EVOLUTION, DIVERSITY, AND BIOGEOGRAPHY IN PELAGIC CALCIFYING PROTISTS

by

HUI LIU

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Oceanography

written under the direction of

Colomban de Vargas and Costantino Vetriani

and approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2009

ABSTRACT OF THE DISSERTATION

EVOLUTION, DIVERSITY, AND BIOGEOGRAPHY IN PELAGIC CALCIFYING PROTISTS

by HUI LIU

Dissertation Directors:

Colomban de Vargas and Costantino Vetriani

For the last ~200 million years, two groups of unicellular eukaryotes have dominated the biomineralization of carbonate in the oceanic plankton: heterotrophic foraminifera and autotrophic coccolithophores. They literally transformed the fate of inorganic and organic carbon in the Earth's biogeochemical system. The study of the evolution and biodiversity of these marine microcalcifiers has a long and venerable history, largely based on geological records and morphological characters. However, obtaining an accurate estimate of their biodiversity and understanding their evolution using only morphology and fossils is difficult due to issues such as dissolution, convergent morphology, and, for coccolithophores, the complicated haplo-diploid life cycle. Recent advances in molecular biology have further challenged the classic morphological studies by highlighting two additional problems: the unknown diversity of poor and non-calcifying forms in the global ocean and the widespread presence of cryptic species. The goal of this thesis was to reassess the evolutionary and ecological complexity of pelagic microcalcifiers at different taxonomic scales using molecular data constrained by morphological, ecological, and biogeographic metadata. I resolved the mode and tempo of the diversification of the

haptophytes using an extensive multigene analysis and interpreted the timing of four key transitions in the evolution of the haptophytes in an ecological and geological context. I used Haptophyta-specific primers and PCR conditions adapted for GC-rich genomes to circumvent the biases inherent in classical genetic approaches. I discovered for the first time that the tiny (<3 μm) unicellular eukaryotes belonging to the haptophyte lineage are dramatically diverse in the planktonic photic realm, where they appear to dominate photosynthesis. I also developed a combined morphological-genetic approach to survey the environmental diversity of coccolithophores and evaluated the diversity level at which phylospecies and morphospecies can be considered equivalent concepts. Finally, I used the *Neogloboquadrinids*, a family of non-spinose planktonic foraminifera, as a model to assess cryptic speciation and global biogeography in the pelagic microcalcifiers.

Acknowledgements

Many people deserve thanks and appreciation for this thesis. First, I would like to express my deep and sincere gratitude to my co-advisors, Dr. Colomban de Vargas and Dr. Costantino Vetriani. Dr. Colomban de Vargas, who was very brave to take me as his first Ph.D. student, was always eager to help me through the toughest challenges during my studies at Rutgers; Dr. Costantino Vetriani was supportive when Dr. Colomban de Vargas moved to France in early 2006.

I give special thanks to Dr. Stéphane Aris-Brosou for his patience and invaluable time. He has always been there to answer my questions, to proofread and edit my chapters, and to ask good questions to help me think through my problems. Without his encouragement and constant guidance, I could not have finished this dissertation. I am very grateful to Professor Marie-Pierre Aubry, who not only showed me the fantastic morphological structures of coccolithophores, but also was willing to spend a lot of her valuable time working with me. I also thank Dr Oscar Schofield for his thoughtful comments and valuable advice about how to organize my time and structure my thesis.

During my studies, I collaborated with many colleagues for whom I have great regard: Stéphane Aris-Brosou, Ian Probert, Miguel Frada, Fabrice Not, Jeremy Young, Martine Couapel, Thibault de Garidel, Swati Narayan-Yadav, Julia Uitz, Hervé Claustre and Marie-Pierre Aubry. In particular, Stéphane Aris-Brosou helped with molecular dating, interpreting genetic data and provided valuable input for Chapter 2-4; Ian Probert isolated and cultured all haptphyte strains (Chapter 2-4); Julia Uitz and Hervé Claustre developed and applied the statistical model for global pigment analysis (Chapter 3); Miguel Frada performed COD-FISH analyses (Chapter 3); Migule Frada and Swati

Naarayan-Yadav participated in part of the DNA sequence acquisition from Haptophyta culture strains (both Swati and Miguel-Chapter 2-4) and from living planktonic foraminifera (Swati-Chapter 5); Fabrice Not collected environmental DNA samples for Chapter 3; Colomban de Vargas, Jeremy Young and Miguel Frada collected environmental samples for Chapter 4; Jeremy Young and Martine Couapel performed morphological analysis and helped to interpret morphological and genetic complex data for Chapter 4; Thibault de Garidel collected part of living foraminifera samples and provided valuable comments for Chapter 5. I also acknowledge Ramon Massana for sharing pico-DNA extractions for Chpater 3. Chapter 3 is published as a research paper in PNAS, which is written by Colomban de Vargas and Ian Probert.

I am very thankful for the opportunity Marie-Pierre Aubry offered me to participate in research on convergent morphology in coccolithophores, which is not part of this thesis, but greatly enhanced my knowledge and understanding about coccolithophore evolution.

I offer my sincere thanks to the director of the graduate program, Professor Paul Falkowski, for his encouragement and support through my six years at IMCS and his critical comments on Chapter 3. I would like to thank everyone at IMCS who has helped me during my tenure here; in particular, I thank Kay, Yair, Marta, and Charlotte. I also would like to acknowledge NSF, IRES, IRND (uOttawa) and the French ANR BOOM project for financial support. Finally, I give special thanks to Mom and to my husband, Il Kyung, for their continuous support and encouragement, especially during the “final run”.

Table of Contents

Lists of tables

List of illustrations

Chapter 1

1.0. Introduction

Planktonic foraminifera and coccolithophores are calcifying protists that inhabit the photic epipelagic zone of the global ocean, which is one of the largest and ecologically most reactive compartments of the Earth's ecosystem. These organisms constitute a substantial portion of the oceanic phytoplankton (coccolithophores) and zooplankton (planktonic foraminifera) and are key players in the marine food web (Landry 2002; Wade and Darling 2002). Because they secrete a hard $CaCO_3$ skeleton, they are known as the pelagic "microcalcifiers" and are responsible for the bulk of oceanic calcification (Milliman 1993); thus, they greatly impact the marine carbonate system and the global carbon cycle.

The skeletons of these microcalcifiers have built the most complete and continuous fossil record of any kind of organism, and they serve as fundamental tools for the study of rates and patterns of evolution (de Vargas and Probert 2004; Norris and de Vargas 2000), as tracers of paleo-oceanic environments (CLIMAP 1976; McIntyre 1967), and as basic markers in biostratigraphy (Berggren 1995; Bukry 1978). Furthermore, some species of coccolithophores, such as *Emiliania huxleyi*, form massive blooms that enhance their role in the marine carbon cycle (Iglesias-Rodriguez et al. 2002). These blooms produce high amounts of dimethyl sulfide (DMS), a gas responsible for cloud nucleation (Malin 2004), and thus may significantly impact local climates on Earth (Westbroek et al. 1993).

By providing a rich suite of morphological characteristics and a uniquely extensive fossil record, these tiny skeletons present a rare opportunity to study evolution and biodiversity in marine plankton. Conventional studies based on morphology have provided

important information bearing on their mode and tempo of evolution and continue to serve as the basis to understand their biodiversity (e.g., Aubry and Bord 2009, Abury 2009). However, the classical morphological view of the evolution and biodiversity of these two groups is largely oversimplified. In particular, extensive studies have shown the widespread presence of cryptic species of planktonic foraminifera (e.g., Darling et al. 2004; Darling et al.2007; Darling et al. 2002; de Vargas et al. 1999; de Vargas et al. 2001). Detailed biogeographic studies and molecular clock estimates of this cryptic or pseudo-cryptic biodiversity have revealed that the majority of the genotypes may in fact represent reproductively isolated species that have been separated for hundreds of thousands to millions of years in a particular habitat (Darling et al. 2007). This suggests that the current morphological criteria used to define planktonic foraminifera species may be too broad and that each morphospecies may be a cluster containing a few sibling species distinguished by subtle morphological characters. The emerging cryptic or pseudo-cryptic biodiversity is even less known in coccolithophores (Geisen et al. 2004; Saez et al. 2003). A few additional problems such as the dissolution of ~70% of morphotypes in the water column; convergent morphology (Aubry et al, in prep); the haplo-diploid life cycle (Billard 1994; Houdan et al. 2004), during which different types of coccoliths/nannoliths are produced have made it difficult to assess the evolution and diversity of coccolithophores merely based on morphology. Furthermore, the hidden or naked (i.e., non-calcifying) or poorly calcified species are virtually inaccessible to observation-based identification.

Molecular techniques have provided new ways to explore evolutionary processes and diversity patterns in calcifying protists. In particular, molecular phylogenetics has made it possible to reconstruct both macroevolution and microevolution of all organisms

as long as their DNA sequences can be obtained. Additionally, molecular survey techniques and metagenomics have been used extensively to estimate at finer scales the biodiversity present in marine ecosystems (Castle et al. 2006; Diez et al. 2001; Guillou et al. 2004; Lopez-Garcia et al. 2001; Rusch et al. 2007; Venter et al. 2004), where the majority of species are small in size and currently impossible to culture. The new tools, initially used in prokaryotes (Chisholm et al. 1988; Giovannoni et al. 1990; Rappe et al. 1998), were soon extended to oceanic protists, mainly those from the pico-planktonic size fraction (cells <2-3 μm e.g., Diez et al. 2000; Moon-Van Der Staay et al. 2001). Hundreds of previously undocumented rDNA phylotypes[1] were revealed that likely will eventually be characterized as phylospecies [2] (Huber et al. 2007; Queiroz and Donoghue 1988). However, such environmental genomic libraries have not yet been constructed for coccolithophores. The promising molecular view of oceanic protistan phylogeny and biodiversity remains ambiguous because of the limited genetic data obtained from culture sequences and the lack of links to morphological diversity.

With increasing attention being paid to the impact of rising ocean acidification on these microcalcifiers, the importance of resolving fundamental questions about their evolution, biodiversity, and biogeography has arisen. Therefore, in this study, I used a top-down approach, from macro- to microevolution, using combinations of morphological and molecular methods to address several key questions relevant to the general task of understanding the evolution and biodiversity of these two groups of marine microcalcifiers on a global scale: such as, what are the important dates when haptophyta diversified and

[1] the cluster of genomes of asexual organisms forms what is called a "phylotypes.
[2] phylospecies is coined to describe those microorganisms, that according to the phylogenetic species concept, form a monophyletic clade at a fundamental level and that are occupying, living and evolving in a specific niche.

evolved specific ecological adaptations, what is the current diversity of pico- and nanno-sized haptophyta, how important is the discrepancy between phenotypic and genotypic information in coccolithophores? Moreover, how are the global biogeographic distributions of a few key species of planktonic foraminifera? The study brought fundamental data (over 1500 LSU and SSU rDNA sequences of environmental planktonic foraminifera and coccolithophores) to understand the evolution, biodiversity and biogeography of these two important microcalcifiers.

1.1. Molecular evolution and biodiversity in coccolithophores

The coccolithophores belong to the phylum Haptophyta. Haptophytes are microalgae characterized by the presence of a flagellum-like appendage called a "haptonema," which is used for attachment or capturing prey (Inouye and Kawachi 1994). The phylum is divided into two classes: the Pavlovophyceae, which are characterized by the anisokont nature of the flagella and the presence of simple organic knob-like scales covering the cells, and the Prymnesiophyceae, which have flagella of more or less equal length and cells bearing organic plate scales (Edvardsen et al. 2000). Coccolithophores are included in the Prymnesiophyceae, and they comprise all haptophyte cells that can precipitate carbonate calcium onto organic scales. Non-calcifying species belong to the clade that includes the coccolithophores, and a new subclass, the Calcihaptophycidae, was proposed to include all potentially calcifying haptophytes (de Vargas et al. 2007). With one exception (*Hymenomonas roseola*, which resides solely in freshwater), all coccolithophores are marine species.

1.1.1. Macroevolution

The phylum Haptophyta is one of the deepest branching groups in the phylogeny of eukaryotes (Baldauf 2003). Based on a molecular clock estimation (Yoon et al 2004) using multiple chloroplast genes, haptophytes originated 1050–1100 Ma. Biological, phylogenetic, and paleontological data tend to support the scenario that haptophytes evolved from coastal or neritic heterotrophs/mixotrophs to oceanic autotrophs since their origination in the Proterozoic (Bown 1987; Farrimond et al. 1986; Yoon et al. 2004).

