A new and diverse plastid-bearing microbial eukaryote and its position on the eukaryotic tree of life.

Submitted by James William Harrison to the University of Exeter

as a thesis for the degree of

Masters by Research in Biosciences

In September 2010

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

Signature: ……..

Supervisor

Dr. T. A. Richards

Acknowledgements.

I would like to thank Dr. M. D. M. Jones for her training and assistance over my time at Exeter. Dr. E. Kim and Dr. J. M. Archibald for collaboration regarding clone library analysis and TSA-FISH. Dr. A. Z. Worden, Dr. S. Sudek and Dr. H. M. Wilcox for access to environmental samples and qPCR analysis. My heartfelt thanks go to Dr. G. Littlejohn and G. Leonard for their invaluable assistance with this thesis. I would like to thank all in our research group for their help and patience. Grateful thanks to Society for General Microbiology, Systematics Association and University of Exeter for funding my research. My special thanks go to Dr. T. A. Richards for putting up with me for over two years, always providing whatever help and guidance was necessary, raising my standards and opening my scientific horizons. Finally I would like to thank Rachel Mundy without whose belief and support I would have fallen at a much smaller hurdle long before this.

Index.

		Page
	Index of Figures.	5
	Index of Tables.	6
	Abstract.	7
1.	**Introduction.**	9
1.1	Identification of new microbial diversity using molecular methods.	10
1.2	The molecular expansion of eukaryotic microbial diversity.	12
1.3	Limitations of environmental clone library analysis.	15
1.4	rRNA encoding genes as taxonomic identification tags for eDNA analysis.	16
1.5	Ancestry of eukaryotic photosynthesis.	18
1.6	Red algal secondary endosymbiosis.	23
1.7	Green algal secondary endosymbiosis.	27
1.8	How many plastid endosymbiotic events?	27
1.9	Identification of novel putatively plastid bearing lineage.	30
1.10	Specific aims of this project.	31
2.	**Materials and Methods.**	32
2.1	Sampling.	32
2.1.1	Oceanic sampling.	32
2.1.2	UK coastal and freshwater sampling.	33
2.2	DNA extraction, clone library construction and sequencing.	34
2.2.1	Oceanic samples.	34
2.2.2	UK coastal and freshwater samples.	34
2.2.3	Oligonucleotide primer design.	35
2.2.4	PCR of separate environmental DNA samples.	35
2.2.5	Testing ecological overlap of target plastid 16S SSU rDNA sequences and biliphytes.	37
2.2.6	Comparison of the presence of target plastid 16S SSU rDNA sequences and biliphytes in different environmental size fractions.	38
2.2.7	Amplification of near full-length target group plastid rDNA gene cluster sequences.	38
2.2.8	Amplification of nucleus encoded rRNA gene cluster of biliphytes.	39
2.2.9	Identification of chimaeric sequences.	40
2.3	Molecular phylogenetics.	41
2.3.1	Alignment construction.	41
2.3.1.1	Environmental sequence alignment.	41

2.3.1.2 Clustering of environmental sequences. 41
2.3.1.3 Near complete target group plastid rDNA gene cluster sequence alignment. 42
2.3.1.4 Biliphyte nucleus encoded rRNA gene cluster alignment. 43
2.3.2 Methods of analysis. 43
2.3.3 Comparative topology test. 45

2.4 Quantitative PCR. 46

2.5 Tyramide signal amplification fluorescence in situ hybridization. 48

3. Results and Discussion. 53

3.1. PCR amplification of rappemonad sequences from environmental DNA. 53
3.2 Test of the ecological overlap of rappemonads and biliphytes. 54
3.3 Clone library surveys to assess the presence of rappemonads and biliphytes in different size-fractionated marine samples. 54
3.4 Phylogenetic analysis of clustered rappemonad environmental plastid 16S SSU rDNA sequences. 55
3.5 Phylogenetic analysis of rappemonad plastid rDNA gene cluster. 58
3.6 Phylogenetic analysis of biliphyte nuclear rDNA gene cluster. 59
3.7 Comparative topology test. 61
3.8 Quantitative PCR. 62
3.9 Tyramide signal amplification fluorescence in situ hybridization. 64

4. Conclusion. 67

5. References. 71

6. Appendix. 78

Index of Figures

		Page
Figure 1	Comparative arrangement of the rDNA gene cluster of eukaryotic nuclear and plastid genomes.	17
Figure 2	Representation of endosymbiotic theory and endosymbiotic gene transfer during primary and secondary endosymbiosis.	19
Figure 3	Schematic diagram showing a hypothesis for the origin and spread of photosynthesis in eukaryotes.	21
Figure 4	Tree showing diversity of environmental sequences clustered at 99% similarity and map showing global sampling locations (Inset).	57
Figure 5	Phylogenetic tree showing branching position of near full length rappemonad plastid rDNA gene cluster.	59
Figure 6	Phylogenetic tree showing branching position of concatenated biliphyte nuclear 18S SSU and 28S LSU rDNA sequences.	60
Figure 7	Results of alternative topology comparison tests.	62
Figure 8	Rappemonad distributions in the Sargasso Sea in 2003 as revealed by qPCR assays.	63
Figure 9	Fluorescence micrographs of rappemonads in the North Pacific.	65
Appendix Figure 1	Phylogenetic trees showing branching position of both rappemonad plastid 16S SSU and 23S LSU sequences separately.	85
Appendix Figure 2	Further TSA-FISH results showing rappemonad cells.	86

Index of Tables.

		Page
Table 1	Locations and characteristics of environmental samples used for eDNA clone libraries.	33
Table 2	PCR primers used to amplify rappemonad sequences from environmental samples.	36
Table 3	PCR primers used to target the biliphyte 18S rRNA gene.	38
Table 4	PCR primers used for amplifying near complete rappemonad plastid rDNA Genecluster.	39
Table 5	Oligonucleotide TSA-FISH probes.	49
Appendix Table 1	Table of sequences gained from clone libraries constructed from freshwater environmental DNA using biliphyte specific primers.	78
Appendix Table 2	PCR primers used for the amplification of the near full length biliphyte nucleus-encoded rRNA gene cluster.	79
Appendix Table 3	Sequences contained in collapsed environmental sequence identity clusters shown in figure 1.	80
Appendix Table 4	Table of Quantitative PCR sample sites and data.	83

Abstract

Using group-specific environmental clone libraries to target a section of the plastid 16S SSU rDNA gene we have identified a novel plastid-bearing eukaryotic lineage and have named this group “rappemonads” in anticipation of formal taxonomic description. Rappemonad sequences were amplified from aquatic environmental DNA samples collected from a wide variety of marine and freshwater sites suggests the rappemonads display a broad ecophysiology and wide geographical habitat distribution. Phylogenetic analysis was carried out on rappemonad rDNA sequences. Firstly, analyzing plastid SSU rDNA sequences suggesting this group form a diverse, strongly supported monophyletic clade encompassing numerous subclades. Secondly, phylogenetic analysis of a near full-length plastid rDNA gene cluster suggests the rappemonads represent an evolutionarily distinct lineage branching within the cryptomonad/haptophyte radiation, specifically as a sister group to the haptophytes. This suggests that the rappemonad plastid shares common ancestry with the red algal derived secondary plastids of the haptophytes and cryptomonads. Consequently, rappemonad plastid sequences appear to represent the plastid of either a novel highly divergent haptophyte or of an entirely new group. Phylogenetic comparisons suggest that the rappemonad lineage is not representative of the plastid of the biliphytes, a recently discovered microbial eukaryote. The biliphytes have been shown to display a weakly supported relationship with the cryptomonads in nuclear 18S SSU rDNA phylogenies although our analysis based on nuclear-encoded SSU and LSU rRNA genes does not support a biliphyte/cryptomonad relationship. Analyses using qPCR methodology demonstrate that the rappemonads can form transient blooms in the Sargasso Sea. Fluorescence *in situ* hybridization

revealed rappemonad cells were 6.6 ± 1.2 x 5.7 ± 1.0 µm, and appear to contain 2 to 4 plastids. We have shown the rappemonads to be a novel, widespread, microbial algae and potentially an important component in global photosynthetic ecosystems and therefore likely a player in oceanic geochemical cycles. This major algal lineage has so far remained absent from ecosystem models and is a significant new addition to the tree of life.

Key words: plastid evolution, phytoplankton, uncultured eukaryotes, rappemonads.

Publications from this thesis.

Kim, E.*, **J. W. Harrison***, et al. "Newly identified and diverse plastid-bearing branch on the eukaryotic tree of life." P. Natl. Acad. Sci. USA.

1. Introduction.

The study of the diversity of the tree of life is a dynamic discipline constantly changing our understanding of the way organisms are related to one another with respect to ecology and evolution. In marine environments microscopic phototrophs play a pivotal role in the development and maintenance of the Earths' ecosystems. Geological evidence suggests the ancient atmosphere was oxygenated by the ancestors of extant cyanobacteria 2 billion years ago (Kaufman, Johnston *et al.* 2007). Present day oceanic microscopic organisms perform approximately 50% (Field, Behrenfeld *et al.* 1998) of global primary production, maintaining marine food webs, and providing a vital component in many geochemical cycles (Falkowski, Scholes *et al.* 2000). It is becoming increasingly clear that the eukaryotic element of marine microbial ecosystems is important and that these microbes play an active role in primary production through-out the marine food web (Falkowski, Scholes *et al.* 2000; Lopez-Garcia, Rodriguez-Valera *et al.* 2001; Moon-van der Staay, De Wachter *et al.* 2001; Guillou, Viprey *et al.* 2008; Massana and Pedrós-Alió 2008; Liu, Probert *et al.* 2009).

Traditional methods for investigating microbial diversity have been based on microscopy, isolation, and culture, combined with morphology and biochemical analysis and have generally proved inadequate for understanding the diversity and evolutionary complexity of microbes (Amann, Ludwig *et al.* 1995; Pace 1997; Hugenholtz, Goebel *et al.* 1998). The lack of discernible morphological variation in natural microbial communities along with the suggestion that a great proportion (up to 99%) (Pace 1997; Hugenholtz, Goebel *et al.* 1998) of the microbial taxa have as yet proved to be unculturable has

biased our understanding of biodiversity towards a subset of taxa amenable to culture.

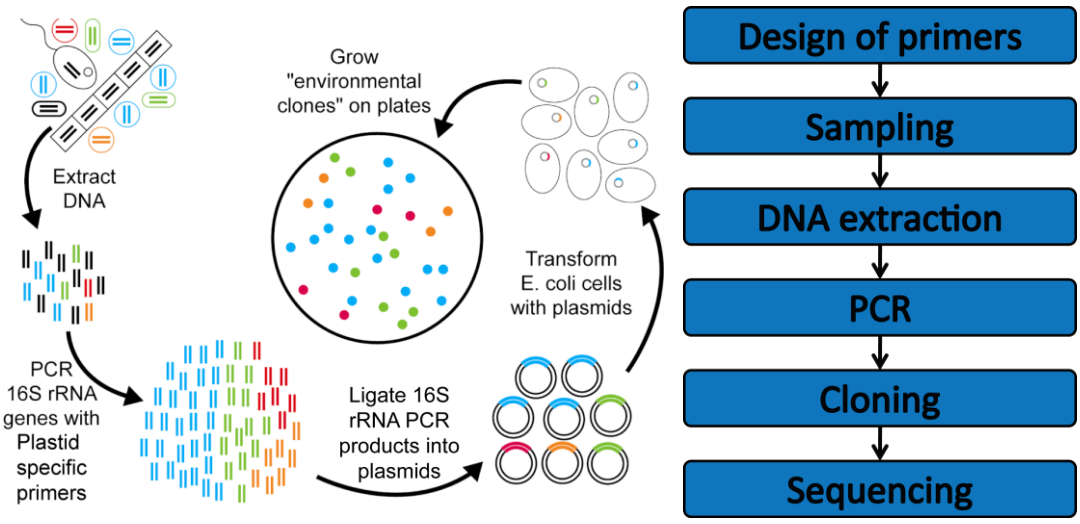
1.1 Identification of new microbial diversity using molecular methods.

Our understanding of microbial diversity was revolutionised by the use of environmental clone libraries to investigate microbial diversity (Giovannoni, Britschgi *et al.* 1990). This methodology involves using oligonucleotide primers targeted to a group of microbes for polymerase chain reaction (PCR) on environmental DNA samples (eDNA). The resulting amplicon populations are then cloned and individual colonies are picked and sequenced. The sequence data are then compared to a database of known sequences using either similarity search algorithms or phylogenetic analyses (Olsen, Lane *et al.* 1986; Pace 1997). These approaches have allowed the inclusion of, as yet, uncultured organisms in diversity studies (Box 1) (Olsen, Lane *et al.* 1986; Giovannoni, Britschgi *et al.* 1990). Using these methods it was demonstrated that the diversity of the prokaryotes was largely underestimated (Giovannoni, Britschgi *et al.* 1990). The best estimation prior to the use of these novel methods divided the bacteria into 12 higher level taxonomic groups, now however it is thought that the number of divisions could be over 40, with a large number of these higher level taxonomic groups only represented by uncultured organisms (Hugenholtz, Goebel *et al.* 1998). Some of these new discoveries have proved to be very important, demonstrating large gaps in our knowledge of the microbial biosphere. For example the discovery of the SAR11 phylogenetic cluster (Giovannoni, Britschgi *et al.* 1990), which have been shown to dominate marine bacterioplankton communities (Morris, Rappe *et al.*

2002) and play a significant role in marine primary production and carbon cycling (Malmstrom, Cottrell *et al.* 2005).

Our understanding of archaeal diversity was also re-evaluated using similar approaches. Studies showing increased diversity of known groups such as the crenarchaeota and suggesting new archaeal phylogenetic groups branching close to thermophilic taxa are present in oligotrophic ocean samples (Fuhrman 1992). Such approaches have identified further novel archaeal groups such as the putative novel group named the korarchaeota (Barns, Delwiche *et al.* 1996).

Box 1. Environmental clone library construction.



I. The process of environmental clone library construction initially requires the design of PCR primers targeted to the sequence of interest, a particular gene or a sub-group of such a gene, for example the plastid 16S genes.
II. An environmental sample - for example a marine water sample - is collected which contains a mixed community of microbial organisms. The DNA is extracted resulting in a mixed sample of DNA representative of the composition of the original environmental sample.
III. PCR is performed using the targeted primers (I) which results in a mixed sample of amplicons representative of the composition of the original sample.
IV. These amplicons are then ligated into plasmids, transformed into competent cells which are then grown in culture.
V. These cultures are then spread on selection plates each colony consisting of cells containing plasmids with an insert representing an amplicon from the PCR reaction and thus an organism from the original sample.

1.2 The molecular expansion of eukaryotic microbial diversity.

Phylogenetic analysis of sequences obtained from environmental clone libraries has suggested that eukaryotic diversity had been underestimated in a similar way as that of the prokaryotes. Studies targeting, for example, the bacterial 16S rRNA gene had picked up an unexpected level of diversity in the eukaryotic plastid encoded 16S rRNA gene (Rappe, Suzuki *et al.* 1998). It was not until these molecular and phylogenetic techniques were used to directly investigate eukaryotic diversity (Staay, Staay *et al.* 2000; Lopez-Garcia, Rodriguez-Valera *et al.* 2001; Moon-van der Staay, De Wachter *et al.* 2001) that an improved understanding of eukaryotic diversity became apparent. Environmental clone libraries constructed using primers targeting the eukaryotic ribosomal SSU gene (18S) were used to assess eukaryotic biodiversity and results confirmed that the diversity of the eukaryotes had been largely underestimated (Lopez-Garcia, Rodriguez-Valera *et al.* 2001; Moon-van der Staay, De Wachter *et al.* 2001; Richards, Vepritskiy *et al.* 2005). These studies demonstrated high levels of novel diversity within known taxonomic groups such as diatoms, prymnesiophytes, prasinophytes, stramenopiles, and alveolates (Diez, Pedros-Alio *et al.* 2001; Richards, Vepritskiy *et al.* 2005; Groisillier, Massana *et al.* 2006; Guillou, Viprey *et al.* 2008). There were also a number of sequences recovered which suggested novel groups that could not be assigned to already identified phyla and were therefore characterized purely on the basis of sequences recovered from environmental gene libraries (Berney, Fahrni *et al.* 2004; Not, Valentin *et al.* 2007).

The expansion of known diversity within the alveolate group provides an example of how the use of molecular phylogenetic methods has uncovered novel diversity within already recognised groups, specifically two diverse deep-

branching alveolate groups (Lopez-Garcia, Rodriguez-Valera *et al.* 2001; Moon-van der Staay, De Wachter *et al.* 2001; Moreira and López-García 2002; Richards and Bass 2005; Massana and Pedrós-Alió 2008). These groups were initially named ‘marine alveolate group I’ and ‘marine alveolate group II’ and were identified in environmental gene libraries taken from the euphotic zone of Antarctic marine waters (Lopez-Garcia, Rodriguez-Valera *et al.* 2001). The novel groups appear to be geographically wide spread and have been shown to inhabit extreme environments (Groisillier, Massana *et al.* 2006; Guillou, Viprey *et al.* 2008), suggesting they are genuine sequences derived from uncultured organisms rather than artefacts from the sequencing or phylogenetic processes. Fluorescent *in situ* hybridisation analyses were used to demonstrate that this group represents a large diversity of intracellular parasites that infect a wide range of organisms from marine invertebrates to eukaryotic algae (Chambouvet, Morin *et al.* 2008). Recent work has placed these novel groups as part of the parasitic sister group to the dinoflagellates known as the syndiniales (Guillou, Viprey *et al.* 2008).

Similarly, eukaryote specific eDNA studies uncovered unexpected levels of diversity within the stramenopile group (Moon-van der Staay, De Wachter *et al.* 2001; Richards and Bass 2005), increasing the known diversity of already characterized phototrophic groups such as the pelagophytes and ochromonads (Lovejoy, Massana *et al.* 2006) and identifying numerous novel branches within the stramenopile phylogeny. These novel groupings have been named MAST (MArine STramenopile) groups with the most frequently recovered phylotypes being in MAST clusters 1, 3 and 7 (Moon-van der Staay, De Wachter *et al.* 2001; Massana, Guillou *et al.* 2002; Lovejoy, Massana *et al.* 2006). FISH analysis of these novel organisms demonstrated that many of

these novel clades encompassed bacterio-trophic protists (Massana, Guillou *et al.* 2002).

These studies demonstrate that better appreciation of the ecological importance and diversity of groups such as the alveolates and stramenopiles result in an improved understanding of community diversity and marine food webs.

Primary production in the open ocean photic zone was previously thought to be dominated by prokaryotic cyanobacteria such as *Prochlorococcus* (Goericke and Welschmeyer 1993). However, current analyses suggest it is in fact dominated by microscopic eukaryotes of the haptophyte group. This was suggested by the high levels of the photosynthetic pigment 19′-hexanoyloxyfucoxanthin, only currently known from the plastids of haptophytes, observed in the world's oceans using analysis of satellite imagery (Liu, Probert *et al.* 2009). The relatively low abundance of these organisms so far discovered in marine environmental clone libraries suggests a major discrepancy between environmental abundance and representation in clone libraries. This discrepancy is thought to be a product of PCR amplification bias against haptophyte representation in standard environmental clone libraries, because haptophytes have long and GC-rich nuclear-encoded rRNA genes. When protocols were adapted to account for this methodological bias, a significant level of novel haptophyte diversity was uncovered most of which form deep diverging clusters among haptophyte groups such as the Prymnesiales (Liu, Probert *et al.* 2009). This research suggests that haptophytes are the most diverse and prevalent photoautotrophic group in global marine environments, contributing between 30% – 50% of marine photosynthetic activity. Firstly, this makes the study of this and other related groups of great importance and

secondly suggests that there is a huge level of undiscovered biodiversity within the ocean photic ecosystem which requires in depth study.

1.3 Limitations of environmental clone library analysis.

These eDNA methodologies are not without problems and care needs to be taken to account for these when interpreting eDNA clone library data. For example, chimeric sequence can be generated during the PCR sampling procedure and which gives a false picture of lineage diversity and identifying artifactual sequences as highly divergent phylogenetic groups (Dawson and Pace 2002; Berney, Fahrni *et al.* 2004; Cavalier-Smith 2004; Richards and Bass 2005) This can be detected using chimera detection software (e.g. Bellerophon (Huber, Faulkner *et al.* 2004)) or by manually analysing the alignment patterns of environmental sequences (e.g. (Berney, Fahrni *et al.* 2004).

An additional source of error when interpreting eukaryotic clone library analyses is the variability between rRNA cistron copy number within a single nucleus - where multiple copies of the target gene can be present with variation - therefore giving an inflated idea of actual taxonomic diversity. The level of intra-individual variation is variable but can be high enough to cause systematic problems when inferring close evolutionary relationships (Richards and Bass 2005).

A final example is the individual and systematic artefacts produced by inappropriate phylogenetic methods (Richards and Bass 2005) such as:

(1) Incomplete or inappropriate taxon sampling, which can impose a false topology on a phylogenetic tree.

