Towards an atlas of green microalgae (Chlorophyta) in the ocean

Margot Tragin

THESE DE DOCTORAT DE L'UNIVERSITE PIERRE ET MARIE CURIE

Ecole Doctorale 227 : Science de la nature et de l'Homme

Présentée par

M^me^ Margot Tragin

Pour obtenir le grade de

DOCTEURE DE L'UNIVERSITE PIERRE ET MARIE CURIE

Sujet de la thèse :

Towards an atlas of green microalgae (Chlorophyta) in the ocean

Soutenue le vendredi 15 décembre 2017, devant un jury composé de :

M^me^. Wenche Eikrem, Associate professor UiO University Oslo, Norvège	Rapporteur
M. Frederik Leliaert, Scientific director Botanic Garden Meise, Belgique	Rapporteur
M.Stein Frederiksen, Professor UiO University Oslo, Norvège	Examinateur
M. Christophe Destombe, Professeur Station Biologique de Roscoff, UPMC/CNRS	Examinateur
M. Daniel Vaulot, Directeur de Recherche Station Biologique de Roscoff, UPMC/CNRS	Directeur de Thèse

“On ne peut pas connaître un pays par la simple science géographique… On ne peut, je crois, rien connaître par la simple science. C’est un instrument trop exact et trop dur. Le monde a mille tendresses dans lesquelles il faut se plier pour les comprendre avant de savoir ce que représentent leur somme.”

Jean Giono, *L’Eau Vive*. (1943)

“ […] La mer prodiguait incessamment ses plus merveilleux spectacles. Elle les variait à l’infini. Elle changeait son décor et sa mise en scène pour le plaisir de nos yeux, et nous étions appelés non seulement à contempler les œuvres du Créateur au milieu de l’élément liquide, mais encore à pénétrer les plus redoutables mystères de l’océan. […] Je reconnus immédiatement cette région merveilleuse dont, ce jour-là, le capitaine Némo nous faisait les honneurs. […] Et rien ne pouvait être plus intéressant pour moi que de visiter l’une de ces forêts [1] que la nature a plantées au fond des mers.”

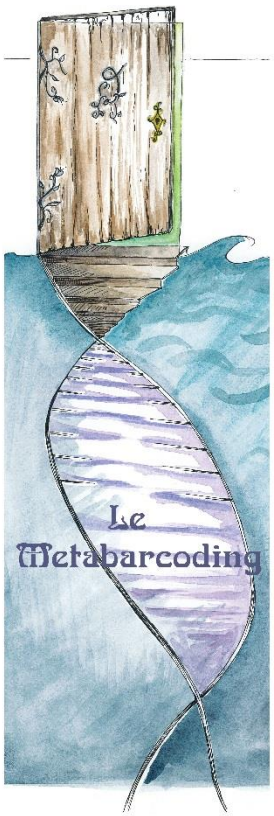
Jules Verne, *Vingt-mille Lieues sous les Mers*. (1869)

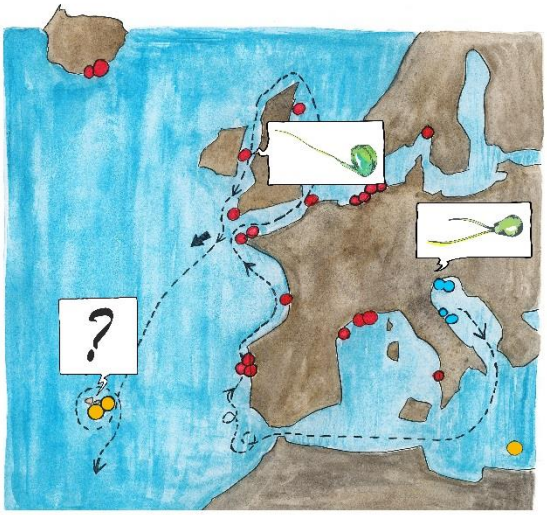
[1] L’adjectif *pétrifiées* désignait initialement les forêts, car Jules Verne parle ici du Corail.

RÉSUMÉS en français :

Ma thèse en 180 secondes, My thesis in 3 minutes

Auditorium de l'Université Pierre et Marie Curie, 18 avril 2017





L'Atlantide et ses jardins :
Empires perdus ?

L'Atlantide et ses jardins : empires perdus ?

"Mais dans le temps qui suivit, il y eut des tremblements de terre et des inondations extraordinaires, et, dans l'espace d'un seul jour et d'une seule nuit néfaste, tout ce que vous aviez de combattants fut englouti d'un seul coup dans la terre et l'île de l'Atlantide, s'étant abîmée dans la mer, disparue de même. "

Platon, *Timée*, (vers 360 avant JC) traduction Emile Chambry.

"Certains se battent contre des moulins à vent, d'autres poursuivent des chimères et moi, je compte… Depuis deux ans, j'explore la diversité botanique dans les jardins de la cité perdue de l'Atlantide. Ces jardins, c'est l'océan et les plantes auxquelles je m'intéresse, ce sont des microalgues vertes appelées Chlorophyta. Pour ouvrir les portes de ce monde englouti invisible, la clé est un mot : le métabarcoding.

Il existe des morceaux d'ADN présents chez tous les organismes vivant avec quelques différences qui nous permettent de les utiliser comme marqueurs pour étudier la diversité d'organismes, trop petits pour être distingués à l'œil nu ou même au microscope.

En pratique, il faut avant tout choisir où l'on veut ouvrir les portes de l'océan. Alors on organise des missions océanographiques et chercheurs de *toute écaille* embarquent pour l'aventure. Une fois sur place, nous capturons d'abord les micro-organismes en filtrant de l'eau de mer, puis après avoir extrait tout leur ADN, nous récupérons les marqueurs qui nous intéressent. Ensuite, les machines interviennent pour traduire les marqueurs, qui sont sous forme de molécules d'ADN, en séquences de lettres que nous, humains, pouvons lire. Cette étape peut produire plusieurs millions de séquences. La dernière étape du métabarcoding consiste à comparer ce que l'on a trouvé avec des herbiers mis au point par de précédents explorateurs de la diversité. Parmi ces résultats, il n'y a que les Chlorophyta qui m'intéressent. Enfin, je sais : qui vit où ! en quelle proportion ! qui était déjà connu et surtout… qui est nouveau ! Et ces tout petits nouveaux, je peux, à mon tour, les ajouter aux herbiers.

A présent que la clé est tournée, prenez une grande inspiration… Et suivez-moi à travers les portes ouvertes par le projet européen Ocean Sampling Day…

Tout commence en Méditerranée, berceau des mythes et légendes antiques. Laissez-moi vous présenter les timides *Pycnococcus* en bleu. Ils dominent l'Adriatique. Attention ! Restons groupés, car le courant nous emporte au-delà du Détroit de Gibraltar… Nous dérivons à présent le long des côtes et ici, nous faisons la connaissance des Mamiellophyceae en rouge. Elles sont très fières de nous présenter leur superstar *Micromonas*, toute petite, mais omniprésente jusque dans les eaux polaires. Egarés par tant de liesse, nous demandons notre chemin et on nous indique que le palais se situe au centre des

jardins. Alors, nous continuons notre équipée vers les eaux océaniques… Là, nous ne rencontrons plus que le mystérieux et insaisissable groupe IX en orange.

Notre voyage s'arrête ici car je cherche toujours… mais, je vous ai présenté mes premiers pas vers Atlas, roi qui fit de l'Atlantide la cité la plus prospère du monde et vers *un* atlas des Chlorophyta de l'océan."

RÉSUMÉ de la thèse (en français)

Vers un atlas des micro algues vertes (Chlorophyta) dans l'océan.

La lignée verte (i.e. les végétaux), qui domine sur terre grâce aux plantes terrestres, est représentée dans l'océan par les algues de la division des Chlorophyta. Celles-ci contribuent en moyenne à 25% des séquences photosynthétiques (Dinoflagellés exclus) retrouvées dans les inventaires moléculaires pan-océanique. Plusieurs lignées de Chlorophyta (i.e. les prasinophytes) partagent des caractères morphologiques ancestraux, tels que la présence d'écailles et sont considérés comme pouvant être proche de l'ancêtre commun de la lignée verte (i.e. « le flagellé vert ancestral », l'AGF, the ancestral green flagellate). Bien que les Chlorophyta jouent un rôle important dans l'écologie de l'océan et nous permettent de comprendre l'histoire évolutive des plantes terrestres, leur diversité et leur distribution dans les eaux marines du globe a été fort peu documentée.

Les objectifs de cette thèse de doctorat sont d'étudier la diversité environnementale des Chlorophyta marines et de décrire leur distribution grâce à des données globales déjà existantes issues du métabarcoding.

Dans un premier temps, toutes les séquences publiques du gène de l'ARNr 18S appartenant à des Chlorophyta ont été rassemblées dans une base de données. L'assignation taxonomique de ces séquences de référence a été soigneusement vérifiée. Dans un second temps, j'ai procédé à l'analyse des jeux de données de métabarcodes produits par le projet européen Ocean Sampling Day (OSD) qui a échantillonné en 2014 un grand nombre de sites marins, principalement côtiers. Le consortium OSD a fourni des données utilisant 2 régions hypervariables du gène de l'ARNr 18S appelées V4 et le V9, communément utilisées comme métabarcodes. Une comparaison a été menée sur un sous ensemble d'échantillons dans le but d'étudier les différences potentielles de diversité du phytoplanctonique estimée par ces deux marqueurs, en s'appuyant sur les Chlorophyta pour lesquelles la littérature était déjà maitrisée. Cette comparaison a illustré l'influence de la base de données de référence sur l'image de la diversité aux niveaux taxonomiques faibles (genre, espèce). Ensuite, l'ensemble des données utilisant la région V4 comme marqueur a été utilisé pour analyser la distribution des Chlorophyta dans l'océan mondial. L'assignation automatique des OTUs (Operationnal Taxonomic Unit) grâce à la base de référence produite lors de la première étape de la thèse et la vérification de ces assignations par reconstruction phylogénétique ont permis de confirmer l'existence de nouvelles lignées environnementales de prasinophytes et de décrire des patterns écologiques. Ces analyses ont confirmé que la classe des Mamiellophyceae dominait les eaux côtières et mis en lumière que les clades VII et IX

des prasinophytes dominaient les milieux océaniques oligotrophiques échantillonnés pendant l'OSD. Elles ont aussi permis de montrer l'écart entre la diversité présente dans les bases de données de référence et la diversité environnementale en particulier pour les genres *Ostreococcus* et *Micromonas* (Mamiellophyceae) qui sont les Chlorophyta marines les plus étudiées.

Ces travaux soulignent ainsi l'importance négligée des Chlorophyta dans le milieu marin et suggèrent de nouvelles pistes pour poursuivre de futures recherches.

Mots clefs : Métabarcoding, Marqueurs, gène 18S de l'ARNr, régions V4 et V9, Chlorophyta, Prasinophytes, diversité, distribution géographique, phylogénie, milieu marin.

Table of content

Introduction

The planktonic compartment

The Ocean hosts a wide range of organisms from the largest known on earth (the whale *Balenoptera musculus*, 30-meter-long and 170 tons) to the smallest ones (bacteria). Some are so light or so tiny that they drift with currents: they constitute the plankton… It is composed of numerous life forms that are very diverse in morphology, size (from less than 1µm up to several meters for the largest jellyfishes), physiology and ecology. Planktonic organisms and in particular the unicellular eukaryotes (i.e. the protists) are distributed throughout all branches of the tree of life (Baldauf, 2008; Burki, 2014) and can be heterotrophic, mixotrophic or photosynthetic (i.e. the phytoplankton). Inside protists, five major photosynthetic divisions dominate in the ocean:

- The Dinophyta (Alveolata, Fig.1) where 50% of the species have plastid (Gómez, 2012) among which some truly photosynthetic species such as the harmful blooming algae *Alexandrium minutum* and mixotrophic ones. The other half are heterotrophic such as the bioluminescent predator *Noctiluca scintillans*, which hosts green microalgal endosymbionts *Pedinomonas noctilucae* (Gómez, 2012; Wang *et al.*, 2016).
- The Ochrophyta (Stramenopiles, Fig.1) are mostly represented by the diatoms (i.e. microalgae with silicate extracellular test) such as the toxic diatom *Pseudo-nitzschia* (Anderson *et al.*, 2011) or the chain forming diatoms *Chaetoceros* and contain less abundant classes such as the Bolidophyceae (Guillou *et al.*, 1999).
- The Chlorophyta (Archaeplastida, Fig.1) are the marine representatives of the green lineage which dominates on earth (cf. the first chapter)
- The Haptophyta are autotrophic and mixotrophic algae (Unrein *et al.*, 2014), some species are calcified (coccolithophorids) such as the well-studied *Emiliana huxleyi* (Frada *et al.*, 2012). Coccolithophores calcified test morphologies are currently particularly investigated to understand the responds to marine environments acidification (Langer *et al.*, 2009; Meyer and Riebesell, 2015).
- The Cryptophyta are flagellated microalgae containing phycoerythrin accessory pigments (Novarino, 2003) such as the *Teleaulax* genus (Hill, 1991) or the *Plagioselmis prolongate* species (Novarino *et al.*, 1994).

Microalgae in the larger size fractions (nanophytoplankton from 3 µm to 20 µm and microphytoplankton from 20 µm to 200 µm) frequently get themselves noticed through colored, and sometimes toxic, water events such as red tides (for example dinoflagellates blooms) or milky turquoise blue waters (calcified Haptophytes blooms). Colored water events could be recorded and further analyze by remote sensing imaging (https://science.nasa.gov/earth-science/oceanography/living-ocean/ocean-color) and are the indicator of phytoplanktonic bloom dynamics, which corresponds to short time rapid growing periods, when environmental conditions get favorable, especially during spring and summer.

In contrast, the smallest size fraction (i.e. the picophytoplankton from 0.2 µm to 3 µm) combine bloom dynamics and continuous presence during the whole year. Some pico-phytoplanktonic algae present all-year-long in marine waters (for example the green micro algae *Micromonas*) complement marine cyanobacteria (*Synechococcus* and *Prochlorococcus*) forming the background of primary production and carbon fixation in the ocean (Li, 1994; Jardillier *et al.*, 2010). The role of eukaryotic picophytoplankton is especially important at high latitudes, where marine cyanobacteria are absent (Lovejoy *et al.*, 2007; Balzano *et al.*, 2012; Flombaum *et al.*, 2013).

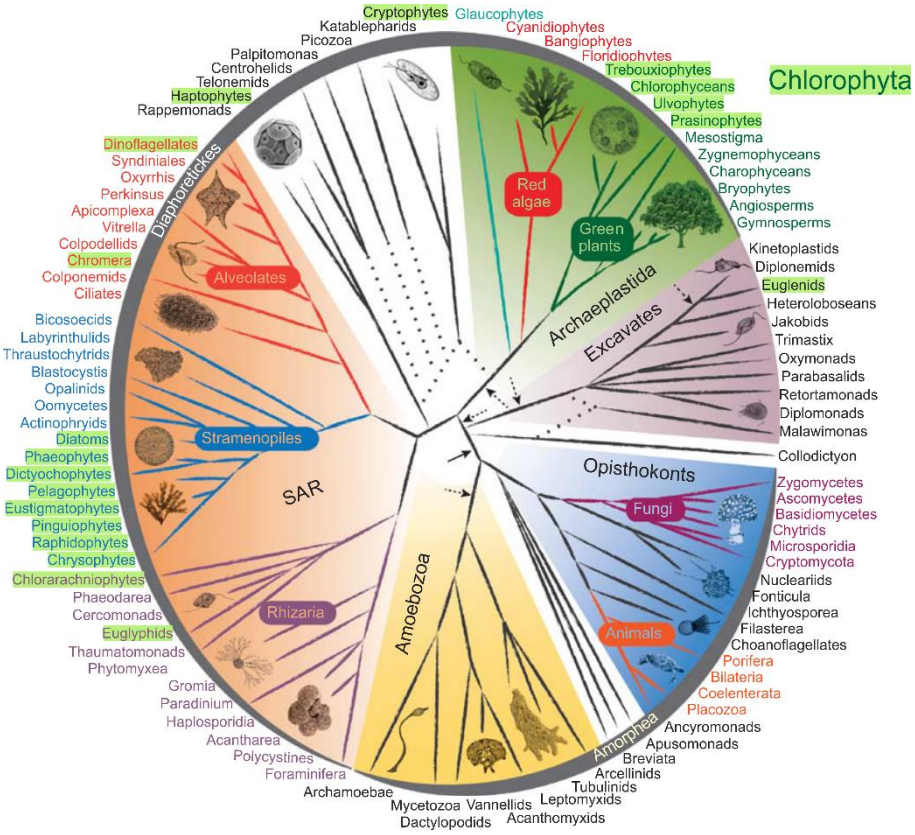


Fig.1: Global tree of eukaryotes from a consensus of phylogenetic evidence (in particular phylogenomics), rare genomic signatures, and morphological characteristics from (Burki, 2014). Photosynthetic lineages found in phytoplankton are highlighted in green.

The oceanic distribution of photosynthetic divisions in the ocean is far from being understood. For Chlorophyta for example, most of distribution survey focused on limited regions such as the Mediterranean Sea (Viprey *et al.*, 2008) or Arctic Ocean (Balzano *et al.*, 2012) or studies focused on specific Chlorophyta lineages (in general the Mamiellophyceae, Monier *et al.*, 2016).

Delimiting biogeographic units in the Sea

Biogeography explores biological diversity in relation with space units. In order to study the ecology of specific organisms, the most accepted approach for delimiting environment is the concept of

ecological niches (Hutchinson and MacArthur, 1959), which are defined by a combination of environmental parameters. In the continuously moving ocean, this concept is more difficult to put into practice and biogeographic generalizations to the oceans are not easy.

In terrestrial ecology, the continental Earth has been separated into several big units named biomes (Clements, 1916) such as steppes, savanna, mountains, temperate or rain forests. These ecological units are an ensemble of ecosystems common to the same biogeographical region which delimitation is based on environmental and biological parameters (for example temperature, rain falls and vegetation…, Shelford, 1929). In 1975, a UNESCO report (Udvardy, 1975) dealing with a classification of biomes in the world only mentioned one for the hydrosphere: lake biome . Aquatic biomes have now been defined for both freshwater (ponds and lakes, rivers and wetlands) and oceanic waters (oceans, coral reefs and estuaries, http://www.ucmp.berkeley.edu/glossary/gloss5/biome/aquatic.html). In comparison to land, the diversity of aquatic biomes and especially oceanic ones, which represents around 70% of the earth surface, looks however particularly small and imprecise.

However, Ekman (1935) and then Briggs (1974) already saw the ocean as an ensemble of regions, sub regions and provinces. Ekman introduced the division of the marine environment into warm, temperate and polar waters. These authors both based their work on the marine fauna to delimit consistent space units. This delimitation was however subjective, partly depending on the organisms studied, and is challenged each time new diversity or distribution data become available (Briggs and Bowen, 2012). For these reasons, they have not been universally used by marine biologists and ecologists. In fact, the use of fixed biogeographic regions is difficult to transpose to a continuously moving environment: the ocean is largely impacted by seasonal cycles and atmospheric conditions. For example, the wind, induces strong water column mixing in winter, changing the environmental conditions. Longhurst (1995, 2007) described the oceanic biogeography in link with seasonal cycles and separated the marine waters into four major biomes (polar, westerly wind, trade wind and coastal) and 56 provinces (Table 1). More recently, Oliver and Irwin (2008) proposed modeling of oceanic biomes based on remote sensing data (surface temperature and ocean color), which allowed to deal with the moving oceanic biogeographic province boundaries in time and space and provided objective limits to biogeographical provinces in the Sea. Reygondeau *et al*. (2013) modelled ideal biogeographic regions: for each unit, environmental conditions should be distinguishable from the others and unique at a global scale. Based on Longhurst work these authors implemented four parameters (bathymetry, Chlorophyll *a*, surface temperature and salinity) in their model to delimit the 56 biogeographic regions in space and time (Fig.2).

While these models provide working hypotheses for biogeographical partitioning of the world ocean, biodiversity studies are often conducted locally at small scales where these divisions are not

always easy to put into practice. As an example, in coastal environments, the impact of rivers, the lagoons and the marshes create a space and time mosaic within described biogeographic regions. In the ocean, results are often interpreted in terms of water masses (Hamilton *et al.*, 2008; Kirkham *et al.*, 2011, for example). In highly dynamical regions such as in the Arctic ocean (Dunbar, 1953) for example, water masses from the Pacific and Atlantic Oceans often co-occurs before getting mixed. Each oceanic region has their own signatures in terms of nutrient concentrations, salinity or temperature, which allow to deduce the origin and history of water masses.

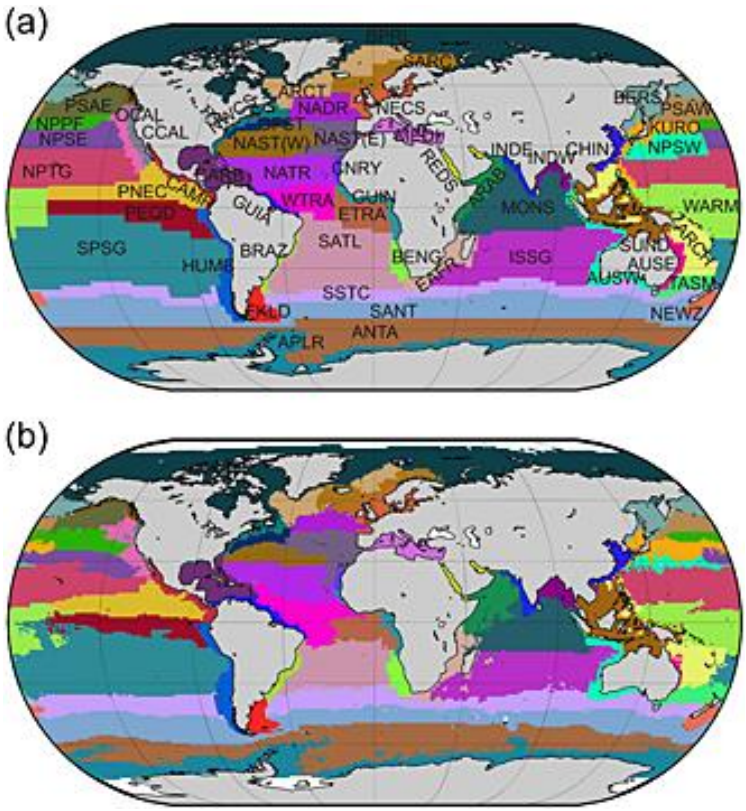


Fig.2: Biogeography of the global ocean (i.e. biogeochemical provinces) proposed by (a) Longhurst (2007) and (b) calculated in (Reygondeau et al., 2013) for the average period from January 1998 to December 2007. Letters corresponds to the abbreviation code summarized in Table 1. From (Reygondeau et al., 2013).

Questions addressed by the thesis

The aim of this thesis is to study the biogeographic patterns of Chlorophyta lineages in the marine waters: where and in which proportion are the Chlorophyta lineages occurring in the global ocean? Do different Chlorophyta lineages co-occur or not? Is there a relation between observed distributions and environmental parameters? Are species of green microalgae distributed in coherent manner in the global ocean?

In order to meet this general aim and answer these questions, we chose to use the most widely used marker gene today, the nuclear encoded18S rRNA gene for which both a large number of reference sequences from described species but also from the environment are available. My first step was to

review the current knowledge on Chlorophyta distribution and diversity in marine waters based on 18S rRNA. This step allowed setting up an accurate 18S database, needed for the subsequent steps. The second step was to validate metabarcoding approaches in particular comparing how two different markers (V4 and V9 regions of the 18S rRNA) compared to assess Chlorophyta diversity on a large data set, the Ocean Sampling Day (OSD). The third step was to describe oceanic Chlorophyta communities at wide taxonomic levels (class) and to document environmental preferenda using the OSD dataset. The fourth and final step was to look at the diversity of specific Chlorophyta groups (Mamiellophyceae) at lower taxonomic levels (species and below) and to determine whether these units had coherent oceanic distributions.

Investigating marine protist diversity with molecular biology approaches

Optical microscopy allows to describe to a certain extent the morphological diversity of representatives of the larger phytoplankton size fractions (for example the diatoms), but does not allow to determine the full extent phytoplankton diversity, especially for the smallest size classes such as picoplankton. As an example, the smallest photosynthetic protist known *Ostreococcus tauri*, is a picophytoplanktonic green microalgae belonging to the Chlorophyta has the size of a bacterium around 1 µm (Courties *et al.*, 1994; Chrétiennot-Dinet *et al.*, 1995). The development of electron and epifluorescence (Hobbie *et al.*, 1977) microscopical techniques improved the detection, quantification and morphological description of the smallest phytoplankton size fraction (Manton, 1959; Melkonian and Preisig, 1986; Moestrup and Throndsen, 1988). However, microscopy based inventories remain time consuming and taxonomically imprecise since many microalgal species consist of complex of cryptic species (Šlapeta *et al.*, 2006). Phytoplanktonic organisms possesses a collection of pigments, which are different between photosynthetic lineages (i.e. fucoxanhin in diatoms, prasinoxanthin in some green microalgae, Wright and Jeffrey). Pigment composition detected for example by High Pressure Liquid Chromatography (HPLC for example see Coupel *et al.*, 2014) and flow cytometry (Marie *et al.*, 1999) have been used to distinguish different phytoplankton populations. The separation brought by these methods remain however quite coarse (e.g. at best the Class level for pigments, or the size class for flow cytometry with additional discrimination based on presence of specific pigments such as phycoerythrin).

Throughout evolutionary time, life have been diversified, but all life forms share “universal” genes and proteins presenting certain degrees of variability (Chenuil, 2006), which allow them to be used as molecular markers to access biological diversity. Molecular markers can be split into three categories (Schlötterer, 2004): the proteins variants (i.e. allozymes), the DNA (Desoxyribo Nucleic Acid) sequence variations (i.e. polymorphisms) and the DNA repeat variation. The development of DNA manipulation tools such as the isolation of restriction enzymes in the 1960s by Arber, Smith and Nathans, the sequencing techniques (for example the Sanger method, Sanger and Coulson, 1975) and gene amplification using polymerase chain reaction (PCR, Saiki *et al.*, 1985) brought remarkable

progress in molecular biology and allowed the development of the use of DNA marker polymorphism for diversity surveys.

Several DNA approaches were developed to assess protist diversity, which can be presented as quantitative, qualitative and semi quantitative methods. Qualitative methods allow the precise quantification of a small number of known taxa. For example, FISH (Fluorescent In Situ Hybridization) and qPCR (quantitative PCR) rely on the design of oligonucleotide probes from known reference sequences. FISH uses marked fluorescent oligonucleotide to hybridize DNA inside cells; as a result fluorescent cells can be counted under a microscope (for example in Not *et al.*, 2002). QPCR uses both primers and fluorescent probes in order determine the number of copies of a given gene in a natural sample (for example in Zhu *et al.*, 2005; Demir-Hilton *et al.*, 2011). In contrast, qualitative approaches are capable of unveiling unexpected new diversity. Among these, DGGE (Denaturing Gradient Gel Electrophoresis, Díez *et al.*, 2004) and TRFLP (Restricted Fragment Length Polymorphism, for example Balzano *et al.*, 2012; Treusch *et al.*, 2012) rely on DNA fragments size differences. The construction of clone library coupled to Sanger sequencing (Moon-van der Staay *et al.*, 2001; for examples Guillou *et al.*, 2004; Viprey *et al.*, 2008; Massana *et al.*, 2015) allows to assess the taxonomic diversity present in a sample producing high quality long sequences that constitute reference database. However, these methods cannot reliably relate the number of sequence and the abundance of organisms in the sample, because the number of sequences sampled always remain low.

In recent years, the so-called "metabarcoding" method has been more and more used to make molecular inventories. It consists in amplifying and sequencing the same marker gene from all organisms using DNA sampled in the environment. The sequences found in environmental samples are then compared to the sequences available in reference database, for which the organisms of origin are known. This step is named the assignation and call on several bioinformatics methods such as Basic Local Alignment Search Tool (BLAST, Altschul *et al.*, 1990) or software intrinsic function such as the *classify.otu* in the mothur software (Schloss *et al.*, 2009). Metabarcoding relies on the comparison of unknown versus known sequences. The PCR amplification step is biased by the number of marker gene copies per organisms (Prokopowich *et al.*, 2003), which is linked to the organisms biomass (Zhu *et al.*, 2005), within photosynthetic organisms the range of copy number is quite limited. For example, the picophytoplanktonic *Ostreococcus tauri* (Chlorophyta) only have four copies of the commonly used 18S rRNA marker gene (Derelle *et al.*, 2006). For photosynthetic groups (Dinoflagellates excluded) metabarcoding is considered as a semi quantitative method and the relative contribution of each groups is commonly considered to provide a good image of the community

Metabarcoding

In the last decade, the development of high throughput sequencing (HTS) technologies such as 454 pyrosequencing (based on the work of Nyren *et al.*, 1993; Ronaghi *et al.*, 1998) in 2005 (Margulies

et al., 2005) and then the Illumina sequencing (based on the work of Canard and Sarfati, 1994) in 2007 allowed the transition between clone library sequenced by Sanger method and the large metabarcoding datasets. Sanger sequencing method provides a relative low number of long high-quality sequences, while HTS provides a large amount of medium-quality sequences and allows only small fragments to be sequenced. Recurrent sequencing errors were rapidly noticed in 454 sequencing leading to the development of denoising bioinformatics tools such as denoiser.py (Reeder and Knight, 2010) implemented in commonly used pipeline QIIME (Caporaso *et al.*, 2010) and the *shhh.seqs* function in Mothur (Schloss *et al.*, 2009). Bioinformatics software, such as QIIME or Mothur, consist of a set of bioinformatics tools used to treat large sequences datasets (for example length and quality filters, chimera detections, alignments, clustering programs…). Both Mothur and QIIME are currently used to deal with metabarcoding datasets and provide equivalent possibilities (Nilakanta *et al.*, 2014). The software, the order in which programs are called, the parameters etc. can impact the final results depending on the questions which are addressed. The pipeline used in this thesis is explained in the second chapter material and methods part and parameters choice are discussed in relation to our objectives. Several studies compared bioinformatics protocols (Majaneva *et al.*, 2015; Ferrera *et al.*, 2016), but no universal solution exists.

Initially, the limited size of reads sequenced by HTS (around 110 bp for 454 and 40 for Illumina, van Dijk *et al.*, 2014) forced the scientific community to use small markers (for example the V9 hypervariable regions of the 18S rRNA gene as in Amaral-Zettler *et al.*, 2009). In recent years, longer reads became possible (up to 700 bp, but commonly around 500 bp with the 454 and 2*300 bp with current Illumina technology, van Dijk *et al.*, 2014) allowing the use of more diverse markers such as the V4 region of the 18S rRNA gene (Massana *et al.*, 2014), rbcL (large subunit of the ribulose-1,5-biphosphate carboxylase-oxygenase) encoded in plastid genomes and cox1 (cytochrome c oxidase subunit I) genes encoded in mitochondrial genomes (Kermarrec *et al.*, 2013). Marker choice depends on the groups targeted, available reference sequences and the goals of the study. For example, photosynthetic protist diversity can be accessed with the plastidial 16S gene (Lepère *et al.*, 2009; Choi *et al.*, 2017). Coupling metabarcoding with sorting by flow cytometry improves the resolution of phytoplankton diversity (Shi *et al.*, 2011; Li *et al.*, 2017). In order to explore microdiversity at the species level or below, hypervariable regions such as the ITS (internally transcribed spacer of the rRNA operon) seems to be more suitable (Coleman, 2003; Rodríguez-Martínez *et al.*, 2013).

The development of metabarcoding technics led to ambitious project such as Tara *Ocean* http://oceans.taraexpeditions.org/m/qui-est-tara/les-expeditions/tara-oceans/, Ocean Sampling Day project https://www.microb3.eu/osd.html) or Moorea Biocode project (http://biocode.swala.org/). This method allows the rapid acquisition of large dataset and invites the marine scientific community to start working at the pan oceanic scale. Metabarcoding allowed the first mapping of plankton distribution at

the global ocean (de Vargas *et al.*, 2015) and revealed wide unexpected diversity (for the Haptophytes for example, Egge *et al.*, 2014). By providing a large number of sequences, metabarcoding also to investigate phylogenies of specific groups. Since sequences are often short, for example 400 bp for the V4 region of the 18S rRNA gene, and represent only one locus, phylogenies can be hard to reconstruct. However, in the case of large scale diversity studies, the one-locus approach has been proven to be effective enough to propose species delineation hypotheses (Leliaert *et al.*, 2014). In this thesis, I chose to define species or hypothetical clades within described species as monophyletic groups of sequences (i.e. a group of sequences composed of one ancestors corresponding to a node and all the branches after it) that were supported by bootstrap values (which is a quantification of the robustness of a phylogenetic construction) higher than 70%, with 2 or 3 different phylogenetic methods such as maximum likelihood or Bayesian methods (Groisillier *et al.*, 2006; Guillou *et al.*, 2008).

This definition matches the unified concept of species (De Queiroz, 2007) and the definition of Samadi and Barberousse (2006): the only necessary properties of a species is to be divergent from all other organisms and all other features (such as morphological feature, monophylly…) are optional criteria used to delimit species in practice. This unified concept of species allows to free the biologists from traditional criteria and encourage them to develop new methods of species delimitation to demonstrate divergence: each optional criterium can be used as to support the divergence of a lineage (De Queiroz, 2007).

Thesis organization

The first Chapter introduces the taxonomy of marine Chlorophyta and summarizes current knowledge about their oceanic distribution based in particular on available 18S rRNA sequences. This review led to the curation of a reference database of 18S rRNA gene sequences, a critical tool to study Chlorophyta communities using metabarcoding. The second chapter focuses on the comparison of two commonly used markers for metabarcoding, the V4 and the V9 regions of the 18S rRNA gene. I investigated how the choice of marker influences of view of Chlorophyta diversity based a subset on the OSD samples. The third chapter analyses the distribution and ecological patterns of Chlorophyta classes in marine coastal waters using the OSD dataset. The fourth chapter describes phylogenetic diversity inside Mamiellophyceae based on environmental metabarcodes and explore the biogeography of key Mamiellophyceae species.

List of Figures

List of Tables

Table 1: The 56 biogeochemical provinces identified by Longhurst (2007).

Province Name	Code	Biome	Ocean
Northwest Atlantic subtropical gyral	NAST W	Westerly	Atlantic
Southwest Atlantic shelves	FKLD	Coastal	Atlantic
Brazilian current coast	BRAZ	Coastal	Atlantic
Benguela current coast	BENG	Coastal	Atlantic
Guinea current coast	GUIN	Coastal	Atlantic
Canary current coast	CNRY	Coastal	Atlantic
Guianas coast	GUIA	Coastal	Atlantic
Northeast Atlantic shelves	NECS	Coastal	Atlantic
Northwest Atlantic shelves	NWCS	Coastal	Atlantic
Atlantic Arctic	ARCT	Polar	Atlantic
Atlantic sub-Arctic	SARC	Polar	Atlantic
South Atlantic gyral	SATL	Trade wind	Atlantic
Eastern tropical Atlantic	ETRA	Trade wind	Atlantic
Western tropical Atlantic	WTRA	Trade wind	Atlantic
Caribbean	CARB	Trade wind	Atlantic
North Atlantic tropical gyral	NATR	Trade wind	Atlantic
Northeast Atlantic subtropical gyral	NAST E	Westerly	Atlantic
Mediterranean Sea	MEDI	Westerly	Atlantic
Northwest Atlantic subtropical gyral	NAST W	Westerly	Atlantic
Gulf Stream	GFST	Westerly	Atlantic
North Atlantic Drift	NADR	Westerly	Atlantic
Humboldt current coast	HUMB	Coastal	Pacific
East Australian coast	AUSE	Coastal	Pacific
Sunda-Arafura shelves	SUND	Coastal	Pacific
China Sea	CHIN	Coastal	Pacific
Central American coast	CAMR	Coastal	Pacific
Alaska coastal downwelling	ALSK	Coastal	Pacific
New Zealand coast	NEWZ	Coastal	Pacific
Coastal Californian current	CCAL	Coastal	Pacific
North Pacific epicontinental sea	BERS	Polar	Pacific
Archipelagic deep basins	ARCH	Trade wind	Pacific
Pacific equatorial divergence	PEQD	Trade wind	Pacific
North Pacific equatorial counter current	PNEC	Trade wind	Pacific
North Pacific Tropical gyre	NPTG	Trade wind	Pacific
California current	C(O)CAL	Trade wind	Pacific
South Pacific gyre	SPSG	Trade wind	Pacific
Western Pacific warm pool	WARM	Trade wind	Pacific
Tasman Sea	TASM	Westerly	Pacific
Kuroshio current	KURO	Westerly	Pacific
Eastern Pacific subarctic gyres	PSAE	Westerly	Pacific
Western Pacific subarctic gyres	PSAW	Westerly	Pacific
North Pacific polar front	NPPF	Westerly	Pacific
Northwest Pacific subtropical	NPSW	Westerly	Pacific
Northeast Pacific subtropical	NPSE	Westerly	Pacific
Eastern India coast	EAFR	Coastal	Indian
Western Australian and Indonesian coast	AUSW	Coastal	Indian
Eastern India coast	IND E	Coastal	Indian
Red Sea, Arabian Gulf	REDS	Coastal	Indian

Province Name	Code	Biome	Ocean
Western India coast	IND W	Coastal	Indian
Indian South subtropical gyre	ISSG	Trade wind	Indian
Indian monsoon gyre	MONS	Trade wind	Indian
Northwest Arabian Sea upwelling	ARAB	Westerly	Indian
South subtropical convergence	SSTC	Westerly	Antarctic
Subantarctic water ring	SANT	Westerly	Antarctic
Antarctic	ANTA	Polar	Antarctic
Austral polar	APLR	Polar	Antarctic
Boreal polar	BPRL	Polar	Arctic

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.

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

Anderson, C.R., Kudela, R.M., Benitez-Nelson, C., Sekula-Wood, E., Burrell, C.T., Chao, Y., et al. (2011) Detecting toxic diatom blooms from ocean color and a regional ocean model. *Geophys. Res. Lett.* **38**: n/a-n/a.

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

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

Briggs, J.C. (1974) Marine zoogeography.

Briggs, J.C. and Bowen, B.W. (2012) A realignment of marine biogeographic provinces with particular reference to fish distributions. *J. Biogeogr.* **39**: 12–30.

Burki, F. (2014) The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb. Perspect. Biol.* **6**: a016147.

Canard, B. and Sarfati, R.S. (1994) DNA polymerase fluorescent substrates with reversible 3′-tags. *Gene* **148**: 1–6.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., et al. (2010) Access : QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**: 335–336.

Chenuil, A. (2006) Choosing the right molecular genetic markers for studying biodiversity: from molecular evolution to practical aspects. *Genetica* **127**: 101–20.

Choi, C., Bachy, C., Jaeger, G.S., Poirier, C., Sudek, L., Sarma, V.V.S.S., et al. (2017) Newly discovered deep-branching marine plastid lineages are numerically rare but globally distributed. *Curr. Biol.* **27**: R15–R16.

Chrétiennot-Dinet, M.-J., Courties, C., Vaquer, A., Neveux, J., Claustre, H., Lautier, J., and Machado, M.C. (1995) A new marine picoeucaryote: *Ostreococcus tauri* gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia* **34**: 285–292.

Clements, F.E. (1916) Plant succession: an analysis of the development of vegetation. *Carnegie Inst. Washingt.*

Coleman, A.W. (2003) ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet.* **19**: 370–375.

Coupel, P., Matsuoka, A., Ruiz-Pino, D., Gosselin, M., Claustre, H., Marie, D., et al. (2014) Pigment signatures of phytoplankton communities in the Beaufort Sea. *Biogeosciences Discuss.* **11**: 14489–14530.

Courties, C., Vaquer, A., Troussellier, M., Lautier, J., Chrétiennot-Dinet, M.J., Neveux, J., et al. (1994) Smallest eukaryotic organism. *Nature* **370**: 255–255.

Demir-Hilton, E., Sudek, S., Cuvelier, M.L., Gentemann, C.L., Zehr, J.P., and Worden, A.Z. (2011) Global distribution patterns of distinct clades of the photosynthetic picoeukaryote *Ostreococcus*. *ISME J.* **5**: 1095–1107.

Derelle, E., Ferraz, C., Rombauts, S., Rouzé, P., Worden, A.Z., Robbens, S., et al. (2006) Genome analysis of

the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features. *Proc. Natl. Acad. Sci. U. S. A.* **103**: 11647–52.

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

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

Dunbar, M.J. (1953) Arctic and Subarctic marine ecology: immediate problems University of Calgary.

Egge, E.S., Eikrem, W., and Edvardsen, B. (2014) Deep-branching Novel Lineages and High Diversity of Haptophytes in the Skagerrak (Norway) Uncovered by 454 pyrosequencing. *J. Eukaryot. Microbiol.* 1–20.

Ekman, S. (1935) Tiergeographie des Meeres Akademisch. Leipzig.

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.

Flombaum, P., Gallegos, J.L., Gordillo, R.A., Rincón, J., Zabala, L.L., Jiao, N., et al. (2013) Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl. Acad. Sci. U. S. A.* **110**: 9824–9.

Frada, M.J., Bidle, K.D., Probert, I., and de Vargas, C. (2012) In situ survey of life cycle phases of the coccolithophore *Emiliania huxleyi* (Haptophyta). *Environ. Microbiol.* **14**: 1558–1569.

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

Groisillier, A., Massana, R., Valentin, K., Vaulot, D., and Guillou, L. (2006) Genetic diversity and habitats of two enigmatic marine alveolate lineages. *Aquat. Microb. Ecol.* **42**: 277–291.

Guillou, L., Chrétiennot-Dinet, M.-J., Medlin, L.K., Claustre, H., Loiseaux de-Goër, S., and Vaulot, D. (1999) *Bolidomonas* : a new genus with two species belonging to a new algal class , the Bolidophyceae (Heterokonta). *J. Phycol.* **35**: 368–381.

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

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

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.

Hill, D.R.A. (1991) A revised circumscription of Cryptomonas (Cryptophyceae) based on examination of Australian strains. *Phycologia* **30**: 170–188.

Hobbie, J.E., Daley, R.J., and Jasper, S. (1977) Use of Nucleopore filters fore counting bacteria by fluorescence microscopy. *Appl. Environ. Microbiol.* **33**: 1125–1128.

Hutchinson, G.. and MacArthur, R.H. (1959) A theoretical ecological model of size distributions among species of animals. *Am. Nat.* **93**: 117–125.

Jardillier, L., Zubkov, M. V, Pearman, J., and Scanlan, D.J. (2010) Significant CO2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *ISME J.* **4**: 1180–1192.

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J.F., and Bouchez, A. (2013) Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Mol. Ecol. Resour.* **13**: 607–19.

Kirkham, A.R., Jardillier, L.E., Tiganescu, A., Pearman, J., Zubkov, M. V., and Scanlan, D.J. (2011) Basin-scale distribution patterns of photosynthetic picoeukaryotes along an Atlantic Meridional Transect. *Environ. Microbiol.* **13**: 975–990.

Langer, G., Nehrke, G., Probert, I., Ly, J., and Ziveri, P. (2009) Strain-specific responses of *Emiliania huxleyi* to changing seawater carbonate chemistry. *Biogeosciences* **6**: 2637–2646.

Leliaert, F., Verbruggen, H., Vanormelingen, P., Steen, F., López-Bautista, J.M., Zuccarello, G.C., and Clerck, O. De (2014) DNA-based species delimitation in algae DNA-based species delimitation in algae. *Eur. J. Phycol.* **49**: 179–196.

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

Li, S., Bronner, G., Lepère, C., Kong, F., and Shi, X. (2017) Temporal and spatial variations in the composition of freshwater photosynthetic picoeukaryotes revealed by MiSeq sequencing from flow cytometry sorted samples. *Environ. Microbiol.* **19**: 2286–2300.

Li, W.K.W. (1994) Primary production of prochlorophytes, cyanobacteria, and eucaryotic ultraphytoplankton: Measurements from flow cytometric sorting. *Limnol. Oceanogr.* **39**: 169–175.

Longhurst, A. (1995) Seasonal cycles of pelagic production and consumption. *Prog. Oceanogr.* **36**: 77–167.

Longhurst, A.R. (2007) Ecological geography of the sea Academic Press.

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

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

Manton, I. (1959) Electron microscopical observations on a very small flagellate: the problem of *Chromulina pusilla* Butcher. *J. Mar. Biol. Assoc. United Kingdom* **38**: 319–333.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*.

Marie, D., Brussaard, C.P.D., Partensky, F., and Vaulot, D. (1999) Flow cytometric analysis of phytoplankton, bacteria and viruses. *Curr. Protoc. Cytom.* **11**: 1–15.

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**: 854–66.

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.

Melkonian, M. and Preisig, H.R. (1986) A Light and Electron Microscopic Study of *Scherffelia dubia*, a New Member of the Scaly Green Flagellates (Prasinophyceae). *Nord. J. Bot.* **6**: 235–256.

Meyer, J. and Riebesell, U. (2015) Reviews and Syntheses: Responses of coccolithophores to ocean acidification: a meta-analysis. *Biogeosciences* **12**: 1671–1682.

Moestrup, Ø. and Throndsen, J. (1988) Light and electron microscopical studies on *Pseudoscourfieldia marina*, a primitive scaly green flagellate (Prasinophyceae) with posterior flagella. *Can. J. Bot.* **66**: 1415–1434.

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**: 607–10.

Nilakanta, H., Drews, K.L., Firrell, S., Foulkes, M.A., and Jablonski, K.A. (2014) A review of software for analyzing molecular sequences. *Biomed. reseach note*.

Not, F., Simon, N., Biegala, I., and Vaulot, D. (2002) Application of fluorescent in situ hybridization coupled with tyramide signal amplification (FISH-TSA) to assess eukaryotic picoplankton composition. *Aquat. Microb. Ecol.* **28**: 157–166.

Novarino, G. (2003) A companion to the identification of cryptomonad flagellates (Cryptophyceae = Cryptomonadea). *Hydrobiologia* **502**: 225–270.

Novarino, G., Ian, L., and Morral, S. (1994) Observation of the genus *Plagioselmis* (Cryptophyceae). *Algologie*.

Nyren, P., Pettersson, B., and Uhlen, M. (1993) Solid Phase DNA Minisequencing by an Enzymatic Luminometric Inorganic Pyrophosphate Detection Assay. *Anal. Biochem.* **208**: 171–175.

Oliver, M.J. and Irwin, A.J. (2008) Objective global ocean biogeographic provinces. *Geophys. Res. Lett.* **35**: L15601.

Prokopowich, C.D., Gregory, T.R., and Crease, T.J. (2003) The correlation between rDNA copy number and genome size in eukaryotes. *Genome* **46**: 48–50.

De Queiroz, K. (2007) Species Concepts and Species Delimitation. *Syst. Biol.* **56**: 879–886.

Reeder, J. and Knight, R. (2010) Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nat. Methods* **7**: 668–669.

Reygondeau, G., Longhurst, A., Martinez, E., Beaugrand, G., Antoine, D., and Maury, O. (2013) Dynamic biogeochemical provinces in the global ocean. *Global Biogeochem. Cycles* **27**: 1046–1058.

Rodríguez-Martínez, R., Rocap, G., Salazar, G., and Massana, R. (2013) Biogeography of the uncultured marine picoeukaryote MAST-4: temperature-driven distribution patterns. *ISME J.* **7**: 1531–43.

Ronaghi, M., Uhlen, M., and Nyren, P. (1998) A sequencing method based on real-time pyrophosphate. *Science (80-.).* **281**: 363–365.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A., and Arnheim, N. (1985) Enzymatic Amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science (80-.).* **230**: 1350–1354.

Samadi, S. and Barberousse, A. (2006) The tree, the network, and the species. *Biol. J. Linn. Soc.* **89**: 509–521.

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**: 7537–41.

Schlötterer, C. (2004) The evolution of molecular markers — just a matter of fashion? *Nat. Rev.* **5**: 63.

Shelford, V.E. (1929) Laboratory and Field Ecology. The Responses of Animals as Indicators of correct Working Methods. London, Baillière, Tindall & Cox.

Shi, X.L., Lepère, C., Scanlan, D.J., and Vaulot, D. (2011) Plastid 16S rRNA gene diversity among eukaryotic picophytoplankton sorted by flow cytometry from the South Pacific Ocean. *PLoS One* **6**: e18979.

Šlapeta, J., López-García, P., and Moreira, D. (2006) Global dispersal and ancient cryptic species in the smallest marine eukaryotes. *Mol. Biol. Evol.* **23**: 23–29.

Treusch, A.H., Demir-Hilton, E., Vergin, K.L., Worden, A.Z., Carlson, C.A., Donatz, M.G., et al. (2012) Phytoplankton distribution patterns in the northwestern Sargasso Sea revealed by small subunit rRNA genes from plastids. *ISME J.* **6**: 481–92.

Udvardy, M.D.F. (1975) A Classification of the Biogeographical Provinces of the World. *Contrib. to UNESCO's Man Biosph. Program. Proj. N°.8* **18**:.

Unrein, F., Gasol, J.M., Not, F., Forn, I., and Massana, R. (2014) Mixotrophic haptophytes are key bacterial grazers in oligotrophic coastal waters. *ISME J.* **8**: 164–76.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., et al. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science (80-.).* **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, L., Lin, X., Goes, J.I., and Lin, S. (2016) Phylogenetic Analyses of Three Genes of *Pedinomonas noctilucae*, the Green Endosymbiont of the Marine Dinoflagellate *Noctiluca scintillans*, Reveal its Affiliation to the Order Marsupiomonadales (Chlorophyta, Pedinophyceae) under the Reinstate. *Protist* **167**: 205–216.

Wright, S.W. and Jeffrey, S.W. Pigment Markers for Phytoplankton Production. In, *Marine Organic Matter: Biomarkers, Isotopes and DNA*. Springer-Verlag, Berlin/Heidelberg, pp. 71–104.

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

Chapter 1

Diversity and ecology of green microalgae in marine systems: an overview based on 18S rRNA gene sequences

Diversity and ecology of green microalgae in marine systems: an overview based on 18S rRNA gene sequences

Margot Tragin[1], Adriana Lopes dos Santos[1], Richard Christen[2, 3], Daniel Vaulot[1*]

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

[2] CNRS, UMR 7138, Systématique Adaptation Evolution, Parc Valrose, BP71. F06108 Nice cedex 02, France

[3] Université de Nice-Sophia Antipolis, UMR 7138, Systématique Adaptation Evolution, Parc Valrose, BP71. F06108 Nice cedex 02, France

Perspectives in Phycology 3:3 p.141–154 (2016)

DOI: 10.1127/pip/2016/0059

Perspectives in Phycology Vol. 3 (2016), Issue 3, p. 141–154
Published online June 2016

Article

Diversity and ecology of green microalgae in marine systems: an overview based on 18S rRNA gene sequences

Margot Tragin[1], Adriana Lopes dos Santos[1], Richard Christen[2,3] and Daniel Vaulot[1*]

[1] Sorbonne Universités, UPMC Univ Paris 06, CNRS, UMR 7144, Station Biologique, Place Georges Teissier, 29680 Roscoff, France
[2] CNRS, UMR 7138, Systématique Adaptation Evolution, Parc Valrose, BP71. F06108 Nice cedex 02, France
[3] Université de Nice-Sophia Antipolis, UMR 7138, Systématique Adaptation Evolution, Parc Valrose, BP71. F06108 Nice cedex 02, France
[*] Corresponding author: vaulot@sb-roscoff.fr

[*] Corresponding author: vaulot@sb-roscoff.fr

Keywords

Chlorophyta, Prasinophytes, diversity, distribution, 18S rRNA gene, phylogeny, ecology, marine systems

Acknowledgments

Financial support for this work was provided by the European Union projects MicroB3 (UE-contract-287589) and MaCuMBA (FP7-KBBE-2012-6-311975) and the ANR Project PhytoPol. MT was supported by a PhD fellowship from the Université Pierre et Marie Curie and the Région Bretagne. We would like to thank Adriana Zingone, Bente Edvardsen, Fabrice Not, Ian Probert and two anonymous reviewers for their constructive comments during the preparation of this review.

Abstract

Green algae (Chlorophyta) are an important group of microalgae whose diversity and ecological importance in marine systems has been little studied. In this review, we first present an overview of Chlorophyta taxonomy and detail the most important groups from the marine environment. Then, using public 18S rRNA Chlorophyta sequences from culture and natural samples retrieved from the annotated Protist Ribosomal Reference (PR^2) database, we illustrate the distribution of different green algal lineages in the oceans. The largest group of sequences belongs to the class Mamiellophyceae and in particular to the three genera *Micromonas*, *Bathycoccus* and *Ostreococcus*. These sequences originate mostly from coastal regions. Other groups with a large number of sequences include the Trebouxiophyceae, Chlorophyceae, Chlorodendrophyceae and Pyramimonadales. Some groups, such as the undescribed prasinophytes clades VII and IX, are mostly composed of environmental sequences. The 18S rRNA sequence database we assembled and validated should be useful for the analysis of metabarcode datasets acquired using next generation sequencing.