Because no fossils exist for the non-calcifying members and the majority of the calcifying members (dissolution and/or poor preservation) of the Haptophyta, reconstructing a comprehensive molecular phylogeny is of crucial importance in understanding the major evolutionary steps that took place in this phylum. Early attempts to reconstruct the molecular phylogeny of the haptophytes using 18S rDNA (Edvardsen et al 2000) and rbcL (Fujiwara et al 2001) were based on a limited number of taxa and certain key coccolithophore orders were not represented at all.

In the macroevolution of the haptophytes, fossil records show that calcification originated about 225 Ma (Bown et al. 2004) although an earlier origins have been supported (see de Vargas et al. 2007). Why calcification occurred is a fundamental question in the evolution of this group. Why have coccoliths evolved? Which evolutionary forces drove their genesis? Did coccolithogenesis occur once or multiple times? Which genetic changes allowed coccolithogenesis? Calcification in coccolithophores appears to be a complex process and it is very poorly understood at present. In a recent review, de Vargas et al (2007) suggested that calcification in coccolithophores is a highly modulated process, which was partially re-invented, shutdown, and then differentially evolved within

the many lineages of coccolithophores. Another important and unresolved question is the evolution of the trophic mode in haptophyta. The haptophyta have been proposed to have a deep origin (Baldauf 2003) and the plastid genome of the haptophyta may have diverged later via secondary endosymbiosis (Falkowski et al, 2004), then haptophyta may have been primarily heterotrophs at the time they evolved. This hypothesis is currently supported by two arguments: (i) haptophytes are characterized by the presence of a flagellar apparatus, the haptonema, which is supposed to act as a hunting device (Aubry 2009; Yoshida et al. 2006); (ii) the early-branching lineages in nuclear molecular phylogenies of extant heterokonts, cryptophytes, and alveolates are systematically represented by heterotrophic (aplastidial) taxa. However, no direct reconstruction has ever been attempted to estimate the ancestral trophic types. Therefore, in the first part of the study (Chapter 2), I used multiple gene dataset to reconstruct and date the key innovations in the evolution of the Haptophyta. In particular, four important discrete traits, i.e. the calcification, trophic mode, transition from coastal to oceanic and the emergence of organic scales were mapped on phylogenies to infer the paleo-ecology and evolution in this group.

1.1.2. Biodiversity in coccolithophores

Currently, ~280 morphological species of coccolithophores are recognized (Young et al. 2003). However, the corresponding genetic diversity data is strikingly scarce due to the difficulties of maintaining most of the diversity in culture. Moreover, a significant portion of coccolithophore biodiversity may lie in the diversity of naked, tiny, or poorly calcified taxa (see Chapter 3), for which the morphology is entirely unknown and in the diversity of cryptic biological species within the classical morphospecies. To date, few studies have addressed cryptic speciation in coccolithophores (Saez and Lozano 2005;

Saez et al. 2003) and no environmental genomic surveys have been reconstructed for coccolithophores.

Molecular survey is an ideal tool to explore the environmental biodiversity of coccolithophores, however two important technical limitations must be considered: the formation of chimeras and the significant taxonomic bias introduced in PCR amplification of eukaryotic genomes. Chimeras are genes made of fragments from the genomes of different species. When performing PCR amplification of rRNA genes on total environmental DNA extracts, chimeric sequences easily form because highly conserved regions of ribosomal genes can anneal, even between sequences from distantly related organisms. Therefore, chimeras represent up to 32% of environmental sequences reported in previous studies (Berney et al. 2004; Hugenholtz and Huber 2003; Robison-Cox et al. 1995; Wang and Wang 1997). While it is relatively easy to detect chimeras made of large fragments from widely divergent species, it is much more difficult to unveil micro-chimeric patterns between related species, genera, or families. In fact, such taxonomically restricted micro-chimeras may significantly—and artificially—increase the diversity of ribotypes amplified from natural populations.

The second problem relates to the differential nature of eukaryotic genomes and rDNAs. Environmental PCR using "universal," "prokaryotic," or "eukaryotic" primers introduces biases in patterns of diversity, principally due to the different secondary structures and nucleotide compositions of the target gene, together with the exponential dynamics of the chemical reaction. This problem, initially revealed in bacteria (Chandler et al. 1997; Webster et al. 2003), may in fact be much worse in eukaryotes. The rDNA genes of eukaryotes vary greatly in length and GC content. In the foraminifera, the SSU rDNA

can be three to five times longer than in any other eukaryote and is thus inaccessible using classical PCR protocols. Indeed, despite their importance in both planktonic and benthic marine biotopes, foraminifera are virtually absent from all environmental surveys of these environments. In coccolithophores, and haptophytes in general, the GC content is very high, and the use of a modified buffer to untie the DNA strands is necessary to perform amplification reactions.

In the third part of the study (Chapter 4), I combined classical morphological and genetic approaches to studying coccolithophore biodiversity. I estimated and compared morphological and genetic diversity from the environment, tested the relationship between the morphological and genetic species concepts, and evaluated the level of diversity at which phylospecies and morphospecies can be considered equivalent concepts.

1.1.3. Marine pico-haptophytes: The tiny, non-calcifying relatives of coccolithophores

Marine pico-eukaryotes (cells < 3 μm) are fundamental and newly revealed components of microbial food webs, and they play an important role in global mineral cycles (Fogg 1995; Moon-van der Staay et al. 2001; Worden et al. 2004; Zhu et al. 2005). However, due to their small size and simple morphology, the groups that dominate in various oceanic settings are poorly known. A recent application of molecular probes coupled to tyramide signal amplification-fluorescent *in situ* hybridization (TSA-FISH) has made quantification of the eukaryotic component of picoplankton possible (Not 2005). In a series of cruises across the Arctic, Atlantic, and Indian Oceans, Not and collaborators recently reported that a significant abundance of pico-haptophytes inhabits the upper region of the euphotic zone (Biegala et al. 2003; Not 2005). These cells account on average

for 6.2% of the pico-eukaryotes in Arctic waters, and they are even more abundant in the Indian and Atlantic Oceans, where they contribute up to 67% and 35%, respectively, of the total number of pico-eukaryotes. Therefore, they are important players in the microbial food web of the world ocean. Eukaryotic molecular surveys based on environmental 18S rDNA clone libraries suggest that these pico-haptophytes include new and sometimes very divergent phylotypes (Diez et al. 2001; Moon-van der Staay et al. 2001). However, these PCR-based studies were largely biased by problems mentioned above, and the identity and diversity of pico-haptophytes needs to be thoroughly analyzed. Therefore, I used an original PCR protocol to amplify the GC-rich genomes of haptophyte algae using haptophyte-specific primers which overcome such bias and discovered a dramatic diversity of novel haptophyta lineages which are responsible for most global light-harvesting in modern oceans according to global estimation of their specific pigment 19-hexanoyloxyfucoxanthin (Chapter 3).

1.2. Planktonic foraminifera

Planktonic foraminifera constitute a group of globally distributed marine protists with calcareous shells (tests), and they are an important part of the marine zooplankton. They first appeared in the fossil record in the mid-Jurassic period and diversified in the Cretaceous. A major evolutionary radiation occurred during the Paleocene, after major extinctions at the Cretaceous/Tertiary boundary. Their fossil record represents the most robust source of biostratigraphic markers and paleo-environmental proxies.

Planktonic foraminifera provide a rare opportunity to study the evolution and diversity of an entire group of marine protists at the morpho-species level, including all fossilized ancestors. Extremely detailed stratigraphic and taxonomic analyses based on

morphological characteristics have been performed, and recent molecular genetic studies have made further progress in understanding the phylogeny of the foraminifera. At present, nine morphological species have been genetically analyzed over relatively large biogeographic ranges, and each was found to consist of three to six genotypes (Darling 2002; de Vargas 2004). Further phylogenetic and biogeographic studies of these genotypes have indicated that they likely represent sister species separated by hundreds of thousands to millions of years of evolution (Darling et al. 2007). However, the mechanism of diversification among these cryptic species is still poorly understood, although several hypotheses have been proposed (Darling et al. 2004; Pawlowski and Holzmann 2002). One of the major problems in understanding diversification is geographic scaling: All morphospecies of planktonic foraminifera occupy worldwide, bi-hemispherical, biogeographic ranges. It is reasonable to assume that each morphospecies can be split into 5 to 10 biological species with more restricted ranges. However, the range of genetic types, even if more restricted, may be widely distributed among the oceans and the distribution might be seasonal. Thus, *worldwide* and *multi-seasonal* sampling is a prerequisite to assess the geographic and ecological range of modern planktonic foraminifera. This approach has been partially conducted for four morphospecies (de Vargas et al. 2004), and each genetic species seems to occupy a different ecological space than their closest relatives.

Another major limitation in studying the foraminifera occurs in the non-spinose planktonic forms. In this group, the different copies of the rDNA clusters are genetically variable within a single individual (single cell). Even the 18S rDNA, a marker often used as a good proxy for biological species (Darling et al. 2006; Darling et al. 2004), is subject to intra-individual variation. In this case, it is necessary to clone several copies of the gene

and first assess the extent of intra-individual genetic variability to re-define species concepts based on rDNA. This time-consuming approach has been applied to only one species (*Globorotalia truncatulinoides* de Vargas et al. 2001), and the phylogenetics of non-spinose planktonic foraminifera clearly needs re-evaluation.

In the final part of the thesis (Chapter 5), I examined both intraindividual and intraspecific SSU rRNA genetic variations in a family of non-spinose planktonic foraminifera (Neogloboquadrinids) and reinterpreted the phylogeny and biogeography of this family.

Chapter2

2.0. A timeline of the environmental genetics of the haptophytes

2.1. Abstract

The use of genomic data and the rise of phylogenomics have radically changed our view of the eukaryotic tree of life at a high taxonomic level by identifying four to six "supergroups". Yet our understanding of the evolution of key innovations within each of these supergroups is limited because of poor species sampling relative to the massive diversity encompassed by each supergroup. Here we apply a multigene approach that incorporates a wide taxonomic diversity to infer the timeline of the emergence of strategic evolutionary transitions in the haptophytes, a group of ecologically and biogeochemically significant marine protists that belong to the Chromalveolata supergroup. Four genes (*SSU*, *LSU*, *tufA* and *rbcL*) were extensively analyzed under several Bayesian models to assess the robustness of the phylogeny, particularly with respect to (i) data partitioning, (ii) the origin of the genes (host *vs*. endosymbiont), (iii) across-site rate variation and (iv) across-lineage rate variation. We show with a relaxed clock analysis that the origin of haptophytes dates back to 824 MYA (95% highest probability density 1031-637 MYA). Our dating results show that the ability to calcify evolved earlier than previously thought, between 329-291 MYA, in the Carboniferous period, and that the transition from mixotrophy to autotrophy occurred during the same time period. Although these two transitions precede a habitat change of major diversities from coastal / neritic waters to the pelagic realm (291-243 MYA, around the P/Tr boundary event), the emergence of calcification, full autotrophy and oceanic lifestyle seem mutually independent.

2.2. Introduction

Eukaryotes are provisionally subdivided into six supergroups (Opisthokonta, Amoebozoa, Archaeplastida, Excavata, Rhizaria, and Chromalveolata) whose phylogenetic relationships are slowly emerging (e.g., Lane and Archibald 2008). The Chromalveolata, one of the six eukaryotic supergroups, comprise a disputed assemblage made of eukaryotes with red algal-derived plastids that originate ultimately from a common secondary endosymbiosis (Yoon et al. 2004). This potentially paraphyletic or even polyphyletic supergroup is composed of the alveolates (dinoflagellates, apicomplexans and ciliates) and the chromists (stramenopiles, cryptophytes and haptophytes), and accounts for about half of the described diversity of protists (Cavalier-Smith 2004). Recent studies found that four of these six lineages (apicomplexans, ciliates, dinoflagellates, stramenopiles) consistently form a monophyletic assemblage, whereas the remaining two lineages (cryptophytes and haptophytes) form a weakly supported group that remains to be substantiated (e.g., Hackett et al. 2007; Hampl et al. 2009; Harper et al. 2005); but see Patron et al. 2007; Rice and Palmer 2006). These studies provide important insights into the basal relationships between these lineages, but they do not have the taxonomic coverage that would allow us to infer the emergence of key evolutionary transitions within each lineage, in particularly within the haptophytes.