(2) The mutational saturation of certain positions within the DNA sequence of a gene which negate their usefulness for phylogenetic analysis, especially when dealing with ancient evolutionary relationships.

(3) Long-branch attraction when two taxa are attracted together in a phylogenetic tree due to the high rate of evolution rather than shared ancestry. This can be compensated for by increased taxon sampling to reduce branch lengths or by using appropriate phylogenetic methods which account for high mutation rates (Bergsten 2005).

These methodological issues need to be kept in mind when interpreting molecular phylogenetic data of environmental sequences.

1.4 rRNA-encoding genes as taxonomic identification tags for eDNA analysis.

The majority of these molecular diversity studies have used the small subunit rRNA gene to investigate microbial diversity. The architecture of the rRNA gene cluster is generally conserved throughout the tree of life, having a small subunit-encoding gene and a large subunit-encoding gene arranged in tandem. Although there are notable differences between the prokaryotes and the eukaryotes - for prokaryotes the LSU is normally known as 23S and the SSU 16S while for the eukaryotes the nuclear-encoded LSU is known as 28S and the SSU 18S (characterized based on the differential sedimentation rates with the S denoting Svedburg, a non SI unit of sedimentation).

More complexity is introduced into eukaryote cells however with the presence of bacterial-derived organelles which retain vestiges of their bacterial genomes (Fig. 1). Therefore, a photosynthetic eukaryote cell has three functional rRNA operons with three separate evolutionary ancestries, the 18S/28S nuclear-encoded rRNA genes, the 16S/23S plastid-encoded rRNA genes and the 12S/16S mitochondrial-encoded rRNA genes all being transcribed in different discrete cellular compartments (Ishikawa 1977).

All of these copies of the rRNA genes can be used for phylogenetic analysis although they provide evolutionary information relevant to their source, for example analysis of the plastid rRNA genes will suggest an evolutionary history for the plastid organelle, whereas the nuclear encoded genes can be used to investigate the ancestry of the host, which may be markedly different to the phylogenetic ancestry of the endosymbiotic organelles.

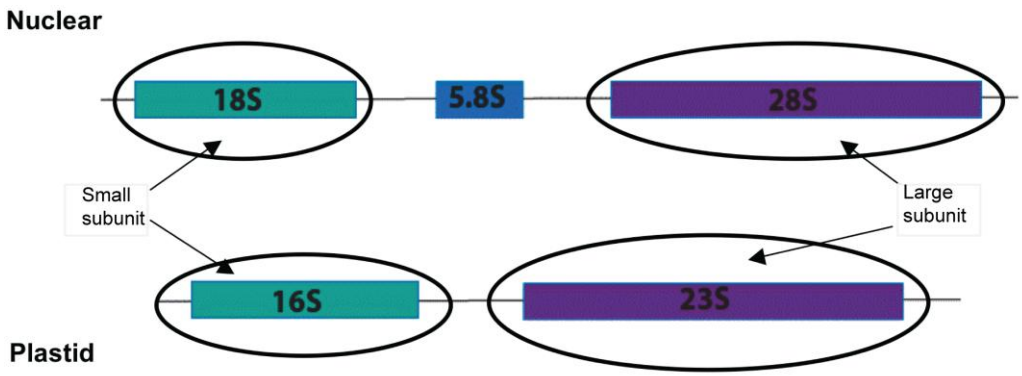


Figure 1. Showing the comparative arrangement of the rRNA gene cluster within the nuclear and plastid genomes of eukaryotic cells. Light blue represents small subunit genes and purple represents large subunit genes.

1.5 Ancestry of eukaryotic photosynthesis.

Photosynthesis first evolved in the prokaryotic ancestor of cyanobacteria between 3.5 and over 2 billion years ago (Nisbet and Sleep 2001; Martin and Russell 2003; Kaufman, Johnston *et al.* 2007). Photosynthesis was then transferred from the prokaryotes to the eukaryotes sometime after the divergence of the Plantae, also known as the Archaeplastida, before this group divided into the red algae, the green algae and the glaucocystophytes (Cavalier-Smith 2000; Yoon, Hackett *et al.* 2004; Adl, Simpson *et al.* 2005; Bhattacharya, Archibald *et al.* 2007). This transfer was achieved via an endosymbiotic event, known as the primary endosymbiosis (Fig. 2), in which a fully evolved heterotrophic eukaryote possessing a nucleus, cytoskeleton, and mitochondria engulfed a photosynthetic cyanobacterium possessing chlorophylls a and b, phycobilisomes, and unstacked thylakoids (Cavalier-Smith 2000; Moreira, Le Guyader *et al.* 2000; Howe, Barbrook *et al.* 2008) and retained it in an endosymbiotic relationship. This cyanobacterial symbiont evolved into the first plastid (Cavalier-Smith 2000; Moreira, Le Guyader *et al.* 2000; McFadden 2001; Howe, Barbrook *et al.* 2008) after which the three major Plantae groups diverged (Fig. 3). The endosymbiotic theory of the origin of the plastid organelle was first suggested around the turn of the twentieth century by theorists such as Mereschkowsky and Altman (Taylor 1974; Martin and Russell 2003). However, it was not until the late 1960s with identification of organellar DNA and the work of Lynn Margulis and others, that the theory began to slowly gain acceptance (Taylor 1974). Prior to the acceptance of the endosymbiotic theory, all organelles had been thought to be of “wholly autogenous” origin; i.e. that the organelles had developed over time through increasing functional compartmentalisation (Taylor 1974). This theory is still the primary explanation

for the development of many subcellular structures such as the Golgi apparatus, endoplasmic reticulum, and peroxisomes (Dacks, Poon *et al.* 2008) .

During the evolutionary path from endosymbiotic bacterium to organelle the plastid has undergone major physiological and genomic reduction. Typically cyanobacterial genomes are between 1.5 and 9 Mbp in length (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) and contain a relatively high number of genes compared to other prokaryotes. For example, the cyanobacterium *Anabeana* sp. Pcc7120 has 5366 genes. In contrast, plastid genomes are rarely more than 200kb (Stoebe, Martin *et al.* 1998; Bhattacharya, Archibald *et al.* 2007) and contain a far smaller complement of genes. For example, the most gene rich plastid genome identified is of the red alga *Porphyra purpurea* which has 251 genes with most plastid genomes having far

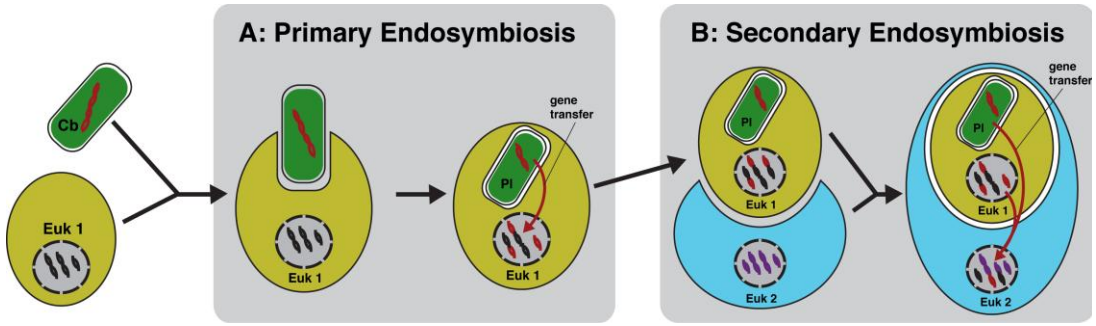


Figure 2. . Representation of endosymbiosis and environmental gene transfer during primary and secondary endosymbiosis. A: shows the engulfment of a cyanobacterial cell (green) by a heterotrophic eukaryote (yellow) becoming a primary endosymbiotic plastid. The red arrow shows the direction of environmental gene transfer. B: Shows the engulfment of a photosynthetic eukaryote (Yellow) bearing a primary plastid (green) by a heterotrophic eukaryote (Blue) becoming a secondary endosymbiotic plastid. Red arrows again showing the direction of environmental gene transfer.

fewer (Gould, Waller *et al.* 2008). This change is explained by a large amount of gene loss and endosymbiotic gene transfer (EGT) from the plastid genome to the host nuclear genome (Fig. 2). The proteins coded by these genes are synthesized in the cytosol of the host cell are targeted back for use in the plastid (Cavalier-Smith 1999). EGT has been shown to be a major influence on

the evolution of the nuclear genome. For example, molecular phylogenies of over 9000 *Arabidopsis thaliana* nuclear coded genes suggested that up 18% of its nuclear genome is of plastid origin (Martin, Rujan *et al.* 2002).

The movement of genes from the plastid genome to the nuclear genome resulted in the cytosolic synthesis of many proteins vital for the function of the plastid – which require relocating back to the plastid. Nuclear-encoded plastid-targeted proteins are synthesized on cytosolic ribosomes and targeted back to the plastid post-translationally using N-terminal targeting transit peptides which are proteolytically removed after import (Soll and Schleiff 2004). Examples of nuclear-encoded proteins which are translocated into the plastid include constituent proteins of the two photosystems, ATP synthase and the CO_2-fixing enzyme ribulose-1,5-biphosphate carboxylase oxygenase (Rubisco). The nuclear localisation of these integral chloroplast proteins suggest there must be stringent cellular control mechanisms to ensure efficient function of this process.

Plastid transit peptides are typically 30-to-100 amino acids in length and provide a recognition signal for the transport of proteins into the organelle (Bhattacharya, Archibald *et al.* 2007). Proteins are translocated across the outer and inner plastid membranes by the TOC (translocator of the outer chloroplast membrane) and TIC (translocator of the inner chloroplast membrane) proteins (Jarvis and Robinson 2004; Soll and Schleiff 2004). This TIC/TOC complex comprises proteins of evolutionarily diverse origins. Channel-forming TOC75 and the proteins TIC62, TIC55, TIC20, TIC22 and TIC21 are likely of cyanobacterial (i.e. endosymbiotic) origin, whereas the remaining

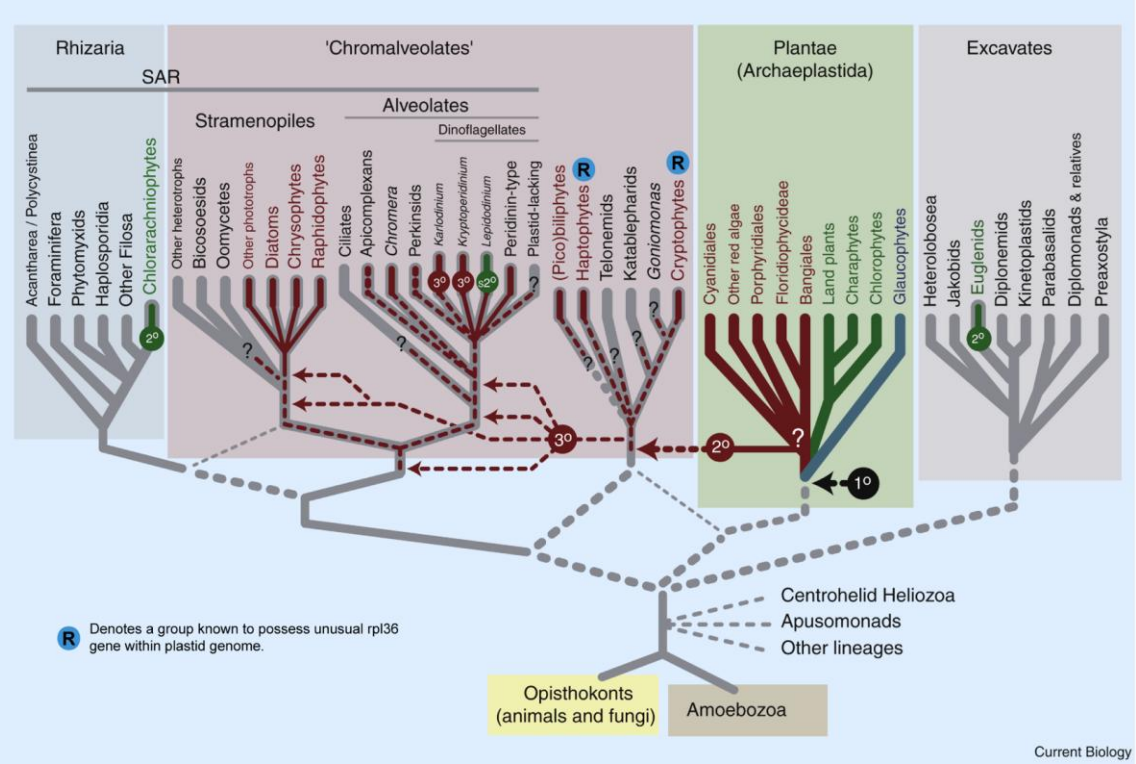


Figure 3. (Reproduced from Archibald J. M. (2009)) Schematic diagram showing a hypothesis for the origin and spread of photosynthesis in eukaryotes. Tree is a schematic branch lengths are uninformative, showing the six putative eukaryotic supergroups. Grey dashed lines show uncertain phylogenetic relationships. Red lines show red algal endosymbionts, green lines show green algal endosymbionts and blue line shows glaucocystophyte symbiont. 1°/2°/3° shows primary/secondary/tertiary endosymbiotic event. Dashed red lines show possible movement of red algal plastid organelles between lineages. Question mark indicates possible secondary loss of plastid. R denotes lineage with unusual HGT of rpl36 gene into plastid genome.

components (TOC159, TOC34, TOC64, TIC40, TIC110, TIC32) probably evolved from co-option of existing host genes or, less likely but still possible, by horizontal gene transfer from other bacterial sources (Bhattacharya, Archibald *et al.* 2007).

Data gained from molecular phylogenies based on both single and multiple gene phylogenies provide is typically consistent with a single plastid origin and these studies provide convincing support for the monophyly of the plastid organelle. It has been shown that plastid genes from divergent Plantae lineages group into one monophyletic clade branching with the cyanobacteria. This suggests, with strong support, that there has only been a single primary endosymbiotic event giving rise to all the extant examples of the plastid

organelle (Delwiche, Kuhsel *et al.* 1995; Turner, Pryer *et al.* 1999). Furthermore, phylogenetic analysis of nuclear-encoded plastid-functioning proteins has confirmed monophyly of the Plantae plastid (Weber, Linka *et al.* 2006). Therefore, the monophyly of the primary plastids of red algae, green algae and land plants, and glaucocystophytes, is widely accepted. However there is some controversial evidence to the contrary, suggesting a more complex ancestry. For example, from phylogenetic analysis of the nucleus-encoded eukaryotic translation elongation factor 2 (EEF 2), suggesting a polyphyletic ancestry (Kim and Graham 2008) and other physiological and biochemical evidence concerning genome organisation, protein import machinery and light harvesting complexes (Larkum, Lockhart et al. 2007).

In contrast to evidence that all eukaryotic photosynthetic functions descend from one endosymbiotic event there is evidence emerging for a very recent and separate photosynthetic acquisition by eukaryotes involving the protist *Paulinella chromatophora.* First characterised in 1895 by Robert Lauterborn*, P. chomatophora* is a thecate, filose amoeba with one or two blue green 'sausage-shaped' inclusions, which were initially thought to be ingested food items demonstrating cyanobacterial characteristics. Closer observation of these structures ruled this out, however, as they have been observed as a stable characteristic in pure cultures of *P. chomatophora* (Marin, M. Nowack *et al.* 2005). *P. chomatophora* has been shown to be non-phagocytic protist and that the division of the protist and the pigmented organellar body is tightly co–regulated (Marin, M. Nowack *et al.* 2005). The putative organelles of *P. chomatophora* have been shown not to group with other plastids in rDNA phylogenies but to group separately with other cyanobacteria in a clade with S*ynechococcus* and *Prochlorococcus* which are species commonly

phagocytosed by *Paulinella ovalis* (Rodríguez-Ezpeleta and Philippe 2006), a close relative of *P. chomatophora.* Taking all this evidence into consideration, it has been surmised that this represents a separate primary endosymbiotic event (Bhattacharya, Helmchen *et al.* 1995). This discovery, whilst being significant in itself also provides the opportunity to study the endosymbiotic process at a much earlier stage as *P. chomatophora* can be compared to its close relatives which do not harbour plastids and the *P. chomatophora* endosymbiont can be compared with its close relatives such as S*ynechococcus* and *Prochlorococcus* (Bhattacharya, Helmchen *et al.* 1995; Reyes-Prieto, Yoon et al. 2010).

1.6 Red algal secondary endosymbiosis.

Photosynthesis was spread throughout diverse groups of eukaryotes via a series of secondary endosymbiotic events in which a heterotrophic eukaryote engulfed a photosynthetic eukaryote with a plastid of primary endosymbiotic origin, retaining the photosynthetic machinery of the symbiont as a secondary plastid (Fig 2) (Gibbs 1978; Cavalier-Smith 2000; Gould, Waller *et al.* 2008). After a process of physiological and genomic integration, the photosynthetic machinery of the eukaryotic symbiont is retained as a secondary plastid by the host cell. The idea of serial endosymbiotic events giving rise to the more complex photosynthetic organelles observed in many eukaryotic lineages was first suggested in the early 1970's (Taylor 1974). This suggestion was followed by the first examples of secondary endosymbiosis identified by Sarah Gibbs with the identification of a green algal derived endosymbiont of *Euglena* (Gibbs 1978; Gibbs 1981).

Secondary endosymbiosis is thought to have occurred at least twice and possibly as many as seven times (Cavalier-Smith 1999; Cavalier-Smith 2000;

Archibald and Lane 2009) resulting in secondary plastids descended from both green and red algal endosymbionts (Fig. 3).

There are at least six known lineages which display evidence of a red algal secondary symbiont:

Cryptomonads include both non-photosynthetic and photosynthetic algal forms. The phototrophs possess a red algal derived plastid surrounded by four membranes (Keeling 2004). In addition, the cryptomonads include many species, which have retained a reduced relic nucleus acquired from the red algal symbiont known as a nucleomorph. This structure is localised between the outer two and the inner two plastid membranes in the periplastid space (Archibald 2007; Archibald and Lane 2009).

Haptophytes possess a red algal derived plastid bound by four membranes (Keeling 2004) and synthesize a unique pigment, 19′-hexanoyloxyfucoxanthin (Liu, Probert *et al.* 2009). The cryptomonad and haptophyte plastid also posses a shared derived rpl36 gene acquired by horizontal transfer from a separate bacterial ancestor from other photosynthetic eukaryotes and cyanobacteria. This association suggests that the haptophyte and cryptomonad plastid are monophyletic and share a common ancestor which was in receipt of this HGT event (Rice and Palmer 2006; Keeling 2009).

Stramenopiles are a very diverse group containing both photosynthetic members and a wide diversity of non-photosynthetic, heterotrophic protists (Massana, Castresana *et al.* 2004). The phototrophs possess a plastid organelle of red algal origin, again surrounded by four membranes (Andersen 2004; Bolte, Bullmann *et al.* 2009).

Apicomplexa, include medically important groups of intracellular parasites such as *Plasmodium falciparum* (the causal agent of malaria) and include many forms that possess a plastid-derived organelle named the apicoplast which has lost its photosynthetic function (McFadden and Waller 1997). The plastid organelle is retained and has become the site of certain essential metabolic processes e.g. fatty acid and isoprenoid biosynthesis (Ralph, van Dooren *et al.* 2004; Archibald 2009). These organelles also retain a highly reduced plastid genome making their phylogenetic positioning problematic. There have been conflicting studies, which have suggested an alternative origin for this organelle, showing it grouping phylogenetically with the green algae (Kohler, Delwiche *et al.* 1997; Cai, Fuller *et al.* 2003) rather than the widely accepted red algal origin (Funes, Davidson *et al.* 2002; Yoon, Hackett *et al.* 2002).

Chromera velia is thought to be a photosynthetic relative of the non-photosynthetic apicoplast bearing apicomplexans; phylogenetic analysis has shown it to group with the apicomplexans. There is also genomic and biochemical evidence that the *Chomera* plastid shares a common ancestor with the apicoplast of the apicomplexans (Moore, Obornik *et al.* 2008; Janouškovec, Horák et al. 2010).

Dinoflagellates. Approximately half of the dinoflagellate taxa are photosynthetic and encompass a highly diverse array of plastid structures, ancestries, and genomic arrangements. Many dinoflagellate plastids contain the unique photosynthetic pigment peredinin, while the plastid organelle is surrounded by three membranes, which originate from a secondary endosymbiotic event involving a red algal symbiont (Hackett, Anderson *et al.* 2004). However, these plastids are unique in the fact that rather than the

canonical circular plastid genome arrangement, the genome of these dinoflagellate plastids is broken up into a series of mini-circles (Zhang, Green *et al.* 1999; Barbrook and Howe 2000; Howe, Nisbet *et al.* 2008).

Another type of dinoflagellate plastid is evident in *Karenia brevis* which has 19′-hexanoyloxyfucoxanthin pigment instead of peredinin, phylogenetic analysis of the plastid-encoded SSU rRNA-encoding gene demonstrated that this organelle was inherited from the haptophytes (Tengs, Dahlberg *et al.* 2000) suggesting a tertiary endosymbiotic origin involving the acquisition of a secondary plastid (Nosenko, Lidie *et al.* 2006).