Introduction

Throughout history, the Earth has witnessed the appearance and disappearance of organisms adapted to their contemporary environments and sometimes these organisms have deeply modified the environment (Kopp *et al.*, 2005; Scott *et al.*, 2008). The best example is provided by the oxygenation of the ocean and the atmosphere by photosynthetic bacteria that first began about 3,500 million years ago (Yoon *et al.*, 2004). Eukaryotic phytoplankton subsequently acquired a chloroplast, a membrane-bound organelle resulting from the phagocytosis without degradation of a cyanobacterium by a heterotrophic host cell (Margulis, 1975), 1,500-1,600 million years ago (Hedges *et al.*, 2004; Yoon *et al.*, 2004). This event marked the origin of oxygenic photosynthesis in eukaryotes. During the course of evolution, endosymbiosis has been repeated several times, new hosts engulfing a eukaryote with an existing plastid, leading to secondary and tertiary endosymbioses (McFadden, 2001). Early in their evolutionary history photosynthetic eukaryotes separated into two major lineages: the green lineage (which includes green algae and land plants) and the red lineage (including diatoms and dinoflagellates) (Falkowski *et al.*, 2004). These two lineages diverged approximately 1,100 million years ago according to molecular clock estimates (Yoon *et al.*, 2004), marking the beginning of algal diversification in the ocean. A number of fundamental differences exist between the members of these two lineages (Falkowski *et al.*, 2004), in particular with respect to pigment content, cellular trace-element composition and plastid gene composition. Green algae possess chlorophyll *b* as the main accessory chlorophyll, while algae from the red lineage mainly harbour chlorophyll *c* (i.e. their chloroplast evolved from a Rhodophytsa algae after secondary endosymbiosis), influencing their respective light absorption properties and ultimately their distribution in aquatic environments. Algae from the red lineage are often derived from secondary or tertiary endosymbioses and have a chloroplast surrounded by three or four membranes, while algae from the green lineage originate mostly from primary endosymbiosis and have a chloroplast surrounded by only two membranes. The evolutionary history of these lineages is probably much more complex than originally thought since it has been suggested that the nuclear genome of diatoms contain green genes (Moustafa *et al.*, 2009), although this has been challenged (Deschamps and Moreira, 2012). Fossil evidence suggests that during the Palaeozoic Era the eukaryotic phytoplankton was dominated by green algae allowing the colonization of terrestrial ecosystems by charophytes, a branch of the green lineage, ultimately leading to the appearance of land plants (Harholt *et al.*, 2015). However, since the Triassic, the major groups of eukaryotic phytoplankton belong to the red lineage (Tappan and Loeblich, 1973; Falkowski *et al.*, 2004).

Green microalgae constitute the base of the green lineage (Nakayama *et al.*, 1998), leading to the hypothesis that the common ancestor of green algae and land plants could be an ancestral green flagellate (AGF) closely related to Chlorophyta (Leliaert *et al.*, 2012). A detailed knowledge of the

diversity of green microalgae is necessary to reconstruct phylogenetic relationships within the green lineage. In the marine environment, the diversity, ecology and distribution of green phytoplankton is poorly known since most studies have focused on groups such as diatoms or dinoflagellates. Finally, green algae could become economically important because in recent years potential applications have developed in industrial sectors such as aquaculture, pharmacy and biofuels (Gómez and González, 2004; Mishra *et al.*, 2008).

This review summarizes current information on the phylogenetic, morphological and ecological diversity of unicellular marine and halotolerant Chlorophyta (we also include some freshwater groups such as the Monomastigales that are very closely related to marine groups). We used around 9,000 Chlorophyta 18S rRNA sequences from culture and environmental samples available in public databases to assess the extent of their diversity and, based on a subset of 2,400 sequences for which geographical information is available, their oceanic distribution. We first present the current state of green algae taxonomy. Then, we detail what is known about each class, and finally we analyze their distribution in oceanic systems from available18S rRNA sequences. These public sequences were extracted from the annotated and expert validated PR^2 database (Guillou *et al.*, 2013), as detailed in the methodology section at the end of the review.

The present state of Chlorophyta classification

The first description of tiny green cells growing in aquatic environments and the first ideas about the classification of microalgae occurred in the middle of the 19th century (Nägeli, 1849). This was followed by a large number of descriptions of green microalgae, leading scientists to reflect on the ecological significance of these organisms. Gaarder (1933) discovered the importance of green microalgae in the food web by looking at the source of oyster food in Norway. Twenty years later, the first marine picoeukaryotic phytoplankter to be described (*Chromulina pusilla*, later renamed *Micromonas pusilla*) was a tiny green alga (Butcher, 1952).

In the 1960s and early 1970s, Round (Round, 1963, 1971), reviewing available morphological information, divided the green algae into four divisions: Euglenophyta, Charophyta, Chlorophyta and Prasinophyta. While Round classified the Prasinophyta in a separate phylum, other authors (Bourrelly, 1966; Klein and Cronquist, 1967) included them in the order Volvocales within the Chlorophyta. The division Chlorophyta was reorganized by Mattox and Stewart (1975) mainly based on ultrastructural characteristics such as the type of mitosis (Sluiman *et al.*, 1989), presence/absence of an interzonal spindle, the structure of the flagellar apparatus (O'Kelly and Floyd, 1983), and the presence of extracellular features such as scales and thecae. They proposed the division of Chlorophyta into four major groups: the Prasinophyceae, Charophyceae, Ulvophyceae and Chlorophyceae (Stewart and Mattox, 1978). This has been partly confirmed by molecular phylogenetic analyses over the years

(Chapman *et al.*, 1998), although it was recognized from the beginning (Christensen, 1962) that prasinophytes constitute a polyphyletic assemblage (i.e. phylogenetic branches without a common ancestor). Therefore the class name Prasinophyceae is no longer used and the generic term prasinophytes, that has no phylogenetic meaning, has replaced it (Leliaert *et al.*, 2012). At present, the Chlorophyta is viewed as composed of two major groups: the prasinophytes and the "core" chlorophytes (Leliaert *et al.*, 2012; Fučíková *et al.*, 2014).

The prasinophytes currently consist of nine major lineages of microalgae corresponding to different taxonomic levels (order, class, undescribed clades) that will probably all be raised to the class level in the future (Leliaert *et al.*, 2012). These lineages share ancestral features such as flagella and organic scales. The number of prasinophyte lineages has been increasing following the availability of novel environmental sequences. Ten years ago, prasinophyte clade VII was introduced using sequences from cultured strains and environmental clone libraries (Guillou *et al.*, 2004). Four years later, two additional clades, VIII and IX, were reported (Viprey *et al.*, 2008) that are only known so far from environmental sequences. Prasinophytes may be divided into three informal groups (Marin and Melkonian, 2010): a group of "basal" lineages (Prasinococcales, Pyramimonadales, Mamiellophyceae), a group of "intermediate" lineages (Pseudoscourfieldiales, clade VII, Nephroselmidophyceae) and a group of "late" diverging lineages (Pedinophyceae and Chlorodendrophyceae). Recently, the "late" diverging lineages have been merged with the Ulvophyceae-Trebouxiophyceae-Chlorophyceae (UTC) clade into the "core" chlorophytes (Fučíková *et al.*, 2014), the Chlorodendrophyceae based on common features, in particular a mode of cell division mediated by a phycoplast (Mattox and Stewart, 1984; Leliaert *et al.*, 2012), and the Pedinophyceae based on strong phylogenetic support (Marin, 2012; Fučíková *et al.*, 2014).

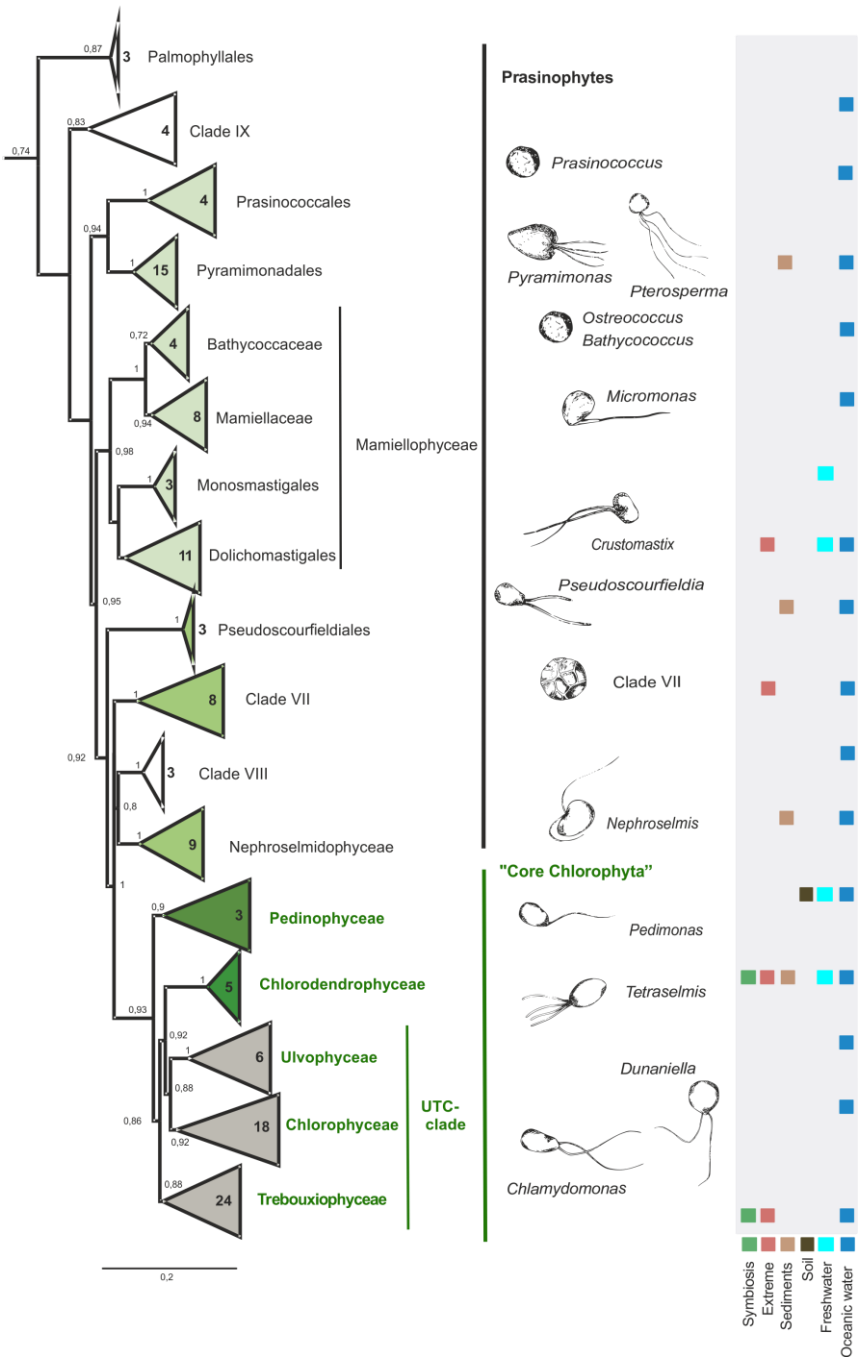


Fig.1: Phylogenetic tree of a set of 132 rRNA 18S reference sequences (Supplementary Table S1) constructed by FastTree (options used: General Time Reverse model, optimized gamma likelihood, rate categories of sites 20), rooted with Oryza sativa (AACV01033636) based on an edited MAFFT 1752 bp alignment stripped at 50% (columns of the alignment counting more than 50% gaps were deleted, Supplementary data). The phylogenetic tree was validated by MrBayes phylogeny which provided a similar result. Fast Tree bootstrap values larger than 70% are reported. The number of references sequences for each group is also reported. Triangle colors correspond to the different groups defined by Marin and Melkonian (2010). Groups labelled in green correspond to "core" chlorophytes. Symbols on the right side of the tree indicate the habitat of each group.

Major lineages within marine Chlorophyta

We extracted a set of 132 reference sequences from the PR^2 database that were used to build a phylogenetic tree of marine Chlorophyta (Fig. 1). In this tree, each triangle (except for the Mamiellophyceae for which we have represented the different families) corresponds to a "lineage" which currently corresponds to either a class, an order or a "clade" *sensu* Guillou et *al.* (2004). In this section, we review what is known about the major Chlorophyta lineages in marine waters following the order used in Guillou et *al*. (2004).

Pyramimonadales (prasinophyte clade I, Guillou et al. 2004) are pyramidal, oval or heart-shaped cells (10 to 400 µm long on average) with generally 4, rarely, 8 or even 16 flagella (Chadefaud, 1950; Hori *et al.*, 1985). In *Pyramimonas*, cells possess three layers of different organic scales on the cell body, two layers on the flagella (Pennick, 1982, 1984) and flagellar hairs (Moestrup, 1982). Twenty-two genera have been described, with *Pyramimonas*, *Pterosperma* and *Halosphaera* containing most species. For the genus *Pyramimonas*, almost 50 species (Suda *et al.*, 2013; Harðardottir *et al.*, 2014; Bhuiyan *et al.*, 2015) and six sub-genera (Hori *et al.*, 1995) have been described, but the low number of ribosomal RNA sequences from described species in public sequence databases is an obstacle to the resolution of the phylogeny of this genus (Table S1, Suda et al. 2013). Novel species have recently been described from isolates from the North Pacific Ocean (Fig.3A, Suda et al. 2013, Bhuiyan et al. 2015) and polar regions (Moro *et al.*, 2002; Harðardottir *et al.*, 2014). In Disko Bay (Greenland), *Pyramimonas* has been found to be important in the sea ice and in the water column and plays an important role in the spring phytoplankton bloom (Harðardottir *et al.*, 2014). Pyramimonadales have been recorded in coastal waters as well as in confined environments such as tide pools (Chisholm and Brand, 1981; Lee, 2008). *Halosphaera* occurs in two forms, one flagellated and one coccoid, the latter that can be up to 800 µm in size and that may sediment quickly. In the Mediterranean Sea, high abundances of *Halosphaera* have been recorded at depths between 1,000 and 2,000 meters (Wiebe *et al.*, 1974).

Mamiellophyceae (clade II, Guillou *et al.*, 2004) are characterized by a wide morphological diversity. They are split into three orders: Mamiellales, which is composed of two families (Mamiellaceae and Bathycoccaceae), Dolichomastigales and Monomastigales (Fig.1, Marin and Melkonian, 2010).

The Mamiellaceae contain three genera that are ecologically important. *Micromonas* are ellipsoid to pyriform naked cells (1 to 3 µm) with a single emergent flagellum (Butcher, 1952). Phylogenetic and ecological studies on the micro-diversity of *Micromonas* suggest that this genus may consists of at least three cryptic species (Šlapeta *et al.*, 2006; Foulon *et al.*, 2008). *Micromonas* is a ubiquitous genus with cultures originating from a wide range of environments extending from the poles to the tropics, but more prevalent in coastal waters. *Mamiella* and *Mantoniella* are reniform cells (up to 10 µm) covered by two

types of body scales: large, more or less square, and small, less regular (Barlow and Cattolico, 1980; Moestrup, 1984). *Mamiella* have two long flagella and spined flagellar scales, while *Mantoniella* has one long and one very short flagella with flagellar scales lacking spines (Marin and Melkonian, 1994). Environmental sequences from the latter two genera have been found in the Arctic Ocean and the Mediterranean Sea using Chlorophyta specific primers or sorted samples (Viprey *et al.*, 2008; Balzano *et al.*, 2012).

Bathycoccaceae are spherical or elliptical coccoid cells and contain two genera, *Bathycoccus*, which is covered by spider-web-like scales (1.5 to 2.5 µm) (Eikrem and Throndsen, 1990), and *Ostreococcus*, which is naked and the smallest known photosynthetic eukaryote to date, with a typical size of 0.8 µm (Chrétiennot-Dinet *et al.*, 1995). *Ostreococcus* was first isolated from a Mediterranean Sea lagoon (Courties *et al.*, 1994) and then from many mesotrophic oceanic regions (Rodríguez *et al.*, 2005; Viprey *et al.*, 2008). Four clades of *Ostreococcus* have been described (Guillou *et al.*, 2004) leading to the formal description of 2 species (Subirana *et al.*, 2013). *Bathycoccus* does not seem to show micro-diversity based on sequences of the 18S rRNA gene (Guillou *et al.*, 2004), although ITS (internal transcribed spacers) sequence and genomic evidence points to the existence of two different ecotypes (Vaulot *et al.*, 2012; Monier *et al.*, 2013). *Bathycoccus* was first isolated from Mediterranean Sea and Norwegian waters (Eikrem and Throndsen, 1990), but sequences have now been recovered from many regions (Viprey *et al.*, 2008).

Monomastigales cells are oblong (3.5 to 15 µm long) and covered by proteinaceous scales. Cells have a single flagellum and a second immature basal body (Heimann *et al.*, 1989). The only member of this order is the freshwater genus *Monomastix*. Sequences have been recorded only in freshwater on four continents (Europe, North America, Asia, Australia) (Scherffel, 1912; Marin and Melkonian, 2010).

Dolichomastigales cells are round or bean-shaped (2 to 5 µm long), biflagellate, naked or covered by spider-web-like scales or a crust. This order regroups the genera *Dolichomastix*, isolated in the Arctic, South Africa and Mediterranean sea (Manton, 1977; Throndsen and Zingone, 1997) and *Crustomastix*, first isolated in the Mediterranean Sea (Nakayama *et al.*, 1998; Zingone *et al.*, 2002; Marin and Melkonian, 2010)

Nephroselmidophyceae (clade III, Guillou *et al.*, 2004) is a class of flattened or bean-shaped cells, with two unequal flagella. The cell body (4,5 to 7 µm long) is covered by 5 different types of scales (squared and stellate) and the flagella by 3 types (Nakayama *et al.*, 2007). Eleven genera have been described in this class and the genus *Nephroselmis* counts the largest number of described species (14 according to AlgaeBase, Table S2) of which 6 new species have recently been described from coastal South African and Pacific waters (Faria et al. 2011, 2012, Yamaguchi et al. 2011, 2013).

Pseudoscourfieldiales (clade V, Guillou *et al.*, 2004) are coccoid cells (1.5 to 5 µm in diameter) without scales but with a cell wall (*Pycnococcus provasolii*, Guillard *et al.*, 1991) or with scales and biflagellate (*Pseudoscourfieldia marina*, Moestrup and Throndsen, 1988). *Pycnococcus* was initially isolated from the North Atlantic ocean (Guillard *et al.*, 1991), but cultures have also been recovered from other environments such as the South-East Pacific Ocean (Le Gall *et al.*, 2008). The 18S rRNA sequences of the two species are 100% identical, leading to the hypothesis that they could represent different life cycle stages, or growth forms, of the same species (Fawley *et al.*, 1999).

Prasinococcales (clade VI, Guillou *et al.*, 2004) is an order composed of coccoid cells (2.5 to 5.5 µm in diameter) without scales, surrounded in general by a multilayer gelatinous matrix made of polysaccharides (Hasegawa *et al.*, 1996; Sieburth *et al.*, 1999). The two main genera are *Prasinococcus* (Miyashita *et al.*, 1993) and *Prasinoderma* (Hasegawa *et al.*, 1996). One species, *Prasinoderma singularis*, lacks the gelatinous matrix (Jouenne *et al.*, 2011). Prasinococcales have been isolated from coastal and open oceanic waters in the North Atlantic (Sieburth *et al.*, 1999) and Pacific Oceans (Miyashita *et al.*, 1993) as well as in the Mediterranean Sea (Viprey *et al.*, 2008). One novel environmental *Prasinoderma* clade has been found using Chlorophyta specific primers (Viprey *et al.*, 2008).

Prasinophyte clade VII has been identified from environmental and culture sequences (Guillou *et al.*, 2004). The first isolate of prasinophyte clade VII, CCMP1205 (=RCC15), was reported by Potter et *al.* (Potter *et al.*, 1997). Since then, the lack of distinct morphological characters has kept these small (3 to 5 µm) coccoid cells without a formal description despite their importance in oceanic waters in particular in the South Pacific Ocean, Mediterranean Sea and South China Sea (Moon-van der Staay *et al.*, 2001; Viprey *et al.*, 2008; Shi *et al.*, 2009; Wu *et al.*, 2014). Prasinophyte clade VII is divided into three well-supported lineages, A, B and C, the latter being formed by *Picocystis salinarum*, a small species found in hypersaline lakes (Lewin *et al.*, 2000; Krienitz *et al.*, 2012). The large number of clade VII strains and environmental sequences now present in public databases has allowed further delineation of at least 10 sub-clades (Lopes dos Santos et al. submitted) within the two major marine lineages A and B described by Guillou et al. (2004).

Prasinophyte clade VIII is a clade known purely from environmental sequences, specifically 3 sequences from the picoplankton size fraction (i.e. cells passing through a 3 µm pore-size filter) found at a single sampling location station in the Mediterranean Sea (Viprey *et al.*, 2008).

Prasinophyte clade IX is also an environmental clade. This clade was initially found using either Chlorophyta-specific primers or from flow cytometry sorted samples (Viprey *et al.*, 2008; Shi *et al.*, 2009). Sequences originate mostly from picoplankton samples collected in oligotrophic areas from the Pacific Ocean (Shi *et al.*, 2009; Wu *et al.*, 2014) and the Mediterranean Sea (Viprey *et al.*, 2008).

Palmophyllales is an order of poorly known colonial algae with a thalli formed by isolated spherical cells in a gelatinous matrix (Zechman *et al.*, 2010; Leliaert *et al.*, 2012). These green algae have been isolated from moderately deep waters. The genus *Palmophyllum* was described from cells (6-7 µm) growing at 70 m depth near New Zealand (Nelson and Ryan, 1986), while *Verdigellas* (Ballantine and Norris, 1994) may live below 100 m and was isolated from the tropical Atlantic Ocean (Zechman *et al.*, 2010). Phylogenetic studies based on the 18S rRNA (3 sequences from isolates are available) and two plastid genes suggested that this lineage is deep branching within the Chlorophyta (Zechman *et al.*, 2010).

Pedinophyceae cells are asymmetrical, ovoid or ellipsoid (about 3 µm long), uniflagellate and naked (Moestrup, 1991). This class consists of two orders, the Pedinomonadales and the Marsupiomonadales (Marin 2012) and six genera (Table S2). One genus of Marsupiomonadales (*Resultomonas*) does not have any 18S sequence available. Marsupiomonadales are marine, while Pedinomonadales live in soil and freshwater (Fig.1, Marin 2012)

Chlorodendrophyceae (clade IV, Guillou *et al.*, 2004) are quadriflagellate elliptical cells (on average 15 to 20 µm long). The cell body is covered by a theca (resulting from the fusion of stellar scales) and the flagella, thick and shorter than the cell, are covered by 2 layers of scales and hairs (Hori, Richard E. Norris, *et al.*, 1982). This class contains four genera (Table S2) (Lee and Hur, 2009). The genus *Tetraselmis*, for which several species have been isolated from brackish lagoons, has been divided into four sub-genera (Hori, Richard E Norris, *et al.*, 1982; Hori *et al.*, 1983, 1986), but molecular studies using the 18S rRNA gene fail to resolve the phylogeny of this genus (Arora *et al.*, 2013).

The UTC (Ulvophyceae, Trebouxiophyceae and Chlorophyceae) clade shows a wide morphologic diversity. Most UTC representatives are macroalgae or originate from freshwater or terrestrial environments. We only focus here on unicellular marine representatives. Unicellular marine Ulvophyceae are represented by one genus, *Halochlorococcum*, with very few sequences from cultures, all originating from Japan (Table S3). Trebouxiophyceae are mostly represented by coccoid cells in coastal marine environments belonging to the genera *Picochlorum* (2 µm in diameter), *Chlorella* (1.5 to 10 µm in diameter), *Elliptochloris* (5 to 10 µm in diameter) and *Chloroidium* (~ 15 µm in diameter) (Andreoli *et al.*, 1978; Henley *et al.*, 2004; Letsch *et al.*, 2009; Darienko *et al.*, 2010). Chlorophyceae are morphologically diverse (de Reviers, 2003) from non-motile coccoid cells to flagellates. Most sequences from marine Chlorophyceae strains, isolated from coastal waters or salt pools, belong to the genera *Asteromonas* (12 to 22 µm long), *Chlamydomonas* (7 to 11 µm long), and *Dunaliella* (8 to 18 µm long) (Hoshaw and Ettl, 1966; Peterfi and Manton, 1968; Preetha, 2012).

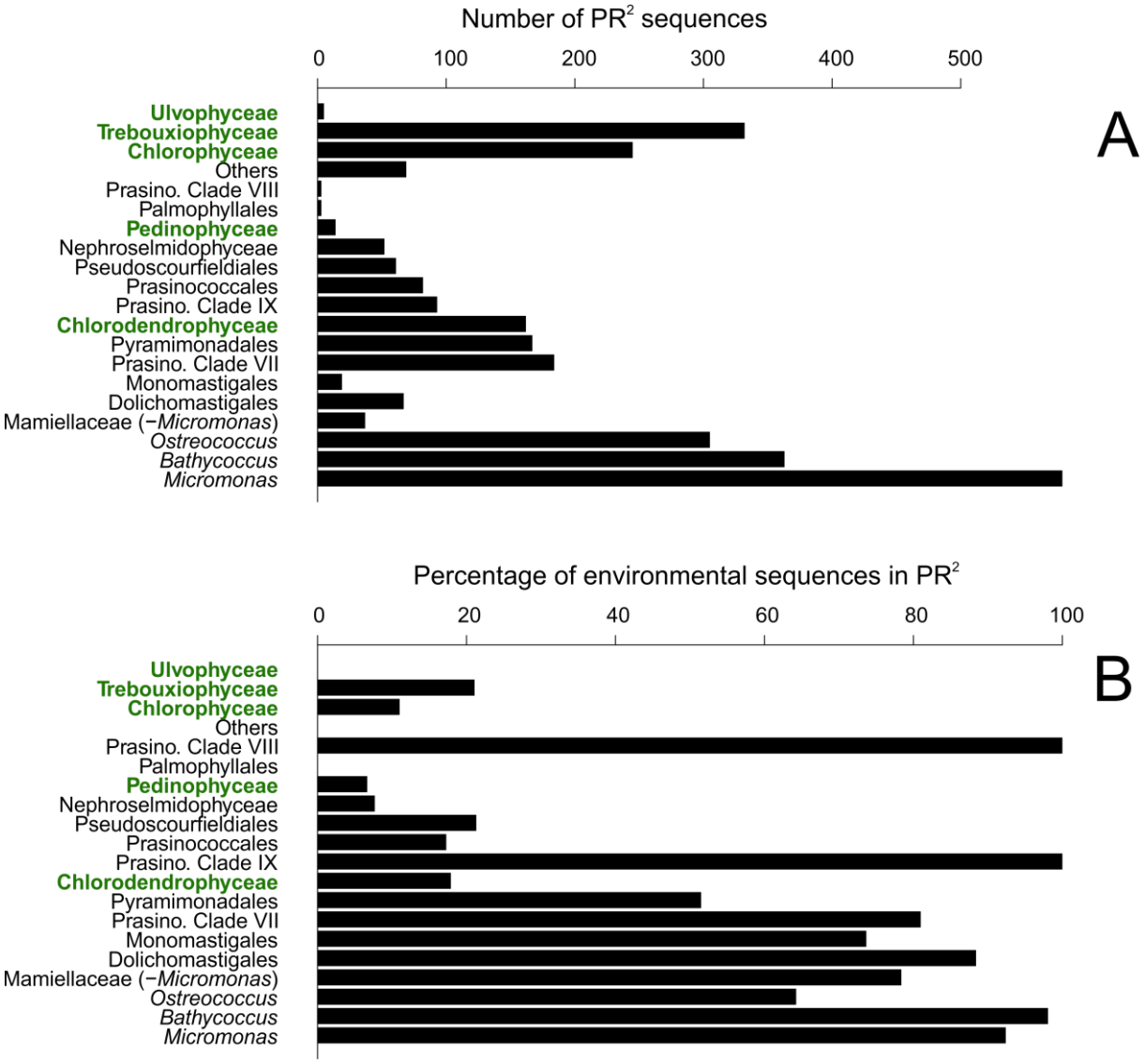


Fig.2: Number of Chlorophyta sequences in the PR2 database (A) and percentage of environmental sequences in PR2 (B) for each clade. The number of sequences for Mamiellaceae does not include Micromonas which is reported separately. Groups labelled in green correspond to "core" chlorophytes (see Fig.1).

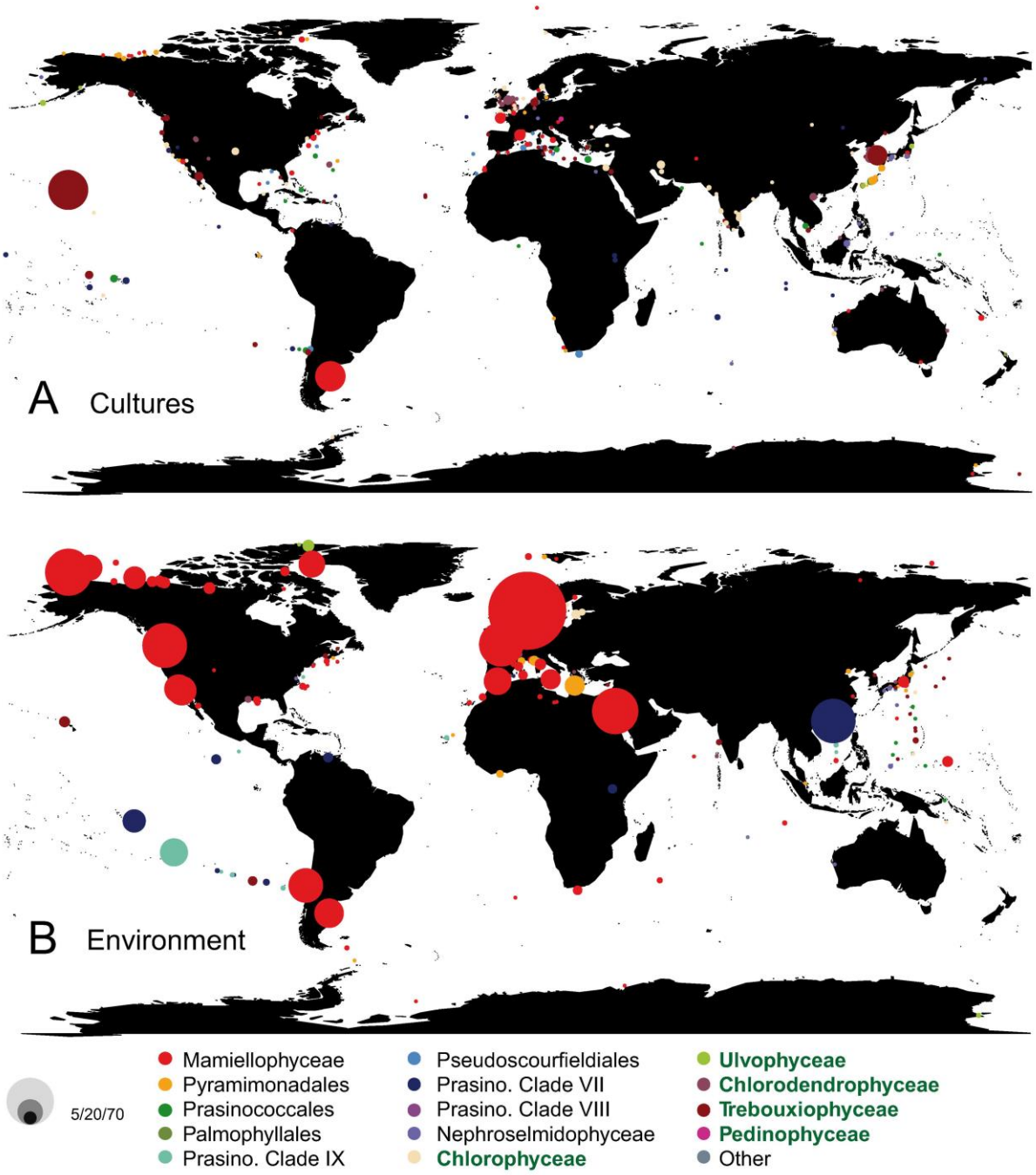


Fig.3: Oceanic distribution of PR2 sequences for major Chlorophyta lineages for cultures (A) and environmental samples (B). The color of the circle corresponds to the most abundant lineage and the surface of the circle is proportional to the number of sequences for this lineage obtained at the location. Groups labelled in green correspond to "core" chlorophytes (see Fig.1).

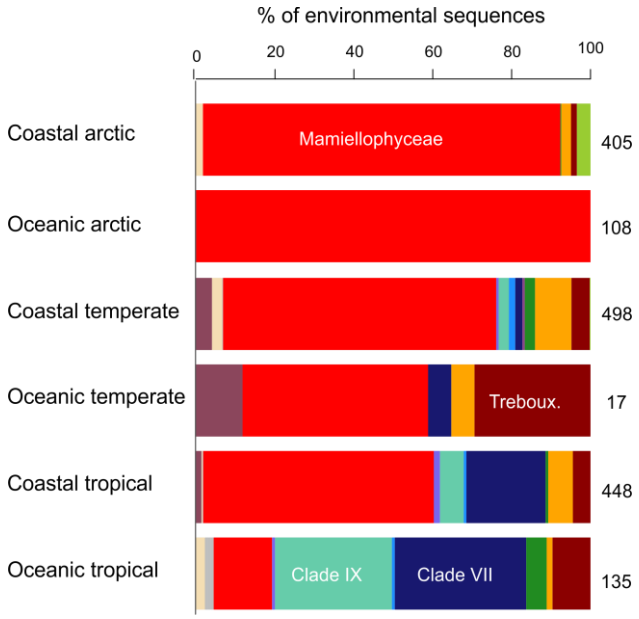


Fig.4: Distribution of PR2 Chlorophyta environmental sequences according to three latitudinal zones (90° to 60°, 60° to 35° and 35° to 0°) and to the distance to the nearest shore (locations closer than 200 km were considered as coastal and the rest as oceanic). Distances to the coast were computed for each sequence using the R packages rgdal and rgeos. Antarctica is not represented because of the very low number of sequences from this area. Colors correspond to Chlorophyta classes and are the same as in Fig.3. The number of sequences in each group is indicated on the right.

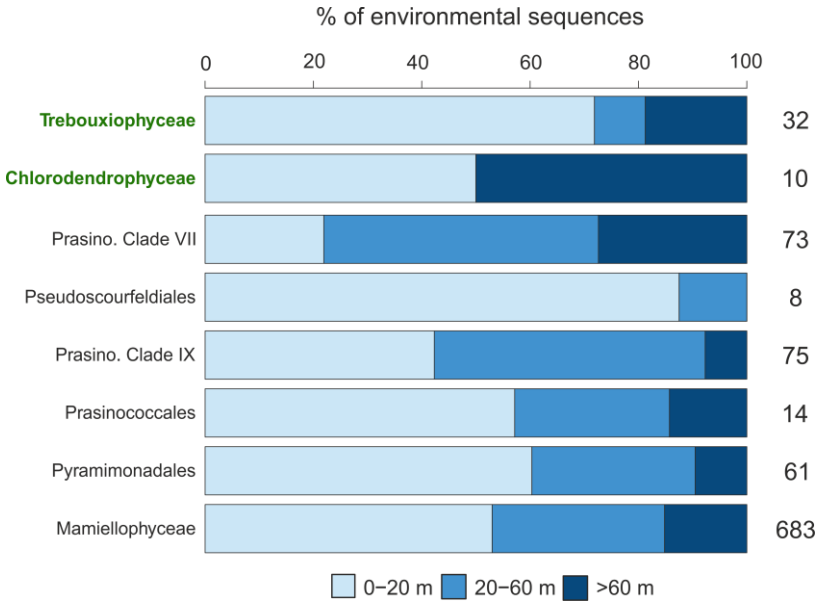


Fig.5: Number of environmental Chlorophyta sequences in PR2 according to depth range for each lineage (only sequences for which depth is reported in the GenBank record are included and lineages for which less than 5 sequences were available were omitted).

Environmental distribution of Chlorophyta in marine ecosystems from 18S rRNA sequences

The number of publicly available sequences (Fig.2A, based on the PR^2 database, Guillou et al. 2013) varies widely between the Chlorophyta groups from 3 for the Palmophyllales up to 560 for the sole genus *Micromonas*. Mamiellophyceae and in particular *Micromonas*, *Bathycoccus* and *Ostreococcus* are the most represented green algal taxa in public sequence databases, followed by Chlorophyceae and Trebouxiophyceae, two groups which were previously mostly seen as continental (Fig.2). The proportion between sequences from cultures and environmental samples is also highly variable (Fig.2B). Some groups are mostly represented by sequences from cultures (e.g. Nephroselmidophyceae and "core" chlorophytes) while others are predominantly or wholly uncultured (e.g. prasinophyte clade IX). The geographic distribution obtained from cultures and from environmental sequences is quite different (compare A and B in Fig.3 and S1). While Mamiellophyceae dominate environmental sequences, this is not true for culture sequences, which offer a better balance between the different Chlorophyta groups (Fig. S1). The contribution of different classes to environmental sequences differs between latitudinal bands and coastal vs. oceanic stations (Fig.4).

Polar waters, whether oceanic or coastal, are totally dominated by Mamiellophyceae (Fig.4), in particular the arctic *Micromonas* clade (Lovejoy *et al.*, 2007; Balzano *et al.*, 2012). The diversity of classes recovered is minimal (Supplementary Fig. S1), with representatives of the Pyramimonadales, Ulvophyceae and Prasinococcales in addition to the Mamiellophyceae. It is noteworthy that few sequences have been recovered from the Southern Ocean in comparison to the Arctic (Supplementary Fig. S1).

The dominance of Mamiellophyceae is less marked for temperate waters where other classes such as the Trebouxiophyceae can be important, especially away from the coast (Fig.4). Indeed, it is in temperate waters that Chlorophyta environmental sequence diversity is maximal, in particular in the North-West Atlantic and North-East Pacific Oceans with more than 10 Chlorophyta classes recovered (Supplementary Fig. S1). Chlorophyceae, Prasinococcales Pseudoscourfieldiales, and clade IX sequences have also been recovered from coastal temperate areas including the Mediterranean Sea (Fig.3B and 4). Nephroselmidophyceae have been repeatedly isolated from Japanese coastal waters (Fig.3). Chlorodendrophyceae, Trebouxiophyceae, Pyramimonadales and clade VII have been found in both coastal and oceanic temperate waters (North Pacific Ocean and Mediterranean Sea, Fig.3 and 4).

The decrease in the dominance of Mamiellophyceae is even more marked in tropical waters. While it shares dominance with prasinophyte clade VII in coastal waters, it becomes a minor component offshore where it is replaced by clade VII and the uncultured clade IX. Trebouxiophyceae,

Prasinococcales, Pyramimonadales and Chlorophyceae have also been found at some locations in the subtropical Pacific and Atlantic Oceans (Fig.3B and Fig.4).

With respect to depth distribution, both Mamiellophyceae, Pyramimonadales, as well as prasinophyte clade VII and IX sequences have been found throughout the photic zone, even below 60 meters (Fig.5). Pseudoscourfieldiales, Trebouxiophyceae and Prasinococcales sequences seem to be restricted to surface waters, while Chlorodendrophyceae sequences appear to be preferentially found at the bottom of the photic zone, below 60 m (Fig.5). The deepest Mamiellophyceae sequences have been recovered from 500 m depth for *Micromonas* and down to 2500 m depth for *Ostreococcus* (Lie *et al.*, 2014).

Mamiellophyceae have been found to dominate environmental sequences in some anoxic waters, as, for example, near Saanich Inlet off Vancouver (Orsi *et al.*, 2012). Sediments also constitute environments where green microalgal sequences have been recovered (Fig.1). For example Dolichomastigales (Mamiellophyceae), Chlorodendrophyceae and prasinophyte clade VII have been found in anoxic sediments (Edgcomb *et al.*, 2011) or in cold methane sediments (Takishita *et al.*, 2007). Cultures of Nephroselmidophyceae, Chlorodendrophyceae, Pseudoscourfieldiales and Pyramimonadales have also been isolated from sediments (Fig.1). However, Chlorophyta found in sediments may not correspond to truly benthic species, but could result from cell sedimentation down the water column.

Advantages and limitations of 18S rRNA as a marker gene for Chlorophyta

Our analysis was based on Chlorophyta sequences for the 18S rRNA gene that are publicly available. Using this gene, we were able to recover (Fig.1) the three diverging groups described by Marin and Melkonian (Marin and Melkonian, 2010) using both nuclear (18S) and plastid (16S) encoded rRNA: the early diverging group (Pyramimonadales, Mamiellophyceae and Prasinococclaes), the intermediate group (Nephroselmidophyceae, Pseudoscourfieldiales and clade VII) and the late-diverging group (Chlorodendrophyceae and Pedinophyceae). Further investigation of the phylogenetic relationships between the different Chlorophyta lineages would require multiple markers. For example, Fučíková et al. (2014) used 8 genes, including *rcbL* (the large subunit of the ribulose-1,5-biphosphate carboxylase-oxygenase gene), *tuf A* (translation unstable factor) and the 18S rRNA to address the relationship within "core" chlorophytes. In order to explore microdiversity at the species level or below, the LSU (large ribosomal subunit) or the ITS seems to be more suitable (Coleman, 2003). For example, the four *Ostreococcus* clades (Mamiellophyceae) are better discriminated with ITS than 18S rRNA (Rodríguez *et al.*, 2005).

In the course of this work, Chlorophyta 18S rRNA sequences were verified and re-annotated. The resulting updated database contains 8554 sequences (Supplementary data S1) and will be useful to

annotate Chlorophyta metabarcode sequences from the V4 or V9 regions of the 18S rRNA gene obtained with High Throughput Sequencing (de Vargas *et al.*, 2015; Massana *et al.*, 2015). The level of similarity within each phylogenetic lineage varies depending on the Chlorophyta lineage and on the region of the gene considered (Supplementary Fig. S2). For the full 18S rRNA gene, it varies from 83.6 % for Ulvophyceae to 99.9 % for Pseudoscourfeldiales.

Within most of the lineages, the V9 region (2,416 sequences) seems more divergent than the V4 region (6,530 sequences), but identity levels are more variable for the former (Supplementary Fig. S2). The V9 region therefore appears to be a good marker for Chlorophyta, although the larger size of the V4 region could be advantageous to reconstruct the phylogeny of novel groups without representatives in the reference database.

A number of caveats have, however, to be considered. Some sequences do not cover the full length of the 18S rRNA gene. For example, only 2,416 sequences (28% of sequences analyzed) cover the full V9 region. Some environmental clades (e.g. prasinophyte clade VIII) are represented only by short sequences, and this clade would be missed in metabarcoding studies using the V9 region. Moreover, not all described species have published 18S sequences. For example, two Pedinophyceae genera are known to live in marine waters, *Resultomonas* and *Marsupiomonas*, but sequences are only available for the latter genus. *Resultomonas* will therefore be "invisible" in surveys based on environmental DNA. Some groups have only cultured sequences, e.g. Nephroselmidophyceae, with an overrepresentation off Japan, because scientists from this country have a keen interest in this group (Nakayama *et al.*, 2007; Faria *et al.*, 2011, 2012, Yamaguchi *et al.*, 2011, 2013). Other groups, such as prasinophyte clades VIII and IX, have completely escaped cultivation and obtaining environmental sequences from these groups is difficult because of competition among different templates when using universal PCR primers. Two methods have been used to increase recovery of Chlorophyta sequences: the use of Chlorophyta specific primers and flow cytometry sorting of photosynthetic organisms (Viprey *et al.*, 2008; Shi *et al.*, 2009). Another limitation is that metadata available in GenBank are far from complete and even sometimes not accurate. For example, only ~2,500 sequences are associated with geographical coordinates (Fig.3 and Fig. S1) and less than 1,000 environmental sequences have depth information (Fig.5).

Conclusion and perspectives

Despite being neglected in comparison to other groups such as diatoms and dinoflagellates, marine green algae are very diverse and are distributed worldwide. Some groups, such as the Mamiellophyceae, are ubiquitous (Fig.4) and are starting to be well characterized from the physiological and genomic points of view, while other groups, such as prasinophyte clade IX, still remain uncultured. In the future, metabarcoding will make it possible to improve our knowledge of the worldwide distribution of each clade and identify their ecological niches.

Methodology

In order to determine the extent of molecular diversity of marine Chlorophyta, we used the Protist Ribosomal Reference (PR^2) database (Guillou et al. 2013). This database contains public eukaryotic 18S rRNA sequences from cultured isolates as well as from environmental samples that have been quality controlled and annotated. All Chlorophyta sequences were extracted, yielding a final dataset of around 9,000 sequences. For each sequence, we extracted metadata from GenBank (such as sampling coordinates and date, publication details), when available. Other metadata were obtained from the literature or from culture collection websites. This information was entered into a Microsoft Access database. In particular, the sampling coordinates were used to map the sequence distribution using the packages maps v2.3-9 and mapdata v2.2-3 of the R3.0.2 software (http://www.R-project.org/). The database and the metadata have been deposited to Figshare (see Supplementary data).

The assignation of sequences was checked down to the species level. For this purpose, we aligned sequences for each phylogenetic group (in general at the class level) using MAFFT v1.3.3 (Katoh, 2002). Phylogenetic trees were constructed using FastTree v1.0 (Price *et al.*, 2009) run within the Geneious software v7.1.7 (Kearse *et al.*, 2012). Phylogenetic trees were compared with those found in the literature. We defined phylogenetic clades as monophyletic groups of sequences that were supported by bootstrap values higher than 70%, with 2 or 3 different phylogenetic methods (Groisillier *et al.*, 2006; Guillou *et al.*, 2008). If more than 2 strain sequences from the same species belonged to a given clade, then the other sequences in this clade were assigned to that species in the database. When the tree was not clear enough, for example for groups represented by a large number of sequences, signatures in the alignment were used to validate the assignation. Chimeric sequences were filtered out by assigning the first 300 and last 300 base pairs of the sequences with the software mothur v1.35.1 (Schloss *et al.*, 2009). If a conflict of assignation between the beginning and the end of the sequences was detected then, they were BLASTed against GenBank to confirm whether they were chimeras and in the latter case, removed from any further analysis.

Reference sequences for each Chlorophyta class were selected and a reference Chlorophyta tree containing 132 sequences was built using Maximum Likelihood and Bayesian methods (Fig.1, Table S1, Supplementary data). When possible, the reference sequences were full-length 18S rRNA sequences from culture strains and already used as references in the literature. Moreover, they were chosen to be distributed in the major clades of each lineage and as a result the number of reference sequences was a function on the micro-diversity within each class.

References

Andreoli, C., Rascio, N., and Casadoro, G. (1978) *Chlorella nana* sp. nov. (Chlorophyceae): a new marine *Chlorella*. *Bot. Mar.* **21**: 253–256.

Arora, M., Anil, A.C., Leliaert, F., Delany, J., and Mesbahi, E. (2013) *Tetraselmis indica* (Chlorodendrophyceae, Chlorophyta), a new species isolated from salt pans in Goa, India. *Eur. J. Phycol.* **48**: 61–78.

Ballantine, D. and Norris, J. (1994) *Verdigellas*, a new deep water genus (Tetrasporales, Chlorophyta) from the Tropical Western Atlantic. *Cryptogam. Bot.* **4**: 368.

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

Barlow, S.B. and Cattolico, R.A. (1980) Fine structure of the scale-covered green flagellate *Mantoniella squamata* (Manton et Parke) Desikachary. *Br. Phycol. J.* **15**: 321–333.

Bhuiyan, M.A.H., Faria, D.G., Horiguchi, T., Sym, S.D., and Suda, S. (2015) Taxonomy and phylogeny of *Pyramimonas vacuolata* sp. nov . (Pyramimonadales, Chlorophyta). *Phycologia* **54**: 323–332.

Bourrelly, P. (1966) Les Algues d'eau douce : Initiation a la systematique. I les algues vertes. Boubée. Paris.

Butcher, R.W. (1952) Contributions to our knowledge of the smaller marine algae. *J. Mar. Biol. Assoc. United Kingdom* **31**: 175.

Chadefaud, M. (1950) Les cellules nageuses des algues dans l'embranchement des Chromophycées. *Compte rendus Hebd. des séances l'académie des Sci.* **231**: 788–790.

Chapman, R.L., Buchheim, M.A., Delwiche, C.F., Friedl, T., Huss, V.A.R., Karol, K.G., et al. (1998) Molecular systematics of the Green Algae. In, Soltis,D.E., Soltis,P.S., and Doyle,J.J. (eds), *Molecular Systematics of Plants II*. Springer US, Boston, MA, pp. 508–540.

Chisholm, S.W. and Brand, L.E. (1981) Persistence of cell division phasing in marine phytoplankton in continuous light after entrainment to light: dark cycles. *J. Exp. Mar. Bio. Ecol.* **51**: 107–118.

Chrétiennot-Dinet, M.-J., Courties, C., Vaquer, A., Neveux, J., Claustre, H., Lautier, J., and Machado, M.C. (1995) A new marine picoeucaryote: *Ostreococcus tauri* gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia* **34**: 285–292.

Christensen, T. (1962) Systematisk Botanik, Alger. In, In Bocher,T.W., Lange,M., and Sorensen,T. (eds), *Botanik*. Munksgaard, Copenhagen, pp. 1–178.

Coleman, A.W. (2003) ITS2 is a double-edged tool for eukaryote evolutionary comparisons. *Trends Genet.* **19**: 370–375.

Courties, C., Vaquer, A., Troussellier, M., Lautier, J., Chrétiennot-Dinet, M.J., Neveux, J., et al. (1994) Smallest eukaryotic organism. *Nature* **370**: 255–255.

Darienko, T., Gustavs, L., Mudimu, O., Menendez, C.R., Schumann, R., Karsten, U., et al. (2010) *Chloroidium*, a common terrestrial coccoid green alga previously assigned to *Chlorella* (Trebouxiophyceae, Chlorophyta). *Eur. J. Phycol.* **45**: 79–95.

Deschamps, P. and Moreira, D. (2012) Reevaluating the green contribution to diatom genomes. *Genome Biol. Evol.* **4**: 683–8.

Edgcomb, V., Orsi, W., Bunge, J., Jeon, S., Christen, R., Leslin, C., et al. (2011) Protistan microbial observatory in the Cariaco Basin, Caribbean. I. Pyrosequencing vs Sanger insights into species richness. *ISME J.* **5**: 1344–56.

Eikrem, W. and Throndsen, J. (1990) The ultrastructure of *Bathycoccus* gen. nov. and *B. prasinos* sp. nov., a non-motile picoplanktonic alga (Chlorophyta, Prasinophyceae) from the Mediterranean and Atlantic. *Phycologia* **29**: 344–350.

Falkowski, P.G., Schofield, O., Katz, M.E., Van de Schootbrugge, B., and Knoll, A.H. (2004) Why is the land green and the ocean red? In, Thierstein,H.R. and Young,J.R. (eds), *Coccolithophores: from Molecular processes to global impact*. Berlin, pp. 427–453.

Faria, D.G., Kato, A., de la Peña, M.R., and Suda, S. (2011) Taxonomy and phylogeny of *Nephroselmis clavistella* sp. nov. (Nephroselmidophyceae, Chlorophyta). *J. Phycol.* **47**: 1388–1396.

Faria, D.G., Kato, A., and Suda, S. (2012) *Nephroselmis excentrica* sp. nov. (Nephroselmidophyceae, Chlorophyta) from Okinawa-jima, Japan. *Phycologia* **51**: 271–282.

Fawley, M.W., Qin, M., and Yun, Y. (1999) The relationship between *Pseudoscourfieldia marina* and *Pycnococcus provasolii* (Prasinophyceae, Chlorophyta): evidence from 18S rDNA sequence data. *J. Phycol.* **843**: 838–843.

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

Fučíková, K., Leliaert, F., Cooper, E.D., Škaloud, P., D'Hondt, S., Clerck De, O., et al. (2014) New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data. *Front. Ecol. Evol.* **2**: 63.

Gaarder, T. (1933) Untersuchungen über Produktions und Lebensbedigungen in norwegischen Austern-Pollen. In, *Naturvidenskapelig Rekke 3. Bergens Museum*., pp. 1–64.

Le Gall, F., Rigaut-Jalabert, F., Marie, D., Garczarek, L., Viprey, M., Gobet, A., and Vaulot, D. (2008) Picoplankton diversity in the South-East Pacific Ocean from cultures. *Biogeosciences* **5**: 203–214.

Gómez, P.I. and González, M.A. (2004) Genetic variation among seven strains of *Dunaliella salina* (Chlorophyta) with industrial potential, based on RAPD banding patterns and on nuclear ITS rDNA sequences. *Aquaculture* **233**: 149–162.

Groisillier, A., Massana, R., Valentin, K., Vaulot, D., and Guillou, L. (2006) Genetic diversity and habitats of two enigmatic marine alveolate lineages. *Aquat. Microb. Ecol.* **42**: 277–291.

Guillard, R.R.L., Keller, M.D., O'Kelly, C.J., and Floyd, G.L. (1991) *Pycnococcus provasolii* gen. et spe. nov., a coccoid prasinoxanthin-containing phytoplankter from the western north Atlantic and Gulf of Mexico. *J. Phycol.* **27**: 39–47.

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.

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

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

Harðardottir, S., Lundholm, N., Moestrup, Ø., and Nielsen, T.G. (2014) Description of *Pyramimonas diskoicola* sp. nov. and the importance of the flagellate *Pyramimonas* (Prasinophyceae) in Greenland sea ice during the winter – spring transition. *Polar Biol.* 1479–1494.

Harholt, J., Moestrup, Ø., and Ulvskov, P. (2015) Why plants were terrestrial from the beginning. *Trends Plant Sci.* **21**: 1–6.

Hasegawa, T., Miyashita, H., Kawachi, M., Ikemoto, H., Kurano, N., Miyachi, S., and Chihara, M. (1996) *Prasinoderma coloniale* gen. et sp. nov., a new pelagic coccoid prasinophyte from the western Pacific Ocean. *Phycologia* **35**: 170–176.

Hedges, S.B., Blair, J.E., Venturi, M.L., and Shoe, J.L. (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. *BMC Evol. Biol.* **4**: 2.

Heimann, K., Benting, J., Timmermann, S., and Melkonian, M. (1989) The flagellar developmental cycle in algae. Two types of flagellar development in uniflagellated algae. *Protoplasma* **153**: 14–23.

Henley, W.J., Hironaka, J.L., Guillou, L., Buchheim, M.A., Buchheim, J.A., Fawley, M.W., and Fawley, K.P. (2004) Phylogenetic analysis of the "*Nannochloris*-like" algae and diagnoses of *Picochlorum oklahomensis* gen. et sp. nov. (Trebouxiophyceae, Chlorophyta). *Phycologia* **43**: 641–652.