The present study focuses on the haptophytes, one of the most abundant groups of oceanic phytoplankton and significant primary producers (Field et al. 1998; Thomsen et al. 1994). Haptophytes, or Haptophyta, differ from other eukaryotes by possessing a unique flagellum-like organelle, the haptonema, that is thought to play a role in prey capture in some species (Kawachi et al. 1991). Another unique feature found in the coccolithophores

or Calcihaptophycideae, the best-known members of this division, is the presence of a calcified exoskeleton consisting of minute, intracellularly formed, calcite platelets (coccoliths) that sediment to the ocean floor upon death of cells, resulting in the formation of limestone and chalk deposits over geological time. Based on morphology, the division Haptophyta is classically subdivided into two classes: the Pavlovophyceae, asymmetrical cells covered by organic knob-like scales and with anisokont (unequal length) flagella, and the Prymnesiophyceae, symmetrical cells covered by organic plate scales (that serve as the matrix for calcification in the coccolithophores) and with isokont flagella. The coccolithophores are presently responsible for the bulk of oceanic calcification (Milliman 1993). Consequently, they heavily influence the marine carbonate system and have a major impact on the global carbon cycle. The fossil archive of the coccolithophores is probably the most complete of any protist lineage, with 20-30% of species leaving a fossil record (Young et al. 2005), and this archive has been intensively studied by biostratigraphers (e.g., Bown 1998). Certain haptophytes, such as members of the genera *Emiliania*, *Gephyrocapsa*, *Phaeocystis*, *Chrysochromulina* and *Prymnesium*, are responsible for extensive blooms that have major biogeochemical, ecological and economic impacts (Brown and Yodar 1994; Edvardsen and Paasche 1998; Lancelot et al. 1998; Robertson et al. 1994). For example, massive blooms of the coccolithophore *Emiliania huxleyi* are thought to affect global climate by increasing water albedo through dimethylsulphide production, and also drive large fluxes of calcium carbonate out of surface waters (Tyrrel and Merico 2004). As focus is increasingly falling on the impacts of rising anthropogenic CO_2 on the carbonate system in the ocean, a better understanding of the diversification of the haptophytes and how this diversification has correlated with past environmental

conditions may help predict how these species will react to future environmental change (Fabry 2008). However, despite their ecological, biogeochemical and geological role, our knowledge of the diversification of this division is still limited to what is known from the coccolithophores, the only members of the haptophytes that leave traces in the fossil record; yet, coccolithophores represent less than half of the existing diversity of the haptophytes (Young et al. 2005).

The molecular studies that pioneered the reconstruction of the diversification of the haptophytes used either a single slowly-evolving nuclear gene such as the 18S rDNA (*SSU*: Edvardsen et al. 2000; Medlin et al. 1997; Simon et al. 1997), or faster evolving plastid genes such as *rbcL* (Daugbjerg and Andersen 1997; Fujiwara et al. 1994; Fujiwara et al. 2001; Inouye 1997) or *tufA* (Saez et al. 2003). These early studies supported the morphological taxonomy by dividing haptophytes into two main clades: the Prymnesiophyceae and the Pavlovophyceae. Yet, phylogenetic resolution beyond this taxonomic level was still limited. In combination with morphological, physiological and ecological data, more recent molecular approaches further recognized four major clades (Prymnesiales, Coccosphaerales, Isochrysidales and Phaeocystales) within the Prymnesiophyceae (Edvardsen et al. 2000). However, the resolution of these molecular studies remained poor, particularly within the coccolithophore clade. The most comprehensive molecular phylogenetic reconstructions of the Haptophyta to date are those of (Medlin et al. 2008) using sequences of the nuclear *SSU* and plastid *tufA* genes from *ca.* 60 cultured species.

With increasing molecular phylogenetic resolution and an outstanding fossil record for the past 220 million years (Bown 1998), the haptophytes are an ideal group for applying

molecular clock methods to date key transitions and unravel the tempo of their evolution. The inadequacy of the strict molecular clock is no longer controversial for most modern datasets, but known limitations can be alleviated by meeting four general conditions (Soltis et al. 2002; Yoon et al. 2004); (i) use of a well-supported and accurate tree that resolves all important nodes (normally entailing the use of large multigene datasets); (ii) use of reliable fossil calibrations; (iii) use of methods that account for substitution rate heterogeneity within and across lineages; and (iv) broad taxon sampling. An early strict molecular clock study based on an *SSU* phylogeny estimated that the haptophytes diverged from other chromists between 1750 and 850 million years ago (MYA; Medlin et al. 1997). Two subsequent studies that did not assume a strict molecular clock, one based on six plastid genes (Yoon et al. 2004) and the other one on a single ribosomal gene *SSU*; Berney and Pawlowski 2006), narrowed down the previous estimate to ~1100-900 MYA. A more recent molecular study based on two genes, *SSU* and *tufA*, did include more representatives of the haptophytes (Medlin et al. 2008), and dated the origin of the Haptophyta at *ca.* 1200 MYA. However, this latter study (i) did not discuss the use of multiple gene partitions to estimate the tree used for dating, (ii) assumed that the two genomes, nuclear and plastid, share the same history, (iii) assumed that this phylogeny is known with an absolute certainty in order to estimate divergence times, and (iv) was still based on a strict molecular clock that limited the analysis to the only gene (*SSU*) following approximately this strict clock hypothesis. Apart from the dating controversy, it was also suggested that extant coccolithophores diversified from a few lineages that survived the major extinction at the Cretaceous/Tertiary (K/T) boundary, whereas non-calcifying haptophytes were not affected by the K/T extinction (Medlin et al. 2008). The adaptation of non-calcifying

haptophytes to eutrophic coastal environments and their ability to switch nutrition modes from autotrophy (photosynthesis only) to mixotrophy (photosynthesis + particle grazing, which requires some phagocytic ability) were posited as possible explanations for their survival during this abrupt global change event. Such a parsimony-driven reconstruction of character states from their observed distribution in contemporary organisms highlights the possibility of using ancestral reconstructions to glimpse the past by discovering how non-fossilizable traits evolved. Statistically robust computational methods are available to reconstruct ancestral characters or states, even in the presence of uncertainty in estimates of the tree and its branch lengths (e.g., Pagel et al. 2004).

Here we resolve the mode and tempo of the diversification of the haptophytes using an extensive multigene analysis that includes both nuclear (*SSU*, *LSU*) and plastid (*tufA*, *rbcL*) gene sequences for a total of 5006 base pairs. Our species sampling includes 34 representative taxa from the Pavlovophyceae and the Prymnesiophyceae, the latter including members of all formally described extant orders. Our analyses show that (i) the haptophytes evolved *ca.* 824 MYA (1031-637 MYA), (ii) the nuclear and plastid genomes share the same history within the haptophytes and (iii) the reconstruction of this history is not plagued by artifacts such as long-branch attraction due to general model misspecification. Moreover, we reconstruct and date four key transitions: the evolution of calcification and organic scales, and the switches from coastal to oceanic dwelling as well as from mixotrophic to autotrophic nutrition mode. The timing of these key evolutionary transitions is interpreted in an ecological and geological context.

2.3. Material and Methods

2.3.1. Taxonomic sampling and culture conditions

About 430 clonal culture strains of haptophytes were isolated and maintained as described in (Probert and Houdan 2004). The majority of these strains are available from the Roscoff Culture Collection (RCC: http://www.sb-roscoff.fr/Phyto/RCC/). Taxonomic identification of cultures was based on Transmission Electron Microscopy (TEM) observation of body scale morphology for non-mineralized taxa, and Scanning Electron Microscopy (SEM) observation of coccolith morphology for mineralized taxa. Taxonomic concepts used here follow those of Young et al. 2003) and Jordan et al. 2004). Partial *LSU* sequences for *ca.* 300 strains were obtained over the course of this study (see Table 2.1 for a list of all sequenced strains). We included four gene sequences (*SSU*, *LSU*, *tufA* and *rbcL*) from each of thirty-four species of haptophyte and six non-haptophyte taxa in our analysis (Table 2.2). Species sampling within the haptophytes included 2-3 representatives of all genera available from the RCC and for each species, sequences of at least three of the four genes included in the analysis (*SSU*, *LSU*, and *tufA*) originated from the exact same culture strain. Our choice of non-haptophyte taxa to root the tree was guided by the availability of the four gene sequences in GenBank. When this data set was assembled (Oct. 2007), the closest outgroup sequences found by BLASTn searches were from six Stramenopiles (Table 2.2).

2.3.2. *LSU* gene sequencing

Exponential phase cultures were harvested by centrifugation (1000 rpm. for 5 minutes) and 100µl of GITC* DNA extraction buffer (4M guanidine thiocyanate, 50mM Tris-HCl (pH

7.6), 2% N-Lauroyl-sarcosine, 0.1M β-mercaptoethanol) were added to the cell pellet. Cells in buffer were stored at -20°C until analysis. Total DNA was extracted using the DNeasy Plant MiniKit following the instructions from the manufacturer. A nuclear *LSU* rDNA fragment of 941 bp containing the D1 and D2 domains was PCR amplified using a set of eukaryotic-primers in forward: Leuk2 (5'- acccgctgaacttaagcatatcact -3') and in reverse: Euk_34r (5'-gcatcgccagttctgcttacc-3'). PCRs were performed using REDTaq DNA polymerase (Sigma-Aldrich) and a PCRx enhancer system (Invitrogen) in order to amplify GC-rich haptophyte sequences. The reaction followed denaturation at 94°C for 30 seconds, annealing at 55°C for 30 seconds and extension at 68°C for two minutes. Thirty-five cycles were performed with initial denaturation and final extension steps. PCR products were purified using Qiaquick PCR purification kit (Qiagen) and then sequenced in both directions using a 3100-Avant Genetic Analyzer. All sequences obtained in this study were deposited in GenBank (see Table 2.2 for accession numbers).

2.3.3. Computational analyses

The four genes, *SSU*, *LSU*, *rbcL* and *tufA*, were aligned individually with `Clustal` ver. 1.83 (Thompson et al. 1997). Alignments were visually inspected and edited where necessary with the Genetic Data Environment ver. 2.2 software (Larsen et al. 1993). Two sets of alignments were analyzed: a "complete alignment" and an alignment where ambiguous regions were removed (*LSU*: positions 344-434, 514-612, 725-777 and 1015-1026; *SSU*: positions 1204-1238, 1270-1295 and 2612-2620). Both alignments are available upon request.

Phylogenetic analyses were based on several Bayesian approaches in order to test the robustness of our results to a number of assumptions. First, the four genes were

concatenated into one single partition that was analyzed under GTR + Γ_4 + I, as selected by `ModelTest` (Posada and Crandall 1998) based on the Akaike Information Criterion. `BEAST` (Drummond and Rambaut 2007), which permits the joint estimation of tree topology and divergence times, was employed. Uncertainty in the mean substitution rate was integrated out along the MCMC samplers. Speciation times were assumed to follow a pure birth (Yule) process, and rates of evolution were assumed to follow an uncorrelated lognormal process (Drummond et al. 2006). Calibration constraints (CCs) were set as minimum divergence ages, represented by the offsets of the exponential prior distributions (Table 2.3). These included five fossil dates based on nannofossil biostratigraphy (e.g. Bown et al. 2004; Perch-Nielsen 1985; Young 1998); see details below) and one additional weak constraint from a previous molecular clock estimate, the divergence of the two haptophyte classes (node 47 of Figure 2.1; estimated to be well > 350 MYA by both Berney and Pawlowski 2006, and Medlin et al. 2008) in order to test whether our estimations were biased by the use of relatively young (< 220 MYA) fossil constraints. To assess the robustness of our results with respect to the type and number of constraint, three models were run: with four, five or six CCs (Table 2.3). The five CC analysis excluded the molecular clock-based constraint at node 47, while the four CC analysis also excluded the sole character-based constraint (node 57 – see 'fossil constraints' below). For each model, four independent MCMC samplers were run. Each sampler was run for 25 million steps with 2000 steps of thinning. Convergence was checked with `Tracer`, which was also used to compute marginal probabilities of the data. The initial two million steps were removed as a burn-in and results from all four runs were merged with an in-house script that

removes burn-in periods and uses BEAST's treeannotator to summarize the results. The final results were analyzed with R (http://cran.r-project.org/).

