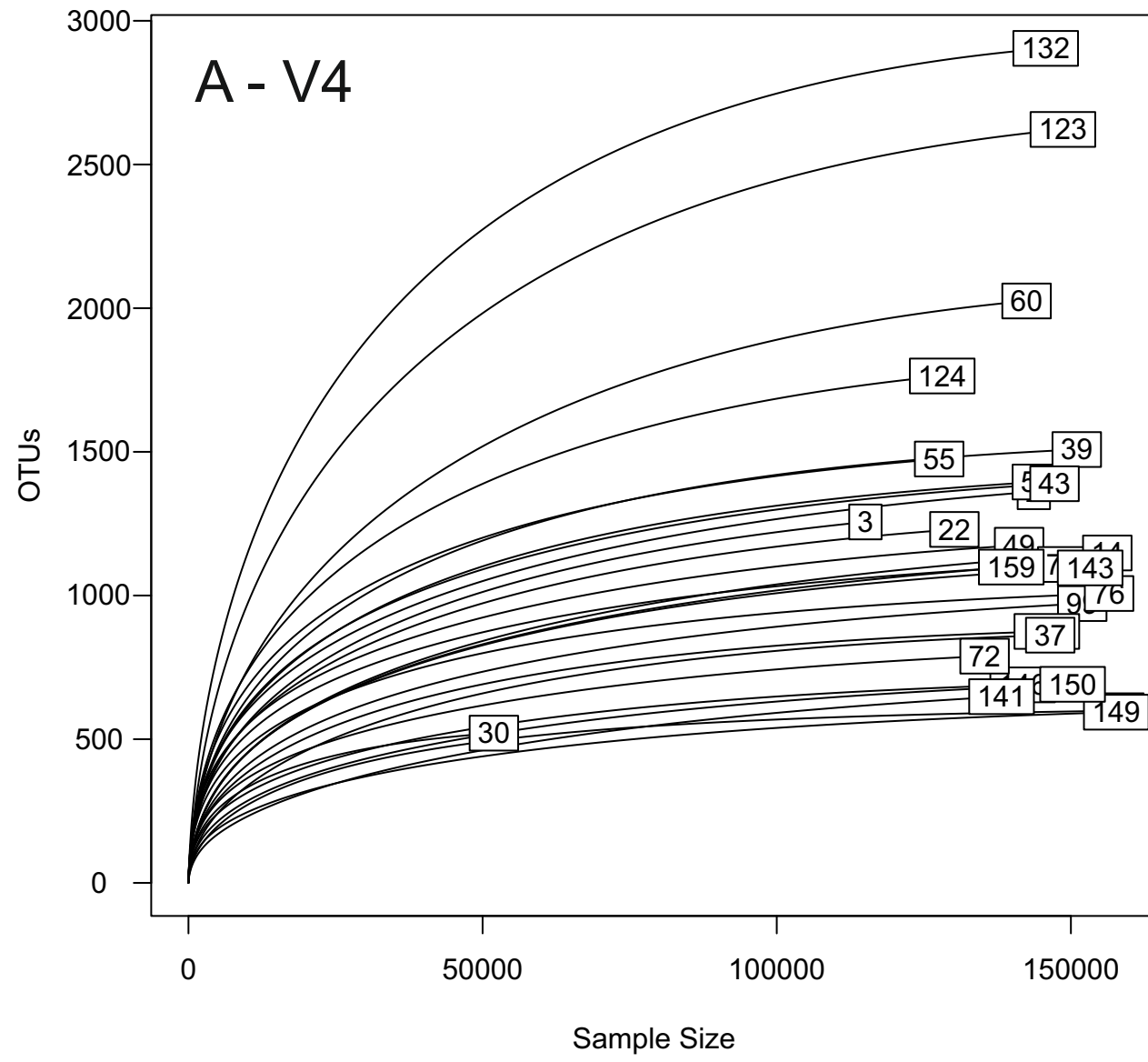


Fig.S4

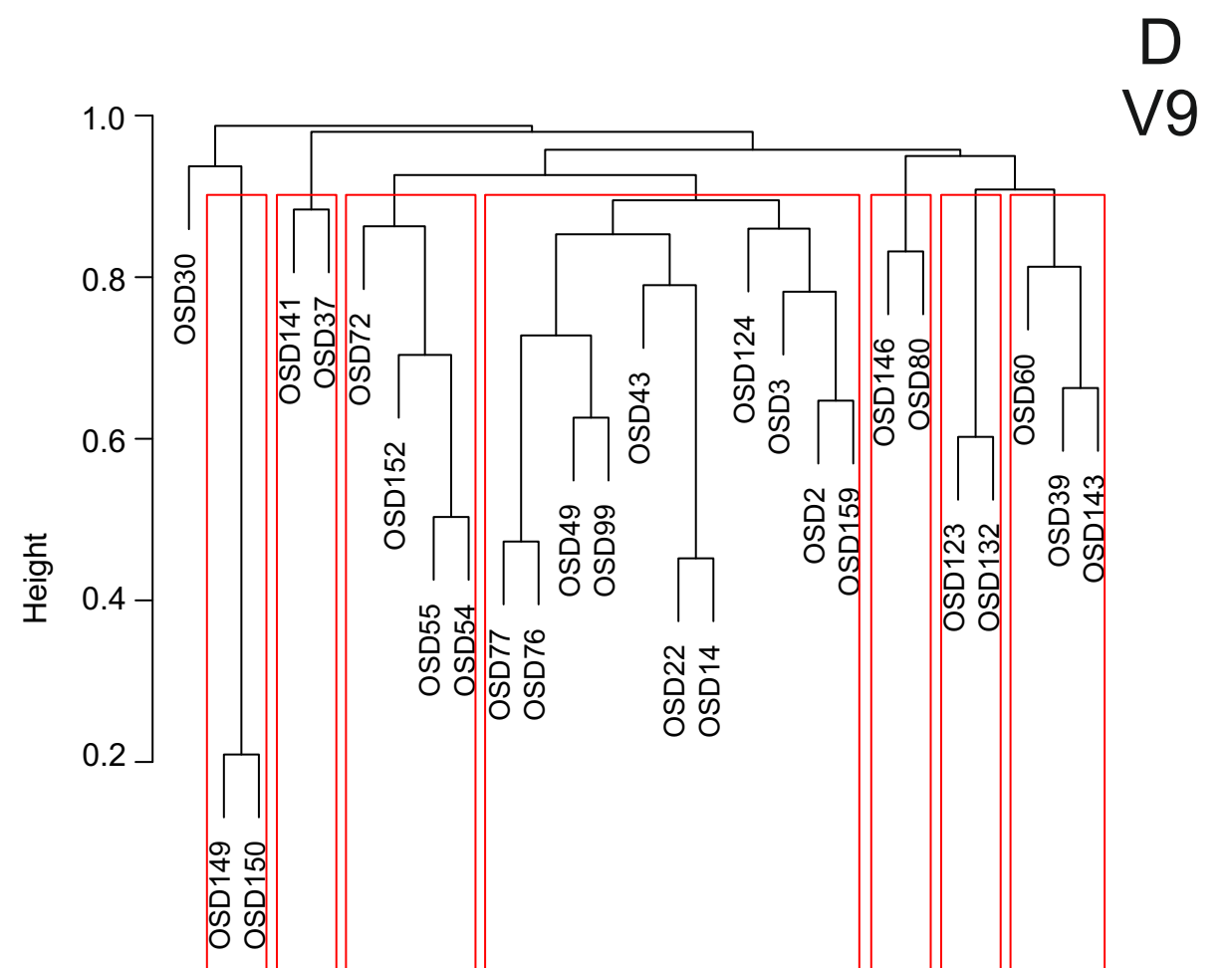