There are at least three other plastid types among the dinoflagellates. (1) The plastid of the genera *Dinophysis* which has been shown to originate from a symbiosis with a cryptomonad, but evidence points to this being a temporary plastid gained from prey cells. The ingested plastids become what are termed as kleptoplastids and therefore demonstrate the very early stages of what may become an endosymbiotic association (Hackett, Anderson *et al.* 2004).

(2) Two further taxa, *Kryptoperidinium foliaceum* and *Durinskia balticum,* may have retained vestiges of their original peredinin containing plastid, which has lost all photosynthetic function (Withers, Cox *et al.* 1977). In this case the photosynthetic function has been taken over by a new plastid which originates from the tertiary endosymbiosis of a diatom (Hackett, Anderson *et al.* 2004). This is a permanent relationship unlike that of *Dinophysis*, but again is an example of a different stage in the endosymbiotic process as the symbiotic diatom retains nucleus, mitochondria, ribosomes and plastids within its membrane. It is thought that this unusual photosynthetic organelle represents a stage of endosymbiosis between the initial engulfment and the reduction of the endosymbiont.

(3) The plastid of the dinoflagellate *Lepidodinium viride* is the only dinoflagellate plastid so far identified not to have originated from a red algal endosymbiosis as this organelle originates from the chlorophyte lineage Prasinophyceae (Hackett, Anderson *et al.* 2004).

1.7 Green algal secondary endosymbiosis.

There are two additional cases of green algal secondary endosymbiosis evident in the euglenids and chlorarachniophytes (Gibbs 1978; McFadden, Gilson *et al.* 1994). Both have been shown to harbour green algal-derived plastids and have three and four membranes surrounding their plastids respectively (Gibbs 1978; McFadden, Gilson *et al.* 1994; Cavalier-Smith 1999; Cavalier-Smith 2000; Keeling 2004), with the chlorarachniophytes possessing a reduced relic eukaryotic nucleus (a nucleomorph) in its periplastid space similar to that of the cryptomonads although of completely different origin (Gilson, Su *et al.* 2006; Archibald 2007).

1.8 How many plastid endosymbiotic events?

The question of how many secondary endosymbiotic events have occurred is hotly debated. There has to have been at least two cases, as there are secondary plastids with red and green algal ancestors (Gould, Waller *et al.* 2008; Archibald 2009). It was previously suggested under the cabozoa hypothesis that the euglenid and chlorarachniophyte plastid shared the same origin and that there had been extensive secondary loss throughout intermittent branching taxa (Cavalier-Smith 2000). However, there is now well documented evidence that the euglenid and chlorarachniophyte plastids have descended

from separate endosymbiotic events with seperate green algae progenitors involved (Rogers, Gilson *et al.* 2007).

The number of red algal secondary endosymbiotic events is an extremely contentious issue. The chromalveolate hypothesis suggested by Cavalier-Smith (Cavalier-Smith 1999) suggests that all the red algal secondary endosymbionts derive from a single secondary endosymbiotic event. As such, the chromalveolate hypothesis groups the chromists (e.g. haptophytes, cryptomonads, stramenopiles and other closely related lineages) with the alveolates (e.g. apicomplexans, dinoflagellates, and ciliates) (Cavalier-Smith 2000). The chromalveolate hypothesis assumes that all non-photosynthetic members of the chromalveolate groups have secondarily lost phototrophic capacity. This suggestion has been supported to some extent with the recent discoveries of plastid-derived elements in the genomes of lineages which were previously thought to lack a plastid e.g. plastid targeted proteins found within the nuclear genome of non photosynthetic dinoflagellates *Oxyrrhis marinia* (Slamovits and Keeling 2008) and *Crypthecodinium cohnii* (Sanchez-Puerta, Lippmeier *et al.* 2007) and the detection of a plastid structure in *Perkinsus atlanticus* (Leonor Teles-Grilo, Tato-Costa *et al.* 2007) which is an alveolate.

Members of chromalveolate groups not possessing any plastid at all e.g. *Cryptosporidium parvum and C. muris.* would seem to counter the theory. However, even though no organelle or plastid-targeted proteins have been identified, genes demonstrating plastid ancestry have been found in the nuclear genome (Huang, Mullapudi *et al.* 2004). Similar cases have been identified in the oomycetes *Phytophthora* sp. (Tyler, Tripathy *et al.* 2006) and some ciliates (Reyes-Prieto, Moustafa *et al.* 2008). Although these observations do suggest a photosynthetic ancestor, which is consistent with the chromalveolate

hypothesis, alternative hypotheses are possible, involving horizontal gene transfer, which has been demonstrated for both of the mentioned groups (Ricard, McEwan *et al.* 2006; Richards, Dacks *et al.* 2006; Stiller, Huang et al. 2009,).

The acceptance of the chromalveolate hypothesis is by no means definite, molecular phylogenies seem to argue against the monophyletic nature of the chromalveolata, with evidence from multi-gene phylogenies suggesting that the chromalveolates are not monophyletic (Hackett, Yoon *et al.* 2007; Patron, Inagaki *et al.* 2007). Some studies group the cryptomonads and the haptophytes separately from the stramenopiles and alveolates, and surprisingly there is evidence that these latter two groups form a clade with the Rhizaria which include the chlorarachniophytes (which harbour a plastid of green algal origin) and together form the SAR clade (Burki, Shalchian-Tabrizi *et al.* 2008; Archibald 2009).

The monophyly of the cryptomonad/haptophyte clade has been supported by molecular phylogenetics e.g. (Patron, Inagaki *et al.* 2007) and by the discovery of an unusual HGT event to the plastid genome which shows a rpl36 gene of bacterial origin was transferred to the plastid of an ancestor of these two groups (Rice and Palmer 2006).

The complex story of secondary (and tertiary) endosymbiosis is far from resolved, one of the competing theories being that the cryptomonad/haptophyte plastid was gained from an ancient secondary endosymbiosis with a red algae, then spread by a tertiary endosymbiotic event giving rise to the plastid of the stramenopiles and alveolates. This event is proposed to have occurred before the HGT event of the rpl36 gene into the cryptomonad/haptophyte plastid lineage (Bodyl, Stiller et al. 2009) (Fig. 3).

The investigation of the complex evolution of the plastid organelle is an ongoing process. New plastid varieties are being discovered regularly, changing our understanding of the spread of photosynthesis among eukaryotes. The use of non-culture-based molecular and phylogenetic techniques has proved to be a powerful method which can more accurately assess microbial diversity and has uncovered many novel groups of organisms. These approaches have aided our understanding of the topology and evolutionary history of the tree of life.

1.9 Identification of novel putatively plastid bearing lineage.

A previously published study (Rappe, Suzuki *et al.* 1998) used eDNA gene libraries targeting the 16S SSU rRNA gene to study eukaryotic plastid diversity in environmental samples taken at sites on the Pacific and Atlantic coasts of the USA. Phylogenetic analysis of the results uncovered several sequences thought to represent novel uncultured lineages of plastid bearing eukaryotes. One of these, OM270 (GenBank U70723), appeared to branch at the base of the cryptomonad/haptophyte clade and appeared to be distinct from any previously characterised lineages according to the phylogenetic analysis performed in the study. Our preliminary phylogenetic analysis seemed to confirm the OM270 sequence represented the plastid encoded 16S SSU rRNA gene of a novel uncultured lineage of eukaryotic putative algae branching with the cryptomonad/haptophyte radiation and as such warranted further study.

The aim of this project is to investigate the evolutionary position and environmental diversity and distribution of a putative novel eDNA plastid SSU rDNA sequence previously recovered using environmental gene library

analysis. We aim to better understand the evolutionary and ecological significance of this uncultured putative algae.

1.10 Specific aims of this project:

1. Demonstrate that the novel sequences are authentic by replicate sampling.
2. Identify the complexity of the novel plastid lineage by multi-primer eDNA clone library analysis from multiple global aquatic sampling sites.
3. Investigate environmental abundance by comparative qPCR analysis in marine environments.
4. Use FISH methodologies to investigate crude cell and plastid morphology.

2.0 Materials and Methods.

This materials and methods section has been taken from Kim, Harrison* et al (manuscript submitted) with the methods performed by myself J. Harrison expanded for this thesis.*

Sections carried out by:

J. Harrison: **2.1.2, 2.2.2, 2.2.3, 2.2.4, 2.2.5, 2.2.9, 2.3.1, 2.3.1.1, 2.3.1.2, 2.3.1.3, 2.3.1.4, 2.3.2, 2.3.3.**

Dr. E. Kim: 2.1.1, 2.2.1, 2.2.3, 2.2.4, 2.2.7, 2.2.8, 2.2.9, 2.3.1, 2.3.1.3, 2.3.2, 2.5.

Dr. S. Sudek: 2.2.6, 2.4.

2.1 Sampling.

2.1.1 Oceanic sampling.

Oceanic samples were collected during 13 research expeditions (Table 1). Two were from the Florida Straits (WS0518 and WS0815) and two were from a transect from coastal to oligotrophic waters in the Pacific Ocean (CN107, the cruise conducted in July 2007 and CN207, the cruise conducted in October 2007). In addition nine cruises were to the Bermuda Atlantic Times-series Study (BATS) sites, eight of which were part of the BATS program and these samples were provided courtesy of S. Giovannoni, representing late winter, spring and summer conditions over three different years. The ninth cruise to BATS was in June 2005 (OC413). For all but BATS program samples, 0.5 - 2 L of seawater was vacuum filtered onto a 0.2 µm Supor filter (Pall Gelman, Port Washington, NY, USA) for DNA harvesting. The BATS program typically filtered between 70 - 90 L of seawater.

2.1.2 UK coastal and freshwater sampling.

Additional surface water samples were collected from 7 sites in the UK: 4 coastal marine sites from Devon, UK (Budleigh Salterton, Sidmouth on 03/11/2008, Lyme Regis, and Seaton on 18/03/2009) and 3 eutrophic freshwater reservoirs located within Dartmoor National Park, Devon, UK (Tottiford, Kennick, and Trenchford on 27/01/2010) (Table 1). The samples were collected from the shallow water using a combination of 1l durans and 10l

Region	Date (d/m/yr)	Station	Latitude (N)	Longitude (W)	Depth (m)	Temp (^{o}C)	Sal (PSU)
N. Pacific	10/07/2007	67-75	35.96	-123.49	30	12.21	33.49
	02/10/2007	67-70	36.129	-123.49	50	12.39	33.14
	06/10/2007	67-155	33.171	-129.257	5	19.03	33.2
	06/10/2007	67-155	33.171	-129.257	86*	13.62	33.12
N. Atlantic	04/11/2008	FS04	25.39143	79.56962	75	26.9	36.08
	03/11/2008	FS01	25.38202	80.03668	1	26.54	36.11
	01/08/2005	FS01	25.60111	80.06639	65	23.5	36.3
	01/06/2005	BATS	31.65555	64.6225	15	25.51	36.68
UK**	03/11/2008	Budleigh	50.63	-3.325	Surf		
	03/11/2008	Sidmouth	50.678	-3.325	Surf		
	18/03/2009	Lyme	50.723	-2.933	Surf		
	18/03/2009	Seaton	50.703	-3.062	Surf		
	18/03/2009	Kennick (FW)	50.644	-3.691	Surf		
	27/01/2010	Tottiford (FW)	50.632	-3.683	Surf		
	27/01/2010	Trenchford (FW)	50.63	-3.688	Surf		

Table 1. Locations and characteristics of environmental samples used for DNA sequencing. Latitude and longitude are given in decimal values. Temperature and salinity were not measured for UK samples. Abbreviations: N, north; FS, Florida Straits; BATS, Bermuda Atlantic Time-series Study; FW, fresh water site; Surf, surface; *, this clone library was constructed using universal 16S rDNA sequence primers (for bacteria and plastids), not the rappemonad specific primer sets. ** Samples collected by myself, J. Harrison with assistance from Dr. M. D. M. Jones.

containers and stored on ice for the short journey back to the laboratory where all samples were processed within four hours. Upon return to the laboratory, 1 L aliquots from each of the marine and freshwater UK samples were passed through a 20 µm mesh (Miracloth, CalBiochem, San Diego, CA, USA) prior to vacuum filtration onto 2.0 µm polycarbonate membrane filters.

2.2 DNA extraction, clone library construction and sequencing.

2.2.1 Oceanic samples.

For most Pacific Ocean, Florida Straits, and BATS clone libraries, DNA was extracted using the DNeasy kit (Qiagen, Hilden, Germany). For qPCR, the DNA extraction protocol was modified from the DNeasy kit (Qiagen) that involves a bead beating step as described previously (Hewson, Poretsky *et al.* 2009). A single Pacific sample, from Station 67-155 at 86 m (October 2007) was extracted using a sucrose protocol (Cuvelier, Allen *et al.* 2010) and these cells had been collected on 3 µm pore size filter (293 mm in diameter). BATS program samples were extracted as described in a previous study (Treusch, Vergin *et al.* 2009).

2.2.2 UK coastal and freshwater samples.

For estuarine and freshwater samples DNA was extracted using the MoBio Powersoil™ DNA kit (MO BIO Laboratories, Inc., Carlsbad, CA, USA) according to the manufacturer's instructions. This method homogenizes and lyses cells using both chemical and mechanical (bead beating) methods, non-DNA contaminants are precipitated from the solution chemically and removed by centrifugation. The resulting supernatant is then centrifuged in a collection tube

with a silica membrane filter. This allows contaminants to pass through whilst the DNA binds to the membrane (with the aid of a high concentration salt solution). An ethanol wash solution is then passed over the filter by centrifugation to remove any remaining contaminants leaving the DNA bound to the silica membrane filter. Finally, the DNA is washed from the filter by centrifugation with a 10mM tris (tris(hydroxymethyl)aminomethane) solution. The DNA was then stored at -20º C.

2.2.3 Oligonucleotide primer design.

To facilitate the design of a range of group specific PCR primers, a comprehensive alignment of 16S plastid encoded rRNA genes was constructed covering a broad range of taxa from across the eukaryotic photosynthetic tree of life. Manual comparison of these 16S rDNA alignments allowed identification of suitable primer sites which were unique to the OM270 (GenBank U70723) (Rappe, Suzuki *et al.* 1998) template sequence. The Primer 1 primer design software (http://www.ncbi.nlm.nih.gov/tools/primer-blast/) was also utilised, using OM270 (GenBank U70723) (Rappe, Suzuki *et al.* 1998) and/or MC622-32 (GenBank EF052198) (McDonald, Sarno *et al.* 2007) as templates (Table 2).

2.2.4 PCR of separate environmental DNA samples.

PCR reactions were performed using TaKaRa Ex Taq™ (TaKaRa Bio Inc., Otsu, Japan) or Promega Master Mix (Promega, Madison, WI, USA) (used by myself J. Harrison). Protocols with annealing temperatures between 51°C and 63°C, with extension times of 1 min were used for primer sets Rappe-16S-0199-5', Rappe-16S-0204-5', Rappe-16S-1278-3', and Rappe-16S-1293-3'.

The primer set Rappe-16S-0677-5′ and Rappe-16S-1119-3′ was used for analysis of the UK marine and freshwater samples using an initial denaturing step of 5 minutes at 95 ºC, followed by 35 cycles of a denaturing step at 95ºC for 1 minute, an annealing step at 63ºC for 1 minute and an extension step at 72ºC for 1 minute followed by a final extension step of 72ºC for 5 minutes. (Primers and protocol used for the UK samples were both developed by myself, J. Harrison).

In addition, 16S rDNA from Station 67-155 at 86 m (October 2007) was amplified using a universal primer set that amplifies bacteria and plastid 16S rRNA genes. Amplicons were gel-purified and cloned using the pGEM®-T Easy Vector System (Promega), the TOPO TA cloning kit (Invitrogen, Carlsbad, CA, USA), or the StrataClone™ PCR cloning kit (used by myself, J. Harrison) (Stratagene, La Jolla, CA, USA). This kit uses a T/A topoisomerase-based method to ligate the purified PCR product into the StrataClone PCR Cloning

Primer name	**Relative positions within *Escherichia coli* sequence**	**Primer sequence (5'→3')**
Rappe-16S-0199-5'	180–199	TAT GCC GCA AGG TGA AAT AC
Rappe-16S-0204-5'	186–204	GCA AGG TGA AAT ACG CAA G
Rappe-16S-0677-5' *	653–677	AGA GTG TGG TAG AGG TAG AGG GAA T
Rappe-16S-1119-3' *	1143–1119	CCC AAC TGA ATG ATG GTA ACT AAA G
Rappe-16S-1278-3'	1304–1278	CGA ACT TAG ACT AAG TTT ATG AGA TTC
Rappe-16S-1293-3'	1313–1293	CCT ACA ATC CGA ACT TAG ACT

Table 2. PCR primers designed and used to amplify rappemonad sequences from environmental samples. Primers marked with * designed by myself, J. Harrison.

Vector pSC-A-amp/kan (Stratagene, UK). The vector was used to transform *Escherichia coli* high competency cells following the manufacturers'

instructions. Transformants were plated on LB plates containing ampicillin (50 μg mL^{-1}) and 5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside (X-gal; 40 μL per plate) and incubated overnight at 37 ^{o}C. White colonies from each sample were selected and grown overnight at 37 ^{o}C with shaking in LB media with added ampicillin. A number of clones from each library were sequenced on a Beckman Coulter CEQ8000 capillary DNA sequencer (Beckman Coulter, Brea, CA, USA) or several different ABI sequencers (Applied Biosystems, Foster City, CA, USA).

2.2.5 Testing ecological overlap of target plastid 16S SSU rDNA sequences and biliphytes.

To test if the presence of target plastid 16S SSU rDNA sequences and biliphyte nuclear 18S SSU rDNA sequences overlapped in the sampled freshwater environments we conducted a two stage nested PCR survey of the three freshwater DNA samples shown to possess the target plastid 16S SSU rDNA sequences. We used group specific primers that target a majority of sampled biliphyte sequences to interrogate the freshwater environmental DNA samples collected from Kennick, Trenchford and Tottiford reservoirs (Table 3). For the PCR we used an initial denaturing step at 95°C for 5 minutes, followed by 35 cycles of 1 minute denaturing at 94°C, 1 minute annealing at 52°C and extension time of 1 minute at 72°C, followed by a final extension time of 5 minutes at 72°C. The initial PCR revealed no identifiable PCR band while the nested amplification generated weak bands of the appropriate size. We replicated this experiment and sequenced ten clones from each PCR amplicon (6 clone libraries in total). This analysis recovered fungal, cryptomonad-like and kathablepharid-like sequences (Appendix Table 1) but did not retrieve any

biliphyte 18S rDNA sequences suggesting that within these freshwater samples the biliphytes were absent while the target plastid 16S SSU rDNA sequences were present.

Primer name	Relative positions within *Arabidopsis thaliana* sequence	Sequence (5'→3')
Bili-18S-134-5'	108–134	CAG TTA TCG TTT ATT TGA TGA TCT CTT G
Bili-18S-151-5'	132–151	TTG CTA CTT GGA TAC CCG TG
Bili-18S-795-3'	820–795	TCC TAT TCT ATT ATT CCA TGC TAA CC
Bili-18S-1073-3'	1097–1073	GAC TTT GAT TTC TCA TAA GGT GCA T

Table 3. PCR primers designed to specifically target the majority of environmental sequences representing the biliphyte 18S rRNA gene.

2.2.6 Comparison of the presence of target plastid 16S SSU rDNA sequences and biliphytes in different environmental size fractions.

Comparison of sequences retrieved using universal 18S and 16S rRNA gene primer sets showed that the biliphyte 18S and target plastid 16S SSU rDNA sequences were recovered from different size fractions at the N. Pacific site where both types of libraries were constructed (Station 67-155).

2.2.7 Amplification of near full-length target group plastid rDNA gene cluster sequences.

Near full-length plastid rDNA operon sequences were obtained from environmental samples using forward PCR primers specific to our target plastid encoded rDNA sequences in combination with ‘universal’ reverse primers (Table 4). PCR conditions were adjusted based on the predicted melting temperature of PCR primers and the expected amplicon size (annealing

Primer name	Relative positions within *E. coli* sequence	Sequence (5'→3')
Rappe-16S-199-5'	180–199	TAT GCC GCA AGG TGA AAT AC
Rappe-16S-845-5'	817–845	CGA TGG ATA CTA GAT GTT GCG TAA CTT GA
Rappe-16S-851-5'	820–851	TGG ATA CTA GAT GTT GCG TAA CTT GAT TAT GC
Rappe-16S-1451-5'	1428–1451	GCC CGA AGT CGT TAC CTT ATC TGG
Rappe-23S-1810-5'	1791–1810	CGC TTA CCT CCA CAC CGA GA
Rappe-23S-2088-5'	2068–2088	GGA GAG CTG GAA GCA AGC ATC
Rappe-23S-2275-5'	2247–2275	GGG ACA AGC CTA AAC TAC TTG TAT TCA AG
23S-2395-3'	2418–2395	TTT AGY CTT ACG AGG TGG TCC TCG
23S-2445-3'	2470–2445	TCT TTT CAC CTT TCC CTC RCG GTA CT
Rappe-23S-2607-5'	2581–2607	GGC AGT RGC AAG GTT AAG RTG TTC ACA
23S-2690-3'	2717–2690	CCT CCA CTT AGT GTT ACC TAA GCT TCA C
23S-3638-3'	3665–3638	TTT GCC GAG TTC CTT AGA GAG AGT TAT C
23S-3959-3'	3978–3959	TCC AGG TGC AGG TAG TCC GC
23S-4506-3'	4530–4506	GAA CTG TCT CAC GAC GTT CTG AAC C
23S-4515-3'	4539–4515	ATA TGG ACC GAA CTG TCT CAC GAC G

Table 4. PCR primers used for amplifying near complete target group plastid rDNA operons.

temperatures between 55°C and 58°C, extension times between 2 and 3 min). The first PCR reactions amplified fragments spanning the 3'-end of 16S rDNA (85–685 nt), Ile tRNA, Ala tRNA, through to the 5'-end of 23S rDNA (425–720 nt). Subsequent PCR runs allowed us to extend these sequences further towards the 3'-end of 23S rDNA locus. Amplified fragments were gel-purified, cloned, and sequenced as described in section 2.2.4.