Hori, T., Inouye, I., Horiguchi, T., and Boalch, G.T. (1985) Observations on the motile stage of *Halosphaera minor* Ostenfeld (Prasinophyceae) with special reference to the cell structure. *Bot. Mar.* **28**: 529–538.

Hori, T., Moestrup, Ø., and Hoffman, L.R. (1995) Fine structural studies on an ultraplanktonic species of *Pyramimonas*, *P. virginica* (Prasinophyceae), with a discussion of subgenera within the genus *Pyramimonas*. *Eur. J. Phycol.* **30**: 219–234.

Hori, T., Norris, R.E., and Mitsuo, A.N. (1982) Studies on the Ultrastructure and Taxonomy of the Genus *Tetraselmis* (Prasinophyceae) - I . Subgenus *Tetraselmis*. 49–61.

Hori, T., Norris, R.E., Mitsuo, A.N., and Chihara, M. (1983) Studies on the ultrastructure and taxonomy of the genus *Tetraselmis* (Prasinophyceae) - II. Subgenus *Prasinocladia*. *Bot. Mag. Tokyo* **96**: 385–392.

Hori, T., Norris, R.E., Mitsuo, A.N., and Chihara, M. (1982) Studies on the ultrastructure and taxonomy of the genus *Tetraselmis* (Prasinophyceae) - I . Subgenus *Tetraselmis* . *Bot. Mag. Tokyo* **96**: 49–61.

Hori, T., Norrm, R.E., and Chihara, M. (1986) Studies on the Ultrastructure and Taxonomy of the Genus *Tetraselmis* - III Subgenus *Parviselmis*. 123–135.

Hoshaw, R.W. and Ettl, H. (1966) *Chlamydomonas smithii* sp. nov. a Chlamydomonad interfertile with *Chlamydomonas reinhardt* . *J. Phycol.* **2**: 93–96.

Jouenne, F., Eikrem, W., Le Gall, F., Marie, D., Johnsen, G., and Vaulot, D. (2011) *Prasinoderma singularis* sp. nov. (Prasinophyceae, Chlorophyta), a solitary coccoid prasinophyte from the South-East Pacific Ocean. *Protist* **162**: 70–84.

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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**: 1647–9.

Klein, R.M. and Cronquist, A. (1967) A consideration of the evolutionary and taxonomic sgnificance of some biochemical, micromorphological, and physiological characters in the Thallophytes. *Q. Rev. Biol.* **42**: 108–296.

Kopp, R.E., Kirschvink, J.L., Hilburn, I.A., and Nash, C.Z. (2005) The Paleoproterozoic snowball Earth: a climate disaster triggered by the evolution of oxygenic photosynthesis. *Proc. Natl. Acad. Sci. U. S. A.* **102**: 11131–6.

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

Lee, H.-J. and Hur, S.-B. (2009) Genetic relationships among multiple strains of the genus *Tetraselmis* based on partial 18S rDNA sequences. *Algae* **24**: 205–212.

Lee, R.E. (2008) Phycology Cambridge. Cambridge University Press, Cambridge.

Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F., and De Clerck, O. (2012) Phylogeny and molecular evolution of the green algae. *CRC. Crit. Rev. Plant Sci.* **31**: 1–46.

Letsch, M.R., Muller-Parker, G., Friedl, T., and Lewis, L.A. (2009) *Elliptochloris marina* sp. nov. (Trebouxiophyceae, Chlorophyta), symbiotic green alga of the temperate Pacific sea anemones *Anthopleura xanthogrammatica* and *A. elegantissima* (Anthozoa, Cnidaria). *J. Phycol.* **45**: 1127–1135.

Lewin, R.A., Krienitz, L., Goericke, R., Takeda, H., and Hepperle, D. (2000) Picocystis salinarum gen. et sp nov (Chlorophyta) - a new picoplanktonic green alga. *Phycologia* **39**: 560–565.

Lie, A. a Y., Liu, Z., Hu, S.K., Jones, A.C., Kim, D.Y., Countway, P.D., et al. (2014) Investigating microbial eukaryotic diversity from a global census: Insights from a comparison of pyrotag and full-length sequences of 18S rRNA genes. *Appl. Environ. Microbiol.* **80**: 4363–4373.

Lopes dos Santos, A., Tragin, M., Gourvil, P., Noël, M.-H., Decelle, J., Romac, S., and Vaulot, D. (2016) Prasinophytes clade VII, an important group of green algae in oceanic waters : diversity and distribution. *ISME J.* **submitted**:

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

Manton, I. (1977) *Dolichomastix* (Prasinophyceae) from arctic Canada, Alaska and South Africa: a new genus of flagellates with scaly flagella. *Phycologia* **16**: 427–438.

Margulis, L. (1975) Symbiotic theory of the origin of eukaryotic organelles; criteria for proof. *Symp. Soc. Exp. Biol.* 21–38.

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

Marin, B. and Melkonian, M. (1994) Flagellar hairs in Prasinophytes (Chlorophyta): ultrastructure and distribution on the flagellar surface. *J. Phycol.* **30**: 659–678.

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

Massana, R., 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.

Mattox, K.R. and Stewart, K.D. (1984) Classification of the green algae: a concept based on comparative cytology. In, Irvine,D.E.G. and John,D.M. (eds), *Systematics of the green algae*. London and Orlando, pp. 29–72.

McFadden, G.I. (2001) Primary and secondary endosymbiosis and the origin of plastids. *J. Phycol.* **37**: 951–959.

Mishra, A., Mandoli, A., and Jha, B. (2008) Physiological characterization and stress-induced metabolic responses of *Dunaliella salina* isolated from salt pan. *J. Ind. Microbiol. Biotechnol.* **35**: 1093–101.

Miyashita, H., Ikemoto, H., Kurano, N., Miyachi, S., and Chihara, M. (1993) *Prasinococcus capsulatus* gen. et sp. nov., a new marine coccoid prasinophyte. *J. Gen. App. Microbio.* **39**: 571–582.

Moestrup, Ø. (1982) Flagellar structure in algae: a review, with new observations particularly on the Chrysophyceae, Phaeophyceae (Fucophyceae), Euglenophyceae, and *Reckertia*. *Phycologia* **21**: 427–528.

Moestrup, Ø. (1991) Further Studies of presumedly primitive green algar, including the description of Pedinophyceae Class. Nov. and *Resultor* gen. nov. *J. Phycol.* **27**: 119–133.

Moestrup, Ø. (1984) Further studies on *Nephroselmis* and its allies (Prasinophyceae). II. *Mamiella* gen. nov., Mamiellaceae fam. nov., Mamiellales ord. nov. *Nord. J. Bot.* **4**: 109–121.

Moestrup, Ø. and Throndsen, J. (1988) Light and electron microscopical studies on *Pseudoscourfieldia marina*, a primitive scaly green flagellate (Prasinophyceae) with posterior flagella. *Can. J. Bot.* **66**: 1415–1434.

Monier, A., Sudek, S., Fast, N.M., and Worden, A.Z. (2013) Gene invasion in distant eukaryotic lineages: discovery of mutually exclusive genetic elements reveals marine biodiversity. *ISME J.* **7**: 1764–1774.

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**: 607–10.

Moro, I., La Rocca, N., Dalla Valle, L., Moschin, E., Negrisolo, E., and Andreoli, C. (2002) *Pyramimonas australis* sp. nov. (Prasinophyceae, Chlorophyta) from Antarctica: fine structure and molecular phylogeny. *Eur. J. Phycol.* **37**: 103–114.

Moustafa, A., Beszteri, B., Maier, U.G., Bowler, C., Valentin, K., and Bhattacharya, D. (2009) Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**: 1724–6.

Nägeli, C. (1849) Gattungen einzelliger Algen physiologisch und systematisch bearbeitet Schulthess,F. (ed) Zürich.

Nakayama, T., Marin, B., Kranz, H.D., Surek, B., Huss, V. a, Inouye, I., and Melkonian, M. (1998) The basal position of scaly green flagellates among the green algae (Chlorophyta) is revealed by analyses of nuclear-encoded SSU rRNA sequences. *Protist* **149**: 367–80.

Nakayama, T., Suda, S., Kawachi, M., and Inouye, I. (2007) Phylogeny and ultrastructure of *Nephroselmis* and *Pseudoscourfieldia* (Chlorophyta), including the description of *Nephroselmis anterostigmatica* sp. nov. and a proposal for the Nephroselmidales ord. nov. *Phycologia* **46**: 680–697.

Nelson, W.A. and Ryan, K.G. (1986) *Palmophyllum umbracola* sp. nov. (Chlorophyta) from offshore islands of northern New Zealand. *Phycologia* **25**: 168–177.

O'Kelly, C.J. and Floyd, G.L. (1983) Flagellar apparatus absolute orientations and the phylogeny of the green algae. *Biosystems* **16**: 227–251.

Orsi, W., Song, Y.C., Hallam, S., and Edgcomb, V. (2012) Effect of oxygen minimum zone formation on communities of marine protists. *ISME J.* **6**: 1586–601.

Pennick, N.C. (1984) Comparative ultrastructure and occurrence of scales in Pyramimonas (chlorophyta, prasinophyceae). *Arch. für Protistenkd.* **128**: 3–11.

Pennick, N.C. (1982) Studies of the External Morphology of Pyramimonas 6. Pyramimonas cirolanae sp. nov. *Arch. für Protistenkd.* **125**: 87–94.

Peterfi, L.S. and Manton, I. (1968) Observations with the electron microscope on *Asteromonas gracilis* Artari emend. (*Stephanoptera gracilis* (Artari) wisl.), with some comparative observations on *Dunaliella* sp. *Br. Phycol. Bull.* **3**: 423–440.

Potter, D., Lajeunesse, T.C., Saunders, G.W., and Andersen, R.A. (1997) Convergent evolution masks extensive biodiversity among marine coccoid picoplankton. *Biodivers. Conserv.* **6**: 99–107.

Preetha, K. (2012) Phenotypic and genetic characterization of *Dunaliella (Chlorophyta)* from Indian salinas and their diversity. *Aquat. Biosyst.*

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009) Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**: 1641–1650.

de Reviers, B. (2003) Biologie et phylogénie des algues Tome 2 Belin. Le Guyader,H. and Laurent,J. (eds).

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

Round, F.E. (1963) The taxonomy of the Chlorophyta. *Br. Phycol. Bull.* **2**: 224–235.

Round, F.E. (1971) The taxonomy of the Chlorophyta. II. *Br. Phycol. J.* **6**: 235–264.

Scherffel, A. (1912) Zwei neue trichocystenartige Bildungen fürhende Flagellaten. *Arch. für Protistenkd.* **27**: 94–128.

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**: 7537–41.

Scott, C., Lyons, T.W., Bekker, A., Shen, Y., Poulton, S.W., Chu, X., and Anbar, A.D. (2008) Tracing the stepwise oxygenation of the Proterozoic ocean. *Nature* **452**: 456–9.

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

Sieburth, J.M., Keller, M.D., Johnson, P.W., and Myklestad, S.M. (1999) Widespread occurence of the oceanic ultraplankter, *Prasinococcus capsulatus* (Prasinophyceae), the diagnostic "Golgi-decapore complex" and the newly described polysaccharide "Capsulan." *J. Phycol.* **35**: 1032–1043.

Šlapeta, J., López-García, P., and Moreira, D. (2006) Global dispersal and ancient cryptic species in the smallest marine eukaryotes. *Mol. Biol. Evol.* **23**: 23–29.

Sluiman, H.J., Kouwets, F.A.C., and Blommers, P.C.J. (1989) Classification and definition of cytokinetic patterns in green algae: Sporulation versus (vegetative) cell division. *Arch. für Protistenkd.* **137**: 277–290.

Stewart, K.D. and Mattox, K.R. (1975) Comparative cytology, evolution and classification of the green algae with some consideration of the origin of other organisms with chlorophylls A and B. *Bot. Rev.* **41**: 104–135.

Stewart, K.D. and Mattox, K.R. (1978) Structural evolution in the flagellated cells of green algae and land plants. *Biosystems* **10**: 145–152.

Subirana, L., Péquin, B., Michely, S., Escande, M.L., Meilland, J., Derelle, E., et al. (2013) Morphology, genome plasticity, and phylogeny in the genus *Ostreococcus* reveal a cryptic species, *O. mediterraneus* sp. nov. (Mamiellales, Mamiellophyceae). *Protist* **164**: 643–659.

Suda, S., Bhuiyan, M.A.H., and Faria, D.G. (2013) Genetic diversity of *Pyramimonas* from Ryukyu Archipelago, Japan (Chlorophyceae, Pyramimonadales). *J. Mar. Sci. Technol.* **21**: 285–296.

Takishita, K., Yubuki, N., Kakizoe, N., Inagaki, Y., and Maruyama, T. (2007) Diversity of microbial eukaryotes in sediment at a deep-sea methane cold seep: Surveys of ribosomal DNA libraries from raw sediment samples and two enrichment cultures. *Extremophiles* **11**: 563–576.

Tappan, H. and Loeblich, A.R. (1973) Evolution of the oceanic plankton. *Earth-Science Rev.* **9**: 207–240.

Throndsen, J. and Zingone, A. (1997) *Dolichomastix tenuilepis* sp. nov., a first insight into the microanatomy of the genus *Dolichomastix* (Mamiellales, Prasinophyceae, Chlorophyta). *Phycologia* **36**: 244–254.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., et al. (2015) Eukaryotic plankton diversity

in the sunlit ocean. *Science (80-.).* **348**: 1261605–1261605.

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

Viprey, M., Guillou, L., Ferréol, M., 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.

Wiebe, P.H., Remsen, C.C., and Vaccaro, R.F. (1974) *Halosphaera viridis* in the Mediterranean sea: size range, vertical distribution, and potential energy source for deep-sea benthos. *Deep Sea Res. Oceanogr. Abstr.* **21**: 657–667.

Wu, W., Huang, B., Liao, Y., and Sun, P. (2014) Picoeukaryotic diversity and distribution in the subtropical-tropical South China Sea. *FEMS Microbiol. Ecol.* **89**: 563–579.

Yamaguchi, H., Nakayama, T., and Inouye, I. (2013) Proposal of *Microsquama subgen* . nov . for *Nephroselmis pyriformis* (Carter) Ettl. *Phycol. Res.* 268–269.

Yamaguchi, H., Suda, S., Nakayama, T., Pienaar, R.N., Chihara, M., and Inouye, I. (2011) Taxonomy of *Nephroselmis viridis* sp. nov. (Nephroselmidophyceae, Chlorophyta), a sister marine species to freshwater *N. olivacea*. *J. Plant Res.* **124**: 49–62.

Yoon, H.S., Hackett, J.D., Ciniglia, C., Pinto, G., and Bhattacharya, D. (2004) A molecular timeline for the origin of photosynthetic eukaryotes. *Mol. Biol. Evol.* **21**: 809–18.

Zechman, F.W., Verbruggen, H., Leliaert, F., Ashworth, M., Buchheim, M.A., Fawley, M.W., et al. (2010) An unrecognized ancient lineage of green plants persists in deep marine waters. *J. Phycol.* **46**: 1288–1295.

Zingone, A., Borra, M., Brunet, C., Forlani, G., Kooistra, W.H.C.F., and Procaccini, G. (2002) Phylogenetic position of *Crustomastix stigmatica* sp. nov. and *Dolichomastix tenuilepis* in relation to Mamiellales (Prasinophyceae, Chlorophyta). *J. Phycol.* **38**: 1024–1039.

List of Figures

List of Supplementary Figures

Fig. S1: Pie chart distribution by oceanic areas of sequences of major Chlorophyta lineages for cultures (A) and environmental samples (B). Sequences were regrouped in rectangular regions using the R software (Table S5). Areas may overlap in regions where no sequences were recorded.

Fig. S2: Range of variation of the percentage of sequence identity for the full 18S rRNA, the V4 and V9 region sequences for each Chlorophyta lineage (maximum, third quartile, median, first quartile, minimum). Identity matrixes were calculated using the function *seqidentity* of the R package bio3d. Number of sequences is indicated between parentheses.

List of Supplementary Tables

Table S1: List of reference sequences used for building the tree of Fig.1.

Table S2: Number of species per described genus for each prasinophyte clade according to AlgaeBase (http://www.algaebase.org/). Number of strain sequences considered for each genus.

Table S3: Member of the UTC clades that are found in the marine environment with the number of sequences considered.

Table S4: Sequences from the Roscoff Culture Collection recently deposited to GenBank and used in this work.

Table S5: Oceanic regions used to regroup sequences presented in Figure S2.

Supplementary data (available from Figshare)

Data S1. Annotated Chlorophyta GenBank sequences. Two files containing sequences in fasta format and their annotated taxonomy. These files can be used to assign metabarcoding data using software such as mothur or Qiime: https://figshare.com/s/c5edff05cd551e466320

Data S2. Metadata for Chlorophyta sequences: https://figshare.com/s/a72fabf6add26f4c98ce

Data S3. Reference alignment for Fig.1: https://figshare.com/s/3ebc93f306f3935cab80

Supplementary Figures

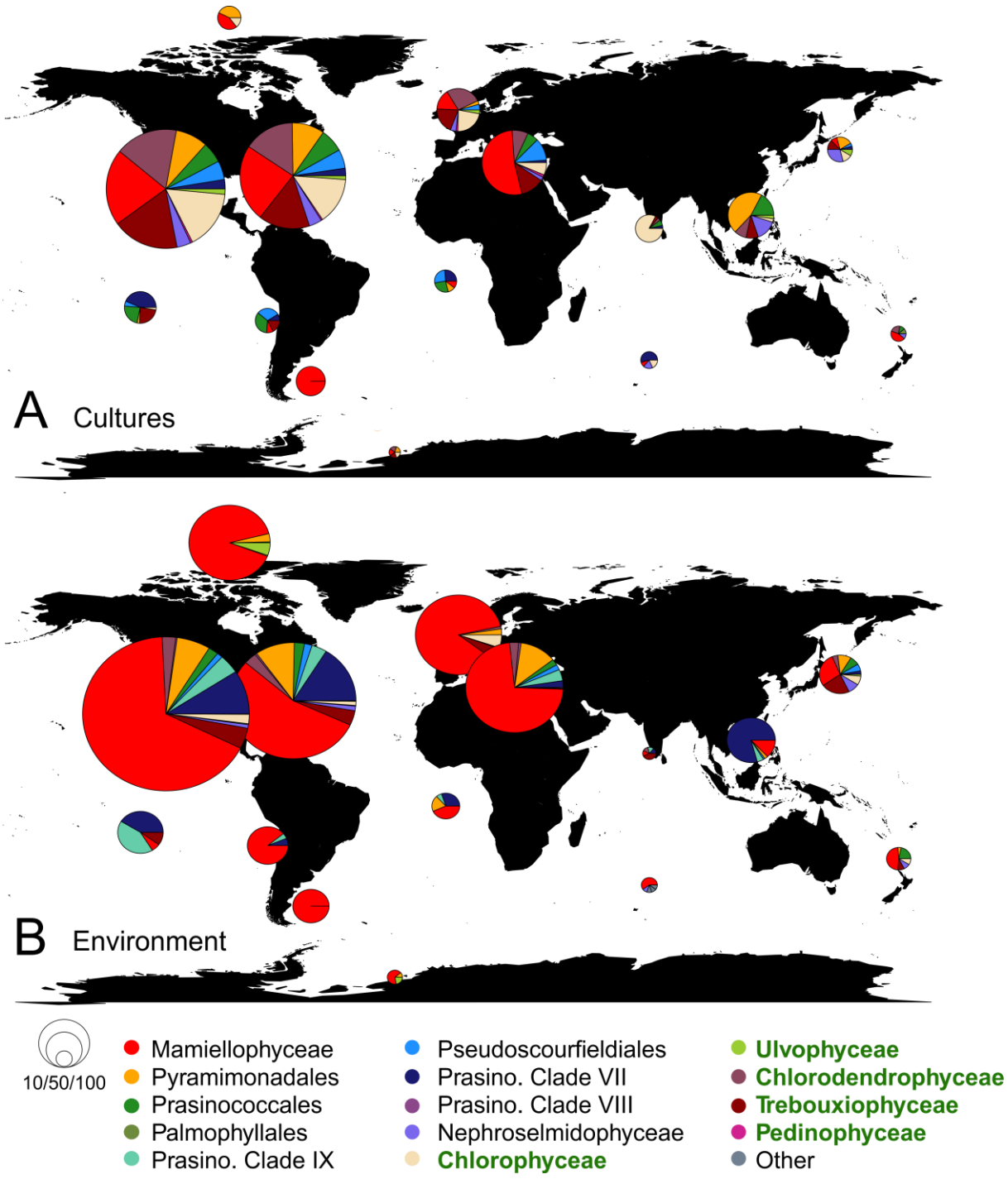


Fig. S1: Pie chart distribution by oceanic areas of sequences of major Chlorophyta lineages for cultures (A) and environmental samples (B). Sequences were regrouped in rectangular regions using the R software (Table S5). Areas may overlap in regions where no sequences were recorded.

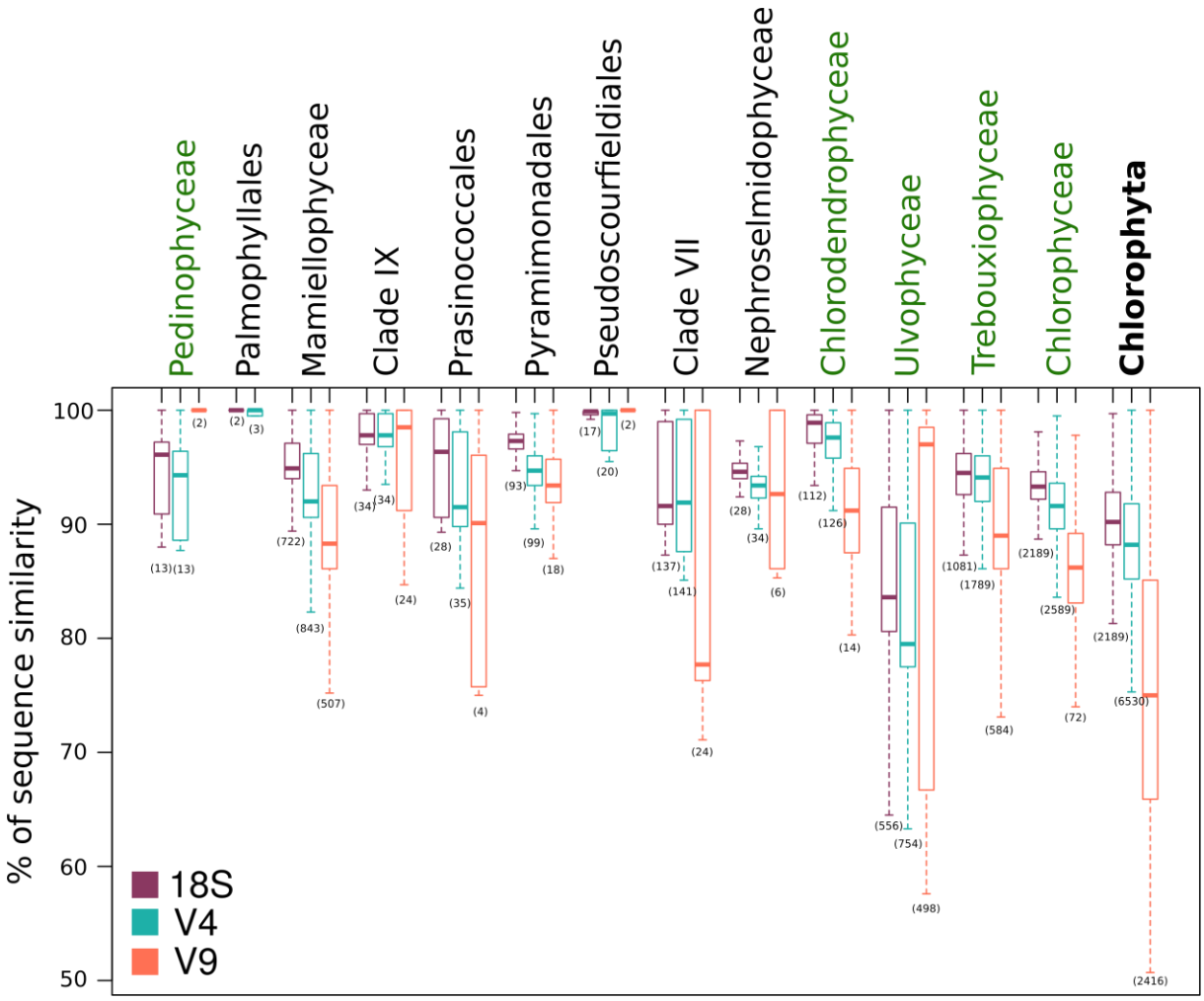


Fig. S2: Range of variation of the percentage of sequence identity for the full 18S rRNA, the V4 and V9 region sequences for each Chlorophyta lineage (maximum, third quartile, median, first quartile, minimum). Identity matrixes were calculated using the function *seqidentity* of the R package bio3d. Number of sequences is indicated between parentheses.

Supplementary Tables

Table S1: List of reference sequences used for building the tree of Fig.1.

Accession	Class/Order	PR² Genus	PR² Species	Strain	Clone
FJ619275	Palmophyllales	*Palmophyllum*	*Palmophyllum_umbracola*	BISH730325	
FJ619277		*Verdigellas*	*Verdigellas_peltata*	HBFH7821, HBFH7822	
EU143487	Prasino-Clade-IX	Prasino-Clade-9-A_X	Prasino-Clade-9-A_X_sp.		PROSOPE.CD-50m.38
FJ537330		Prasino-Clade-9-B_X	Prasino-Clade-9-B_X_sp.		Biosope_T41.151
HM474493		Prasino-Clade-9-B_X	Prasino-Clade-9-B_X_sp.		T41_W01D.013
KJ759349		Prasino-Clade-9_XXX	Prasino-Clade-9_XXX_sp.		SGYO1200
AB058375	Prasinococcales	Prasinococcales-Clade-B_X	Prasinococcales-Clade-B1_X_sp.	MBIC10622	
EU371173		Prasinococcales-Clade-B_X	Prasinococcales-Clade-B2_X_sp.		NPK2_59
FJ997212		*Prasinoderma*	*Prasinoderma_singularis*	RCC927	
FN562437		*Prasinoderma*	*Prasinoderma_coloniale*	CCMP 1413	
AB017121	Pyramimonadales	*Pyramimonas*	*Pyramimonas_disomata*	Singapore	
AB017122		*Pyramimonas*	*Pyramimonas_olivacea*	Shizugawa Bay, Miyagi	
AB017124		*Pyramimonas*	*Pyramimonas_parkeae*	Hachijo	
AB017125		*Halosphaera*	*Halosphaera_sp.*	Shizugawa	
AB017126		*Cymbomonas*	*Cymbomonas_tetramitiformis*	Shizugawa	
AB017127		*Pterosperma*	*Pterosperma_cristatum*	Yokohama	
AB052289		*Pyramimonas*	*Pyramimonas_aurea*	NBRC102947	
AB183649		*Prasinopapilla*	*Prasinopapilla_vacuolata*	NBRC102950	
AB854013		*Pyramimonas*	*Pyramimonas_sp.*	MIY1-8	
AB854017		*Pyramimonas*	*Pyramimonas_sp.*	OD6P1	
AJ404886		*Pyramimonas*	*Pyramimonas_australis*		
EU330223		*Pyramimonas*	*Pyramimonas_mucifera*	WitsPyrami	
FN562441		*Pyramimonas*	*Pyramimonas_tetrarhynchus*	SCCAP K-0002	
HQ111511		*Pyramimonas*	*Pyramimonas_gelidicola*	Ace Lake	
JN934689		*Pyramimonas*	*Pyramimonas_sp.*	RCC2500	
AB017128	Mamiellophyceae	*Mantoniella*	*Mantoniella_antarctica*		
AB183589		*Micromonas*	Micromonas_Clade-A.ABC.1-2	NBRC102743	
AB183628		*Crustomastix*	*Crustomastix_didyma*	MBIC10709	
AB275082		Dolichomastigaceae-B	Dolichomastigaceae-B_sp.		DSGM-82
AB491653		*Monomastix*	*Monomastix_minuta*		
AB721076		Crustomastigaceae-AB	Crustomastigaceae-AB_sp.		RW4_2011
AJ629844		*Crustomastix*	*Crustomastix_stigmata*	CCMP3273, CCMP2493	

Accession	Class/Order	PR² Genus	PR² Species	Strain	Clone
AY046690		Dolichomastigaceae-B	Dolichomastigaceae-B_sp.		
AY425321		RCC391	RCC391_sp.	RCC 391	
AY954994		*Micromonas*	Micromonas_pusilla-clade-C.D.5	CCMP1545	
CP000588		*Ostreococcus*	*Ostreococcus_lucimarinus*	CCE9901	
EU143397		Crustomastigaceae-AB	Crustomastigaceae-AB_sp.		PROSOPE.CD-15m.160
EU143398		Crustomastigaceae-C	Crustomastigaceae-C_sp.		PROSOPE.CM-5m.179
FN562445		*Monomastix*	*Monomastix_opisthostigma*	CCAC0206	
FN562449		*Dolichomastix*	*Dolichomastix_tenuilepis*	CCAC 1680 B	
FN562450		*Mamiella*	*Mamiella_gilva*	PLY 197	
FN562452		*Micromonas*	Micromonas_Clade-B.E.3	CCAC1681-B	
FN562453		*Bathycoccus*	*Bathycoccus_prasinos*	SCCAP K-0417	
HM135071		Monomastigaceae_X	Monomastigaceae_X_sp.		STFeb_352
HQ868900		Dolichomastigaceae-B	Dolichomastigaceae-B_sp.		SHAX386
JN862916		*Ostreococcus*	*Ostreococcus_mediterraneus*	RCC2583	
JN934683		*Micromonas*	Micromonas_Clade-B_arctic	RCC2308	
KJ763360		Dolichomastigaceae-A	Dolichomastigaceae-A_sp.		SGUH865
X73999		*Mantoniella*	*Mantoniella_squamata*	CCAP 1965/1	
Y15814		*Ostreococcus*	Ostreococcus_clade-C	OTTH 0595genome	
AB058377	Pseudoscourfieldiales	Pycnococcaceae-clade1	Pycnococcaceae-clade1-sp.	NBRC102846	
AY425304		*Pseudoscourfieldia*	*Pseudoscourfieldia_marina*	RCC 261	
FR865764		*Pycnococcus*	*Pycnococcus_provasolii*	CCMP1197, CCMP2194	
AJ402345	Prasino-Clade-VII	Prasino-Clade-VII-A-6	Prasino-Clade-VII-A-6_sp.		
FJ537298		Prasino-Clade-VII-B-1	Prasino-Clade-VII-B-1_sp.		Biosope_T123.014
FJ537305		Prasino-Clade-VII-B-2	Prasino-Clade-VII-B-2_sp.		Biosope_T19.017
FJ537346		Prasino-Clade-VII-A-3	Prasino-Clade-VII-A-3_sp.		Biosope_T65.119
KF422632		Prasino-Clade-VII-A-1	Prasino-Clade-VII-A-1_sp.	RCC998	
KF615770		Prasino-Clade-VII-A-4	Prasino-Clade-VII-A-4_sp.	CCMP1998	
KF899843		Prasino-Clade-VII-A-5	Prasino-Clade-VII-A-5_sp.	CCMP2175	
U40921		Prasino-Clade-VII-A-2	Prasino-Clade-VII-A-2_sp.	RCC 15	
EU143504	Prasino-Clade-VIII	Prasino-Clade-VIII_XXX	Prasino-Clade-VIII_XXX_sp.		PROSOPE.C1-30m.214
EU143505		Prasino-Clade-VIII_XXX	Prasino-Clade-VIII_XXX_sp.		PROSOPE.C1-30m.229
EU143506		Prasino-Clade-VIII_XXX	Prasino-Clade-VIII_XXX_sp.		PROSOPE.C1-30m.93
AB058391	Nephroselmidophyceae	*Nephroselmis*	*Nephroselmis_pyriformis*	MBIC11099	
AB158373		*Nephroselmis*	*Nephroselmis_anterostigmatica*	MBIC11158	AB158373

Accession	Class/Order	PR² Genus	PR² Species	Strain	Clone
AB158375		*Nephroselmis*	*Nephroselmis_spinosa*	NIES 935	
AB214975		*Nephroselmis*	*Nephroselmis_clavistella*	MBIC11149	
AB533370		*Nephroselmis*	*Nephroselmis_viridis*	Fij7	
AB601448		*Nephroselmis*	*Nephroselmis_excentrica*	BS2-3	
AB605798		*Nephroselmis*	*Nephroselmis_astigmatica*	SS21	
FN562435		*Nephroselmis*	*Nephroselmis_rotunda*	M0932	
X74754		*Nephroselmis*	*Nephroselmis_olivacea*	SAG 40.89	
HE610136	Pedinophyceae	*Marsupiomonas*	*Marsupiomonas_pelliculata*	PLY 441	
JN592588		*Pedinomonas*	*Pedinomonas_minor*	SAG 1965-3	
JN592589		*Pedinomonas*	*Pedinomonas_tuberculata*	SAG 42.84	
HE610130	Chlorodendrophyceae	*Tetraselmis*	*Tetraselmis_cordiformis*	SAG 26.82	
HE610131		*Tetraselmis*	*Tetraselmis_marina*	CCMP898	
JN903999		*Tetraselmis*	*Tetraselmis_chuii*	SAG 1.96	
X68484		*Scherffelia*	*Scherffelia_dubia*	SAG 17.89	
X70802		*Tetraselmis*	*Tetraselmis_striata*	Ply 443	
AB058346	Ulvophyceae	Ulvophyceae_XXX	Ulvophyceae_XXX_sp.	MBIC10461	
AF015279		*Monostroma*	*Monostroma_grevillei*		
AJ416104		*Dangemannia*	*Dangemannia_microcystis*	SAG 2022, CCAP 233/3	
FN562431		*Oltmannsiellopsis*	*Oltmannsiellopsis_viridis*	NIES 360, 8280G41-2	
JF932253		*Caulerpa*	*Caulerpa_sp.*		
JQ315653		*Klebsormidium*	*Klebsormidium_subtilissimum*	KMMCC 1083	
U41102		*Pseudoneochloris*	*Pseudoneochloris_marina*	UTEX 1445	
AB058336	Chlorophyceae	*Chlorococcum*	*Chlorococcum_littorale*	NBRC 102761	
AJ410454		*Chloromonas*	*Chloromonas_augustae*	SAG 13.89	
DQ009744		*Asteromonas*	*Asteromonas_gracilis*	CCMP 813	
DQ009751		*Chlamydomonas*	*Chlamydomonas_sp.*	CCMP 219	
DQ009769		*Dunaliella*	*Dunaliella_sp.*	CCMP 367	
DQ009778		*Dunaliella*	*Dunaliella_peircei*	UTEX LB 2192	
DQ324021		*Dunaliella*	*Dunaliella_sp.*	SPMO 601-1	
FN690715		CW-Chlamydomonadales_X	CW-Chlamydomonadales_X_sp.		7-D1
FN824388		*Chaetophora*	*Chaetophora_elegans*	CCAP 413/2	
FR865748		*Chaetopeltis*	*Chaetopeltis_orbicularis*	CCAP 412/1	
GQ122365		*Chlorococcum*	*Chlorococcum_minutum*		
JN934685		*Carteria*	*Carteria_sp.*	RCC2487	
JQ315503		*Chlamydomonas*	*Chlamydomonas_hedleyi*	KMMCC 188	
JQ315546		*Monoraphidium*	*Monoraphidium_sp.*	KMMCC 1137	JQ315546

Accession	Class/Order	PR² Genus	PR² Species	Strain	Clone
JQ315598		*Scenedesmus*	*Scenedesmus_sp.*	KMMCC 373	
JQ315599		*Scenedesmus*	*Scenedesmus_sp.*	KMMCC 1297	
JQ315600		*Scenedesmus*	*Scenedesmus_sp.*	KMMCC 245	
JX413790		*Coelastrum*	*Coelastrum_sp.*	HA-1	
U70787		*Chlamydomonas*	*Chlamydomonas_noctigama*	SAG 19.73	
AB058309	Trebouxiophyceae	*Picochlorum*	*Picochlorum_sp.*	NBRC 102739	
AB080301		*Marvania*	*Marvania_coccoides*	CCAP 251/1b	
AB080304		*Picochlorum*	*Picochlorum_eukaryotum*	SAG 55.87	
AB183575		*Chloroidium*	*Chloroidium_saccharophilum*	NBRC102700	
AJ131691		*Picochlorum*	*Picochlorum_sp.*	RCC 11	
AJ311569		*Koliella*	*Koliella_antarctica*	SAG 2030	
AJ431572		*Desmococcus*	*Desmococcus_olivaceus*	SAG 35.83	
EF526889		*Trebouxia*	*Trebouxia_sp.*		NA2_1H8
EU127469		*Coccomyxa*	*Coccomyxa_parasitica*	CCAP 216/18	
FM205834		*Chlorella*	*Chlorella_sorokiniana*	SAG 211-8k	
FM205839		*Didymogenes*	*Didymogenes_anomala*	SAG 18.91	
FM205858		*Chlorella*	*Chlorella_sp.*	CCAP 222/18	
FM205866		*Micractinium*	*Micractinium_pusillum*	SAG 13.81	
FM205879		*Diacanthos*	*Diacanthos_belenophorus*	SAG 42.98	
FM205884		*Actinastrum*	*Actinastrum_hantzschii*	CCAP 200/3	
FR865659		*Chlorella*	*Chlorella_vulgaris*	CCAP 211/11Q	
GQ487236		*Hindakia*	*Hindakia_tetrachotoma*	CCAP 222/56	
GU017647		*Asterochloris*	*Asterochloris_phycobiontica*	SAG 26.81	
JQ315611		*Stichococcus*	*Stichococcus_bacillaris*	KMMCC 169	
KM020066		*Pseudostichococcus*	*Pseudostichococcus_monallantoides*	SAG 380-1	
KM020189		*Heterochlorella*	*Heterochlorella_luteoviridis*	SAG 2196	
Z21553		*Trebouxia*	*Trebouxia_asymmetrica*	SAG 48.88	
Z68695		*Leptosira*	*Leptosira_obovata*	SAG 445-1	

Table S2: Number of species per described genus for each prasinophyte clade according to AlgaeBase (http://www.algaebase.org/). Number of strain sequences considered for each genus.

Class/Order	Genus	Number of described species	Number of 18S sequences from strains in PR²
Pyramimonadales	*Amphoraemonas*	1	
	Angulomonas	2	
	Chloraster	1	
	Coccopterum	1	
	Cymbomonas	3	1
	Gyromitus	2	
	Halosphaera	4	1
	Korschikoffia	2	
	Kuzminia	1	
	Poccillomonas	1	
	Polyasterias	1	
	Polyblepharides	2	
	Prasinochloris	1	
	Prasinopapilla	0	
	Printziella	1	
	Protoaceromonas	1	
	Pterosperma	19	2
	Pyramimonas	41	58
	Stephanoptera	2	
	Sycamina	1	
	Tasmanites	3	
	Trichloridella	1	
Mamiellophyceae	*Bathycoccus*	1	6
	Crustomastrix	2	4
	Dolichomastrix	4	4
	Mamiella	1	1
	Mantoniella	2	3
	Micromonas	1	44
	Monomastrix	5	5
	Ostreococcus	3	78
Nephroselmidophyceae	*Anticomonas*	1	
	Argillamonas	1	
	Bipedinomonas	2	
	Fluitomonas	6	
	Hiemalomonas	1	
	Myochloris	2	
	Nephroselmis	14	29

Class/Order	Genus	Number of described species	Number of 18S sequences from strains in PR²
	Prottractomonas	2	
	Pseudopedinomonas	1	
	Sennia	3	
	Sinamonas	1	
Pseudoscourfieldiales	*Pseudoscourfieldia*	1	14
	Pycnococcus	2	
Prasinococcales	*Prasinococcus*	1	5
	Prasinoderma	2	18
CladeVII	*Picocystis*	1	11
Palmophyllales	*Palmoclathrus*	1	
	Palmophyllum	2	2
	Verdigellas	3	1
Chlorodendrophyceae	*Pachysphaera*	1	
	Prasinocladus	3	
	Scherfellia	8	1
	Tetraselmis	34	95
Pedinophyceae	*Anisomonas*	1	
	Dioriticamonas	1	
	Marsupiomonas	1	
	Pedinomonas	12	9
	Resultomonas	1	4
	Scourfieldia	7	

Table S3: Member of the UTC clades that are found in the marine environment with the number of sequences considered.

Class	Species	Number of sequences
Ulvophyceae	*Halochlorococcum sp.*	1
Trebouxiophyceae	*Chlorella sp.*	49
	Chlorella vulgaris	13
	Chlorella stigmatophora	2
	Chloroidium saccharophila/saccharophilum	17
	Coccomyxa sp.	8
	Elliptochloris marina/sp.	16
	Picochlorum sp.	176
	Pseudostichococcus monallantoides	3
	Stichococcus bacillaris/sp.	23
	Trebouxia sp.	1
	Schizochlamydella capsulata	1
	Phyllosiphon sp.	1
	Parietochloris sp.	1
	Oocystis sp.	3
	Koliella antarctica	1
	Desmococcus olivaceus	1
Chlorophyceae	*Asteromonas gracilis*	3
	Brachiomonas sp.	1
	Carteria sp.	1
	Chlamydomonas asymmetrica	1
	Chlamydomonas concordia	1
	Chlamydomonas hedleyi	5
	Chlamydomonas kuwadae	1
	Chlamydomonas noctigama	1
	Chlamydomonas parkeae	3
	Chlamydomonas raudensis	1
	Chlamydomonas reginae	3
	Chlamydomonas sp.	26
	Chlorococcum sp.	17
	Chloromonas augustae	1
	Chlorosarcinopsis gelatinosa	2
	Coelastrum sp.	1
	Desmodesmus sp.	4
	Dunaliella salina	34
	Dunaliella sp.	62
	Dunaliella tertiolecta	7
	Scenedesmus sp.	22

Class	Species	Number of sequences
Chlorophyceae	*Halosarcinochlamys cherokeensis*	1
	Monoraphidium sp.	2
	Plagiobryum donianum	1
	Hemiflagellochloris kazakhstanica	1

Table S4: Sequences from the Roscoff Culture Collection recently deposited to GenBank and used in this work.

Class	Species	RCC	Accession	length
Chlorodendrophyceae	*Tetraselmis_chui*	128	KT860866	1566
	Tetraselmis_chui	129	KT860867	1601
	Tetraselmis_convolutae	1563	KT860913	1647
	Tetraselmis_convolutae	1564	KT860914	1649
	Tetraselmis_rubens	132	KT860870	1564
	Tetraselmis_rubens	133	KT860871	1565
	Tetraselmis_sp.	1936	KT860727	638
	Tetraselmis_sp.	1942	KT860728	779
	Tetraselmis_sp.	1946	KT860729	726
	Tetraselmis_sp.	1947	KT860730	660
	Tetraselmis_sp.	1949	KT860731	640
	Tetraselmis_sp.	1975	KT860790	486
	Tetraselmis_sp.	1976	KT860791	479
	Tetraselmis_sp.	2604	KT860822	720
	Tetraselmis_sp.	2628	KT860824	626
	Tetraselmis_sp.	2629	KT860825	571
	Tetraselmis_sp.	119	KT860857	1593
	Tetraselmis_sp.	120	KT860858	1643
	Tetraselmis_sp.	121	KT860859	1622
	Tetraselmis_sp.	122	KT860860	1521
	Tetraselmis_sp.	123	KT860861	1079
	Tetraselmis_sp.	124	KT860862	1630
	Tetraselmis_sp.	125	KT860863	1239
	Tetraselmis_sp.	126	KT860864	1626
	Tetraselmis_sp.	127	KT860865	459
	Tetraselmis_sp.	233	KT860876	1629
	Tetraselmis_sp.	571	KT860880	1628
	Tetraselmis_sp.	1755	KT860916	1648
	Tetraselmis_sp.	235	KT860627	546
	Tetraselmis_sp.	348	KT860643	549
	Tetraselmis_striata	130	KT860868	1578
	Tetraselmis_striata	131	KT860869	1625
Chlorophyceae	*Chlamydomonas_concordia*	1	KT860848	1634
	Chlamydomonas_reginae	2	KT860849	1654
	Chlamydomonas_sp.	300	KT860660	1747
	Chlamydomonas_sp.	2512	KT860764	793
	Chlamydomonas_sp.	2607	KT860823	640
	Chlorophyceae_XXX_sp.	666	KT860676	521
	Chlorophyceae_XXX_sp.	2955	KT860803	664
	Dunaliella_sp.	5	KT860850	1621

Class	Species	RCC	Accession	length
	Dunaliella_tertiolecta	6	KT860851	1610
Mamiellophyceae	*Ostreococcus_clade-B*	410	KT860680	813
	Ostreococcus_clade-B	141	KT878663	1034
	Ostreococcus_lucimarinus	1616	KT860704	630
	Ostreococcus_lucimarinus	1565	KT860709	621
	Ostreococcus_lucimarinus	1566	KT860710	606
	Ostreococcus_lucimarinus	1645	KT860712	622
	Ostreococcus_lucimarinus	1662	KT860713	663
	Ostreococcus_lucimarinus	2343	KT860781	640
	Ostreococcus_lucimarinus	2344	KT860782	856
	Ostreococcus_lucimarinus	675	KT860887	
	Ostreococcus_lucimarinus	747	KT860895	1616
	Ostreococcus_lucimarinus	798	KT860898	1616
	Ostreococcus_lucimarinus	343	KT860633	549
	Ostreococcus_lucimarinus	550	KT860634	550
	Ostreococcus_lucimarinus	420	KT878674	676
	Ostreococcus_sp.	468	KT860674	473
	Ostreococcus_sp.	1120	KT860804	671
	Ostreococcus_sp.	429	KT860644	550
	Ostreococcus_sp.	426	KT860647	691
	Ostreococcus_sp.	427	KT860648	653
	Ostreococcus_sp.	428	KT878675	610
	Ostreococcus_sp.	462	KT878676	587
	Ostreococcus_sp.	467	KT860653	674
	Ostreococcus_sp.	453	KT860656	678
	Ostreococcus_sp.	454	KT860657	723
	Ostreococcus_tauri	1560	KT860912	1595
Nephroselmidophyceae	*Nephroselmis_anterostigmatica*	1805	KT860917	1623
	Nephroselmis_anterostigmatica	1806	KT860918	1595
	Nephroselmis_anterostigmatica	1807	KT860919	1623
	Nephroselmis_anterostigmatica	1808	KT860920	1633
	Nephroselmis_clavistella	3064	KT860827	1105
	Nephroselmis_clavistella	3073	KT860837	1031
	Nephroselmis_clavistella	3074	KT860838	1068
	Nephroselmis_pyriformis	2499	KT860763	793
Pseudoscourfieldiales	*Pycnococcus_provasolii*	519	KT860670	669
	Pycnococcus_provasolii	522	KT860678	799
	Pycnococcus_provasolii	730	KT860682	700
	Pycnococcus_provasolii	734	KT860683	496
	Pycnococcus_provasolii	733	KT878683	660
	Pycnococcus_provasolii	731	KT860684	702
	Pycnococcus_provasolii	823	KT860693	470

Class	Species	RCC	Accession	length
Pseudoscourfieldiales	*Pycnococcus_provasolii*	1751	KT860725	859
	Pycnococcus_provasolii	1931	KT860778	679
	Pycnococcus_provasolii	2363	KT860779	658
	Pycnococcus_provasolii	2364	KT860780	548
	Pycnococcus_provasolii	2361	KT860787	630
	Pycnococcus_provasolii	2365	KT860788	647
	Pycnococcus_provasolii	2338	KT860792	533
	Pycnococcus_provasolii	581	KT860810	682
	Pycnococcus_provasolii	709	KT860812	718
	Pycnococcus_provasolii	885	KT860821	688
	Pycnococcus_provasolii	3054	KT860831	1106
	Pycnococcus_provasolii	2336	KT860844	1648
	Pycnococcus_provasolii	3055	KT860845	1637
	Pycnococcus_provasolii	899	KT860909	1124
	Pycnococcus_provasolii	251	KT860629	393
	Pycnococcus_provasolii	432	KT860637	548
	Pycnococcus_provasolii	444	KT860639	548
	Pycnococcus_provasolii	459	KT860651	503
	Pycnococcus_provasolii	245	KT860654	744
	Pycnococcus_provasolii	460	KT860658	664
	Pycnococcus_sp.	345	KT878666	670
Prasino-Clade-VII	Prasino-Clade-VII-A-1_sp.	712	KT860929	648
	Prasino-Clade-VII-A-1_sp.	713	KT860813	340
	Prasino-Clade-VII-A-1_sp.	719	KT860933	621
	Prasino-Clade-VII-A-1_sp.	997	KT860935	2170
	Prasino-Clade-VII-A-2_sp.	717	KT860691	441
	Prasino-Clade-VII-A-2_sp.	138	KT860872	1610
	Prasino-Clade-VII-A-3_sp.	1043	KT860695	686
	Prasino-Clade-VII-A-3_sp.	1019	KT860708	637
	Prasino-Clade-VII-A-4_sp.	726	KT860692	620
	Prasino-Clade-VII-A-4_sp.	1124	KT860719	710
	Prasino-Clade-VII-A-4_sp.	722	KT860817	672
	Prasino-Clade-VII-A-5_sp.	700	KT860686	503
	Prasino-Clade-VII-A-5_sp.	19	KT860855	1606
	Prasino-Clade-VII-A-5_sp.	227	KT860875	1609
	Prasino-Clade-VII-B-2_sp.	696	KT860688	690
	Prasino-Clade-VII-B-2_sp.	2337	KT860793	524
Prasinococcales	*Prasinococcus_capsulatus*	474	KT860675	431
	Prasinococcus_capsulatus	1962	KT860733	858
	Prasinococcus_capsulatus	2359	KT860770	697
	Prasinococcus_capsulatus	2687	KT860772	805

Class	Species	RCC	Accession	length
	Prasinococcus_capsulatus	2349	KT860783	670
	Prasinococcus_capsulatus	2356	KT860784	490
Prasinococcales	*Prasinococcus_capsulatus*	2357	KT860785	650
	Prasinococcus_capsulatus	2359	KT860786	819
	Prasinococcus_capsulatus	896	KT860907	1390
	Prasinococcus_capsulatus	2694	KT860924	1252
	Prasinococcus_sp.	473	KT860667	453
	Prasinococcus_sp.	520	KT860671	590
	Prasinoderma_coloniale	708	KT860685	640
	Prasinoderma_coloniale	711	KT860687	523
	Prasinoderma_coloniale	710	KT860690	720
	Prasinoderma_coloniale	959	KT860706	665
	Prasinoderma_coloniale	961	KT860707	716
	Prasinoderma_coloniale	1957	KT860732	654
	Prasinoderma_coloniale	1967	KT860734	787
	Prasinoderma_coloniale	960	KT860735	857
	Prasinoderma_coloniale	714	KT860814	689
	Prasinoderma_coloniale	3066	KT860834	1107
	Prasinoderma_coloniale	3068	KT860835	1115
	Prasinoderma_coloniale	672	KT860885	794
	Prasinoderma_coloniale	673	KT860886	1631
	Prasinoderma_singularis	928	KT860930	470
	Prasinoderma_sp.	2695	KT860796	630
Pyramimonadales	*Pterosperma_sp.*	2503	KT860767	850
	Pyramimonas_parkeae	619	KT860881	
	Pyramimonas_sp.	669	KT860677	708
	Pyramimonas_sp.	2047	KT860756	807
	Pyramimonas_sp.	2048	KT860757	668
	Pyramimonas_sp.	2296	KT860798	690
	Pyramimonas_sp.	2500	KT860922	797
	Pyramimonas_sp.	2501	KT860923	1622
Trebouxiophyceae	*Chlorella_sp.*	288	KT860661	422
	Chlorella_sp.	537	KT860672	1667
	Chlorella_sp.	533	KT860679	896
	Chlorella_sp.	664	KT860884	904
	Chlorella_sp.	347	KT860631	942
	Chlorella_sp.	396	KT860642	551
	Chlorella_stigmatophora	661	KT860883	952
	Coccomyxa_sp.	891	KT860906	1651
	Coccomyxa_sp.	903	KT860910	1653
	Phyllosiphon_sp.	2979	KT860839	820
	Picochlorum_sp.	475	KT860662	421

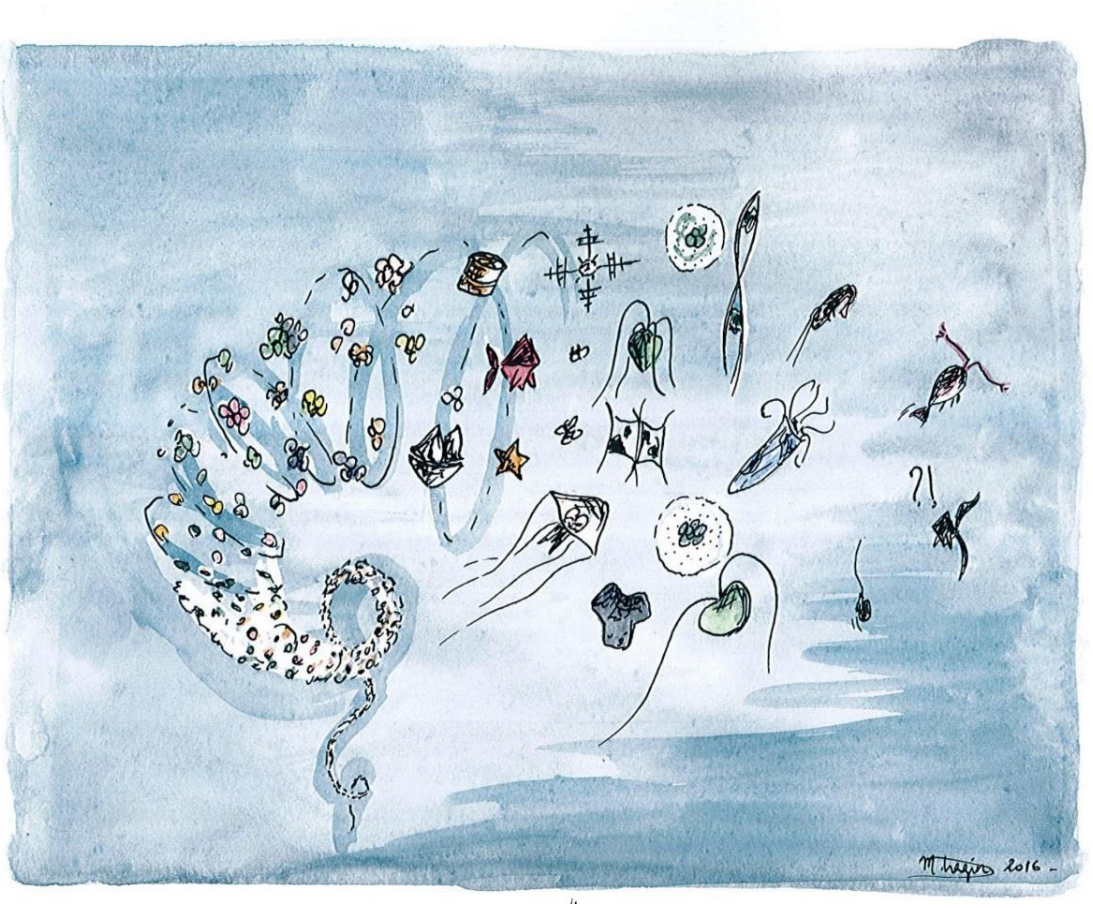
Class	Species	RCC	Accession	length
Trebouxiophyceae	*Picochlorum_sp.*	633	KT860663	512
	Picochlorum_sp.	632	KT860665	503
	Picochlorum_sp.	637	KT860666	549
	Picochlorum_sp.	484	KT860668	727
	Picochlorum_sp.	246	KT860805	660
	Picochlorum_sp.	485	KT860806	743
	Picochlorum_sp.	846	KT860820	641
	Picochlorum_sp.	3065	KT860832	1101
	Picochlorum_sp.	3067	KT860833	1104
	Picochlorum_sp.	3071	KT860836	1105
	Picochlorum_sp.	944	KT860842	1641
	Picochlorum_sp.	3070	KT860847	642
	Picochlorum_sp.	9	KT860852	1484
	Picochlorum_sp.	13	KT860853	1633
	Picochlorum_sp.	14	KT860854	1629
	Picochlorum_sp.	99	KT860856	1633
	Picochlorum_sp.	140	KT860873	1284
	Picochlorum_sp.	142	KT860874	1576
	Picochlorum_sp.	250	KT860877	1346
	Picochlorum_sp.	683	KT860889	1631
	Picochlorum_sp.	684	KT860890	1635
	Picochlorum_sp.	689	KT860892	1631
	Picochlorum_sp.	748	KT860896	1630
	Picochlorum_sp.	897	KT860908	1114
	Picochlorum_sp.	2935	KT860926	1541
	Picochlorum_sp.	236	KT860628	509
	Picochlorum_sp.	289	KT860649	682
	Picochlorum_sp.	237	KT860655	667
	Stichococcus_sp.	1055	KT860696	670
	Trebouxiophyceae_XXX_sp.	2942	KT860826	712
Ulvophyceae	*Desmochloris_leptochaete*	2960	KT860928	1574
	Ulvophyceae_XXX_sp.	2980	KT860840	673
	Ulvophyceae_XXX_sp.	2981	KT860841	781
	Ulvophyceae_XXX_sp.	2945	KT860927	1613

Table S5: Oceanic regions used to regroup sequences presented in Figure S2.