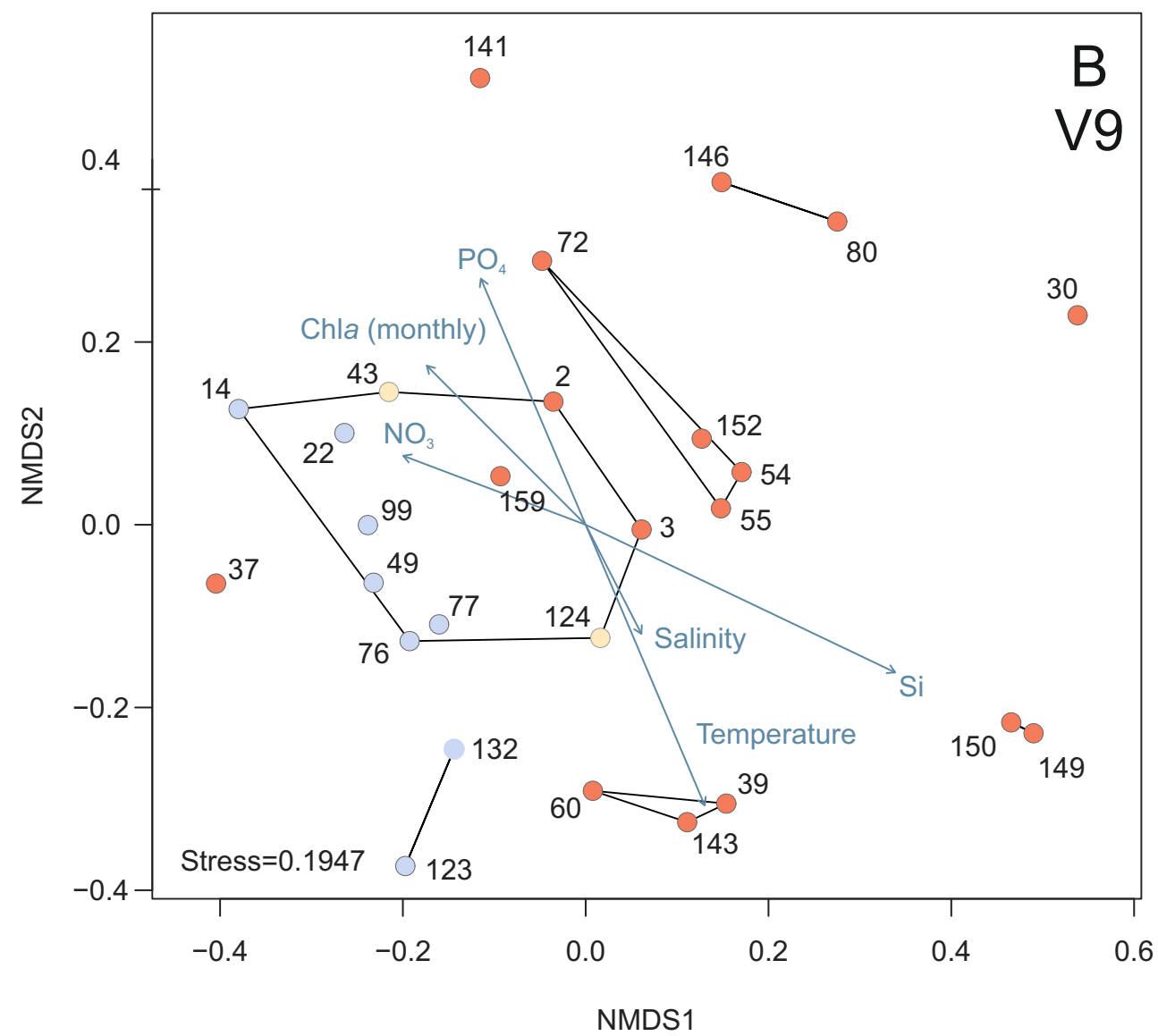
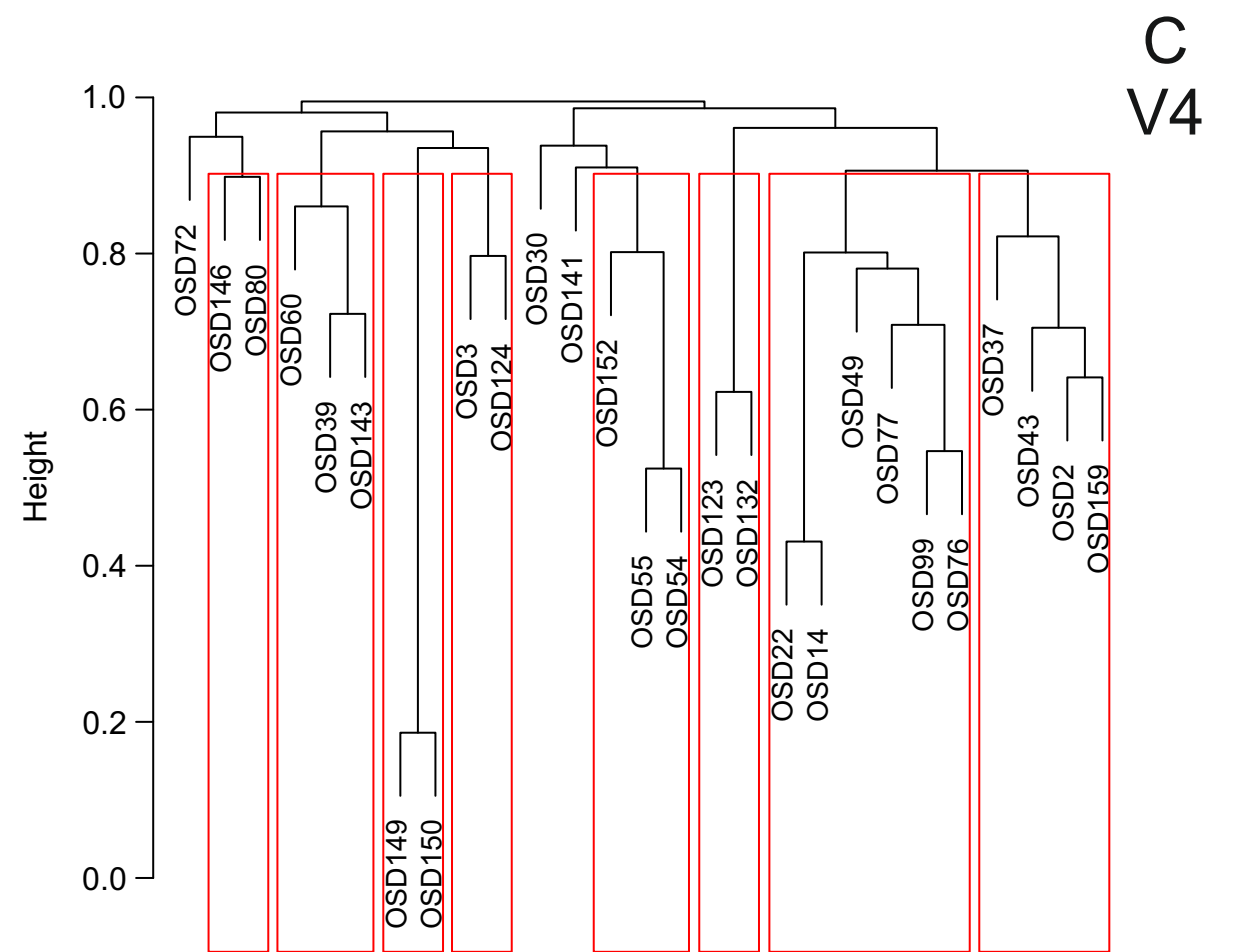
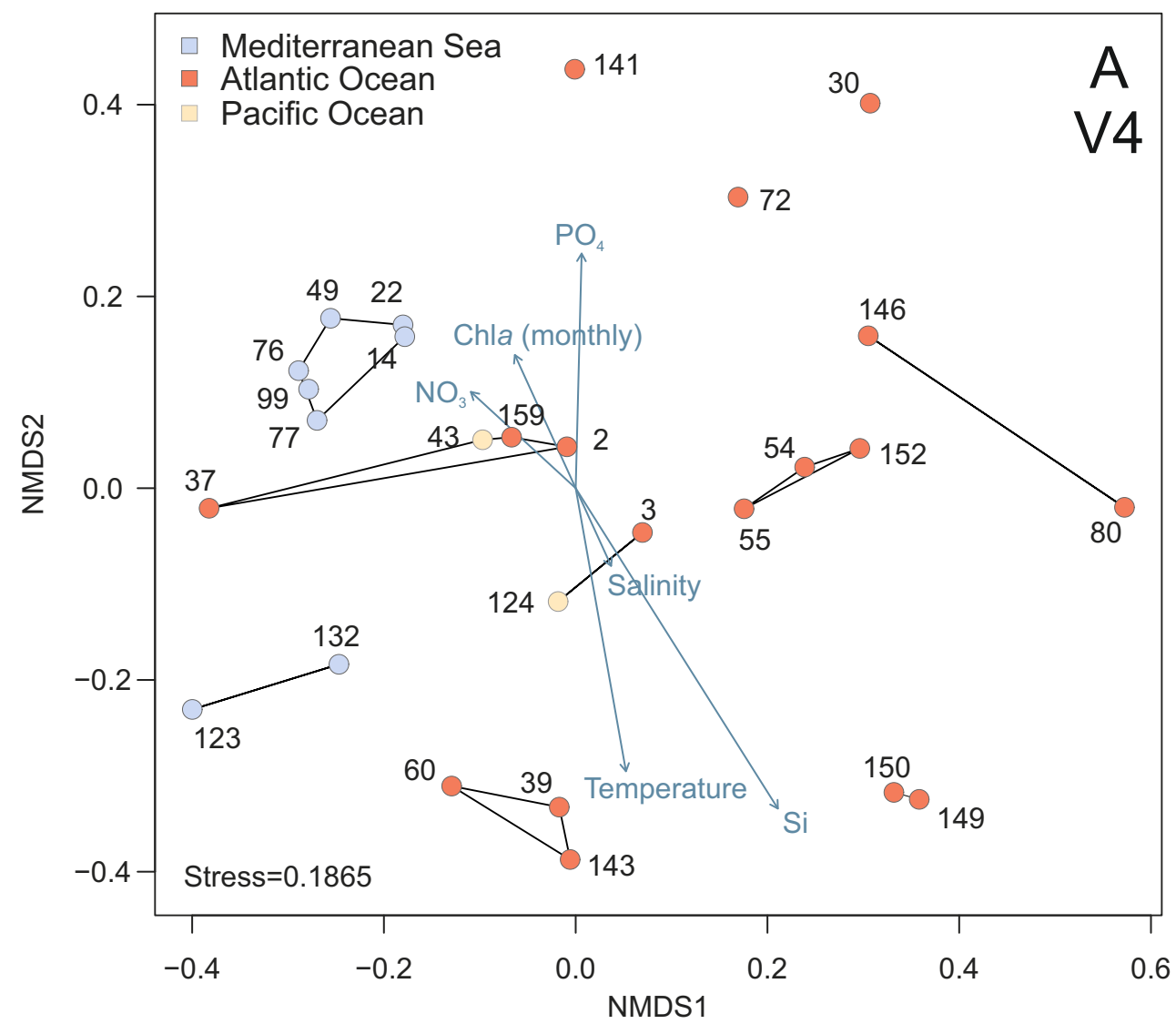