2.2.8 Amplification of nucleus-encoded rRNA gene cluster of biliphytes.

The nucleus-encoded rDNA operon sequences of biliphytes, including 18S rDNA, ITS1, 5.8S rDNA, ITS2, and 28S rDNA, were also obtained using the protocol described in 2.2.7 using the primers listed in appendix table 2.

2.2.9 Identification of chimaeric sequences.

A number of approaches were adopted to identify and eliminate chimaeric sequences. For all extended rRNA gene array sequencing (target group plastid rDNA and biliphyte nuclear rDNA locus sequences), multiple PCR reactions were performed using different primer sets and two different environmental DNA samples, station FS01 surface waters and FS04 deep chlorophyll maximum (75 m) waters (collected in November 2008, Table 1). This allowed us to independently verify each section of the putative contig (by performing PCRs in two samples). For all environmental 16S rDNA clone library sequences, a stepwise approach was used to identify chimaeras. Primarily, all sequences that were grouped into 99% cluster groups and which included sequences generated from independent PCR reactions were classified as true sequences. The remaining sequences were treated as possible chimera artefacts and were investigated in two ways. First, the diversity alignment was taken and split in half generating two sub-alignments. These sub-alignments were then used to repeat the phylogenetic bootstrap analysis (detailed in section 2.3.2). This analysis did not identify any sequences that appeared to re-position across the phylogeny between the total alignment analysis and the two half subsection analyses suggesting that no one sequence was composed of mixed phylogenetic signal. Second, the unmasked sequence alignment was studied in detail to identify molecular synapomorphies such as insertion/deletions and specific character motifs. We did not find any such synapomorphies inconsistently distributed across the 16S rDNA alignment, suggesting, given the data currently available, that the sequences reported here are not chimaeric.

2.3 Molecular phylogenetics.

2.3.1 Alignment construction.

2.3.1.1 Environmental sequence alignment.

A plastid 16S SSU rDNA alignment was constructed focusing on a 478 bp region of the gene sampled by all our target plastid group specific primer sets. This alignment was constructed with a view to investigating the diversity of this novel group across multiple aquatic environments (Table 1). Please note that this analysis included a comprehensive sampling of sequences representing our target plastid 16S SSU group from GenBank with only two being found in this public database. Taxa were also selected from haptophytes as preliminary phylogenetic analysis had suggested the target environmental 16S SSU rRNA sequences grouped as sister to this clade. Cryptomonads were included in the alignment as they have been shown to be a close neighbour of the haptophytes in multigene phylogenies (Patron, Inagaki *et al.* 2007) and their inclusion should help resolve the branching position of the target plastid 16S SSU sequence within this clade. The rhodophytes were included as an outgroup as the putative endosymbiotic ancestor of the haptophyte/cryptomonad plastid (Cavalier-Smith 2000).

2.3.1.2 Clustering of environmental sequences.

Prior to alignment, all sequences were clustered into 99% identity groups using Sequencher (Gencode) and one representative of each 99% cluster group was used in the final alignment (Appendix table 3). The alignment included a reduced sampling of outgroup taxa. The selection of these outgroup taxa was

based on the phylogenetic analysis detailed section 2.3.1.3 which had more complete taxon sampling covering the photosynthetic tree of life and used the cyanobacteria as an outgroup. In this analysis however, the cyanobacteria were replaced as an outgroup and red algae used instead as this was a better reflection of the taxa represented in this analysis.

2.3.1.3 Near complete target group plastid rDNA gene cluster sequence alignment.

Near complete plastid rDNA operon sequences encompassing the 16S rDNA, two tRNAs, and the 23S rDNA regions were used to investigate the evolutionary branching position of the novel plastid sequences. The taxa included in the alignment were carefully chosen from available full length rRNA gene cluster sequences. Haptophyte and cryptomonad taxa were chosen as preliminary phylogenetic analysis suggested they represent the closest branching groups to the target group and it was hoped this analysis would resolve their branching order. Newly acquired sequences of *Rebecca salina* and *Exanthemachrysis gayraliae* were included to represent the Pavlovales a divergent haptophyte group known to branch at the base of the haptophyte clade, the inclusion of which aids the resolution of the target group position with respect to the haptophyte clade. A number of stramenopile taxa were included as these are thought to possess a red algal-derived plastid and to branch closely with the haptophyte/cryptomonad radiation (Cavalier-Smith 2000). Primary plastid-bearing rhodophyta were included as it is thought to be a secondary endosymbiotic event involving an ancestral rhodophyte which gave rise to the plastid of the cryptomonads, haptophytes and stramenopiles among others. The two further primary plastid-bearing lineages of the viridiplantae and

the glaucophyta were also included and the cyanobacteria were included as an outgroup.

2.3.1.4 Biliphyte nucleus-encoded rRNA gene cluster alignment.

A concatenated alignment was constructed from biliphyte nucleus-encoded 18S and 28S rDNA sequences obtained from environmental samples to re-examine the branching position of biliphytes. A comprehensive range of taxa were selected including representatives of the major groups from across the eukaryotic tree of life. When analyzing near-complete rDNA operon sequences, we acknowledge that some of the sequences assembled together from environmental amplicons (section 2.2.8) could be derived from distinct but very closely related species or strains. However, minor intra-operon heterogeneity should not impact the placement of the target plastid encoded or biliphyte nuclear encoded gene cluster sequences relative to those of other lineages. Some taxa or groups that demonstrated long branch lengths were excluded from the analysis such as the excavates and some amoebazoa. These taxa were removed in order to avoid the effect of long branch attraction artefacts (Bergsten 2005) in the situation where a set of more closely related taxa could not be included in order to interrupt the branch length. Long branch attraction is a phylogenetic artefact which results in erroneous groupings based on high evolutionary rate rather that actual evolutionary relationships.

2.3.2 Methods of analysis.

For each analysis sequences were manually aligned with a wide sampling of homologues from GenBank using the MUSCLE (Edgar 2004) alignment algorithm in Seaview ver 4.2.6 (Gouy, Guindon *et al.* 2010). Ambiguously

aligned regions were removed. Fast maximum likelihood (ML) trees were inferred based on 100 random maximum-parsimony starting trees and with the GTRMIX model using RAxML ver. 7.04 (Stamatakis 2006). RAxML is a program for sequential and parallel fast maximum likelihood based inference of large phylogenetic trees. It starts with an initial parsimony tree calculated from the provided alignment data, as this is a relatively fast method for building a tree. The process is repeated a number of times, in this case 100, for one analysis resulting in multiple starting parsimony trees. The best starting tree is selected and then subjected to maximum likelihood re-analysis which is then used to compute a final consensus tree. RAxML can also be used to perform bootstrap analysis. To expedite the process of creating RAxML tree topologies the Perl script EasyRax was used, which is available from http://projects.exeter.ac.uk/ceem/easyRAx.html. Bootstrap analyses were based on 1,000 replicates. Because compositional bias can cause phylogenetic reconstruction artefacts (Foster and Hickey 1999) we ran an additional bootstrap analysis using Log-Det methods (Lockhart, Steel *et al.* 1994). Log-Det method compensates for compositional bias among sequences during phylogenetic analysis. The method utilises the determinant of the observed comparative positional divergence matrices producing an assessment of evolutionary relationships by recovery of an additive distance between sequences regardless of compositional bias (Charleston, Hendy *et al.* 1994). The Log-Det analysis was used with stepwise addition (10 random starting trees per replicate) and tree-bisection-reconnection branch swapping algorithm using the program PAUP* (Swofford 2003).

2.3.3 Comparative topology test.

Comparative alternative phylogenetic topology tests of both a nuclear rDNA alignment, that included biliphyte sequences, and a plastid rDNA alignment, that included sequences from the target 16S SSU plastid group, were used to investigate, within an identical taxon set, if we could exclude that these two groups represented equivalent branching positions. For topology tests, four alignments were constructed: 1) plastid 16S rDNA, 2) nuclear 18S rDNA, 3) concatenated plastid 16S and 23S rDNA sequences and 4) concatenated nuclear 18S and 28S rDNA. Taxa were selected from the cryptomonads and the haptophytes as the biliphytes have been shown in previous studies to branch weakly with the cryptomonads (Rappe, Suzuki *et al.* 1998) and preliminary phylogenetics has shown the target plastid sequences to branch with the haptophytes and it is these topologies which are being assessed. The stramenopiles were included as a close phylogenetic neighbour and the rhodophyta as an outgroup. The taxon sampling was amended so that alignment 1 and 2 had equivalent taxon sampling and alignments 3 and 4 had equivalent taxon sampling apart from in the case of the plastid rDNA alignments (alignment 1 and 3), target plastid rDNA sequences were included, while for the nuclear rDNA alignments (alignment 2 and 4) billiphyte rDNA sequences were included. A number of topologies were calculated under 4 constraints (see Fig. 7). A) Candidate sequences branching with the haptophytes. B) Candidate sequences branching with the cryptomonads. C) Candidate sequences branching as a basal branch to the cryptophyte/haptophyte clade and D) Candidate sequences branching with the stramenopiles. Using both parsimony and distance methods using PAUP* (Swofford 2003) the program consel v0.1k was used to compare topologies using Kishino-Hasegawa (KH) and

approximately unbiased (AU) tests (Shimodaira and Hasegawa 2001; Shimodaira 2002). These tests assess the likelihood of a previously generated phylogenetic tree topology accurately representing the specified alignment used to create the tree and therefore can be used to compare alternative tree topologies from the same data set. The p-value generated represents the statistical significance of the specified tree topology accurately representing the given data normal confidence boundaries of 95% can be assumed.

2.4 Quantitative PCR.

The following primer pair was designed from manual inspection of 82 rappemonad sequences, and comparison to outgroups, and used for qPCR on environmental samples: Rappe-16S-1257-5' (5' – ACA ATG GCT AAG ACA AAG AGC - 3') and Rappe-16S-1293-3' (Table 2). qPCR reactions used 12.5 µl PowerSybrGreen 2x Mastermix (Applied Biosystems), 2.5 µl of each primer (final concentration, 50 nM), 5.5 µl H_2O and 2 µl template (plasmid standards or diluted environmental samples, see below). Thermal cycling conditions for qPCR were 10 min at 95°C, followed by 45 cycles of 15 s at 95°C and 1 min at 60°C using an AB7500 instrument (Applied Biosystems). Data were collected during the annealing phase. Plasmid standards (see below) and no-template controls were run in triplicate, as were the environmental samples themselves, with an additional well for an inhibition test (see below). For each qPCR run, the threshold was set to 0.2 and baseline values were automatically calculated using the 7500 software package (Applied Biosystems).

The specificity of the qPCR primer pair was tested by qPCR on genomic DNA derived from cultures of the cryptophytes *Rhodomonas salina* CCMP1319 and *Rhodomonas abbreviata* CCMP1178, as well as environmental samples.

The cultured phytoplankton did not yield qPCR product and all amplifications of environmental samples in which rappemonads were detected yielded a single peak dissociation curve (T_m 73.8°C), suggesting the absence of primer-dimers or other unspecific products. To further verify qPCR specificity, we cloned qPCR products from Pacific and Atlantic environmental samples using pCR2.1-TOPO (Invitrogen) and sequenced 11 and 10 clones, respectively. Sequencing was performed on a 3130xl genetic analyzer using Big Dye terminater v3.1 chemistry (Applied Biosystems). All sequences retrieved belonged exclusively to the target group.

Absolute quantification of the qPCR reactions was achieved using a plasmid standard curve that was also used to ascertain the efficiency of the qPCR primer set. Plasmid NP67-155D3Bb026_6Oct07_86m, containing a rappemonad 16S rDNA template from Station 67-155 at 86 m that is identical to OM270, was used as a standard. This plasmid was quantified spectrophotometrically (NanoDrop) and serially diluted with TE buffer (1 to 10 Tris-Cl to EDTA at pH 8). Eight dilutions with copy numbers from 10 to 10^8 copies per 2 µl (the volume added to each qPCR reaction) were used for the standard curve. C_t values of these dilutions versus the respective copy numbers on a log scale were plotted and a linear regression fitted. Efficiency of the primer set was 85% (r^2=0.99) and all dilutions of the standard curve were within the dynamic range. For enumeration of environmental samples, C_t values of the samples were converted to copy numbers $well^{-1}$ using the standard curve regression line and then back-calculated to copies ml^{-1} seawater using DNA extraction elution volume, template dilution and volume of seawater filtered. Standard deviation of the data derived from the three sample wells were used to estimate technical error.

For each environmental sample, inhibition of the qPCR reaction was tested by adding 10^4 copies of the plasmid containing a rappemonad 16S rDNA insert to a sample well which also contained the environmental template (while reducing the amount of water in the overall master mix to maintain volume). The observed cycle threshold (C_t) was compared to the C_t of the 10^4 standard. Using the formula Efficiency=1-(C_t^{sample} – C_t^{10000})/C_t^{10000}, we rejected a sample as "inhibited" if Efficiency<0.96. Samples were then re-run at a greater dilution. The theoretical detection limit calculated from volume of seawater filtered, quantity of DNA extracted, or volume eluted, and extent of dilution of the extract was between 1 and 38 copies ml^{-1} seawater for BATS 2003 samples and all Pacific samples and between 53 and 647 copies ml^{-1} seawater for BATS samples from two prior years, which exhibited considerable inhibition. "Zero" values derived from most of these samples were considered unreliable due to the high levels of inhibition. For statistical analyses only those samples with detection at 38 copies ml^{-1}, or better, were used to allow fair comparison across samples. In addition it should be noted that phosphorus and nitrate plus nitrate were sometimes below detection limits at BATS.

2.5 Tyramide signal amplification fluorescence in situ hybridization.

TSA-FISH was performed on CN207 cruise samples, using two rappemonad-specific probes (RappeA and RappeB; Table 5), and the hybridized cells were detected by epifluorescence microscopy. To prepare and store samples for hybridization, seawater (~180 ml) was preserved with paraformaldehyde (1%, final concentration) for a minimum of 1 hour at 4°C in the dark. The seawater was filtered onto a 0.2 µm Anodisc (25 mm, Whatman, Maidstone, UK), the filters were dried with an ethanol series (50%, 80% and 100% ethanol diluted in

autoclaved 18.2 MΩ H_2O for 3 min each) and stored at -80°C prior to hybridization. FISH was performed on replicate filter pieces in conjunction with tyramide signal amplification (TSA) using a modification of previously published methods (Not, Valentin *et al.* 2007; Cuvelier, Ortiz *et al.* 2008; Wendeberg 2010). The horseradish peroxidase (HRP)-conjugated probes that specifically bind to the target group (Table 5) were purchased from biomers.net (Ulm, Germany). The rappemonad-specific probes (RappeA and RappeB) contained at least 5 mismatches to non-target organisms. The RappeB probe was designed to target a majority of the rappemonad clade I (86%, or 89 out of 103 sequences analyzed and had a single base mismatch or no mismatches to the RappeB probe). The RappeA probe also targets at least some of the sub-groups within clade I, although the extent to which all members of this group are labelled by this probe could not be evaluated because many of the sequences obtained for this group did not span the region targeted by the probe. Due to limited sample availability, FISH was not used to systematically enumerate cells and hybridizations using both probes simultaneously were not performed due to differing optimal hybridization conditions (see below). Probe specificity was empirically verified by hybridization against two non-target algal species, the cryptophyte alga *Rhodomonas salina* CCMP1319 and the

Probe	***E. coli* target position**	**Sequence (5'→3')**	**Formamide %**
RappeB-Helper	814–837	CGC AAC ATC TAG TAT CCA TCG TTT	40
RappeB	838–862	GCT AAA ACA CTG CAT AAT CRA GTT A	40
RappeA	1438–1464	CGT CCC CCW GAT AAG GTA AC	25
RappeA-Helper	1467–1488	CCY AGT TAT CAG CTC TGC CTT A	25

Table 5. Oligonucleotide probes designed in this study and used for TSA-FISH.

haptophyte alga *Rebecca salina*. A culture strain of *R. salina* was kindly provided by T. Nakayama (Tsukuba University).

Immediately prior to beginning each TSA-FISH experiment, individual filters were cut into 6–8 pieces. To inactivate any endogenous peroxidase, the filter pieces were immersed in 0.02 M HCl for 15 min, followed by a 5 sec rinse in PBS solution and incubation in sterile water for 10 min at room temperature. Each filter piece was then covered with 20 µl hybridization buffer (900 mM NaCl, 20 mM Tris pH 7.5, 0.02% SDS, 25 or 40% formamide [depending on the probe used], 1% blocking reagent, 10% dextran sulfate) that included a HRP-conjugated probe and an unlabeled helper oligonucleotide (0.15 ng μl^{-1} final concentration each). A helper oligonucleotide was added to the hybridization mixture as it can increase the *in situ* accessibility of the target regions (Fuchs, Glockner *et al.* 2000). The concentration of formamide was empirically determined: 40% for the RappeA probe and 25% for the RappeB probe. Hybridizations were carried out for 3 hours in a humidity chamber at 46°C, followed by a brief rinse in pre-warmed wash buffer (20 mM Tris pH 7.5, 56 or 159 mM NaCl depending on the formamide concentration used for hybridization, 0.01% SDS, 5 mM EDTA) and 15 min incubation in wash buffer at 46°C. Filter pieces were then incubated in TNT buffer (0.1 M Tris pH 7.5, 150 mM NaCl, 0.074% Tween 20) for 15 min at room temperature. A TSA™ plus fluorescence kit (PerkinElmer, Inc., Waltham, MA, USA) was used for signal amplification; the amplification reagent was comprised of 1 volume of the fluorescein tyramide stock solution, 65 volumes of 1X plus amplification diluent, and 35 volumes of 40% dextran sulfate. All the following steps were carried out in the dark. Each filter piece was overlain with 15 µl of the amplification reagent and incubated for 45 min at room temperature. Unbound fluorescein labelled

tyramides were removed using two sequential washes (20 min and 15 min each) with TNT buffer at 46°C. After hybridization, FISH filters were counter-stained with 4', 6-diamidino-2-phenylindole (DAPI). This was performed by counterstaining with 1.5 µg ml^{-1} for 5 min, rinsing for 5 min at room temperature in sterile H_2O, dipping in 80% ethanol for 2 min, air drying for approximately 10 min, and finally applying 40 µl of mounting solution [antifading solution AF1 (Citifluor, London, UK)]. The cover slip was then sealed to the slide with nail polish and filters counted within two days.

Several controls were performed alongside each of the rappemonad hybridizations of field samples under the same conditions as for the positive probes. The bacterial antisense NON338 probe (5' –ACT CCT ACG GGA GGC AGC - 3') (Worden, Chisholm *et al.* 2000) was used as a negative control for all hybridizations, this probe should not hybridize to any cells and can be used to account for non-specific hybridization. A probe that targets an unrelated freshwater bacterium (5' – GCA CCA ATT TCA AAT AAA GTC AAC - 3') (Kim, Park *et al.* 2010) was also used as a negative control. A no-probe control was also added at least once for each sample. Biliphyte cells were labelled using previously published probes (Not, Valentin *et al.* 2007) using the same controls as above. We tested methods for sequential TSA-FISH hybridizations, necessary for the use of two different labels with horseradish peroxidase based protocols, to allow evaluation of potential co-localization of biliphyte and rappemonad probes. However, sequential hybridizations rendered a weak signal for whichever test probe was applied in the second round of hybridization; we could thus not apply such an approach with confidence to environmental samples.

Hybridized cells were observed using a Zeiss Axiovert 200M

epifluorescence microscope (Jena, Germany) with an X-Cite® 120PC Q lamp (EXFO Life Sciences, Mississauga, ON, Canada). Probe signal was detected in the FITC channel and associated DAPI fluorescence (showing the cell nucleus) was verified. Photographs of cells were taken using a Hamamatsu ORCA-R2 digital camera (Jena, Germany) and were used for cell size measurements.

