PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy

Johan Decelle, Sarah Romac, Rowena F Stern, El Mahdi Bendif, Adriana Zingone, Stéphane Audic, Michel D Guiry, Laure Guillou, Désiré Tessier, Florence Le Gall, et al.

Received Date : 12-Nov-2014

Revised Date : 20-Feb-2015

Accepted Date : 02-Mar-2015

Article type : Resource Article

PhytoREF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy

Johan Decelle[1,2], Sarah Romac[1,2], Rowena F. Stern[3], El Mahdi Bendif[4], Adriana Zingone[5], Stéphane Audic[1,2], Michael D. Guiry[6], Laure Guillou[1,2], Désiré Tessier[7,8], Florence Le Gall[1,2], Priscillia Gourvil[1,2], Adriana Lopes dos Santos[1,2], Ian Probert[1,2], Daniel Vaulot[1,2], Colomban de Vargas[1,2*], Richard Christen[7,8*]

1- UMR 7144 - Sorbonne Universités, UPMC Univ Paris 06, Station Biologique de Roscoff, 29680 Roscoff, France
2- CNRS, UMR 7144, Station Biologique de Roscoff, 29680 Roscoff, France
3- Sir Alister Hardy Foundation for Ocean Science, The Laboratory, Citadel Hill, Plymouth, PL1 2PB.
4- Marine Biological Association, The Laboratory, Citadel Hill, Plymouth, PL1 2PB.
5- Stazione Zoologica Anton Dohrn, Villa Comunale, 80121 Naples, Italy
6- The AlgaeBase Foundation, c/o Ryan Institute, National University of Ireland, University Rd., Galway, Ireland
7- CNRS, UMR 7138, Systématique Adaptation Evolution, Parc Valrose, BP71. F06108 Nice cedex 02, France
8- Université de Nice-Sophia Antipolis, UMR 7138, Systématique Adaptation Evolution, Parc Valrose, BP71. F06108 Nice cedex 02, France

Keywords: plastidial 16S rRNA gene, photosynthesis, high-throughput sequencing, metabarcoding, phytoplankton, protists

*To whom correspondence should be addressed. Email: vargas@sb-roscoff.fr (Tel:+33 2 98 29 25 37) & Richard.christen@unice.fr (Tel: 33 601 830 144).

Running title: 16S rDNA sequences of microalgae in PhytoREF

Abstract

Photosynthetic eukaryotes have a critical role as the main producers in most ecosystems of the biosphere. The ongoing environmental metabarcoding revolution opens the perspective for holistic eco-systems biological studies of these organisms, in particular the unicellular microalgae that often lack distinctive morphological characters and have complex life cycles. In order to interpret environmental sequences, metabarcoding necessarily relies on taxonomically-curated databases containing reference sequences of the targeted gene (or barcode) from identified organisms. To date, no such reference framework exists for photosynthetic eukaryotes. In this study, we built the PhytoREF database that contains 6,490 plastidial 16S rDNA reference sequences that originate from a large diversity of eukaryotes representing all known major photosynthetic lineages. We compiled 3,333 amplicon sequences available from public databases and 879 sequences extracted from plastidial genomes, and generated 411 novel sequences from cultured marine microalgal strains belonging to different eukaryotic lineages. 1,867 environmental Sanger 16S rDNA sequences were also included in the database. Stringent quality filtering and a phylogeny-based taxonomic classification were applied for each 16S rDNA sequence. The database mainly focuses on marine microalgae, but sequences from land plants (representing half of the PhytoREF sequences) and freshwater taxa were also included to broaden the applicability of PhytoREF to different aquatic and terrestrial habitats. PhytoREF, accessible via a web interface (http://phytoref.org), is a new resource in molecular ecology to

foster the discovery, assessment and monitoring of the diversity of photosynthetic eukaryotes using high-throughput sequencing.

Introduction

Eukaryotes that acquired photosynthesis through endosymbiosis with cyanobacteria or plastid-bearing eukaryotes are distributed across most eukaryotic super-groups and exhibit a bewildering morphological diversity across more than eight orders of magnitude in organism size (Archibald 2012; Not *et al.* 2012). Most photosynthetic eukaryotes are unicellular (referred to as protists), but a few lineages, essentially macroalgae (e.g. the rhodophyte class Florideophyceae or the chlorophyte class Ulvophyceae) and the embryophyte land plants, have evolved into multicellular forms. The radiation of photosynthetic marine protists during the Neoproterozoic arguably led to a major oxidation event in the history of the Earth system (Knoll 2014). Today, eukaryotic microalgae are key players in aquatic food webs and global biogeochemical processes. In the marine ecosystem, they are the major contributors to primary production through their capacity to perform oxygenic photosynthesis (Falkowski *et al.* 2004; Worden *et al*. 2004; Jardillier *et al*. 2010), and to export and sequester organic carbon to the deep ocean and sediments (Richardson & Jackson 2007). In addition, evidence is growing that many eukaryotic microalgal taxa are mixotrophs, being able to both photosynthesize and feed on various microbial prey (McKie-Krisberg & Sanders 2014; Unrein *et al.* 2014). Their contribution to bacterivory can even exceed that of strict heterotrophs in oceanic waters (Zubkov & Tarran 2008; Hartmann *et al.* 2012). In coastal areas, some microalgal species can be toxic and/or form harmful blooms, which can be highly detrimental to marine life and human activities such as

fisheries, aquaculture and tourism (Zingone & Wyatt 2005; Chambouvet *et al.* 2008; Anderson *et al.* 2012).

Despite their ecological and economic importance, it remains difficult to assess the total diversity of photosynthetic eukaryotes in the natural environment using classical microscopy-based techniques. For most taxa, taxonomic identification is greatly hindered by their minute size (as small as 0.8 µm for the prasinophyte *Ostreococcus*; Courties *et al.* 1994; Vaulot *et al*. 2008), lack of distinctive morphological features, and fragility when classical fixatives are used (Vaulot *et al*. 1989). The complex life cycles of many microalgal species are additional obstacles that render detection in the environment very difficult with traditional microscopy. Many taxa undergo a succession of morphologically distinct forms (e.g. sexual morphotypes, resting cysts; Montresor & Lewis 2006; Gaebler-Schwarz *et al.* 2010) or can be "hidden" within a host cell as a parasitic or mutualistic symbiont (Skovgaard *et al.* 2012; Decelle *et al.* 2012). In this context, environmental DNA metabarcoding (high-throughput sequencing of DNA markers), which has unveiled a vast and unsuspected diversity of microorganisms in recent years, provides a powerful new tool to assess the composition and ecological function of microalgal communities (Bik *et al.* 2012; Bittner *et al.* 2013). Environmental metabarcoding approaches have also been proposed for bio-assessment and bio-monitoring of sentinel or indicator species, including microalgae (Taberlet *et al.* 2012; Kermarrec *et al.* 2013; Pawlowski *et al.* 2014), and the study of diet regimes in predators (Pompanon *et al.* 2012; Piñol *et al.* 2014). For marine protists, variable regions of the nuclear ribosomal RNA genes (particularly the small subunit, 18S rRNA) are traditionally used as "universal" markers in environmental surveys (Stoeck *et al.* 2010; Logares *et al.* 2012,2014). However, several drawbacks limit the use of these nuclear markers to assess the biodiversity of photosynthetic eukaryotes: (i) some 18S rDNA clone library-based surveys

have been shown to be biased towards heterotrophic eukaryotes, and consequently tend to overlook phototrophs in complex community assemblages (Vaulot *et al*. 2002; Kirkham *et al.* 2011); (ii) ribosomal DNA of large protist cells (mainly heterotrophs and potentially multinucleated) or metazoans tends to be preferentially PCR-amplified because of the relatively higher copy number of ribosomal genes in these organisms (Zhu *et al.* 2005; Godhe *et al.* 2008); (iii) distinction between phototrophic and heterotrophic taxa is very often not possible in complex multi-functional protistan groups, such as dinoflagellates. In addition, given the extreme genetic diversity of eukaryotes, "universal" DNA markers cannot detect all lineages with a high taxonomic resolution (CBOL Protist: Pawlowski *et al.* 2012). Therefore, barcoding systems with narrower taxonomic and/or functional focus need to be developed to provide a better picture of the taxonomic and functional composition of eukaryotes in complex ecosystems.