Oceanic zone	Longitude range	Latitude range
Northern high latitude		70°N-90°N
North Sea and English Channel	13°W-26.6°E	49°N-70°N
Mediterranean and Red Sea	13°W-40°E	13°N-49°N
North Eastern Atlantic Ocean	38°E-100°E	10°N-50°N
South Atlantic Ocean	47°W-47°E	56°S-20°S
South America Eastern coast	47°W-67°W	56°S-34°S
Chile coast	67°W-90°W	56°S-13°S
South Eastern Pacific Ocean	90°W-180°W	56°S-10°N
North Eastern Pacific Ocean	100°W-150°W	10°N-65°N
North Western Pacific Ocean	130°E-180°E	13.3°N-90°N
South Western Pacific Ocean	130°E-180°E	56°S-13.3°N
South Western Asian archipelago	94°E-130°E	7°S-30°N
North Indian Ocean	49°E-94°E	7°S-30°N
South Indian Ocean	47°E-130°E	56°S-7°S
Southern Ocean		90°S-53°S

Chapter 2

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



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[1], Adriana Zingone[2], 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

For ***Environmental Microbiology*** accepted September 2017

DOI: 10.1111/1462-2920.13952

Keywords

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

Acknowledgments

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 paper.

* Corresponding author: vaulot@sb-roscoff.fr

Abstract

High Throughput Sequencing (HTS) approaches are getting more and more used to investigate the diversity and community structure of microbial eukaryotes in marine waters. Partial 18S rRNA gene regions (V4 and V9 regions) are commonly used as genetic barcodes. Selecting between the V4 and V9 regions remains a matter of debate. Here, we compared the composition of communities estimated using these two genetic markers at 27 sites sampled during Ocean Sampling Day 2014, with a focus on photosynthetic groups and, more specifically, Chlorophyta. Globally, the V4 and V9 regions of the 18S rRNA gene provided similar images of alpha diversity and ecological patterns. However, the V9 dataset provided 20% more OTUs built at 97% identity than the V4 dataset and 39% and 56% of the genera were found only in one dataset, respectively V4 and V9. For photosynthetic groups, the V4 and V9 regions also 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) unveiling the lack of V9 sequences for this group and the importance of the reference database in the metabarcoding process. We conclude that, to address specific 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.

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, 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*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 remains a matter of debate. Hu *et al.* (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 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 impacts: using the 454 technology, Behnke *et al.* (2011) showed that the sequencing error rate was taxon dependent, but V4 error rates were higher than for V9. Analysis of mock communities have 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 vs. 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 *et al.* (2010) found that the V9 region recovered more diversity at higher taxonomic levels than the V4 region: for example 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 *et al.* (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 stations) 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 constructed (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).

Material and Methods

Water samples were collected from 0-2 meter 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, USA) 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: 2x250 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" fasta files (https://owncloud.mpi-bremen.de/index.php/s/RDB4Jo0PAayg3qx?path=/2014/silva-ngs/18s/lifewatch/).

All subsequent sequence analyses (Supplementary Data and Fig. S1, Table 2) were done with Mothur v 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 v7.1.7 (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 v 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 (Supplementary Data). 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 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 PR² database (Guillou *et al.*, 2013) for which the Chlorophyta sequences had been checked against the latest taxonomy (Tragin *et al.*, 2016). The PR² database considers 8 taxonomic levels (from Kingdom to Species). OTUs are considered as assigned when their last 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 (for example 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 (Supplementary Data) 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* varies between 0 and 1 and is null when *S* is equal to 1. *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 stats R package. 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 2-dimensional space with the iterative ordination method Nonparametric Multi-Dimensional Scaling (NMDS) plot using the *metaMDS* function of Vegan. The

axis scale is arbitrary defined by the calculation, but groups of dots and their relative position inform us on the differences between communities at the station. The more closely related the points are, the more they are similar. 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 were projected onto the NMDS plots using the *envfit* function from the Vegan with the p.max option set as 0.95.

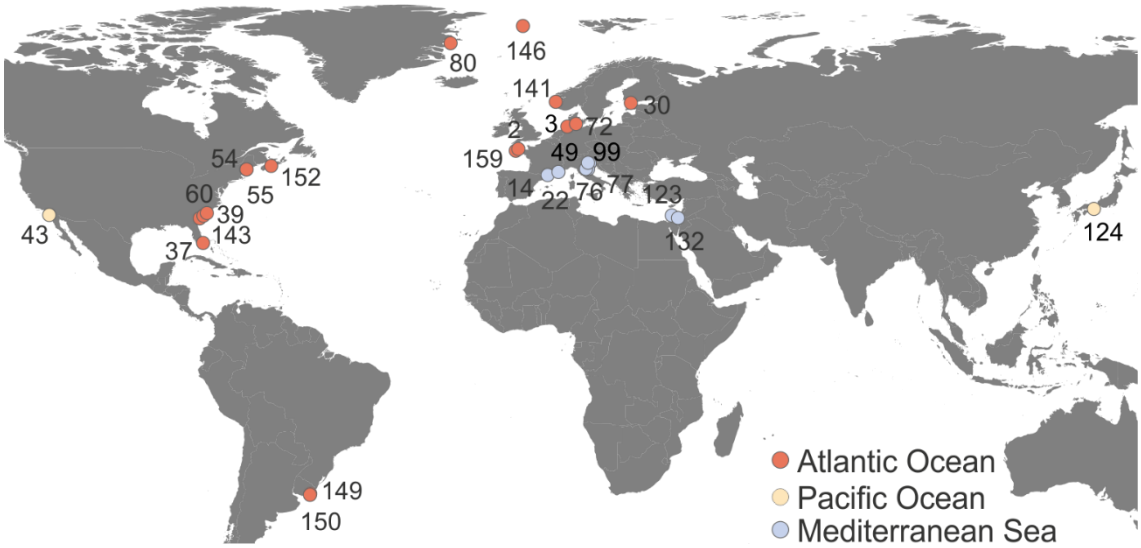


Fig.1: Map of the 27 OSD stations sampled 2014 for which both V4 and V9 sequences were available.

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	Atlantic	English Channel	343 626	59,0	28,1	387 351	52,3	25,5
3	Helgoland	Atlantic	North Sea	315 340	64,3	27,7	257 957	78,6	34,3
14	Banyuls	Med. Sea	Western Basin	311 053	65,2	18,2	406 871	49,8	11,2
22	Marseille - Solemio SOMLIT	Med. Sea	Western Basin	302 687	67,0	7,6	353 503	57,3	7,8
30	Tvärminne	Atlantic	Gulf of Finland	296 892	68,3	8,8	346 294	58,5	4,3
37	Port Everglades	Atlantic	USA coast	338 053	60,0	33,6	361 524	56,1	27,7
39	Charleston Harbor	Atlantic	USA coast	332 841	60,9	73,1	296 868	68,3	49,0
43	SIO Pier	Pacific	USA coast	320 295	63,3	10,9	388 996	52,1	21,0
49	Vida	Med. Sea	Adriatic Sea	*202 710*	*100,0*	14,4	302 436	67,0	14,5
54	Maine Booth Bay	Atlantic	USA coast	290 311	69,8	14,5	365 441	55,5	36,9
55	Maine Damariscotta River	Atlantic	USA coast	237 919	85,2	30,7	276 076	73,4	43,3
60	South Carolina 2 - North Inlet	Atlantic	USA coast	268 351	75,5	50,3	353 390	57,4	33,9
72	Boknis Eck	Atlantic	Kattegat	356 529	56,9	26,5	475 461	42,6	23,3
76	Foglia	Med. Sea	Adriatic Sea	242 825	83,5	11,5	386 655	52,4	13,2
77	Metauro	Med. Sea	Adriatic Sea	303 448	66,8	26,5	377 917	53,6	26,4
80	Young Sound	Atlantic	Greenland Sea	349 267	58,0	17,4	436 165	46,5	23,0
99	C1	Med. Sea	Adriatic Sea	339 739	59,7	14,2	449 242	45,1	12,7
123	Shikmona	Med. Sea	Eastern Basin	286 203	70,8	8,9	416 420	48,7	8,3
124	Osaka Bay	Pacific	Japan Sea	237 367	85,4	34,8	478 261	42,4	31,0
132	Sdot YAM	Med. Sea	Eastern Basin	285 592	71,0	26,4	399 001	50,8	17,7
141	Raunefjorden	Atlantic	Coast of Norway	308 267	65,8	0,8	402 413	50,4	1,6
143	Skidaway Institute of Oceanography	Atlantic	USA coast	328 039	61,8	81,4	410 937	49,3	65,3
146	Fram Strait	Atlantic	Greenland Sea	369 221	54,9	44,4	447 907	45,3	41,7
149	Laguna Rocha Norte	Atlantic	Coast of Uruguay	324 063	62,6	52,2	323 981	62,6	44,0
150	Laguna Rocha Sur	Atlantic	Coast of Uruguay	338 373	59,9	50,8	367 936	55,1	44,9
152	Compass Buoy Station	Atlantic	Baedford Basin	327 454	61,9	9,8	407 377	49,8	17,0
159	Brest - SOMLIT	Atlantic	Celtic Sea	327 901	61,8	30,7	443 747	45,7	22,9

Results

Global protist community

Twenty-seven stations were selected for which both V4 and V9 metabarcodes were obtained. 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 for V4 than for V9 (Table 2). About forty times more chimeras were found for V4 than for V9 (about 7,500 against 170). Following taxonomic assignment, all eukaryotic groups were retained, not just protists.

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

Rarefaction curves computed for the global datasets as well as for each station (Fig. S2A and Fig. S3) reached saturation, suggesting that the sequencing effort was sufficient. Global maximum richness varied between the datasets: V9 reached a total of 16,383 OTUs (4,311 distinct assignations) against 13,169 OTUs (3,412 distinct assignations) for V4. The two datasets yielded similar rank abundance curves (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 3,000 with respective averages of 1,200 and 1,600 for V4 and V9 (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 of the Simpson's index (0.6 to 0.9), it 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, 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 3,669, among which 39% (for V4) and 56% (for V9) were recovered only in one dataset. On average, 4 OTUs were assigned to the same genus and the maximum number of OTUs per genus reached 128 for V4 and 187 for V9. 98 % of genera found only in one dataset were represented by less than 10 OTUs.

Non-parametric multidimensional scaling analysis (NMDS, Fig. S4A-B) was used to visualize the V4 and V9 communities based on OTUs (final stress values were respectively 0.187 and 0.195). For both V4 and V9, stations grouped together in a similar way and each group of stations was geographically coherent. Stations from the Mediterranean Sea (OSD14, 22, 49, 76, 77, 99, 123, 132), the North Atlantic (OSD2, 3, 54, 55, 152, 159) and a South Atlantic lagoon (OSD149, 150) grouped together, respectively. Other stations clustering together included OSD30, 72, 80, 141 and 146 located in the northern high latitudes, and OSD39, 60 and 143 from the subtropical Atlantic coast of the United States (Fig. S4and Fig.1). Northern high latitude stations did not form a tight cluster suggesting that either communities were very diverse or these stations were not well represented on the two main NMDS axes. OSD37 (South Florida) stood apart from the other Atlantic Ocean stations. 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 (Fig. S4). Only the effect of silicates was statistically significant.

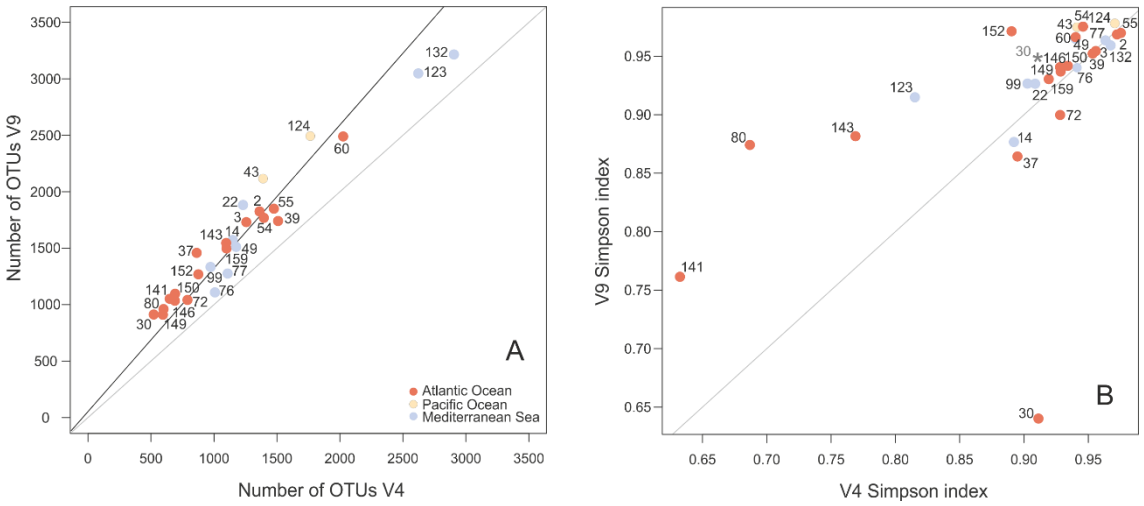


Fig.2: A - "Species richness": number of OTUs per stations for V4 versus V9. The grey line corresponds to y=x, and the black line corresponds to the regression y=1.27x+53 (R^2 = 0.996). B – Simpson's diversity index per stations for V4 vs. V9. Grey star corresponds to the OSD30 Simpson's index without the metazoanV9 OTU.

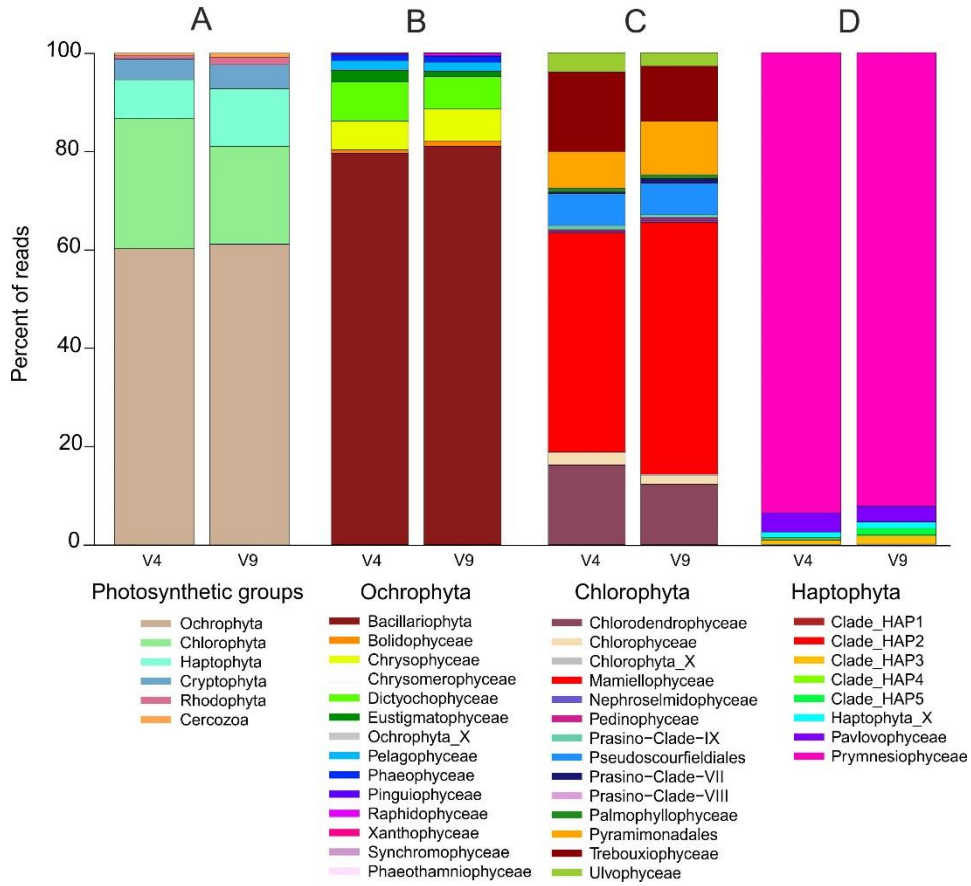


Fig.3: A. Contribution of divisions to photosynthetic metabarcodes (Dinophyceae were excluded) for V4 and V9. B-D. Distribution of reads among classes for the three major photosynthetic divisions for V4 and V9: B. Ochrophyta, C. Chlorophyta D. Haptophyta.

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). 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% vs. 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, 91 and 573, 100, 241 in V4 versus V9). For these 3 photosynthetic groups, average pairwise identity between the OTUs reference sequences was higher for V4 than V9 region (Table S.1), meaning that V9 had higher genetic variability for these groups, and therefore was more 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 (Fig. S5A). Chlorophyta contribution varied between 5% (OSD76, 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 (Fig. S5B). Haptophyta contribution varied across stations from a few percent up to 40% (OSD22, 49, 146) and was larger for V9 than for V4 (Table 3) except for OSD3 (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 (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, 146) to 90% (OSD159, 60). Diatom 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 (Fig. S6B). Dictyochophyceae contribution was slightly higher with V4 at these five stations. Chrysophyceae-Synurophyceae relative contribution was below 10% except for OSD76 (25%, Fig. S6C and Table 3) and V4 and V9 contribution were similar except at OSD30, where V9 was higher and OSD49 and 76 where V4 was higher (Fig.

S6C). Pelagophyceae relative contribution was below 10% at individual stations but V4 and V9 were similar (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 the two environmental clades HAP3 and HAP4 (Edvardsen *et al.*, 2016) were also recovered (Fig.3D). For photosynthetic groups, the percentage 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 (Fig. S7). Within Ochrophyta, more genera were found using V9 in 5 out of 8 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 (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 (Fig. S7).

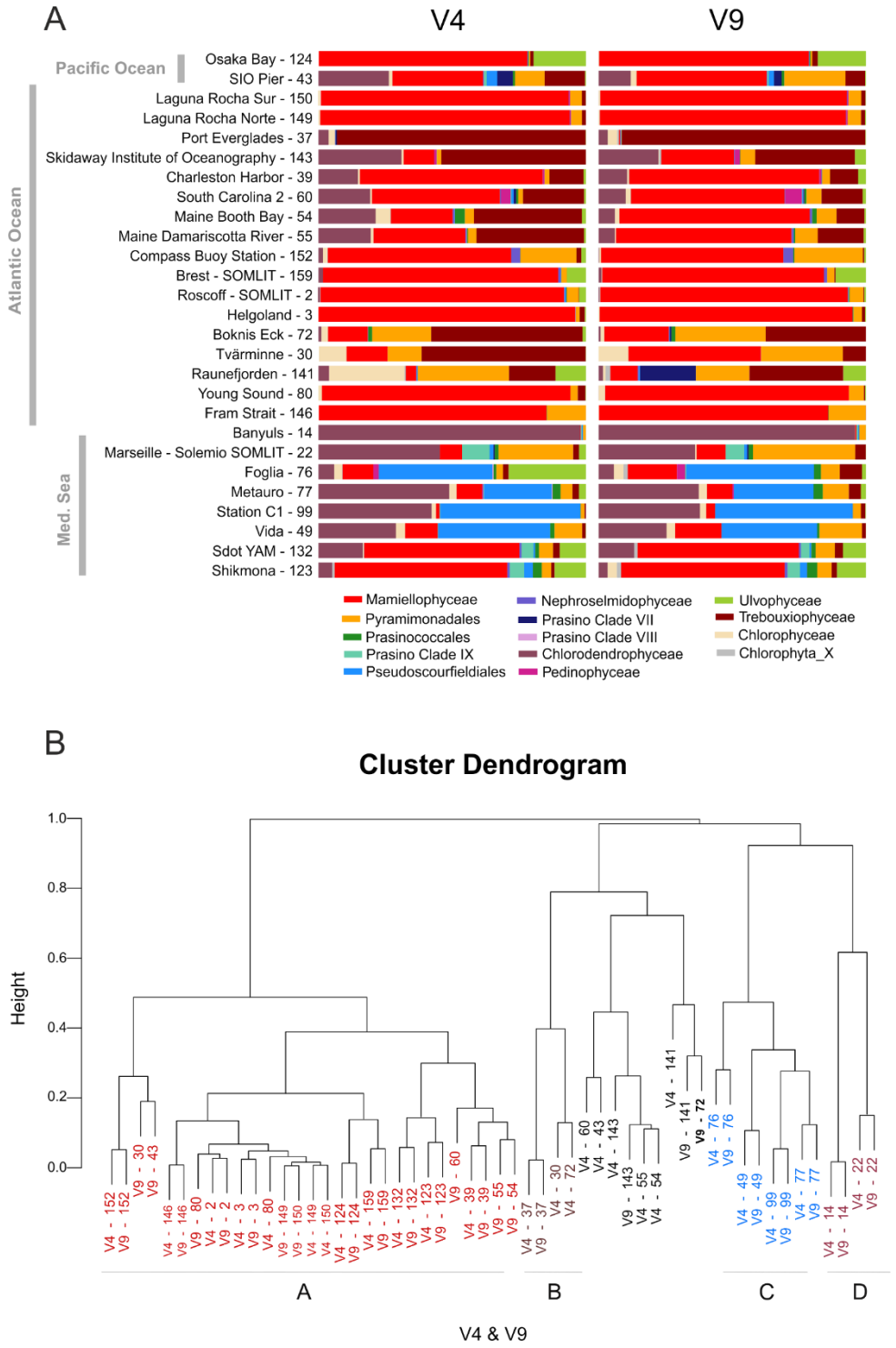


Fig.4: A. Comparison of Chlorophyta read distribution (assigned at the class level) for 27 OSD2014 stations. B. Comparison of Chlorophyta communities at the class level based hierarchical clustering for V9 and V4. The dissimilarity matrix was computed using Bray Curtis distance. The stations were labelled 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).

Chlorophyta classes

The relative contributions of the 6 major Chlorophyta groups (Mamiellophyceae, Trebouxiophyceae, Chlorodendrophyceae, Pyramimonadales, Ulvophyceae and Pseudoscourfieldiales) in V4 and V9 were similar at most stations (Fig.4A and 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, 141, all located in North Atlantic coastal waters presented differences in Chlorophyta classes contribution recovered with V4 and V9 (Fig.4A and Fig. S9A). For the first three, the Mamiellophyceae contribution in V9 was partially replaced in V4 by classes of the "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 (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 versus 47 reads, respectively).

In general, Chlorophyta OTUs were well assigned by the Wang approach implemented in the Mothur software (Wang et al., 2007) compared to the results of BLAST (Supplementary data 6 and 7). However, some V9 reads initially assigned as Chlorophyta by the Wang approach hit bacterial sequences and were not considered any further. Some V9 Chlorophyta OTUs were also linked to several different Chlorophyta genera with 100% identity (mostly in the Ulvophyceae, Trebouxiophyceae and Chlorophyceae clade, UTC clade) suggesting that the V9 region might not have the appropriate resolution to investigate UTC clade diversity. A number of genera within Ulvophyceae, Trebouxiophyceae, Chlorophyceae (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 (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).

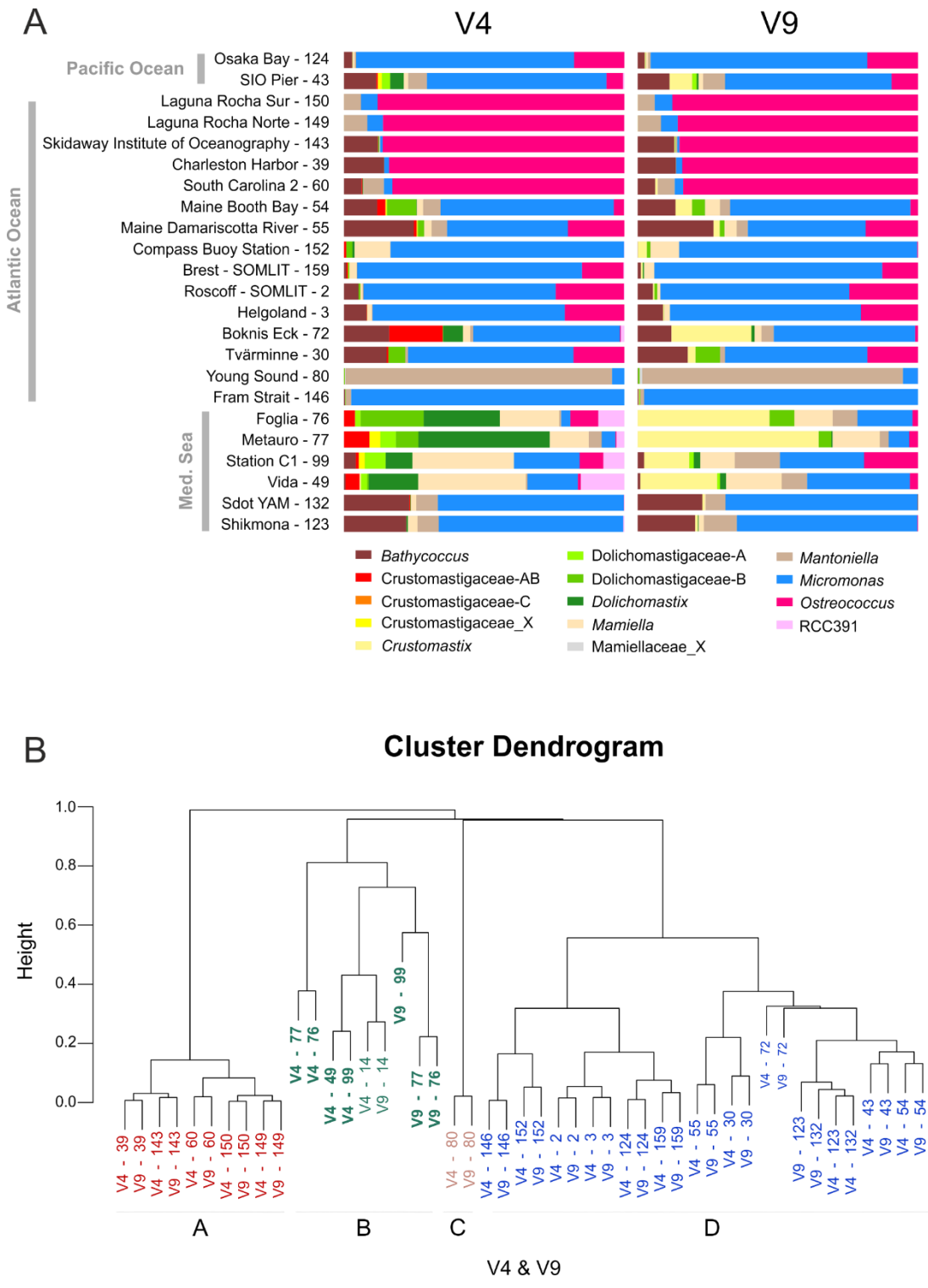


Fig.5: A. Comparison of Mamiellophyceae read distribution (assigned at the genus level) for 23 OSD2014 stations. Stations, where the number of reads assigned to Mamiellophyceae was lower than 100 were removed (OSD14, 22, 37 and 141). B. Comparison of Mamiellophyceae communities at the genus level by 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).

Mamiellophyceae genera

Mamiellophyceae dominated at most OSD stations and 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 Supplementary data), 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* (Fig. S10) was statistically similar in the two datasets (Table 3). Stations located in the Adriatic Sea (OSD49, 76, 77, 99) showed a different pattern in the heatmap (Fig. S9B) because V9 failed to discriminate the Dolichomastigales clades at the genus level. V9 recorded only *Crustomastix* contribution while V4 found 4 to 6 different clades of Crustomastigaceae and Dolichomastigaceae (Fig.5A). The relative contribution of the four Mamiellophyceae genera *Micromonas*, *Mamiella*, *Ostreococcus* and *Bathycoccus* was similar in V4 and V9 except at some stations (OSD22, 49, 132, 123) for *Mamiella*. Bray-Curtis distances always clustered together V4 and V9 (Fig.5B). Four groups of stations were observed depending on the Mamiellophyceae genus dominant at the station: *Micromonas*, *Ostreococcus*, Dolichomastigales or *Mantoniella* (Fig.5B).

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 Fig. S5, Fig. S6, Fig. S8 and Fig. S10). 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.

	V4				V9				
Parameter	max.	min.	mean	SD	max.	min.	mean	SD	P value
OTU number	2906	522	1216	578.9	3216	911	1620	617.65	*0.00000592*
Simpson Index	0.63	0.99	0.91	0.08	0.63	0.99	0.92	0.073	0.049
Ochrophyta (photo. %)	91.11	19.31	60.26	18.12	87.58	20.43	64.13	19.8	**0.4**
Chlorophyta (photo. %)	69.59	3.94	26.44	17.21	55.22	2.31	19.87	15.48	0.0000633
Haptophyta (photo. %)	30.2	0.11	7.9	8.64	37.67	0.28	11.73	10.53	0.000000819
Cryptophyta (photo. %)	13.61	0.03	4.21	3.58	19.69	0.28	4.91	4.22	0.008
Bacillariophyta (photo. %)	89.53	8.59	49.12	21.56	86.17	8.06	51.77	23.36	0.095
Dictyochophyceae (photo. %)	35.24	0.006	4.4	7.52	30.12	0.002	3.34	6.09	0.0029
Chryso-Synurophyceae (photo. %)	24.28	0.07	3.43	4.83	16.66	0.12	3.16	3.67	**0.5**
Pelagophyceae (photo. %)	7.02	0	0.74	1.56	7.99	0	0.74	1.63	***0.56***
Mamiellophyceae (photo. %)	49.58	0.04	12.09	14.19	45.69	0.19	10.93	12.94	**0.25**
Trebouxiophyceae (photo. %)	53.49	0	4.9	11.24	39.3	0.01	2.33	7.48	0.023
Chlorodendrophyceae (photo. %)	68.09	0	4.89	13.01	53	0	3.07	10.11	*0.000025*
Pyramimonadales (photo. %)	6.36	0.006	1.66	1.98	7.61	0.04	1.7	1.92	0.0096
Ulvophyceae (photo. %)	4.85	0	0.64	1.16	3.91	0	0.42	0.86	0.046
Pseudoscourfieldiales (photo. %)	17.11	0	1.15	3.46	9.02	0	0.61	1.83	*0.001*
Micromonas (Chloro. %)	34.94	0	4.8	7.61	31.98	0.03	4.27	6.8	**0.54**
Mamiella (Chloro. %)	2.45	0	0.22	0.47	1.77	0	0.15	0.34	***0.26***
Ostreococcus (Chloro. %)	35.79	0	4	9.57	40.03	0	4.36	10.64	***0.47***
Bathycoccus (Chloro. %)	4.04	0	0.68	1.14	3.69	0	0.63	1.01	***0.64***

Discussion

The OSD V4 and V9 datasets

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. Also 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 2,200 full 18S sequences of protist (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 size (V4: 450 bp - 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).

In the OSD dataset, photosynthetic groups (Dinophyceae excluded) varied widely from 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).

Comparison of the photosynthetic communities assessed by the V4 vs. V9 regions

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 such as the Naples times series results (Piredda *et al.*, 2017). Piredda *et al.* (2017) also found 20% more OTUs built at 97% identity for V9 than for V4. This could be linked to the size difference between V4 and V9 as discussed above. Interestingly, these authors showed that the number of OTUs built at 95% identity was similar for 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 were 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 to 40 more than for OSD) which increases the occurrence of the rare OTUs.

Mamiellophyceae dominated 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 *et al.* (2009) found *Micromonas* to be the most prevalent genus in the world ocean coastal waters and at more local scale *Micromonas* dominates coastal picoplankton in the Western English Channel (Not *et al.*, 2004). Rii *et al.* (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 *et al.* (2008) and Monier *et al.* (2016) made similar observations in oligotrophic Mediterranean surface waters and in the Tara *Oceans* survey, respectively.

At six stations (OSD30, 80, 123, 141, 143, 152) the same species richness (OTU number) was observed but Simpson index was different between V4 and V9 (Fig.2 A 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).

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 (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 (Fig.5A). 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.

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, 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 (e.g. A1 and A2) defined to date (Lopes dos Santos *et al.*, 2016), 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 have the same similarity average and should be 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 (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 V4, because many sequences in public databases are short

and do not extend to the end of the 18S rRNA gene. For example, Viprey *et al.* (2008) discovered one novel prasinophyte group (clades VIII) by using Chlorophyta specific primers that only amplified a short (around 910 base pairs) sequences 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) allows to investigate phylogenetic diversity without assignation against a reference database. Phylogenetic approach however would be impacted by the lack of reference sequences and could be completed by statistical testing of the consistency of phylogenetic signals (Kembel, 2009; Stegen *et al.*, 2012). Still overall, our analyses demonstrate that in most cases V4 and V9 provide similar images of the distribution of specific groups such as the Chlorophyta and therefore that global studies using either of these markers are comparable.

References

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W., and Huse, S.M. (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**: 340–9.

Berry, M.A., White, J.D., Davis, T.W., Jain, S., Johengen, T.H., Dick, G.J., et al. (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., Moorthi, S.D., et al. (2009) Defining DNA-based operational taxonomic units for microbial-eukaryote ecology. *Appl. Environ. Microbiol.* **75**: 5797–808.

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. (2006) Choosing the right molecular genetic markers for studying biodiversity: from molecular evolution to practical aspects. *Genetica* **127**: 101–20.

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., and Lovejoy, C. (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., Sogin, M., et al. (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**: 2194–200.

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., and Edvardsen, B. (2013) 454 Pyrosequencing to Describe Microbial Eukaryotic Community Composition, Diversity and Relative Abundance: A Test for Marine Haptophytes. *PLoS One* **8**:.

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.

Fosso, B., Santamaria, M., Manzari, C., Lionetti, C., Erchia, A.M.D., Gissi, C., et al. (2016) Characterization of the eukaryotic microbiome by 18S rRNA metabarcoding data analysis and assessment of the relative resolution of V4 and V9 regions. *Rapp. la Commision Int. pour la Mer Méditerranée* **41**: 259.

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.* 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**: 1647–9.

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., Lehmann, K., et al. (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. (2016) 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., and Blomster, J. (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**: 1868–80.

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**: 854–66.

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.

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**: 607–10.

Not, F., del Campo, J., Balagué, V., de Vargas, C., and Massana, R. (2009) New insights into the diversity of marine picoeukaryotes. *PLoS One* **4**:.

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**: 4064–72.

Pernice, M.C., Logares, R., Guillou, L., Massana, R., and Franz, M. (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**:.

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**: 7537–41.

Simmons, M.P., Sudek, S., Monier, A., Limardo, A.J., Jimenez, V., Perle, C.R., et al. (2016) Abundance and Biogeography of Picoprasinophyte Ecotypes and Other Phytoplankton in the Eastern North Pacific Ocean. *Appl. Environ. Microbiol.* **82**: 1693–705.

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**: 1653–64.

Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M.D.M., Breinner, H.-W., and Richards, T.A. (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., Mahe, 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.

List of Figures

List of Tables

List of Supplementary Figures

Fig. S1: A. Bioinformatics pipeline use to build and assigned OTUs from V4 and V9 datasets. Reference alignment was SILVA seed release 119 and the Chlorophyta curated PR² database (Tragin *et al.*, 2016) was used as taxonomic references. 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 order of size.

Fig. S3: Rarefaction curves per station A- V4; 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. Cryptophyceae.

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 5 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*.

List of Supplementary Tables

Table S1: Percentage of identity within OTUs reference sequences from photosynthetic groups

List of Supplementary Data

The data are deposited on Figshare 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

Supplementary Data 1. Mothur script for sequence analysis (on Figshare and in the next pages)

Supplementary Data 2. Fasta file of Chlorophyta OTUs for V4

Supplementary Data 3. Fasta file of Chlorophyta OTUs for V9

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

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

Supplementary Data 6. 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.

Supplementary Data 7. 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

Supplementary Figures

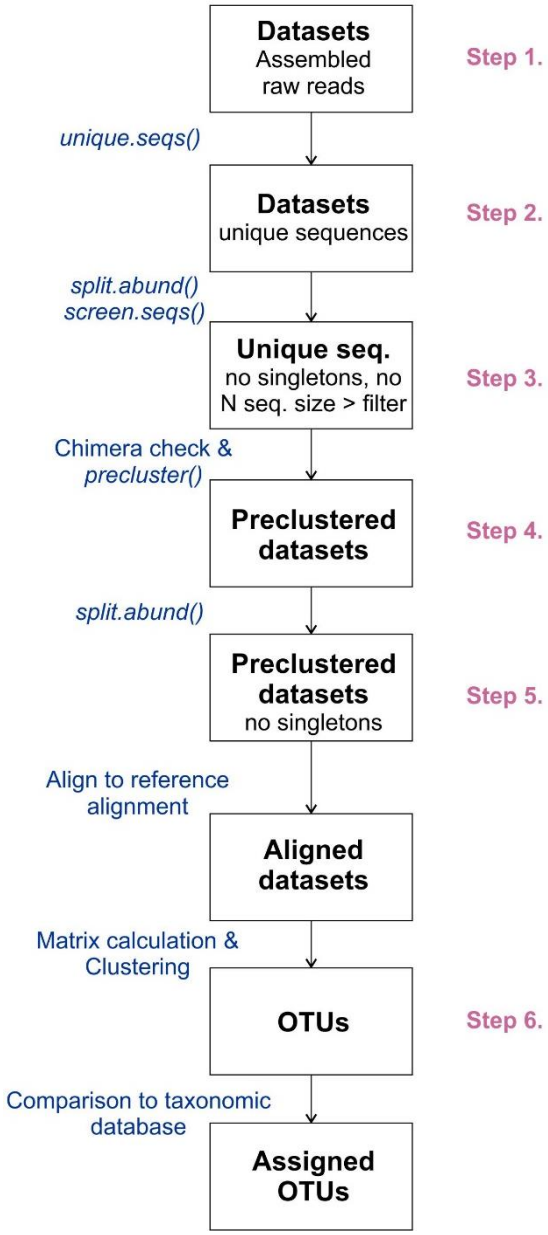


Fig. S1: A. Bioinformatics pipeline use to build and assigned OTUs from V4 and V9 datasets. Reference alignment was SILVA seed release 119 and the Chlorophyta curated PR² database *(Tragin et* al., 2016) was used as *taxonomic* references. The number of sequences at each step appears in Table 2.

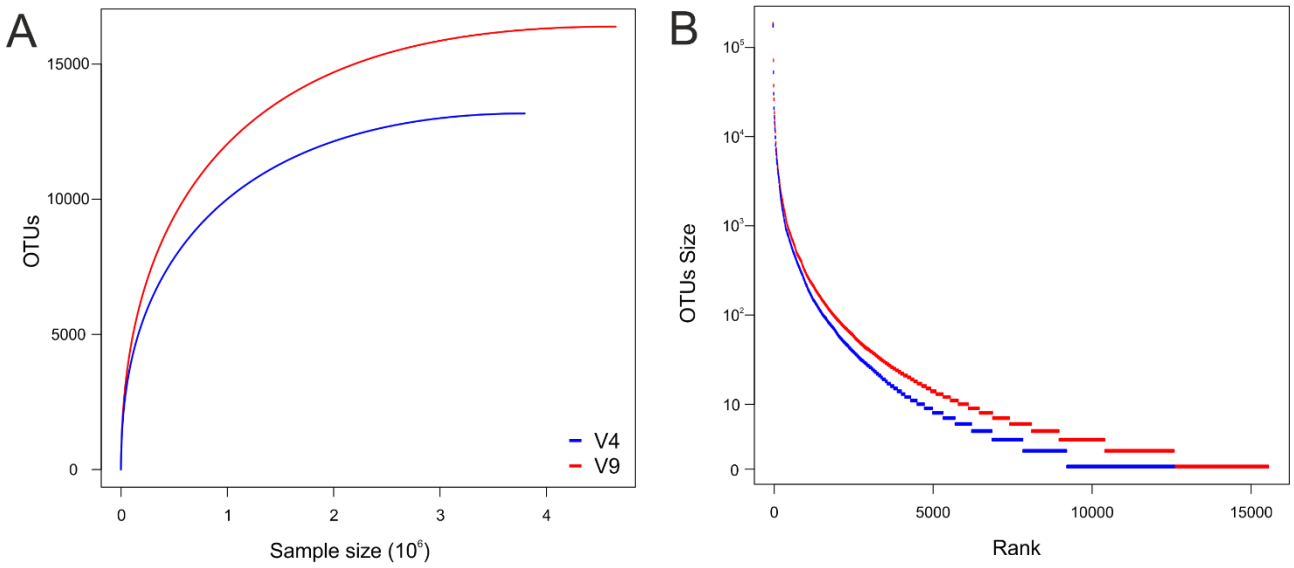


Fig. S2: A - Rarefaction curves; B - Rank abundance distribution. x-axis represents OTUs by decreasing order of size.

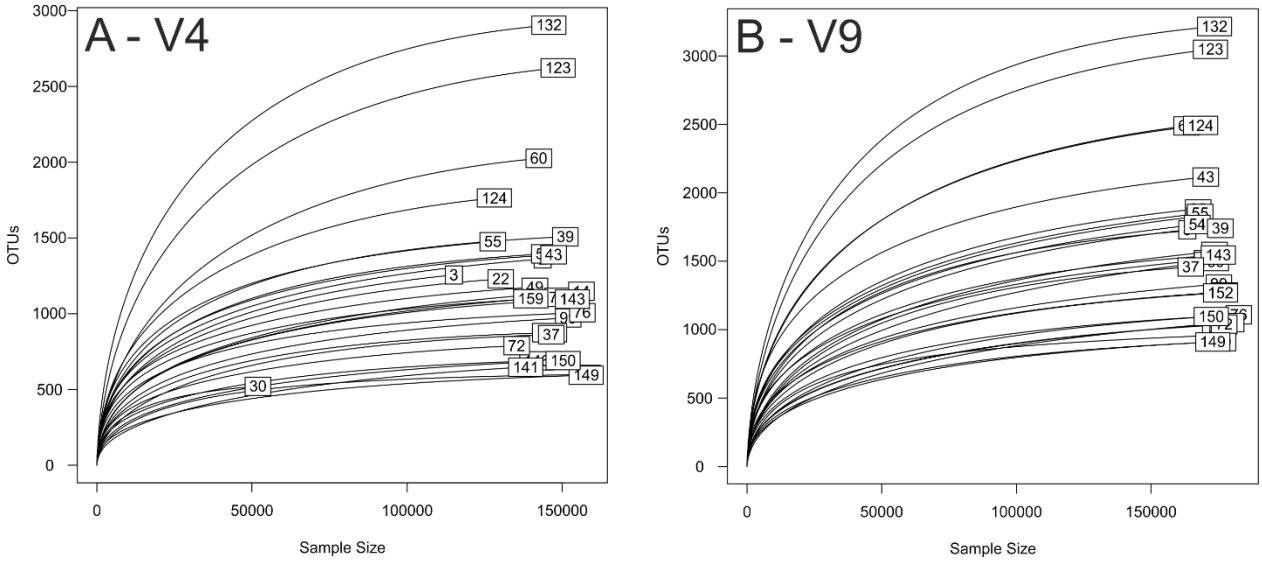


Fig. S3: Rarefaction curves per station A- V4; B - V9.

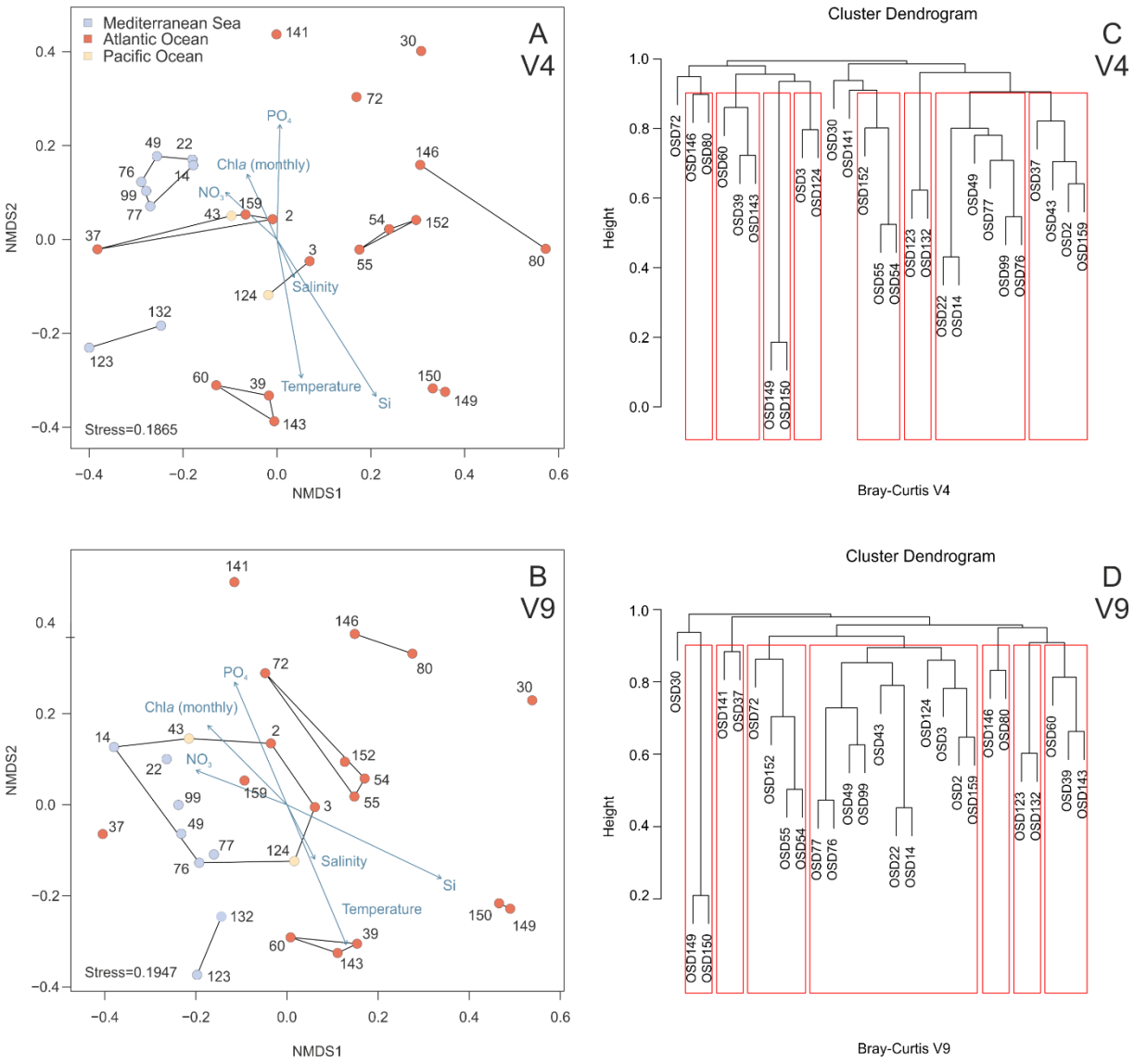


Fig. S4: A and B. Non-metric Multi-Dimensional Scaling (NMDS) representation of communities based on lowest taxonomic level (OTUs) for V4 (A) and V9 (B). The dissimilarity matrix was computed using Bray Curtis distance. C and D. Hierarchical cluster analysis based on the Bray Curtis matrix for V4 (C) 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).

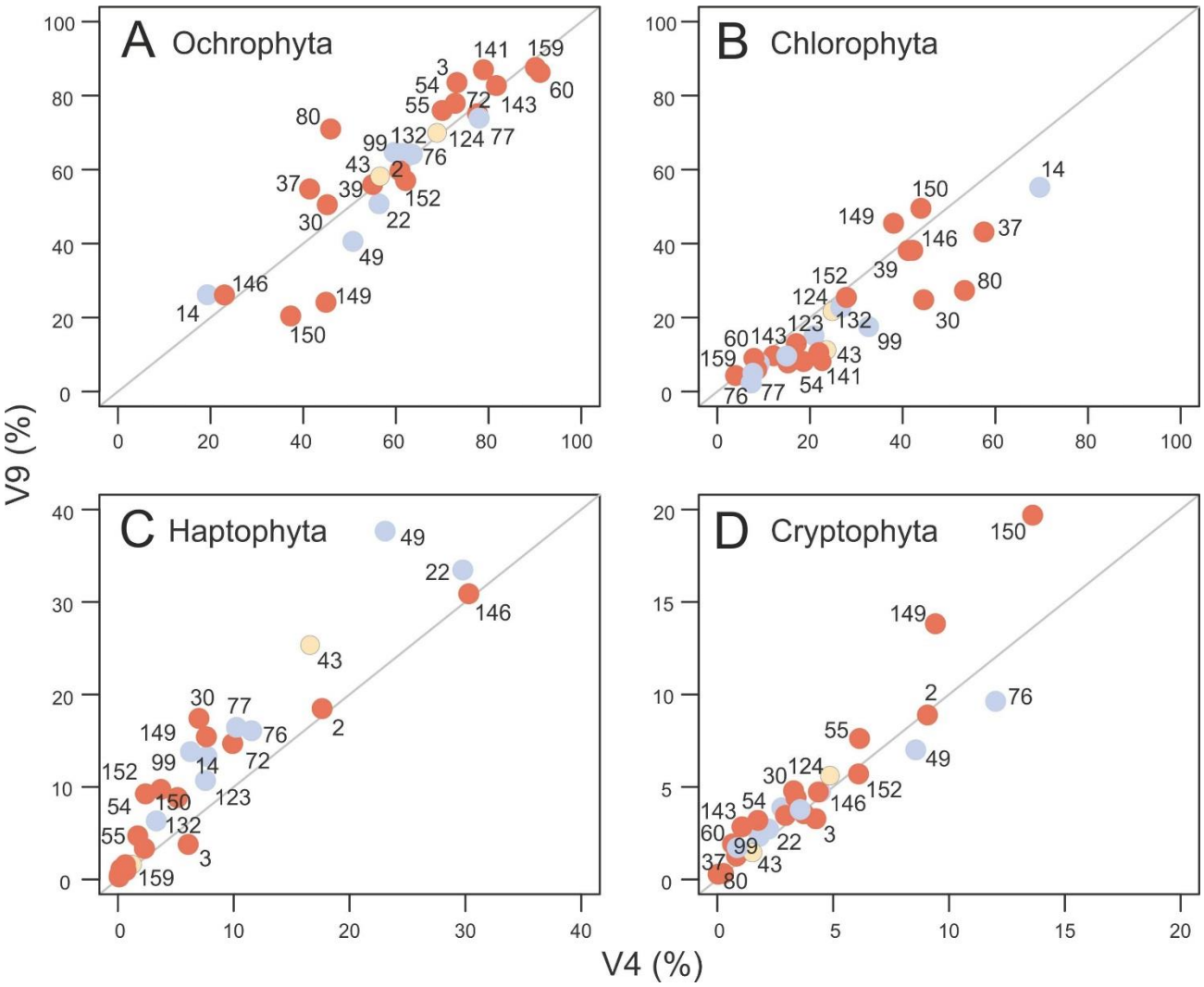


Fig. S5: Correlation between V4 and V9 relative contribution to photosynthetic metabarcodes in major photosynthetic phyla: A. Ochrophyta, B. Chlorophyta, C. Haptophyta, D. Cryptophyceae.