3.0 Results and Discussion.

This results and discussion section has been adapted from Kim, Harrison* et al (manuscript submitted). With the sections produced by myself J. Harrison expanded for this thesis.*

Sections produced by:

J. Harrison: **3.1, 3.2, 3.4, 3.5, 3.6, 3.7**

Dr. E. Kim: 3.1, 3.5, 3.9

Dr. S. Sudek: 3.3, 3.8

A novel plastid 16S SSU rDNA sequence OM270 (Genbank U70723) was identified from a study using plastid-encoded 16S SSU rRNA environmental gene libraries to investigate ultraplankton diversity at two sites on the Atlantic and Pacific coasts of the USA (Rappe, Suzuki *et al.* 1998). Preliminary phylogenetic analyses seemed to confirm the OM270 sequence represented a novel red algal secondary plastid-bearing organism branching at the base of the cryptomonad haptophyte radiation. In the absence of detailed morphological data, we name this novel group the "rappemonads", in reference to the first author of the publication initially reporting the OM270 clone sequence (Rappe, Suzuki *et al.* 1998).

3.1 PCR amplification of rappemonad sequences from environmental DNA.

A diverse set of Plastid 16S SSU rDNA sequences was recovered using primer sets designed to target the OM270 Plastid rDNA sequence (Fig. 4). Sequences were recovered from eDNA samples collected at Pacific and Atlantic open ocean and coastal marine locations and oligotrophic freshwater locations in the

UK (Fig.4, Fig 4 inset, Appendix Tab 4) demonstrating that the rappemonad group is present in a wide geographical range of aquatic habitats. The presence of rappemonad sequences in both marine and freshwater habitats indicates either that the group features both marine and freshwater ecotypes or that individuals within the group demonstrate a flexible physiology adaptable to different aquatic habitats, allowing them to survive in both marine and freshwater habitats.

3.2 Test of the ecological overlap of rappemonads and biliphytes.

Nested PCR using primers specific to the biliphyte nuclear 18S SSU rDNA (targeting the majority of known biliphyte diversity) was used to assess whether biliphytes were present in freshwater locations previously demonstrated to contain rappemonad plastid 16S SSU rDNA sequences (shown in section 3.1). This protocol recovered nuclear 18S SSU rDNA sequences of kathablepharid-like (kathablepharids being non-photosynthetic members of the cryptomonad/haptophyte clade), cryptomonad-like and fungal taxa but did not recover any biliphyte sequences from eDNA clone libraries constructed from these freshwater samples (Appendix Tab. 1). This result supports the hypothesis that rappemonads and biliphytes are two distinct groups as the biliphyte sequences do not occur in the oligotrophic freshwater environment shown to contain rappemonad sequences.

3.3 Clone library surveys to assess the presence of rappemonads and biliphytes in different size-fractionated marine samples.

Rappemonads and biliphytes were recovered in plastid 16S SSU and nuclear 18S SSU rDNA clone libraries, respectively, from the same depth and site

(Station 67-155, Appendix Tab. 4) for which 768 clones were sequenced per size fraction and primer pair. However, rappemonad plastid 16S SSU rDNA sequences were only found in the 3-20 µm size fraction, while biliphyte nuclear 18S SSU rDNA sequences were recovered in the 0.8-3 µm size fraction. This result again supports the hypothesis that the rappemonads and biliphytes represent distinct groups as biliphytes are present in the smaller size fraction while rappemonads are not.

3.4 Phylogenetic analysis of clustered rappemonad environmental plastid 16S SSU rDNA sequences.

When the rappemonad plastid 16S SSU rDNA sequences recovered by this study along with the 2 sequences recovered from the Genbank database were clustered at 99% identity, 6 distinct clusters were formed with 5 sequences remaining distinct from any of the cluster groups (Appendix Tab. 3). Two of the cluster groups, 99% cluster 0002 and 99% cluster 0006 included sequences from multiple marine sample sites, suggesting these clusters are present in diverse geographical locations. 99% cluster 0016 contained both freshwater and UK coastal marine sequences suggesting this cluster represents a subgroup whose physiology is flexible to different aquatic habitats, (Fig. 4). However, it is unclear how plastid 16S SSU rDNA sequence variability relates to genome or ecological diversity within eukaryotic algae; these 99% cluster groups may encompass numerous distinct ecotypes.

The phylogenetic tree produced from the clustered sequence data showed the rappemonad plastid 16S SSU rRNA sequences forming a distinct monophyletic group with strong bootstrap support (96/99). The group appears to be divided into two distinct clades (clade I and clade II). However both clades

are weakly supported by both phylogenetic methods used, however the support could be improved with the use of longer (>478bp) 16S rDNA sequences in the analysis. Neither clade appears to be geographically specific containing sequences from varied sample locations (Fig. 4).

Figure 4. (Following page) Environmental diversity and sample locations. Maximum likelihood phylogenetic tree of environmental 16S rDNA sequences obtained herein (bold) as well as OM270. Clone library sample sites (Table 1) and the number of sequences obtained are shown to the right, including marine samples from the North Pacific (NP), Florida Straits (FS), Bermuda Atlantic Time-series Study (BATS), as well as United Kingdom (UK) coastal and freshwater samples. The scale bar indicates the inferred number of nucleotide substitutions per site. Bootstrap support values (≥ 50%) are from RaxML and Log-Det distance analyses, respectively. Inset map shows the approximate positions of sites sampled.

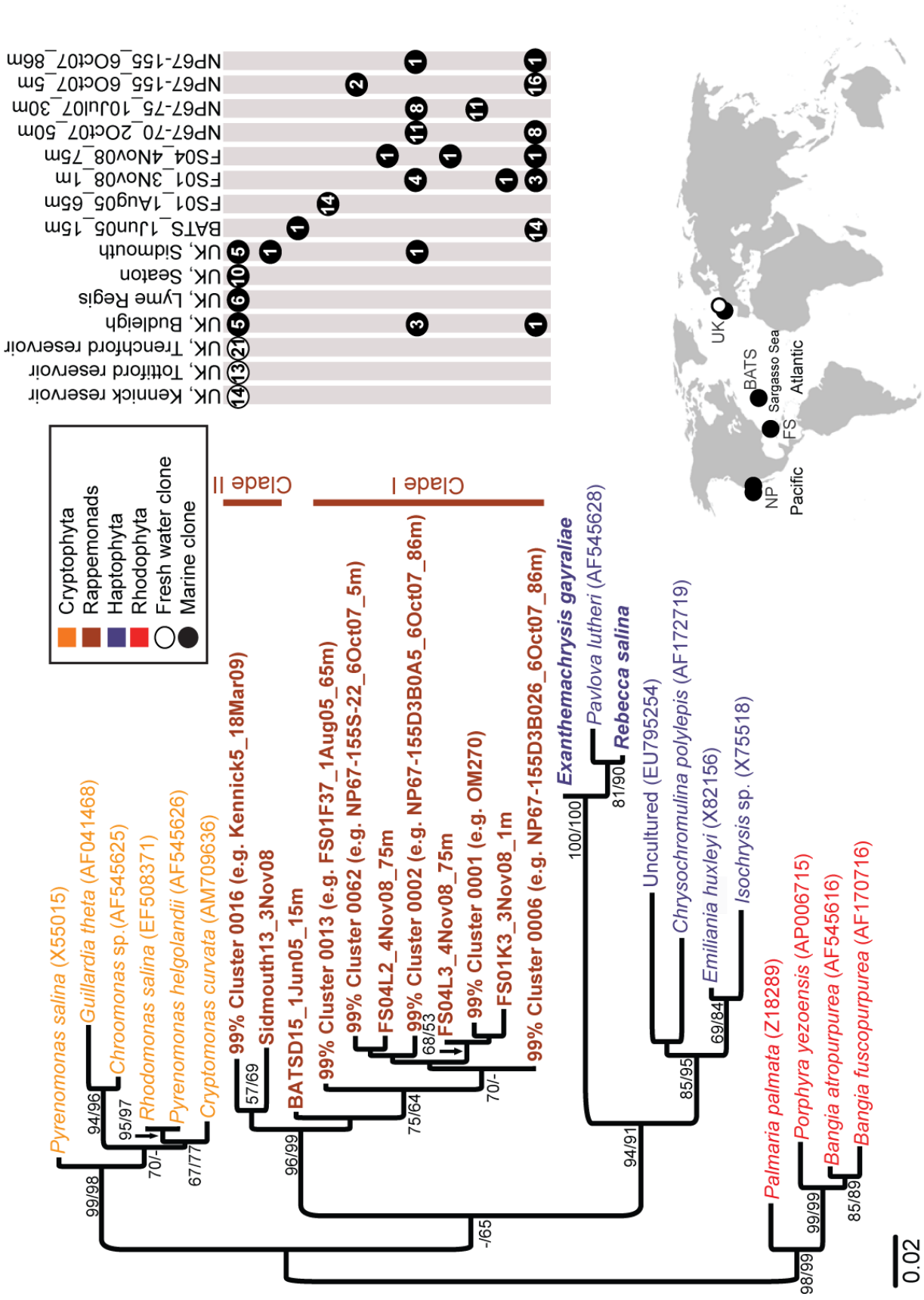


Figure 4 (see previous page for legend)

3.5 Phylogenetic analysis of rappemonad plastid rDNA gene cluster.

To resolve the phylogenetic position of the rappemonads, a region spanning the plastid 16S SSU rRNA gene through to the 23S LSU locus (including the intergenic transcribed spacers and two tRNAs) was amplified and sequenced. This allowed us to establish the group's evolutionary position relative to known red algae and algae with red algal-derived secondary plastids (specifically cryptophytes, haptophytes and stramenopiles). Analyses of these near-complete environmental plastid rDNA gene clusters (for separate SSU and LSU analysis see Appendix Figure 1), together with newly acquired sequences from the haptophytes, representing the pavlovales, *Rebecca salina* and *Exanthemachrysis gayraliae*, showed that the novel group is a unique lineage that branches deeply within the haptophyte and cryptophyte radiation. The rappemonads demonstrate a strongly supported (89/93) sister relationship to the haptophytes using full length gene cluster phylogenetic analysis (Fig. 5). This relationship suggests that the rappemonads possess a secondary endosymbiotic plastid of red algal origin which shares a common ancestor with the haptophyte and cryptomonad plastid (Cavalier-Smith 2000).

This analysis shows the rappemonads to branch with the haptophytes which are a diverse and anciently diverged lineage (Medlin, Saez *et al.* 2008) of measurable environmental importance (Liu, Probert *et al.* 2009; Cuvelier, Allen *et al.* 2010). Based on bootstrap analyses, it is reasonable to assert that rappemonads represent a deeply diverged, previously unrecognized haptophyte or haptophyte-like group (Fig. 5) occupying a deep branching position, within the cryptophyte /haptophyte radiation based on the rDNA gene cluster analysis.

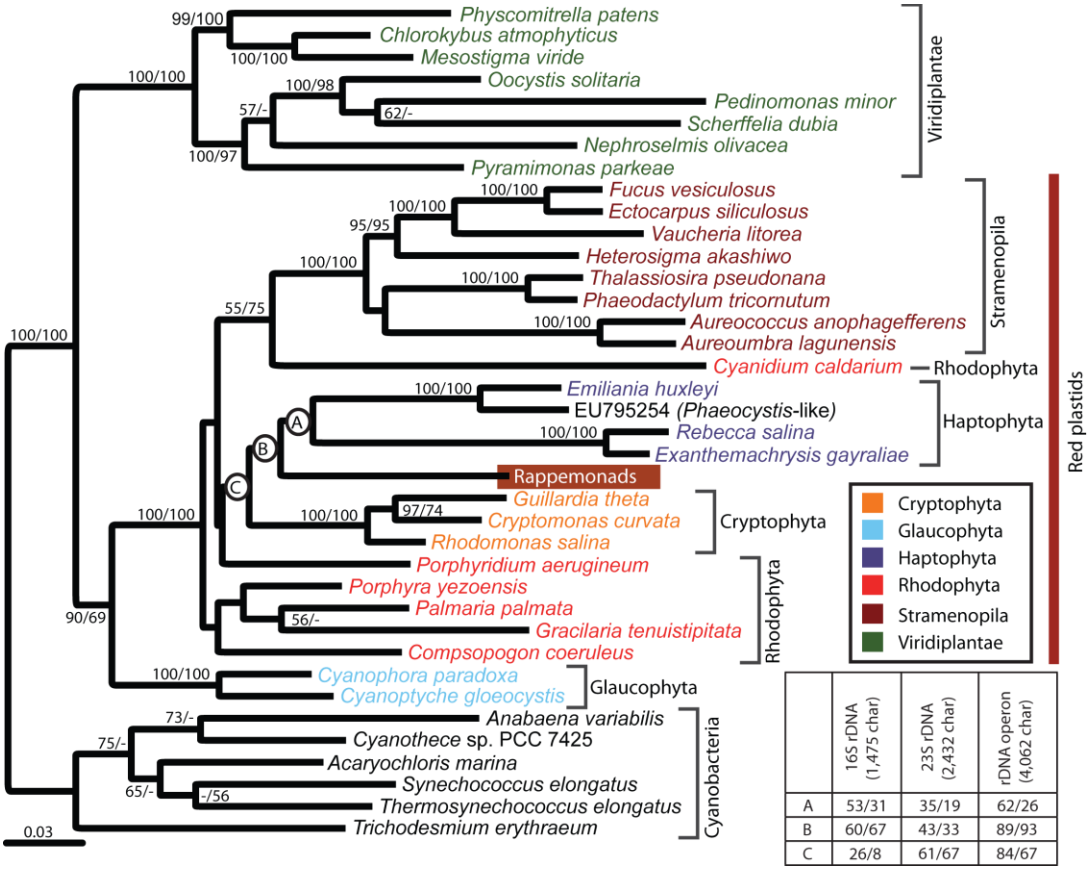


Figure 5. Maximum likelihood (ML) phylogenetic tree of plastid 16S-Ile tRNA-Ala tRNA- 23S rDNA sequences rooted with select cyanobacteria. Sequences in bold were generated as part of this study, including newly obtained Pavlophyceae sequences (Moreira, von der Heyden *et al.* 2007). The scale bar indicates the inferred number of nucleotide substitutions per site. Bootstrap support values (≥ 50%) are from ML and Log-Det distance analyses, respectively. Inset shows bootstrap support for nodes labeled A, B and C using 16S rDNA, 23S rDNA and near complete rDNA operon alignments (see Appendix Fig. 1 for representative 16S and 23S rDNA tree topologies).

3.6 Phylogenetic analysis of biliphyte nuclear rDNA gene cluster.

The multi-gene phylogenetic analysis of concatenated nuclear 18S SSU and 28S LSU rDNA genes including the biliphyte nuclear rRNA gene cluster recovered in this study was used to re-test the biliphyte branching position within the eukaryote tree of life. This analysis used more phylogenetic information and improved taxon sampling than previously published (Not, Valentin *et al.* 2007; Cuvelier, Ortiz *et al.* 2008). As in previous studies, the resulting phylogeny lacked bootstrap support above 50% for placement of

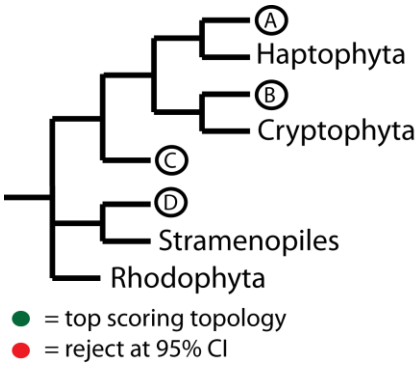
biliphytes with respect to known eukaryotic groups. Furthermore, the phylogeny demonstrated that biliphytes are separated from cryptophyte and haptophyte algae by multiple branches resolved with weak bootstrap support, although the maximum likelihood (ML) bootstrap analyses demonstrated that glaucophytes formed a moderately supported clade (71%) with cryptophytes and

Figure. 6. Maximum likelihood (ML) phylogenetic tree of nuclear 18S and 28S rDNA sequences including a comprehensives ampling of eukaryotic taxa, focused on algal lineages and eukaryotes with shorter branching sequences. The aim of this phylogeny was to investigate the branching position of the biliphyte lineage (in bold and highlighted in yellow). This tree demonstrated weak bootstrap support for the placement of the biliphyte branch but in contradiction to previous studies based only on 18S rDNA analysis this 18S/28S rDNA analysis demonstrated no support for the placement of the biliphytes as sister to the katablepharid/cryptophyte clade. Bootstrap support values (≥ 50%) are from ML and Log-Det distance analyses, respectively. The scale bar indicates the inferred number of nucleotide substitutions per site.

katablepharids, to the exclusion of biliphytes (Fig. 6). This suggests a different branching relationship than that from 18S rDNA analyses alone (Not, Valentin *et al.* 2007; Cuvelier, Ortiz *et al.* 2008).

3.7 Comparative topology test.

Alternative phylogenetic topology tests of the nuclear biliphyte and plastid rappemonad alignments were used to investigate whether we could reject the hypothesis that the two groups represented equivalent branching positions on the plastid and nuclear rDNA trees. The SSU rDNA analysis showed the best scoring branching position for the rappemonads as (A), sister to the haptophytes whereas the highest scoring topology for the biliphytes is (B), branching with the cryptomonads. The complete rDNA gene cluster analysis showed the best scoring position for the rappemonads to be either (A), branching with the haptophytes or (C), branching in a basal position to the cryptomonad/haptophyte clade. Interestingly this analysis significantly rejects the rappemonads branching with the cryptomonads. The highest scoring topology for the biliphytes is (C), branching in a basal position to the cryptomonad/haptophyte clade (Fig. 7). Whilst it is not possible to reject many alternative topologies in this analysis the results suggest differing branching positions for the two groups. This data suggests the rappemonad plastid and biliphyte nucleus have incongruent ancestries, and therefore the 16S rappemonad plastid sequences do not represent sequences from the plastid of the biliphytes (Fig. 7). However this may result from secondary or tertiary endosymbioses, or potentially methodological artefacts.



Branching position	Novel plastid 16S AU test	Novel plastid 16S KH test	Biliphyte 18S AU test	Biliphyte 18S KH test	Novel plastid 16S-23S AU test	Novel plastid 16S-23S KH test	Biliphyte 18S -28S AU test	Biliphyte 18S -28S KH test
Ⓐ	0.847	0.573	0.311	0.177	0.618	0.477	0.257	0.237
Ⓑ	0.304	0.135	0.687	0.503	0.001	0.017	0.532	0.341
Ⓒ	0.233	0.089	0.147	0.503	0.590	0.523	0.691	0.659
Ⓓ	0.039	0.027	0.086	0.077	0.039	0.027	0.414	0.309

Figure 7. Results of alternative topology comparison tests. The simplified phylogeny demonstrates the four alternative branching positions for the rappemonad 16S rDNA and biliphyte 18S rDNA sequences tested. These branching positions are marked A, B, C, D on the tree. The best topology score for each branching position for each alignment is included in the grid using the approximately unbiased (AU) and Kishino-Hasegawa (KH) tests. Topology values coloured red could be excluded at the 95% confidence interval. Best scoring topologies are marked green. This analysis demonstrates the rappemonads are consistently placed with the haptophyte/cryptophyte clade while there is no consistent signal for the placement of the biliphytes.

3.8 Quantitative PCR.

To investigate distributions in the environment, a rappemonad specific qPCR assay was developed. The rappemonads were detected in 23 of 48 marine euphotic-zone samples ranging from 15 ± 14 to 4318 ± 38 gene-copies ml^{-1} (Appendix Tab.4). Sixteen samples in which cells were not detected showed high inhibition (an issue frequently seen with environmental DNA extractions), and required dilution to levels that detection limits were poor (in one case minimum detection being 647 copies ml^{-1}, although typically much better than this; Appendix Table 4); these data were not included in additional statistical analyses comparing abundance with environmental parameters. High rappemonad 16S rRNA gene-copies ml^{-1} were detected in what appeared to be a late-winter bloom in surface waters at the Bermuda Atlantic Time-series Site (BATS) (Fig. 8). The water column at this time showed several fluorescence maxima and rappemonads were concentrated at the shallowest of these, indicating deeper maxima were composed of other taxa. Very few or no

rappemonads were detected in stratified summer-time conditions, when there was a pronounced deep chlorophyll maximum (Fig. 8). In addition, 11 of 12 samples from a North Pacific anti-cyclonic eddy, in which colder more nutrient rich waters, akin to the higher nutrient availability in late-winter BATS samples, were brought to the surface, resulting in a shallower mixed layer, had measureable copy numbers (averaging 186 ± 78 gene copies ml^{-1}). Other samples from the 500-mile transect had fewer (Appendix Tab.4).