In order to focus on the phototrophic compartment of eukaryotic communities, the photosynthetic protein-coding *psbA* (protein D1 of photosysytem-II reaction center) and *rbc*L genes (large subunit of the Ribulose-l,5-diphosphate carboxylase/oxygenase, RuBisCO) have been used as markers for phytoplankton communities (Paul *et al*. 2000; Zeidner *et al.* 2003; Man-Aharonovich *et al*. 2010). However, the primers targeted essentially cyanobacteria and cyanophages (viruses), and to a lesser extent photosynthetic eukaryotes, and the same species can have different sequence types (e.g. forms IA, IB for *rbc*L). By contrast, the plastidial 16S rRNA gene has been successfully employed in several marine surveys since it contains sufficiently conserved regions to use generalist primers to target all plastid-bearing eukaryotes and can distinguish major eukaryotic lineages with a relatively good taxonomic resolution (Fuller *et al.* 2006; MacDonald *et al.* 2007; Lepère *et al.* 2009; Shi *et al.* 2011; Kirkham *et al.* 2011, 2013). However, annotation and interpretation of the plastidial 16S rDNA clone libraries

obtained in these studies have been hindered by the lack of reference sequences of taxonomically well-identified organisms. Although a number of curated reference databases are publicly available for ribosomal RNA genes of eukaryotes and prokaryotes, such as the Protist Ribosomal Reference Database (PR2; Guillou *et al.* 2013), SILVA (Pruesse *et al*. 2007), Ribosomal Database Project (Cole *et al.* 2005), and Greengenes (DeSantis *et al.* 2006), no reference database exists for the plastidial 16S rRNA gene of photosynthetic eukaryotes. Here, we describe an extensive reference database of the plastidial 16S rRNA gene including sequences from all major lineages of photosynthetic eukaryotes, comprising terrestrial, freshwater and marine organisms. This database, named PhytoREF, has been built through the compilation of all of the publicly available plastidial 16S rDNA sequences (amplicons and sequences extracted from plastidial genomes), as well as novel Sanger amplicons that we obtained from a wide taxonomic spectrum of cultured microalgal strains. PhytoREF is not only a new resource to explore, evaluate and monitor the diversity of photosynthetic eukaryotes in aquatic and terrestrial ecosystems, but is also useful to taxonomically identify new plastidial 16S rDNA sequences and design primers and probes to target specific lineages of photosynthetic eukaryotes. PhytoREF will pave the way for a range of applications in bio-monitoring photosynthetic eukaryotes in various habitats (e.g. water, sediments and ice), paleoecological studies of primary producers in past environments, and dietary studies in unicellular and multicellular herbivores.

Data sources

Retrieval of plastidial 16S rDNA sequences from public databases

Plastidial 16S rDNA sequences were first retrieved from the International Nucleotide Sequence Database Collaboration (INSDC: http://www.insdc.org) using various keywords (e.g. plastidial,

plastid, chloroplast, 16S, small subunit), and BLAST searches with different query sequences of distinct photosynthetic eukaryotic lineages. Additional sequences were retrieved from the PR2 database (May 2014) (Guillou *et al.* 2013; http://ssu-rrna.org/). When available in GenBank (release 201), the source literature for each sequence was searched to compile and/or verify their specific features (e.g. taxon names, culture strains), resulting in a bibliographic database of 565 source publications. 16S rDNA sequences were also extracted from all plastidial genomes available at http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=2759&opt=plastid. All plastidial 16S rDNA sequences that originated from an identified organism (e.g. culture strain or isolated organism) were defined as reference sequences in the PhytoREF database. Environmental Sanger 16S rDNA sequences obtained in clone libraries were also retrieved from INSDC and included in PhytoREF. Those sequences that lacked taxonomic identification were assigned at the class level based on sequence similarity scores with the references sequences of PhytoREF. Finally, all of the 16S rDNA sequences of Cyanobacteria were extracted from SILVA (release 115, Quast *et al.* 2013) in order to root the phylogenetic trees and unambiguously annotate and classify eukaryotic sequences. The cyanobacterial sequences are available as separate files at http://phytoref.org.

Newly generated 16S rDNA sequences from microalgal cultures

We generated 411 novel plastidial 16S rDNA sequences from eukaryotic microalgal strains from the Roscoff Culture Collection (RCC, http://roscoff-culture-collection.org/), the NCMA (formerly CCMP; https://ncma.bigelow.org/), and the culture collection of the Stazione Zoologica Anton Dohrn of Naples (Table S1). Cultured cells were harvested in exponential growth phase and concentrated by centrifugation. Total nucleic acids were extracted using the

Nucleospin RNA II kit (Macherey-Nagel) and quantified using a Nanodrop ND-1000 Spectrophotometer (Labtech International). An 850 bp fragment was PCR-amplified with a generalist photosynthetic eukaryote primer set biased against cyanobacteria: PLA491F: 5'- GAG GAA TAA GCA TCG GCT AA -3′ (Fuller *et al.* 2006), and OXY1313R: 5′- CTT CAY GYA GGC GAG TTG CAG C -3′ (West *et al.* 2001). PCR amplifications were performed with the Phusion high-fidelity DNA polymerase (Finnzymes) in a 25-µL reaction volume, using the following PCR parameters: 30 s at 98°C; followed by 35 cycles of 10 s denaturation at 98°C, 30 s annealing at 60°C, and 30 s extension at 72 °C; with a final elongation step of 10 min at 72 °C. PCR products were purified by either EXOSAP-IT (GE Healthcare Bio-Sciences Corp.) or the NucleoSpin® Extract II kit (Macherey-Nagel, Hoerdt, France), and sequenced in both forward and reverse directions using the ABI-PRISM Big Dye Terminator Cycle Sequencing Kit (Applied Biosystems). Raw Sanger sequences were edited and assembled with ChromasPro v1.7.5 (Gene Codes), and primer sequences were trimmed off. The new plastidial 16S rDNA sequences were deposited in GenBank under the accession numbers LN735194 to LN735532 (Table S1), and can also be retrieved on the PhytoREF web interface at http://phytoref.org.