Fig.S5

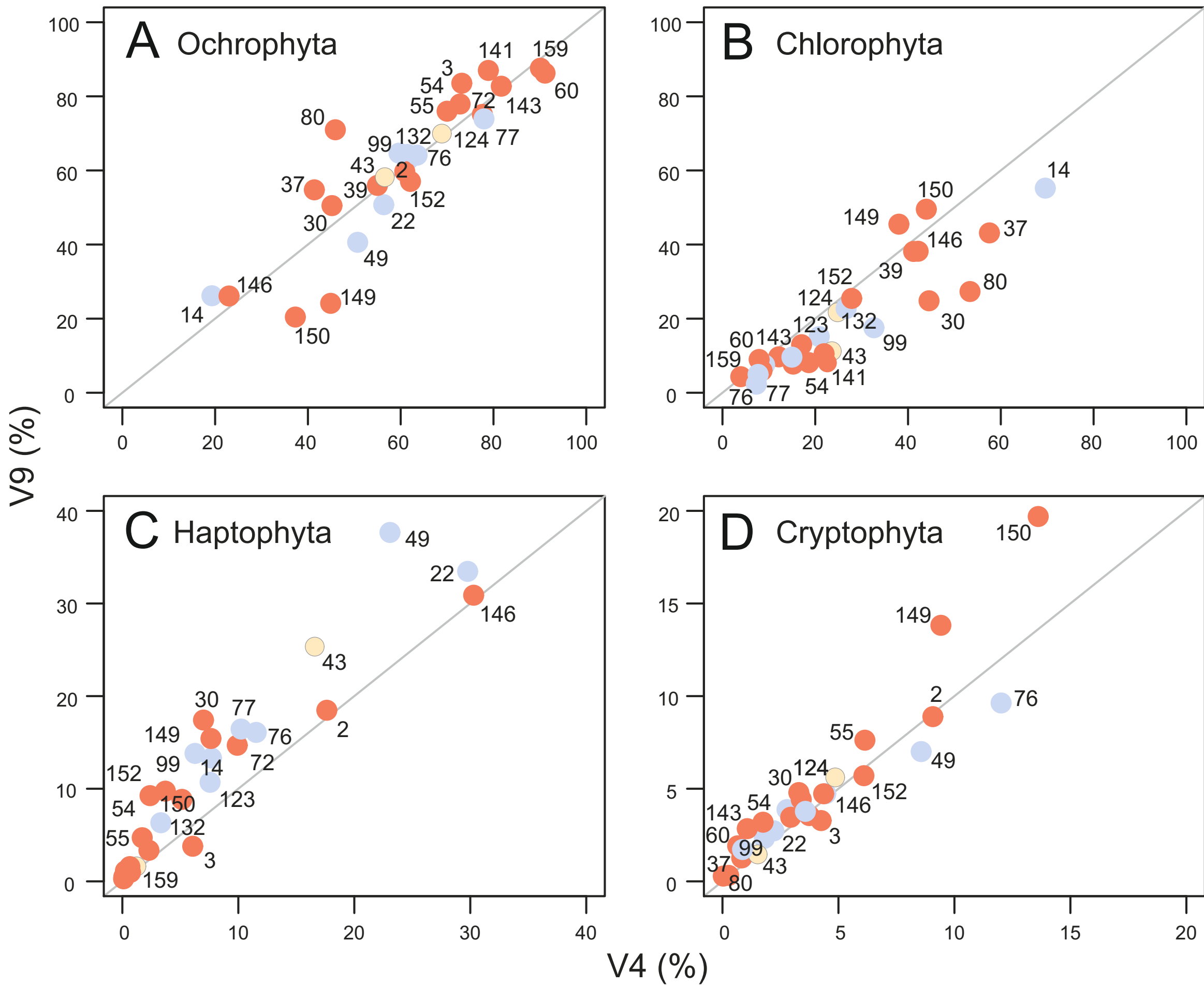
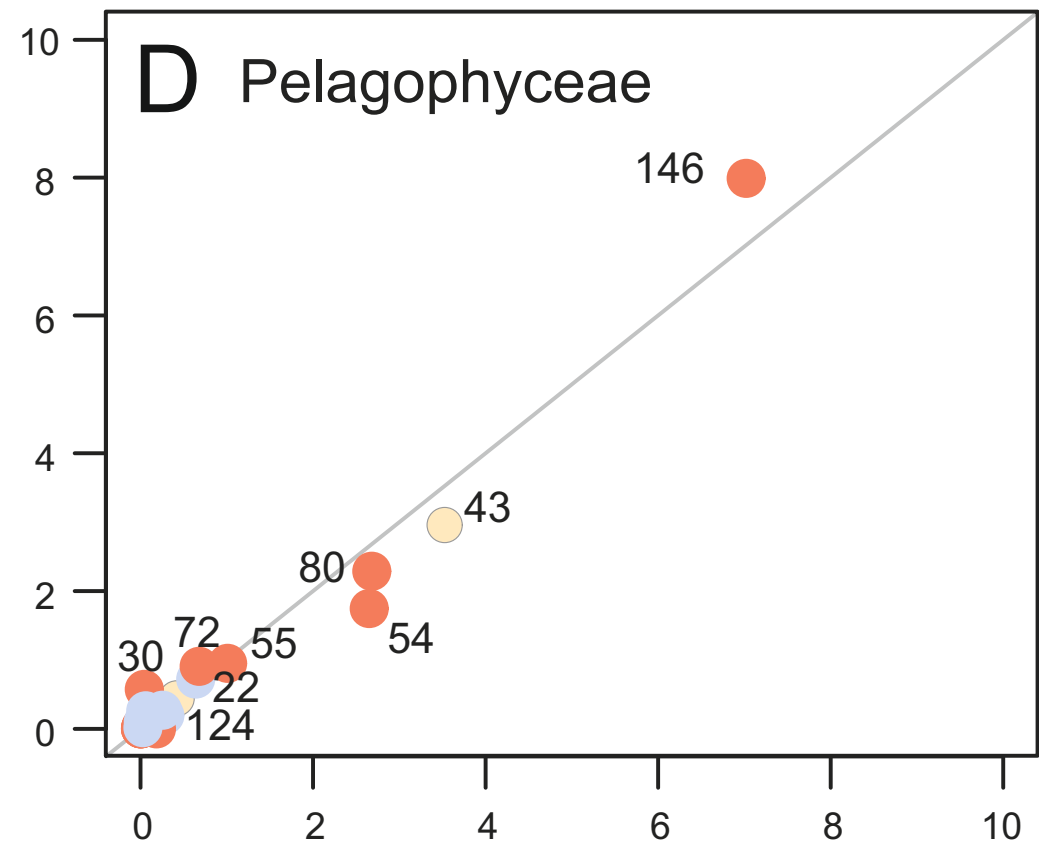