Although depth, temperature, salinity, phosphate, chlorophyll *a* and nitrate plus nitrite were measured, no statistically significant differences were identified between samples where rappemonads were detected and those samples where none were detected (t-tests, Mann-Whitney Rank Sum), considering only those in which nutrient concentrations were above detection limits. The upper range of phosphate concentrations in rappemonad-positive samples was 0.69 µM, lower than for all samples (1.19 µM). In addition,

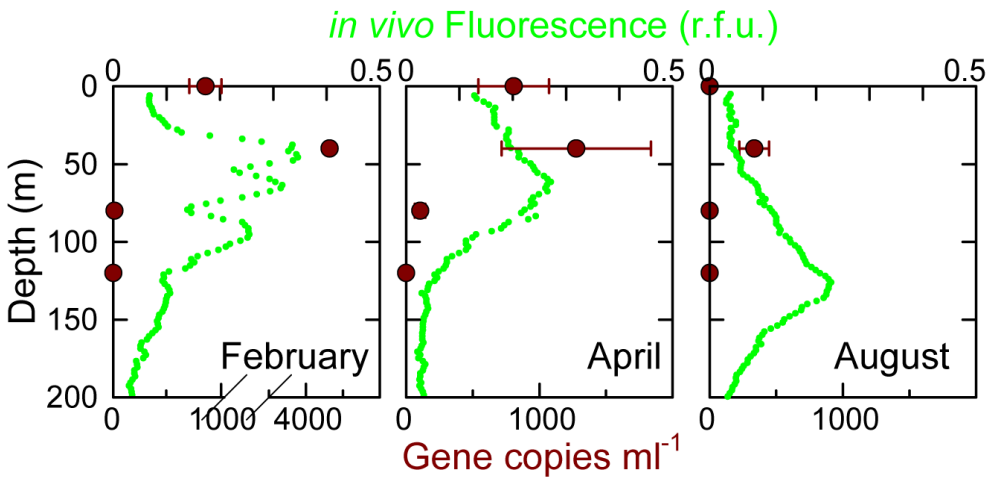


Figure 8. Rappemonad distributions in the Sargasso Sea in 2003. Seasonal transitions are shown for BATS as revealed by qPCR assays for the 16S rRNA gene (dark red, gene 17 copies ml-1). *In vivo* chlorophyll fluorescence is shown in relative fluorescence units (green, r.f.u.). Note differences in X-axis scales.

chlorophyll *a* ranged from 0.070 to 0.690 µg l^{-1} for rappemonad-containing samples and 0.03 to 2.71 µg l^{-1} for all samples. Average temperatures of the samples investigated, and those that contained rappemonads were identical (17

± 4°C). Rappemonad sequences were detected in waters ranging from 11-24°C water, with sequences also recovered from 26°C waters (Table 1), although this sample was not screened by qPCR, as the DNA was not extracted in a quantitative manner. The temperature range of rappemonad-containing samples again indicates this lineage may have a broad ecophysiological range.

3.9 Tyramide signal amplification fluorescence *in situ* hybridization identification of the rappemonads.

We also characterized rappemonads morphologically. Oligonucleotide probes targeting two different 16S rRNA regions of distinct rappemonad sub-groups within clade I (Fig. 4) were designed for use with tyramide signal amplification fluorescence *in situ* hybridization (TSA-FISH), and verified for specificity on a series of non-target controls. The probes were applied to 2 samples for which FISH filters were available and notable gene copies ml^{-1} were detected by qPCR. Rappemonads (n=88) measured 5.7 ± 1.0 (SD) µm in width (shortest dimension) and 6.6 ± 1.2 (SD) µm in length (longest dimension). Each cell appeared to contain two, three or four plastids (Fig. 9; Appendix Fig. 2) with four being the most common (52% or 46 of 88 cells); it is conceivable that instances of three or four organelles associated with a single nucleus correspond to dividing stages of the cell. Alternatively, plastids can be bi-lobed, giving the appearance of multiple plastids when only one is present and cell orientation can bias imaging. The microscopy analyses also revealed a faint reddish fluorescence co-localized with the hybridized plastid compartments using a DAPI filter set (excitation: G365, emission: LP420), presumably derived from residual chlorophyll pigments. Average cell bio-volume was 112 μm^3 and

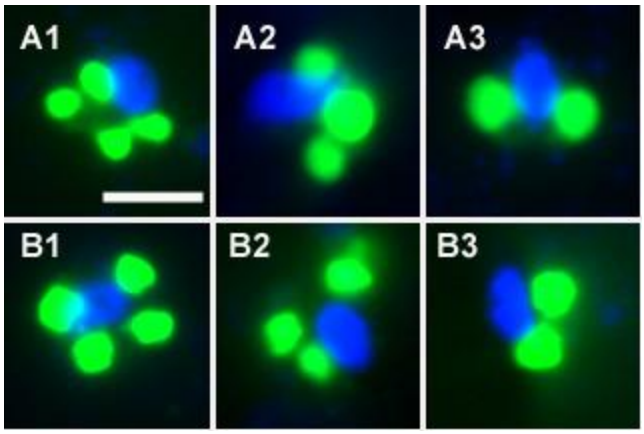


Figure 9. Fluorescence micrographs of rappemonads in the North Pacific. The DAPI-stained nucleus (blue) was often slightly elongated with a tapering end. Two to four plastids appeared present per cell (green, TSA-FISH labeled). Cells shown in A1–A3 were detected using the RappeA probe while B1–B3 were detected with the RappeB probe. The scale bar represents 5 µm.

carbon content 27 pg C cell^{-1}, based on cellular dimensions and an established carbon conversion factor (Worden, Nolan *et al.* 2004). This is significantly greater cellular carbon content than the picophytoplankton (<2-3 µm diameter) that dominate such regions, for example small haptophytes which range from 1 to 3 pg C cell^{-1} (Cuvelier, Allen *et al.* 2010). Large cell size may also be responsible for the rarity of reported rappemonad sequences. The majority of environmental 16S (and 18S) rDNA clone libraries, especially those using primer sets targeting plastid-16S rRNA genes, are constructed from water pre-filtered through 3 µm pore-sized filters e.g. (Fuller, Campbell *et al.* 2006; Lepere, Vaulot *et al.* 2009), which would select against these cells. In addition high abundance of heterotrophic bacteria (10^5-10^6 ml^{-1}) could effectively swamp 16S rDNA libraries constructed using universal primers, so that few plastid-derived sequences are attained.

This method was also used to observe biliphyte cells which appeaed to be smaller than rappemonads, with two biliphyte clades being 3.5 ± 0.9 x 3.0 ± 0.9 and 4.1 ± 1.0 x 3.5 ± 0.8, respectively (Cuvelier, Ortiz *et al.* 2008). Biliphytes detected here in the North Pacific, using the same FISH probes (Not, Valentin

et al. 2007; Cuvelier, Ortiz *et al.* 2008), were similar in size to those described by Cuvelier *et al.* (Cuvelier, Ortiz *et al.* 2008). However, the highly punctate phycobilin-like (orange) fluorescence reported previously (Not, Valentin *et al.* 2007; Cuvelier, Ortiz *et al.* 2008) was not seen co-localized with the hybridized (North Pacific) cells and biliphytes were not present at levels significantly above background counts for negative controls.

4.0 Conclusion.

In this study we have used molecular and phylogenetic methods combined with microscopy to characterise a major new plastid lineage and therefore add a unique branch on the plastid tree of life.

We have demonstrated that this newly characterised plastid lineage, the rappemonads, are diverse and are present in a wide range of open ocean, coastal marine, and oligotrophic freshwater environments with qPCR analysis suggesting that they form transient seasonal blooms in open ocean environments.

Our phylogenetic analysis, of the rappemonad plastid 16S SSU rDNA sequences suggests the rappemonads form a monophyletic clade. We show the rappemonad clade forming a number of subclades within the main radiation containing various ecotypes from various geographic locations. These results suggest the rappemonads are an ecologically diverse and geographically widespread. Along with the rDNA operon analysis these results show the rappemonads form a clade which branches deeply within the cryptomonad/haptophyte radiation as a sister to the haptophyte branch. This branching position indicates that they could either be a highly novel haptophyte lineage or an entirely new group. However, the sequences used for the plastid 16S eDNA analysis were only 478 bp in length, which could affect the phylogenetic signal gained from the analysis. The results of this analysis could also be influenced by methodological artefacts of molecular phylogenetics, such as inappropriate taxon sampling and Long Branch attraction which can impose false topology onto a phylogenetic tree. Mutational saturation can also influence the interpretation of phylogenetic results, and can invalidate phylogenetic

analysis of certain positions within a sequence, especially when dealing with ancient divergences.

The biliphytes are a recently discovered uncultured group of eukaryotes identified from environmental nuclear 18S rDNA clone libraries (Not, Valentin *et al.* 2007; Cuvelier, Ortiz *et al.* 2008). Phylogenetic analysis based on nuclear encoded 18S SSU rDNA genes, conducted in these previous studies, suggested that the biliphytes branch with the cryptomonads. We conducted alternative topology tests for both biliphyte nuclear rDNA genes and rappemonad plastid rRNA genes to investigate if the rappemonad plastid sequences represent a novel plastid lineage distinct from the biliphytes; we demonstrated that, although our results are not conclusive, they do suggest that it is unlikely that the rappemonads represent the plastid of the biliphyte group, although further work is required to confirm this result and rule out a separate tertiary endosymbiotic event. The identification and phylogenetic analysis of the rappemonad nuclear encoded 18S SSU rDNA sequence could be used to investigate whether the two lineages are indeed distinct or originate from the same cell. This could be achieved using indirect methods using a flow sorted sample combined with environmental eDNA analysis, however, gaining the rappemonads in pure culture would allow characterization of their morphology and biochemistry and allow the direct sequencing of their nuclear rDNA gene cluster. This would allow the phylogenetic analysis of the nuclear encoded rDNA sequences, and to check for convergence between nuclear and plastid phylogenies.

The cellular morphology of rappemonad cells was investigated using TSA-FISH microscopy analysis. The results showed the rappemonads were comparatively large cells showing high cellular carbon content. Rappemonad

cells appeared to possess one bilobed plastid organelle structure or 2, 3, or 4 individual plastid structures.

Our TSA-FISH analysis showed the rappemonad plastid 16S rDNA gene is actively transcribed in various environments suggesting that they are actively phototrophic. To confirm the photosynthetic function of the rappemonad plastid organelle a larger portion of the plastid genome needs to be sequenced. The discovery of genes encoding a core component of photosynthetic protein complexes located on the rappemonad plastid genome would strongly suggest the rappemonads have retained active photosynthetic function. Furthermore, such data would allow qPCR assays to be designed to assess the phototrophic contribution of rappemonads in aquatic environments.

The cryptomonads and haptophytes possess an unusual rpl36 gene in their plastid genome gained from horizontal gene transfer (Rice and Palmer 2006). Further analysis of the rappemonad plastid genome would allow the characterization of the rappemonad rpl36 plastid gene. This would indicate whether the rappemonads share the same unusual rpl36 gene HGT as the cryptomonads and the haptophytes. The presence of this rpl36 HGT would support the hypothesis that the plastids of rappemonads, cryptomonads and haptophytes share a common ancestor and thus suggest that rappemonads do branch within this clade. Furthermore, the haptophytes possess an unusual photopigment, 19′-hexanoyloxyfucoxanthin, and has been shown in a recent study to be abundant in the open ocean (Liu, Probert *et al.* 2009). Confirmation that rappemonads possess 19′-hexanoyloxyfucoxanthin would support the close relationship between the rappemonads and the haptophytes and could potentially indicate that the rappemonads contribute to 19′-

hexanoyloxyfucoxanthin abundance in the open ocean or other aquatic environments (Liu, Probert *et al.* 2009).

The coccolithophores are an ecologically important subgroup of the haptophytes, which utilise environmental carbon to form calcium carbonate scales. The coccolithophores form huge open ocean blooms. These large populations collapse, usually due to viral predation (Bratbak, Egge J. K. *et al.* 1993) sequestering this carbon to the deep-sea to form geological deposits (Jordan and Chamberlain 1997). It is possible that as a close evolutionary relation of the haptophytes the rappemonads also share a scale-forming trait. If this is the case it would suggest that scale formation was a characteristic present in the rappemonad/haptophyte common ancestor and has been secondarily lost by several haptophyte subgroups.

Finally, it would be informative to investigate morphological characteristics of the rappemonads. For example given their evolutionary position as close relatives of the haptophytes, they may also possess the distinctive ‘haptonema’ an organelle which is distinct to the haptophyte lineage (Manton 1967), which would provide further evidence for this phylogenetic position.

This study has expanded upon previously published work (Rappe, Suzuki *et al.* 1998; Not, Valentin *et al.* 2007; Cuvelier, Ortiz *et al.* 2008) showing that our understanding of microbial biodiversity and community composition is far from complete. The rappemonads appear to be an important part of the marine photosynthetic biosphere potentially overlooked by sampling artefact, and are therefore likely to play a influential role in marine geochemical cycling. The study of these organisms would make a valuable contribution to marine ecosystem modelling.

5.0 References.

Adl, M. S., A. G. B. Simpson, *et al.* (2005). "The new higher level classification of eukaryotes with emphasis on the taxonomy of protists." J. Eukaryot. Microbiol. **52**(5): 399-451.

Amann, R., W. Ludwig, *et al.* (1995). "Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation." Microbiol. Rev. **59**(1): 143-169.

Andersen, R. A. (2004). "Biology and systematics of heterokont and haptophyte algae." Am. J. Bot. **91**(10): 1508-1522.

Archibald, J. M. (2007). "Nucleomorph genomes: structure, function, origin and evolution." BioEssays **29**(4): 392-402.

Archibald, J. M. (2009). "The puzzle of plastid evolution." Curr. Biol. **19**(2): 81-88.

Archibald, J. M. and C. E. Lane (2009). "Going, going, not quite gone: nucleomorphs as a case study in nuclear genome reduction." J. Hered. **100**(5): 582-590.

Barbrook, A. C. and C. J. Howe (2000). "Minicircular plastid DNA in the dinoflagellate *Amphidinium operculatum*." Mol. Gen. Genet. **263**(1): 152-158.

Barns, S. M., C. F. Delwiche, *et al.* (1996). "Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences." P. Natl. Acad. Sci. USA **93**(17): 9188-9193.

Bergsten, J. (2005). A review of long-branch attraction, Blackwell Science Ltd. **21:** 163-193.

Berney, C., J. Fahrni, *et al.* (2004). "How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys." BMC Biology **2**(1): 1-13.

Bhattacharya, D., J. M. Archibald, *et al.* (2007). "How do endosymbionts become organelles? Understanding early events in plastid evolution." BioEssays **29**(12): 1239-1246.

Bhattacharya, D., T. Helmchen, *et al.* (1995). "Comparisons of nuclear-encoded small-subunit ribosomal RNAs reveal the evolutionary position of the glaucocystophyta." Mol. Biol. Evol. **12**(3): 415-420.

Bodyl, A., J. W. Stiller, *et al.* (2009). "Chromalveolate plastids: direct descent or multiple endosymbioses?" Trends Ecol. Evol. **24**(3): 119-121.

Bolte, K., L. Bullmann, *et al.* (2009). "Protein targeting into secondary plastids." J. Eukaryot. Microbiol. **56**(1): 9-15.

Bratbak, G., Egge J. K., *et al.* (1993). "Viral mortality of the marine alga *Emiliania huxleyi* (Haptophyceae) and termination of algal blooms." Mar. Ecol.-Prog. Ser. **93**: 39-48.

Burki, F., K. Shalchian-Tabrizi, *et al.* (2008). "Phylogenomics reveals a new 'megagroup' including most photosynthetic eukaryotes." Biology Letters **4**(4): 366-369.

Cai, X., A. L. Fuller, *et al.* (2003). "Apicoplast genome of the coccidian *Eimeria tenella.*" Gene **321**: 39-46.

Cavalier-Smith, T. (1999). "Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree." J. Eukaryot. Microbiol. **46**(4): 347-366.

Cavalier-Smith, T. (2000). "Membrane heredity and early chloroplast evolution." Trends. Plant. Sci. **5**(4): 174-182.

Cavalier-Smith, T. (2004). "Only six kingdoms of life." P. Roy. Soc. Lond. B Bio. **271**(1545): 1251-1262.

Chambouvet, A., P. Morin, *et al.* (2008). "Control of toxic marine dinoflagellate blooms by serial parasitic killers." Science **322**(5905): 1254-1257.
Charleston, M. A., M. D. Hendy, *et al.* (1994). "The effects of sequence length, tree topology, and number of taxa on the performance of phylogenetic methods." J. Comput. Biol. **1**(2): 133-151.
Cuvelier, M. L., A. E. Allen, *et al.* (2010). "Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton." P. Natl. Acad. Sci. USA **107**(33): 14679–14684.
Cuvelier, M. L., A. Ortiz, *et al.* (2008). "Widespread distribution of a unique marine protistan lineage." Environ. Microbiol. **10**(6): 1621-1634.
Dacks, J. B., P. P. Poon, *et al.* (2008). "Phylogeny of endocytic components yields insight into the process of nonendosymbiotic organelle evolution." P. Natl. Acad. Sci. USA **105**(2): 588-593.
Dawson, S. C. and N. R. Pace (2002). "Novel kingdom-level eukaryotic diversity in anoxic environments." P. Natl. Acad. Sci. USA **99**(12): 8324-8329.
Diez, B., C. Pedros-Alio, *et al.* (2001). "Study of genetic diversity of eukaryotic picoplankton in different oceanic regions by small-subunit rRNA gene cloning and sequencing." Appl. Environ. Microbiol. **67**(7): 2932-2941.
Edgar, R. C. (2004). "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic Acids Res. **32**(5): 1792-1797.
Falkowski, P., R. J. Scholes, *et al.* (2000). "The global carbon cycle: a test of our knowledge of Earth as a system." Science **290**(5490): 291-296.
Field, C. B., M. J. Behrenfeld, *et al.* (1998). "Primary production of the biosphere: integrating terrestrial and oceanic components." Science **281**(5374): 237-240.
Foster, P. G. and D. A. Hickey (1999). "Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions." J. Mol. Evol. **48**(3): 284–290.
Fuchs, B. M., F. O. Glockner, *et al.* (2000). "Unlabeled helper oligonucleotides increase the *in situ* accessibility to 16S rRNA of fluorescently labeled oligonucleotide probes." Appl. Environ. Microb. **66**(8): 3603–3607.
Fuhrman, J. A., K. Mccallum, *et al.* (1992). "Novel major archaebacterial group from marine plankton." Nature **356**(6365): 148-149.
Fuller, N. J., C. Campbell, *et al.* (2006). "Analysis of photosynthetic picoeukaryote diversity at open ocean sites in the Arabian Sea using a PCR biased towards marine algal plastids." Aquat. Micro. Ecol. **43**(1): 79–93.
Funes, S., E. Davidson, *et al.* (2002). "A green algal apicoplast ancestor." Science **298**(5601): 2155.
Gibbs, S. P. (1978). "The chloroplasts of *Euglena* may have evolved from symbiotic green algae." Can. J. Bot. **56**(22): 2883–2889.
Gibbs, S. P. (1981). "The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae." Ann. N.Y. Acad. Sci. 361: 193-208.
Gilson, P. R., V. Su, *et al.* (2006). "Complete nucleotide sequence of the chlorarachniophyte nucleomorph: Nature's smallest nucleus." P. Natl. Acad. Sci. USA **103**(25): 9566-9571.
Giovannoni, S. J., T. B. Britschgi, *et al.* (1990). "Genetic diversity in Sargasso Sea bacterioplankton." Nature **345**(6270): 60-63.
Goericke, R. and N. A. Welschmeyer (1993). "The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and

primary production in the Sargasso Sea." Deep Sea Res. Pt.I **40**(11-12): 2283-2294.
Gould, S. B., R. F. Waller, *et al.* (2008). "Plastid evolution." Annu. Rev. Plant Biol. **59**(1): 491-517.
Gouy, M., S. P. Guindon, *et al.* (2010). "SeaView Version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building." Mol. Biol. Evol. **27**(2): 221-224.
Groisillier, A. S., R. Massana, *et al.* (2006). "Genetic diversity and habitats of two enigmatic marine alveolate lineages." Aquat. Micro. Ecol. **42**(3): 277-291.
Guillou, L., M. Viprey, *et al.* (2008). "Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata)." Environ. Microbiol. **10**(12): 3349-3365.
Hackett, J. D., D. M. Anderson, *et al.* (2004). "Dinoflagellates: a remarkable evolutionary experiment." Am. J. Bot. **91**(10): 1523-1534.
Hackett, J. D., H. S. Yoon, *et al.* (2007). "Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates." Mol. Biol. Evol. **24**(8): 1702-1713.
Hewson, I., R. S. Poretsky, *et al.* (2009). "*In situ* transcriptomic analysis of the globally important keystone N_2 fixing taxon *Crocosphaera watsonii*." ISME J. **3**(5): 618–631.
Howe, C. J., A. C. Barbrook, *et al.* (2008). "The origin of plastids." Philos. T. Roy. Soc. B **363**(1504): 2675-2685.
Howe, C. J., R. E. R. Nisbet, *et al.* (2008). "The remarkable chloroplast genome of dinoflagellates." J. Exp. Bot. **59**(5): 1035-1045.
Huang, J., N. Mullapudi, *et al.* (2004). "Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*." Genome Biol. **5**(11): R88.
Huber, T., G. Faulkner, *et al.* (2004). "Bellerophon: a program to detect chimeric sequences in multiple sequence alignments." Bioinformatics **20**(14): 2317-2319.
Hugenholtz, P., B. M. Goebel, *et al.* (1998). "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity." J. Bacteriol. **180**(18): 4765-4774.
Ishikawa, H. (1977). "Evolution of ribosomal RNA." Comp. Biochem. Phys. B. **58**(1): 1-7.
Janouškovec, J., A. Horák, *et al.* (2010). "A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids." P. Natl. Acad. Sci. USA **107**(24): 10949-10954.
Jarvis, P. and C. Robinson (2004). "Mechanisms of protein import and routing in chloroplasts." Curr. Biol. **14**(24): R1064-R1077.
Jordan, R. W. and A. H. L. Chamberlain (1997). "Biodiversity among haptophyte algae." Biodivers. and Conserv. **6**(1): 131-152.
Kaufman, A. J., D. T. Johnston, *et al.* (2007). "Late archean biospheric oxygenation and atmospheric evolution." Science **317**(5846): 1900-1903.
Keeling, P. J. (2004). "Diversity and evolutionary history of plastids and their hosts." Am. J. Bot. **91**(10): 1481-1493.
Keeling, P. J. (2009). "Chromalveolates and the evolution of plastids by secondary endosymbiosis." J. Eukaryot. Microbiol. **56**(1): 1-8.
Kim, E. and L. E. Graham (2008). "EEF2 analysis challenges the monophyly of archaeplastida and chromalveolata." PLoS ONE **3**(7): e2621.