Construction of the PhytoREF database

The core content of the database is composed of the reference plastidial 16S rDNA sequences from public databases with unambiguous taxonomic assignation, and the novel sequences obtained from duly identified cultures. Each reference sequence, including taxonomic affiliation, was validated and filtered following different steps: (i) sequences shorter than 400 bp from cultures, and shorter than 800 bp from public sequences (including environmental sequences) were removed; (ii) sequences with more than 10 consecutive non-ACGT characters were also

discarded; (iii) sequence alignments were performed for different well-defined taxonomic groups (e.g. at the class level) using MAFFT v6. 953b with default options (Katoh *et al.* 2002), and visualized to verify the presence of introns or putative chimeric sequences; (iv) poorly aligned or difficult-to-align nucleotide positions were removed for subsequent phylogenetic analyses using the program trimAl v1.4 program (with a -gt value of 0.8, and -st value of 0.001; Capella-Gutierrez 2009); (v) phylogenetic trees were constructed separately for each taxonomic group (i.e. generally at the class level) using FastTree v.2.1.1, a fast and accurate approximate maximum-likelihood method using the GTR model (Price *et al.* 2010), in order to identify mislabeled sequences and other possible conflicts, and to build up the taxonomic framework (see below).

Additional publicly available plastidial 16S rDNA sequences with uncertain taxonomic status were subsequently added to this validated core dataset. These sequences were assigned to a given phylum using a similarity threshold based on global pair-wise alignments (using a Needleman-Wunsch algorithm) against the reference sequences. Sequences of each phylum were then aligned based on conserved 2D structures and sequences of the archaeal 16S, bacterial 16S and eukaryotic 18S small subunit ribosomal RNA using the SSU-align program and Infernal software package, which generate large-scale alignments of up to millions of sequences (Nawrocki *et al.* 2009). 2D-based alignments allowed us to verify whether the new sequences corresponded to the 16S rRNA gene or other ribosomal genes. Phylogenetic trees were then built using BioNJ as implemented in Seaview v.4 (Gouy *et al.* 2010), and visualized using TreeDyn (Chevenet *et al.* 2006). Functions implemented in TreeDyn as well as specific Python scripts allowed us to determine the taxonomic level of each sequence (e.g at the "Family" level). All sequences

included in PhytoREF have two unique identifiers, the GenBank accession number and a PhytoREF ID number.

The taxonomic framework of PhytoREF

For every new validated sequence, we established a standardized and ranked taxonomy with 10 levels: 1- Domain; 2- Super-group; 3- Phylum; 4- Class; 5- Subclass; 6- Order; 7- Sub-order; 8- Family; 9- Genus; and 10- Species. For the "Super-group", "Phylum" and "Class" levels, the taxonomic framework of PhytoREF was derived from the PR2 database (http://ssu-rrna.org/; Guillou *et al.* 2013), which mainly follows a comprehensive recent classification framework of eukaryotes (Adl *et al.* 2012). The "Family" and "Order" levels of terrestrial, marine and freshwater micro- and macroalgae were based on the taxonomic classification system of the AlgaeBase database (Guiry and Guiry 2014; Guiry *et al*. 2014; http://www.algaebase.org/). For the taxa that were not present in AlgaeBase and PR2 (mostly embryophytes), the taxonomic classification of NCBI (May 2014) was followed (taxdump.nodes and taxdump.names files at http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NCBI/metarepresentation.html). Overall, the standardized taxonomic framework established in PhytoREF was designed to assist in the analysis of large datasets of environmental plastidial 16S rDNA amplicons generated by high-throughput environmental metabarcoding.

For some 16S rDNA sequences, it was not possible to define an accurate and/or complete taxonomic identity because the taxonomic description of the corresponding organism is not fully resolved, e.g. only at the "Family" or "Genus" level. In these cases, the sequence was labeled as described for the PR2 database. For instance, the taxonomic path of a sequence identified up to the "Family" level would be: Family, Family_X (for the "Genus" level), and Family_XX (for the

"Species" level). Moreover, some key groups of microalgae have only been classified into informal clades and sub-clades based on published phylogenetic analyses without morphology-based taxonomy (e.g. prasinophyte clades VII, IX; Apicomplexa-related lineages I -V). For the PhytoREF database, information about the molecular clade was verified through specific phylogenetic analyses with the 16S rRNA gene and indicated at different taxonomic ranks. For instance, prasinophytes belonging to clade VII and sub-clade A1 are annotated: clade 7 ("Order" level), clade_7A ("Family" level), clade_7A1 ("Genus" level) and clade_7A1+sp ("Species" level). A confidence level for taxonomic assignation (named Refseq) was given to each PhytoREF sequence, indicating the level at which a given sequence is unambiguously assigned (RefSeq=1: Eukaryota; RefSeq=2: super-group; RefSeq=3: Phylum; RefSeq=4: Class; RefSeq=5:Order; RefSeq=6:Family; RefSeq=7:Genus). Finally, we also included 16S rDNA sequences originating from symbiotic microalgae or kleptoplastids found in hosts. The origin of these sequences that are generally incorrectly assigned to the host in public databases was modified and marked as "symbiont" or "kleptoplastid" in the PhytoREF database.

Results and Discussion

Overview of PhytoREF database

The PhytoREF database (release 1) currently contains 6,490 partial and complete plastidial 16S rDNA sequences (of which 6,051 sequences are > 800 bp long). In total, 411 novel sequences from marine microalgal strains were produced in this study and 6,079 sequences retrieved from public databases (5,200 amplicons from organisms and environmental samples, and 879 sequences extracted from plastidial genomes, Fig. 1). 2D alignments combined with BLAST analyses allowed us to determine that 52 sequences (mostly from streptophytes) considered as

16S rDNA in GenBank were actually nuclear 18S rDNA, and were therefore excluded from PhytoREF. In addition to sequences from identified plastid-bearing organisms, PhytoREF contains 1,867 environmental Sanger sequences from clone libraries, which have been assigned to known eukaryotic lineages based on sequence similarity (combining a Needleman-Wunsch algorithm and phylogenetic analyses). Every PhytoREF sequence was quality-checked, phylogenetically-analyzed and classified following our standardized taxonomy. In addition to the taxonomic path, all sequences were associated to a suite of descriptors, such as the organism, molecular origin (amplicon or extracted from genomes), GenBank accession number, cultured strain, and original publication. Additional categories indicated whether sequences are environmental or belong to morphologically identified organisms, and if they correspond to kleptoplastids, parasitic or mutualistic microalgae (in such cases, the taxonomic name of the host is also provided).

Taxonomic composition of PhytoREF

All of the known major lineages of photosynthetic eukaryotes from terrestrial, freshwater and marine environments are represented in PhytoREF. At the super-group level, the composition of the database is as follows: Archaeplastida (3,834 sequences), Stramenopila (1,704 sequences), Alveolata (144 sequences), Hacrobia (501 sequences), Excavata (288 sequences), and Rhizaria (20 sequences) (Figs. 2 and 3). Although our effort while building PhytoREF (in particular in producing novel plastidial reference sequences) mainly focused on marine microalgal taxa, reference sequences from streptophytes (i.e. from mosses and ferns to gymnosperms and angiosperms), and marine and freshwater macroalgae (e.g., Rhodophyceae, Phaeophyceae, Ulvophyceae) were also included in the final reference database of all known photosynthetic

eukaryotes. Thus, PhytoREF can be used in metabarcoding surveys to study communities of microalgae in different marine and freshwater habitats (e.g. seawater, estuaries, brackish waters, lakes), as well as to detect the presence of macroalgae and streptophytes in aquatic systems as reproductive stages (gametes, pollen) or in the digestive tracts of herbivores.