Second, to assess the impact of concatenating genes that evolve at different rates, we performed two sets of analyses: (i) the four genes were concatenated or (ii) the data were partitioned according to the four sampled genes. Under this latter partitioning scheme, the two protein-coding genes, *rbcL* and *tufA*, were further partitioned across the three coding positions, so that in total eight partitions were considered. Partitions only shared the tree topology, all the other parameters being independent or "unlinked". Here again, the most appropriate model of evolution was selected with ModelTest (Posada and Crandall 1998) based on the Akaike Information Criterion. The resulting model, GTR + Γ_4 + I, was used with MrBayes ver. 3.1.2 (Ronquist and Huelsenbeck 2003). Each Markov chain Monte Carlo (MCMC) sampler was run for five million steps; autocorrelation was decreased by sampling every 1000 step (thinning); mixing was improved by using tempering with three heated chains. Two independent such samplers were run to check convergence under each model of evolution (with or without partition); at stationarity, split frequencies were checked to be < .015. The first two million steps were discarded as burn-in. Trees were compared with the SH test (Shimodaira and Hasegawa 1999) as implemented in PAML 4 (Yang 2007) and by estimating marginal probabilities as in (Suchard et al. 2001) with Tracer (http://tree. bio.ed.ac.uk/ software/ tracer/). For this partition test, eight partitions were assumed and the GTR + Γ_5 nucleotide substitution model was used with all parameters unlinked.

Third, the robustness to long-branch attraction (LBA) artifacts was assessed by successively removing the taxa that showed the longest root-to-tip branch lengths as in

(Brinkmann et al. 2005; Hampl et al. 2009), and rerunning the MrBayes and BEAST analyses as described above. The MrBayes analyses can be construed as "unconstrained", in the sense that the time-dependency of the evolutionary process is not taken into account; on the other hand, the BEAST analyses directly incorporate the time-dependency of the evolutionary process. For all these analyses, we tracked stability in terms of posterior probabilities of five groups, the Stramenopiles (outgroup), the Pavlovales, the Phaeocystales (one species), the Prymnesiales and the coccolithophores, as a function of the number of taxa removed.

Fourth, we tested if both nuclear and plastid genes reconstructed the same phylogeny, as an analysis of deep divergences based on both nuclear and plastid genes might be affected by endosymbiotic events. For assessing this potential effect, we ran two separate analyses with MrBayes. The first included the two nuclear RNA genes with two unlinked partitions. The second analysis included the two protein-coding plastid genes with six unlinked partitions (three codon positions for each gene). For this comparison of nuclear *vs*. plastid trees, species whose plastid genes were not included in our data were removed from the nuclear tree with APE (Paradis 2006).

Finally, we tested for the potential effect of saturation due, on the one hand, to multiple substitutions at highly exchangeable nucleotides and, on the other hand, to variation of the rate of evolution in time. These two substitution processes can be responsible for incorrect phylogenetic reconstructions (Lartillot et al. 2007) due to the long-branch attraction artifact (Felsenstein 1978). The CAT-GTR model (Lartillot and Philippe 2004), abbreviated as CAT here and implemented in PhyloBayes (ver. 2.3c), accounts for spatial variation of substitution rates (across sites). It was used here to assess the potential

impact of across-site rate variation on the reconstructed phylogenetic trees. The CAT-BP model (Blanquart and Lartillot 2008) implemented in `nh_PhyloBayes` accounts for both the spatial (across sites) and the temporal variation of substitution rates (across lineages). It was used here to test for the effect of rate variation in time (BP model) or in space and time (CAT-BP model). The four genes were concatenated into one single partition. Two independent MCMC samplers were run for 10^5 steps under each model, and split frequencies were checked to be < .015 at stationarity.

Ancestral characters and paleo-environments were reconstructed by maximum likelihood with the `R` package `APE` (Pagel et al. 2004; Paradis 2006). All analyses are based on the consensus tree estimated under the model described above and implemented in `BEAST` (see Figure 2.2 for support values). The outgroup species (Stramenopiles) were removed from the `BEAST` consensus tree with `APE`. Characters and environmental features were assumed to follow a model where all rates of change are different.

2.3.4. Fossil constraints

Two approaches can be used to place temporal constraints on the internal nodes of a tree, using either *character-based* constraints or *divergence-based* constraints (Medlin et al. 2008). Character-based constraints refer to the first occurrence (FO) in the fossil record of a shared derived character or synapomorphy; divergence-based constraints refer to the FO of an ancestor from which descendants within a clade evolved. Of the five fossil constraints used in this study, the oldest (node 57 of Figure2.1) was a *character-based* constraint for the FO of heterococcoliths (i.e. coccoliths consisting of cycles of interlocking crystal units produced during the diploid phase of many coccolithophores). Heterococcolith calcification, a highly distinctive character (Young et al. 1999), is present in the entire

coccolithophore clade in our tree (with a secondary loss in the Isochrysidales). In the fossil record, the first heterococcoliths occur in the Norian stage of the Late Triassic, *ca.* 220 MYA (Bown 1998).

The four other constraints employed were *divergence-based*. From the fossil record alone, a number of uncertainties persist as to the phylogenetic relationships between the Syracosphaeraceae (represented in our tree by *Syracosphaera pulchra* and *Coronosphaera mediterranea*) and other members of the order Syracosphaerales, and between this order and the Zygodiscales (Pontosphaeraceae and Helicsphaeraceae; see (Bown 2005a). Molecular phylogenies tend to indicate a more recent link between the Syracosphaeraceae and the Zygodiscales (node 77 of Figure 2.1) than can be confidently inferred from stratigraphic studies. *S. pulchra* is used as a default identification for larger fossil *Syracosphaera* coccoliths, so that we adopted a conservatively young date for this node by setting it at *ca.* 55 MYA.

We followed (Saez et al. 2003) in dating the divergence of *Umbilicosphaera* and *Calcidiscus* (node 63 of Figure 2.1) at 24 MYA, and set the divergence between *Coccolithus pelagicus* and *Umbilicosphaera hulburtiana* (node 62 of Figure 2.1) to 65 MYA. However, (Medlin et al. 2008) suggested to use 65 MYA for the divergence of *Coccolithus* and *Cruciplacolithus* (node 61 of Figure 2.1). We therefore ran a second set of analyses setting node 61, instead of node 62, to 65 MYA.

Coccoliths assigned to the Pontosphaeraceae (including *Scyphosphaera*) occur down to the late Paleocene, *ca.* 55 MYA (Bown 2005a). Medlin et al. (2008) dated the divergence of the Helicosphaeraceae from the Pontosphaeraceae (node 79 of Figure 2.1) at 50 MYA on the basis of interpreting the fossil record of *Helicosphaera* as being continuous down to

this date in the early Eocene. However, Aubry et al. (*in prep*) postulated that the morphological similarity of coccoliths of *H. carteri*, which has a FO *ca.* 25 MYA, with older species assigned to the Helicosphaeraceae is a result of convergent evolution. In light of this uncertainty, we adopted the younger FO of *H. carteri* (25 MYA) as the calibration constraint of this divergence. To assess the impact of this choice on date estimates, we also run an additional set of analyses constraining node 79 to 50 MYA, the older FO of the Helicosphaeraceae (50 MYA).

2.4. Results

2.4.1. Time are robust to alignments, calibration constraints and data partitions

The divergence times of the haptophytes were estimated assuming one single partition under the time-homogeneous GTR + Γ_4 + I substitution model. Note that with our approach, implemented in `BEAST`, the phylogeny and the divergence times are jointly estimated.

The resulting phylogeny estimated from the complete alignment is showed in Figure 2.2. Most of the nodes are highly supported, with almost all clade posterior probabilities (PP) ≥ 0.80 and the vast majority ≥ 0.95. The long branches around the root indicate that early divergences have likely been lost to extinction or not sampled. All order-level groups of taxa according to current taxonomy were resolved in this phylogeny, with the early divergence within the Calcihaptophycideae of the orders Isochrysidales and Syracosphaerales receiving the weakest support (PP = 0.85). Of the cases where two species of the same genus were included in the analysis, only *Hymenomonas* proved to be paraphyletic. All nodes used for calibrating the tree with dates from the fossil record were highly supported. The analysis of the alternative alignment without the ambiguous regions

resulted in a topology where the only difference was the position of Isochrysidales, which branched with a very low support (PP = 0.63) at a basal position right after the divergence of the Pavlovales. As the Isochrysidales belong to the coccolithophores, the placement of this clade is likely the result of a long-branch attraction artifact with the alternative alignment (see section below). In spite of this topological discrepancy between the complete and the alternative alignment, the estimated divergence times are essentially the same irrespective of the alignment used (Fig. 2.3 A). Indeed, the overlapping 95% credibility intervals with the first diagonal (Fig. 2.3 B) suggest that the differences are not significant, except for node 73 that represents the most basal divergence of the Isochrysidales. Because the dating results are robust to the alignment choice, the complete alignment is used throughout the rest of the text.

The influence of the calibration constraints (CCs) appears to be minimal on time estimates (Table 2.4-2.5) as all three series of mean posterior estimates, with four five or six CCs, are highly correlated ($\rho > 0.997$) and, more significantly, marginal probabilities are all with one log-likelihood unit (Table 2.4). Besides, marginal log-likelihood values indicate that, while the model with five CCs was the most likely, the difference with the four and six CC models is not significant (see Table 2.4). Note that our use of relatively vague priors (Table 2.3) ensures that our results are robust to the disputed use of calibration of node 79 at 25 MYA instead of 50 MYA (Tables 2.6 and 2.7), as well as to the potential misidentification of node 62 for node 61 (Tables 2.8 and 2.9).

Because Figure 2.4 C suggests a potential issue with *Undaria pinnatifida*, whose long branch could indicate a misaligned part of the *SSU* gene (this sequence is actually corrupt and consists essentially of *ITS* and *LSU* sequence), we firstly removed this taxon from our

data set and reran the analyses as in Table 2.3 (4, 5 and 6 CCs); secondly, we also moved the CC at node 61 to node 62 (again, using a total of 4, 5 and 6 CCs). The results show very little difference between the estimated dates when *Undaria pinnatifida* is removed from the analysis, be it for the model as in Table 2.3 (Fig. 2.5 C; Tables 2.10-2.11) or when node 62 is misidentified for node 61 (Fig. 2.5 D; Tables 2.12-2.13). In spite of the robustness of our date estimates to the potential misalignment of *Undaria pinnatifida*, we note that deep divergences would be potentially overerestimated if *SSU* and *LSU* were analyzed on their own (Fig. 2.5 E), although the 95% credibility intervals are so large (Fig. 2.6 E) that these differences are rarely significant. Since all the dating results are robust to (i) the CCs employed and (ii) the inclusion of *Undaria pinnatifida*, the results with five CCs (Fig. 2.1) as specified in Table 2.3 are those that are used in the rest of this study.

From this analysis, haptophytes were estimated to have diverged from the other eukaryotes included in the analysis 824 MYA (95% highest probability density: 1017-640 – see Table 2.5) in the mid-Neoproterozoic Cryogenian period. The divergence of the two extant haptophyte classes is estimated to have occurred 543 MYA (823-328) in the early Cambrian. The divergences between the four taxa within the Pavlovophyceae, including representatives of each of the three extant groups within this class defined by Van Lenning. K et al. (2003), are all estimated to be relatively ancient, occurring in the Mesozoic between 230 and 103 MYA. Within the Prymnesiophyceae, the two non-calcifying orders are estimated to have diverged prior to the Mesozoic, the Phaeocystales at *ca.* 329 MYA and then the Prymnesiales at *ca.* 291 MYA, both in the Carboniferous period. According to our estimates, the primary divergence within the Calcihaptophycideae occurred at 243 MYA, around the Permian/Triassic (P/Tr) boundary event. Molecular divergence within

both the Prymnesiales and Calcihaptophycideae is estimated to have occurred throughout the Jurrassic, Cretaceous and Tertiary periods, with ten calcihaptophyte lineages (from the 24 species in this clade as included in our analysis) crossing the K/T boundary, representing a much weaker signal of divergence occurring predominately after the K/T boundary than reported by Medlin et al. (2008).

Because the model used above makes the simplifying assumption that the data can be concatenated, we need to test that our results are not sensitive to the data partitioning scheme or affected by LBA artifacts.

2.4.2. Robustness of the phylogeny to data partitioning

Our first simplifying assumption was that we could combine all four genes into one single partition without affecting the estimated tree. The two partitioning strategies compared were: (i) one single partition where the two RNA genes and the two protein-coding genes were concatenated, and (ii) one partition for each RNA gene plus one partition for each codon position of each protein-coding gene, which amounts to a parameter-rich model with a total of eight partitions. The substitution model selected by `ModelTest` with the Akaike Information Criterion was GTR + Γ_4 + I for all data sets.