Kim, E., J. S. Park, *et al.* (2010). "Complex array of endobionts in *Petalomonas sphagnophila*, a large heterotrophic euglenid protist from *Sphagnum*-dominated peatlands." ISME J. **4**: 1108–1120.

Kohler, S., C. F. Delwiche, *et al.* (1997). "A plastid of probable green algal origin in apicomplexan parasites." Science **275**(5305): 1485-1489.

Larkum, A. W. D., P. J. Lockhart, *et al.* (2007). "Shopping for plastids." Trends Plant Sci **12**(5): 189-195.

Leonor Teles-Grilo, M., J. Tato-Costa, *et al.* (2007). "Is there a plastid in *Perkinsus atlanticus* (phylum perkinsozoa)?" Eur. J. Protistol. **43**(2): 163-167.

Lepere, C., D. Vaulot, *et al.* (2009). "Photosynthetic picoeukaryote community structure in the south east pacific ocean encompassing the most oligotrophic waters on Earth." Environ. Microbiol. **11**(12): 3105–3117.

Liu, H., I. Probert, *et al.* (2009). "Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans." P. Natl. Acad. Sci. USA **106**(31): 12803-12808.

Lockhart, P. J., M. A. Steel, *et al.* (1994). "Recovering evolutionary trees under a more realistic model of sequence evolution." Mol. Biol. Evol. **11**(4): 605–612.

Lopez-Garcia, P., F. Rodriguez-Valera, *et al.* (2001). "Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton." Nature **409**(6820): 603-607.

Lovejoy, C., R. Massana, *et al.* (2006). "Diversity and distribution of marine microbial eukaryotes in the arctic ocean and adjacent Seas." Appl. Environ. Microbiol. **72**(5): 3085-3095.

Malmstrom, R. R., M. T. Cottrell, *et al.* (2005). "Biomass production and assimilation of dissolved organic matter by SAR11 bacteria in the northwest atlantic ocean." Appl. Environ. Microbiol. **71**(6): 2979-2986.

Manton, I. (1967). "Further observations on the fine structure of *Chrysochromulina chiton* with special reference to the haptonema, 'peculiar' golgi structure and scale production." J. Cell Sci. **2**(2): 265-272.

Marin, B., E. C. M. Nowack, *et al.* (2005). "A plastid in the making: evidence for a second primary endosymbiosis." Protist **156**(4): 425-432.

Martin, W., T. Rujan, *et al.* (2002). "Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus." P. Natl. Acad. Sci. USA **99**(19): 12246-12251.

Martin, W. and M. J. Russell (2003). "On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells." Philos. T. Roy. Soc. B **358**(1429): 59-83.

Massana, R., J. Castresana, *et al.* (2004). "Phylogenetic and ecological analysis of novel marine stramenopiles." Appl. Environ. Microbiol. **70**(6): 3528-3534.

Massana, R., L. Guillou, *et al.* (2002). "Unveiling the organisms behind novel eukaryotic ribosomal DNA sequences from the ocean." Appl. Environ. Microbiol. **68**(9): 4554-4558.

Massana, R. and C. Pedrós-Alió (2008). "Unveiling new microbial eukaryotes in the surface ocean." Curr. Opin. Microbiol. **11**(3): 213-218.

McDonald, S. M., D. Sarno, *et al.* (2007). "Genetic diversity of eukaryotic ultraphytoplankton in the Gulf of Naples during an annual cycle." Aquat. Micro. Ecol. **50**(1): 75–89.

McFadden, G. I. (2001). "Chloroplast origin and integration." Plant. Physiol. **125**(1): 50-53.
McFadden, G. I., P. R. Gilson, *et al.* (1994). "Evidence that an amoeba acquired a chloroplast by retaining part of an engulfed eukaryotic alga." P. Natl. Acad. Sci. USA **91**(9): 3690-3694.
McFadden, G. I. and R. F. Waller (1997). Plastids in parasites of humans. Bioessays **19:** 1033-1040.
Medlin, L. K., A. G. Saez, *et al.* (2008). "A molecular clock for coccolithophores and implications for selectivity of phytoplankton extinctions across the K/T boundary." Mar. Micropaleontol. **67**(1-2): 69–86.
Moon-van der Staay, S. Y., R. De Wachter, *et al.* (2001). "Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity." Nature **409**(6820): 607-610.
Moore, R. B., M. Obornik, *et al.* (2008). "A photosynthetic alveolate closely related to apicomplexan parasites." Nature **451**(7181): 959-963.
Moreira, D., H. Le Guyader, *et al.* (2000). "The origin of red algae and the evolution of chloroplasts." Nature **405**(6782): 69-72.
Moreira, D. and P. López-García (2002). "The molecular ecology of microbial eukaryotes unveils a hidden world." Trends in Microbiol. **10**(1): 31-38.
Moreira, D., S. von der Heyden, *et al.* (2007). "Global eukaryote phylogeny: Combined small- and large-subunit ribosomal DNA trees support monophyly of Rhizaria, Retaria and Excavata." Mol. Phylogenet. Evol. **44**(1): 255-266.
Morris, R. M., M. S. Rappe, *et al.* (2002). "SAR11 clade dominates ocean surface bacterioplankton communities." Nature **420**(6917): 806-810.
Nisbet, E. G. and N. H. Sleep (2001). "The habitat and nature of early life." Nature **409**(6823): 1083-1091.
Nosenko, T., K. L. Lidie, *et al.* (2006). "Chimeric plastid proteome in the Florida "red tide" Dinoflagellate *Karenia brevis*." Mol. Biol. Evol. **23**(11): 2026-2038.
Not, F., K. Valentin, *et al.* (2007). "Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes." Science **315**(5809): 253-5.
Olsen, G. J., D. J. Lane, *et al.* (1986). "Microbial ecology and evolution: A ribosomal RNA approach." Annu. Rev. of Microbiol. **40**(1): 337-365.
Pace, N. R. (1997). "A molecular view of microbial diversity and the biosphere." Science **276**(5313): 734-740.
Patron, N. J., Y. Inagaki, *et al.* (2007). "Multiple gene phylogenies support the monophyly of cryptomonad and haptophyte host Lineages." Curr. Biol. **17**(10): 887-891.
Ralph, S. A., G. G. van Dooren, *et al.* (2004). "Tropical infectious diseases: metabolic maps and functions of the *Plasmodium falciparum* apicoplast." Nat. Rev. Micro. **2**(3): 203-216.
Rappe, M. S., M. T. Suzuki, *et al.* (1998). "Phylogenetic diversity of ultraplankton plastid small-subunit rRNA genes recovered in environmental nucleic acid samples from the pacific and atlantic coasts of the United States." Appl. Environ. Microbiol. **64**(1): 294-303.
Reyes-Prieto, A., A. Moustafa, *et al.* (2008). "Multiple genes of apparent algal origin suggest ciliates may once have been photosynthetic." Curr. Biol. **18**(13): 956-962.
Reyes-Prieto, A., H. S. Yoon, *et al.* (2010). "Differential gene retention in plastids of common recent origin." Mol. Biol. Evol. **27**(7): 1530-1537.

Ricard, G., N. McEwan, *et al.* (2006). "Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrates-rich environment." BMC Genomics **7**(1): 22.

Rice, D. and J. Palmer (2006). "An exceptional horizontal gene transfer in plastids: gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters." BMC Biology **4**(1): 31.

Richards, T. A. and D. Bass (2005). "Molecular screening of free-living microbial eukaryotes: diversity and distribution using a meta-analysis." Curr. Opin. Microbiol. **8**(3): 240-252.

Richards, T. A., J. B. Dacks, *et al.* (2006). "Evolution of filamentous plant pathogens: gene exchange across eukaryotic kingdoms." Curr. Biol. **16**(18): 1857-1864.

Richards, T. A., A. A. Vepritskiy, *et al.* (2005). "The molecular diversity of freshwater picoeukaryotes from an oligotrophic lake reveals diverse, distinctive and globally dispersed lineages." Environ. Microbiol. **7**(9): 1413-1425.

Rodríguez-Ezpeleta, N. and H. Philippe (2006). "Plastid origin: replaying the tape." Curr. Biol. **16**(2): 53-56.

Rogers, M. B., P. R. Gilson, *et al.* (2007). "The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts." Mol. Biol. Evol. **24**(1): 54-54.

Sanchez-Puerta, M. V., J. C. Lippmeier, *et al.* (2007). "Plastid Genes in a non-photosynthetic dinoflagellate." Protist **158**(1): 105-117.

Shimodaira, H. (2002). "An approximately unbiased test of phylogenetic tree selection." Systematic Biol. **51**(3): 492–508.

Shimodaira, H. and M. Hasegawa (2001). "CONSEL: for assessing the confidence of phylogenetic tree selection." Bioinformatics **17**(12): 1246–1247.

Slamovits, C. H. and P. J. Keeling (2008). "Plastid-derived genes in the nonphotosynthetic alveolate *Oxyrrhis marina*." Mol. Biol. Evol. **25**(7): 1297-1306.

Soll, J. and E. Schleiff (2004). "Protein import into chloroplasts." Nat. Rev. Mol. Cell. Biol. **5**(3): 198-208.

Staay, S. Y. M.-v. d., G. W. M. v. d. Staay, *et al.* (2000). "Abundance and diversity of prymnesiophytes in the picoplankton community from the equatorial Pacific ocean Inferred from 18S rDNA sequences." Limnol. Oceanogr. **45**(1): 98-109.

Stamatakis, A. (2006). "RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models." Bioinformatics **22**(21): 2688–2690.

Stiller, J., J. Huang, *et al.* (2009). "Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses?" BMC Genomics **10**(1): 484.

Stoebe, B., W. Martin, *et al.* (1998). "Distribution and nomenclature of protein-coding genes in 12 sequenced chloroplast genomes." Plant Mol. Biol. Rep. **16**(3): 243-255.

Swofford, D. L. (2003). PAUP*. Phylogenetic analysis using parsimony (*and other methods). Sunderland, MA, Sinauer Associates, Inc.

Taylor, F. J. R. (1974). "Implications and extensions of the serial endosymbiosis theory of the origin of eukaryotes." Taxon **23**(2/3): 229-258.

Tengs, T., O. J. Dahlberg, *et al.* (2000). "Phylogenetic analyses indicate that the 19'hexanoyloxy-fucoxanthin-containing dinoflagellates have tertiary plastids of haptophyte origin." Mol. Biol. Evol. **17**(5): 718-729.

Treusch, A. H., K. L. Vergin, *et al.* (2009). "Seasonality and vertical structure of microbial communities in an ocean gyre." ISME J. **3**(10): 1148–1163.

Tyler, B. M., S. Tripathy, *et al.* (2006). "Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis." Science **313**(5791): 1261-1266.

Weber, A. P. M., M. Linka, *et al.* (2006). "Single, ancient origin of a plastid metabolite translocator family in plantae from an endomembrane-derived ancestor." Eukaryot. Cell **5**(3): 609-612.

Wendeberg, A. (2010). "Fluorescence *in situ* hybridization for the identification of environmental microbes " Cold Spring Harbor Protocols: doi:10.1101/pdb.prot5366.

Withers, N. W., E. R. Cox, *et al.* (1977). "Pigmemts of dinoflagellate Peridinium balticumand its photosynthetic endosymbiont." J. Phycol. **13**(4): 354-358.

Worden, A. Z., S. W. Chisholm, *et al.* (2000). "*In situ* hybridization of *Prochlorococcus* and *Synechococcus* (marine cyanobacteria) spp. with rRNA-targeted peptide nucleic acid probes." Appl. Environ. Microbiol. **66**(1): 284–289.

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

Yoon, H. S., J. D. Hackett, *et al.* (2004). "A molecular timeline for the origin of photosynthetic eukaryotes." Mol. Biol. Evol. **21**(5): 809-818.

Yoon, H. S., J. D. Hackett, *et al.* (2002). "The single, ancient origin of chromist plastids." P. Natl. Acad. Sci. USA **99**(24): 15507-15512.

Zhang, Z., B. R. Green, *et al.* (1999). "Single gene circles in dinoflagellate chloroplast genomes." Nature **400**(6740): 155-159.

6.0 Appendix.

Appendix Table 1. Table of sequences gained from clone libraries constructed from freshwater environmental DNA using biliphyte specific primers.

Origin	Sample	Top BLAST hit	Likely inference
Kennick reservoir	JH1-1	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH1-2	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH1-3	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH1-4	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH1-5	No significant similarity found	Uncertain
	JH1-7	Uncultured eukaryotic picoplankton clone P1.27	Cryptomonad
	JH1-8	Uncultured eukaryotic picoplankton clone P1.27	Cryptomonad
	JH1-9	Uncultured eukaryotic picoplankton clone P1.27	Cryptomonad
	JH1-10	No significant similarity found	Uncertain
	JH4-1	No significant similarity found	Uncertain
	JH4-3	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH4-4	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH4-6	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH4-7	No significant similarity found	Uncertain
	JH4-8	Uncultured kathablepharid clone GHB30.9	Kathablepharid
	JH4-9	No significant similarity found	Uncertain
	JH4-10	No significant similarity found	Uncertain
	jh1-2	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	jh1-3	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	jh1-4	No significant similarity found	Uncertain
Tottiford reservoir	JH2-2	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH2-4	Expression vector pYPX251	Vector
	JH2-5	No significant similarity	Uncertain
	JH2-7	Allelic replacement vector pJC84	Vector
	JH2-9	No significant similarity found	Uncertain
	JH2-10	Shuttle vector pLV.DsRed	Vector
	JH5-1	No significant similarity found	Uncertain
	JH5-3	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH5-4	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH5-7	No significant similarity found	Uncertain
	JH5-8	No significant similarity found	Uncertain
	JH5-10	Cloning vector pDDB57	Vector
	jh2-2	No significant similarity found	Uncertain
	jh2-3	No significant similarity found	Uncertain
	jh2-4	Uncultured Plakinidae sp. clone Elev_18S_603	Fungi
Trenchford reservoir	JH3-1	Plectosphaerella sp. MH727	Fungi
	JH3-3	Uncultured kathablepharid clone GHB30.9	Kathablepharid
	JH3-4	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH3-5	Plectosphaerella sp. MH727	Fungi
	JH3-6	T7RNA polymerase vector pGemT7cat	Vector
	JH3-7	Plectosphaerella sp. MH727	Fungi
	JH3-8	Uncultured kathablepharid clone EB52.129	Kathablepharid
	JH3-9	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH3-10	Plectosphaerella sp. MH727	Fungi
	JH6-1	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH6-3	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH6-5	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH6-6	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH6-7	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH6-8	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH6-9	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	JH6-10	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	jh3-1	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	jh3-2	Uncultured eukaryotic picoplankton clone VP10	Kathablepharid
	jh3-3	Uncultured Plakinidae sp. clone Elev_18S_603	Fungi

Appendix table 2. PCR primers used for the amplification of the near full length nucleus-encoded rRNA gene cluster from biliphytes.

Primer name	Relative positions within *A. thaliana* sequence	Sequence (5'→3')
Bili-1064-5'	1039–1064	GGG ATG TGG AGK CGT TAA CTT TGT AC
Bili-1075-5'	1050–1075	GCG TTA ACT TTG TAC GAC CCT CCA TG
Bili-1668-5'	1645–1668	TCG TTA CTA CCG ATT GGT GTG CAG
Bili1-2677-5'	2654–2677	ACT TGC GTT CGT CCG GTC TTG TAT
Bili1-2681-5'	2659–2681	CGT TCG TCC GGT CTT GTA TCG AC
Bili2-2931-5'	2911–2931	GCC AGC ATC AGT TCG TTC AGC
Bili2-2993-5'	2972–2993	GCT GTG AGG ACT GAG GTT TTG G
28S-3071-3'	3094–3071	TCC TTG GTC CGT GTT TCW AGA CGG
28S-3078-3'	3101–3078	GTT AGA CTC CTT GGT CCG TGT TTC
28S-4813-3'	4838–4813	CTA GAG TCA AGC TCA ACA GGG TCT TC
28S-5368-3'	5392–5368	AAC TAA CCT GTC TCA CGA CGG TCT A
28S-5522-3'	5545–5522	GGA TTC TGR CTT AGA GGC GTT CAG

Appendix table 3. Sequences in collapsed clusters (Fig. 4). Newly obtained sequences in this study are in bold face.

Cluster ID	Sequence ID (GenBank accession number)
Cluster 0001	OM270 (U70723), **NP67-75D-1_10Jul07_30m (HM594229), NP67-75D-2_10Jul07_30m (HM594230), NP67-75D-4_10Jul07_30m (HM594231), NP67-75D-8_10Jul07_30m (HM594233), NP67-75D-9_10Jul07_30m (HM594234), NP67-75D-13_10Jul07_30m (HM594238), NP67-75D-14_10Jul07_30m (HM594239), NP67-75D-16_10Jul07_30m (HM594240), NP67-75D-20_10Jul07_30m (HM594244), NP67-75D-21_10Jul07_30m (HM594245), NP67-75D-22_10Jul07_30m (HM594246)**
Cluster 0002	**Budleigh1_3Nov08 (HM595080), Budleigh2_3Nov08 (HM595081), Budleigh4_3Nov08 (HM595083), FS01E2L1_3Nov08_1m (HM595089), FS01E2L3_3Nov08_1m (HM595091), FS01E2L4_3Nov08_1m (HM595092), FS01E2L6_3Nov08_1m (HM595094), NP67-70D-4_2Oct07_50m (HM594213), NP67-70D-5_2Oct07_50m (HM594214), NP67-70D-8_2Oct07_50m (HM594215), NP67-70D-9_2Oct07_50m (HM594216), NP67-70D-10_2Oct07_50m (HM594217), NP67-70D-12_2Oct07_50m (HM594218), NP67-70D-14_2Oct07_50m (HM594220), NP67-70D-16_2Oct07_50m (HM594221), NP67-70D-17_2Oct07_50m (HM594222), NP67-70D-20_2Oct07_50m (HM594225), NP67-70D-23_2Oct07_50m (HM594227), NP67-75D-5_10Jul07_30m (HM594232), NP67-75D-10_10Jul07_30m (HM594235), NP67-75D-11_10Jul07_30m (HM594236), NP67-75D-12_10Jul07_30m (HM594237), NP67-75D-17_10Jul07_30m (HM594241), NP67-75D-18_10Jul07_30m (HM594242), NP67-75D-19_10Jul07_30m (HM594243), NP67-75D-23_10Jul07_30m (HM594247), NP67-155D3B0A5_6Oct07_86m (HM594191), Sidmouth12_3Nov08 (HM595134)**
Cluster 0006	MC622-32 (EF052198), **Budleigh3_3Nov08 (HM595082), BATSC1_1Jun05_15m (HM595040), BATSC2_1Jun05_15m (HM595041), BATSC3_1Jun05_15m (HM595042), BATSC5_1Jun05_15m (HM595043), BATSC6_1Jun05_15m (HM595044), BATSC8_1Jun05_15m (HM595045), BATSC9_1Jun05_15m (HM595046), BATSC10_1Jun05_15m (HM595047), BATSD11_1Jun05_15m (HM595048), BATSD13_1Jun05_15m (HM595049), BATSD14_1Jun05_15m (HM595050), BATSD16_1Jun05_15m (HM595052), BATSD18_1Jun05_15m (HM595053), BATSD19_1Jun05_15m (HM595054), FS01E2L2_3Nov08_1m (HM595090), FS01E2L5_3Nov08_1m (HM595093), FS01K6_3Nov08_1m (HM595073), FS04L1_4Nov08_75m (HM595074), NP67-70D-1_2Oct07_50m (HM594210), NP67-70D-2_2Oct07_50m (HM594211), NP67-70D-3_2Oct07_50m (HM594212), NP67-70D-13_2Oct07_50m (HM594219), NP67-70D-18_2Oct07_50m (HM594223), NP67-70D-19_2Oct07_50m (HM594224), NP67-70D-22_2Oct07_50m (HM594226), NP67-70D-24_2Oct07_50m (HM594228), NP67-155D3B026_6Oct07_86m (HM594190),**