Land plants (streptophytes) are numerically the dominant group in PhytoREF with 2,973 sequences, representing 373 families and 796 genera. The macroalgae from the classes Rhodophyceae (61 genera), Phaeophyceae (7 genera), and Ulvophyceae (18 genera) are represented by 161, 46, and 82 sequences, respectively (Fig. 3). Diatoms (Bacillariophyta), green algae (Chlorophyta) and Haptophyta (coccolithophores and their relatives) are numerically the most important microalgae in the database with 1094, 653 and 369 sequences, respectively, covering a wide taxonomic diversity with 109, 79, and 34 described genera and 268, 113, and 67 described species, respectively (Fig. 3). Within the Haptophyta, one environmental sequence found in the Pacific Ocean, named "S25_1200" (EF574856), was included as it has been identified to form a novel photosynthetic lineage (Janouškovec *et al.* 2011). For the green algae, freshwater taxa are less represented than their marine relatives and represent an obvious target for future reference sequencing. Amongst the Hacrobia, there are 126 sequences of cryptophytes covering 36 described species from marine (e.g. *Rhodomonas* sp.), brackish (e.g. *Chroomonas* sp. and *Geminigera* sp.) and fresh (e.g. *Cryptomonas* sp.) waters. Because genetic diversity does not correspond well with taxonomic features and life-stages are likely to be complex, the systematics of cryptophytes are still under revision, except for the genera *Cryptomonas* and *Hemiselmis* that have been relatively well delineated (Hoef-Emden & Melkonian 2003; Hoef-Emden 2005). The euglenozoans from the super-group Excavata are also well represented in PhytoREF with 115 described species from freshwater and marine habitats, such as species of

Euglena, *Monomorphina* and *Trachelomonas*. Of note, PhytoREF also contains 6 sequences of the recently discovered rappemonads (Hacrobia), an uncultured microalgal group widely distributed in marine and fresh waters, but taxonomically undescribed (Kim *et al.* 2011). Since no nuclear ribosomal (18S rRNA gene) and genomic sequences are available for the rappemonads, the plastidial 16S rRNA gene is currently the only genetic marker available for evolutionary and environmental studies of this lineage.

During the course of evolution, photosynthesis has been lost in several lineages of plants and single-celled eukaryotes, but a vestigial plastid containing a 16S rRNA gene has been retained in some taxa (Williams & Keeling 2003). Some of these non-photosynthetic organisms present in PhytoREF are very often parasites, such as the holoparasitic angiosperm *Epifagus virginiana*, the heterotrophic euglenid *Euglena longa*, and the green alga *Helicosporidum*. In particular, 33 sequences correspond to the non-photosynthetic alveolate apicomplexans (e.g. *Plasmodium*, *Toxoplasma*, *Babesia*), which are obligate intracellular parasites of metazoans and protists, but which have kept a relict plastid, known as the apicoplast (Lim & MacFadden 2010; MacFadden 2014). Apicomplexan-related lineages, called ARLs (class Colpodellid), which include the microalgae *Chromera* (Moore *et al.* 2008) and *Vitrella* (Oborník *et al.* 2012), are also represented in PhytoREF by 77 sequences (mostly environmental), and classified according to the framework proposed by Janouškovec *et al.* (2011) i.e. ARL I, II etc.

As in apicomplexans and ARLs, the plastidial 16S rRNA gene of photosynthetic dinoflagellates is rapidly-evolving and their sequences are very difficult to align. This may be related to the unique genomic organization of plastid genes in dinoflagellates that can be found separately in small minicircles (Zhang *et al.* 2002; Green 2011). This extreme genetic divergence may explain the very low PCR amplification success rates we obtained during the present study on different

cultures of photosynthetic dinoflagellates. Consequently, one shortcoming of PhytoREF is the limited number of dinoflagellate sequences (34 sequences representing 15 genera), a caveat to consider when interpreting metabarcoding datasets using PhytoREF. It is important to note that several plastidial 16S rDNA sequences from GenBank were re-assigned in the database because they were mislabeled as "dinoflagellate" when in fact they correspond to plastids of photosynthetic eukaryotes "stolen" by dinoflagellate hosts (kleptoplastids). For instance, the dinoflagellate *Dinophysis* can sequester plastids of different microalgal prey, such as cryptophytes, raphidophytes and chlorophytes (Kim *et al.* 2012). This issue was also found in all other organisms present in PhytoREF that can either establish kleptoplastidy (e.g. the Ciliata *Mesodinium rubrum* and benthic Foraminifera), or photosymbiosis with microalgal cells (e.g. the katablepharid *Hatena arenicola*). Re-assignment of these sequences was necessary to avoid biases in annotation of the metabarcoding reads. Finally, 2700 16S rDNA cyanobacterial sequences have also been included in PhytoREF as separate files to avoid any ambiguities in the taxonomic assignation of query sequences. These sequences were clustered at different similarity levels (from 98 to 80%) and the longest sequences of each cluster are available for download at http://phytoref.org/.

PhytoREF: a new tool to explore the ecology of photosynthetic eukaryotes

To date, PhytoREF is the only tool in molecular ecology specifically designed to explore the total diversity of photosynthetic eukaryotes from complex marine and terrestrial ecosystems using metabarcoding or metagenomics approaches. Although the taxonomic resolution of the plastidial 16S rDNA barcode is not as high as that of established barcodes like the mitochondrial cytochrome *c* oxydase I gene for animals (Herbert *et al.* 2003) and the large subunit of ribulose

1,5-bisphosphate carboxylase gene (*rbc*L) for plants (CBoL Plant Working Group 2009), it can recover and distinguish all photosynthetic eukaryotes at the class level, family level (e.g. Cryptophyta; Stern *et al*. 2014), and down to the genus and sometimes species level for most major lineages, such as the haptophytes (Edvardsen *et al*. 2011), euglenozoans (Linton *et al.* 2010; Na *et al*. 2012), and diatoms (Pillet *et al*. 2011). As proposed by the CBoL Protist Working Group for the 18S rRNA (Pawlowski *et al.* 2012), the 16S rRNA gene can be used as a "pre-barcode" to explore the diversity of photosynthetic eukaryotes in the environment.

In this study, we found that the copy number of the 16S rRNA gene in plastid genomes can range from 1 to 10 (e.g. 4 and 6 copies in the euglenophyte *Euglena gracilis* and the prasinophyte *Pedinomonas minor*, respectively). However, in about 80% of the plastid genomes of eukaryotes (mainly streptophytes) sequenced so far, only 2 copies of the 16S rRNA were found (Table S2), which is in accordance with the plastid genome structure with two inverted repeats that duplicate ribosomal RNA genes (Green 2011). The copy number variation of the plastidial 16S rRNA gene seems therefore to be much less important than that of the nuclear 18S rRNA gene, which correlates with genome size, cell size and biovolume and can vary by up to four orders of magnitude (e.g., the green algae *Prasinococcus* sp. and *Ostreococcus* sp. have 2 and 4 copies of the 18S RNA gene, respectively, while the diatoms *Ditylum* sp. and *Coscinodiscus* sp. have >30,000 copies; Zhu *et al.* 2005; Godhe *et al.* 2008). Thus, the plastidial 16S rRNA gene has the potential to be a suitable proxy in metabarcoding studies for assessing the relative abundance of eukaryotic phototrophs in the environment. Nevertheless, one has to consider that biological biases may also occur with the plastidial 16S rRNA gene. The number of plastids can vary (hence the number of 16S copies per individual) not only within one cell among eukaryotes, but also throughout the life cycle of a species (e.g. before and after cytokinesis). Although most

species in many microalgal groups (e.g. haptophytes, cryptophytes, chlorophytes, pennate diatoms) have only one or a few plastids, some taxa can harbour more than 100 plastids (e.g. centric diatoms). Less is known about the number of plastid genome copies in microalgal species, which can also alter the 16S rDNA copy number per individual. Photosynthetic eukaryotes typically maintain 50-100 copies of the plastid genomes per plastid. This number varies greatly in land plants from tens to hundreds during the plant development (Oldenburg & Bendich 2004). In microalgae, the plastid of the chlorophyte *Chlamydomonas reinhardtii* contains about 75 genome copies (Armbrust 1998), but continuous replication and accumulation of plastid DNA throughout the cell cycle has been shown for this taxon and for the dinoflagellate *Amphidinium operculatum* and the chrysophyte *Ochromonas* (Coleman & Nerozzi 1999; Hiramatsu *et al*. 2006; Koumandou & Howe 2007).