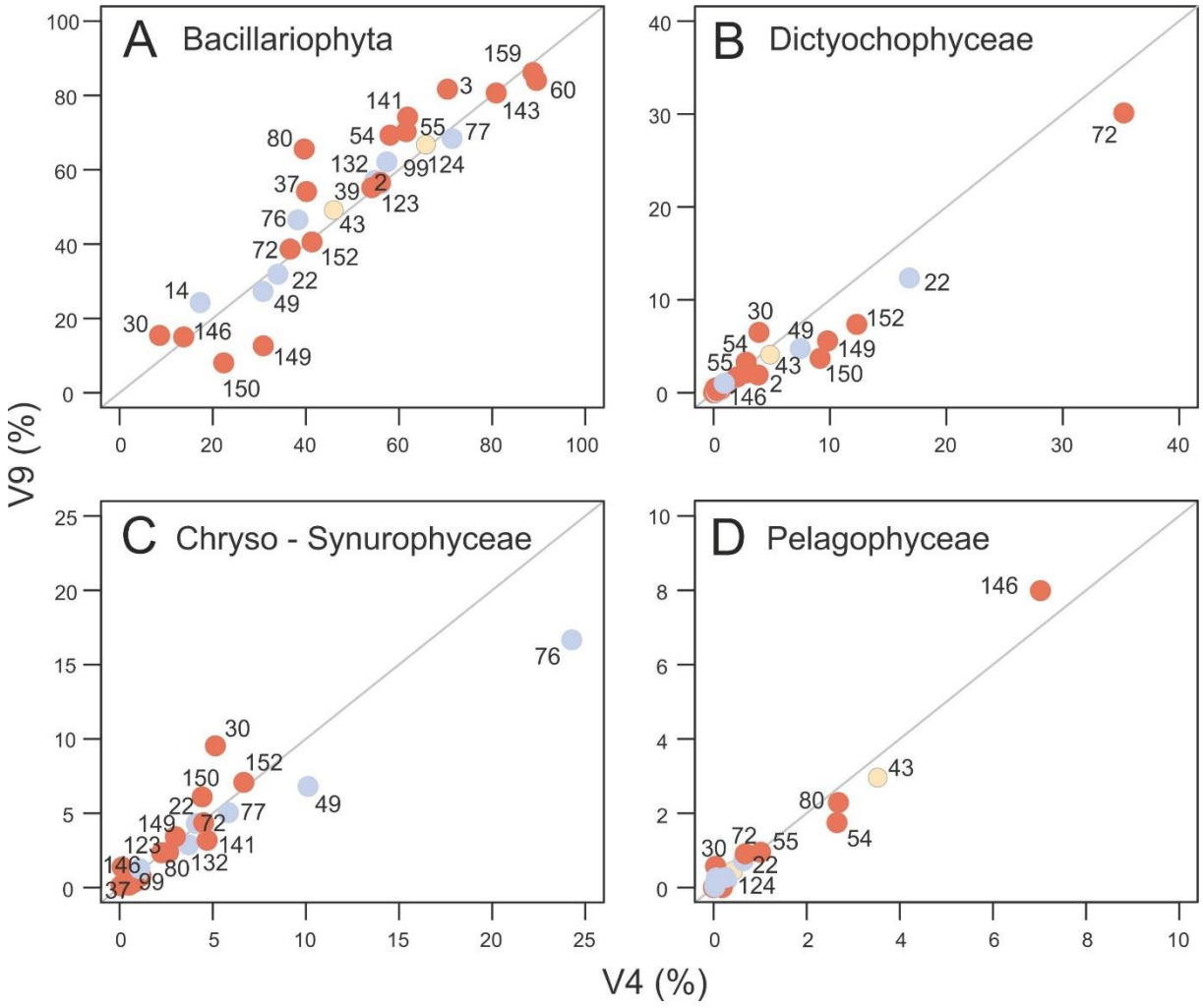


Fig. S6*:* Correlation between V4 and V9 relative contribution of the four major Ochrophyta Classes: A. Bacillariophyta, B. Dictyochophyceae, C. Chrysophyceae-Synurophyceae, D. Pelagophyceae.

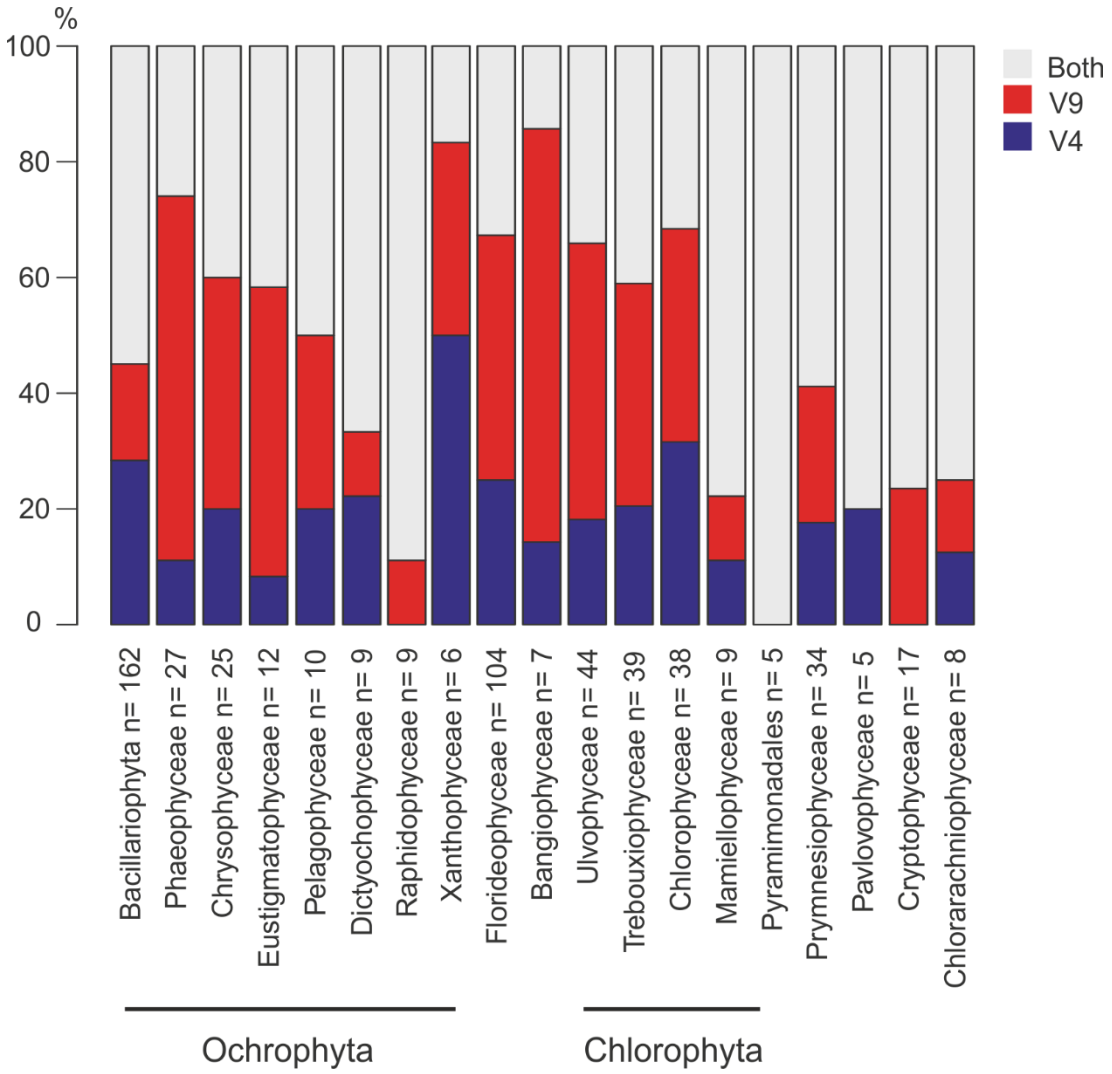


Fig. S7: Percentage of genera from photosynthetic groups found either only in V4 (blue), or only in V9 (red), or in both datasets (grey). Only taxonomically valid genera and only Classes with at least 5 genera were taken into account. Numbers below each group indicate the total number of genera recorded.

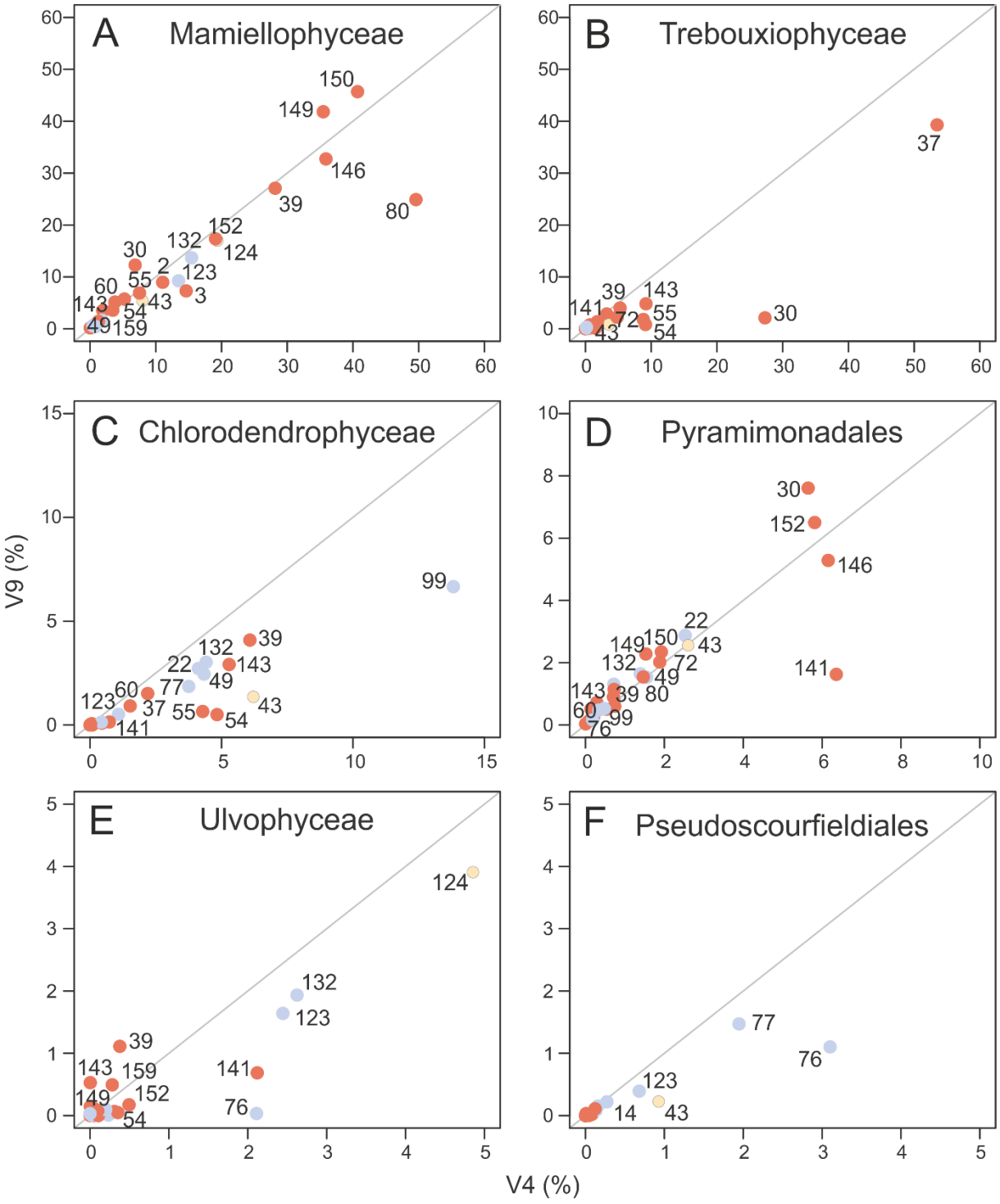


Fig. S8: Correlation between V4 and V9 relative contribution to photosynthetic metabarcodes for major Chlorophyta Classes: A. Mamiellophyceae, B. Trebouxiophyceae, C. Chlorodendrophyceae (OSD14 is not represented on the scatter plot with 65% and 60% for V4 and V9 respectively), D. Pyramimonadales, E.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

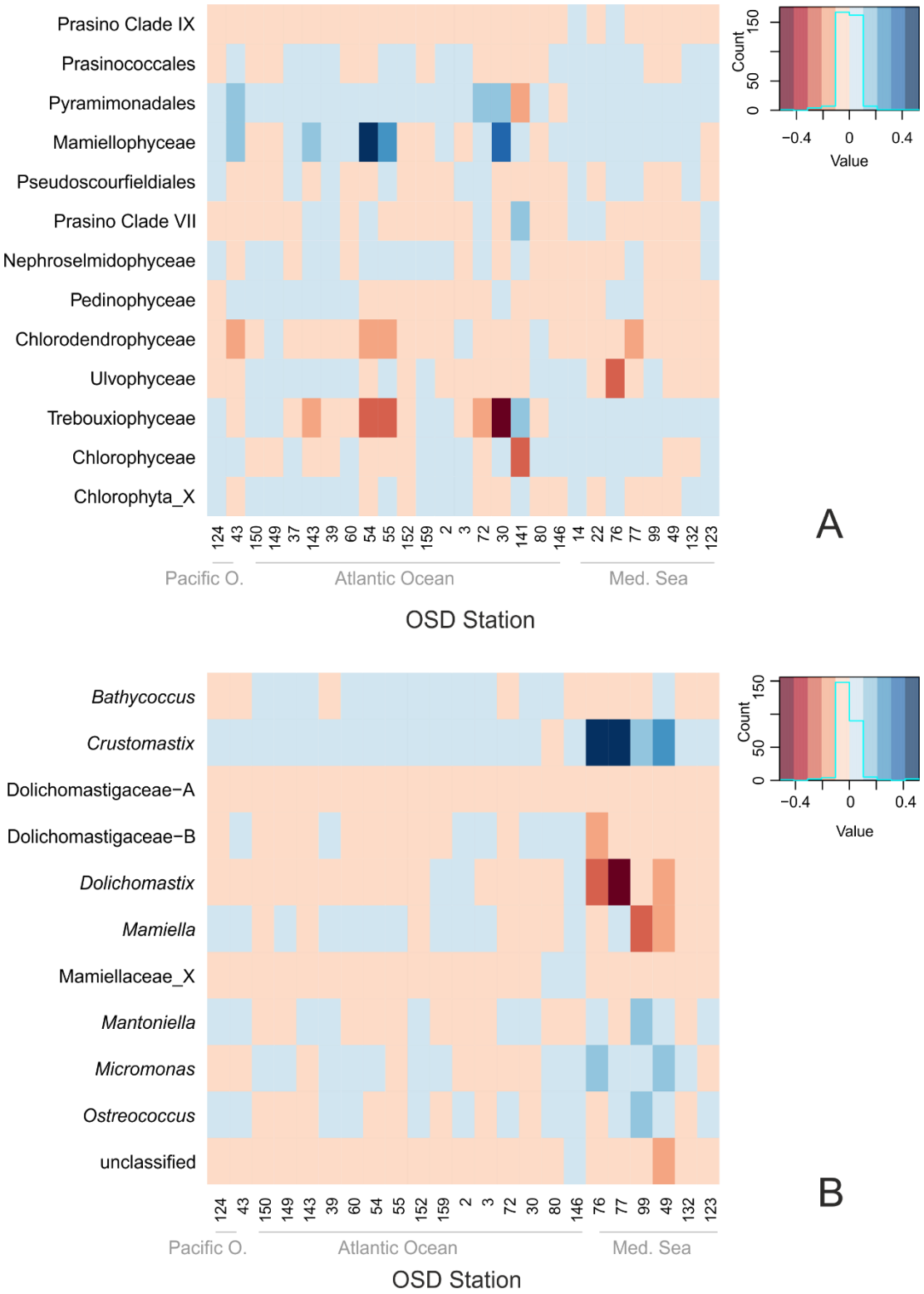


Fig. S9: Heatmap of differences between V9 and V4 (V9-V4) relative contribution: A- Chlorophyta classes B- Mamiellophyceae genera. The colors correspond to the difference from - 50% (- 0.5) to + 50 % (0.5).

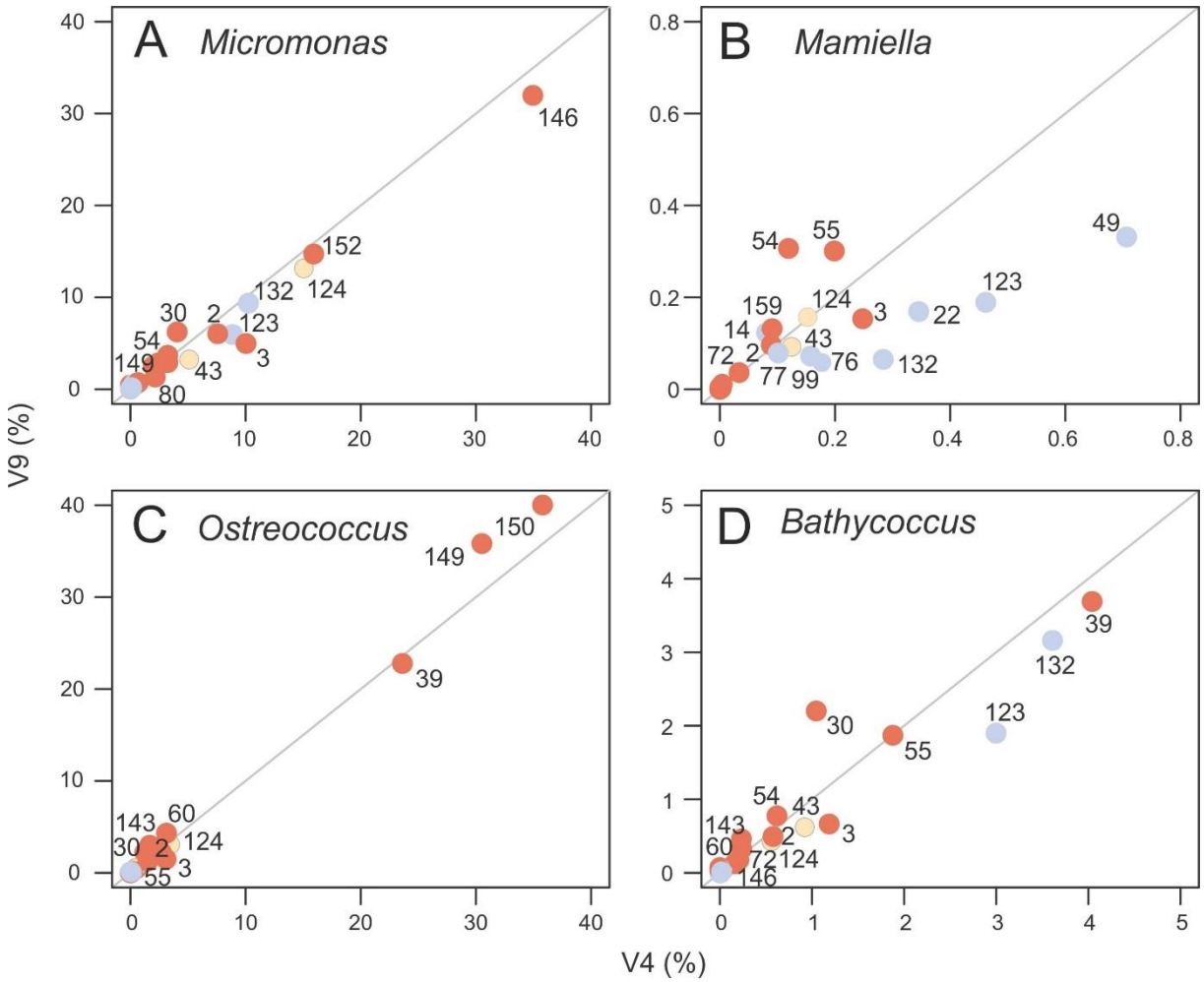


Fig. S10: Correlation between V4 and V9 relative contribution to Chlorophyta metabarcodes for major Mamiellophyceae genera: A. Micromonas, B. Mamiella, C. Ostreococcus, D. Bathycoccus.

Supplementary Table

Table S1: Percentage of identity within OTUs reference sequences from photosynthetic groups.

Identity %	V4	V9
Chlorophyta	76.3	72.1
Cryptophyta	84.1	75.7
Haptophyta	86	77.4

Supplementary Data

Supplementary Data 1: Mothur script for sequence analysis.

```
set.dir(input=/projet/sbr/osd/pathtoinputfiles,
output=/projet/sbr/osd/pathtooutputfiles)

# database cleaning

unique.seqs(fasta=database.fasta)

list.seqs(fasta=database.unique.fasta)

get.seqs(accnos=database.unique.accnos,taxonomy=database.taxo)

# ref alignment SILVA

summary.seqs(fasta=/projet/sbr/osd/database/silva.seed_v119.Euk.V2.align)

# Sub sampling dataset

       #-----Stations------#

get.groups(fasta=dataset.fasta, group=dataset.groups, groups=OSD123-OSD132-
OSD49-OSD99-OSD77-OSD76-OSD22-OSD14-OSD146-OSD80-OSD141-OSD30-OSD72-OSD3-
OSD2-OSD159-OSD152-OSD55-OSD54-OSD60-OSD39-OSD143-OSD37-OSD149-OSD150-
OSD43-OSD124)

         #-----Reads------#

sub.sample(fasta=dataset.pick.fasta, size=202710,
group=dataset.pick.groups, persample=T)
```

```
# dataset

unique.seqs(fasta=dataset.fasta)

summary.seqs(fasta=dataset.unique.fasta)

count.seqs(name=dataset.names, group=dataset.groups, processors=1)

split.abund(count=dataset.count_table, fasta=dataset.unique.fasta,
cutoff=1, accnos=T)

screen.seqs(fasta=datasetV4.unique.abund.fasta, minlength=170, maxambig=0,
count=dataset.abund.count_table)             #V4 length filter

screen.seqs(fasta=datasetV9.unique.abund.fasta, minlength=90, maxambig=0,
count=dataset.abund.count_table)              #V9 length filter


# align to SILVA and filter

align.seqs(fasta=dataset.unique.abund.good.fasta,
reference=/projet/sbr/osd/database/silva.seed_v119.Euk.pcr.V2.align,
flip=T, processors=1)

filter.seqs(fasta=dataset.unique.abund.good.align)

unique.seqs(fasta=dataset.unique.abund.good.filter.fasta,
count=dataset.abund.good.count_table)

# Preclustering and Chimeras checking

pre.cluster(fasta=dataset.unique.abund.good.filter.unique.fasta,
count=dataset.unique.abund.good.filter.count_table, diffs=2, processors=1)

chimera.uchime(fasta=dataset.unique.abund.good.filter.unique.precluster.fas
ta, count=dataset.unique.abund.good.filter.unique.precluster.count_table,
processors=1)

remove.seqs(fasta=dataset.unique.abund.good.filter.unique.precluster.fasta,
accnos=dataset.unique.abund.good.filter.unique.precluster.uchime.accnos,
count=dataset.unique.abund.good.filter.unique.precluster.count_table)

# classification with PR²

split.abund(count=dataset.unique.abund.good.filter.unique.precluster.pick.c
ount_table,
fasta=dataset.unique.abund.good.filter.unique.precluster.pick.fasta,
cutoff=1, accnos=T)
```

```
classify.seqs(fasta=dataset.unique.abund.good.filter.unique.precluster.pick
.abund.fasta,
count=dataset.unique.abund.good.filter.unique.precluster.pick.abund.count_t
able, template=database.unique.fasta, taxonomy=database.pick.taxo,
processors=1, probs=T)

summary.tax(taxonomy=dataset.unique.abund.good.filter.unique.precluster.pic
k.abund.pick.wang.taxonomy,
count=dataset.unique.abund.good.filter.unique.precluster.pick.abund.count_t
able)

# Clustering and OTUs

unique.seqs(fasta=dataset.unique.abund.good.filter.unique.precluster.pick.a
bund.fasta,
count=dataset.unique.abund.good.filter.unique.precluster.pick.abund.count_t
able)

dist.seqs(fasta=dataset.unique.abund.good.filter.unique.precluster.pick.abu
nd.unique.good.fasta, cutoff=0.05, countends=F, processors=1)

cluster(column=dataset.unique.abund.good.filter.unique.precluster.pick.abun
d.unique.good.dist,
count=dataset.unique.abund.good.filter.unique.precluster.pick.abund.unique.
good.count_table, method=nearest)

make.shared(list=dataset.unique.abund.good.filter.unique.precluster.pick.ab
und.unique.good.nn.unique_list.list,
count=dataset.unique.abund.good.filter.unique.precluster.pick.abund.unique.
good.count_table, label=0.03)

classify.otu(taxonomy=dataset.unique.abund.good.filter.unique.precluster.pi
ck.abund.pick.wang.taxonomy,
count=dataset.unique.abund.good.filter.unique.precluster.pick.abund.unique.
good.count_table,
list=dataset.unique.abund.good.filter.unique.precluster.pick.abund.unique.g
ood.nn.unique_list.list, label=0.03, probs=F, basis=sequence)

get.oturep(fasta=dataset.unique.abund.good.filter.unique.precluster.pick.ab
und.unique.good.fasta,
column=dataset.unique.abund.good.filter.unique.precluster.pick.abund.unique
.good.dist,
count=dataset.unique.abund.good.filter.unique.precluster.pick.abund.unique.
good.count_table,
```

```
list=dataset.unique.abund.good.filter.unique.precluster.pick.abund.unique.good.nn.unique_list.list, cutoff=0.03, sorted=number)

create.database(shared=dataset.unique.abund.good.filter.unique.precluster.pick.abund.unique.good.nn.unique_list.shared, label=0.03,
repfasta=dataset.unique.abund.good.filter.unique.precluster.pick.abund.unique.good.nn.unique_list.0.03.rep.fasta,
count=dataset.unique.abund.good.filter.unique.precluster.pick.abund.unique.good.nn.unique_list.0.03.rep.count_table,
constaxonomy=dataset.unique.abund.good.filter.unique.precluster.pick.abund.unique.good.nn.unique_list.0.03.cons.taxonomy)
```

Chapter 3

Communities of green microalgae in marine coastal waters: the OSD dataset

Communities of green microalgae in marine coastal waters: the OSD dataset

Margot Tragin[1], Daniel Vaulot[1]*

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

For: Frontiers in Aquatic Microbiology (Marine sciences/microbiology)

Keywords

Chlorophyta, Prasinophytes, distribution, 18S rRNA gene, V4 region, ecology, coastal marine systems, Ocean Sampling Day 2014

Acknowledgments

Financial support for this work was provided by the European Union projects MicroB3 (UE-contract-287589). MT was supported by a PhD fellowship from the Université Pierre et Marie Curie and the Région Bretagne. We would like to thank the Ocean Sampling Day consortium for providing sequence data and the ABIMS platform in Roscoff for access to bioinformatics resources.

* Corresponding author: vaulot@sb-roscoff.fr

Abstract

The ecology and distribution of green phytoplankton (Chlorophyta) in the ocean is poorly known since most studies have focused on abundant groups such as diatoms or dinoflagellates. The analysis of the Ocean Sampling Day metabarcoding dataset, which uses the V4 region of the 18S rRNA gene as a marker and sampled quasi-simultaneously 145 marine stations, mostly in coastal waters reveals that, Chlorophyta are ubiquitous and can be locally dominant. In this dataset, they represented 29% of the global photosynthetic reads (Dinoflagellates excluded) and their contribution was especially high in oligotrophic stations (up to 94%) and along the European Atlantic coast. Mamiellophyceae dominated most of coastal stations. At coastal stations where Mamiellophyceae were not dominating, they were replaced by Chlorodendrophyceae, Ulvophyceae, Trebouxiophyceae or Chlorophyceae, while oligotrophic stations were dominated either by prasinophytes clade VII (Chloropicophyceae) or IX. Several Chlorophyta classes showed preferenda in terms of nitrate concentration, distance to the coast, temperature and salinity. For example, Chlorophyceae preferred coastal northern high latitudes cold, low salinity waters, or prasinophytes clade IX warm, high salinity, oligotrophic oceanic waters.

Introduction

Marine waters are inhabited by a heterogeneous assemblage of organisms that includes a large diversity of unicellular eukaryotes, the protists. Protists are found in all branches of the tree of life (Baldauf, 2008). They are highly diversified with respect to size (from a few microns to several hundreds), morphology and trophic types (from photosynthetic to parasitic). This work focused on green micro algae from the Chlorophyta division. Green algae originate from primary endosymbiosis, they have a chloroplast surrounded by only two membranes and possess chlorophyll *b* as the main accessory chlorophyll.

The ecology and distribution of green phytoplankton in the ocean is poorly known since most studies have focused on groups that are easily identified by microscopy and cause massive blooms such as diatoms or dinoflagellates. Green algae representatives are found in several size fractions, in particular the picophytoplankton (cells from 0.2 to 2 µm) and nanophytoplankton (cells from 2 to 20µm), which are key primary producers in central oceanic regions (Worden *et al.*, 2004). Chlorophyta constitute the base of the green lineage (Nakayama *et al.*, 1998), leading to the hypothesis that the common ancestor of green algae and land plants could be an ancestral green flagellate (AGF) closely related to Chlorophyta (Leliaert *et al.*, 2012). The Chlorophyta division is composed of two major groups: the prasinophytes and the "core" Chlorophytes (Leliaert *et al.*, 2012; Fučíková *et al.*, 2014). The prasinophytes consist currently of nine major lineages of microalgae corresponding to different taxonomic levels (Order, Class, undescribed clades). The prasinophytes lineages share ancestral features such as flagella and organic scales. The number of prasinophytes lineages has been increasing following the availability of novel cultures and environmental sequences. Ten years ago, prasinophyte clade VII was introduced using sequences from cultured strains and environmental clone libraries (Guillou *et al.*, 2004). Four years later, two additional clades, VIII and IX, were reported (Viprey *et al.*, 2008) that are only known so far from environmental sequences. Prasinophytes clade VIII reference sequences were assigned as clade VIII for the first time in reference databases (Tragin *et al.*, 2016) and so this clade was looked for in metabarcoding datasets for the first time. Many prasinophytes clades should be raised to the Class level in the future (Leliaert *et al.*, 2012). As an example, Leliaert *et al.* (2016) recently used multigenic phylogenies to establish the new Palmophyllophyceae class, which gathers the Prasinococcales and the Palmophyllales orders. Clade VII has just been included into 2 new classes, Chloropicophyceae and Picocystophyceae (Lopes dos Santos *et al.*, 2017). The "late" diverging lineages (Pedinophyceae and Chlorodendrophyceae) have been merged with the Ulvophyceae-Trebouxiophyceae-Chlorophyceae (UTC) clade into the "core" Chlorophytes (Fučíková *et al.*, 2014).

Differences in the distribution of major classes or clades have already been demonstrated between coastal and oceanic waters. Mamiellophyceae are the major Chlorophyta contributors in coastal water, while the prasinophytes clade VII (Lopes dos Santos *et al.*, 2016) and IX (Rii *et al.*, 2016)

dominate oceanic waters. However, no global analysis of the relative importance and distribution of the different green algal groups in the ocean has yet been performed.

High Throughput Sequencing (HTS) methods provide large metabarcoding datasets which allow to explore the diversity and distribution of protist groups in the ocean. The Ocean Sampling Day project (OSD, Kopf *et al.*, 2015) has sampled in 2014 the global ocean, mostly at coastal stations, at the boreal summer solstice (June 21) and sequenced at each station the V4 region of the 18S rRNA gene. In this paper, we analyze the OSD V4 metabarcoding datasets to describe the distribution in the global coastal ocean of major classes of Chlorophyta for which a reference sequence database has been recently validated (Tragin *et al.*, 2016).

Materials and Methods

Sampling and sequencing

157 water samples from 145 marine locations (Table S1) were filtered on 0.22 µm pore size Sterivex without prefiltration and frozen at -80°C. Metadata (Temperature, Salinity, Nutrients and Chlorophyll *a*) are available at https://github.com/MicroB3-IS/osd-analysis/wiki/Guide-to-OSD-2014-data. Temperature and salinity were measured *in situ* during the sampling, while nutrients concentration were historical data uploaded from the World Ocean Database 2013 (Boyer *et al.*, 2013, https://www.nodc.noaa.gov/OC5/WOD13/) and the Chlorophyll *a* data were estimated from remote sensing ocean color from the MODIS AQUA database (Moderate Resolution Imaging Spectroradiometer, http://oceancolor.gsfc.nasa.gov/cgi/l3). In this paper, we only considered 145 samples corresponding to the surface layer.

DNA was extracted using the Power Water isolation kit (MoBio, Carlsbad, CA, USA) following the manufacturer instructions. V4 was amplified using TAReuk454FWD1 (5′-CCA GCA SCY GCG GTA ATT CC-3′) as forward primer and the modified TAReukREV3_modified (5′-ACT TTC GTT CTT GAT YRA TGA-3′) as reverse primer (Stoeck *et al.*, 2010; Piredda *et al.*, 2017). The Illumina libraries were prepared using the Ovation Rapid DR Multiplex System 1- 96 (NuGEN, link to protocol: https://owncloud.mpi-bremen.de/index.php/s/RDB4Jo0PAayg3qx?path=/2014/protocols). Sequencing (2x250 paired end) was done with Illumina technology MiSeq using V3 chemistry by the LGC genomics GmbH (Germany, http://www.lgcgroup.com/).

Data processing

R1 and R2 were filtered on quality and length and assembled by the OSD consortium which provided the so-called "workable" fasta files (https://owncloud.mpi-bremen.de/index.php/s/RDB4Jo0PAayg3qx?path=%2F2014%2Fsilva-ngs%2F18s). This dataset provided around 5 million workable V4 region of the 18S rRNA gene metabarcodes.

All subsequent sequence analyses were done with Mothur v 1.35.1 (Schloss *et al.*, 2009). Reads were filtered to be longer than 300 bp and without ambiguities (N). Then, reads were aligned on SILVA seed release 123 alignment (Pruesse *et al.*, 2007) corrected by hand with the Geneious software v7.1.7 (Kearse *et al.*, 2012): gaps at the beginning and at the end were deleted. The aligned datasets were filtered by removing columns containing only insertions. Chimeras were checked using Uchime v 4.2.40 (Edgar *et al.*, 2011) as implemented in Mothur. The datasets were pre-clustered using Mothur. After distance matrix calculation, the sequences were clustered using the Nearest Neighbor method and Operational Taxonomic Units (OTUs) were built at 99% similarity. OTUs represented by only one sequence (singletons) were deleted. OTUs were finally assigned using the Wang approach (Wang *et al.*, 2007) and the PR² database (Guillou *et al.*, 2013), available at https://figshare.com/articles/PR2_rRNA_gene_database/3803709, for which the Chlorophyta

sequences had been checked against the latest taxonomy (Tragin *et al.*, 2016). Assignation supported lower than 80% bootstrap were not taken into account. Each OTU is linked to a reference sequence and an OTU is considered to be assigned when the lowest taxonomic level ("Species" level in PR^2) differs from "unclassified". All OTU reference sequences were further BLASTed against GenBank nt database using megablast: max 10 best hit were recorded with a 97% identity and 0.001 e-value cutoff threshold. The BLAST hits allowed checking OTU assignation.

Statistical analyses

Graphics and ecological analyses were performed using the R 3.0.2 software (http://www.R-project.org/). We used the following packages: Treemap, Gplots, Mapdata and Maps. Distance to the coast was calculated for each station using Rgal and Rgeos packages and the coastline file available (http://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-coastline/). The Vegan package was used to compute rarefaction curve slopes (using the function *rareslope*), Bray-Curtis dissimilarity matrices (function *vegdist*) and to perform Nonparametric Multi-Dimensional Scaling (NMDS using the *metaMDS* function). OSD metadata were projected onto the NMDS plots using the *envfit* function from the Vegan with the p.max option set as 0.95.

Results

The OSD dataset

All OSD stations (Table S1) were sampled around the same date, June 21, 2014, the boreal summer solstice. In contrast to other global surveys such as Tara *Oceans* (Pesant *et al.*, 2015), OSD stations were mostly coastal: distance from the coast varied from a few meters (OSD43 off Scripps Institute of Oceanography in California was calculated to be 10 m from the coast) to more than 300 km (OSD146 Fram Strait in Greenland Sea). However, some stations located offshore oceanic islands such as OSD7 (Moorea - Tiahura) in French Polynesia corresponded to truly oceanic waters. Sample sites corresponded to a wide range of temperature and salinity: from polar (minimal water temperature was -1.6°C at OSD146, Fram Strait in Greenland Sea) to tropical waters (max. water temperature was 31.3°C at OSD39), from freshwater (OSD10 was located in Lake Erie with 0.14 PSU) and brackish waters (for example OSD35 in Chesapeake Bay with 8.9 salinity) to marine (for example OSD57 salinity 34) or hypersaline waters (max. salinity was 100 at OSD145). Nitrates ranged from below the detection limit (OSD6 and 14 in Mediterranean Sea, OSD56, 57 and 144 off Hawaii, OSD46 in the Gulf of Mexico and OSD147, Bay of Bengal) to 9 µM in a coastal lagoon in Uruguay (OSD149, 150 and 151) with an average 2.3±3.2 µM were calculated for the OSD stations excluding the 21 stations without data (in which the OSD7, Moorea in French Polynesia). The phosphate concentration was in average 0.23±0.22 µM with bimodal distribution with a wide peak around 0.05 µM and a second around 0.55 µM. Concentrations ranged from less than 0.005 µM in OSD24, 25 and 94 (off Morroco), OSD28 (Belize) and 45 (Tampa Bay in the Gulf of Mexico) to 1.55 µM in OSD71 (Otago in New Zealand).

Chorophyta contribution to photosynthetic phytoplankton in coastal waters

The global OSD dataset provided 1,103,675 reads of the 18S rRNA V4 regions that could be assigned to photosynthetic organisms, Dinoflagellates excluded because about 50% of the species are not photosynthetic (Gómez, 2012). Among these, Chlorophyta represented on average more than a quarter (28±24% Fig.2A) and constituted the second most represented photosynthetic division in terms of percent of reads and number of OTUs after Ochrophyta (Diatoms, Fig.1). The number of reads per station assigned to Chlorophyta ranged from 4 at OSD172 (off Belgium) to 18,570 at OSD111 (Ria de Aveiro in Portugal, Table S1) with an average of 2,041±3,076 reads. Less than 100 Chlorophyta were recovered in 20 surface stations (Table S1). In terms of percentage of photosynthetic reads, it varied from less than 1% at OSD41 (Alaska,), 128 (Eyafjordur 3 off Iceland), 155 (Oslofjord off Norway), 157 (Skaggerak off Norway) and 187 (Palmer Station in Antarctica) to 94% at OSD7 (Moorea in French Polynesia). The percentage of Chlorophyta decreased from the equator (around 40% of Chlorophyta reads in average) to 60°N (circa 10 %) and increased again up to 20 % in the Northern high latitudes (Fig.3A). It was maximum between 0.5 and 1 km of the coast, decreasing in the near shore areas to increase again further away to almost 40% (Fig.3B).

745 OTUs built at 99% identity were assigned to Chlorophyta. The slope of Chlorophyta based OTU rarefaction curves was inversely proportional to the number of reads (Fig. S1) and reached saturation (slope <0.1) for 92% of the stations. Saturation slope did not appear to be linked to the geographic origin of the samples (Fig. S1).

On average, 36±20 OTUs were found per station, ranging, considering only stations with more than 100 Chlorophyta reads, from around 10 in OSD80 (off Greenland) and 174 (off Belgium) to 98 OTUs in OSD92 (off Morocco) (Fig.2B). No correlation was found between the percentage of Chlorophyta and the number of OTUs at the same station (R^2=0.06, p-value = 0.002, data not shown). At some stations, a high percent of Chlorophyta corresponded to a low number of OTUs (Fig.2) such as at OSD7 (Moorea, 94%, 20 OTUs, Table S1), 50 (Spain, 90%, 31 OTUs), 80 (Greenland, 40%, 28 OTUs), 105 (Cambridge Bay, Canada, 54%, 16 OTUs) and 146 (47%, 28 OTUs). At these stations, the Chlorophyta community was dominated by one or very few species such as *Micromonas polaris* (OSD105 and 146) or *Carteria* sp. and *Pyramimonas* sp. (OSD80) at the Northern high latitude stations. For OSD7, the dominant OTUs were assigned to prasinophytes clade IX and VIIB2 (now Chloroparvula sp., Lopes dos Santos *et al.*, 2017) and in OSD146 the main OTUs was assigned to an unknown Chlorodendrophyceae and to *Micromonas bravo*. In contrast, for other stations such as OSD10 (lake Erie, 5%, 41 OTUs), OSD48 (Gulf of Venice, Italy, 4%, 33 OTUs), 72 (Baltic Sea, 6%, 42 OTUs), 95 (Singapore, 19%, 40 OTUs) and OSD178 (Belgium, 6%, 39 OTUs) a low contribution of Chlorophyta to photosynthetic reads corresponded to a large number of OTUs (Fig.2).

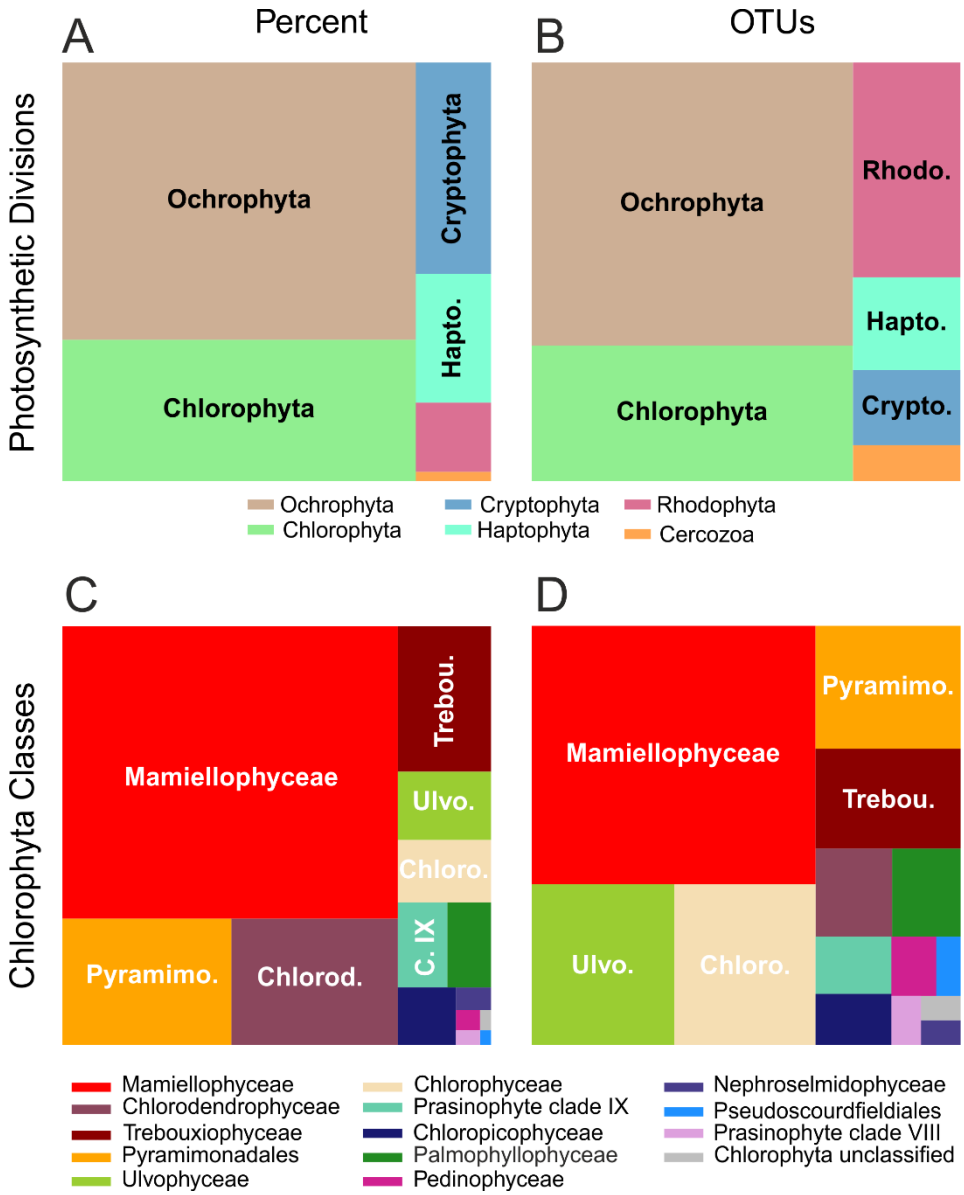


Fig.1: Overview of the contribution (number of reads) and diversity (number of OTUs) of photosynthetic group at OSD stations. A. Reads per photosynthetic divisions (Total = 1,103,675). B. Idem for OTUs (Total = 3069). C. Reads per Chlorophyta classes (Total = 320,481). D. Idem for OTUs (Total = 745).

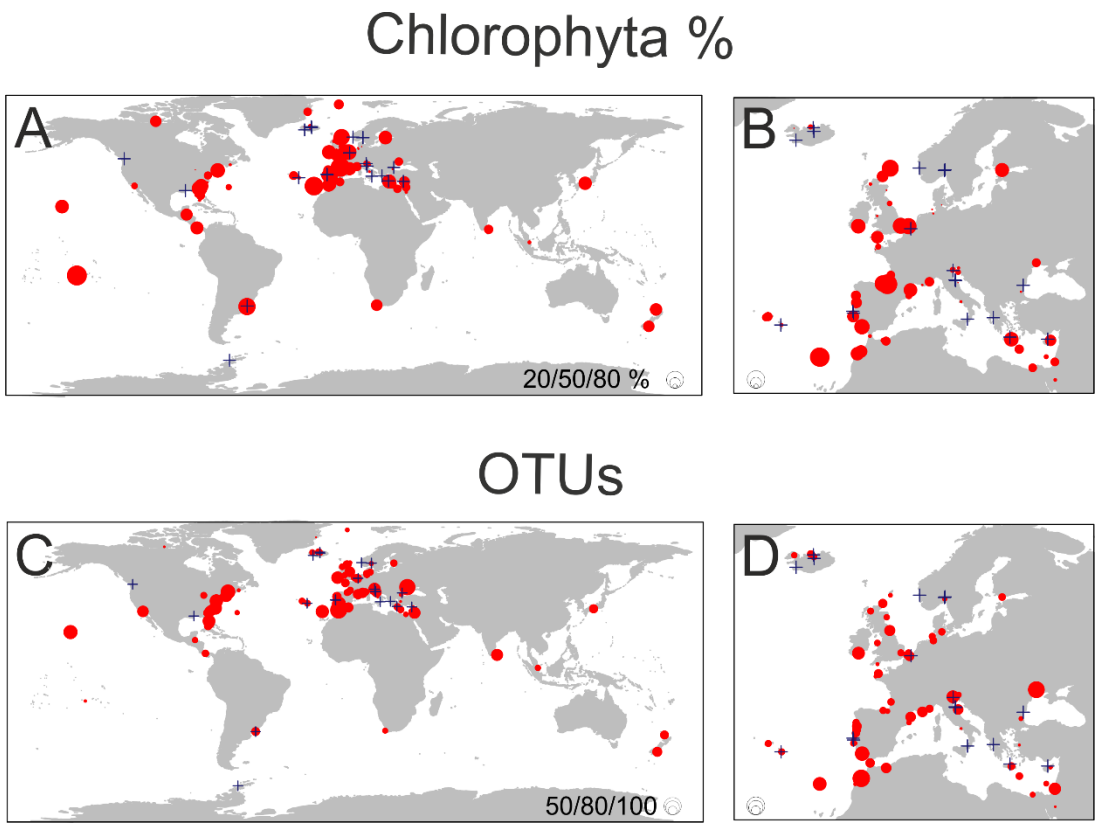


Fig.2:A Map of the contribution of Chlorophyta to OSD photosynthetic reads (dinoflagellates excluded) B**.** Idem Europe. C and D**.** Idem for number of OTUs based on 99% similarity. Stations where less than 100 Chlorophyta reads were recorded are represented by blue crosses.

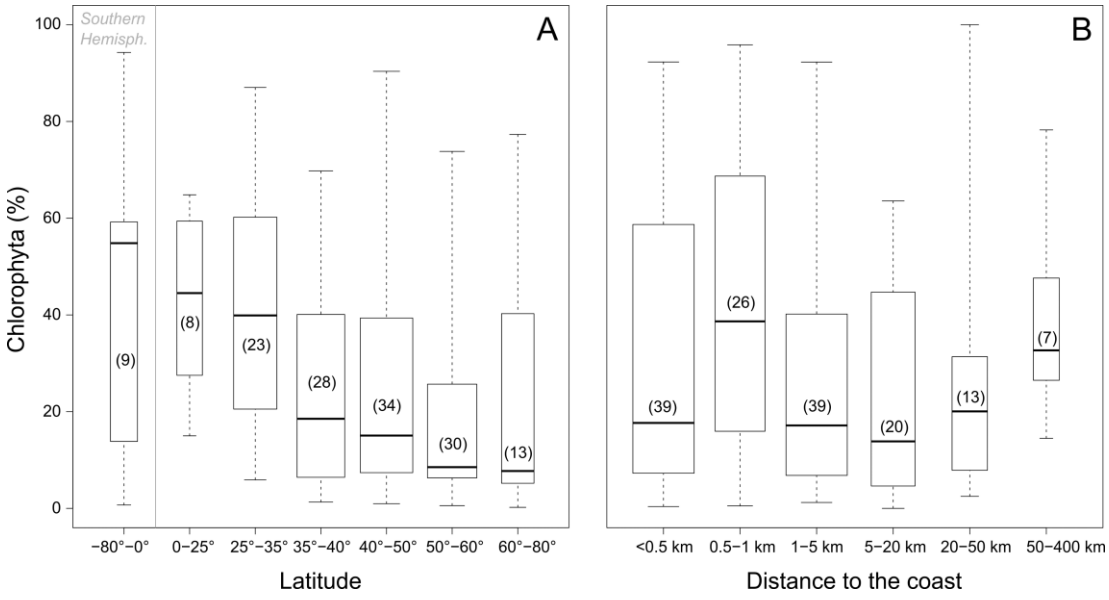


Fig.3: Boxplots of Chlorophyta contribution to photosynthetic reads (Dinoflagellates excluded) per range of metadata **A**. Latitude. B. Distance to the coast. Number in brackets are the number of stations in the range also represented by the boxplot width. The value represented from top to bottom: maximum, third quartile, median (in bold), first quartile and minimum.

Relative abundance and diversity of the different Chlorophyta classes in coastal waters

Overall, Mamiellophyceae dominated Chlorophyta in terms of read abundance (55%, Fig.1) and had the highest number of OTUs (304, Fig.1). They were followed by Pyramimonadales (12%), Chlorodendrophyceae (12%), and the UTC clade (Ulvophyceae, Trebouxiophyceae and Chlorophyceae: 3.5%, 7.5% and 3.2% respectively, Fig.1). The distribution of OTUs among the different classes was almost the same as their abundance (Fig.1). However, Ulvophyceae and Chlorophyceae had more OTUs (respectively 95 and 94 OTUs) than could be expected from their relative contribution (respectively 3.5 and 3.2%). Ulvophyceae representatives were mostly corresponding to macroalgae and could have originated from gametes or unicellular stages. Pyramimonadales and Chlorodendrophyceae both represented 12% of the Chlorophyta reads but three time more OTUs belonged to Pyramimonadales (74) than to Chlorodendrophyceae (28) OTUs (Fig.1). Chlorodendrophyceae were dominated by OTUs with large number of reads (the larger one corresponding to 29,899 reads and was assigned to undescribed environmental reference sequences), while Pyramimonadales OTUs has a smaller number of reads, the two major one corresponding to 5089 and 2627 reads, respectively. Several classes with low overall contributions had a quite large number of OTUs: for example, the Palmophyllophyceae, the prasinophytes clade VII and clade IX contributed to about 2% of the Chlorophyta reads, but had respectively 25, 16 and 18 OTUs (respectively 3.4, 2.2 and 2.4% of the Chlorophyta OTUs, Fig. 1). Pedinophyceae represented less than 0.3% of the Chlorophyta reads but 11 OTUs (1.5 % of the OTUs, Fig.1).

Distribution of specific Chlorophyta classes in coastal waters

Mamiellophyceae were recovered at almost all stations (Fig.4) sometimes in very high proportion (up to 99% at OSD183 off Belgium). Mamiellophyceae were also recovered in Lake Erie (OSD10, Fig.4). The major Mamiellophyceae OTUs were assigned to the three genera *Ostreococcus* (80,988 reads), *Micromonas* (47,778 reads) and *Bathycoccus* (22,305 reads). Remarkably, no Mamiellophyceae reads were recorded from the oligotrophic station OSD7 and OSD28, as well as at OSD90 (Etoliko lagoon in Greece), 96 (one of the 3 stations in Azores) and 114 (Portugal, Fig.4).

Pyramimonadales contribution ranged between 90 (OSD108, Portugal coast) to 0% especially in OSD28 (Belize), in Azores (OSD97 and 98) and in OSD124 (Japan). Pyramimonadales were sporadically spread in the global ocean (Fig.4). No clear distribution patterns appeared at this taxonomic level, which could be linked to the large amount of OTUs assigned to this class.