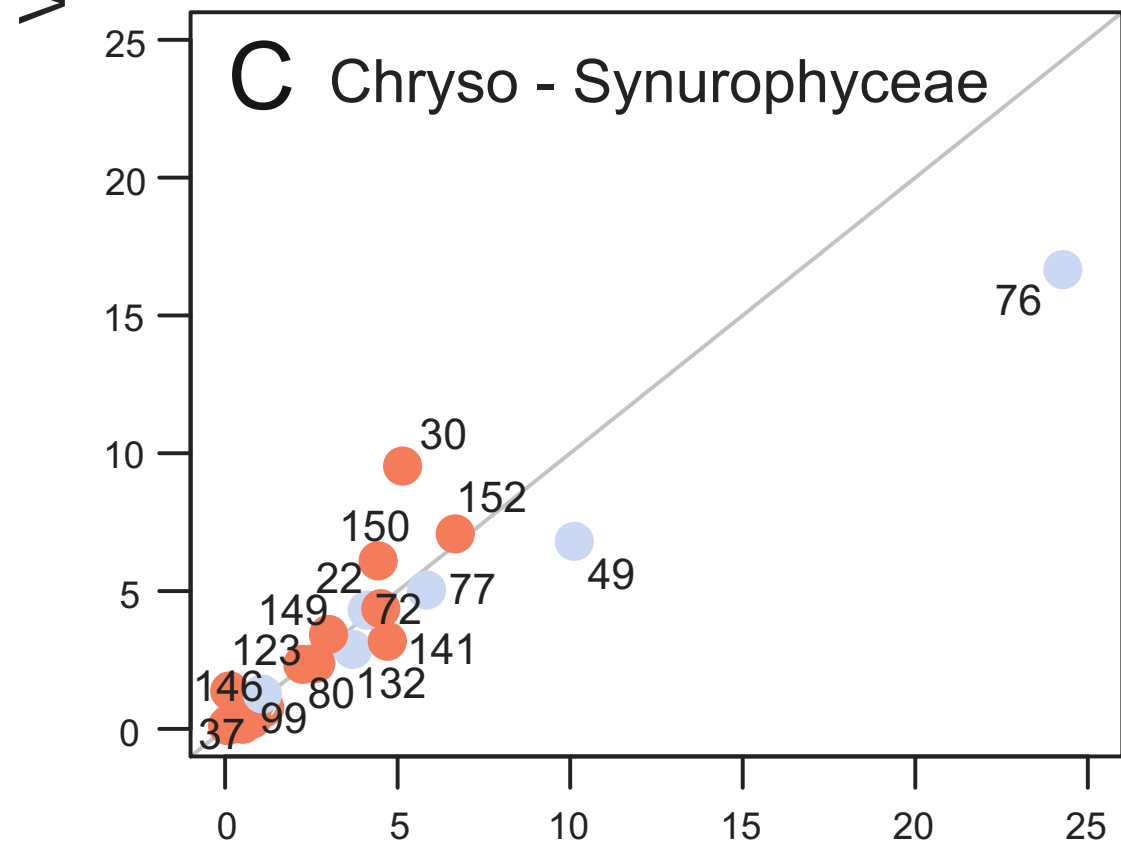
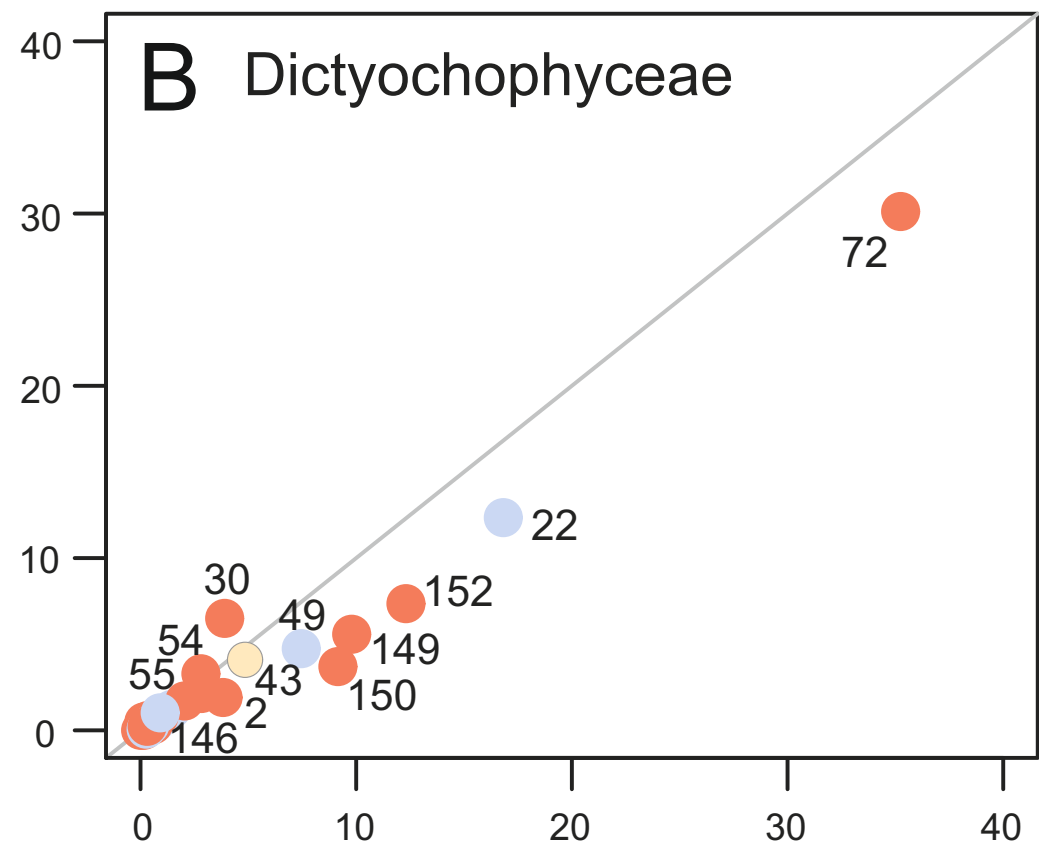
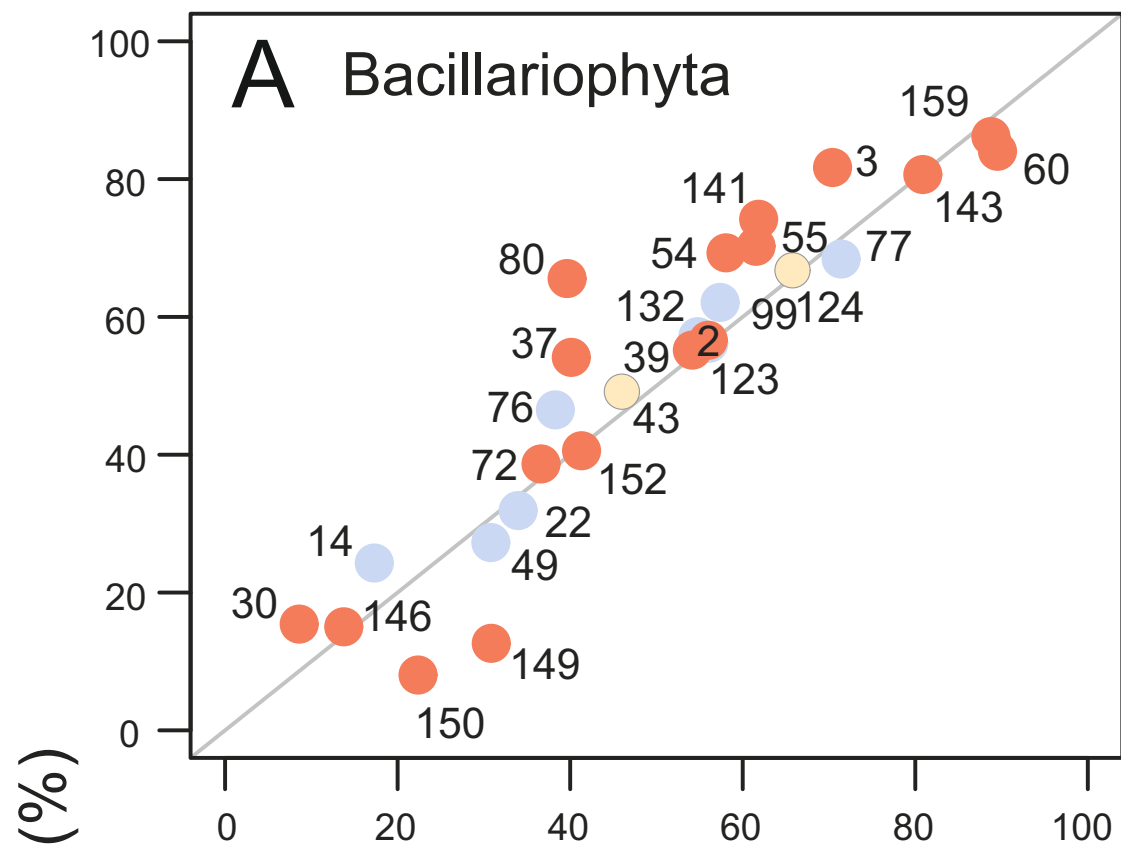