Description of the PhytoREF web interface

The PhytoREF web interface provides easy and rapid access to all reference plastidial 16S rDNA sequences, and allows users to explore the database with interactive graphs (e.g. Krona pie charts) and perform different search options. Sequences can be retrieved in Fasta format either by taxonomic rank (e.g phylum, genus, species) using a taxonomy browser, or through specific identifiers such as the GenBank accession number or the culture strain code (e.g. AY702161 or RCC393). Information associated with each sequence can be also downloaded as a tab-separated file, and different formats of the database are proposed to be used by the QUIIME, Mothur or TreeDyn programs. In addition, for each sequence, publication metadata such as title, authors and abstract are available on the web site. Web links to the Roscoff Culture Collection and GenBank database provide more information about the taxonomy and the origin of each

plastidial 16S rDNA sequence. Finally, a BLAST interface is available on the web site allowing users to identify individual or multiple plastidial 16S rDNA sequences against all PhytoREF reference sequences, and download selected hit sequences. In order to improve future releases of PhytoREF, users are encouraged to indicate errors and suggest better taxonomic placements for reference sequences in a dedicated page.

Conclusion and perspectives

PhytoREF is the first resource allowing exploration of the total diversity of photosynthetic eukaryotes in any given ecosystem. It can be used for a range of purposes, such as: (i) annotation and classification of new plastidial 16S rDNA sequences; (ii) taxonomic assignation of environmental sequences from massive metabarcoding and metagenomic datasets; and (iii) design of primers and probes to target any group of photosynthetic eukaryotes. All of the main eukaryotic lineages that have a functional or relict plastid are represented in PhytoREF, including free-living, mutualistic or parasitic organisms from aquatic and terrestrial habitats. Some organisms are under-represented in the current PhytoREF version, such as dinoflagellates and freshwater microalgae, which may lead to coarse taxonomic assignations. PhytoREF has therefore the potential to be used for many applications from biomonitoring of photosynthetic eukaryotes in past and present environments (water, sediment, ice) to feeding selectivity studies. Updates of the database will be performed every six months by adding new reference sequences, expert validation of new public sequences from GenBank, and inclusion of novel taxonomic features from the literature, such as the description of novel algal classes..

Acknowledgments

This work was supported by the OCEANOMICS project, funded by the French Government and managed by the *Agence Nationale de la Recherche*, under the grant agreement "*Investissements d'Avenir*" ANR-11-BTBR-0008. Financial support for this work was also provided by the European Union programs MicroB3 (UE-contract-287589) and MaCuMBA (FP7-KBBE-2012-6-311975) and the French "*Investissements d'Avenir*" program EMBRC-France. This work was granted access to the HPC and visualization resources of "*Centre de Calcul Interactif*" hosted by "*Université Nice Sophia Antipolis*". We are grateful to Ales Horák, Jan Janouškovec, Chris Jackson, Akira Kuwata, Camille Poirier and Daphné Grulois for providing 16S rDNA sequences and microalgal strains.

Data Accessibility

All of the plastidial 16S rDNA sequences in the PhytoREF database can be downloaded via the web interface http://phytoref.org/, and the accession numbers of the newly produced reference 16S rDNA sequences are provided in Table S1.

Supporting Information

Table S1: List of the eukaryotic microalgal strains from which the plastidial 16S rDNA sequence have been obtained in this study by DNA extraction and PCR.

Table S2. Number of 16S rRNA copies found in public plastidial genomes in photosynthetic eukaryotes (essentially land plants).

Author Contributions

JD, SR and CdV conceived and designed the research, and wrote the paper. JD, SR, RS, MB, AZ, FG, PG, ALS performed the experiments. JD, SR, DV, RC analyzed data. MG, LG, IP provided resources and databases. RC and DT designed the web site platform of PhytoREF.

References

Adl SM, Simpson AGB, Lane CE, Lukes J, Bass D, Bowser SS, Brown MW, Burki F, Dunthorn M, Hampl V et al. (2012) The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology*, **59**, 429‐493.

Armbrust EV (1998) Uniparental inheritance of chloroplast genomes. In The Molecular Biology of Chloroplasts and Mitochondria in *Chlamydomonas*, S. Merchant, ed (Amsterdam: Kluwer), pp. 93–113.

Anderson DM, Cembella AD, Hallegraeff GM (2012) Progress in understanding harmful algal blooms: paradigm shifts and new technologies for research, monitoring, and management. *Annual Review of Marine Science*, **4**, 143-176.

Archibald J (2012) The evolution of algae by secondary and tertiary endosymbiosis. *Advances in Botanical Research*, **64**, 87-118.

Bik HM, Porazinska D L, Creer S, Caporaso JG, Knight R, Thomas WK (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, **27**, 233-243.

Bittner L, Gobet A, Audic S, Romac S, Egge ES, Santini S, Ogata H, Probert I, Edvardsen B, de Vargas C (2013) Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. *Molecular Ecology*, **22**, 87-101.

Burki F, Keeling PJ (2014) Rhizaria. *Current Biology*, **24**, R103-R107.

Capella- Gutiérrez S, Silla- Martínez JM, Gabaldón T (2009) TrimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972-1973.

CBoL Plant Working Group (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 12794-12797.

Chambouvet A, Morin P, Marie D, Guillou L (2008) Control of toxic marine dinoflagellate blooms by serial parasitic killers. *Science*, **322**, 1254-1257.

Chevenet F, Brun C, Banuls AL, Jacq B, Christen R (2006) TreeDyn: towards dynamic graphics and annotations for analyses of trees. *BMC Bioinformatics*, **7**, 439.

Cole JR Chai B Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Research,* **33**, 294-296.

Coleman AW, Nerozzi AM (1999) Temporal and spatial coordination of cells with their plastid component. *International Review of Cytology*, **193**, 125-164.

Courties C, Vaquer A, Trousselier M, Lautier J, Chrétiennot-Dinet M-J, Neveux J, Machado C, Claustre H (1994) Smallest eukaryotic organism. *Nature*, **370,** 255.

Decelle J, Probert I, Bittner L, Desdevises Y, colin S, de Vargas C, Gali M, Simo R, Not F (2012) An original mode of symbiosis in open ocean plankton. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 18000-18005.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, **72**, 5069-5072.

Edvardsen B, Eikrem W, Throndsen J, Sáez AG, Probert I, Medlin LK (2011) Ribosomal DNA phylogenies and a morphological revision provide the basis for a revised taxonomy of the Prymnesiales (Haptophyta). *European Journal of Phycology*, **46**, 202-228.

Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJ (2004) The evolution of modern eukaryotic phytoplankton. *Science*, **305**, 354-360.

Fuller NJ, Tarran GA, Cummings DG, Woodward EMS, Orcutt KM, Yallop M et al. (2006a). Molecular analysis of photosynthetic picoeukaryote community structure along an Arabian Sea transect. *Limnology and Oceanography*, **51**, 2502-2514.

Fuller NJ, Campbell C, Allen DJ, Pitt FD, Zwirglmaier K, Le Gall F et al. (2006b). Analysis of photosynthetic picoeukaryote diversity at open ocean sites in the Arabian Sea using a PCR biased towards marine algal plastids. *Aquatic Microbial Ecology*, **43**, 79-93.

Gaebler-Schwarz S, Davidson A, Assmy, P, Chen JX, Henjes J, Nothig EM, Lunau M, Medlin LK (2010) A new cell stage in the haploid-diploid life cycle of the colony-forming haptophyte *Phaeocystis antarctica* and its ecological implications. *Journal of Phycology*, **46**, 1006-1016.

Godhe A, Asplund ME, Härnström K, Saravanan V, Tyagi A, et al. (2008) Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Applied and Environmental Microbiology*, **74**, 7174-7182.

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, **27**, 221-224.

Green BR (2011) Chloroplast genomes of photosynthetic eukaryotes. *The Plant Journal*, **66**, 34-44.

Guillou L, Bachar D, Audic S, Bass D, Berney C et al. (2012) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, 1-8: doi:10.1093/nar/gks1160

Guiry MD, Guiry GM (2014) *AlgaeBase*. World-wide electronic publication, National University of Ireland, Galway. http://www.algaebase.org; searched on 18 April 2014.

Guiry MD, Guiry, GM, Morrison L, Rindi F, Valenzuela Miranda S, Mathieson AC, Parker BC, Langangen A, John DM, Bárbara I, Carter CF, Kuipers P, Garbary DJ (2014) AlgaeBase: an on-line resource for algae. *Cryptogamie Algologie*, **35**, 105-115.

Hartmann M, Grob C, Tarran GA, Martina AP, Burkill PH, Scanlan DJ et al. (2012) Mixotrophic basis of Atlantic oligotrophic ecosystems. *Proceedings National Academy Science USA*, **109,** 5756-5760.

Herbert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Series B-Biological Sciences*, **270**, 313-321.

Hiramatsu T, Nakamura S, Misumi O, Kuroiwa T, Nakamura S (2006) Morphological changes in mitochondrial and chloroplast nucleoids and mitochondria during the *Chlamydomonas reinhardtii* (*Chlorophyceae*) cell cycle. *Journal of Phycology*, **42**, 1048-1058.

Hoef-Emden K (2005) Multiple independent losses of photosynthesis in the genus *Cryptomonas* (Cryptophyceae) - combined phylogenetic analyses of DNA sequences of the nuclear and the nucleomorph ribosomal operons. *Journal of Molecular Evolution*, **60**, 183-195.

Hoef-Emden K, Melkonian M (2003) Revision of the genus *Cryptomonas* (Cryptophyceae): a combination of molecular phylogeny and morphology provides insights into a long-hidden dimorphism. *Protist*, **154**, 371-409.

Janouškovec J, Horák A, Barott KL, Rohwer FL, Keeling PJ (2012) Global analysis of plastid diversity reveals apicomplexan-related lineages in coral reefs. *Current Biology*, **22**, R518-R519.

Jardillier L, Zubkov MV, Pearman J, Scanlan DJ (2010) Significant CO2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *The ISME Journal*, **4**, 1180-1192.

Katoh M, Kuma M (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059-3066.

Kermarrec L, Franc A, Rimet F, Chaumeil P, Humbert JF, Bouchez A (2013) Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Molecular Ecology Resources*, **13**, 607-619.

Kim E, Harrison JW, Sudek S, Jones MDM, Wilcox HM, Richards TA, Worden AZ, Archibald JM (2011) Newly identified and diverse plastid-bearing branch on the eukaryotic tree of life. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 1496-1500.

Kim M, Kim S, Yih W, Park MG (2012) The marine dinoflagellate genus *Dinophysis* can retain plastids of multiple algal origins at the same time. *Harmful Algae*, **13**, 105-111.

Kirkham AR, Jardillier LE, Tiganescu A, Pearman J, Zubkov MV, Scanlan DJ (2011) Basin-scale distribution patterns of photosynthetic picoeukaryotes along an Atlantic Meridional Transect. *Environmental Microbiology*, **13**, 975-990.

Kirkham AR, Lepère C, Jardiller LE, Not F, Bouman H, Mead A, Scanlan DJ (2013) A global perspective on marine photosynthetic picoeukaryote community structure. *The ISME Journal*, **7**, 922-936.

Knoll AH (2014) Paleobiological perspectives on early eukaryotic evolution. *Cold Spring Harb Perspect Biol*, **6**, a016121.

KoumandouVL, Howe CJ (2007) The copy number of chloroplast gene minicircles changes dramatically with growth phase in the dinoflagellate *Amphidinium operculatum*. *Protist*, **158**, 89-103.

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

Lim L, MacFadden GI (2010) The evolution, metabolism and functions of the apicoplast. *Philosophical transactions of the Royal Society B*, **365**, 749-763.

Linton EW, Karnkowska-Ishikawa A, Kim JI, Shin W, Bennett MS, Kwiatowski J, Zakryś B, Triemer RE (2010) Reconstructing euglenoid evolutionary relationships using three genes: nuclear SSU and LSU, and chloroplast SSU rDNA sequences and the description of *Euglenaria* gen. nov. (Euglenophyta). *Protist*, **161**, 603-619.

Logares R, Audic S, Bass D, Bittner L, Boutte, C, Christen R et al. (2014) Patterns of rare and abundant marine microbial eukaryotes. *Current Biology*, **24**, 813-821.

Logares R, Audic S, Santini S, Pernice MC, de Vargas C, Massana (2012) Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. *The ISME Journal*, **6**, 1823-1833.

MacFadden GI (2014) Apicoplast. *Current Biology*, **24**, R262-263.

Man-Aharonovich D, Philosof A, Kirkup BC, Le Gall F, Yogev T, Berman-Frank I, Polz MF, Vaulot D, Béjà O (2011) Diversity of active marine picoeukaryotes in the Eastern Mediterranean Sea unveiled using photosystem-II *psbA* transcripts. *The ISME journal*, **4**, 1044-1052.

McKie-Krisberg ZM, Sanders RW (2014) Phagotrophy by the picoeukaryotic green alga *Micromonas*: implications for Arctic Oceans. *The ISME Journal*, **8**, 1953-1961.

McDonald SM, Sarno D, Scanlan DJ, Zingone A (2007) Genetic diversity of eukaryotic ultraphytoplankton in the Gulf of Naples during an annual cycle. *Aquatic Microbial Ecology*, **50**, 75-89.

Montresor M and Lewis J (2006) Phases, stages, and shifts in the life cycles of marine phytoplankton. In Subba Rao, D. (ed.), Algal Cultures, Analogues of Blooms and Applications. Science Publishers, Enfield, USA, pp. 91-129

Moore RB, Oborník M, Janouškovec J, Chrudimsky T, Vancova M, et al. (2008) A photosynthetic alveolate closely related to apicomplexan parasites. *Nature*, **451**, 959-963.