Chlorodendrophyceae represented up to 99% of Chlorophyta reads at OSD93 (Morocco) and were abundant at Mediterranean stations (OSD4 with 91%, 6 with 58%, 14 with 81%, 24 with 82%, 94 with 43% for example, Fig.4). Chlorodendrophyceae percent were less abundant along the North American coasts (OSD28 with 16%, 41 with 3.9%, 58 with 4.6%, 60 with 12% for example, Fig.4) and absent in

the sub-polar North Atlantic (stations around Iceland, Greenland or Fram Strait Fig.4). Chlorodendrophyceae were also recovered in Lake Erie (OSD, Fig.4).

Ulvophyceae maximal contribution was recorded in OSD169 (North Sea off UK, 70%). Ulvophyceae were mostly present along the North Atlantic European coast, in some stations of Mediterranean Sea (OSD78 in Adriatic Sea and OSD 123 off Israel for example), in equatorial stations (OSD28,124 and 147) and in Antarctica (OSD187, Fig.4).

Trebouxiophyceae represented up to 80% of Chlorophyta reads in OSD45 (Gulf of Mexico). They were recorded in temperate coastal waters, especially in the USA East coast, North Europe Coast and in Uruguay coastal lagoon (OSD151, Fig.4). Trebouxiophyceae were not recorded at high latitudes nor oligotrophic stations (such as Hawaiian, French Polynesian or Azores stations).

Chlorophyceae were always minor contributors to Chlorophyta and represented more than 1% of Chlorophyta reads only at 35 stations located in the North hemisphere (Fig.4). Their maximal contribution was reached in Greenland (OSD80, 95%), in Lake Erie (62%) and in the Black Sea (OSD13, 47%) and Mediterranean Sea (OSD90, Etoliko lagoon, Greece, 57%).

The uncultivated prasinophytes clade IX represented more than 1% of the Chlorophyta reads at 17 stations mostly located in oligotrophic tropical and temperate stations (Fig S2 and Fig. S3). The highest prasinophytes clade IX contribution were found in the Pacific Ocean (OSD7, French Polynesia, 78%), Mediterranean Sea (OSD52 and 53, respectively 70 and 78%) and at OSD28 (Belize, 34%, Fig S2 and Fig.5).

Within Palmophyllophyceae only OTUs assigned to Prasinococcales were found and none to Palmophyllales. They contributed to more than 1% at 39 stations (Fig. S3) mostly in the Mediterranean Sea and along North Europe coasts (Fig S2). Maxima were recorded off Cyprus, OSD18 and 19 respectively 32 and 77%) and in the Skaggerak (OSD157, 37%).

Prasinophytes clade VII (now Chloropicophyceae, Lopes et al. 2017) represented more than 1% at 25 stations (Fig. S3) mostly located in tropical oceanic waters. Prasinophytes clade VII reached their highest contribution at 2 of the Azores stations OSD96 and 97 (respectively 96% and 97%) and off Bermuda (OSD8, 29%, Fig.4).

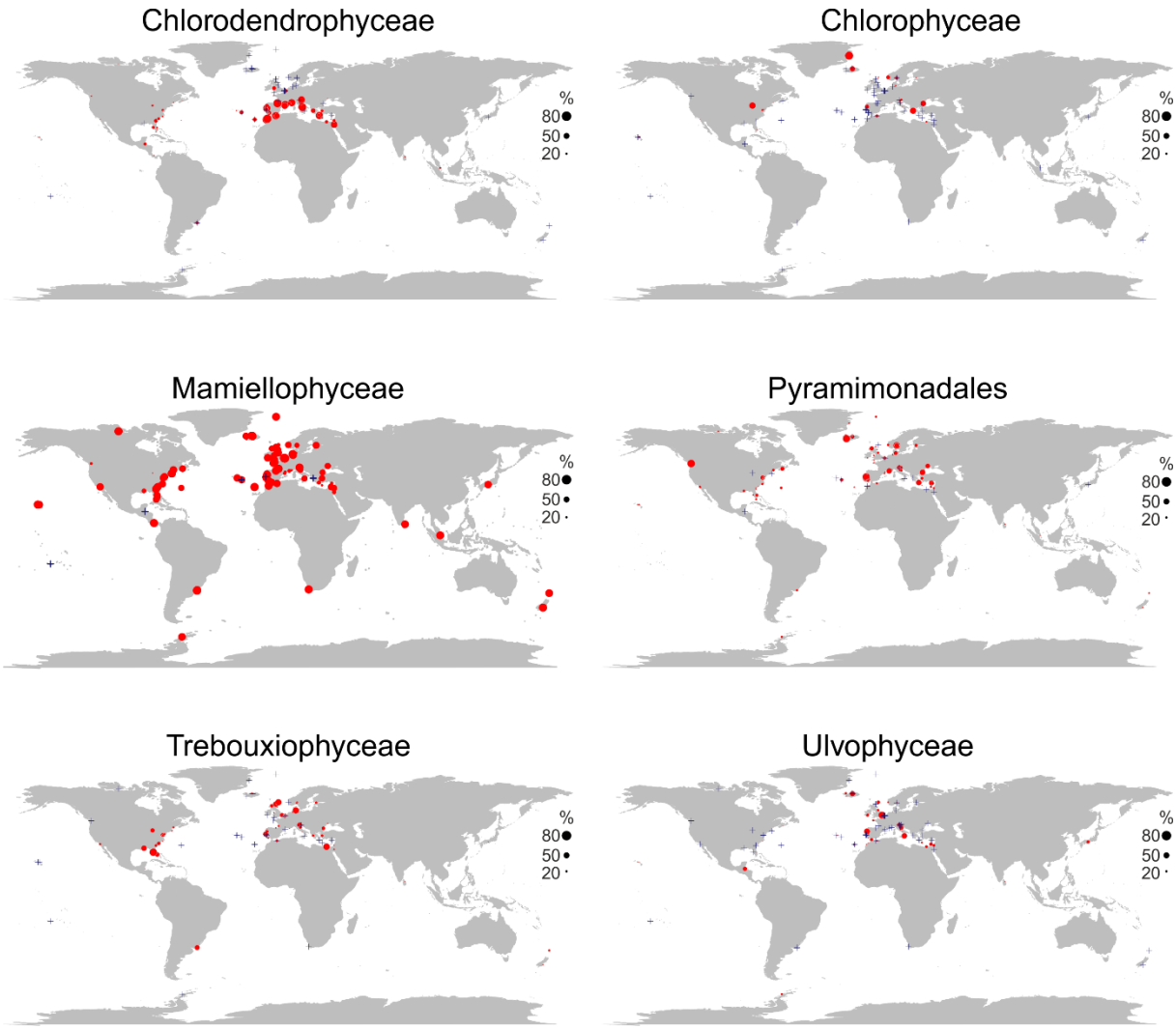


Fig.4: Contribution of the 6 major classes of Chlorophyta at OSD stations in surface. Stations where a given class was not recorded are represented by blue crosses. The circle surface is proportional to the percent of class versus Chlorophyta reads.

Nephroselmidophyceae represented more than 1% at 19 stations (Fig. S3) and their maximal contribution between 5 and 6% of the Chlorophyta reads were recorded in the coastal north Atlantic Ocean (OSD106 and 129 in Iceland, 152 in Canada and 157 in Norway, Fig S2). The Nephroselmidophyceae also reached 2% in several stations in the Mediterranean Sea Eastern Basin (such as OSD123 in Israel, Fig S2). No Nephroselmidophyceae reads were sampled in OSD133 (South Africa) and their contribution to the stations located in the Pacific Ocean were very low, although several Nephroselmidophyceae strains were isolated and described from these areas (Faria *et al.*, 2011, 2012, Yamaguchi *et al.*, 2011, 2013).

Pedinophyceae represented more than 1% of the Chlorophyta reads at 10 stations (Fig. S3) and were mostly present at stations located off the USA Atlantic coast (OSD35, 46, 143, 186) and in the Mediterranean and Black Seas (OSD64 and 78, Fig S2). The highest contribution (7.1%) was recorded in Chesapeake Bay (OSD35).

Prasinophytes clade VIII represented more than 1% at 4 stations (Fig. S3) in Mediterranean Sea and North Europa coast (Fig S2). The maximal contributions were found off Iceland (OSD20, 8%) and in the Adriatic Sea (OSD76 and 77, respectively 4.3 and 2%).

Pseudoscourfieldiales was the least represented class in this dataset. This class represented more than 1% of the Chlorophyta reads only at 3 stations (Fig S2 and Fig. S3), in particular at OSD46 (4.5%, Horn Island in Gulf of Mexico), two stations located in Adriatic Sea (OSD 48 and 99, 1.8% and 1%, respectively).

Finally, at 16 stations (Fig. S3), more than 1% of the reads could not be classified in any Chlorophyta class (undetermined Chlorophyta, Fig S2). The maximal fraction of unclassified sequences was found in the Mediterranean Sea off Cyprus (OSD18,16%), off Belize (OSD 28,8.1%), off the East Coast of the US (OSD58, 7.5). Other unclassified reads were recovered from the Mediterranean Sea stations and off Iceland (OSD128 and 129).

Chlorophyta community structure in coastal waters

In coastal waters, several types of Chlorophyta communities could be clearly defined (Fig.5). A group of 103 stations was dominated by Mamiellophyceae. At these stations Mamiellophyceae were often complemented by a second Chlorophyta class, either Pyramimonadales, Chlorodendrophyceae or Trebouxiophyceae. A second group of 14 stations was dominated only by Chlorodendrophyceae with virtually no other class present. Finally, a group of 28 stations were dominated by one of the other classes. In this group, stations sampled in oligotrophic waters were dominated by prasinophytes environmental clade IX (OSD7, 28, 52 and 53) or prasinophytes clade VII (OSD7, 96 and 97, Fig.4 and Fig.5). Interestingly, these two prasinophytes clade VII and IX were rarely found at the same stations

(Fig.5). In coastal waters, the UTC classes, the Chlorodendrophyceae or the Pyramimonadales very rarely co-occured (Fig.5).

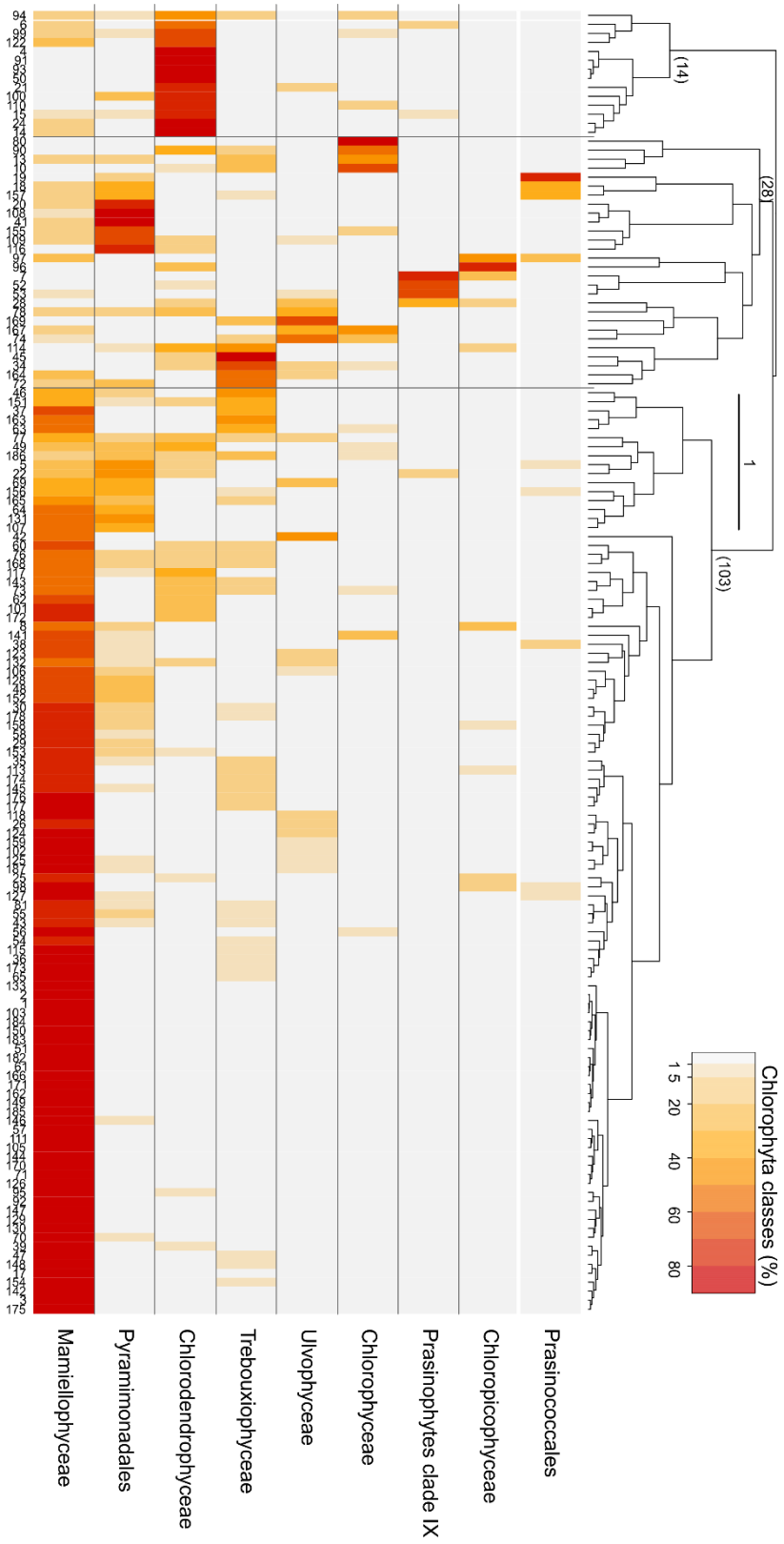


Fig.5: Heatmap of the OSD Chlorophyta communities. Chlorophyta classes representing on average less than 1% in the 145 stations were excluded. The dendrogram hierarchically clustered the Chlorophyta relative contribution per station based on a dissimilarity matrix. Number in brackets are the number of stations in each cluster. Colors refer to the percent of reads in each class related to the total number of Chlorophyta reads.

Relation with environmental parameters

Mamiellophyceae did not seem to have any marked preferendum with respect to the environmental parameters recorded or estimated at the OSD stations (Fig.6), except that they seem to be less dominant at salinities between 37 and 40 PSU, typical of the Mediterranean Sea. The contribution of Pyramimonadales and Ulvophyceae was also similar under most environmental conditions. In contrast, some groups had marked preferenda. For example, Chlorophyceae and Chlorodendrophyceae were bigger contributors at low NO_3 and PO_4 and close to the coast but the formers were contributing more at low temperature and low salinity while it was the opposite for the latter. Two groups were typically found in oligotrophic oceanic waters, clade VII and IX, as reflected by their preference for high salinity, very low nutrients (NO_3) and large distances from the coast. However, clade VII extends a bit more towards the coast and has a slightly wider range of temperature, compared to clade IX which is mostly found in waters above 25°C but not beyond 30°C. As clades VII and IX, Pedinophyceae were mostly observed in low nitrate waters above 15°C but they could be found very close to the coast.

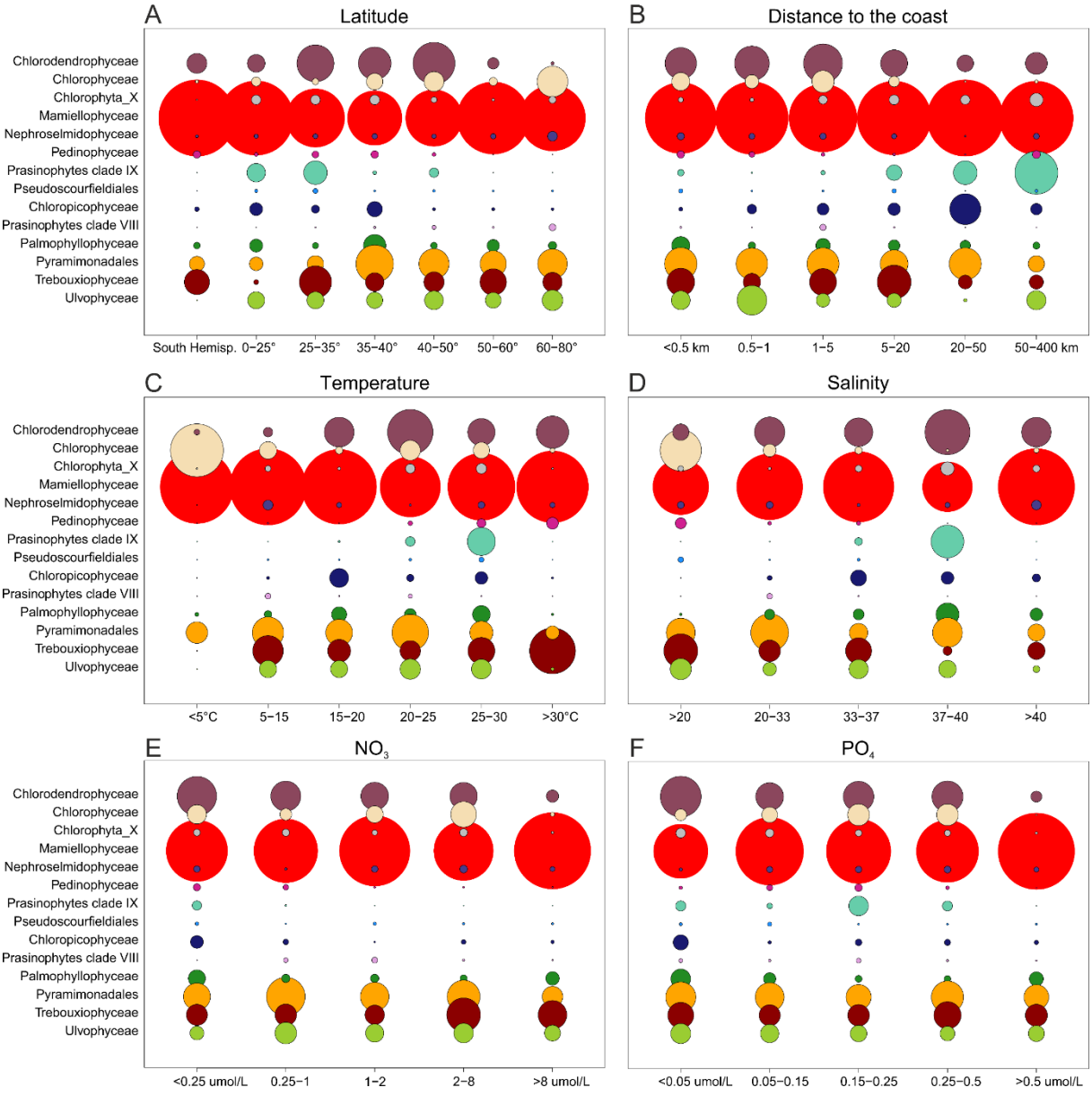


Fig.6: Contribution of Chlorophyta classes per range of metadata: A. Latitude (OSD metadata), B- distance to the coast (calculated), C- Water temperature (measured in situ), D- Salinity (measured in situ), E- Nitrates (World Ocean database 2013), F- Phosphates (World Ocean database 2013). Circles are proportional to the average contribution of a given class to Chlorophyta. For salinity, OSD10 was not taken into account since it is located in a freshwater lake.

Discussion

Green algae are clearly significant photosynthetic contributors in coastal waters. Chlorophyta constituted the second major photosynthetic group (dinoflagellates excluded) both in terms of read contribution and number of OTUs (Fig.1). The importance of Chlorophyta had already been highlighted previously. In European coastal, the contribution of Chlorophyta to photosynthetic 18S rRNA clones fell in the same range found during OSD (42%, Massana and Pedrós-Alió, 2008). In the English Channel and North Sea, 47% of the eukaryotic cells hybridized by TSA-FISH were Chlorophyta, more precisely Mamiellophyceae (Not *et al.*, 2002, 2004, 2007; Masquelier *et al.*, 2011). A recent HTS survey in coastal European surface waters found Chlorophyta as the major photosynthetic group after diatoms (Massana *et al.*, 2015). Similar contributions were also observed in two other OSD data sets for a smaller number of stations using the V4 and the V9 marker (26% and 20%, respectively, Tragin *et al.*, 2017). In comparison, Chlorophyta have a lower overall contribution (13% in average) in the Tara *Ocean* V9 data set from oceanic waters (Lopes dos Santos *et al.*, 2016). The number of Chlorophyta 99% OTUs (745) was in the same range than for other studies using different thresholds : 314 V4 OTUs at 97% similarity in European coastal waters (Massana *et al.*, 2015) or 1420 to build the V9 OTUs built with SWARM (de Vargas *et al.*, 2015).

In the OSD dataset the percentage of Chlorophyta was maximum in tropical oceanic waters (e.g. 94% OSD7 off Moorea, Fig.3). Such high Chlorophyta contributions in tropical waters have been recently observed in the Tara *Ocean* data set (Lopes dos Santos *et al.*, 2016) but also previously in clone libraries studies (Shi *et al.*, 2009). In contrast, low Chlorophyta contributions (less than 100 Chlorophyta reads) were recovered in Northern Europe Fjord, Mediterranean Sea and at the Palmer station in Antarctica.

In the OSD dataset, Mamiellophyceae (especially *Micromonas*, *Ostreococcus* and *Bathycoccus*) were omnipresent in coastal waters exhibiting a wide range of environmental conditions, as previously reported by many studies in a wide range of coastal and nutrient-rich environments from the Arctic Ocean to the Mediterranean Sea through the Pacific and Indian Oceans (Not *et al.*, 2004, 2005, 2008; Marie *et al.*, 2006; Lovejoy *et al.*, 2007; Masquelier *et al.*, 2011; Lin *et al.*, 2016). Not *et al.* (2009) found *Micromonas* to be the most prevalent genus in the world ocean coastal waters and at more local scale, *Micromonas* dominates coastal picoplankton in the Western English Channel (Not *et al.*, 2004). Collado-Fabri *et al* (2011) and Rii *et al* (2016) found that Mamiellophyceae (*Micromonas*, *Ostreococcus* and *Bathycoccus* mostly) were dominant in the upwelling-influenced coastal waters off Chile and FISH analyses during a 2 years-long study showed that they accounted for the totality of Chlorophyta cell in the Chile coastal upwelling (Collado-Fabbri *et al.*, 2011). Using quantitative PCR, Marie *et al.* (2006) found *Bathycoccus* to be dominant in a transect through the Mediterranean Sea.

In contrast, the contribution of Mamiellophyceae was low at oceanic OSD stations, which had been previously demonstrated by Lopes dos Santos *et al.* (2016) based on the Tara *Ocean* oceanic dataset in which only 17% of the reads assigned to Chlorophyta belong to Mamiellophyceae. Nutrient depleted environments such as the oligotrophic Pacific Ocean have been previously reported to host Chloropicophyceae (Lopes dos Santos *et al.*, 2016) and clade IX (Shi *et al.*, 2011; Wu *et al.*, 2014; Rii *et al.*, 2016). Clade IX distribution pattern in OSD stations was consistent with the fact that clade IX was initially discovered in the Mediterranean Sea (Viprey *et al.*, 2008). Despite the fact that both prasinophytes clade IX and Chloropicophyceae are characteristics of oligotrophic oceanic regions, they rarely co-occurred in OSD stations (Fig.5). These classes could be differently distributed in a continuum of low nutrients gradient, prasinophytes clade IX preferring more oligotrophic areas such as the south China Sea (Wu *et al.*, 2014) or the Pacific gyre (Shi *et al.*, 2009) than Chloropicophyceae. Moreover, the clade IX (as well as the prasinophytes clade VIII) could have been underestimated in previous HTS studies based on the the PR² database (Guillou *et al.*, 2013) because, before the database curation for Chlorophyta sequences (Tragin *et al.*, 2016), clade VIII and IX reference sequences were badly assigned as Nephroselmidophyceae and prasinophytes clade I, respectively.

Pyramimonadales were recovered everywhere in OSD and were the second most abundant Chlorophyta class as found in the Tara *Oceans* dataset (de Vargas *et al.*, 2015). They were particularly prevalent in the Mediterranean Sea and North Atlantic Ocean, where microplankton microscopy inventories previously recorded the presence of the genera *Halosphaera* and *Pterosperma* (Wiebe *et al.*, 1974; Kimor and Wood, 1975; Jenkinson, 1986; Sarno *et al.*, 1993; Gomez and Gorsky, 2003; Balkis, 2009). In the OSD dataset Pyramimonadales did not show any environmental preferendum supporting the observation made by Viprey *et al.* (2008) that Pyramimonadales were found in almost all metadata categories they sampled in the Mediterranean Sea. Pyramimonadales strains have been isolated from a large range of environments (Moestrup and Hill, 1991): coastal waters (Chisholm and Brand, 1981; Pienaar and Sym, 2002), polar regions (Daugbjerg and Moestrup, 1992; Moro *et al.*, 2002; Harðardottir *et al.*, 2014), Mediterranean Sea (Wiebe *et al.*, 1974; Zingone *et al.*, 1995). Surprisingly, Pyramimonadales were not recovered from the Japan station (OSD124), while numerous strain or natural samples sequences from GenBank originate from this area (Suda *et al.*, 2013; Tragin *et al.*, 2016), nor in South Africa coastal waters, where a wide diversity of *Pyramimonas* was isolated (Pienaar and Sym, 2002).

In OSD dataset, Chlorodendrophyceae were well represented in Mediterranean Sea and contributed for a small part of Chlorophyta off the US coast and in the Indian Ocean. The major Chlorodendrophyceae genus *Tetraselmis* has been reported in several microscopic inventories in the Mediterranean Sea and Northern Atlantic Ocean (Marshall, 1980; Samanidou *et al.*, 1987; Sarno *et al.*, 1993; Harzi *et al.*, 1998) and strains have been isolated in a wide range of environments (Lee and Hur, 2009). Arora *et al.* (2013) isolated *Tetraselmis* strains from Indian salt pan suggesting that species from

this genus can thrive in high salinity environments as observed in OSD. At some OSD stations Mamiellophyceae were replaced by Chlorodendrophyceae. This group has been overlooked from 18S rRNA surveys because most of these focused on the picophytoplankton size fraction (Not *et al.*, 2004, 2009; Wu *et al.*, 2014; Giner *et al.*, 2016; Limardo *et al.*, 2017), while Chlorodendrophyceae species, such as those from the genus *Tetraselmis*, are rather nanoplanktonic (Tragin *et al.*, 2016). Some 18S rRNA sequences have been retrieved from the Mediterranean Sea from surface, low nutrients, high temperature and high salinity samples (Guillou *et al.*, 2004; Viprey *et al.*, 2008), which corroborate the pattern observed in the OSD data.

At some other OSD stations, one of the classes from the UTC clade dominated the Chlorophyta communities. The Chlorophyceae especially showed clear environmental preferendum for low salinity and low temperature waters in this dataset. Chlorophyceae (e.g. *Dunaliella*) have been shown to be tolerant to a large salinity range from freshwater to marine water (Margulis *et al.*, 1980; Borowitzka and Huisman, 1993) and have been already recorded in coastal Arctic, Southern Ocean and Northern Europe samples (Marshall, 1980; Harzi *et al.*, 1998; Majaneva *et al.*, 2012; Tragin *et al.*, 2016). In contrast, Trebouxiophyceae and Ulvophycae did not show environmental preferenda in OSD. The Ulvophyceae contribution should be investigated at the OTUs level in order to be carefully interpreted because sequences may correspond to unicellular stage (e.g. gametes…) of macroalgae.

Conclusion

This paper offers a first insight into the contribution and distribution of Chlorophyta classes in marine waters. It highlights that Chlorophyta can be the main photosynthetic group in some ecosystems. Although most of the work on this division in the last decade has focused on Mamiellophyceae and more specifically three genera *Ostreococcus*, *Bathycoccus* and *Micromonas*, other classes of green algae can be locally important in specific environments and probably play a key role in the ocean ecosystems such as Chlorodendrophyceae or Pyramimonadales.

References

Arora, M., Anil, A.C., Leliaert, F., Delany, J., and Mesbahi, E. (2013) *Tetraselmis indica* (Chlorodendrophyceae, Chlorophyta), a new species isolated from salt pans in Goa, India. *Eur. J. Phycol.* **48**: 61–78.

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

Balkis, N. (2009) Seasonal variations of microphytoplankton assemblages and environmental variables in the coastal zone of Bozcaada Island in the Aegean Sea (NE Mediterranean Sea). *Aquat. Ecol.* **43**: 249–270.

Borowitzka, M.A. and Huisman, J.M. (1993) The Ecology of *Dunaliella salina* (Chlorophyceae, Volvocales): Effect of Environmental Conditions on Aplanospore Formation. *Bot. Mar.* **36**: 233–244.

Boyer, T.P., Antonov, J.I., Baranova, O.K., Coleman, C., Garcia, H.E., Grodsky, A., et al. (2013) World Ocean Database 2013 Technical Ed. Mishonov,A. (ed) Sydney Levitus.

Chisholm, S.W. and Brand, L.E. (1981) Persistence of cell division phasing in marine phytoplankton in continuous light after entrainment to light: dark cycles. *J. Exp. Mar. Bio. Ecol.* **51**: 107–118.

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

Daugbjerg, N. and Moestrup, Ø. (1992) Ultrastructure of *Pyramimonas cyrtoptera* sp.nov. (Prasinophyceae), a species with 16 flagella from northern Foxe Basin, Arctic Canada, including observations on growth rates. *Can. J. Bot.* **70**: 1259–1273.

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**: 2194–200.

Faria, D.G., Kato, A., de la Peña, M.R., and Suda, S. (2011) Taxonomy and phylogeny of *Nephroselmis clavistella* sp. nov. (Nephroselmidophyceae, Chlorophyta). *J. Phycol.* **47**: 1388–1396.

Faria, D.G., Kato, A., and Suda, S. (2012) *Nephroselmis excentrica* sp. nov. (Nephroselmidophyceae, Chlorophyta) from Okinawa-jima, Japan. *Phycologia* **51**: 271–282.

Fučíková, K., Leliaert, F., Cooper, E.D., Škaloud, P., D'Hondt, S., Clerck De, O., et al. (2014) New phylogenetic hypotheses for the core Chlorophyta based on chloroplast sequence data. *Front. Ecol. Evol.* **2**: 63.

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.

Gomez, F. and Gorsky, G. (2003) Annual microplankton cycles in Villefranche Bay, Ligurian Sea, NW Mediterranean. *J. Plankton Res.* **25**: 323–339.

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.

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

Harðardottir, S., Lundholm, N., Moestrup, Ø., and Nielsen, T.G. (2014) Description of *Pyramimonas diskoicola* sp. nov. and the importance of the flagellate *Pyramimonas* (Prasinophyceae) in Greenland sea ice during the winter – spring transition. *Polar Biol.* 1479–1494.

Harzi, A.M., Tackx, M., Daro, M.H., Kesaulia, I., Caturao, R., and Podoor, N. (1998) Winter distribution of phytoplankton and zooplankton around some sandbanks of the Belgian coastal zone. *J. Plankton Res.* **20**: 2031–2052.

Jenkinson, I.R. (1986) *Halosphaera viridis, Ditylum brightwellii* and other phytoplankton in the north-eastern North Atlantic in spring: Sinking, rising and relative abundance. *Ophelia* **26**: 233–253.

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**: 1647–9.

Kimor, B. and Wood, E.J.F. (1975) A plankton study in the eastern Mediterranean Sea. *Mar. Biol.* **29**: 321–333.

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

Lee, H.-J. and Hur, S.-B. (2009) Genetic relationships among multiple strains of the genus *Tetraselmis* based on partial 18S rDNA sequences. *Algae* **24**: 205–212.

Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F., and De Clerck, O. (2012) Phylogeny and molecular evolution of the green algae. *CRC. Crit. Rev. Plant Sci.* **31**: 1–46.

Leliaert, F., Tronholm, A., Lemieux, C., Turmel, M., DePriest, M.S., Bhattacharya, D., et al. (2016) Chloroplast phylogenomic analyses reveal the deepest-branching lineage of the Chlorophyta, Palmophyllophyceae class. nov. *Sci. Rep.* **6**: 25367.

Limardo, A.J., Sudek, S., Choi, C.J., Poirier, C., Rii, Y.M., Blum, M., et al. (2017) Quantitative biogeography of picoprasinophytes establishes ecotype distributions and significant contributions to marine phytoplankton. *Environ. Microbiol.* **19**: 3219–3234.

Lin, Y.-C., Chung, C.-C., Chen, L.-Y., Gong, G.-C., Huang, C.-Y., and Chiang, K.-P. (2016) Community Composition of Photosynthetic Picoeukaryotes in a Subtropical Coastal Ecosystem, with Particular Emphasis on Micromonas. *J. Eukaryot. Microbiol.* **64**: 349–359.

Lopes dos Santos, A., Gourvil, P., Tragin, M., Noël, M.-H., Decelle, J., Romac, S., and Vaulot, D. (2016) Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. *ISME J.* **11**: 512–528.

Lopes dos Santos, A., Pollina, T., Gourvil, P., Corre, E., Marie, D., Garrido, J.L., et al. (2017) Chloropicophyceae, a new class of picophytoplanktonic prasinophytes. *Sci. Rep.* **7**: 14019.

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

Majaneva, M., Rintala, J.M., Piisilä, M., Fewer, D.P., and Blomster, J. (2012) Comparison of wintertime eukaryotic community from sea ice and open water in the Baltic Sea, based on sequencing of the 18S rRNA gene. *Polar Biol.* **35**: 875–889.

Margulis, L., Barghoorn:, E.S., Ashendorf, D., Banerjee, S., Chase, D., Francis, S., et al. (1980) The microbial community in the layered sediments at laguna Figueroa, Baja Californmia Mexico: Does it have Precambrian analogues? *Precambrian Res.* **11**: 93–123.

Marie, D., Zhu, F., Balagué, V., Ras, J., Phine, Vaulot, D., et al. (2006) Eukaryotic picoplankton communities of the Mediterranean Sea in summer assessed by molecular approaches (DGGE, TTGE, QPCR). *FEMS Microbiol. Ecol.* **55**: 403–415.

Marshall, H.G. (1980) Seasonal Phytoplankton Composition in the Lower Chesapeake Bay and Old Plantation Creek, Cape Charles, Virginia. *Estuaries* **3**: 207.

Masquelier, S., Foulon, E., Jouenne, F., Ferréol, M., Brussaard, C.P.D., and Vaulot, D. (2011) Distribution of eukaryotic plankton in the English Channel and the North Sea in summer. *J. Sea Res.* **66**: 111–122.

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.

Moestrup, Ø. and Hill, D.R.A. (1991) Studies on the genus *Pyramimonas* (Prasinophyceae) from Australian and

European waters: *P. propulsa* sp. nov. and *P. mitra* sp. nov. *Phycologia* **30**: 534–546.

Moro, I., La Rocca, N., Dalla Valle, L., Moschin, E., Negrisolo, E., and Andreoli, C. (2002) *Pyramimonas australis* sp. nov. (Prasinophyceae, Chlorophyta) from Antarctica: fine structure and molecular phylogeny. *Eur. J. Phycol.* **37**: 103–114.

Nakayama, T., Marin, B., Kranz, H.D., Surek, B., Huss, V. a, Inouye, I., and Melkonian, M. (1998) The basal position of scaly green flagellates among the green algae (Chlorophyta) is revealed by analyses of nuclear-encoded SSU rRNA sequences. *Protist* **149**: 367–80.

Not, F., del Campo, J., Balagué, V., de Vargas, C., and Massana, R. (2009) New insights into the diversity of marine picoeukaryotes. *PLoS One* **4**: e7143.

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**: 4064–72.

Not, F., Latasa, M., Scharek, R., Viprey, M., Karleskind, P., Balagué, V., et al. (2008) Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. *Deep Sea Res. Part I Oceanogr. Res. Pap.* **55**: 1456–1473.

Not, F., Massana, R., Latasa, M., Marie, D., Colson, C., Eikrem, W., et al. (2005) Late summer community composition and abundance of photosynthetic picoeukaryotes in Norwegian and Barents Seas. *Limnol. Oceanogr.* **50**: 1677–1686.

Not, F., Simon, N., Biegala, I., and Vaulot, D. (2002) Application of fluorescent in situ hybridization coupled with tyramide signal amplification (FISH-TSA) to assess eukaryotic picoplankton composition. *Aquat. Microb. Ecol.* **28**: 157–166.

Not, F., Valentin, K., Romari, K., Lovejoy, C., Massana, R., Töbe, K., et al. (2007) Picobiliphytes: A Marine Picoplanktonic Algal Group with Unknown Affinities to Other Eukaryotes. *Science (80-.).* **315**: 253–255.

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

Pienaar, R.N. and Sym, S.D. (2002) The genus *Pyramimonas* (Prasinophyceae) from southern African inshore waters. *South African J. Bot.* **68**: 283–298.

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.

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.

Samanidou, V., Fytianos, K., Vasilikiotis, G., Marino, D., Mazzochi, M.G., Modigh, M., et al. (1987) Distribution of nutrients in the Thermaikos Gulf, Greece. *Sci. Total Environ.* **65**: 181–189.

Sarno, D., Zingone, A., Saggiomo, V., and Carrada, G.C. (1993) Phytoplankton biomass and species composition in a Mediterranean coastal lagoon. *Hydrobiologia* **271**: 27–40.

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**: 7537–41.

Shi, X.L., Lepère, C., Scanlan, D.J., and Vaulot, D. (2011) Plastid 16S rRNA gene diversity among eukaryotic picophytoplankton sorted by flow cytometry from the South Pacific Ocean. *PLoS One* **6**: e18979.

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

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

Suda, S., Bhuiyan, M.A.H., and Faria, D.G. (2013) Genetic diversity of *Pyramimonas* from Ryukyu Archipelago, Japan (Chlorophyceae, Pyramimonadales). *J. Mar. Sci. Technol.* **21**: 285–296.

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.

Tragin, M., Zingone, A., and Vaulot, D. (2017) 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. *Environ. Microbiol.* **in press**:

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., et al. (2015) Eukaryotic plankton diversity in the sunlit ocean. *Science (80-.).* **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.

Wiebe, P.H., Remsen, C.C., and Vaccaro, R.F. (1974) *Halosphaera viridis* in the Mediterranean sea: size range, vertical distribution, and potential energy source for deep-sea benthos. *Deep Sea Res. Oceanogr. Abstr.* **21**: 657–667.

Worden, A.Z., Nolan, J.K., and Palenik, B. (2004) Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. Limnol. Oceanogr., 49(1), 2004, 168–179. *Limnol. Ocean.* **49**: 168–179.

Wu, W., Huang, B., Liao, Y., and Sun, P. (2014) Picoeukaryotic diversity and distribution in the subtropical-tropical South China Sea. *FEMS Microbiol. Ecol.* **89**: 563–579.

Yamaguchi, H., Nakayama, T., and Inouye, I. (2013) Proposal of *Microsquama subgen* . nov . for *Nephroselmis pyriformis* (Carter) Ettl. *Phycol. Res.* 268–269.

Yamaguchi, H., Suda, S., Nakayama, T., Pienaar, R.N., Chihara, M., and Inouye, I. (2011) Taxonomy of *Nephroselmis viridis* sp. nov. (Nephroselmidophyceae, Chlorophyta), a sister marine species to freshwater *N. olivacea*. *J. Plant Res.* **124**: 49–62.

Zingone, A., Throndsen, J., and Forlani, G. (1995) Pyramimonas oltmannsii (Prasinophyceae) reinvestigated. *Phycologia* **34**: 241–249.

List of Figures

List of Supplementary Tables

Table S1: OSD stations with number of photosynthetic and Chlorophyta reads, fraction of Chlorophyta reads, number of OTUs and metadata.

List of Supplementary figures

Fig. S1: Slope at the extremity of the Chlorophyta OTU (99%) rarefaction curve for OSD surface stations. Stations with less than 100 Chlorophyta reads (Table S1) are in italic. Color of the dots refers to oceanic region: Atlantic Ocean – red, Pacific Ocean – ocher, Mediterranean Sea – blue.

Fig S2: Contribution of the 8 minor classes of Chlorophyta at OSD stations in surface. Stations where a given class was not recorded are represented by blue crosses. The circle surface is proportional to the percent of class versus Chlorophyta reads.

Fig. S3: Number of OSD surface stations were more than 1% of the Chlorophyta classes was recovered.

List of Supplementary data (available on Figshare)

The data are deposited on Figshare at https://figshare.com/s/36f84f77790367e48e1c

Data S1: Mothur script for sequence analysis.

Data S2: Chlorophyta OTU table: number of reads assigned to each OTUs per OSD 2014 stations.

Data S3: Fasta file of OSD 2014 Chlorophyta OTUs representative sequences.

Data S4: Taxonomy file of OSD 2014 Chlorophyta OTUs representative sequences.

Supplementary Figures

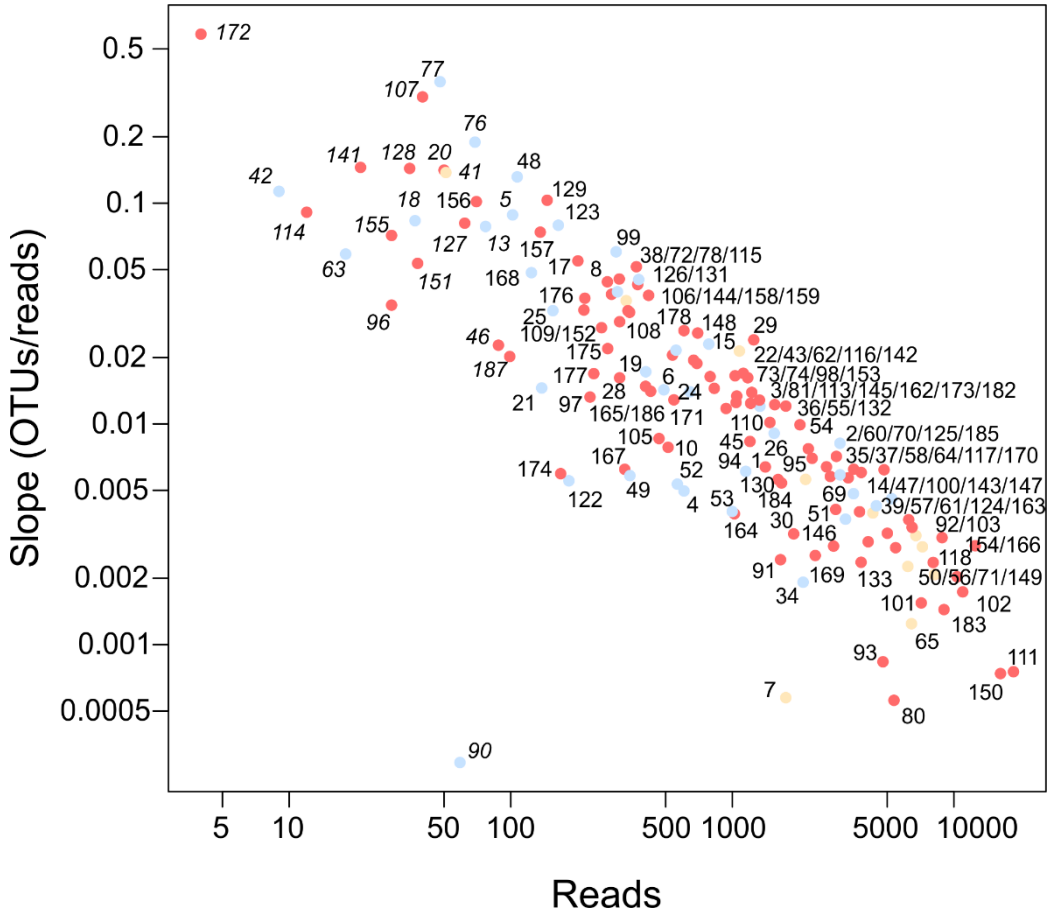


Fig. S1: Slope at the extremity of the Chlorophyta OTU (99%) rarefaction curve for OSD surface stations. Stations with less than 100 Chlorophyta reads (Table S1) are in italic. Color of the dots refers to oceanic region: Atlantic Ocean – red, Pacific Ocean – ocher, Mediterranean Sea – blue.

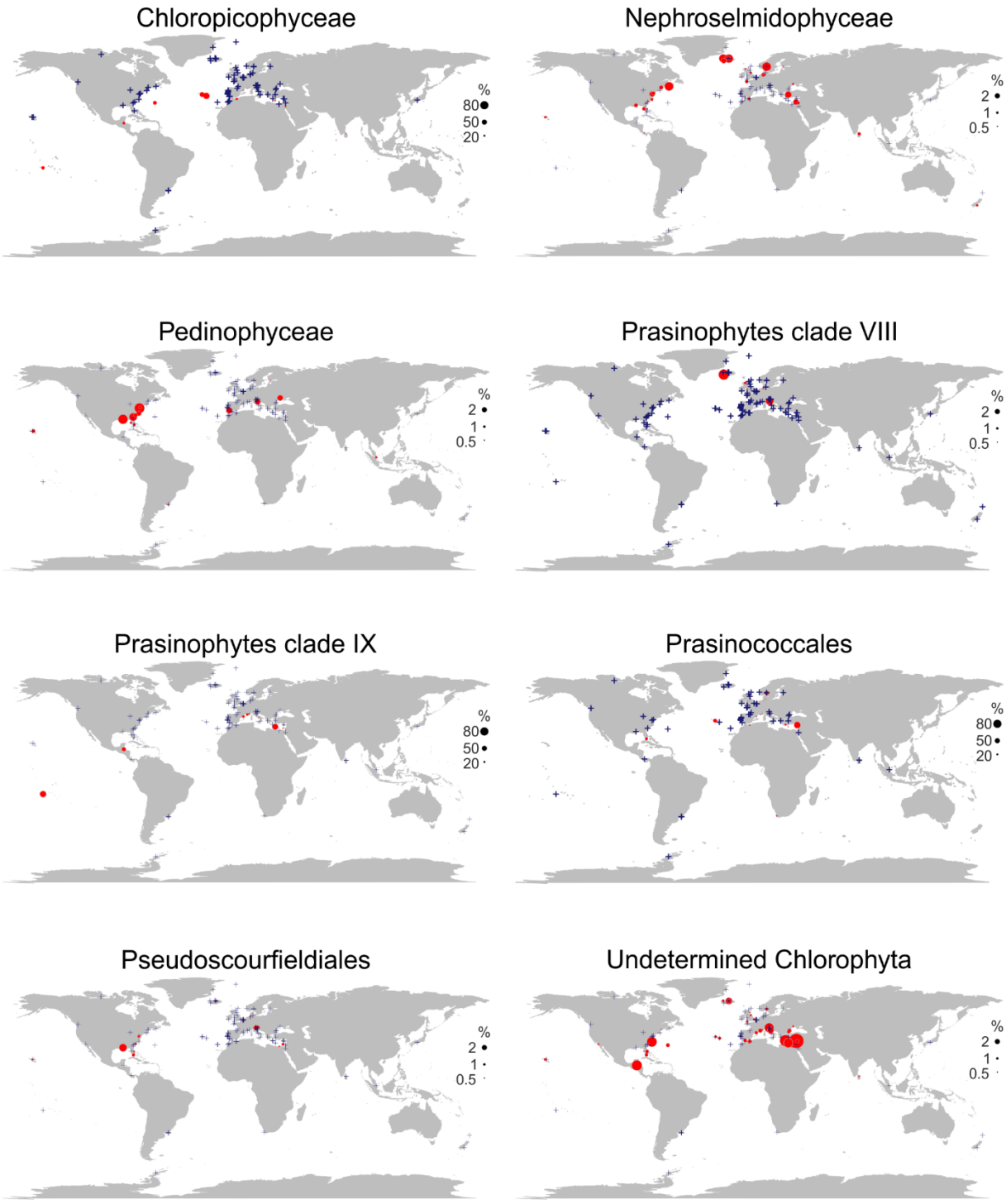


Fig S2: Contribution of the 8 minor classes of Chlorophyta at OSD stations in surface. Stations where a given class was not recorded are represented by blue crosses. The circle surface is proportional to the percent of class versus Chlorophyta reads.

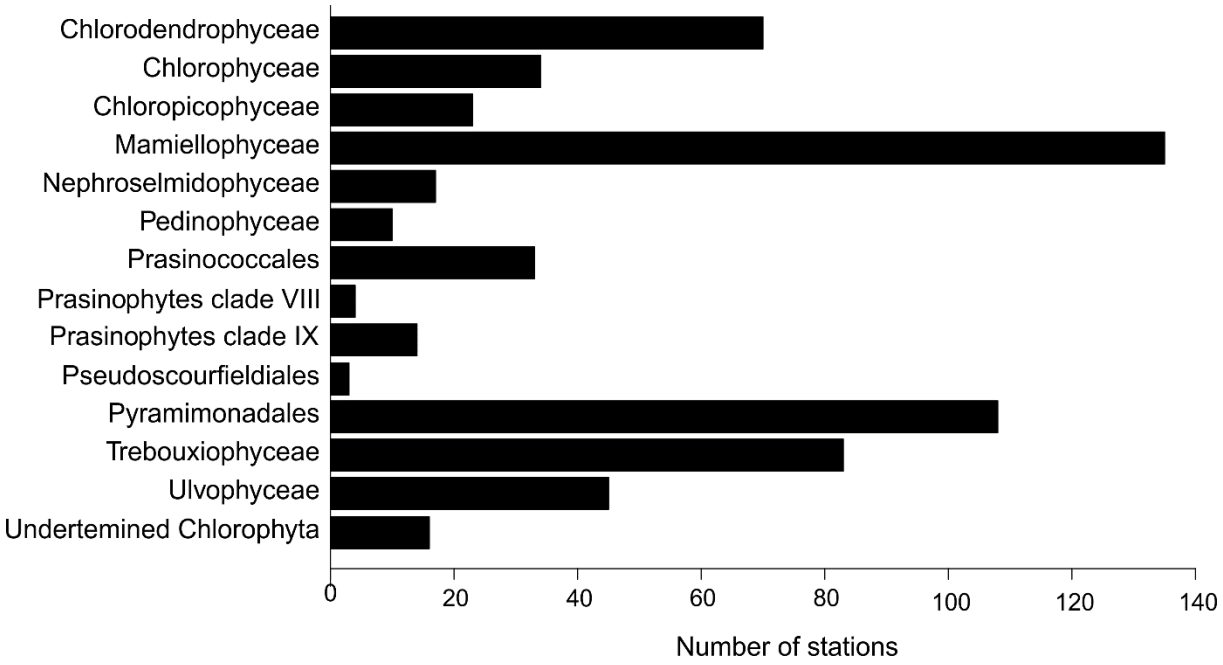


Fig. S3: Number of OSD surface stations were more than 1% of the Chlorophyta classes was recovered.

Supplementary Table

Table S1: OSD stations with number of photosynthetic and Chlorophyta reads, fraction of Chlorophyta reads, number of OTUs and metadata.