Fig.S6



V4 (%)



# Fig.S7

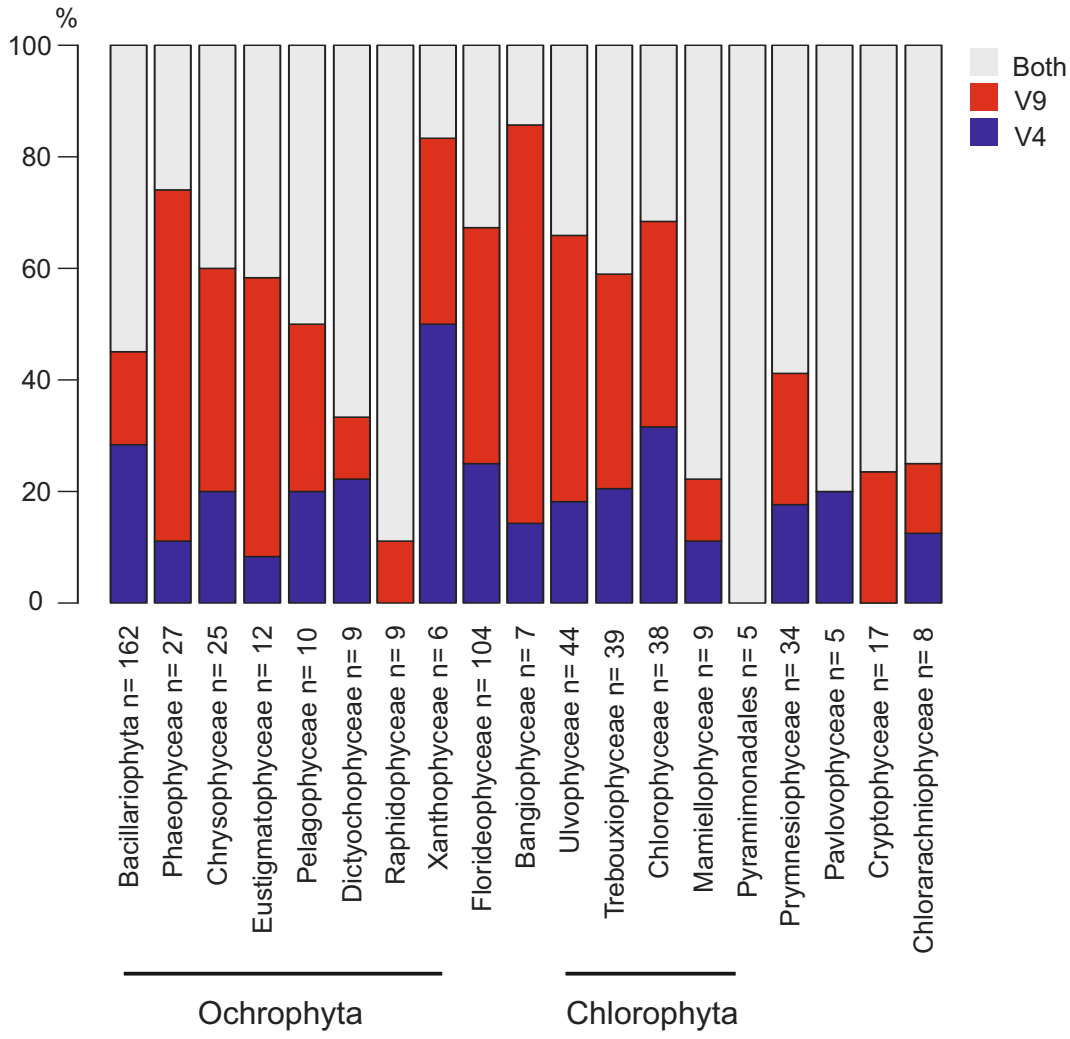


Fig.S8

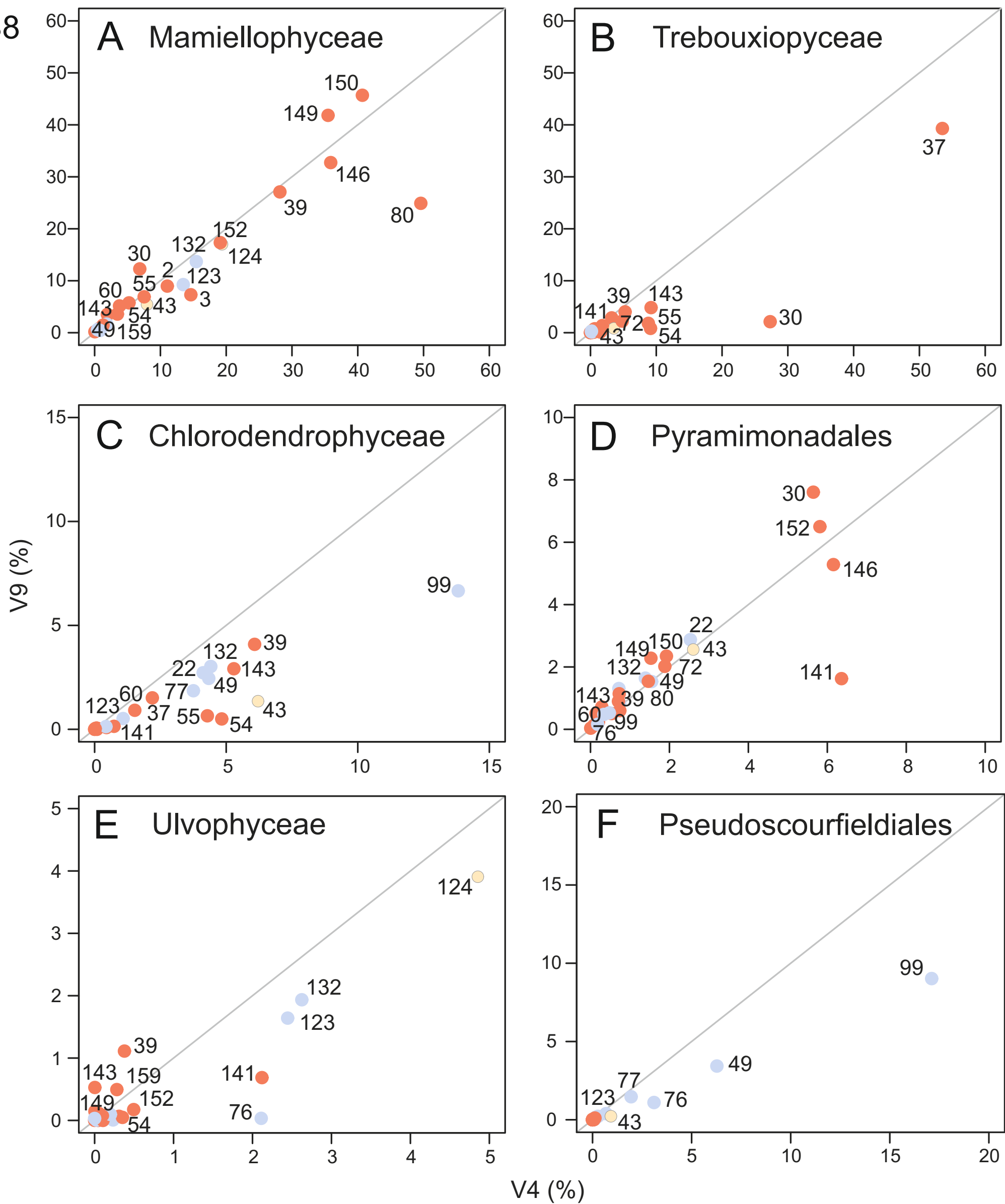
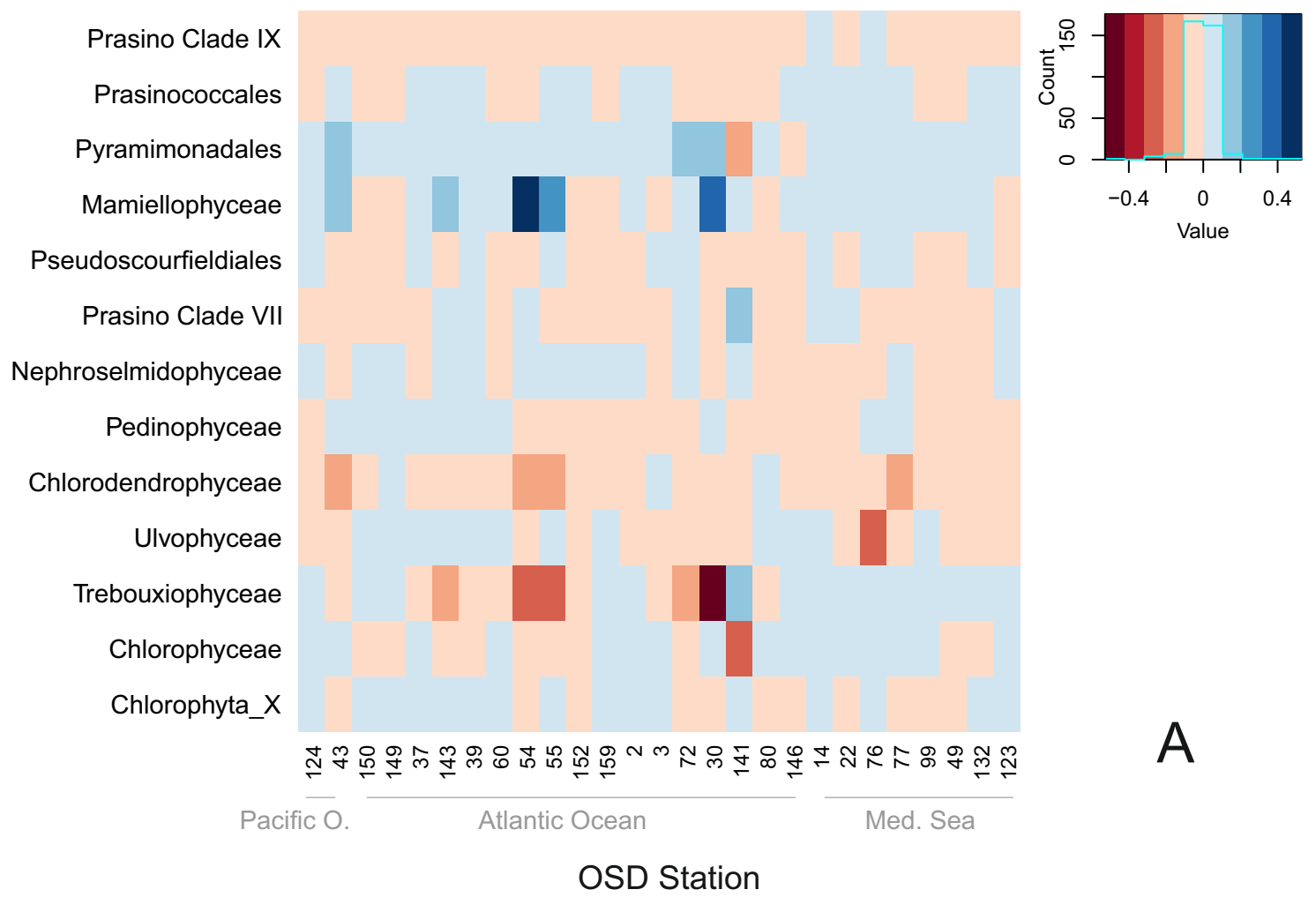
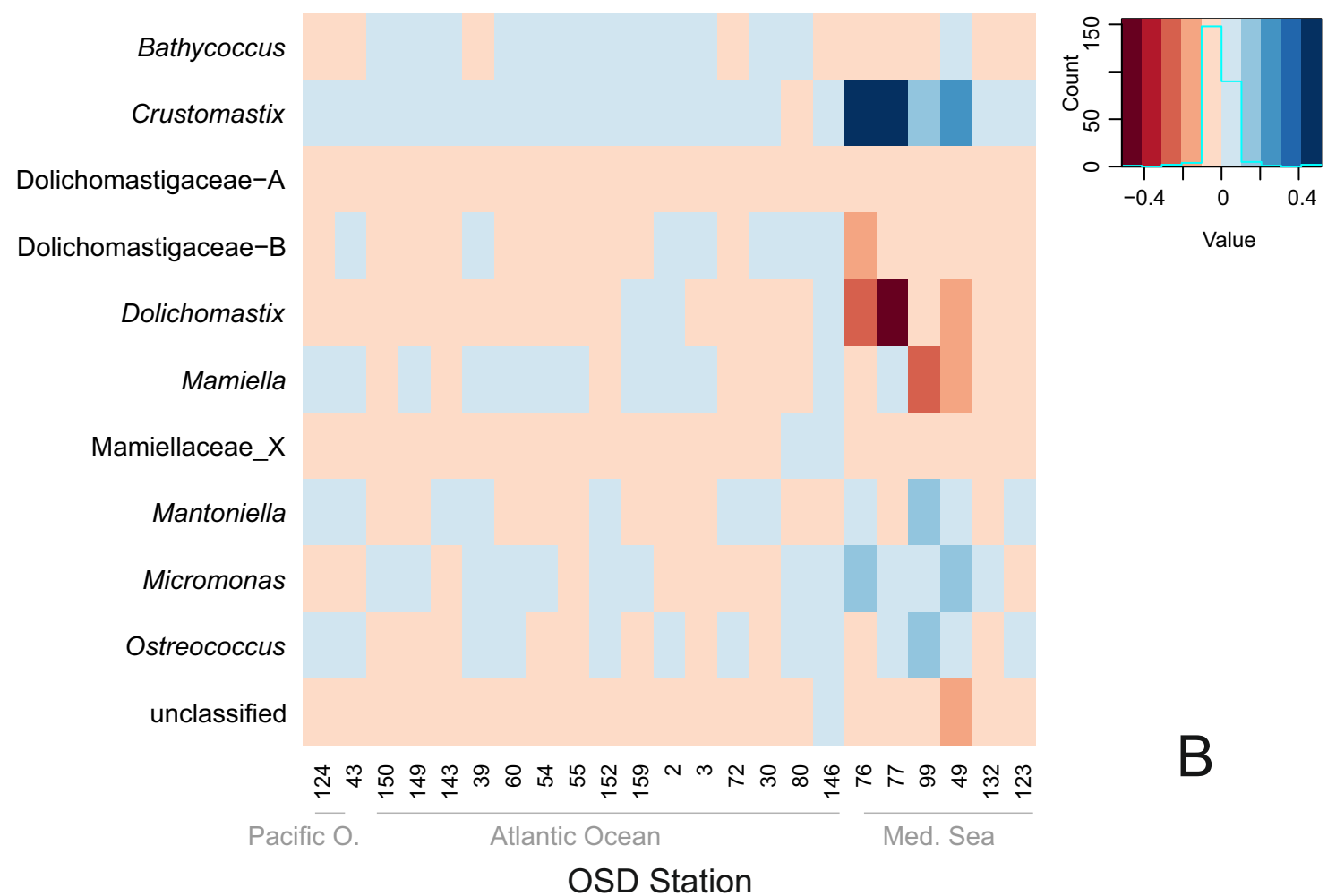