Figure 2.4 A-B shows the trees obtained under these two partitioning schemes. In both cases, the Pavlovophyceae were resolved and branched off first. Within the Prymnesiophyceae, intermediate branching orders (Phaeocystales / Prymnesiales / Calcihaptophycideae) were identical, even though these were the least well-supported nodes of each tree (PP slightly less than 0.80 or $\in$ [0.80,0.90]). Note also that in both partitioning schemes, the genus *Hymenomonas* is paraphyletic with high PP (= 1). The main differences between the two trees (Fig. 2.4 A-B) occurs within the coccolithophores,

notably in resolving the early branching between the Isochrysidales and Syracosphaerales as well as in the exact position of two (out of twenty-four) taxa, *Algirosphaera robusta* and to a lesser extent *Cruciplacolithus neohelis*. Note that this "`MrBayes` tree" with one single partition is not significantly different from the "`BEAST` tree" estimated above (SH: p = 0.208).

The initial motivation for partitioning was to account for the fact that some of the genes or partitions evolve much faster than others. Indeed, the relative rates of the different partitions, as estimated by maximum likelihood with `PAML`, are: 1.00 (*LSU*); 0.37 (*SSU*); 0.25 (*rbcL*, codon position 1); 0.07 (*rbcL*, codon position 2); 1.54 (*rbcL*, codon position 3); 0.48 (*tufA*, codon position 1); 0.17 (*tufA*, codon position 2); 119.44 (*tufA*, codon position 3). The third codon positions are therefore likely to be noisy. In spite of all these differences and potential issues about noise, the two trees are not significantly different at the 1% level (SH: p = 0.196), which suggests that the data can be analyzed under the simplest model that contains one single partition without significantly affecting the reconstructed tree. However, the more appropriate computation of marginal probabilities, *m*, suggests that the more complex model (m(8 partitions) = -39,168) outperforms the simpler model (m(1 partition) = -40,553). In spite of the inaccuracy of these *m* estimates (Lartillot and Philippe 2006), these latter results suggest that a stability analysis of the trees and of the estimated divergence times is required.

2.4.3. Robustness of the phylogeny to long-branch attraction

To assess the stability of the reconstructed tree (Fig. 2.1), in particular with respect to long-branch attraction (LBA) artifacts caused by model misspecification, we successively

removed the taxa that showed the longest root-to-tip branch lengths (Brinkmann et al. 2005; Hampl et al. 2009). Two approaches were used. In the first approach, tree topologies were completely unconstrained in the sense that the time-dependency of the evolutionary process was not taken into account. With this approach, estimated PPs for the Stramenopiles, the Pavlovales and the Prymnesiales were unaffected and consistently close to 1 (Fig. 2.7 A), suggesting that LBA is not an issue for these groups. On the other hand, progressive taxon removal changed the position of Phaeocystales from being sister to the Prymnesiales and the coccolithophores, to being sister to the Prymnesiales with high PP, while support for the coccolithophores collapsed completely (Fig. 2.7 A) due to the position of Isochrysidales. This suggests that the position of Phaeocystales as sister to Prymnesiales and coccolithophores, as in Figure 2.2, is probably the result of an LBA artifact. However, the effect of LBA on the non-monophyly of coccolithophores is quite intriguing, as their monophyly has repeatedly been supported by previous studies (e.g., de Vargas et al. 2007; Edvardsen et al. 2000; Fujiwara et al. 2001; Medlin et al. 2008).

As noted above, one very general cause of LBA is model misspecification, which is known to impact posterior probabilities (e.g., Buckley 2002; Lemmon and Moriarty 2004; Yang and Rannala 2005). In particular, unconstrained analyses as performed above with `MrBayes` implicitly assume that the tree topology provides no information about relative branch lengths. (Drummond et al. 2006) suggested that modeling the time-dependency of the evolutionary process should improve tree reconstruction. To assess this proposition here, we incorporated time-dependency by setting calibration constraints in the taxon-removal analysis. `BEAST` was used as described in the Material and Methods section; in particular, the monophyly of the different groups was not enforced. The results

(Fig. 2.7 B) show that all five groups studied here are monophyletic and highly stable, to the exception of Phaeocystales that exhibit signs of LBA and tend to become sister to the Prymnesiales only when > 13 taxa are removed from the analysis. Therefore, the enforcement of time-dependency stabilizes the tree reconstruction process, minimizing the impact of highly divergent taxa, and thereby, appears to remove most LBA artifacts from the analysis.

In the face of this result, it is desirable to know whether a particular calibration constraint or set of constraints has a major stabilizing effect, or if the mitigation of LBA is mainly due to the time-dependency structure of the model. To address this question, we reran the taxon-removal analyses with select calibration constraints (node 47, or nodes 47 & node 57, or node 57, or nodes 77 & 79), or only with the extremely diffuse prior on the root (see Table 2.3). To simplify the presentation of the results, we focus on the two clades that showed evidence for LBA in the unconstrained analysis: the coccolithophores and the Phaeocystales. Figure 2.8 A shows that in the case of the coccolithophores, the introduction of time-dependency into the model is solely responsible for the stabilizing effect, irrespective of the calibration constraints used. On the other hand, PP stabilization for Phaeocystales depends on the calibration constraints included: when no constraints other than the vague root prior are incorporated into the model, LBA apparently disappears when a small number of taxa (8) are removed. Alternatively, in the presence of ("internal") calibration constraints, LBA removal requires that more taxa be removed (Fig. 2.8 B). Therefore, the introduction of time-dependency into a model might help mitigate some of the LBA artifacts, but is not eliminating them all.

2.4.4. Both nuclear and plastid genes share the same phylogeny

Our fourth intrinsic assumption was that both the nuclear and the plastid genes share the same history. This need not be so as endosymbiotic gene transfers postdating the divergence of haptophytes could have affected the history of these genomes.

Figure 2.4 C-D shows the trees obtained for the nuclear and for the plastid genes, respectively. Note that the branch lengths were longer for the plastid tree, reflecting the fact that the protein-coding genes evolve much faster than the RNA genes (see above). Some differences were observed in the relative positions of certain taxa within the coccolithophore clade, notably for *Algirosphaera robusta*, and PPs were generally lower within this clade in the plastid gene tree. In spite of these differences, the nuclear and the plastid trees were not significantly different at the 1% level (SH: $p = 0.020$). Again, this test might not be the most appropriate in the context of hypotheses derived from Bayesian analyses, but it nonetheless indicates that (i) there is no strong evidence that the phylogenetic signal has been blurred by horizontal gene transfer, endosymbiotic gene transfer (or replacement), or by a "tertiary transfer" from which the haptophytes would have received their plastid (Hackett et al. 2007), and (ii) the data can be analyzed under the simplest model that contains one single partition without significantly affecting the reconstructed tree.

2.4.5 Robustness of the phylogeny to the evolutionary process

Our last assumption was that the evolutionary processes assumed here are time-homogeneous, that is, do not change in time across the different lineages.

Figure 2.9 shows the trees estimated under a rate-across-site model (CAT, panel A), a rate-across-lineage model (BP, panel B) and a rate-across-site and lineage model

(CAT-BP, panel C). Under the GTR + Γ_5 substitution model, the best (maximum likelihood) tree was the one estimated under CAT-BP, and the two other trees were not significantly different from this one at the 1% level (SH test: $p_{CAT} = 0.415$; $p_{BP} = 0.549$). This result suggests that it is safe to ignore spatiotemporal variation of substitution rates and that saturation is not an issue.

2.4.6. Reconstruction of ancestral characters and of paleo-environments

Because the phylogeny obtained is relatively well supported (Fig. 2.2), a simple maximum likelihood reconstruction of ancestral characters is appropriate, and does not require integrating over topological uncertainty. Our reconstruction of the evolution of the ability to calcify is represented in Figure 2.10A. Calcification is shared by most coccolithophores, and has clearly been secondarily lost in the genus *Isochrysis*. The model predicts that while the ability to calcify had not evolved in the earliest haptophyte (with a probability of 0.828), the most recent common ancestor of the Calcihaptophycideae and Prymnesiales (node 52 of Fig. 2.1) has a 0.815 probability of having possessed the ability to calcify. Intracellular calcification may thus have evolved early, before the divergence of the Calcihaptophycideae and Prymnesiales (between 329-291 MYA). Calcification was later lost twice, along the branches leading to (i) the Prymnesiales (between 291-171 MYA) and (ii) the Isochrysidaceae (between 119-37 MYA).

The ability to calcify required the presence of organic plate scales, but these scales were probably not a sufficient condition. Figure 2.10 B shows that the cenancestor of the haptophytes had a high probability (0.906) of possessing organic plate scales. This suggests that knob scales evolved on the branch leading to the Pavlovales, i.e. between 543-230 MYA.

Similarly, the model indicates that the cenancestor of the haptophytes inhabited a coastal environment (probability 1.000; Fig. 2.7 C). The cenancestor of the Prymnesiales may not have left coastal environments (P = 0.645), in which case only one transition towards oceanic environments occurred, presumably after the divergence of coccolithophores and Prymnesiales, between 291-243 MYA around the time of the P/Tr boundary event (251 MYA). Some taxa then returned to a coastal environment: first with the most recent common ancestor of the Hymenomonadaceae and Pleurochrysidaceae (between 181-130 MYA), and a second time, independently, prior to the divergence of the Isochrysidaceae (between 119-37 MYA).

Finally, Figure 2.10 D shows the reconstructed trophic modes and suggests that autotrophy evolved from mixotrophy probably only once (P = 0.677) by losing the phagocytic ability around the split Phaeocystales / Prymnesiales, between 329-291 MYA. Our model suggests that mixotrophy was then regained along the branch leading to the Prymnesiales, between 291-171 MYA.

2.5. Discussion

Molecular clocks represent a potentially powerful tool for generating insights into the major events in the evolutionary history of groups of organisms, provided they are applied and interpreted with appropriate caution. Here we attempt to further develop these insights by combining a relaxed molecular clock analysis based on a statistically sound Bayesian phylogenetic reconstruction of the haptophytes, with a maximum likelihood reconstruction of probable ancestral character states. The resulting estimate of the timeline of phenotypic and ecological evolution in this important group of photosynthetic protists can then be

interpreted in a geological context.

2.5.1. Phylogeny of the Haptophytes

Our extensive analysis of a large four-gene haptophyte dataset indicates that the phylogenetic tree shown in Figure 2.2 is robust to (i) data partitioning, (ii) the genomic origin of the genes (host *vs*. endosymbiont), (iii) across-site rate variation, and (iv) across-lineage rate variation. Overall, this haptophyte tree is highly consistent with existing taxonomic schemes and with previous molecular phylogenies, notably the single gene Haptophyta phylogenies presented by (Medlin et al. 2008). Although their analysis included almost twice the number of taxa, the data generated and assembled here have a comparable taxonomic range, and underwent a more comprehensive analysis.

The order of the early divergences within the Calcihaptophycideae remains the most contentious part of the molecular phylogeny of the Haptophyta, but our extensive analyses suggest that the results based on models that include time-dependency are robust to LBA with the complete alignment. The earliest divergence within the Calcihaptophycideae has most often placed the Isochrysidales as a sister group to all other coccolithophore orders (de Vargas et al. 2007; Edvardsen et al. 2000), *TufA* phylogeny of Medlin et al. (2008). This scenario would imply that holococcolith, a structure that results from a unique calcification process in haploid phase cells and that is not present in extant Isochrysidales, evolved after this divergence, along the branch leading to all other coccolithophores (Syracosphaerales / Zygodiscales / Coccolithales). The placement of the divergence of the nannolith-bearing genus *Braarudosphaera* as basal to the entire calcihaptophyte clade by (Takano et al. 2006) would not alter this interpretation since this genus is not known to produce a holococcolith-bearing haploid phase. In contrast, our consensus tree gives the

earliest divergence within the Calcihaptophycideae by placing the Coccolithales as a sister group to a clade that includes Isochrysidales and Syracosphaerales + Zygodiscales. The *SSU* phylogeny of Medlin et al. (2008) also groups *Coronosphaera*, a genus probably affiliated to the Syracosphaerales, with the Isochrysidales. Our results and those of Medlin et al. (2008) would therefore both support an early origin of holococcoliths. This is consistent with a number of observations: (i) a number of cytological features of the Isochrysidales, such as the structure of plate scales (when present) or that of flagellar roots, are relatively simple; (ii) while the oldest fossil holococcoliths date back to *ca.* 185 MYA Bown (1998) instead of the expected 220 MYA under the scenario of an early divergence of Isochrysidales, holococcoliths are more fragile and significantly less well preserved in the fossil record than heterococcoliths, so that an earlier origin cannot be ruled out on the grounds of absence of fossils; (iii) secondary loss of holococcoliths is known to have occurred in a clade of coastal Coccolithales species (Young et al. 2005), so that an additional secondary loss in the Isochrysidales is conceivable as this clade possesses a number of other lost features such as the haptonema. Because holococcoliths are formed by a complex calcification process that is quite unlikely to have evolved more than once (Young et al. 1999), our result of an early origin of calcification with a secondary loss of holococcoliths in Isochrysidales (Fig. 2.10 A) is reasonable.