Na X, Shaojun P, Tifeng S, Feng L, Xiaobo Z, Suqin G (2012) Molecular identification and culture trials of *Eutreptiella gymnastica* (Eutreptiales, Euglenophyceae). *Chinese Journal of Oceanology and Limnology,* **30**, 446-455.

Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: Inference of RNA Alignments. *Bioinformatics*, **25**, 1335-1337.

Not F, Siano R, Kooistra WHCF, Simon N, Vaulot D, Probert I (2012) Diversity and ecology of eukaryotic marine phytoplankton. *Advances in Botanical Research*, **64**, 1-53.

Oborník M, Modry D, Lukeš M, Cernotikova-Stribrna E, Cihlar J, et al. (2012) Morphology, ultrastructure and life cycle of *Vitrella brassicaformis* n. sp., n. gen., a novel chromerid from the Great Barrier Reef. *Protist*, **163**, 306-323.

Oldenburg DJ, Bendich AJ (2004) Changes in the structure of DNA molecules and theamount of DNA per plastid during chloroplast development in maize. *Journal of Molecular Biology*, **344**, 1311-1330.

Paul JH, Alfreider A, Wawrik B (2000) Micro- and macrodiversity in *rbc*L sequences in ambient phytoplankton populations from the southeastern Gulf of Mexico. *Marine Ecology Progress Series*, **198**, 9-18.

Pawlowski J, Esling P, Lejzerowicz F, Cedhagen T, Wilding TA (2014) Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Molecular Ecology Resources*, **14**, 1129-1140.

Pawlowski J, Audic S, Adl S, Bass D, Belbahri L et al. (2012) CBOL Protist Working Group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology* 10(11): e1001419.

Pillet L, De Vargas C, Pawlowski J (2011) Molecular identification of sequestered diatom chloroplasts and kleptoplastidy in Foraminifera. *Protist*, **162**, 394-404.

Piñol J, San Andrés V, Clare EL, Mir G, Symondson WOC (2014) A pragmatic approach to the analysis of diets of generalist predators: the use of next-generation sequencing with no blocking probes. *Molecular Ecology Resources*, **14**, 18-26.

Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P (2012) Who is eating what: diet assessment using next generation sequencing. *Molecular Ecology*, **21**, 1931-1950.

Price MN, Dehal PS, Arkin AP (2010) FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**(3), e9490.

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

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, **41**, D590-D596.

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

Skovgaard A, Karpov SA, Guillou L (2012) The parasitic dinoflagellates *Blastodinium* spp. inhabiting the gut of marine, planktonic copepods: morphology, ecology and unrecognized species diversity. *Frontiers in microbiology*, **3**, 305.

Shi XL, Lepère C, Scanlan DJ, Vaulot D (2011) Plastid 16S rRNA gene diversity among eukaryotic picophytoplankton sorted by flow cytometry from the South Pacific Ocean. *PLoS One,* **6**(4), e18979.

Stern RF, Amorim AL, Bresnan E (2014) Diversity and plastid types in *Dinophysis acuminata* complex (Dinophyceae) in Scottish waters. *Harmful Algae*, **39**, 223-231.

Stoeck T, Bass D, Nebel M, Christen R, Jones MD, et al. (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Molecular Ecology*, **19**, 21-31.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E (2012) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045-2050.

Unrein F, Gasol JM, Not F, Forn I, Massana R (2014) Mixotrophic haptophytes are key bacterial grazers in oligotrophic coastal waters. *The ISME Journal*, **8**, 164-176.

Vaulot D, Romari K, Not F (2002) Are autotrophs less diverse than heterotrophs in marine picoplankton? *Trends in Microbiology***, 10**, 266-267.

Vaulot D, Courties C, Partensky (1989) A simple method to preserve oceanic phytoplankton for flow cytometric analyses. *Cytometry*, **10** ,629-635.

Vaulot D, Eikrem W, Viprey M, Moreau H (2008) The diversity of small eukaryotic phytoplankton (≤3 μm) in marine ecosystems. *FEMS Microbiology Review*s **32**,795-820.

West NJ, Schönhuber WA, Fuller NJ, Amann RI, Rippka R, Post AF, Scanlan DJ (2001) Closely related *Prochlorococcus* genotypes show remarkably different depth distributions in two oceanic regions as revealed by in situ hybridization using 16S rRNA- targeted oligonucleotides. *Microbiology* **147**, 1731-1744.

Williams BAP, Keeling PJ (2003) Cryptic organelles in parasitic protists and fungi. *Advances in Parasitology*, **54**, 9-67.

Worden AZ, Nolan JK, Palenik B (2004) Assessing the dynamics and ecology of marine picophytoplankton: the importance of the eukaryotic component. *Limnology and Oceanography*, **49,** 168-179.

Zeidner G, Preston CM, DeLong EF, Massana R, Post AF, Scanlan DJ, Béjà O (2003) Molecular diversity among marine picophytoplankton as revealed by *psbA* analyses. *Environmental Microbiology* **5**, 212-216.

Zhang Z, Cavalier-Smith T, Green VR (2002) Evolution of dinoflagellate unigenic minicircles and the partially concerted divergence of their putative replicon origins. *Molecular Biology and Evolution*, **19**, 489-500.

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

Zingone A, Wyatt T (2005) Harmful algal blooms: keys to the understanding of the phytoplankton ecology. *In*: Robinson, A. R. & Brink, K. H. [Eds.] *The Sea.* Harvard University Press, Harvard, pp. 867-926.

Zubkov MV, Tarran GA (2008) High bacterivory by the smallest phytoplankton in the North Atlantic Ocean. *Nature*, **455**, 224-226.

Figures and Legends

Figure 1: Treemap and histograms showing the origin and number (A), and length (B) of the plastidial 16S rDNA sequences compiled into the PhytoREF database. A: PhytoREF is composed of 3,333 amplicons from identified organisms, 1,867 environmental amplicons produced from Sanger clone libraries, 879 sequences extracted from plastidial genomes, and 411 novel amplicons that have been generated in this study from cultures of marine microalgae. B: Most 16S rDNA sequences in PhytoREF are distributed in two peaks: the one with 700-900 bp-long sequences containing the novel amplicons obtained here from cultured microalgal strains, and the other one with full-length (ca. 1500 bp) sequences from public databases.