OSD sample code	OSD sta.	Stations	Ocean	Region	Photosynt reads	Chloroph reads	Chloro phyta %	Chloro phyta OTUs	Latitude	Longitude	Distance to the coast (km)	Water Temperature (°C)	Salinity (PSU)	Nitrates (µM)	PO_4^{2-} (µM)	Chl*a* *(monthly)*
OSD1	1	Plymouth - L4	Atl. Ocean	English Channel	2450	1403	57.3	23	50.15	-4.13	16.79	16.66	35.22	0.13	0.16	0.84
OSD2	2	Roscoff - SOMLIT	Atl. Ocean	English Channel	9631	2660	27.6	50	48.78	-3.94	5.62	14.38	35.13	1.10	0.11	5.19
OSD3	3	Helgoland	North Sea		14537	1224	8.42	40	54.18	7.90	0.15	14.00	32.59	11.68	0.19	6.97
OSD4	4	LTER-MC	Med. Sea	Tyrrhenia n Sea	4590	605	13.2	16	40.81	14.25	3.28	23.30	36.88	0.22	0.23	6.64
OSD5	*5*	*Crete*	Med. Sea	*Aegean Sea*	*458*	*92*	*20.1*	*21*	*35.66*	*24.99*	*27.15*	*24.46*	*39.29*	*0.54*	*0.04*	*0.09*
OSD6	6	Blanes	Med. Sea	Balearic Sea	1512	487	32.2	31	41.67	2.80	0.24	20.66	37.83	0.00	0.29	0.29
OSD7	7	Moorea - Tiahura	Pacific Ocean		1849	1743	94.3	20	-17.29	-149.54	26.55	27.00	37.00			0.10
OSD8	8	BATS	Atl. Ocean		909	270	29.7	31	32.16	-64.50	22.19					
OSD10	10	Lake Erie W4			9748	513	5.26	41	41.84	-83.19	467.15	22.40	0.14	1.02	0.41	18.54
OSD13	*13*	*Varna Bay*	*Black Sea*		*5593*	*76*	*1.36*	*19*	*43.18*	*27.91*	*1.37*	*22.87*	*12.60*	*2.26*	*0.39*	*33.69*
OSD14	14	Banyuls	Med. Sea	Western Basin	8386	5234	62.4	59	42.49	3.15	0.39	20.42	37.83	0.00	0.01	0.52
OSD15	15	Villefranche - SOMLIT	Med. Sea	Adriatic Sea	1791	781	43.6	45	43.69	7.32	0.04	21.78	37.97	0.12	0.26	0.19
OSD17	17	VLIZ	North Sea		7880	201	2.55	23	51.43	2.81	21.37	16.68	34.03	8.56	0.57	9.72
OSD18	*18*	*Kyrenia*	Med. Sea	*Eastern Basin*	*725*	*33*	*4.55*	*11*	*35.36*	*33.29*	*2.35*	*20.70*	*38.50*	*0.12*	*0.10*	*0.11*
OSD19	19	Famagusta	Med. Sea	Eastern Basin	768	402	52.3	25	35.19	33.90	0.21	29.00	39.10	0.03	0.01	
OSD20	*20*	*Faxafloi*	Atl. Ocean	*Iceland*	*285*	*50*	*17.5*	*17*	*64.21*	*-22.02*	*3.76*	*11.00*	*31.20*	*1.41*	*0.19*	*9.91*
OSD21	21	Croatia	Med. Sea	Adriatic Sea	530	138	26.0	12	45.08	13.61	1.35	19.50	37.10	1.78	0.03	0.62
OSD22	22	Marseille Solemio SOMLIT	Med. Sea	Western Basin	4816	553	11.5	60	43.23	5.75	7.64	22.00	38.09	0.12	0.26	0.59
OSD24	24	Marchica	Med. Sea	Alboran Sea	2770	644	23.3	27	35.19	-2.88	2.62	26.50	36.00	0.10	0.00	0.52
OSD25	25	Saidia Rocher	Med. Sea	Alboran Sea	368	153	41.6	20	35.09	-2.21	0.57	23.10	30.00	0.10	0.00	0.31
OSD26	26	Tangier	Atl. Ocean	Strait of Gibraltar	8509	1530	18.0	53	35.82	-5.75	0.88	28.00	34.19	0.10	0.04	1.18
OSD28	28	Belize	Atl. Ocean	Caribbean Sea	505	288	57.0	37	16.80	-88.08	15.86	29.50	35.00		0.00	0.21
OSD29	29	Florida	Atl. Ocean		2935	1239	42.2	74	27.47	-80.28	0.30	26.90	35.70	2.00	0.05	1.17
OSD30	30	Tvärminne	North Sea	Gulf of Finland	2948	1892	64.2	43	59.88	23.25	1.96	10.50	5.63	0.09	0.08	34.33
OSD34	34	Alexandria	Red Sea		5227	2085	39.9	33	31.22	29.97	3.72	27.00	36.00		0.01	3.89
OSD35	35	Cheasapeake Bay	Atl. Ocean		9889	3818	38.6	78	38.68	-76.17	0.86	26.32	8.97	0.18	0.16	35.67
OSD36	36	Delaware	Atl. Ocean		14733	1554	10.5	47	39.33	-75.47	0.39	23.55	7.42	0.09	0.18	11.23
OSD37	37	Port Everglades	Atl. Ocean		21490	2769	12.9	60	26.10	-80.09	1.72	27.70	33.82		0.05	0.14
OSD38	38	Long Key	Atl. Ocean		2461	369	15.0	50	24.74	-80.78	7.32	29.60	36.25		0.08	4.53
OSD39	39	Charleston Harbor	Atl. Ocean		9231	6249	67.7	62	32.75	-79.90	0.38	31.30	24.30	0.08	0.12	9.31
OSD41	*41*	*Sequim Bay Park*	*Pacific Ocean*	*Southeast Alaska*	*5290*	*51*	*0.964*	*12*	*48.04*	*-123.03*	*0.55*	*15.90*	*24.73*	*4.26*	*0.86*	*4.77*
OSD42	*42*	*Faro Lake*	Med. Sea	*Tyrrhen. Sea*	*81*	*9*	*11.1*	*4*	*38.27*	*15.64*	*0.81*	*20.00*	*36.80*	*0.10*	*0.16*	*0.15*
OSD43	43	SIO Pier	Pacific Ocean		3371	1073	31.8	67	32.87	-117.26	0.00	20.04	33.54	0.10	0.33	0.97
OSD45	45	Tampa Bay	Atl. Ocean	Gulf of Mexico	7197	1201	16.7	25	27.62	-82.73	5.67	31.20	33.84	2.00	0.00	11.03
OSD46	*46*	*Horn Island*	*Atl. Ocean*	*Gulf of Mexico*	*1493*	*88*	*5.89*	*10*	*30.25*	*-88.75*	*0.41*	*29.80*	*17.50*	*0.00*	*0.10*	*18.75*

OSD sample code	OSD sta.	Stations	Ocean	Region	Photosynt reads	Chlorophyta reads	Chlorophyta %	Chlorophyta OTUs	Latitude	Longitude	Distance to the coast (km)	Water Temperature (°C)	Salinity (PSU)	Nitrates (µM)	PO_4^{2-} (µM)	Chl*a* *(monthly)*
OSD47	47	Venice Lagoon	Med. Sea	Adriatic Sea	26538	4465	16.8	44	45.50	12.42	0.42	25.30	27.42	1.78	0.03	4.83
OSD48	48	Venice Gulf	Med. Sea	Adriatic Sea	2457	100	4.07	33	45.41	12.53	5.98	22.20	33.14	1.78	0.03	6.13
OSD49	49	Vida	Med. Sea	Adriatic Sea	3075	345	11.2	31	45.33	13.33	16.74	22.00	34.00	1.78	0.03	1.85
OSD50	50	Pasaia	Atl. Ocean		9050	8178	90.4	31	43.33	-1.93	0.51	20.00	34.30	0.12	0.26	0.51
OSD51	51	Bocas del Toro	Atl. Ocean	Caribbean Sea	4740	2928	61.8	44	9.35	-82.27	0.85	29.10	34.60	1.35	0.53	2.62
OSD52	52	Abu Hashish	Red Sea		1234	527	42.7	34	33.91	27.03	157.22	27.00	38.33		0.23	0.13
OSD53	53	Ras Disha	Red Sea		3022	935	30.9	43	33.91	27.04	158.63	27.28	38.35		0.23	0.13
OSD54	54	Maine Booth Bay	Atl. Ocean		15397	2017	13.1	85	43.84	-69.64	0.10	11.90	31.00		0.02	
OSD55	55	Maine Damariscotta River	Atl. Ocean		14876	1742	11.7	61	43.86	-69.58	0.16	12.50	32.00	1.00		36.40
OSD56	56	Hawaii Kakaako	Pacific Ocean		13937	6195	44.4	51	21.29	-156.86	13.41	26.06	35.00	0.01	0.08	0.09
OSD57	57	Hawaii Oahu	Pacific Ocean		10382	6731	64.8	82	21.29	-157.84	0.57	27.58	34.00	0.01	0.08	0.08
OSD58	58	PICO	Atl. Ocean		23151	3264	14.1	76	34.72	-76.67	1.37	25.80	35.80		0.05	2.73
OSD60	60	South Carolina 2 - North Inlet	Atl. Ocean		16156	2943	18.2	64	33.32	-79.17	0.28	27.77	35.13	0.10	0.35	9.14
OSD61	61	Vineyard Sound	Atl. Ocean		9409	6474	68.8	71	41.52	-70.67	0.20	19.20	30.70	1.02	0.41	7.52
OSD62	62	Manai Straits	Atl. Ocean	Irish Sea	6101	692	11.3	37	53.23	-4.16	0.03	16.00	34.00	4.82	0.45	
OSD63	*63*	*Venice Acqua Alta*	Med. Sea	*Adriatic Sea*	*48*	*18*	*37.5*	*5*	*45.31*	*12.51*	*13.20*	*21.78*	*32.77*	*1.78*	*0.03*	*7.77*
OSD64	64	Odessa	Black Sea		7514	3054	40.6	94	46.44	30.78	1.11	20.30	17.63	0.73	0.45	4.85
OSD65	65	Leigh Marine Laboratory (NZ)	Pacific Ocean		10958	6434	58.7	52	-36.29	174.82	0.10	16.00	34.00	0.20	0.15	0.82
OSD69	69	Marghera	Med. Sea	Adriatic Sea	23153	3521	15.2	44	45.46	12.26	0.53	25.70	29.41	1.78	0.03	9.14
OSD70	70	Lido	Med. Sea	Adriatic Sea	11061	3057	27.6	72	45.41	12.44	0.59	23.40	31.67	1.78	0.03	11.14
OSD71	71	Otago	Pacific Ocean		15003	8174	54.5	58	-45.74	170.77	4.06	10.99	35.20		1.55	
OSD72	72	Boknis Eck	North Sea	Baltic Sea	5940	374	6.30	42	54.83	10.00	3.49	13.99	14.25	11.68	0.19	47.88
OSD73	73	Lima Estuary	Atl. Ocean		2248	1029	45.8	41	41.68	-8.83	0.58	18.40	32.30	0.71	0.07	2.39
OSD74	74	Douro Estuary	Atl. Ocean		25158	1174	4.67	55	41.14	-8.67	0.70	20.20	13.75	0.71	0.07	1.18
OSD76	*76*	*Foglia*	Med. Sea	*Adriatic Sea*	*1975*	*69*	*3.49*	*28*	*43.95*	*12.94*	*3.29*	*23.61*	*27.78*	*0.77*	*0.04*	*5.23*
OSD77	*77*	*Metauro*	Med. Sea	*Adriatic Sea*	*1761*	*48*	*2.73*	*25*	*43.85*	*13.07*	*2.67*	*24.10*	*26.29*	*1.45*	*0.05*	*3.99*
OSD78	78	CONISMA	Med. Sea	Adriatic Sea	5465	366	6.70	59	43.57	13.60	0.00	24.25	34.33	0.77	0.04	2.49
OSD80	80	Young Sound	Arctic Ocean	Greenland Sea	13324	5364	40.3	11	74.31	-20.30	3.26	-0.10	5.00	0.03	0.22	
OSD81	81	Ria Formosa Lagoon	Atl. Ocean		10662	1047	9.82	37	37.01	-7.97	0.08	22.20	34.30	0.40	0.09	1.04
OSD90	*90*	*Etoliko Lagoon*	Med. Sea	*Ionian Sea*	*2563*	*59*	*2.30*	*7*	*38.48*	*21.32*	*0.22*	*26.29*	*15.09*	*0.10*	*0.16*	
OSD91	91	Oualidiya	Atl. Ocean		6970	1650	23.7	22	32.75	-9.04	0.83	19.00	27.24	0.50	0.17	7.00
OSD92	92	Casablanca	Atl. Ocean		21589	12478	57.8	98	33.58	-7.70	0.05	24.00	30.75	0.20	0.12	21.13

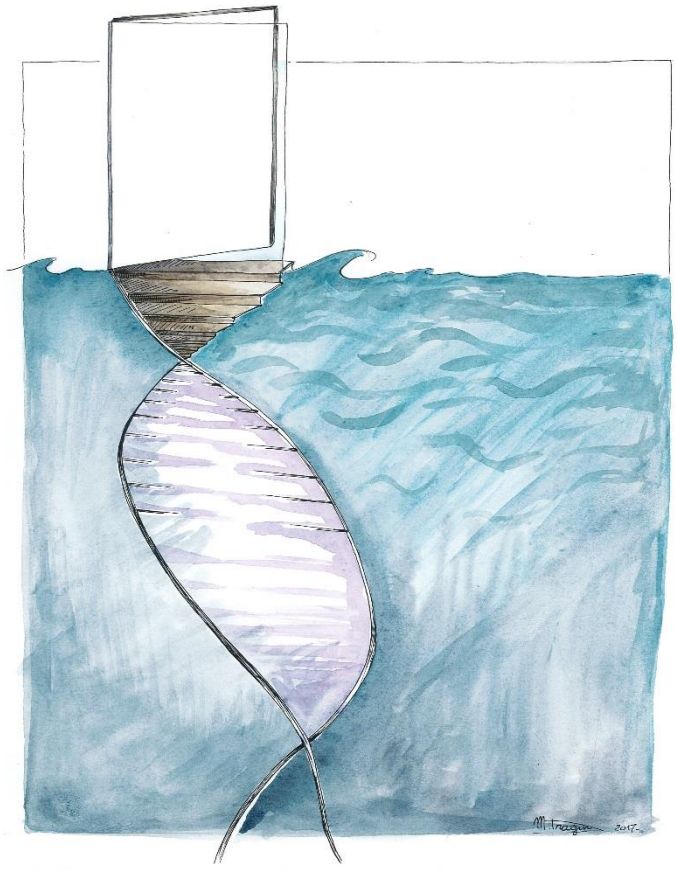
OSD sample code	OSD sta.	Stations	Ocean	Region	Photosynt reads	Chloroph yta reads	Chloro phyta %	Chloro phyta OTUs	Latitude	Longitude	Distance to the coast (km)	Water Temperature (°C)	Salinity (PSU)	Nitrates (µM)	PO_4^{2-} (µM)	Chl*a* *(monthly)*
OSD93	93	Eljadida	Atl. Ocean		7743	4783	61.8	15	33.26	-8.50	0.12	19.00	32.88	0.20	0.12	6.31
OSD94	94	Saidia Marina	Med. Sea	Alboran Sea	3314	1145	34.5	59	35.09	-2.21	0.57	23.60	31.00	0.10	0.00	0.31
OSD95	95	Singapore Indigo_V	Pacific Ocean	Singapore Strait	11320	2141	18.9	37	1.27	103.92	4.95	31.00	31.03	0.90	0.16	
OSD96	*96*	*Sao Miguel Azores I*	*Atl. Ocean*		*455*	*29*	*6.37*	*6*	*37.43*	*-25.32*	*32.81*	*18.50*	*35.00*			*0.12*
OSD97	97	Faial Azores	Atl. Ocean		632	228	36.1	18	38.53	-28.60	2.13	16.90	35.60	0.20	0.03	0.17
OSD98	98	Sao Jorge Azores	Atl. Ocean		3136	1319	42.1	40	38.64	-28.13	0.86	18.70	35.60	0.20	0.03	0.18
OSD99	99	C1	Med. Sea	Adriatic Sea	2063	298	14.4	34	45.70	13.71	1.07	20.82	33.93	1.78	0.03	7.15
OSD100	100	Crete - GOS	Med. Sea	Aegean Sea	4870	3240	66.5	45	35.35	25.29	1.06	24.21	39.05	0.17	0.04	0.13
OSD101	101	Quinta do Lorde	Atl. Ocean		8185	7123	87.0	54	32.74	-16.71	0.47	20.50	37.00			0.17
OSD102	102	Marina do Funchal	Atl. Ocean		18687	10960	58.6	65	32.65	-16.91	0.40	20.80	36.00			0.12
OSD103	103	Porto da Cruz	Atl. Ocean		10163	8843	87.0	77	32.77	-16.83	1.69	20.20	37.00			0.17
OSD105	105	Cambridge Bay, Nunavut	Arctic Ocean		861	467	54.2	16	69.02	-105.34	3.74	-0.72	26.91	2.10	0.19	
OSD106	106	REYKIS	Atl. Ocean	Iceland	5879	342	5.82	36	65.94	-22.42	1.62	7.50	29.20	1.53	0.21	
OSD107	*107*	*Lisboa*	*Atl. Ocean*		*1735*	*40*	*2.31*	*21*	*39.14*	*-9.38*	*0.05*	*20.20*	*30.00*	*0.52*	*0.05*	*0.70*
OSD108	108	Alcochete	Atl. Ocean		4803	310	6.45	16	38.76	-8.97	0.36	20.10	30.00	0.95	0.28	
OSD109	109	Rosario	Atl. Ocean		3352	214	6.38	25	38.68	-9.01	3.82	20.50	30.00	0.95	0.28	
OSD110	110	Figueira da Foz	Atl. Ocean		7270	1478	20.3	47	40.15	-8.87	1.13	23.00	22.50	2.55	0.07	1.39
OSD111	111	Ria de Aveiro_1	Atl. Ocean		37249	18569	49.8	50	40.66	-8.70	0.65	25.20	30.00	0.10	0.24	0.17
OSD113	113	Cascais Watch	Atl. Ocean		1452	794	54.7	37	38.67	-9.44	4.38	18.00	35.27	0.95	0.28	4.72
OSD114	*114*	*Berlengas Watch*	*Atl. Ocean*		*20*	*12*	*60.0*	*5*	*39.41*	*-9.51*	*9.22*	*18.50*	*33.57*	*0.05*		*0.10*
OSD115	115	Santa Cruz	Atl. Ocean		2707	308	11.4	26	39.13	-9.38	0.10	20.30	40.00	0.52	0.05	0.70
OSD116	116	Lagoa de Obidos	Atl. Ocean		11526	668	5.80	27	39.42	-9.22	0.74	24.70	22.50	0.52	0.05	1.66
OSD117	117	Tavira Beach	Atl. Ocean		6934	4837	69.8	81	37.17	-7.50	0.75	23.64	37.93	0.40	0.09	3.37
OSD118	118	Lough Hyne	Atl. Ocean	Celtic Sea	12052	8062	66.9	72	51.74	-8.31	0.67	18.00	38.00	0.55	0.06	1.86
OSD122	122	Station A Gulf Of Eilat	Red Sea	Gulf of Aqaba	1072	183	17.1	15	29.47	34.93	3.46	24.00	40.50		0.21	0.15
OSD123	123	Shikmona	Med. Sea	Eastern Basin	718	164	22.8	33	32.82	32.95	164.24	27.00	39.40	0.16	0.07	0.07
OSD124	124	Osaka Bay	Pacific Ocean	Japan Sea	11412	7207	63.1	53	34.32	135.12	0.37	21.15	33.19	1.19	0.21	19.62
OSD125	125	Cullercoats Beach	North Sea		8894	2284	25.7	62	55.03	-1.43	0.22	16.04	26.40	0.43	0.17	6.68
OSD126	126	Eyafjordur_1	Arctic Ocean	Greenland Sea	5470	284	5.19	24	66.01	-18.20	3.99	12.20	42.30	1.30	0.15	
OSD127	*127*	*Eyafjordur_2*	*Arctic Ocean*	*Greenland Sea*	*1917*	*62*	*3.23*	*13*	*66.01*	*-18.20*	*4.11*	*11.24*	*59.70*	*1.55*	*0.24*	
OSD128	*128*	*Eyafjordur_3*	*Arctic Ocean*	*Greenland Sea*	*16057*	*34*	*0.21*	*10*	*65.49*	*-18.06*	*14.49*	*12.00*	*14.00*	*3.15*	*0.33*	
OSD129	129	Eyafjordur_4	Arctic Ocean	Greenland Sea	8778	141	1.61	33	65.82	-18.10	1.95	9.90	61.50	3.15	0.33	
OSD130	130	Eyafjordur_5	Arctic Ocean	Greenland Sea	6775	1609	23.7	43	66.13	-18.79	0.92	10.10	48.00	1.55	0.24	1.02

OSD sample code	OSD sta.	Stations	Ocean	Region	Photosynt reads	Chlorophyta reads	Chlorophyta %	Chlorophyta OTUs	Latitude	Longitude	Distance to the coast (km)	Water Temperature (°C)	Salinity (PSU)	Nitrates (µM)	PO_4^{2-} (µM)	Chl*a* *(monthly)*
OSD131	131	Zlatna ribka	Black Sea		2676	301	11.2	28	42.24	27.40	20.26				0.09	14.09
OSD132	132	Sdot YAM	Med. Sea	Eastern Basin	3304	1324	40.1	69	32.07	34.84	7.64	27.30	39.35	0.20	0.05	
OSD133	133	Robben Island	Atl. Ocean		6952	3812	54.8	35	-33.90	18.39	1.50	15.06	35.16	3.57	0.81	3.25
OSD141	*141*	*Raunefjorden*	*North Sea*		*272*	*21*	*7.72*	*8*	*60.16*	*5.12*	*2.97*	*10.13*	*30.67*		*0.11*	
OSD142	142	Gray's Reef National Marine Sanctuary	Atl. Ocean		734	537	73.2	28	31.38	-80.87	26.61	27.43	35.85	0.02	0.16	3.23
OSD143	143	Skidaway Institute of Oceanog.	Atl. Ocean		27114	3751	13.8	52	-64.50	-81.02	371.59	30.79	26.58	0.02	0.16	
OSD144	144	Maunalua Bay O'ahu	Pacific Ocean		914	330	36.1	28	21.27	-157.72	1.14	25.80	35.00	0.01	0.08	0.07
OSD145	145	North Sea - Blankenberge	North Sea		14033	1210	8.62	40	51.36	3.12	3.31	17.00	100.00	8.56	0.57	6.82
OSD146	146	Fram Strait	Arctic Ocean	Greenland Sea	6054	2863	47.3	28	78.45	-2.83	306.87	-1.61	33.78	1.50	0.26	
OSD147	147	Rajarata	Indian Ocean	Bay of Bengal	9636	4295	44.6	70	8.52	81.05	7.96	28.80	33.00	0.00	0.27	0.23
OSD148	148	Wadden Sea	North Sea		6982	698	9.99	43	53.58	8.15	1.54	17.77	31.14	1.27	0.09	10.10
OSD149	149	Laguna Rocha Norte	Atl. Ocean		17462	10340	59.2	52	-34.37	-54.16	22.95	10.98	14.30	9.00	0.39	52.78
OSD150	150	Laguna Rocha Sur	Atl. Ocean		19853	16225	81.7	44	-34.68	-54.28	0.57	9.99	18.00	9.00	0.39	75.00
OSD151	*151*	*South Atlantic Microbial Observatory*	*Atl. Ocean*		*847*	*38*	*4.49*	*10*	*-34.42*	*-54.16*	*17.83*	*11.67*	*32.86*	*9.00*	*0.39*	*75.00*
OSD152	152	Compass Buoy Station	Atl. Ocean	Bedford Basin	1725	257	14.9	23	44.69	-63.64	0.90	12.80	28.90	0.05	0.36	
OSD153	153	Faro Island	Atl. Ocean		5804	1121	19.3	50	37.00	-7.97	0.51	21.10	34.40	0.33		1.04
OSD154	154	Arcachon-SOMLIT	Atl. Ocean	Bay of Biscay	10419	4100	39.4	42	44.67	-1.17	0.07	20.70	32.40	0.33	0.01	
OSD155	*155*	*Steilene Oslofjord*	*North Sea*	*Skaggerak*	*5092*	*29*	*0.56*	*8*	*59.82*	*10.60*	*0.29*	*18.70*	*30.05*	*0.12*	*0.08*	
OSD156	*156*	*Hvaler Tisler Site*	*North Sea*	*Skaggerak*	*2850*	*70*	*2.46*	*22*	*59.90*	*10.72*	*0.05*	*17.80*	*30.05*	*0.12*	*0.08*	
OSD157	157	ELLEIm2	North Sea	Skaggerak	23524	135	0.57	30	59.62	10.63	0.40	18.00	26.12	8.98	0.67	
OSD158	158	Sao Miguel Azores II	Atl. Ocean		2181	416	19.1	40	37.43	-25.19	33.87	19.20	35.70			0.11
OSD159	159	Brest-SOMLIT	Atl. Ocean		4575	338	7.39	23	48.36	-4.55	0.43	16.00	34.78	1.10	0.11	
OSD162	162	Stonehaven	North Sea		10325	828	8.02	36	56.96	-2.10	5.96	12.20	34.32	0.34	0.38	2.38
OSD163	163	Scapa	North Sea		9439	5007	53.0	52	58.96	-2.97	0.25	11.70	34.45	0.77	0.26	
OSD164	164	Scalloway	North Sea		1322	1022	77.3	27	60.14	-1.28	0.21	12.20	35.14	6.44	0.34	
OSD165	165	Loch Ewe	Atl. Ocean	West Coast of Scotland	2620	428	16.3	39	57.85	-5.65	1.80	14.30	32.45	5.01	0.15	4.82
OSD166	166	Armintza	Atl. Ocean		7201	5456	75.8	41	43.43	-2.90	0.78	17.88	35.05	0.12	0.12	0.28
OSD167	167	Eyafjordur_6	Arctic Ocean	Greenland Sea	6136	322	5.25	38	65.71	-18.12	0.16	11.00	1.10	3.46	0.32	
OSD168	168	IMST_izmir	Med. Sea	Aegean Sea	9500	124	1.31	17	38.41	27.03	0.46	25.74	38.30	0.11	0.06	2.58

OSD sample code	OSD sta.	Stations	Ocean	Region	Photosynt reads	Chloroph yta reads	Chloro phyta %	Chloro phyta OTUs	Latitude	Longitude	Distance to the coast (km)	Water Temperature (°C)	Salinity (PSU)	Nitrates (µM)	PO_4^{2-} (µM)	Chl*a* *(monthly)*
OSD169	169	Brightlingsea Creek, Essex	North Sea		3236	2358	72.9	40	51.80	1.01	0.86	18.50	35.20	8.56	0.57	6.11
OSD170	170	Belgium - 130	North Sea		18311	3335	18.2	53	51.27	2.90	2.80	18.62	32.26	8.56	0.57	8.52
OSD171	171	Belgium - 230	North Sea		11771	545	4.63	24	51.31	2.85	8.24	18.26	32.81	8.56	0.57	9.72
OSD172	*172*	*Belgium - 700*	*North Sea*		*54*	*4*	*7.41*	*4*	*51.37*	*3.22*	*2.64*	*18.48*	*32.34*	*8.56*	*0.57*	*16.20*
OSD173	173	Belgium - 710	North Sea		13322	937	7.03	41	51.44	3.14	10.89	18.18	33.13	8.56	0.57	8.70
OSD174	174	Belgium - 780	North Sea		9711	168	1.73	10	51.47	3.06	15.64	17.94	33.37	8.56	0.57	11.88
OSD175	175	ZG02	North Sea		3596	274	7.62	19	51.33	2.50	23.38	17.22	34.29	8.56	0.57	8.81
OSD176	176	Belgium - 215	North Sea		5388	216	4.01	20	51.28	2.61	13.74	18.12	33.37	8.56	0.57	15.49
OSD177	177	Belgium - 120	North Sea		3454	237	6.86	14	51.19	2.70	2.01	18.74	32.58	8.56	0.57	9.09
OSD178	178	Belgium - 435	North Sea		9379	605	6.45	39	51.58	2.79	35.40	16.76	34.50	8.56	0.57	5.48
OSD182	182	W08	North Sea		4507	1040	23.1	28	51.46	2.35	40.46	15.71	35.00	8.56	0.57	2.33
OSD183	183	W09	North Sea		12218	9018	73.8	43	51.75	2.70	54.07	15.80	35.01	8.56	0.57	1.32
OSD184	184	W10	North Sea		6084	1667	27.4	27	51.68	2.42	58.14	15.83	35.03	8.56	0.57	1.54
OSD185	185	Belgium - 421	North Sea		11886	2203	18.5	48	51.48	2.45	38.33	16.11	34.90	8.56	0.57	3.56
OSD186	186	SERC Rhode River Maryland			3322	406	12.2	30	38.89	-76.54	0.49	26.80	7.20	0.18	0.16	
OSD187	*187*	*Palmer station*	*South Ocean*	*Antarctica*	*14363*	*99*	*0.68*	*9*	*-64.77*	*-64.05*	*0.53*					

Chapter 4

Novel diversity within *Micromonas* and *Ostreococcus* (Mamiellophyceae) unveiled by metabarcoding analyses



Introduction

Mamiellophyceae consist of four orders: Mamiellales, Bathycoccales, Dolichomastigales and Monomastigales. The latter is confined to freshwater environments (Marin and Melkonian, 2010). Mamiellales and Bathycoccales host some of the most common Chlorophyta microalgae such as the ubiquitous *Micromonas*, the smallest known eukaryotes *Ostreococcus* or the coccoid *Bathycoccus* (Marin and Melkonian, 2010). Within Mamiellales, *Micromonas pusilla* (Butcher, 1952) was recently split into four species: *Micromonas bravo* (previous clade B.E.3), *Micromonas commoda* (previous clade A.ABC.1-2), *Micromonas polaris* (previous clade B arctic), *Micromonas pusilla* (previous clade C.D.5) and two clades mentioned as candidate species 1 (clade B._.4) and candidate species 2 (clade B warm) (Simon *et al.*, 2017). The history of clade nomenclature is detailed in Simon *et al.* (2017). In Bathycoccales, four *Ostreococcus* clades have been delineated with two species formerly described: *O. tauri* (Chrétiennot-Dinet *et al.*, 1995), *O. mediterraneus* (Subirana *et al.*, 2013), *"O. lucimarinus"* (clade A) and clade B, which both lack of formal taxonomic description (Guillou *et al.*, 2004). Pigment analyses of *Ostreococcus* strains allowed to distinguish two ecotypes (Rodríguez *et al.*, 2005) OI (corresponding to *O. tauri* and *"O. lucimarinus"*) and OII (corresponding to *Ostreococcus* clade B). These two ecotypes have been targeted by qPCR primers and probes (Demir-Hilton *et al.*, 2011).

As mentioned in the previous chapter, Mamiellophyceae represented 55% of the Chlorophyta reads found in the OSD2014 surface water, this Chlorophyta lineage did not show any geographic distribution patterns or environmental preferenda and were recovered in all coastal environments. We hypothesize than in order to detect distribution patterns this class should be investigated at lower taxonomic levels such as the species level. Fourteen Mamiellophyceae genera were recovered in OSD dataset (Fig.1) among them, *Ostreococcus* represented 45% (105,500 reads), *Micromonas* 34% (80,301 reads), *Bathycoccus* 10% (24,371 reads) and *Mantoniella* 8.7% (20,505 reads). This chapter focus on the taxonomic diversity and the distribution of the well-studied *Ostreococcus* and *Micromonas* genera. Both these genera already saw their distribution documented using several methods such as clone libraries (Guillou *et al.*, 2004; Viprey *et al.*, 2008), qPCR (Demir-Hilton *et al.*, 2011; Clayton *et al.*, 2017; Limardo *et al.*, 2017), available public sequences (Simon *et al.*, 2017) and metabarcoding using the V9 region of the 18S rRNA gene in oceanic water (Monier *et al.*, 2016), where Mamiellophyceae are replaced by other Chlorophyta lineages, such as Chloropicophyceae (cf. Chapter 3). The OSD dataset offers the opportunity to investigate their diversity and distribution in coastal waters using the metabarcoding approach.

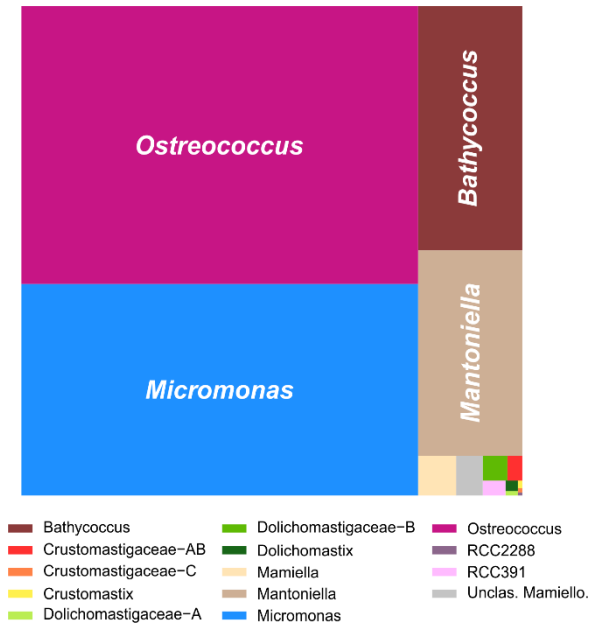


Fig.1: Treemap of the Mamiellophyceae genera contribution in the OSD2014 dataset.

Methodology

The OSD consortium provided 2 metabarcoding datasets for 2014 using the V4 region of the 18S rRNA gene: the LGC dataset (introduced in Chapter 3) and the Life Watch dataset (LW, introduced in Chapter 2). These datasets were sampled in the same areas and date, but different filtration and sequencing protocols were used to produce the reads. Finding the same reads in both dataset can be used to confirm the existence of novel sequences not detected previously. The LGC and LW datasets were analyzed with the same bioinformatic pipeline based on the mothur software (Schloss *et al.*, 2009) as detailed in the previous chapters up to the read clustering step: LW reads were clustered and OTUs were built at 99% identity (as in Chapter 3), while LGC unique sequences were extracted before the clustering (Fig. S1). 99% OTUs and unique sequences were automatically assigned using the curated PR² reference database (Chapter 1). OTUs reference sequences from LW and unique sequences from LGC were aligned to reference sequences and phylogenies were built using the Geneious software (Kearse *et al.*, 2012). Distribution throughout the OSD stations was computed using only the LGC dataset, while phylogenies relied on both datasets. Further analyses and graphics were computed using the R software version 3.0.2 (http://www.R-project.org/) and the same packages than in previous chapters.

Ostreococcus

349 LGC unique sequences were assigned to *Ostreococcus* genus, among which 10 were represented more than 100 times. These sequences constituted five clades (Fig.2A and B), four already described and a new one we named clade E to follow up the initial clade description (Guillou *et al.*, 2004). The same topology of tree was recovered with Fast Tree (Price *et al.*, 2009, 2010) and Bayesian building method (Ronquist *et al.*, 2012). Alignments (Fig.2B) confirmed, that clear signatures existed in the V4 to delineate the five *Ostreococcus* clades. In the V4 region, genetic variation between clades (Table 1) does not allow to discriminate all clades for OTUs built at 99% identity threshold. For example, the novel clade E cannot be distinguished from *O. tauri*, *"O. lucimarinus"* and clade B if OTUs are built at 99%.

	"O. lucimarinus"	clade B	*O. tauri*	*O.mediterraneus*	clade E
"O. lucimarinus"		98.9	99.4	98.0	99.1
O. clade B			99.1	97.7	99.4
O. tauri				98.3	99.3
O.mediterraneus					97.9
clade E					

Table 1: Matrix of the pairwise identity percent between *Ostreococcus* clades. The calculation was done on sequences from the alignment in Fig.2B.

Ostreococcus clade E unique sequences had clear signatures in the V4 region alignment, which resembles that of clade B (Fig.2B). This clade contained both unique LGC sequences as well as LW 99% OTUs but no reference sequences from Genbank either from cultures or from environmental clone libraries. Clade E represented up to 91% (OSD111) of the Mamiellophyceae reads and dominated coastal subtropical stations (Fig.3) from the US Southern Atlantic coast (OSD39, 58 and 143), the South European coast (Portugal OSD81, 111, 117, 153 and France OSD154) and the Adriatic (Venice, OSD69, 18% of the Mamiellophyceae). Clade E did not co-occur with other *Ostreococcus* or *Micromonas* clades underlying the specificity of its distribution pattern and ecological putative importance (Fig. S2). Previous studies on *Ostreococcus* distribution (Demir-Hilton *et al.*, 2011; Clayton *et al.*, 2017) missed this new environmental clade because they focused on qPCR approaches based on primers and probes designed on available V4 sequences from strains in culture. The probes aimed at discriminating 2 clades (OI and OII), but sequences from clade E differ at two positions from clade OII (targeting *Ostreococcus* clade B) which are located inside the qPCR probe (Fig. S3).

"O. lucimarinus" represented up to 87% of the Mamiellophyceae reads in South Africa (Robben Island, OSD133). It dominated Atlantic and North Sea European coastal stations (off Belgium OSD182 70%, 183 84% and 184 61%, off Portugal OSD115 69%) and represented 60% of Mamiellophyceae reads in one of the 3 Azores stations (OSD98, Fig.3). *"O. lucimarinus"* was totally absent from the

Mediterranean Sea and tropical waters (Fig.3). These results agree with previous literature: *"O. lucimarinus"* targeted by the OI clade probes was described as a cold mesotrophic coastal clade (Demir-Hilton *et al.*, 2011).

The third best represented unique sequence (22,950 reads) was assigned to the species *O. mediterraneus*. This unique sequence reached a maximal contribution to the Mamiellophyceae reads in the Uruguay lagoon stations (OSD149 76% and 150 83%) and was recorded to a lower extent in North European stations (Finland, OSD30 28%) and in the Black Sea (off Ukraine OSD 64 21% and off Bulgaria 8%, Fig.3). *O. mediterraneus* initially isolated from the French coast of Mediterranean Sea (Subirana *et al.*, 2013) was surprisingly not recovered in Mediterranean Sea although OSD sampled few stations in the particular area of Mediterranean Sea from which it was isolated (only 3 stations OSD6 Blanes, 14 Banyuls and 22 Marseilles).

Ostreococcus clade B unique sequence was represented by 7269 sequences and reached 83% of the Mamiellophyceae reads in the station off Panama (OSD51, Fig.3). Clade B contribution to Mamiellophyceae was higher than 10% at 7 tropical stations from a range of oceans (OSD25, 37, 51 Florida, 95 Singapore, 122 Red Sea, 144 and 147 Sri Lanka, Fig.3). These results agreed with previous qPCR studies. The OII clade which encompasses *Ostreococcus* clade B was recovered in oligotrophic and warm oceanic waters (Demir-Hilton *et al.*, 2011).

Three unique sequences were assigned to the species *Ostreococcus tauri* with between 4500 and 1500 identical reads, which did not gather in a monophyletic clade in the phylogenetic FastTree and Bayesian trees (Fig.2A). The 18S phylogeny of *O. tauri* reference sequences should be investigated more carefully, in particular sequence Y15814 which was obtained more than 20 years ago and probably from a mixed culture. In the V4 alignment (Fig.2B), *O. tauri* showed a dual signature and was more than 99% similar to the other *Ostreococcus* clades (except *O. mediteraneus* 98.3% sequence similarity, Table 1). Distribution patterns were not similar for the three unique sequences. The two major ones were HWI-M02024_112_000000000-ACJ3F_1_1105_25414_3189 (referred as OT_3189) and HWI-M02024_112_000000000-ACJ3F_1_1116_6656_8393 (OT_8393) both corresponding to the *O. tauri* genome sequence (mutations occurred in the non-showed part of the alignment). OT_3189 contributed to Mamiellophyceae up to 81% Delaware (OSD 36), but reached 10% in only 4 stations from the the US Atlantic coast (OSD 35, 36, 39 and 186, Fig.3). OT_8393 contributed to up to 33% of Mamiellophyceae reads (Scotland, OSD163) and reached 10% in 12 stations from the North Europe coast (Scotland OSD163, off Belgium OSD176 26%) and Adriatic Sea (Venice gulf OSD 48 30%, Fig.3). Just as, *O. tauri* was targeted as *"O. lucimarinus"* by OI qPCR probes (Demir-Hilton *et al.*, 2011). *O. tauri* was absent from Mediterranean Sea (except one Adriatic station OSD48) despite the fact that *O. tauri* was initially isolated from the Thau lagoon (Chrétiennot-Dinet *et al.*, 1995) and recorded in the Mediterranean Sea in several diversity studies (Guillou *et al.*, 2004; Subirana *et al.*, 2013).

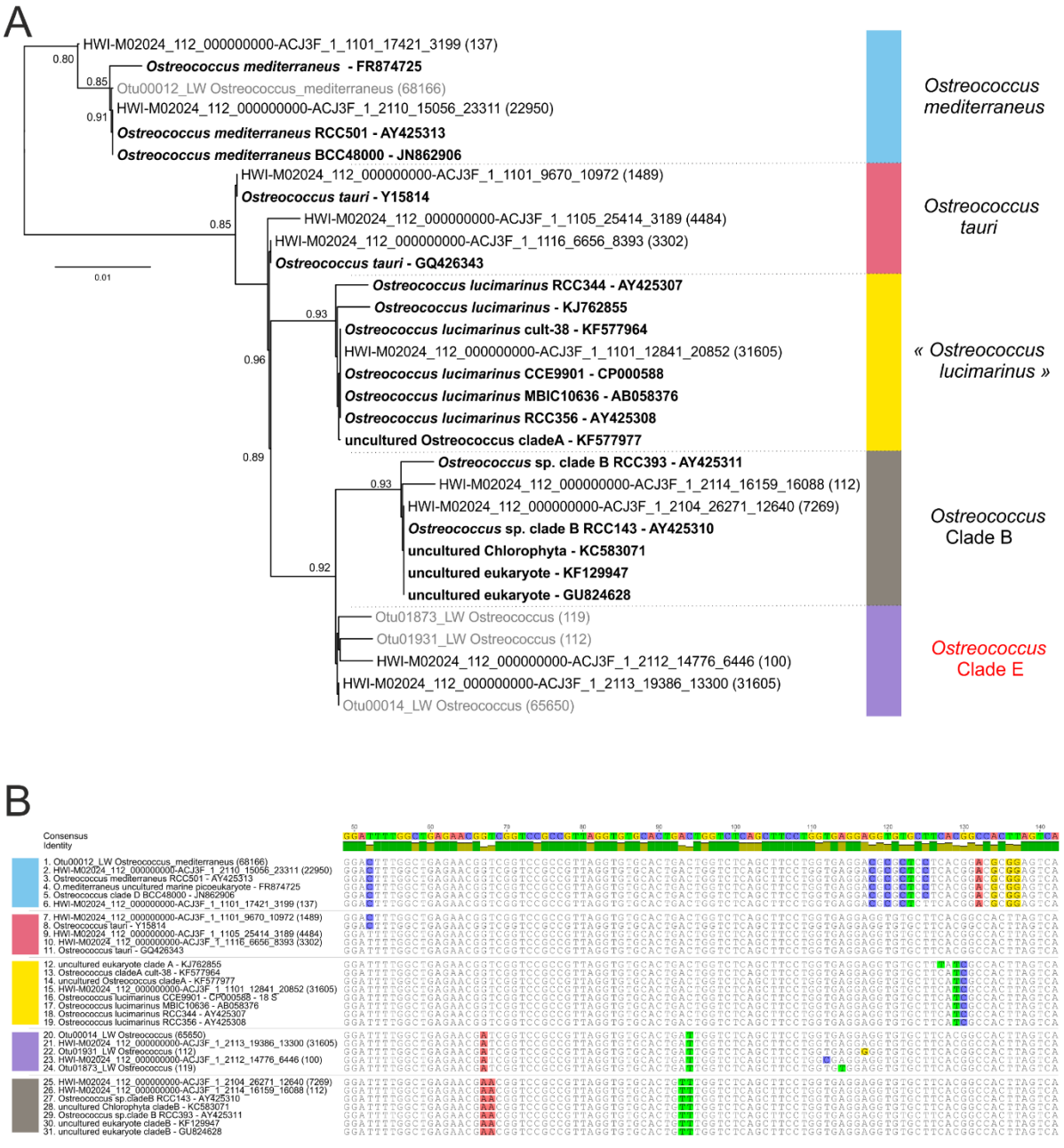


Fig.2: Phylogenetic diversity inside *Ostreococcus* genus. A- phylogenetic FastTree of 31 *Ostreococcus* V4 regions of the 18S rRNA gene, the tree was rooted with *Bathycoccus prasinos* (AY425315, FN562453, JX625115, KF501036) and only bootstrap values higher than 70% were represented. Reference sequences from GenBank were in bold, representative sequences of the Life Watch (LW, see Chapter 2) OTUs built at 99% identity were in grey. Number into brackets in the sequence names is the number of reads either of the unique sequences or inside the LW OTUs 99% and only unique sequences and OTUs represented by more than 100 reads were taken into account. Red legends refer to new diversity unveiled by OSD2014 datasets. B- Alignment of 31 *Ostreococcus* V4 regions, the alignment was 344 base pairs, but only the main signatures were shown (around the 40th and 150th position of the original alignment).

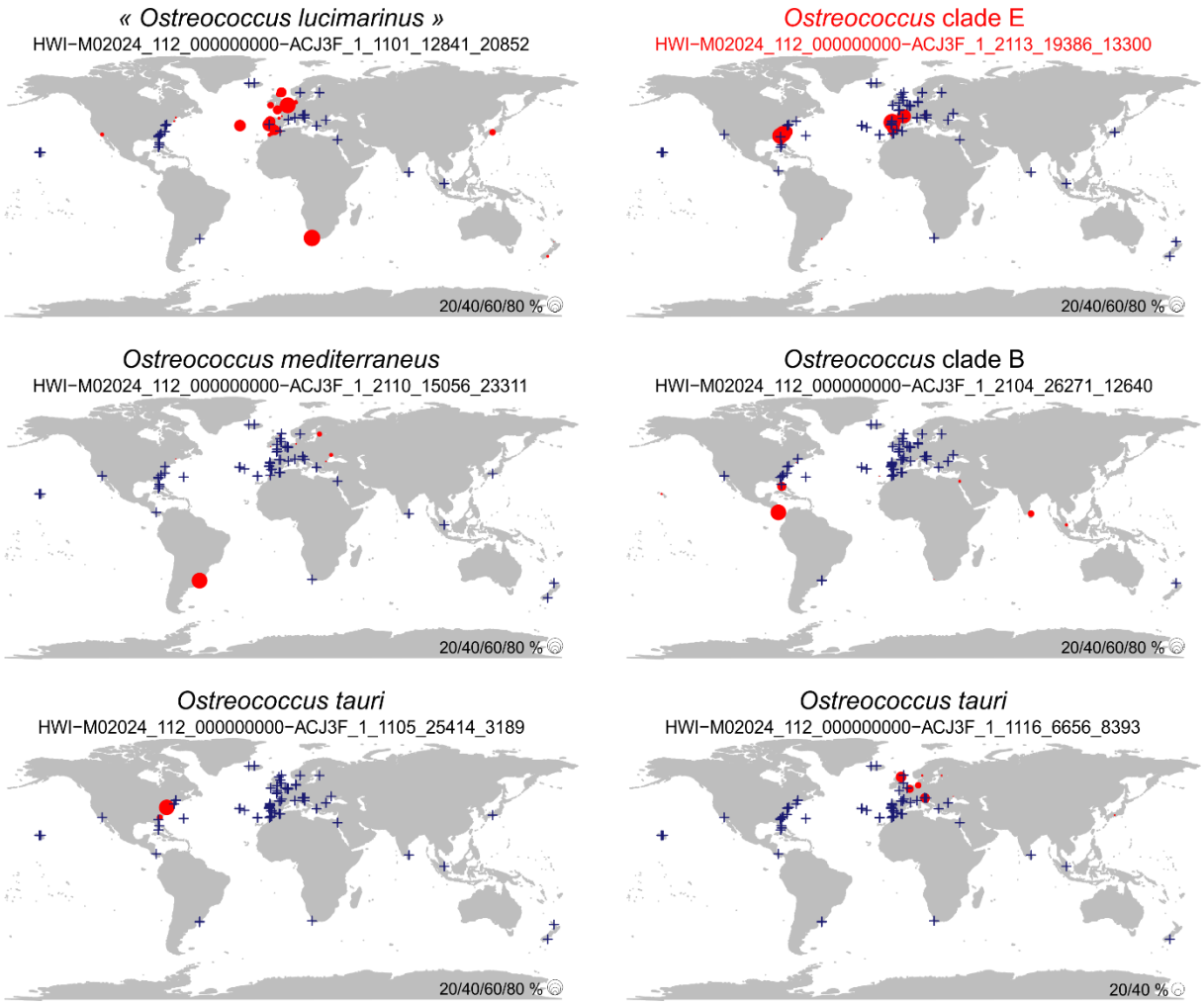


Fig.3: *Ostreococcus* 6 major unique sequences distribution in OSD2014 (LGC). Stations where the sequences were not recorded are represented by blue crosses. The circle surface corresponds to the percent of read versus the Mamiellophyceae read number at this station. Legend in red refers to the new *Ostreococcus* clade (Fig.2).

Micromonas

467 unique sequences were assigned to *Micromonas* genus, among which 17 were represented more than 100 times in the OSD LGC dataset. These sequences can be divided into 7 clades (Fig.4A and B): the four recently described species (Simon *et al.*, 2017), the two candidate species described by Simon *et al.* (2017) and a new one named clade B sub-arctic in reference to its phylogenetic place in the tree and geographical distribution (Fig.4A and B). According to signatures in the alignment and phylogenies *Micromonas commoda* (Van Baren *et al.*, 2016) and *Micromonas bravo* (Simon *et al.*, 2017) were split into two subclades (Fig.4A and B): *Micromonas commoda* AB and C in reference to previous literature (Šlapeta *et al.*, 2006; Worden, 2006) and *Micromonas bravo* I and II. The same well supported tree topology was recovered with both Fast Tree and Bayesian building method. The genetic divergence between clades is larger than 1 % for almost all the clade pairs (Table 2) and allows to distinguish all clades using an identity threshold of 99% except *M. commoda* AB and C (99.2% identity), *M. polaris* and the new clade B sub-arctic (99.2%) and *M. commoda* and the clade B warm (Table 2).

	M. polaris	*M. bravo*	B sub-arctic	B._.4	*M. pusilla*	B warm	*M. commoda*(C)	*M. commoda*(AB)
M. polaris		98.6	99.2	98.1	97.9	98.0	98.0	97.7
M. bravo			98.2	97.5	97.7	97.9	97.7	97.4
B sub-arctic				97.0	96.9	98.0	97.3	97.3
B._.4					97.2	98.6	97.4	97.3
M. pusilla						98.9	97.5	96.9
B warm							99.0	98.3
M. commoda(C)								99.2
M. commoda(AB)								

Table 2: Matrix of the pairwise identity percent between *Micromonas* clades. The calculation was done on sequences from the alignment in Fig.*4B.*

The main unique sequence was represented by 41,275 reads and were assigned to *M. bravo* II (Fig.4A and B). *M. bravo* II unique sequence represented up to 85% of the Mamiellophyceae reads off Morocco (OSD93) and off Portugal (OSD102). It dominated most of Mediterranean Sea stations (Fig.5A), some north European stations and to a lower extent Pacific Ocean coastal (OSD41 16%, 43 25% and the Hawaiian OSD144 38%) stations and South Atlantic station (Uruguay coastal lagoon OSD151 26%). It was the only *Micromonas* representative recorded in Japan (OSD124 41%, Fig.5A).

Micromonas bravo I unique sequence was represented by 4813 reads and represented more than 10% of Mamiellophyceae at 13 stations spread out along the European coast and in contrast to *M. bravo* II was almost absent from the Mediterranean Sea (Fig.5A). This unique sequence contributed up to 65% of Mamiellophyceae in the English Channel (Plymouth, OSD1). Previous work in the English Channel

lead to the conclusion that *M. bravo* (previously non artic B.E.3 clade) dominated the *Micromonas* community in summer and should be adapted to warm well lighted coastal waters (Foulon *et al.*, 2008) which is consistent with the. OSD dataset since sampling was done in June. Analyses of environmental and strain sequences (Chapter 1 and Simon *et al.*, 2017) led to the conclusion that this clade is ubiquitous, while in the OSD metabarcoding dataset *M. bravo* showed clear distribution patterns (Fig.5A). Although *M.bravo* II and I seem to have distinct distributon (Fig. S2), the separation between these two sub-clades should be confirmed using finer resolution markers such as the ITS.

Micromonas commoda C unique sequence was represented by 14429 reads (Fig.4A and B) and contributed to up to 79% of Mamiellophyceae in Iceland (OSD128). *M. commoda* C was found in North Sea (United Kingdom coast OSD62 53%, Norway fjord OSD141 54%), North Atlantic (Iceland OSD128 and Canada OSD152 50%, Fig.5A) and contributed to around 20% in New Zealand stations (OSD65 and 71). *Micromonas commoda* AB was represented by 7394 100% identical sequences and contributed to 100% of the Mamiellophyceae in the Adriatic Sea off Croatia (OSD21). *M. commoda* AB reads were distributed in tropical and subtropical waters (Fig.5A) especially in the Gulf of Mexico (USA, OSD46 80%), Florida (USA, OSD38 70%), off Sri Lanka (OSD147 55%), Singapore (OSD95 48%) and Hawaii (OSD56, 57 around 25% and 144 10%). *M. commoda* was firstly described by Van Baren *et al.* (2016) who just mentionned that this species was not recorded in high latitudes yet (beyond 60°North and South). The species was then revised by Simon *et al.* (2017), who described the distribution of this species as ubiquitous using available reference sequences. The high genetic variability within this newly described species was already highlighted in the literature (Šlapeta *et al.*, 2006; Worden *et al.*, 2009; Simon *et al.*, 2017). Simon *et al.* proposed the hypothesis that speciation events should be ongoing inside *M. commoda*. The OSD dataset provided specific patterns for both *M. commoda* AB and C (Fig.5 and Fig. S2) suggesting that, these two clades are not ubiquitous (*M. commoda* was not recovered in Mediterranean Sea Fig.5A) and that, *M. commoda* AB and C could be in fact different species. Interestingly, *M. commoda* AB OSD distribution particularly fitted the clade A strains distribution in Šlapeta *et al.* (Šlapeta *et al.*, 2006).

Micromonas environmental clade B-warm (candidate species 2) unique sequence corresponded to 7299 reads (Fig.4A and B) and contributed to more than 10% of the Mamiellophyceae reads at 7 stations from tropical waters (Fig.5B). Clade B-warm reached 52% of the Mamiellophyceae reads off Hawaii (OSD56). This clade was recorded at the 3 Hawaii stations (OSD56, 57 29%, 144 23%), in Florida (OSD37 31%), off Portugal (OSD101 30%), in Mediterranean Sea off Egypt (OSD53 22%) and in Singapore (OSD95 12%). The OSD distribution agreed with previous data (Simon *et al.*, 2017). One representative strain was isolated in Mediterranean Sea in summer in 2006 (RCC1109, http://roscoff-culture-collection.org/rcc-strain-details/1109) and then, recovered from clone libraries in the Red Sea

(Acosta *et al.*, 2013), South China Sea ("unknown clade", Wu *et al.*, 2014) and off Taiwan (*Micromonas* clade VI, Lin *et al.*, 2016).

Micromonas polaris unique sequence was represented by 2852 reads and contributed to more than 10% of the Mamieollophyceae reads at 4 stations. *M. polaris* reached 93% in Nunavut (Canada OSD105) and was essentially hosted in the cold waters of Northern (Arctic Ocean OSD146 86% and off Greenland OSD80 42%) and Southern latitude (Palmer station OSD187 21%). *M. polaris* was firstly isolated from the Arctic Ocean (CCMP2099, Lovejoy *et al.*, 2007) and shown to be the dominant pico-eukaryote there in the summer (Balzano *et al.*, 2012), but recently recorded in the Southern Ocean (Simmons *et al.*, 2015).

A new *Micromonas* clade close to *M. polaris* was observed. Reference sequences already exist for this clade and full sequence alignments and phylogeny will be necessary to clarify it positions (Fig.4A and B). This new clade was named B sub-arctic given the distribution of the 1329 corresponding reads. It contributed to more than 10% of Mamiellophyceae reads at 6 stations located in north temperate waters from the North Atlantic Ocean (Canada OSD 152 66%, off USA OSD54 15%, off Iceland OSD106 and 130 respectively 33% and 19%) and in the North Sea (off United Kingdom OSD125 and 169 around 20%).

Micromonas clade B._.4 (candidate species 1) unique sequence was recovered in 2116 copies (Fig.4A and B). This clade contributed to more than 10 % of Mamiellophyceae reads in 4 tropical stations (Mediterranean Sea off Egypt OSD52 and 53 respectively 13% and 17%, Red Sea off Israël OSD122 13%) and reached 18% in Florida (OSD29). Environmental clade B._.4 was little represented in the OSD coastal dataset but showed a tropical distribution (Fig.5A). t, Available reference sequences originate from the Mediterranean, Red Sea and Pacific Ocean (Chapter 1 and Simon *et al.*, 2017).