2.5.2. Timing of the diversification of the haptophytes

The reliability of molecular clock estimates is obviously related to the accuracy of fossil calibration constraints. In this context the fossil record of coccolithophores is unique in providing many well-documented fossil dates. Ongoing work on their biostratigraphy is likely to continually refine the accuracy of fossil dates in the future. In the current context

however, our analysis proved to be robust to the specification of calibration constraints, as the removal or the addition of a constraint to the five that were employed did not affect the estimated dates significantly (Tables 2.5, 2.7 and 2.9).

Since in practice, fossil dates are defined with varying levels of accuracy, there is often a trade-off in choosing the number of constraints to be employed in a molecular clock analysis between the degree of coverage (of the phylogeny and the time range) and the degree of confidence in constrained dates. The Bayesian modeling adopted here elegantly accommodates both sides of the trade-off, first with an increased coverage by employing more CCs within the coccolithophore clade than any previous study (de Vargas et al. 2007; Medlin et al. 2008), and second with the placement of relatively vague prior distributions on these CCs. As the taxonomic range of multigene datasets of the coccolithophores increases, notably within the undersampled Syracosphaerales and Zygodiscales clades, a number of additional fossil dates known with relative confidence could be used to calibrate relaxed molecular clocks to extend our work and further test our results.

Reciprocally, the predictive power of our analysis can be assessed by checking congruence between our time estimates and nodes that are unconstrained in our analysis but for which fossil dates do exist. One such example of congruence with an older, character-based constraint is the date for the origin of alkenones. These are distinctive lipids produced exclusively by members of the Isochrysidales and best known for their use as paleo-indicators of surface water temperature (Brassell et al. 1986; Conte et al. 1998; Prahl and Wakeham 1987). The known geological record of alkenones extends down to the mid-Cretaceous at *ca.* 120 MYA (Brassell and Dumitrescu 2004), a first occurrence that is not very well constrained and could conceivably have a much earlier age (Medlin et al.

2008). Since alkenones are produced by all members of the Isochrysidales, they must have evolved some time between the divergence of this order from other coccolithophores and the first divergence within the order, a range that we estimated at 203-119 MYA. Given that molecular divergences always predate morphological divergences, there is quite a remarkable degree of congruence between our time estimates and external fossil dates in this very particular example. On existing evidence, an interpretation would be that alkenones evolved shortly before the crown divergence of species within this order, although in the absence of information relating to the function of alkenones there is no reason to believe that the two events were causally linked.

Despite overall consistency of the reconstructed phylogeny with previous studies and the use of comparable fossil constraints, our results did depart from those of previous studies, in particular with respect to the estimated times of early divergences within the Haptophyta division. Firstly, the divergence between Stramenopiles and the haptophytes was here dated *ca.* 824 MYA (1031-637), which is significantly more recent than the 1100 MYA average often estimated (e.g., Medlin et al. 1997; Medlin et al. 2008; Yoon et al. 2004). Secondly, we dated the divergence of the Pavlovales at 543 MYA (823-328), while Medilin et al. (2008) estimated it around 800 MYA. Thirdly, the divergence of the Phaeocystales, which Medilin et al. (2008) estimated at *ca.* 490 MYA, was dated at 329 MYA (428-248) in our analysis. In each of these cases, our divergence time estimates were younger, and significantly so in two cases out of three, than previous molecular estimates reported in the literature. Several factors could explain these differences.

Firstly, we used a Bayesian approach that integrates over all uncertainties of the model parameters, including the tree topology. However, because the estimated topology proved

to be highly supported, this factor is unlikely to explain the important difference in time estimates compared with previous studies. Secondly, previous studies were mainly based on the slowly evolving *SSU* RNA gene. The incorporation of all three codon positions of protein-coding genes is often criticized for incorporating noise into phylogenetic analyses, but a theoretical study suggests otherwise (Seo and Kishino 2008). Therefore, it might be important to incorporate these rapidly evolving positions to help discriminate otherwise poorly resolved nodes and dates. Again, since most of the nodes were well resolved, incorporating rapidly evolving genes is unlikely to explain completely why our time estimates are younger than previous molecular studies. Thirdly, we used a relaxed uncorrelated clock model to estimate divergence times. Simulation studies show that this class of models outperforms all other dating methodologies (Aris-Brosou 2007), including the penalized likelihood approaches (Sanderson 2002; Sanderson 2003) that have been used in most previous molecular studies (Berney and Pawlowski 2006). Finally, note that relaxed clock models are specifically designed to account for among-lineage rate variation, either by means of an autocorrelated process (Sanderson 2002; Sanderson 2003), or like here with an uncorrelated process (Drummond et al. 2006). Therefore, the inclusion of genes that exhibit "non-clocklike" behavior such as *tufA* (Medlin et al. 2008) is unlikely to affect our analysis, as these models account for these non-linear effects (Aris-Brosou 2007).

The phylogenetic relationships between haptophytes and other chromalveolate lineages remain unresolved, but there is general consensus that the crown divergence leading to modern lineages occurred early in the history of this supergroup, some few hundred million years after the origin of the eukaryotes (e.g., Cavalier-Smith 2006). A

recent molecular study (Berney and Pawlowski 2006) dated the cenancestral eukaryote at *ca.* 1126 (948-1357) MYA and the origin of haptophytes at slightly later than 900 MYA. Considering that 'Bayesian algorithms' can miss rapid rate variation (but see Aris-Brosou 2007), Cavalier-Smith 2006) proposed that eukaryotes originated 900 ± 100 MYA, with chloroplasts and Plantae evolving between 570-850 MYA, and chromalveolates, opisthokonts, Rhizaria and excavates originating '*ca.* 570 MYA' when the Proterozoic snowball Earth melted. Our clock estimate for the origin of haptophytes at 824 (1031-637) MYA is not consistent with this theory that chromalveolates originated shortly before the Cambrian explosion.

Cavalier-Smith (2006) also argues for the sudden origin of many phyla near the Precambrian boundary, followed by the origin of novel classes and/or orders in the early Mesozoic and early Cenozoic, presumably by exploiting niches or whole adaptive zones emptied by the greatest mass extinctions. Our estimated date for the divergence of the two extant haptophyte classes puts this event near Precambrian boundary (543 MYA). Our analysis indicates that the primary radiation within the Prymnesiophyceae (the divergence of Phaeocystales from other prymnesiophytes) occurred in the Carboniferous period, a considerable time before the P/Tr boundary event (end of the Paleozoic / start of the Mesozoic). Alternatively, widespread shallow epicontinental seas persisted throughout much of the Carboniferous, a period that was preceded by an important extinction event at the end of the Devonian. Our estimates for the timing of the next two divergences within the Prymnesiophyceae, the divergence of the Prymnesiales and the primary radiation of the Calcihaptophycideae, both coincide with important Earth system transitions, early in the Permian and the Triassic respectively.

2.5.3. Timing of the environmental adaptations of the haptophytes

Our reconstructions of four key evolutionary transitions (calcification, nature of organic scales, oceanic environment and trophic mode) suggest that calcification evolved along the same lineage where the phagocytic ability was lost, just after the divergence of the Phaeocystales (ca. 230 MYA). Although conceivably calcification might hamper predation, a strict dependence or ecological link between these two transitions does not seem to be supported by the situation in *Isochrysis*, a genus that has lost its calcification capacity without regaining mixotrophy.

Likewise, the transition to an oceanic environment took place after the two previous transitions to calcification and full autotrophy, *ca.* 230-172 MYA. Yet, it is difficult to argue that both calcification and autotrophy were prerequisites for the transition to an oceanic environment. Indeed, the two clades that returned to coastal environments either remained autotrophic calcifiers (*Hymenomonadaceae* and *Pleurochrysidaceae*) or never completely lost the phagocytic ability in the first place and lost the ability to calcify (Prymnesiales). As a result, calcification, trophic mode and transitions to an oceanic environment seem to be mutually independent transitions. This result is consistent with previous studies that suggested that the interaction between calcification and photosynthesis may not be direct (Brownlee and Taylor 2004).

Explaining the origin of algal plastids continues to be a major challenge in evolutionary biology (Yoon et al. 2002). As the most recent common ancestor of all haptophytes was most probably mixotrophic, photosynthesis must have evolved before the origin of this group, i.e. before 824 MYA. This time interval is fully consistent with the 1072-767 MYA time window estimated by Douzery et al. (2004) with 129 proteins from 34 eukaryotes, and

supports their claim that this secondary event occurred shortly after the primary endosymbiosis. However, these time windows are extremely wide, so that it is difficult to argue about the exact point in time when a red-algal plastid was acquired by secondary endosymbiotic event (*contra* Medlin et al. 1997). Consequently, we cannot rule out the possibility that the haptophytes led an early heterotrophic life before acquiring their plastid. Only the inclusion of an early-branching organism to break up the long branch that leads to haptophytes, or the evidence that the haptophyte plastid is shared with another lineage such as the Cryptophytes, as suggested by Rice and Palmer (2006) from the existence of a bacterial gene in both plastids, can help resolve this issue.

Our reconstruction of the emergence of organic scales supports an alternative scenario to what was previously proposed de Vargas et al. (2007). Indeed, our results show that the proto-haptophytes may have already evolved the ability to produce organic plate scales, and that plate scales reverted to a simple and less elaborate knoblike scale on the branch leading to the Pavlovales. Therefore, the ability to control the intracellular precipitation of calcite on the plate scale most likely emerged in the prymnesiophyte ancestor of the coccolithophores.

2.6. Conclusions

We have presented the first robust and extensive phylogeny of the haptophytes. Our results are consistent with previous work based on morphology (Young et al. 1999) or on the fossil record (Bown et al. 2004). Although we dated the most recent common ancestor of calcifying haptophytes to 243 MYA, our analyses suggest that the ability to calcify evolved much earlier than this, probably between 329-243 MYA, in the Carboniferous / early Triassic. As this innovation was shortly followed by the transition of these organisms to the

global ocean in the Permian / Triassic, our results imply that global carbon cycles were probably impacted by the haptophytes much earlier than previously thought (Fabry 2008; Ridgwell and Zeebe 2005)

2.7. Tables

Table 2.1: List of the strains from the Roscoff Culture Collection for which 28S rDNA was sequenced.

Species	Strain
Coccolithus pelagicus	AC613
Calcidiscus quadriperforatus	RCC1147
Calcidiscus leptoporus	RCC1157
Umbilicosphaera hulburtiana	RCC1474
Umbilicosphaera foliosa	RCC1470
Umbilicosphaera sibogae	RCC1468
Cruciplacolithus neohelis	RCC1206
Calyptrosphaera sphaeroidea	RCC1178
Helladosphaera sp	RCC1182
Hymenomonas coronata	RCC1339
Jomonlithus litoralis	RCC1354
Hymenomonas globosa	RCC1338
Ochrosphaera neapolitana	AC94
Ochrosphaera sp.	CCMP2002
Pleurochrysis carterae	RCC1418
Pleurochrysis dentata	RCC1400
Gephyrocapsa oceanica	RCC1289
Isochrysis galbana	RCC1348
Isochrysis littoralis	RCC1346
Algirosphaera robusta	AC503
Coronosphaera mediterranea	RCC1204
Syracosphaera pulchra	RCC1460
Helicosphaera carteri	RCC1333
Scyphosphaera apsteinii	RCC1455
Prymnesium parvum	RCC1434
Prymnesium sp.	RCC1443
Platychrysis pigra	RCC1390
Imantonia rotunda	RCC1343
Phaeocystis sp.	AC618
Pavlova virescens	RCC1535
Pavlova pinguis	RCC1538
Rebecca salina	RCC1545
Exanthemachrysis gayraliae	RCC1523

Table 2.2: Accession numbers of the sequences included in this study.