Fig.S9



A



B

Fig.S10

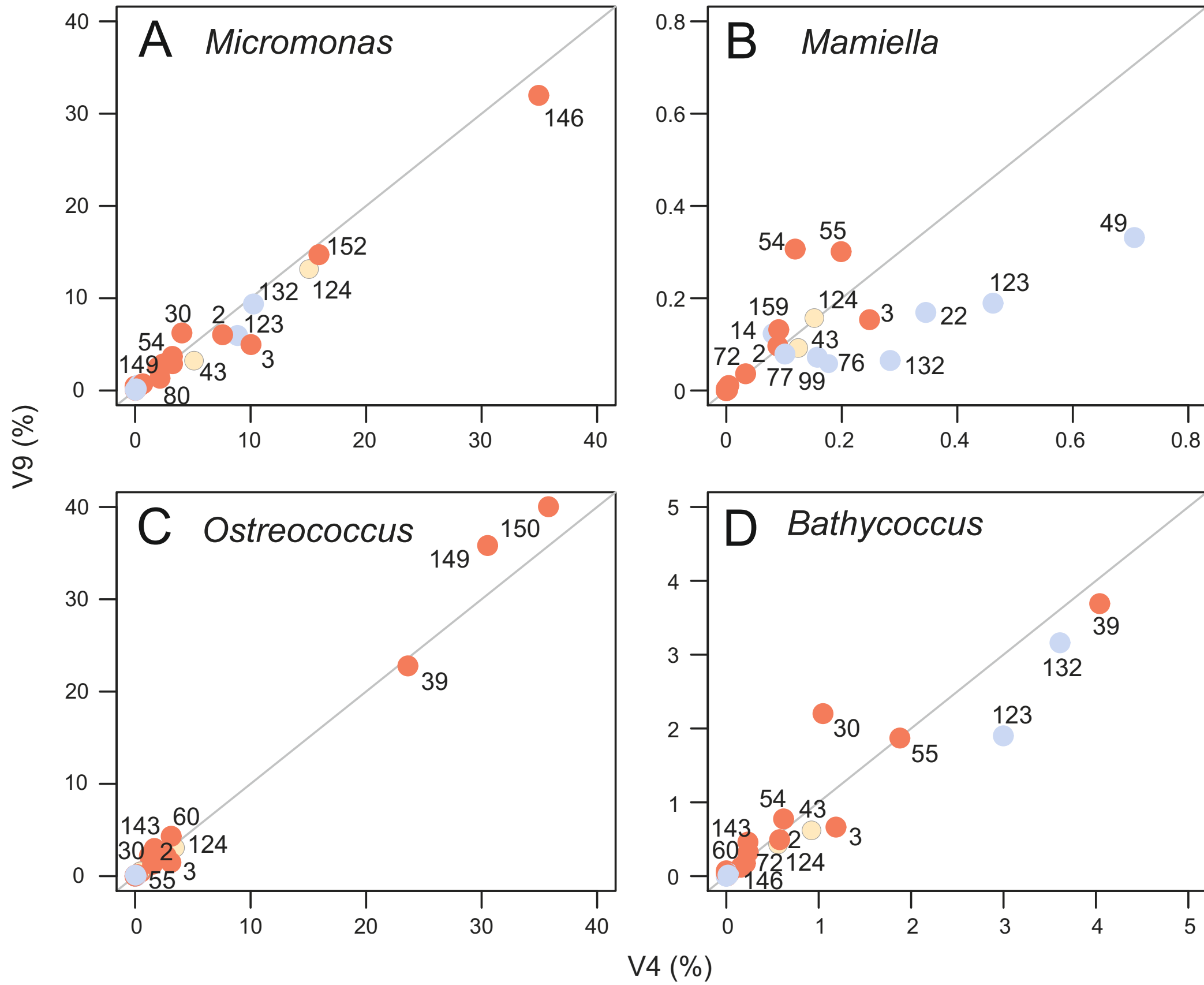


Fig.1

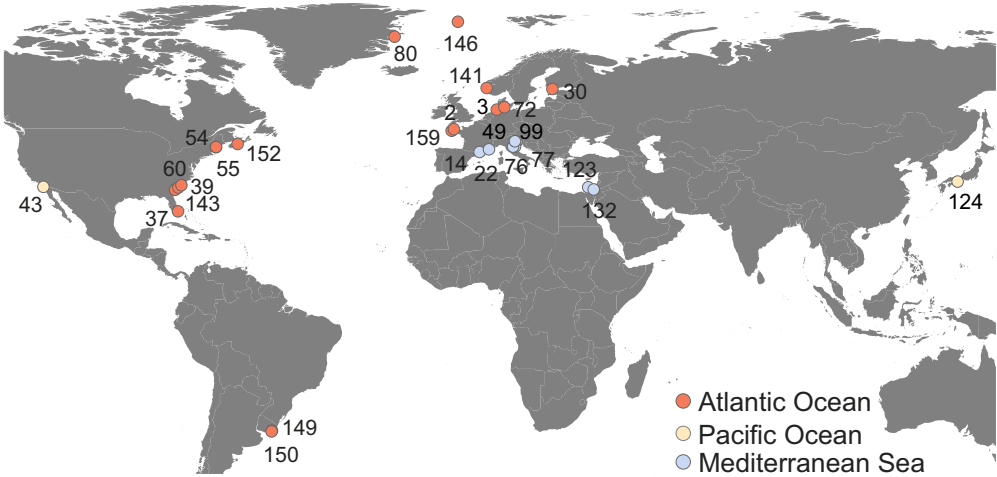


Fig.2

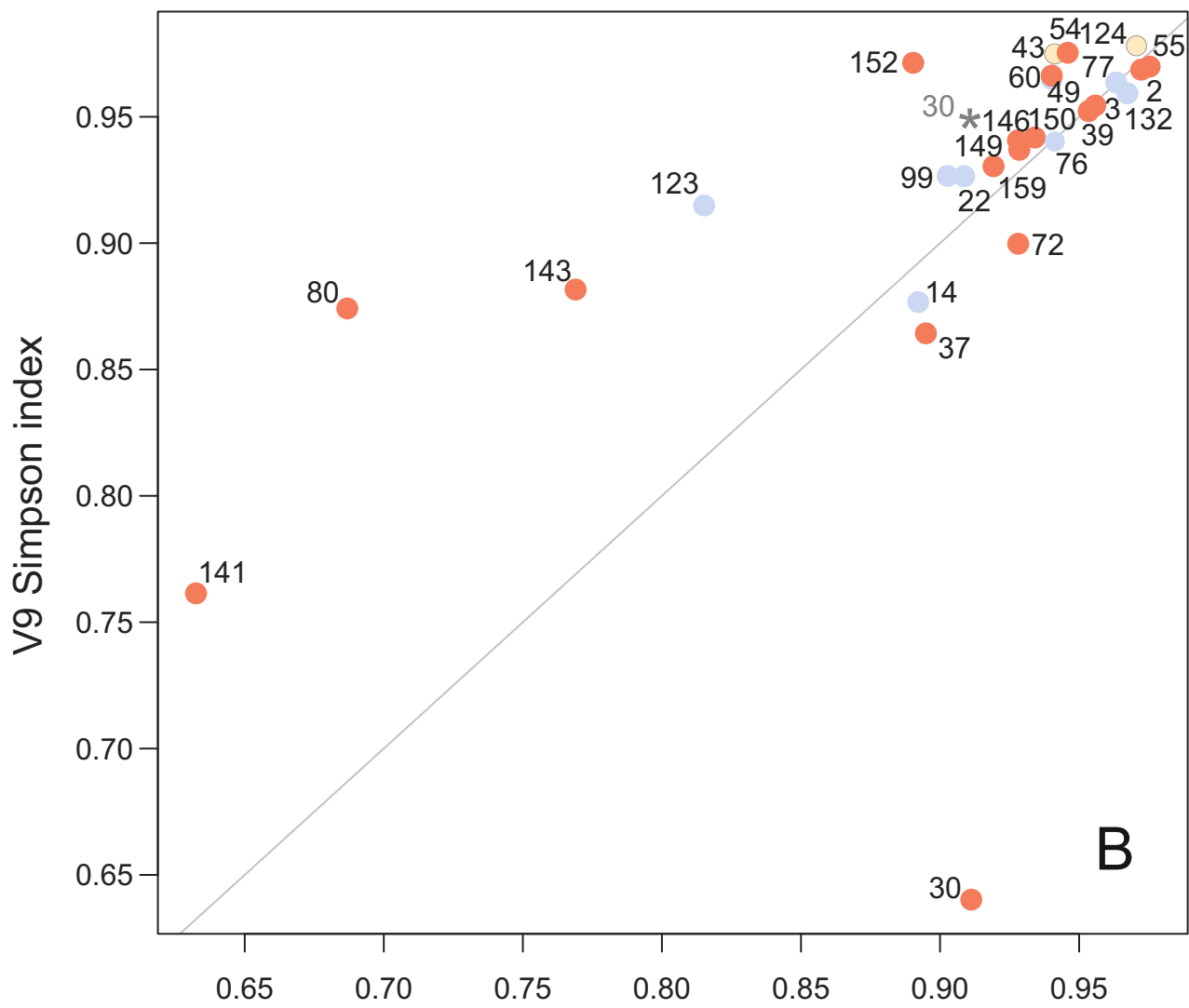
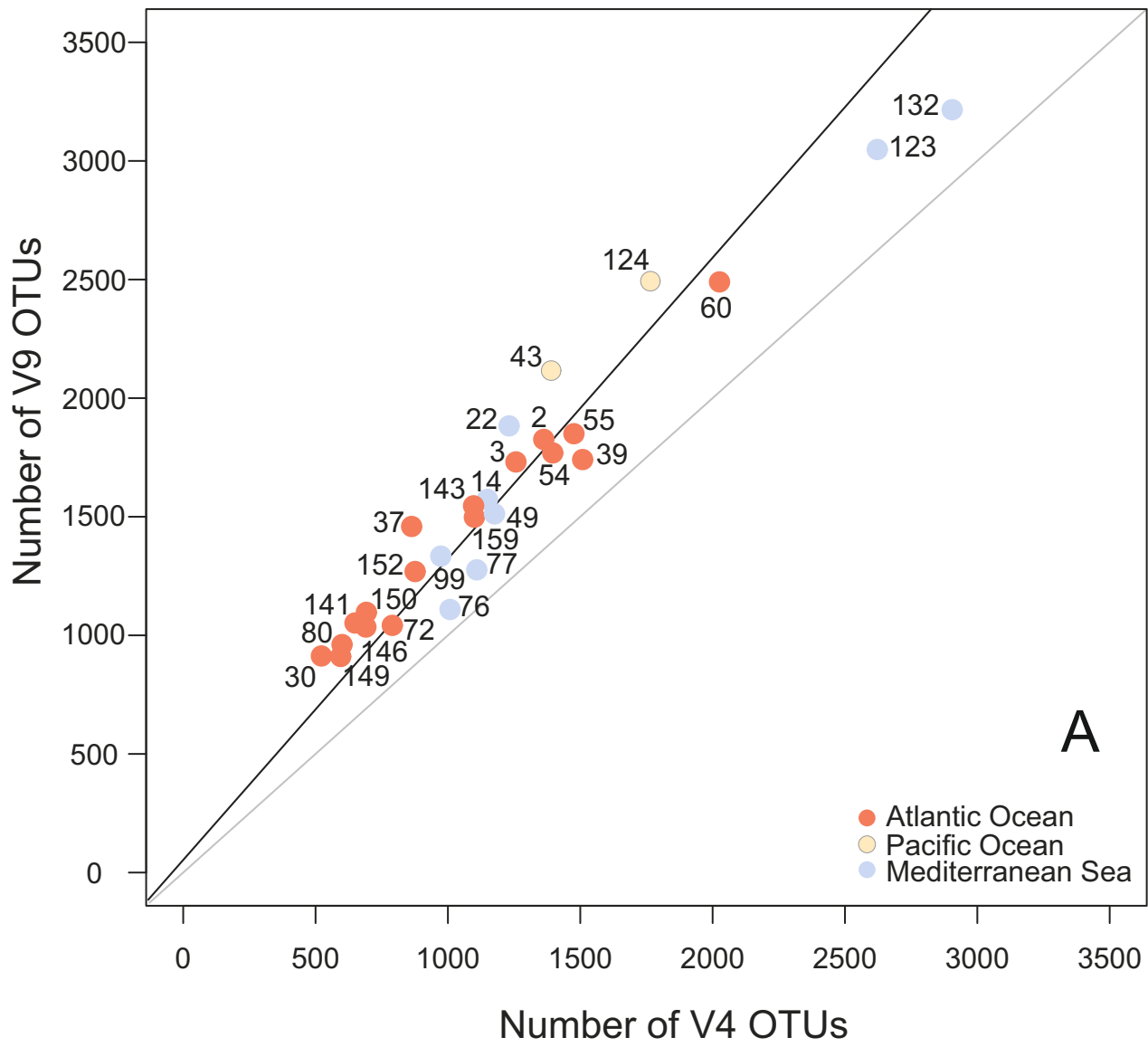