M. pusilla unique sequences was the least represented of the *Micromonas* clades with 1619 sequences (Fig.4A and B). *M. pusilla* contributed to more than 10% to the Mamiellophyceae at only 2 stations and was recovered in Uruguay coastal lagoon (OSD151 28%) and in Baltic Sea (OSD72 13%) and off Belgium (OSD170 and 174 respectively 7 and 9.3%). *M. pusilla* strains have been isolated from the coast of Chile (Le Gall *et al.*, 2008) and in the North Sea (Šlapeta *et al.*, 2006).

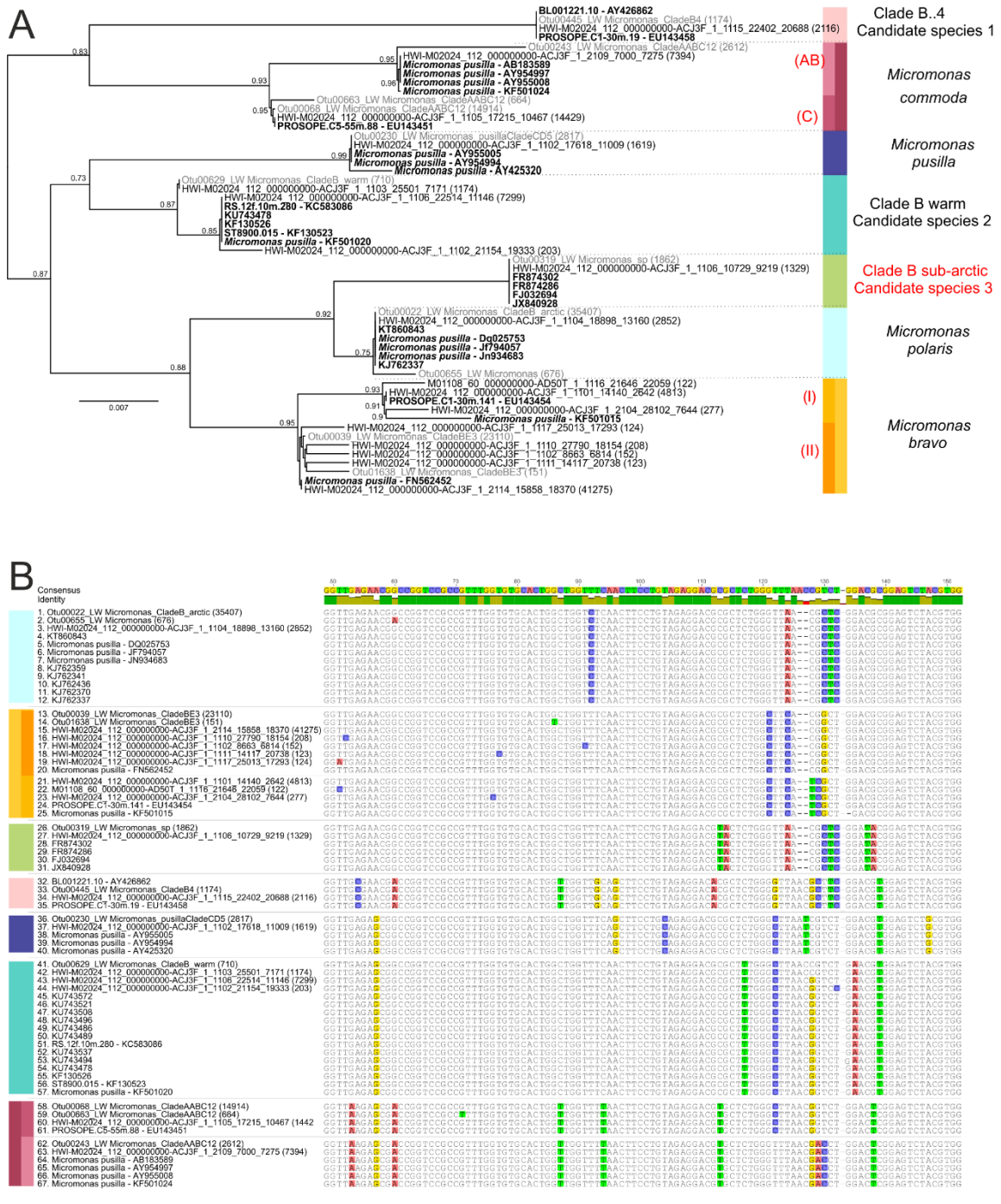


Fig.4: Phylogenetic diversity inside *Micromonas* genus. A- phylogenetic FastTree of 55 Micromonas V4 regions of the 18S rRNA gene, the tree was rooted with Mamiellales (RCC391, AY425321 and *Mamiellla gilva*, FN562450) and only bootstrap values higher than 70% were represented. Reference sequences from GenBank were in bold, representative sequences of the Life Watch (LW, see Chapter 2) OTUs built at 99% identity were in grey. Number into brackets in the sequence names is the number of reads either of the unique sequences or inside the LW OTUs 99% and only unique sequences and OTUs represented by more than 100 reads were taken into account. Red legends refer to new diversity unveiled by OSD2014 datasets. B- Alignment of 55 *Micromonas* V4 regions, the alignment was 327 base pairs, but only the main signatures were shown (around the 40th and 150th position of the original alignment).

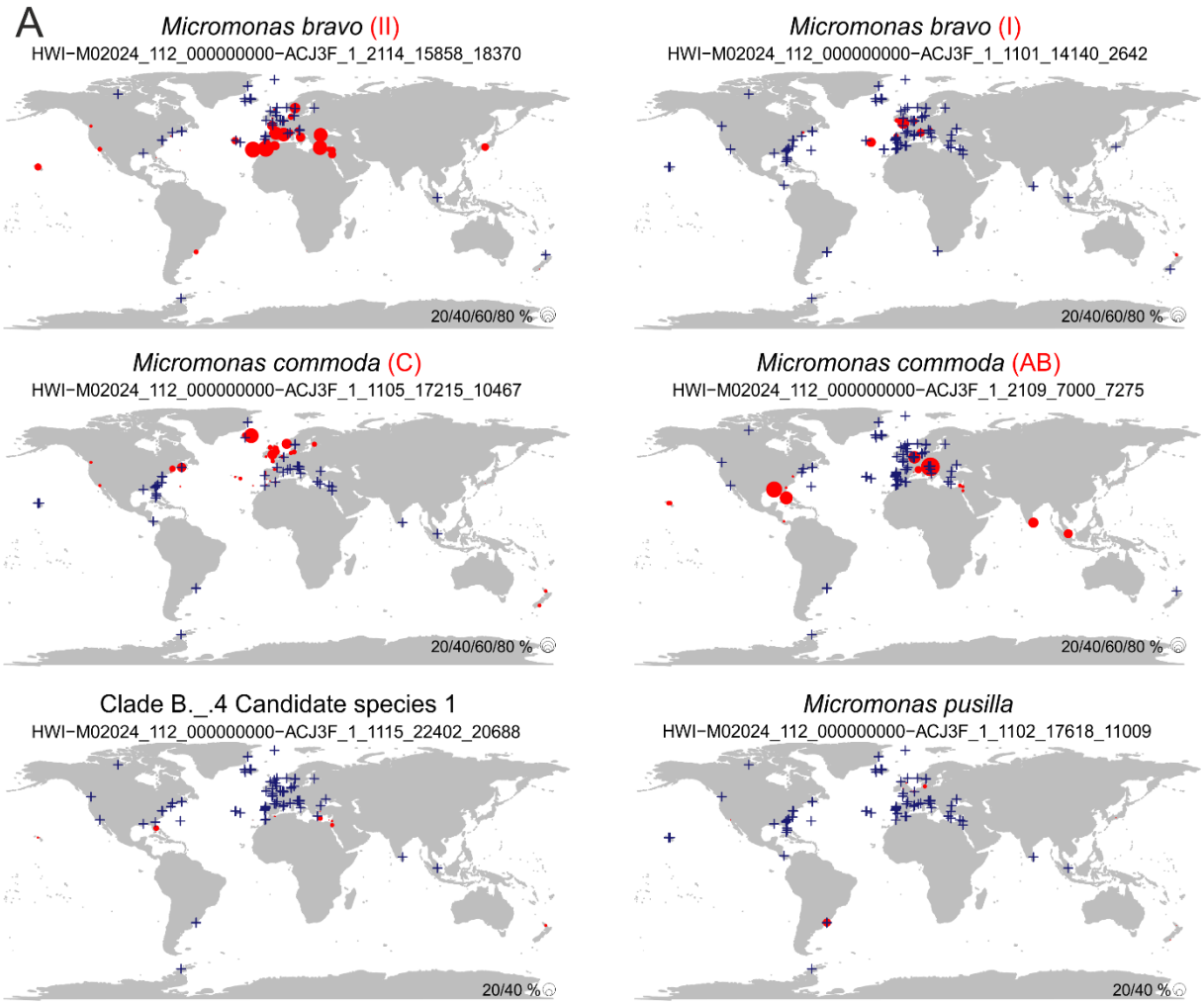


Fig.5: A and B, *Micromonas* 9 major unique sequences distribution in OSD2014 (LGC). Stations where the sequences were not recorded are represented by blue crosses. The circle surface corresponds to the percent of read versus the Mamiellophyceae read number at this station. Legend in red refers to the new *Micromonas* clades (Fig.4).

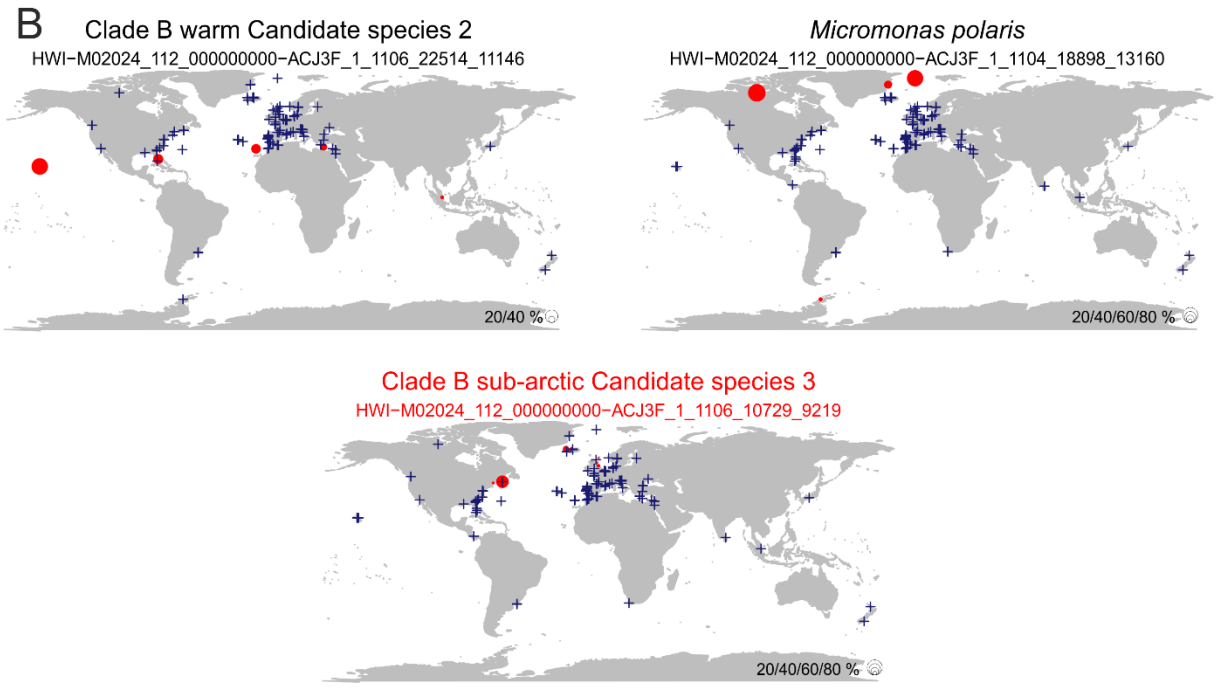


Fig.5: A and **B**, Micromonas 9 major unique sequences distribution in OSD2014 (LGC). Stations where the sequences were not recorded are represented by blue crosses. The circle surface corresponds to the percent of read versus the Mamiellophyceae read number at this station. Legend in red refers to the new Micromonas clades (Fig.4).

Conclusions

The OSD dataset allowed to study the biogeography distribution of species and clades within *Ostreococcus* and *Micromonas*. Each taxonomic unit (species and clades) that had been described previously was found in the OSD data set. A new clade for *Ostreococcus* (clade E) was uncovered, which seems to have a high contribution to Mamiellophyceae as well as a specific geographic distribution. This chapter underlines the importance of detailed phylogenetic analyses to assign correctly metabarcodes in relation to existing sequences. The use of unique sequences analyses provides the most detailed picture of diversity and distribution of a specific lineage

In the future, similar analysis should be performed for all Chlorophyta classes detected in OSD dataset, setting as a priority Chlorophyta classes, that did not show clear distribution patterns (see Chapter 3), such as the Pyramimonadales.

References

Acosta, F., Ngugi, D.K., and Stingl, U. (2013) Diversity of picoeukaryotes at an oligotrophic site off the Northeastern Red Sea Coast. *Aquat. Biosyst.* **9**: 1:16.

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

Van Baren, M.J., Bachy, C., Reistetter, E.N., Purvine, S.O., Grimwood, J., Sudek, S., et al. (2016) Evidence-based green algal genomics reveals marine diversity and ancestral characteristics of land plants. *BMC Genomics* **17**: 22.

Butcher, R.W. (1952) Contributions to our knowledge of the smaller marine algae. *J. Mar. Biol. Assoc. United Kingdom* **31**: 175.

Chrétiennot-Dinet, M.-J., Courties, C., Vaquer, A., Neveux, J., Claustre, H., Lautier, J., and Machado, M.C. (1995) A new marine picoeucaryote: *Ostreococcus tauri* gen. et sp. nov. (Chlorophyta, Prasinophyceae). *Phycologia* **34**: 285–292.

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.

Demir-Hilton, E., Sudek, S., Cuvelier, M.L., Gentemann, C.L., Zehr, J.P., and Worden, A.Z. (2011) Global distribution patterns of distinct clades of the photosynthetic picoeukaryote *Ostreococcus*. *ISME J.* **5**: 1095–1107.

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

Le Gall, F., Rigaut-Jalabert, F., Marie, D., Garczarek, L., Viprey, M., Gobet, A., and Vaulot, D. (2008) Picoplankton diversity in the South-East Pacific Ocean from cultures. *Biogeosciences* **5**: 203–214.

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

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**: 1647–9.

Limardo, A.J., Sudek, S., Choi, C.J., Poirier, C., Rii, Y.M., Blum, M., et al. (2017) Quantitative biogeography of picoprasinophytes establishes ecotype distributions and significant contributions to marine phytoplankton. *Environ. Microbiol.* **19**: 3219–3234.

Lin, Y.-C., Chung, C.-C., Chen, L.-Y., Gong, G.-C., Huang, C.-Y., and Chiang, K.-P. (2016) Community Composition of Photosynthetic Picoeukaryotes in a Subtropical Coastal Ecosystem, with Particular Emphasis on Micromonas. *J. Eukaryot. Microbiol.* **64**: 349–359.

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

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

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.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2009) Fasttree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**: 1641–1650.

Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**: e9490.

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

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., et al. (2012) MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst. Biol.* **61**: 539–542.

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**: 7537–41.

Simmons, M.P., Bachy, C., Sudek, S., van Baren, M.J., Sudek, L., Ares, M., et al. (2015) Intron Invasions Trace Algal Speciation and Reveal Nearly Identical Arctic and Antarctic *Micromonas* Populations. *Mol. Biol. Evol.* **32**: 2219–2235.

Simon, N., Foulon, E., Grulois, D., Six, C., Desdevises, Y., Latimier, M., et al. (2017) Revision of the genus *Micromonas* Manton et Parke (Chlorophyta, Mamiellophyceae), of the species *M. pusilla* (Butcher) Manton et Parke, of the species *M.commoda* van Baren, Bachy et Worden and description of two new species, based on the. *Protist* **accepted**:

Šlapeta, J., López-García, P., and Moreira, D. (2006) Global dispersal and ancient cryptic species in the smallest marine eukaryotes. *Mol. Biol. Evol.* **23**: 23–29.

Subirana, L., Péquin, B., Michely, S., Escande, M.L., Meilland, J., Derelle, E., et al. (2013) Morphology, genome plasticity, and phylogeny in the genus *Ostreococcus* reveal a cryptic species, *O. mediterraneus* sp. nov. (Mamiellales, Mamiellophyceae). *Protist* **164**: 643–659.

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.

Worden, A. (2006) Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquat. Microb. Ecol.* **43**: 165–175.

Worden, A.Z., Lee, J.-H., Mock, T., Rouzé, P., Simmons, M.P., Aerts, A.L., et al. (2009) Green Evolution and Dynamic Adaptations Revealed by Genomes of the Marine Picoeukaryotes Micromonas. *Science (80-.).* **324**:.

Wu, W., Huang, B., Liao, Y., and Sun, P. (2014) Picoeukaryotic diversity and distribution in the subtropical-tropical South China Sea. *FEMS Microbiol. Ecol.* **89**: 563–579.

List of Figures

List of Tables

List of Supplementary Figures

Supplementary Figures

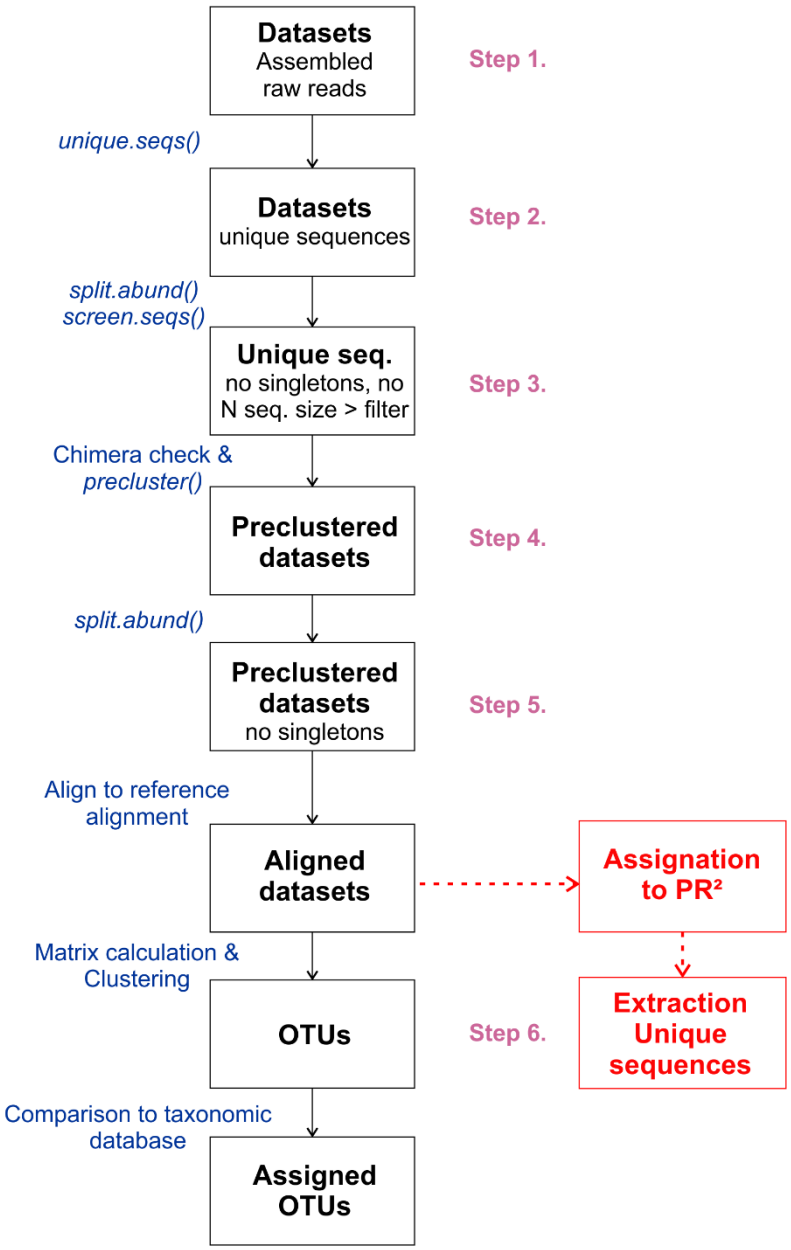


Fig. S1: Bioinformatics pipeline under mothur software schema modified from chapter 2 to work with unique sequences.

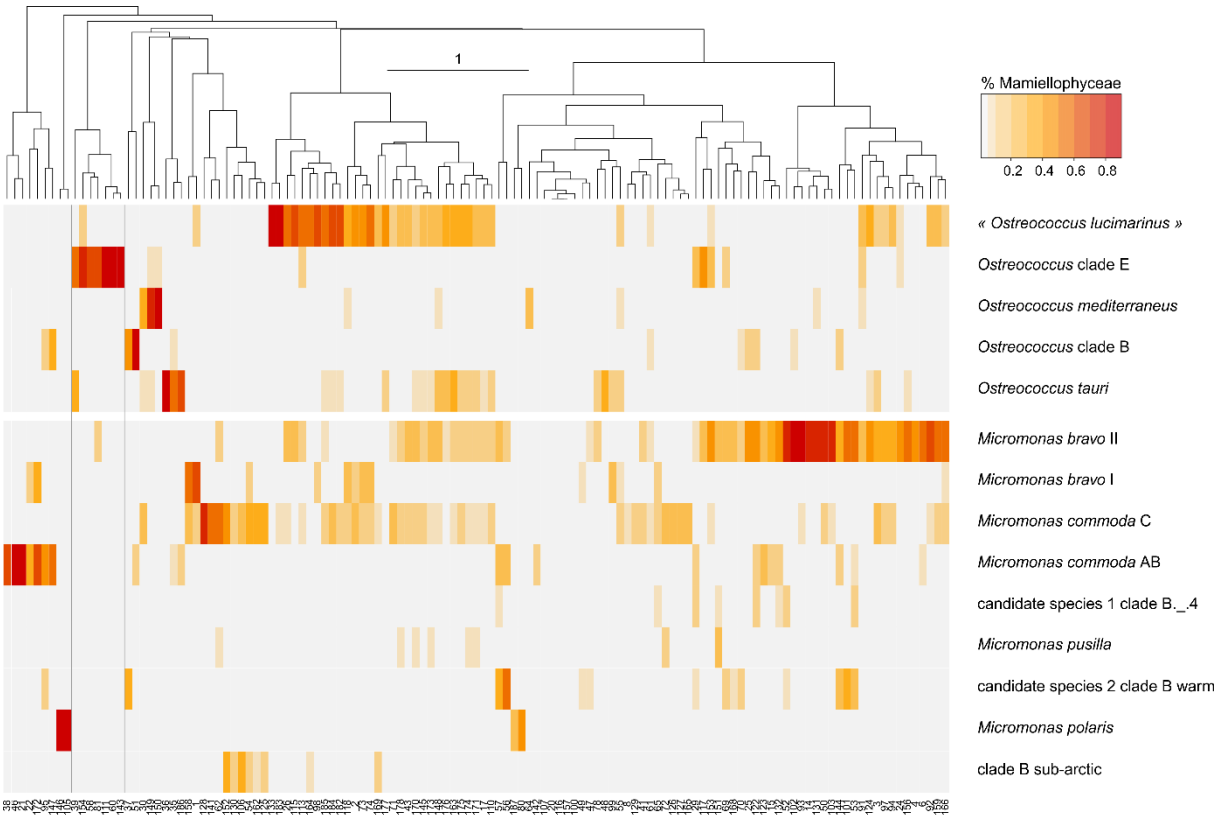


Fig. S2: Heatmap of the *Ostreococcus* and *Micromonas* communities in OSD surface stations. Unique sequences represented by more than 100 reads (sequence list in Fig.2B and Fig.4B) were aggregated by clades and normalized by the number of Mamiellophyceae reads per station.

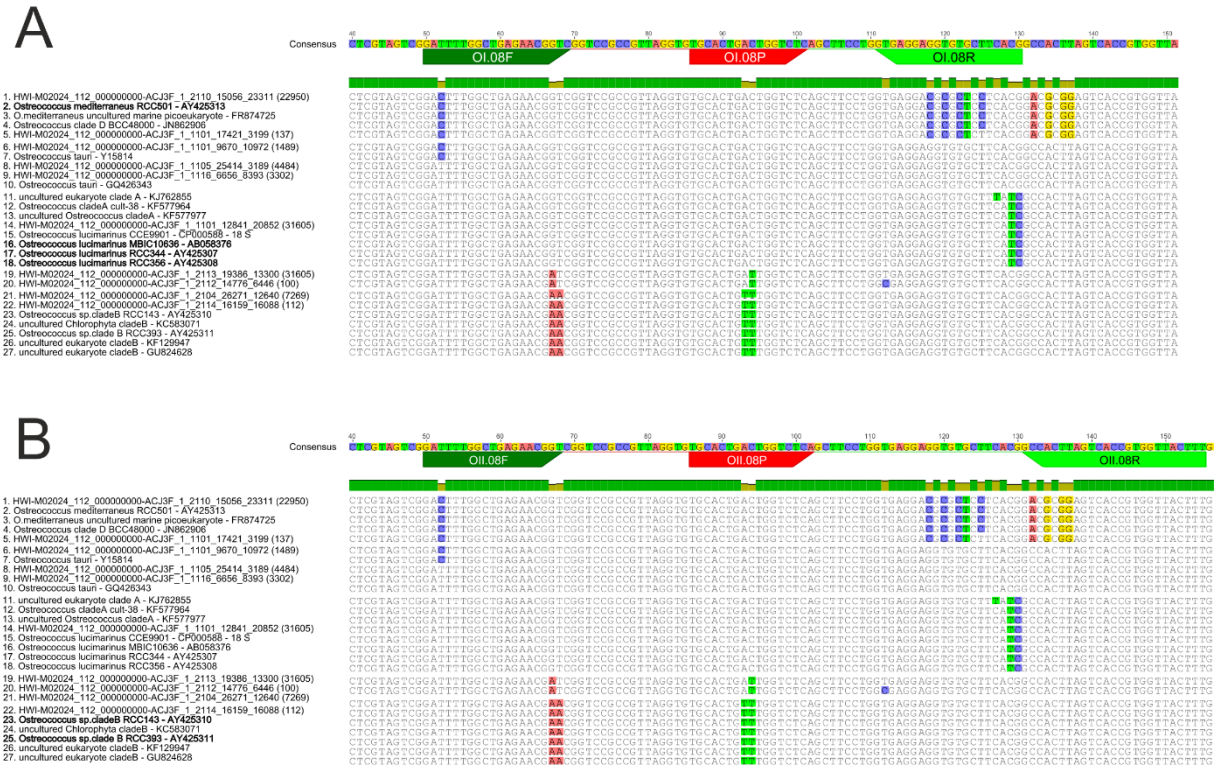


Fig. S3: Alignments of *Ostreococcus* V4 regions and localization of the qPCR primers and probes (Demir-Hilton et al., 2011) used to quantify A- OI clade and B- OII clade. Strains sequences initially used to describe *Ostreococcus* clades are in bold (Guillou et al., 2004).

Conclusion and perspectives

Improvements in Chlorophyta metabarcoding

The first chapter provided an accurate database of reference sequences needed to identify environmental reads. This allowed for example to assign prasinophytes clade VIII and IX in metabarcoding datasets, since the reference sequences belonging to these clades were badly assigned in previous versions of the PR^2 database. The addition of metadata to the database such as the isolation and sampling geographical coordinates allowed to draw maps of Chlorophyta distribution in the ocean from publicly available sequences.

The second chapter offered the confirmation that the V4 region of the 18S rRNA gene allows to assign Chlorophyta reads up to the species taxonomic level, while V9 lacked reference sequences for some taxonomic groups, preventing these groups to be correctly detected. This chapter highlighted the importance of databases in metabarcoding analyses and in the resulting images of diversity.

In Chapter 4, checking automatic assignation using phylogenies based on alignments of OTUs sequences with high quality reference sequences is very critical. The presence of clear signatures in alignments for species or clades is also important.

The addition in reference databases of clearly identified new clades defined in metabarcode studies such as OSD could improve assignation for future datasets. Since these sequences are small (around 400 bp for the V4 region), however, these new clades should be confirmed by full gene high quality sequences (e.g. obtained by the Sanger method) before these new clades can make their way to reference databases.

Chlorophyta lineages distribution and relation to environmental conditions

The third chapter provided the first global distribution of the 13 existing Chlorophyta classes. Some classes showed very specific distribution patterns and environmental preferenda such as the Chlorophyceae, Chloropicophyceae and the environmental clade prasinophytes clade IX. This work unveiled the importance of the Ulvophyceae, Trebouxiophyceae and Chlorophyceae (UTC clade) in coastal waters. Before, these three classes were mostly thought to be restricted to freshwaters and macroalgae (Ulvophyceae). In contrast, other classes are ubiquitous in coastal waters (such as the Mamiellophyceae, Pyramimonadales) and do not show any specific patterns at the class level. Lower taxonomic levels need to be investigated to obtain specific distributions (see for example chapter 4).

Unfortunately, the OSD dataset did not sampled extensively oceanic waters in contrast to coastal ones, which limits possible interpretations of switch between Chlorophyta communities in these two environments. Moreover, biotic interactions (such as parasitism, predation, competition…) should be taken into account in order to completely understand biogeography and ecology of Chlorophyta classes. These interactions might be the investigated through interaction networks, but it will be necessary to

curate metabarcoding references databases for known eukaryotic parasites and to include information for viruses, which are not amenable metabarcoding since they rarely share universal genes and for which metagenomics are a better approach (Hingamp *et al.*, 2013).

New environment clades for the well-studied Mamiellophyceae genera *Micromonas* and *Ostreococcus*

In the fourth chapter, I analyzed the diversity of the two Mamiellophyceae genera *Micromonas* and *Ostreococcus* that have been extensively studied in the recent decade. Phylogenetic analysis of OSD unique sequences allowed to demonstrate the existence of a new environmental clades for both of these genera.

This is a good example that, provided that appropriate bioinformatics methods are used, metabarcoding is powerful enough to describe diversity down to the species level and to unveil new diversity inside some well described taxa. Other markers such as the internally transcribed spacer (ITS) would be probably more suitable for investigating lower taxonomic levels, but reference databases are not yet as well developed that for other markers such as nuclear 18S or plastidial 16S. Moreover, it is often difficult to design primers to amplify regions compatible with current sequencing technology because ITS sequences align poorly.

The more accurately diversity will be described, the easiest the link between distribution and environmental will be understood. In marine ecosystems, available environmental variables are often limited to a small set of parameters (temperature, salinity, nutrients concentration, depth…), but the acquisition of other data such as concentration of iron or of other trace metals may allow to better understand protist distribution.

Perspectives

In the future, the diversity and distribution of unique sequences should be studied for Chlorophyta lineages to better diversity and distribution at low taxonomic levels and possibly highlight new diversity especially in little studied lineages. For example, one could focus on lineages with no obvious distribution patterns such as the Pyramimonadales, to delimitate phylogenetic units, which may have specific distribution patterns. The OSD dataset will hopefully provide information on where to sample in order to put new diversity into culture. Having cultures also provide resources to study the physiology or interactions of representative organisms, to look active molecules for medical or industry purposes, but even more importantly to establish taxonomic descriptions that rely on morphological descriptions.

Literature cited

Hingamp, P., Grimsley, N., Acinas, S.G., Clerissi, C., Subirana, L., Poulain, J., et al. (2013) Exploring nucleo-cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *ISME J.* **7**: 1678–95.

List of publications

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

Lopes dos Santos A., **Tragin M.**, Gourvil P., Noël M.-H., Decelle J., Romac S. & Vaulot D. 2016. Diversity and oceanic distribution of prasinophytes clade VII, the dominant group of green algae in oceanic waters. **ISME Journal** 11 (2) p.512-528

Simon N., Foulon E., Grulois D., Six C., Desdevises Y., Latimier M., Le Gall F., **Tragin M.**, Houdan A., Derelle E., Jouenne F., Marie D., Le Panse S., Vaulot D., Marin B. 2017. Revision of the genus *Micromonas* Manton et Parke (Chlorophyta, Mamiellophyceae), of the type species *M. pusilla* (Butcher) Manton & Parke and of the species *M. commoda* van Baren, Bachy and Worden and description of two new species based on the genetic and phenotypic characterization of cultured isolates. **Protist**. 168 p.612-635

Tragin M., Zingone A., Vaulot D. 2017. 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** (accepted)

Tragin M., Vaulot D. Communities of green microalgae in marine coastal waters: the OSD dataset. **Frontiers in Aquatic Microbiology** (in prep.)

Tragin M., Vaulot D. Novel diversity within *Micromonas* and *Ostreococcus* (Mamiellophyceae) unveiled by metabarcoding analyses. (in prep.)

Kuwata, A., Yamada, K., Lopes dos Santos, A., **Tragin, M.**, Ichinomiya, M., Yoshikawa, S. & Vaulot, D. Biology and ecology of Parmales (Bolidophyceae), a picophytoplankton group, sister of diatoms. **Frontiers in Microbiology** (submitted December 2017).

Tragin M., Vaulot D. Environmental diversity of Chlorophyta microalgae in coastal water. **Journal of Phycology** (in prep.)

Appendix: Scientific cruise onboard the Amundsen icebreaker and lab work for Green Edge project.

While tropical and temperate marine primary production is dominated by cyanobacteria and picophytoplankton (cells<3µm), cyanobacteria are completely absent in Arctic waters (Lovejoy *et al.*, 2007). In contrast to tropical waters, small green algae such as *Micromonas polaris* and *Bathycoccus prasinos* are dominating picophytoplankton and persist throughout all seasons in the Arctic Ocean (Lovejoy *et al.*, 2007; Balzano, Marie, *et al.*, 2012). Green microalgae such as *Pyramimonas* are also important in the larger size classes (Balzano, Gourvil, *et al.*, 2012; Balzano, Marie, *et al.*, 2012).

Every year, when the ice melts in Arctic Ocean, a massive bloom begins to take place under the ice and then propagate to the open ocean. The French-Canadian Green Edge project aims at understanding the phenology of this ice edge bloom. In 2015 and 2016, physical, optical, chemical and biological parameters have been recorded during three months at an Ice Camp in Qikiqtarjuaq (2015 and 2016). In 2016, a six-week cruise took place in Baffin Bay onboard the Canadian Amundsen ice breaker. Roscoff scientists are implicated in the Green Edge project to investigate biological diversity (mostly eukaryotic) and I participated personally to the Amundsen cruise.

Material and methods

A range of approaches were used to investigate the spring bloom diversity: flow cytometry, molecular biology, scanning electron microscopy and cultures.

The 2015 Ice Camp samples were analyzed on the Canto flow cytometer for natural fluorescence autotrophic populations and after labelling with SYBR green for bacterial populations. During the cruise, 1 880 samples from 125 stations were analyzed on board using an Accuri C6 flow cytometer. At the same time, samples were collected using DMSO fixation for flow cytometry sorting back in the laboratory.

For Molecular biology, three liters from 6 depths were filtered on 20 µm, 3 µm and 0.2 µm filters. These samples will be analyzed by metabarcoding, metagenomics or metatranscriptomics. Two liters were also filtered on 0.8 µm for quantitative PCR.

Three methods were used to culture phytoplankton organisms. Seawater filtered through 3 or 0.8 µm was enriched different media for eukaryotes and cyanobacteria. Cells contained in 2 liters of seawater were concentrated to small volumes using tangential filtration. Finally, enriched seawater was diluted into 96-deep-well plates in order to get statistically 1 or 10 cells per well. Filters were also collected for morphological analysis by scanning electron microscopy.

First results

Bloom dynamics off Baffin Island in 2015 and 2016

The Ice camp (in Qikiqtarjuaq, Nunavut) took place from March to early July 2015. Water was sampled in the euphotic layer at a fixed location under the ice (67.4°N, 68.8°W) every two days. I analyzed all 2015 samples by flow cytometry with the help of D. Marie. Picophytoplankton developed earlier than Nanophytoplankton in late June (Fig.1) but unfortunately, sampling was stopped too early and only allowed to catch the beginning of the phytoplankton spring bloom. In 2016, sampling could be performed until late July allowing to recover the peak of the bloom (around day 190, Fig.2).

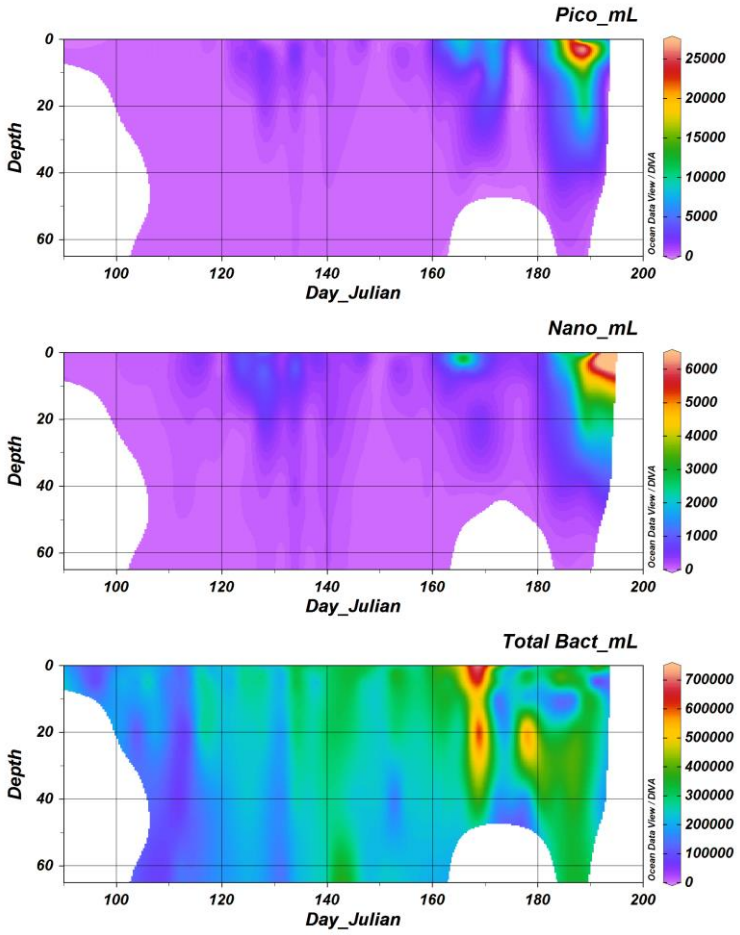


Fig.1: Evolution of pico and nano-phytoplankton as well as bacteria under the ice in 2015 measured by flow cytometry.

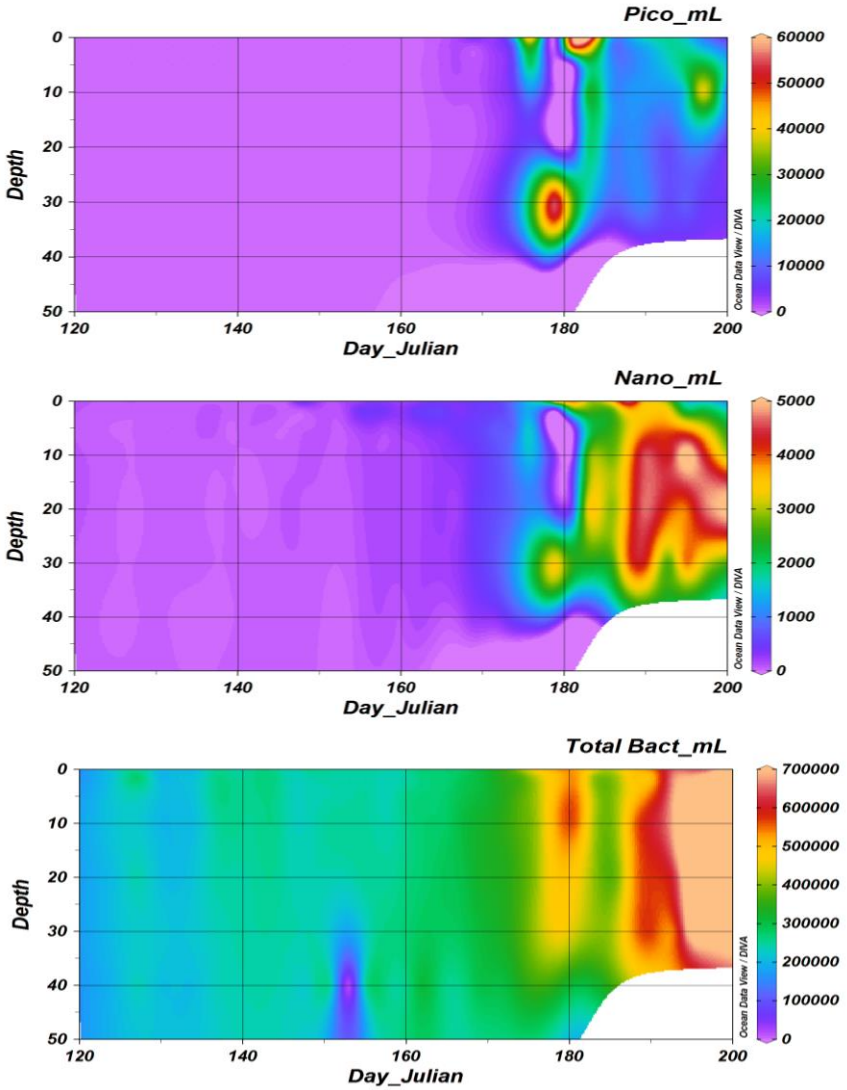


Fig.2: Evolution of pico and nano-phytoplankton as well as bacteria under the ice in 2016 measured by flow cytometry.

The 2016 Amundsen cruise

The Amundsen performed 7 transects between the latitudes 68°N and 70.5°N. Each transect went from open water to more or less compact sea ice. The ice edge moved rapidly from the East coastal of Greenland to the West coast of Canada during the cruise due to the environmental conditions (waves, temperature…).

Surface and subsurface blooms of Pico- and Nano-phytoplankters were caught during the cruise (Fig.3).

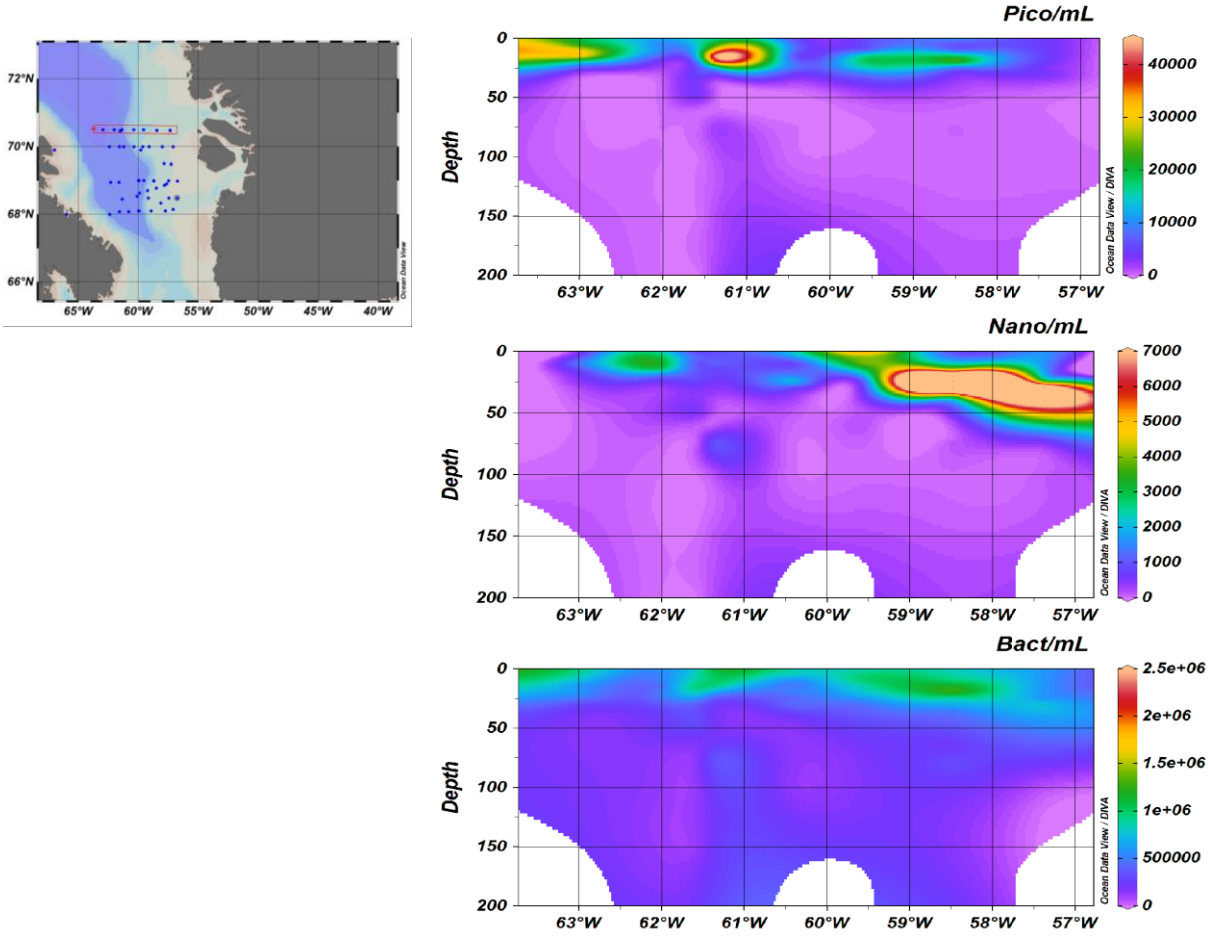


Fig.3: Abundance of bacteria, pico and nano-phytoplankton measured by flow cytometry on transect 6 (70.5°N) during the Amundsen 2016 cruise.

Communication and outreaches

During the Green Edge cruise, I was required to participate to a communication project with French school, which lead to the publication of videos. I had to answer the question: "How are the sampled preserved along a six-week scientific cruise?" (available here http://www.greenedgeproject.info/icebreaker.php). And in my free time I made some water color paintings, which were published as communication open access common support for the Green Edge project (here http://www.greenedgeproject.info/drawing.php).

Literature cited

Balzano, S., Gourvil, P., Siano, R., Chanoine, M., Marie, D., Lessard, S., et al. (2012) Diversity of cultured photosynthetic flagellates in the northeast Pacific and Arctic Oceans in summer. *Biogeosciences* **9**: 4553–4571.

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

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

Remerciements

Écrire des remerciements demande de se replonger trois ans en arrière, dans ses souvenirs et d'en ressortir les éléments clefs, ce qui peut se compliquer un peu lorsque la mémoire atteint à peine celle du poisson rouge… Ensuite, il faut prendre le temps - vous savez ce Graal que l'on cherche tous - de revivre ces moments à la lueur de nos réminiscences entachées de nos sentiments actuels. Enfin, il faut choisir ses mots pour les offrir aux lecteurs, il faut choisir jusqu'à quel point on souhaite se dévoiler. Et il y a la peur… cette peur d'avoir oublié. L'inévitable oubli. Je prends donc le parti de présenter dès à présent, mes plus plates excuses à toutes personnes qui se sentiraient lésées par la suite : Merci à vous tous !

Me voici débarquée d'une longue traversée, où j'ai pu naviguer sur de nombreux voiliers, à chacun son équipage, à chacun son odyssée (Ulysse, c'est une spéciale dédicace), à chacun son équipée. De prime abord, je souhaite remercier le capitaine de cette traversée, Daniel Vaulot, qui m'a menée à bon port, tout en me laissant naviguer, ainsi que tous les conseillers qui ont accepté de suivre ce voyage scientifique : A. Zingone de Napoli, B. Edvardsen d'Oslo, F. Viard, F. Not de Roscoff et ceux qui ont accepté d'en apprécier le dénouement : W. Eikrem, F. Leliaert, S. Frederiksen et C. Destombe.

Merci à tous les *planktonautes* (passés et présents) pour leur accueil et leur aide, tout particulièrement aux orateurs Charles, R. Siano, N. Simon, A.C. Beaudoux, C. Jeanthon pour les discussions qu'ils m'ont offertes. Merci aux soleils brésiliens Adriana et Catherine dont même l'hiver roscovite (bien qu'il soit là) n'a pas su ternir la joie de vivre. Aux autres mousses, apprentis-chercheurs et chercheuses, présents mais aussi passés car ils nous ont montré la voie : Greg, Justine…, merci de même. Je souhaiterais aussi adresser plus de remerciements que l'océan ne compte de gouttes d'eau à l'amiral administratif de l'unité, Céline Manceau sans qui nous tomberions tous de Charybde en Scylla. Et comment ne pas remercier tous mes colocataires de bureau, car un bureau, finalement, c'est presque une fratrie : merci à Théophile, Delphine, Klervi, Marie, Charles et Ruibo pour le 331 puis à ma demi fratrie d'adoption dans le 309 Théophile, Weiting et PYM.

A propos de naviguer, car tel était bien l'allusion en introduction, merci aussi à tous les *Admundsenautes* qui ont contribué à m'extraire du temps pendant les six semaines de la campagne. En particulier merci à Domi (ce n'est pas tout mais puisque tu es sensé être mon père à Québec, tu aurais dû te retrouver dans la section famille) et Pris pour m'avoir initiée au côté Tétris des missions de terrain. De la courte mais intense régate qui m'a permis de reprendre rapidement la mer, je souhaiterais adresser un humble salut aux *Tresconautes* du petit Pogo : Sarah, Fabrice, Charles (encore lui…), Gurvan (co-équipier de choc pour l'Aquathlon) et au skipper X. mais aussi à ceux de l'imposant « Maison Kervran » et en particulier à Franck.

De mes anciennes aventures, je tiens à écrire merci aux Sanctoludoviciens et -ciennes, AgroParisTechniciens et -ciennes, vétérinaires de tout poil et plume (mais surtout poil) qui m'ont fait

l’honneur de suivre mes aventures roscovites ou de visiter notre petite cité de caractère battue par les vents et marées en ma compagnie : en particulier Morgane, Tiphaine, Najda (dite Nadège ou bien est-ce l’inverse ?) et Lucas.

Car la thèse et la vie sont certes de merveilleux voyages, mais aussi des jeux, je remercie tous ceux qui ont accepté de jouer avec moi au cours de ces trois ans et trois mois, parmi les plus notables : Miriam, Solène, Joanne, Hugo et Laura, Zujaila ; Valérian, Caroline, Fred et Lôh pour les sessions de ping pong ; François, Harold, Erwann, Guillaume et tous les garçons et les filles du foot le mercredi, Loïc pour m’avoir initiée à la slackline et entraînée au foot. Et puis il y a ceux qui jouent avec les notes, merci beaucoup à tous les *Roskalonautes* en particulier Léna, Virginie, Miryam L., Laurent et à Benoît, j’espère que vous continuerez à arpenter la bibliothèque, car c’est bien plus amusant de vivre en chantant. Merci à tous les participants de l’édition 2017 de MT180 Sorbonne Universités pour l’ambiance exceptionnelle de la préparation et de la soirée de présentation et à Sofia-Elena pour m’avoir encouragé à me lancer dans cette aventure.

Je souhaite adresser un grand merci de même à Amélie, Gilda et Benoît M. de l’école d’arts martiaux du Phoenix Celtique pour tout ce que vous m’avez appris et qui me permet d’aller libre.

Merci à tous les auteurs, scientifiques ou non, à tous les compositeurs et musiciens, qui ont guidé mes pas le long des spires de la thèse. Merci aussi à tous ces inconnus, qui découvrent toujours avec des yeux émerveillés notre quotidien et se bercent de l’illusion que nous sommes tous, nous chercheurs, les héros de nos propres contes.

Pour finir, merci à ma famille surtout P&M, Armel et Suzanne, Gabrielle pour ce qu’ils sont et parce qu’ils acceptent ce que moi, je suis.

ABSTRACT

Towards an Atlas of Green micro-algae (Chlorophyta) in the ocean.

In the world ocean, the green algal lineage that dominates on land is represented by Chlorophyta which account in average for 25% of photosynthetic sequences (Dinoflagellates excluded) in global marine molecular inventories. Several lineages of Chlorophyta (especially prasinophytes) share ancestral morphological features such as the presence of scales and are considered to be close to the common ancestor of the green lineage. Although Chlorophyta are major keys for ecological understanding of the ocean, as well as the evolutionary story understanding of land plants, their diversity and distribution in marine waters has been understudied. This thesis aims at investigating the environmental diversity of marine Chlorophyta and describing their distributions based on available large scale metabarcoding datasets. First, a reference database of publicly available 18S rRNA sequences of Chlorophyta was assembled and critically curated. The next steps relied on the analysis the Ocean Sampling Day (OSD) 18S metabarcode datasets that focuses mostly on coastal waters. Autotrophic and more specifically Chlorophyta diversity was compared for a limited sample set based on two regions of the 18S rRNA gene commonly used as barcodes, the V4 and V9 regions. Then, a global analysis of Chlorophyta distribution was done using the full OSD V4 dataset. Careful taxonomic investigations using both automatic and hand checked assignation of OTUs using alignments and phylogenies allowed to confirm the existence of new environmental prasinophytes clades and to describe ecological patterns at low taxonomic levels. These analyzes confirmed that the Mamiellophyceae were the major group in coastal waters, but also highlighted that prasinophytes Clade VII and IX were dominating the oceanic oligotrophic stations. Comparing V4 and V9 regions illustrated the influence of the reference database on the diversity pictures at low taxonomic levels (for example genus or species levels) provided by different markers. The taxonomic investigation highlighted the diversity gaps between reference databases and environmental datasets. This work emphasizes the neglected importance of Chlorophyta in marine waters and provides some suggestions for future research.

Keywords: Chlorophyta, prasinophytes, Metabarcoding, 18S rRNA gene, V4 and V9 regions, diversity, distribution, phylogeny, marine environment