Species	LSU	SSU	tufA	rbcL
Coccolithus pelagicus	EU729464*	AJ246261	AJ544128	X
Calcidiscus quadriperforatus	EU502878*	AJ544115	AJ544124	X
Calcidiscus leptoporus	EU729460*	AJ544116	AJ544126	AB043690
Umbilicosphaera hulburtiana	EU729463*	AM490993	AM502981	X
Umbilicosphaera foliosa	EU729462*	AJ544119	AJ544130	AB043629
Umbilicosphaera sibogae	EU729461*	AJ544118	AJ544129	AB043691
Cruciplacolithus neohelis	EU729467*	AB058348	X	AB043689
Calyptrosphaera sphaeroidea	EU729466*	AM490990	X	AB043628
Helladosphaera sp	EU729465*	AB183607	X	X
Hymenomonas coronata	EU819083	AM490981	X	X
Jomonlithus litoralis	EU502875*	AM490979	X	X
Hymenomonas globosa	EU502872	AM490982	X	X
Ochrosphaera neapolitana	EU729469*	AM490980	X	X
Ochrosphaera sp.	EU819082	AB183615	X	X
Pleurochrysis carterae	EU819084	AJ246263	AJ544131	D11140
Pleurochrysis dentata	EU729468*	AJ544121	AJ544132	AB043688
Gephyrocapsa oceanica	EU729476*	AB058360	AF545609	D45844
Isochrysis galbana	EU729474*	AJ246266	AF545610	AB043693
Isochrysis littoralis	EU819085	AM490996	X	X
Algirosphaera robusta	EU729470*	AM490985	Algirosp	X
Coronosphaera mediterranea	EU729471*	AM490986	Coronosp	X
Syracosphaera pulchra	EU502879*	AM490987	X	X
Helicosphaera carteri	EU729473*	AM490983	AJ544134	X
Scyphosphaera apsteinii	EU729472*	AM490984	X	X
Prymnesium patelliferum	AF289038	L34671	X	X
Prymnesium parvum	EU729443*	AJ246269	X	AB043698
Prymnesium sp.	EU729445*	U40923	X	X
Platychrysis pigra	EU729458*	AM491003	X	X
Imantonia rotunda	EU729457*	AJ246267	X	X
Phaeocystis sp.	EU729477*	X77475	X	X
Pavlova virescens	EU729477*	AJ515248	AF545612	X
Pavlova pinguis	EU502883*	AF106047	X	X
Rebecca salina	EU729478*	L34669	X	AB043633
Exanthemachrysis gayraliae	EU729479*	AF106060	X	X
Vaucheria bursata	AF409127	U41646	U09448	AF476940
Tribonema aequale	Y07979	M55286	AF038002	AF084611
Undaria pinnatifida	AY851528	AF319007	AF038003	AY851535
Costaria costata	AY851522	AB022819	U09429	AY851541
Heterosigma akashiwo	AF409124	AB183667	AF545613	AB176660
Skeletonema costatum	EF433522	AY684947	AF015569	AF545615

Notes—Missing genes are indicated by a cross (X). An asterisk (*) denotes the accession numbers of the sequences obtained in this study (see Supplementary Table 2.1 for corresponding RCC identifiers).

Table 2.3: Specification of calibration constraints (CC) used in BEAST. Times are in billion years. A lognormal process was assumed for modeling evolution of the rates of

evolution across lineages. Plus signs (+) indicate the offset applied to each prior (minimum age setting).

	root	node 47	node 57	node 62	node 63	node 77	node 79
4 CC	LN(0.0, 0.5) + 0.5	∅	∅	E(0.01) + 0.065	E(0.01) + 0.024	E(0.01) + 0.055	E(0.01) + 0.025
5 CC	LN(0.0, 0.5) + 0.5	∅	E(0.1) + 0.220	E(0.01) + 0.065	E(0.01) + 0.024	E(0.01) + 0.055	E(0.01) + 0.025
6 CC	LN(0.0, 0.5) + 0.5	E(0.2) + 0.350	E(0.1) + 0.220	E(0.01) + 0.065	E(0.01) + 0.024	E(0.01) + 0.055	E(0.01) + 0.025

Notes—Node identifiers represent the following divergences (see Figure 2.1 for details): Exanthemachrysis gayraliae and Helicosphaera carteri (node 47); Coccolithus pelagicus and Helicosphaera carteri (node57); Coccolithus pelagicus and Umbilicosphaera hulburtiana (node 62); Calcidiscus leptoporus and Umbilicosphaera foliosa (node 63); Coronosphaera mediterranea and Scyphosphaera apsteinii (node 77); Helicosphaera carteri and Scyphosphaera apsteinii (node 79). LN(x,y): lognormal distribution with mean x and variance y; E(x): exponential distribution with parameter x; ∅: no CC specified.

Table 2.4: Marginal log-likelihoods $P(X \mid M)$ of the data X under model M (BEAST analysis with CCs as in Table 2.3).

.

Model	$P(X \mid M)$	SE
4 CCs	-43496.868	0.154
5 CCs	-43495.906	0.164
6 CCs	-43496.116	0.164

Table 2.5: Posterior means of node-specific divergence times (in billion years) with their 5% and 95% bounds; three time constraints are considered with four (4CCs), five (5CCs) or six (6CCs) calibrations. Node identifiers refer to Figure 2.1. (BEAST analysis with CCs as in Table 2.3).

	4 CCs			5 CCs			6 CCs		
	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD
root	0.621	0.792	0.980	0.637	0.824	1.031	0.640	0.815	1.017
node 42	0.276	0.515	0.829	0.289	0.518	0.799	0.277	0.554	0.857
node 43	0.216	0.414	0.627	0.232	0.424	0.645	0.233	0.446	0.675
node 44	0.160	0.314	0.474	0.170	0.316	0.478	0.168	0.332	0.516
node 45	0.066	0.102	0.140	0.069	0.107	0.147	0.070	0.107	0.145
node 46	0.081	0.184	0.309	0.084	0.187	0.321	0.083	0.194	0.326
node 47	0.269	0.478	0.765	0.328	0.543	0.823	0.350	0.480	0.659
node 48	0.102	0.211	0.341	0.105	0.230	0.370	0.106	0.221	0.351
node 49	0.067	0.159	0.272	0.070	0.172	0.286	0.072	0.168	0.279
node 50	0.032	0.094	0.174	0.036	0.103	0.188	0.035	0.100	0.179
node 51	0.173	0.280	0.394	0.248	0.329	0.428	0.247	0.320	0.406
node 52	0.155	0.243	0.344	0.232	0.291	0.363	0.231	0.285	0.349
node 53	0.065	0.147	0.237	0.085	0.171	0.257	0.080	0.166	0.255
node 54	0.034	0.085	0.147	0.039	0.095	0.158	0.036	0.094	0.159
node 55	0.016	0.053	0.101	0.018	0.059	0.108	0.017	0.059	0.110
node 56	0.001	0.005	0.012	0.001	0.006	0.014	0.001	0.006	0.013
node 57	0.125	0.197	0.273	0.220	0.243	0.285	0.220	0.241	0.278
node 58	0.099	0.154	0.215	0.124	0.181	0.236	0.128	0.180	0.232
node 59	0.030	0.072	0.116	0.028	0.075	0.123	0.035	0.078	0.124
node 60	0.001	0.007	0.016	0.001	0.008	0.018	0.001	0.008	0.018

node 61	0.073	0.104	0.143	0.077	0.113	0.154	0.077	0.113	0.154
node 62	0.065	0.075	0.092	0.065	0.077	0.096	0.065	0.077	0.096
node 63	0.035	0.053	0.072	0.037	0.056	0.074	0.036	0.054	0.073
node 64	0.012	0.029	0.048	0.012	0.031	0.051	0.012	0.030	0.049
node 65	0.023	0.040	0.057	0.025	0.043	0.061	0.024	0.041	0.060
node 66	0.008	0.020	0.034	0.008	0.022	0.036	0.008	0.021	0.035
node 67	0.065	0.112	0.165	0.078	0.130	0.183	0.082	0.129	0.180
node 68	0.043	0.087	0.134	0.054	0.100	0.149	0.055	0.099	0.146
node 69	0.012	0.045	0.087	0.013	0.051	0.096	0.012	0.051	0.093
node 70	0.019	0.055	0.095	0.022	0.062	0.105	0.021	0.060	0.103
node 71	0.004	0.020	0.042	0.004	0.023	0.047	0.004	0.022	0.046
node 72	0.013	0.045	0.081	0.014	0.052	0.095	0.017	0.053	0.092
node 73	0.101	0.178	0.276	0.126	0.203	0.301	0.124	0.203	0.305
node 74	0.047	0.105	0.171	0.044	0.119	0.193	0.053	0.119	0.193
node 75	0.008	0.034	0.067	0.008	0.037	0.075	0.008	0.037	0.074
node 76	0.065	0.123	0.198	0.073	0.140	0.229	0.075	0.140	0.229
node 77	0.055	0.068	0.091	0.055	0.072	0.098	0.055	0.071	0.095
node 78	0.012	0.029	0.050	0.012	0.032	0.054	0.012	0.032	0.054
node 79	0.025	0.031	0.042	0.025	0.031	0.043	0.025	0.031	0.043

Table 2.6: Marginal log-likelihoods $P(X \mid M)$ (BEAST analysis with node 79 at 50MYA).

Model	$P(X \mid M)$	SE
4 CCs	-43496.005	0.174
5 CCs	-43496.044	0.192
6 CCs	-43495.699	0.165

Table 2.7: Posterior means of node-specific divergence times (in billion years) with their 5% and 95% bounds; three time constraints are considered with four (4CCs), five (5CCs) or six (6CCs) calibrations. Node identifiers refer to Figure 2.1(BEAST analysis with node 79 at 50MYA).

.

	4 CCs			5 CCs			6 CCs		
	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD
root	0.622	0.806	1.008	0.645	0.838	1.058	0.633	0.826	1.039
node 42	0.257	0.514	0.859	0.293	0.561	0.892	0.297	0.529	0.803
node 43	0.191	0.408	0.657	0.245	0.450	0.703	0.242	0.427	0.646
node 44	0.144	0.305	0.494	0.179	0.334	0.518	0.176	0.320	0.492
node 45	0.061	0.100	0.141	0.071	0.112	0.157	0.067	0.110	0.154
node 46	0.080	0.178	0.312	0.089	0.198	0.325	0.086	0.189	0.319
node 47	0.270	0.512	0.807	0.343	0.568	0.851	0.350	0.508	0.710
node 48	0.089	0.214	0.358	0.105	0.235	0.381	0.106	0.224	0.358
node 49	0.063	0.161	0.286	0.072	0.179	0.302	0.068	0.168	0.279
node 50	0.031	0.096	0.182	0.036	0.108	0.199	0.038	0.102	0.191
node 51	0.180	0.291	0.418	0.248	0.335	0.445	0.245	0.325	0.413
node 52	0.165	0.255	0.362	0.233	0.296	0.378	0.231	0.289	0.361
node 53	0.064	0.150	0.238	0.084	0.174	0.267	0.080	0.168	0.255
node 54	0.031	0.085	0.147	0.040	0.099	0.168	0.038	0.094	0.157
node 55	0.015	0.053	0.099	0.019	0.061	0.117	0.017	0.057	0.105
node 56	0.001	0.005	0.012	0.001	0.006	0.014	0.001	0.006	0.014
node 57	0.133	0.207	0.286	0.220	0.245	0.292	0.220	0.243	0.286
node 58	0.103	0.161	0.226	0.127	0.186	0.243	0.127	0.181	0.235
node 59	0.026	0.071	0.115	0.027	0.078	0.126	0.025	0.076	0.124
node 60	0.001	0.007	0.017	0.001	0.009	0.020	0.001	0.008	0.019
node 61	0.073	0.106	0.145	0.078	0.116	0.160	0.077	0.114	0.158
node 62	0.065	0.076	0.094	0.065	0.078	0.099	0.065	0.078	0.097
node 63	0.035	0.054	0.073	0.038	0.056	0.075	0.037	0.055	0.075
node 64	0.011	0.030	0.049	0.013	0.031	0.050	0.012	0.031	0.049
node 65	0.023	0.040	0.058	0.025	0.043	0.061	0.024	0.042	0.061

node 66	0.008	0.020	0.034	0.008	0.022	0.036	0.008	0.021	0.036
node 67	0.066	0.117	0.173	0.076	0.133	0.190	0.082	0.130	0.186
node 68	0.047	0.090	0.143	0.052	0.100	0.154	0.053	0.099	0.149
node 69	0.011	0.046	0.085	0.014	0.051	0.094	0.013	0.051	0.094
node 70	0.020	0.055	0.096	0.023	0.061	0.108	0.021	0.060	0.103
node 71	0.004	0.020	0.042	0.004	0.023	0.048	0.004	0.023	0.047
node 72	0.013	0.046	0.083	0.016	0.055	0.101	0.016	0.052	0.094
node 73	0.106	0.192	0.301	0.132	0.209	0.309	0.131	0.215	0.329
node 74	0.042	0.116	0.198	0.054	0.123	0.195	0.051	0.124	0.205
node 75	0.008	0.036	0.073	0.009	0.038	0.077	0.009	0.039	0.079
node 76	0.078	0.138	0.215	0.085	0.147	0.231	0.083	0.150	0.237
node 77	0.063	0.083	0.107	0.062	0.085	0.113	0.064	0.086	0.112
node 78	0.013	0.035	0.061	0.013	0.037	0.063	0.012	0.036	0.064
node 79	0.050	0.054	0.063	0.050	0.055	0.063	0.050	0.054	0.063

Table 2.8: Marginal log-likelihoods *P(X | M)* (BEAST analysis with CC at node 61 instead of node 62).