Fig.3

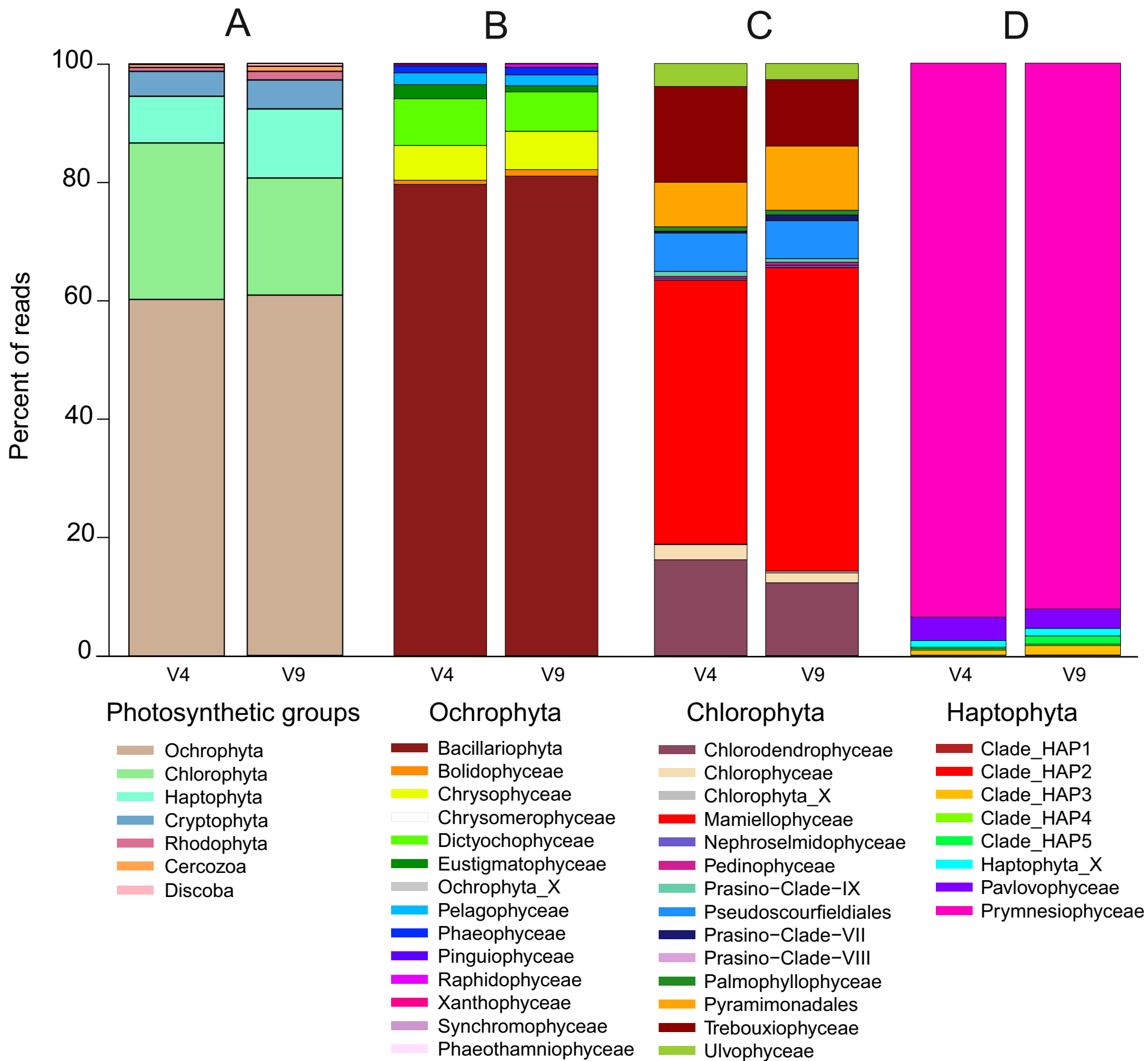
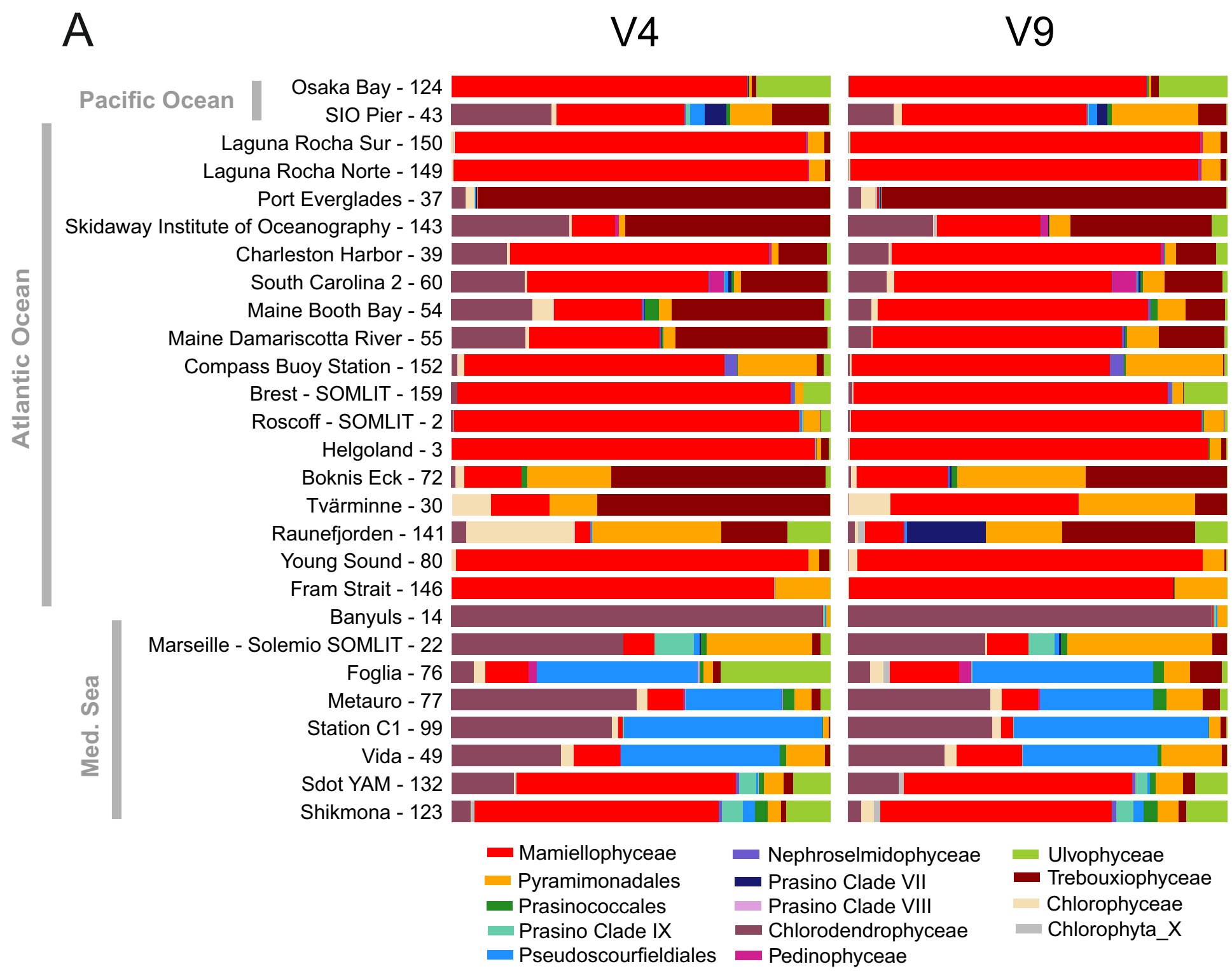