	NP67-155S-02_6Oct07_5m (HM594192), NP67-155S-03_6Oct07_5m (HM594193) , NP67-155S-04_6Oct07_5m (HM594194), NP67-155S-05_6Oct07_5m (HM594195), NP67-155S-06_6Oct07_5m (HM594196), NP67-155S-07_6Oct07_5m (HM594197), NP67-155S-08_6Oct07_5m (HM594198), NP67-155S-10_6Oct07_5m (HM594199), NP67-155S-12_6Oct07_5m (HM594200), NP67-155S-13_6Oct07_5m (HM594201), NP67-155S-14_6Oct07_5m (HM594202), NP67-155S-16_6Oct07_5m (HM594203), NP67-155S-18_6Oct07_5m (HM594205), NP67-155S-21_6Oct07_5m (HM594206), NP67-155S-23_6Oct07_5m (HM594208), NP67-155S-24_6Oct07_5m (HM594209)
Cluster 0013	**FS01E22_1Aug05_65m (HM595058), FS01E23_1Aug05_65m (HM595059), FS01E24_1Aug05_65m (HM595060), FS01E25_1Aug05_65m (HM595061), FS01E26_1Aug05_65m (HM595062), FS01E27_1Aug05_65m (HM595063), FS01E28_1Aug05_65m (HM595064), FS01F31_1Aug05_65m (HM595065), FS01F32_1Aug05_65m (HM595066), FS01F33_1Aug05_65m (HM595067), FS01F35_1Aug05_65m (HM595068), FS01F37_1Aug05_65m (HM595069), FS01F38_1Aug05_65m (HM595070), FS01F39_1Aug05_65m (HM595071)**
Cluster 0016	**Budleigh5_3Nov08 (HM595084), Budleigh6_3Nov08 (HM595085), Budleigh7_3Nov08 (HM595086), Budleigh8_3Nov08 (HM595087), Budleigh9_3Nov08 (HM595088), Kennick1_Mar09 (HM595095), Kennick5_18Mar09 (HM595098), Kennick7_18Mar09 (HM595100), Kennick9_18Mar09 (HM595102), Kennick10_18Mar09 (HM595103), Kennick11_18Mar09 (HM595104), Kennick13_18Mar09 (HM595105), Kennick14_18Mar09 (HM595106), Kennick15_18Mar09 (HM595107), Kennick18_18Mar09 (HM595109), Kennick19_18Mar09 (HM595110), Kennick20_18Mar09 (HM595111), Kennick21_18Mar09 (HM595112), Kennick3F23_18Mar09 (HM595113), Lyme1_18Mar09 (HM595114), Lyme5_18Mar09 (HM595115), Lyme8_18Mar09 (HM595116), Lyme9_18Mar09 (HM595117), Lyme11_18Mar09 (HM595118), Lyme12_18Mar09 (HM595119), Seaton1_18Mar09 (HM595120), Seaton2_18Mar09 (HM595121), Seaton3_18Mar09 (HM595122), Seaton4_18Mar09 (HM595123), Seaton5_18Mar09 (HM595124), Seaton6_18Mar09 (HM595125), Seaton8_18Mar09 (HM595126), Seaton10_18Mar09 (HM595127), Seaton11_18Mar09 (HM595128), Seaton18_18Mar09 (HM595129), Sidmouth2_3Nov08 (HM595130), Sidmouth3_3Nov08 (HM595131), Sidmouth4_3Nov08 (HM595132), Sidmouth6_3Nov08 (HM595133), Sidmouth15_3Nov08 (HM595136), Tottiford4_27Jan10 (HM595138), Tottiford5_27Jan10 (HM595139), Tottiford6_27Jan10 (HM595140), Tottiford8_27Jan10 (HM595141), Tottiford10_27Jan10 (HM595142), Tottiford12_27Jan10 (HM595143), Tottiford13_27Jan10 (HM595144),**

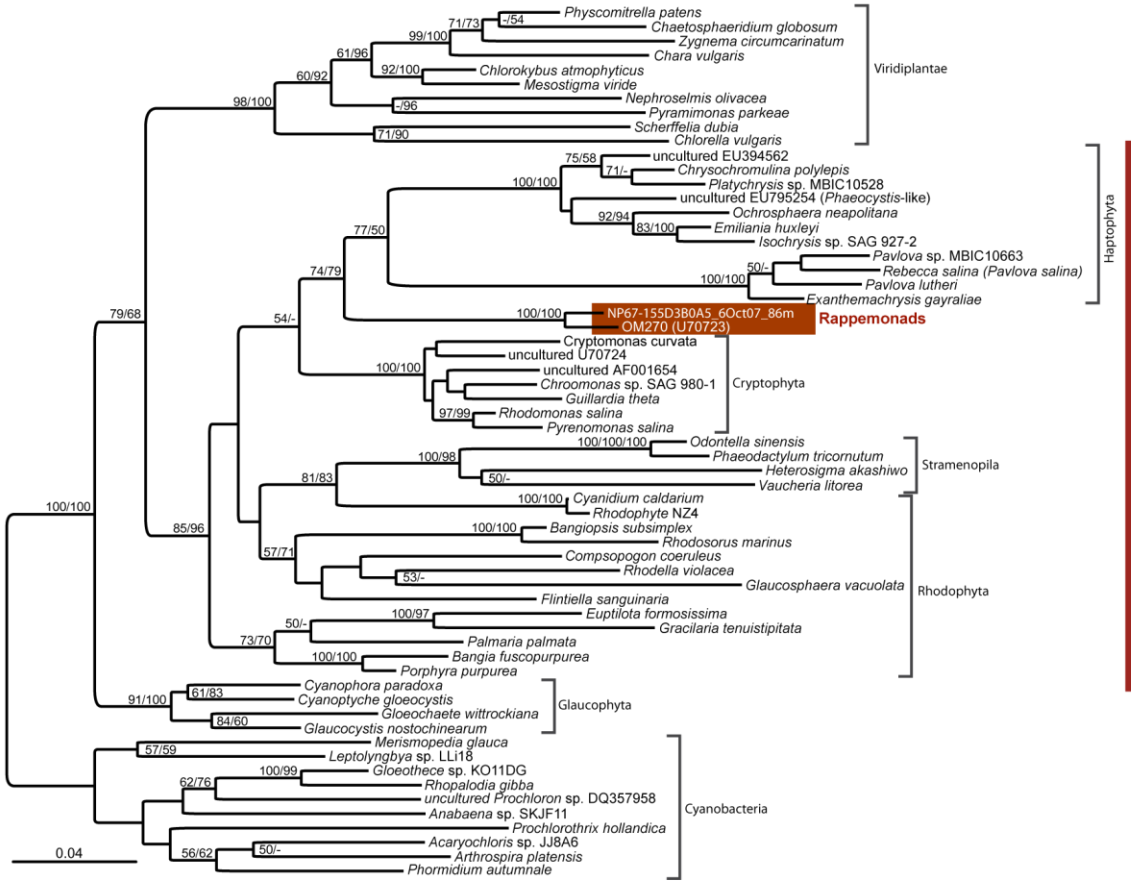
	Tottiford15_27Jan10 (HM595145), Tottiford17_27Jan10 (HM595146), Tottiford18_27Jan10 (HM595147), Tottiford19_27Jan10 (HM595148), Tottiford20_27Jan10 (HM595149), Tottiford24_27Jan10 (HM595150), Trenchford1_27Jan10 (HM595151), Trenchford2_27Jan10 (HM595152), Trenchford4_27Jan10 (HM595154), Trenchford5_27Jan10 (HM595155), Trenchford6_27Jan10 (HM595156), Trenchford7_27Jan10 (HM595157), Trenchford10_27Jan10 (HM595158), Trenchford12_27Jan10 (HM595159), Trenchford14_27Jan10 (HM595160), Trenchford15_27Jan10 (HM595161), Trenchford16_27Jan10 (HM595162), Trenchford17_27Jan10 (HM595163), Trenchford18_27Jan10 (HM595164), Trenchford19_27Jan10 (HM595165), Trenchford20_27Jan10 (HM595166), Trenchford21_27Jan10 (HM595167), Trenchford22_27Jan10 (HM595168), Trenchford23_27Jan10 (HM595169), Trenchford24_27Jan10 (HM595170), Trenchford2N23_27Jan10 (HM595171), Trenchford3P23_27Jan10 (HM595172)
Cluster 0062	**NP67-155S-17_6Oct07_5m (HM594204), NP67-155S-22_6Oct07_5m (HM594207)**
Non-clustered sequences	**BATSD15_1Jun05_15m (HM595051), FS01K3_3Nov08_1m (HM595072), FS04L2_4Nov08_75m (HM595075), FS04L3_4Nov08_75m (HM595076), Sidmouth13_3Nov08 (HM595135)**

Appendix Table 4. Quantitative PCR samples sites and data. Latitude and longitude are given in decimal values. Standard deviations are calculated from technical replicates. Detection limits were calculated for consistent detection of 10 copies well^{-1}. Abbreviations: N, northern; BATS, Bermuda Atlantic Time-series Station; temp, temperature; st dev, standard deviation; n.d., not detected; NO_3+NO_2, nitrate+nitrite; PO_4, phosphate; Chl *a*, chlorophyll *a.*

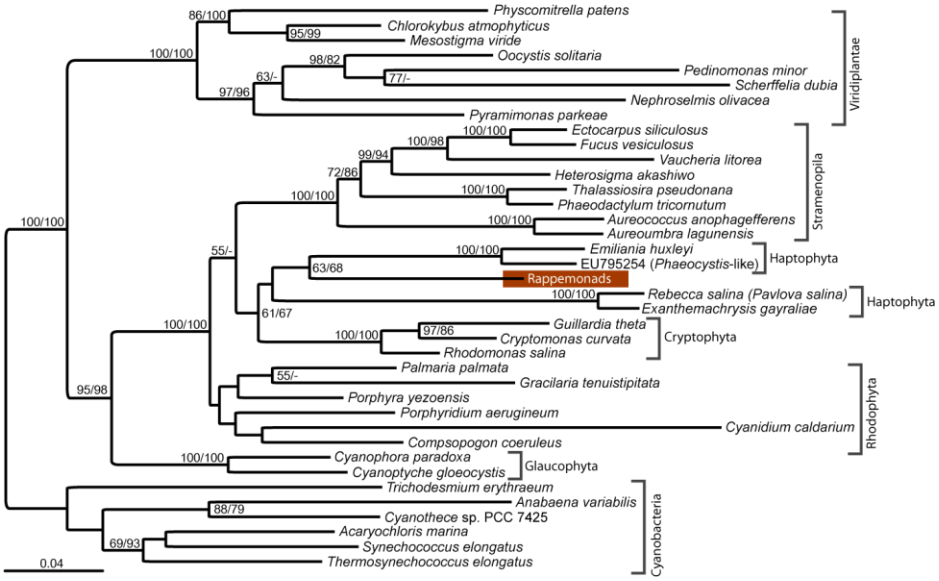
region	date (d/m/yr)	Station	Latitude (N)	Longitude (W)	depth (m)	Temp. (°C)	Salinity (PSU)	copies ml^{-1}	st dev	Detection limit copies ml^{-1})	NO3 + NO2 (uM)	PO4 (uM)	ChlA (ug L^{-1})
N.	02/10/07	67-70	36.126	-123.49	0	16.1	33.05	n.d.		15	31.10	0.045	0.360
Pacific	02/10/07	67-70	36.126	-123.49	50	12.4	33.14	187	42	30	42.40	0.693	0.720
	05/10/07	67-135	33.950	-128.04583	0	18.85	33.02	n.d.		5	23.85	0.065	0.070
	05/10/07	67-135	33.950	-128.04583	80	14.32	33.06	n.d.		5	19.32	0.208	0.160
	05/10/07	67-135	33.950	-128.04583	100	13.23	33.08	n.d.		2.6	15.83	0.234	0.200
	06/10/07	67-155	33.287	-129.42833	0	19.03	33.2	363	25	10	29.03	0.104	0.070
	06/10/07	67-155	33.287	-129.42833	86	13.62	33.12	171	74	15	28.62	0.185	0.270
	08/10/07	Eddy1	34.500	-127.1	0	17.28	32.73	182	46	5.2	22.48	0.162	0.150
	08/10/07	Eddy1	34.500	-127.1	80	12.51	32.86	156	41	5.4	17.91	0.246	0.380
	08/10/07	Eddy2	34.500	-126.5	0	17.37	32.85	97	67	20	37.37	0.139	
	08/10/07	Eddy2	34.500	-126.5	60	11.27	32.68	114	9	20.5	31.77	0.472	0.390
	08/10/07	Eddy3	34.500	-126	0	17.5	33.11	224	99	20	37.50	0.06	0.310
	08/10/07	Eddy3	34.500	-126	40	13.74	33.4	283	262	20.5	34.24	0.581	0.690
	09/10/07	Eddy4	34.500	-125.5	0	17.53	33.33	95	58	28.6	46.13	0.091	0.49
	09/10/07	Eddy4	34.500	-125.5	20	17.48	33.33	n.d.		20	37.48	0.18	0.64
	09/10/07	Eddy5	34.500	-125	0	17.42	32.87	245	132	20	37.42	0.051	0.150
	09/10/07	Eddy5	34.500	-125	60	11.45	32.76	74	97	23.5	34.95	0.465	0.540
	09/10/07	Eddy6	34.500	-124.4	0	17.98	32.92	295	104	20	37.98	0.063	0.110
	09/10/07	Eddy6	34.500	-124.4	70	12.54	32.88	213	52	10.5	23.04	0.362	0.440
N.	14/03/00	BATS	31.583	-64.134	0	19.36	36.65	n.d.		53.2	0.11	0	0.349
Atlantic	14/03/00	BATS	31.583	-64.134	40	19.24	36.64	n.d.		98.1	0.11	0	0.481
	14/03/00	BATS	31.583	-64.134	80	19.09	36.64	n.d.		87.3	0.95	0.04	0.264
	14/03/00	BATS	31.583	-64.134	120	19.08	36.64	n.d.		647.4	1.06	0.04	0.083
	10/07/00	BATS	31.603	-64.182	0	26.51	36.49	n.d.		300	0	0	0.049
	10/07/00	BATS	31.603	-64.182	40	24.26	36.54	n.d.		193.2	0	0	0.048

	10/07/00	BATS	31.603	-64.182	80	22.13	36.72	15	14	56.9	0		0.163
	10/07/00	BATS	31.603	-64.182	120	20.84	36.7	n.d.		110.8	0.77	0	0.237
	20/02/01	BATS	31.667	-64.197	0	19.93	36.66	157	189	64.6	0	0	0.321
	20/02/01	BATS	31.667	-64.197	40	19.89	36.65	n.d.		182.5	0	0	0.406
	20/02/01	BATS	31.667	-64.197	80	19.89	36.65	n.d.		206.9	0	0	0.403
	20/02/01	BATS	31.667	-64.197	120	19.06	36.64	n.d.		392	1.11	0.06	0.033
	06/08/01	BATS	31.660	-64.239	0	27.55	36.53	n.d.		90.2	0	0	0.063
	06/08/01	BATS	31.660	-64.239	40	23.71	36.66	n.d.		107	0	0	0.118
	06/08/01	BATS	31.660	-64.239	80	20.18	36.67	n.d.		106.7	0	0	0.292
	06/08/01	BATS	31.660	-64.239	120	18.99	36.62	n.d.		107	0.42	0	0.259
	21/02/03	BATS	31.682	-64.173	0	20.6	36.73	871	152	2.9	0	0	0.134
	21/02/03	BATS	31.682	-64.173	40	20.24	36.7	431 8	38	2.2	0	0	0.347
	21/02/03	BATS	31.682	-64.173	80	19.89	36.67	9	15	38.4	0.27	0	0.369
	21/02/03	BATS	31.682	-64.173	120	19.85	36.67	n.d.		2.2	0.49	0	0.157
	02/04/03	BATS	31.706	-64.212	0	19.99	36.67	806	265	7.2	0	0	0.278
	03/04/03	BATS	31.706	-64.212	40	19.94	36.67	902	763	8.6	0	0	0.327
	03/04/03	BATS	31.706	-64.212	80	19.54	36.64	104	39	2.3	0.73	0	0.249
	03/04/03	BATS	31.706	-64.212	120	19.33	36.64	n.d.		28	1.38	0.05	0.035
	13/08/03	BATS	31.660	-64.155	0	28.86	36.64	n.d.		17.3	0	0	0.037
	13/08/03	BATS	31.660	-64.155	40	24.52	36.46	335	112	1.2	0	0	0.105
	13/08/03	BATS	31.660	-64.155	80	21.2	36.61	n.d.		4	0	0	0.200
	13/08/03	BATS	31.660	-64.155	120	19.6	36.62	n.d.		8.8	0.72	0	0.381

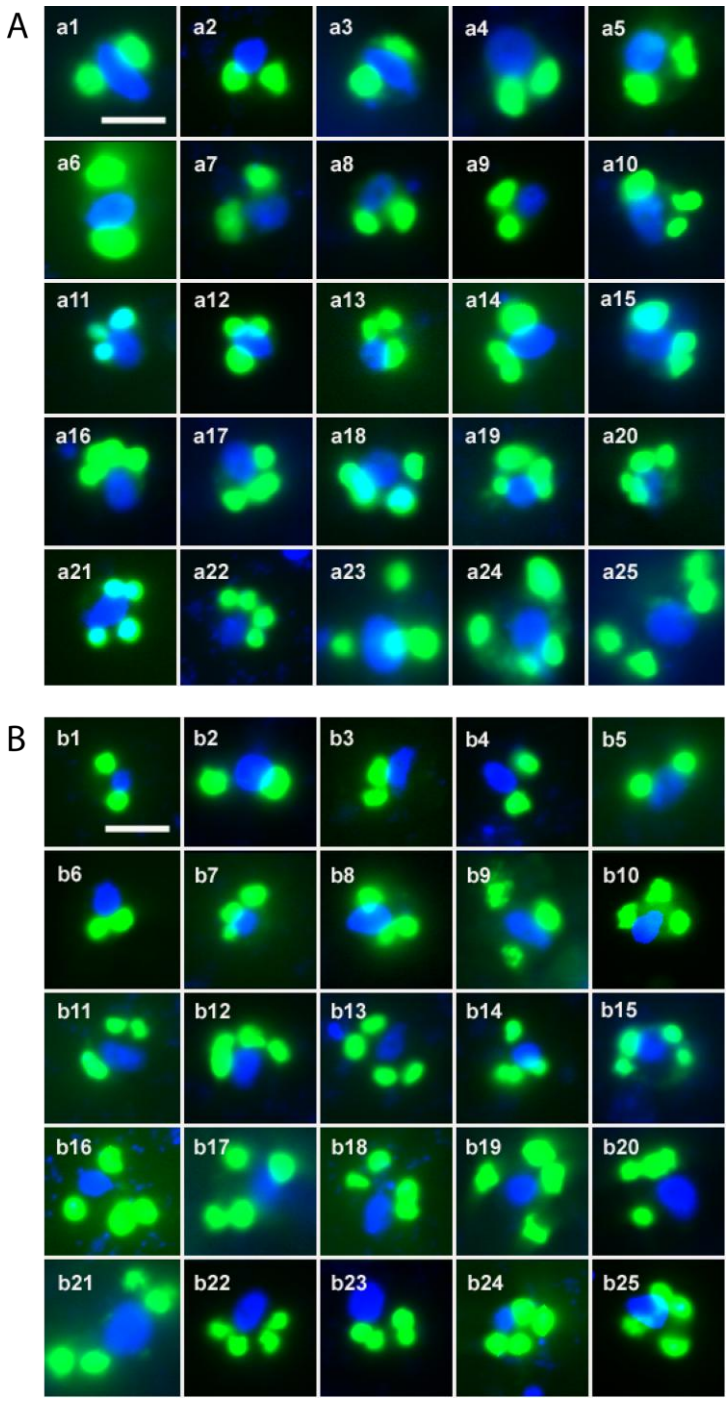
A. 16S rDNA tree (1,447 characters)



B. 23S rDNA tree (2,432 characters)



Appendix Figure 1. Maximum likelihood trees based on 16S rDNA (A) and 23S rDNA (B). Bootstrap values ≥ 50% (ML/Log-Det distance) are indicated. Newly obtained sequences are in bold. The scale bar indicates the inferred number of nucleotide substitutions per site.



Appendix figure 2. TSA-FISH results showing rappemonad cells. Cells shown in A (a1–a25) were identified using the RappeA probe and B (b1-b25) with the RappeB probe. Deposited fluorescein dye is shown in green. DAPI-stained nuclei are shown in blue. All scale bars = 5 µm.