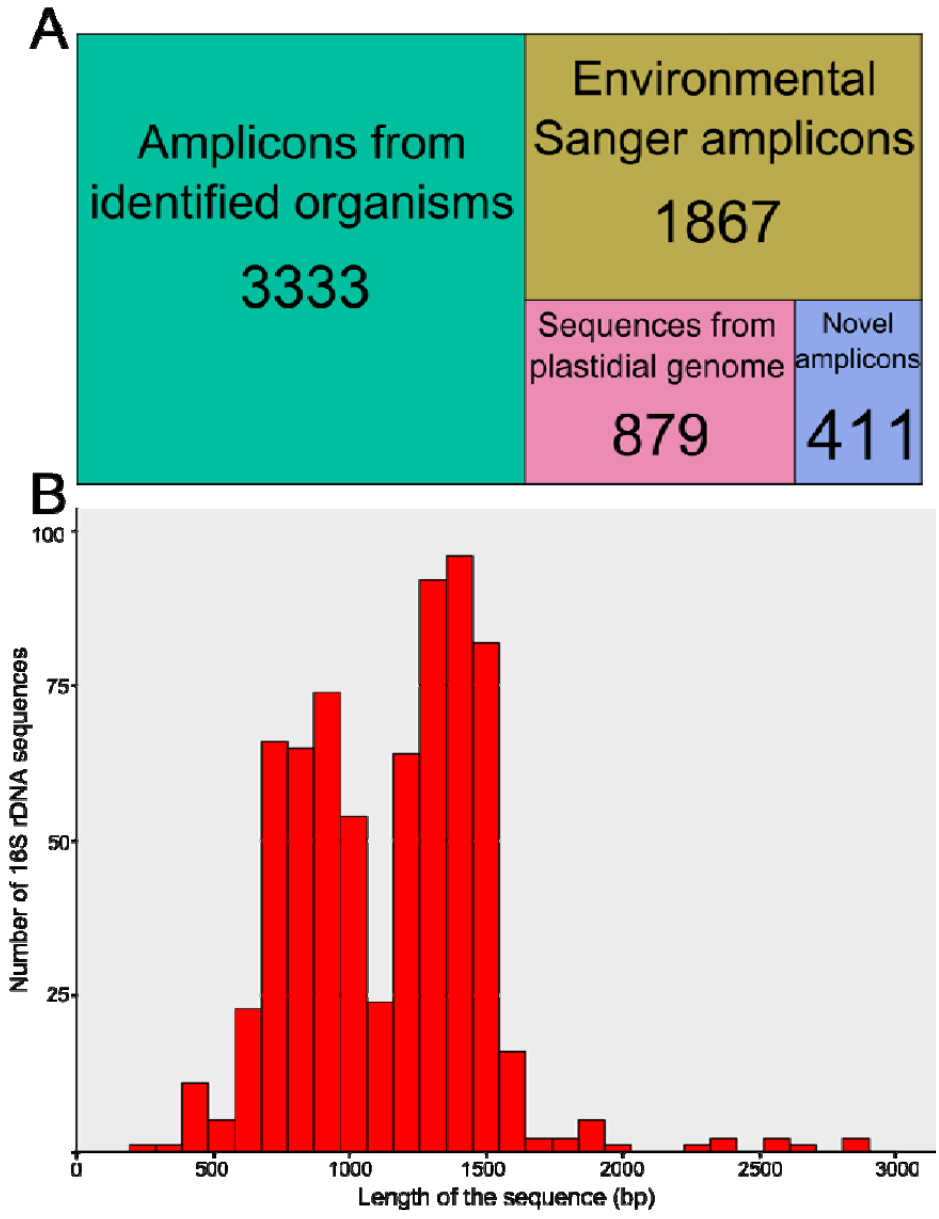

Figure 2: Distribution and number of PhytoREF plastidial 16S rDNA sequences in the tree of eukaryotic life. The schematic phylogenetic tree is based on up-to-date phylogenomics and morphological evidence (Burki & Keeling 2014). Each plastid-containing eukaryotic lineage is highlighted in green, and the number of plastidial 16S rDNA sequences available in the PhytoREF database is indicated in small grey circles.

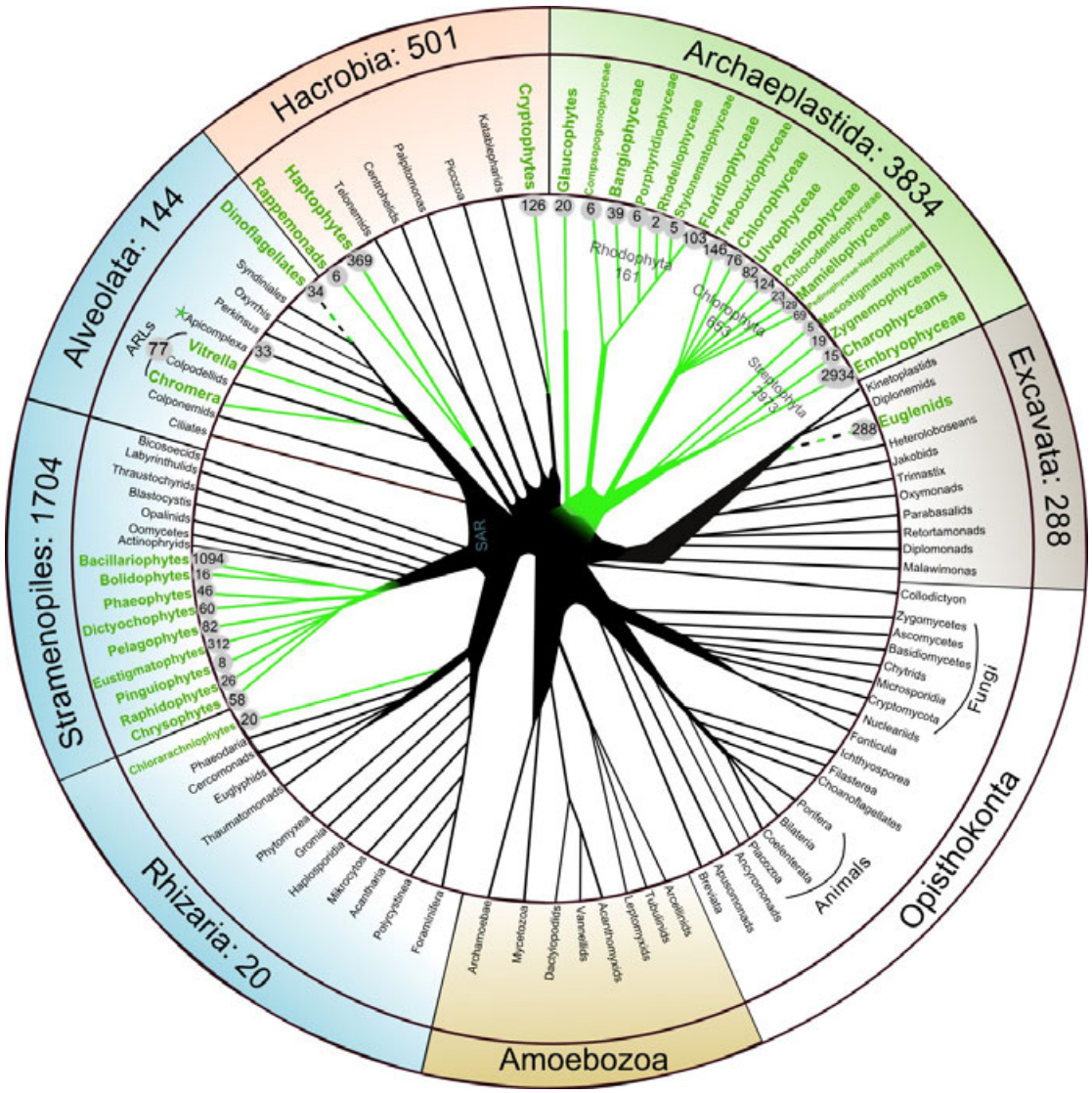

Figure 3: Taxonomic composition of the PhytoREF database at the class level. Bar charts represent the number of PhytoREF plastidial 16S rDNA sequences and taxonomically-described families, genera and species that are present in a given class. Several key groups of microalgae lack full taxonomic description, such as the prasinophytes (clade VII) and the rappemonads. Streptophytes (land plants) that are represented by 2,973 sequences (373 families and 796 genera) were not considered here for a better clarity.