Fig.4



**B**

**Cluster Dendrogram**

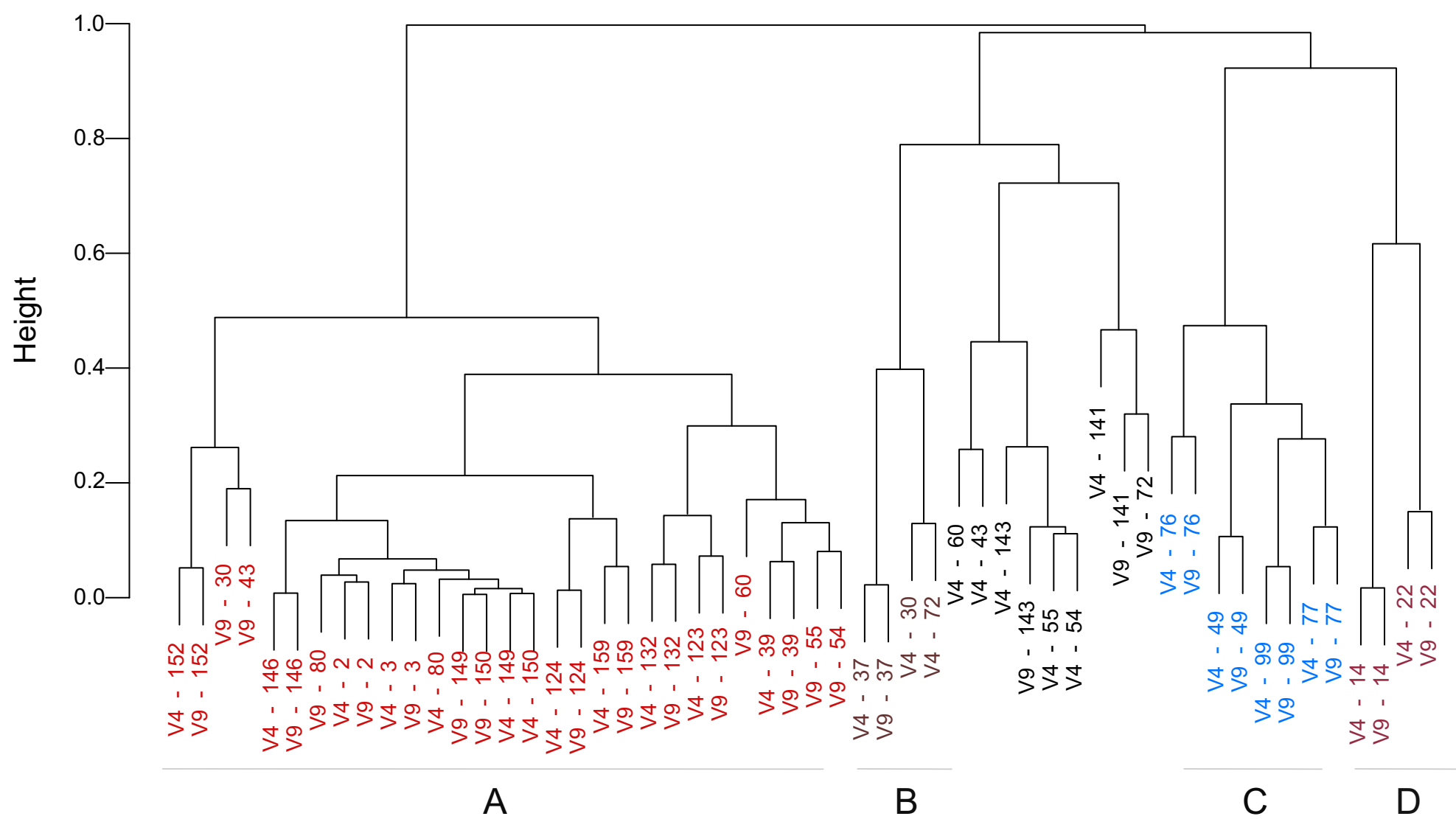
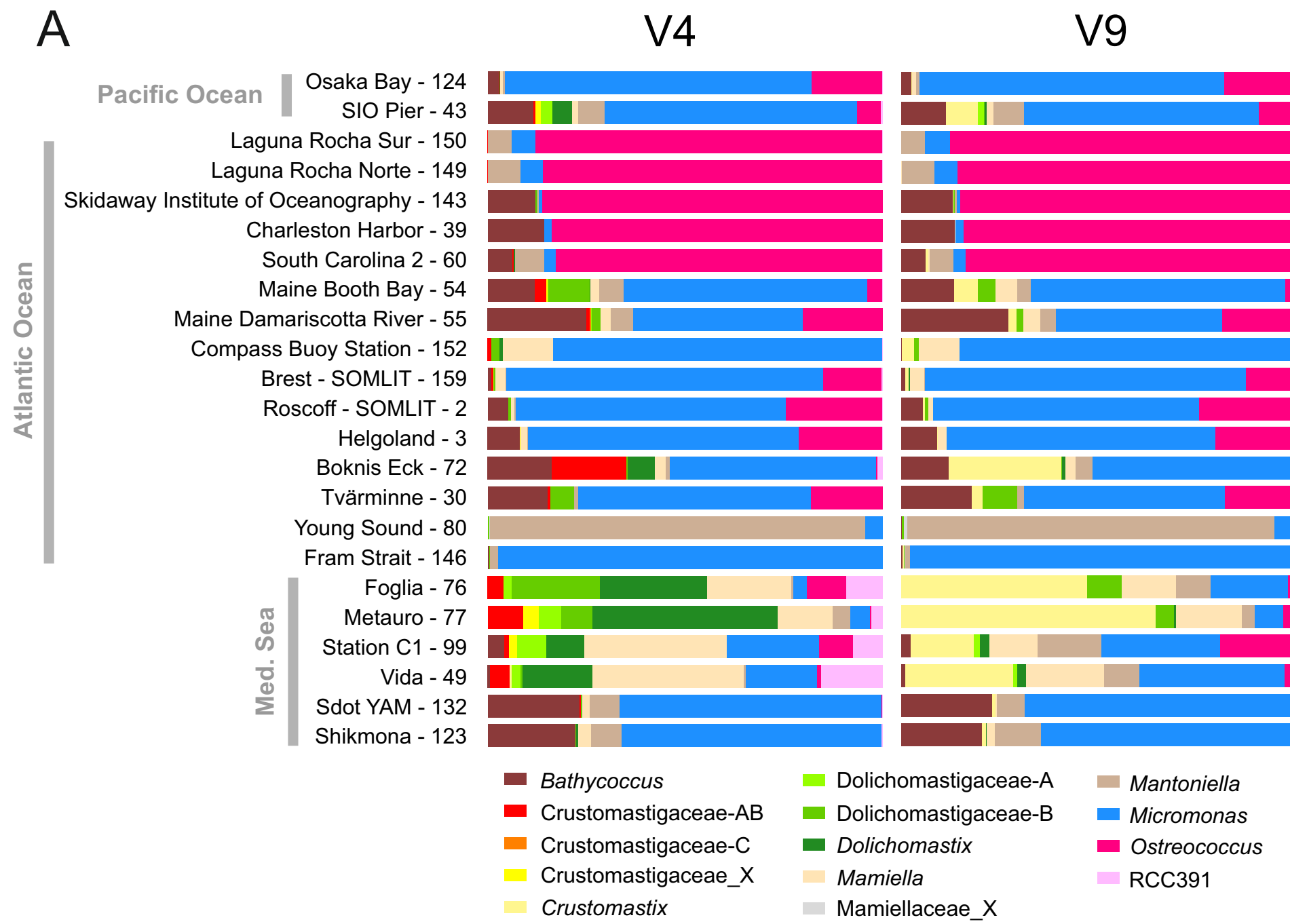




Fig. 5



**B**

**Cluster Dendrogram**

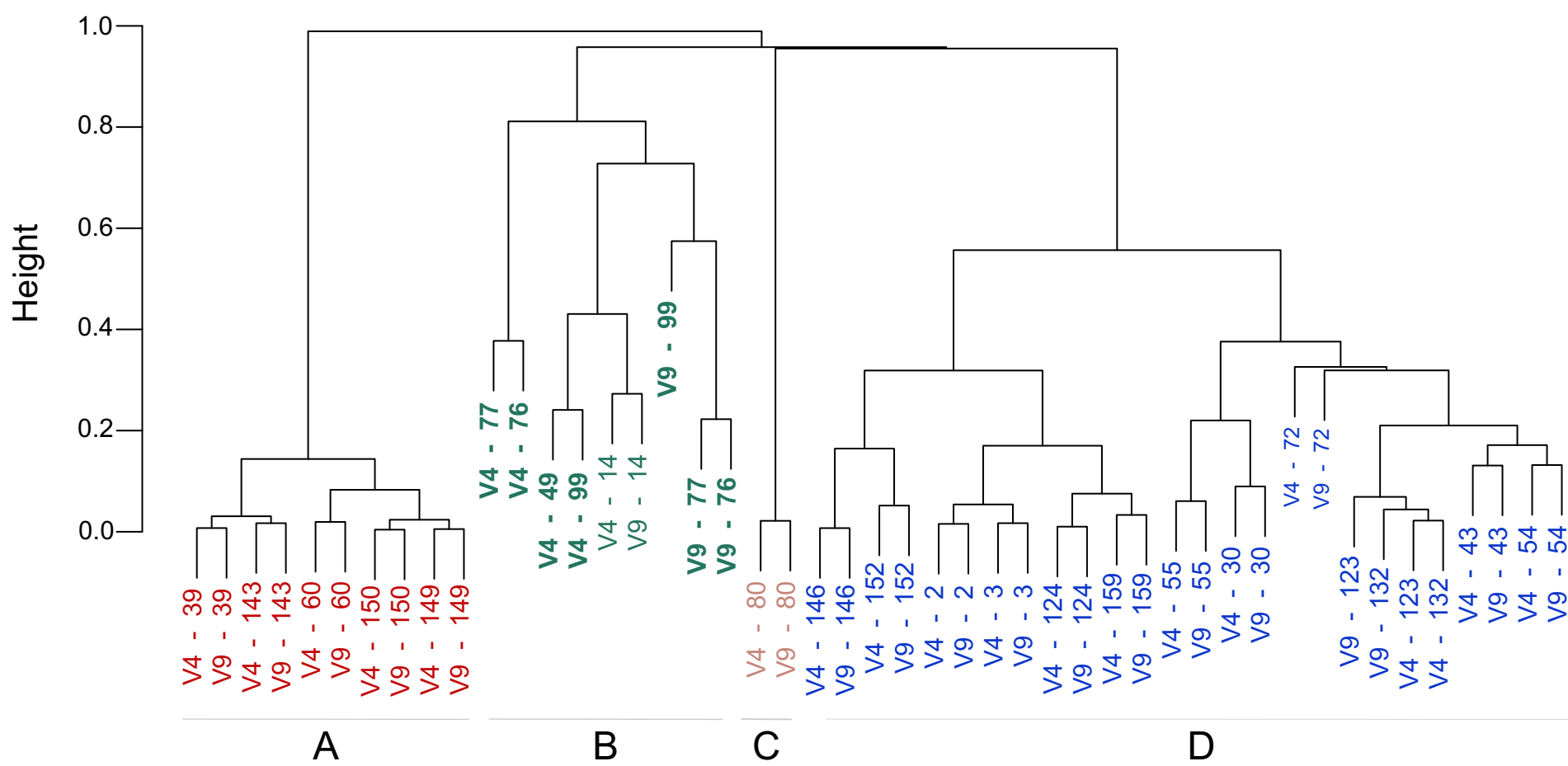


TABLE 1

OSD	Station	Ocean	Region	V4			V9		
				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	353 390	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	Baeford 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

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 finally	3 796 476	4 651 851
6	OTUs (97% similarity)	13 169	16 383

TABLE 3

Parameter	V4				V9				
	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)									
Ochrophyta	91.1	19.3	60.3	18.1	87.6	20.4	64.1	19.8	<b>0.4</b>
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	<b>0.5</b>
Pelagophyceae	7.0	0.0	0.7	1.6	8.0	0.0	0.7	1.6	<b>0.56</b>
Mamiellophyceae	49.6	0.0	12.1	14.2	45.7	0.2	10.9	12.9	<b>0.25</b>
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)									
<i>Micromonas</i>	34.9	0.0	4.8	7.6	32.0	3.00E-02	4.3	6.8	<b>0.5</b>
<i>Mamiella</i>	2.5	0.0	0.2	0.5	1.8	0.0	0.2	0.3	<b>0.3</b>
<i>Ostreococcus</i>	35.8	0.0	4.0	9.6	40.0	0.0	4.4	10.6	<b>0.5</b>
<i>Bathycoccus</i>	4.0	0.0	0.7	1.1	3.7	0.0	0.6	1.0	<b>0.6</b>