Model	*P*(*X* | *M*)	SE
4 CCs	-43496.215	0.172
5 CCs	-43495.200	0.153
6 CCs	-43495.574	0.174

Table 2.9: Posterior means of node-specific divergence times (in billion years) with their 5% and 95% bounds; three time constraints are considered with four (4CCs), five (5CCs) or six (6CCs) calibrations. Node identifiers refer to Figure 2.1(BEAST analysis with CC at node 61 instead of node 62).

	4 CCs			5 CCs			6 CCs		
	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD
root	0.620	0.779	0.967	0.629	0.815	1.024	0.633	0.805	1.007
node 42	0.210	0.447	0.694	0.259	0.489	0.782	0.260	0.490	0.764
node 43	0.184	0.361	0.552	0.205	0.391	0.608	0.211	0.399	0.603
node 44	0.140	0.265	0.414	0.142	0.291	0.467	0.161	0.296	0.461
node 45	0.053	0.091	0.134	0.058	0.100	0.142	0.064	0.101	0.142
node 46	0.067	0.159	0.265	0.077	0.168	0.293	0.082	0.178	0.300
node 47	0.248	0.473	0.759	0.336	0.563	0.848	0.350	0.477	0.665
node 48	0.083	0.189	0.312	0.101	0.232	0.375	0.100	0.215	0.338
node 49	0.060	0.143	0.250	0.073	0.174	0.298	0.072	0.161	0.265
node 50	0.028	0.084	0.159	0.037	0.105	0.190	0.032	0.098	0.169
node 51	0.160	0.261	0.372	0.246	0.327	0.428	0.245	0.315	0.402
node 52	0.141	0.229	0.323	0.230	0.288	0.363	0.230	0.282	0.347
node 53	0.062	0.140	0.226	0.076	0.166	0.259	0.083	0.163	0.250
node 54	0.028	0.078	0.133	0.037	0.092	0.157	0.035	0.088	0.150
node 55	0.013	0.048	0.090	0.017	0.057	0.109	0.016	0.054	0.104
node 56	0.000	0.005	0.011	0.001	0.006	0.014	0.001	0.006	0.013
node 57	0.120	0.185	0.259	0.220	0.242	0.284	0.220	0.240	0.277
node 58	0.088	0.139	0.195	0.116	0.174	0.233	0.116	0.173	0.230
node 59	0.022	0.055	0.088	0.028	0.061	0.095	0.026	0.062	0.097
node 60	0.001	0.007	0.015	0.001	0.008	0.018	0.001	0.008	0.018
node 61	0.065	0.082	0.108	0.065	0.090	0.117	0.065	0.090	0.118
node 62	0.045	0.065	0.090	0.047	0.072	0.097	0.047	0.072	0.099
node 63	0.030	0.047	0.065	0.033	0.052	0.073	0.032	0.051	0.072
node 64	0.010	0.026	0.043	0.012	0.029	0.049	0.011	0.029	0.048
node 65	0.020	0.036	0.052	0.022	0.040	0.058	0.021	0.039	0.057
node 66	0.006	0.018	0.030	0.008	0.020	0.034	0.008	0.019	0.033
node 67	0.059	0.103	0.151	0.071	0.124	0.180	0.073	0.124	0.178

node 68	0.040	0.080	0.123	0.050	0.094	0.148	0.051	0.095	0.143
node 69	0.010	0.041	0.077	0.012	0.048	0.091	0.013	0.049	0.090
node 70	0.016	0.049	0.086	0.019	0.057	0.099	0.022	0.058	0.101
node 71	0.003	0.019	0.041	0.004	0.022	0.045	0.004	0.022	0.046
node 72	0.013	0.041	0.073	0.016	0.049	0.088	0.015	0.050	0.091
node 73	0.092	0.163	0.246	0.128	0.212	0.326	0.130	0.206	0.313
node 74	0.037	0.097	0.162	0.052	0.126	0.211	0.049	0.119	0.195
node 75	0.006	0.031	0.063	0.007	0.038	0.081	0.008	0.038	0.078
node 76	0.073	0.124	0.193	0.087	0.149	0.236	0.085	0.148	0.237
node 77	0.060	0.079	0.103	0.063	0.084	0.110	0.062	0.084	0.110
node 78	0.011	0.031	0.057	0.012	0.035	0.061	0.013	0.035	0.062
node 79	0.050	0.054	0.062	0.050	0.054	0.063	0.050	0.054	0.063

Table 2.10: Marginal log-likelihoods $P(X \mid M)$ of the data X under model M (BEAST analysis with CCs as in Table 2 without *Undaria pinnatifida)*.

Model	$P(X \mid M)$	SE
4 CCs	-40587.449	0.170
5 CCs	-40587.366	0.175
6 CCs	-40587.471	0.178

Table 2.11: Posterior means of node-specific divergence times (in billion years) with their 5% and 95% bounds; three time constraints are considered with four (4CCs), five (5CCs) or six (6CCs) calibrations. Node identifiers refer to figure 2.1.

	4 CCs			5 CCs			6 CCs		
	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD
root	0.620	0.768	0.932	0.675	0.843	1.024	0.665	0.828	1.003
node 42	0.347	0.511	0.679	0.383	0.581	0.769	0.402	0.579	0.763
node 43	0.298	0.445	0.599	0.335	0.505	0.684	0.344	0.503	0.675
node 44	0.220	0.352	0.483	0.247	0.401	0.553	0.264	0.399	0.549
node 45	na	na	na	na	na	na	na	na	na
node 46	0.126	0.225	0.332	0.140	0.258	0.376	0.149	0.257	0.375
node 47	0.317	0.465	0.633	0.388	0.536	0.704	0.375	0.506	0.650
node 48	0.109	0.186	0.269	0.131	0.215	0.305	0.127	0.210	0.296
node 49	0.104	0.178	0.261	0.126	0.207	0.299	0.120	0.202	0.288
node 50	0.059	0.111	0.166	0.070	0.128	0.191	0.069	0.126	0.186
node 51	0.176	0.238	0.306	0.239	0.292	0.352	0.240	0.288	0.343
node 52	0.163	0.219	0.280	0.233	0.271	0.318	0.230	0.267	0.310
node 53	0.083	0.132	0.187	0.107	0.163	0.221	0.104	0.158	0.213
node 54	0.046	0.078	0.114	0.057	0.094	0.133	0.058	0.092	0.131
node 55	0.027	0.053	0.081	0.034	0.064	0.097	0.034	0.063	0.094
node 56	0.001	0.005	0.009	0.001	0.005	0.011	0.001	0.005	0.011
node 57	0.140	0.187	0.235	0.220	0.234	0.261	0.220	0.233	0.259
node 58	0.110	0.148	0.189	0.144	0.180	0.218	0.142	0.179	0.214
node 59	0.044	0.074	0.106	0.053	0.087	0.122	0.051	0.086	0.120
node 60	0.002	0.007	0.013	0.002	0.008	0.015	0.002	0.008	0.015
node 61	0.078	0.102	0.129	0.091	0.118	0.149	0.090	0.117	0.148
node 62	0.065	0.080	0.098	0.067	0.090	0.111	0.066	0.089	0.111
node 63	0.045	0.060	0.077	0.049	0.067	0.085	0.049	0.067	0.086
node 64	0.019	0.034	0.049	0.021	0.038	0.054	0.022	0.038	0.056
node 65	0.032	0.047	0.063	0.036	0.053	0.070	0.037	0.053	0.071
node 66	0.013	0.024	0.035	0.015	0.027	0.039	0.015	0.027	0.040
node 67	0.075	0.109	0.146	0.094	0.131	0.169	0.092	0.130	0.167
node 68	0.053	0.081	0.114	0.064	0.096	0.132	0.063	0.096	0.131
node 69	0.022	0.044	0.070	0.025	0.052	0.081	0.026	0.052	0.082
node 70	0.028	0.052	0.078	0.033	0.061	0.090	0.033	0.060	0.090
node 71	0.007	0.018	0.033	0.008	0.022	0.037	0.009	0.022	0.038

node 72	0.021	0.042	0.065	0.026	0.051	0.077	0.027	0.051	0.078
node 73	0.124	0.189	0.271	0.158	0.216	0.309	0.161	0.231	0.317
node 74	0.067	0.116	0.186	0.077	0.131	0.199	0.080	0.140	0.219
node 75	0.014	0.031	0.052	0.017	0.036	0.060	0.015	0.037	0.061
node 76	0.093	0.134	0.179	0.104	0.157	0.207	0.105	0.158	0.213
node 77	0.071	0.094	0.121	0.074	0.103	0.132	0.075	0.103	0.133
node 78	0.017	0.036	0.057	0.019	0.040	0.064	0.019	0.040	0.063
node 79	0.050	0.053	0.060	0.050	0.054	0.062	0.050	0.054	0.062

Table 2.12: Marginal log-likelihoods *P(X | M)* (BEAST analysis with CC at node 61 instead of node 62 without *Undaria pinnatifida)*.

Model	*P*(*X* | *M*)	SE
4 CCs	-40587.578	0.168
5 CCs	-40587.599	0.175
6 CCs	-40587.863	0.183

Table 2.13: Posterior means of node-specific divergence times (in billion years) with their 5% and 95% bounds; three time constraints are considered with four (4CCs), five (5CCs) or six (6CCs) calibrations. Node identifiers refer to Figure 2.1.

	4 CCs			5 CCs			6 CCs		
	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD
root	0.608	0.751	0.905	0.666	0.836	1.013	0.660	0.819	0.994
node 42	0.336	0.495	0.669	0.400	0.572	0.755	0.395	0.561	0.754
node 43	0.278	0.430	0.585	0.335	0.497	0.661	0.326	0.486	0.652
node 44	0.207	0.341	0.475	0.253	0.394	0.536	0.246	0.385	0.530
node 45	na	na	na	na	na	na	na	na	na
node 46	0.120	0.218	0.327	0.147	0.254	0.368	0.135	0.246	0.356
node 47	0.302	0.444	0.606	0.375	0.531	0.694	0.366	0.498	0.638
node 48	0.100	0.176	0.258	0.127	0.213	0.302	0.124	0.206	0.289
node 49	0.097	0.170	0.250	0.123	0.207	0.297	0.116	0.195	0.277
node 50	0.057	0.106	0.161	0.070	0.128	0.191	0.066	0.122	0.180
node 51	0.165	0.227	0.296	0.240	0.289	0.345	0.237	0.285	0.337
node 52	0.152	0.209	0.270	0.232	0.268	0.311	0.231	0.265	0.306
node 53	0.079	0.127	0.179	0.105	0.159	0.217	0.106	0.157	0.215
node 54	0.044	0.075	0.109	0.056	0.092	0.130	0.056	0.091	0.130
node 55	0.026	0.051	0.078	0.032	0.063	0.095	0.033	0.062	0.095
node 56	0.001	0.004	0.009	0.001	0.005	0.011	0.001	0.005	0.010
node 57	0.131	0.177	0.225	0.220	0.233	0.258	0.220	0.232	0.255
node 58	0.100	0.139	0.178	0.138	0.176	0.210	0.138	0.175	0.210
node 59	0.039	0.067	0.095	0.051	0.081	0.112	0.050	0.081	0.112
node 60	0.002	0.006	0.012	0.002	0.008	0.014	0.002	0.008	0.014
node 61	0.066	0.092	0.116	0.084	0.111	0.139	0.081	0.110	0.136
node 62	0.054	0.075	0.098	0.066	0.091	0.116	0.065	0.089	0.114
node 63	0.039	0.056	0.074	0.048	0.067	0.087	0.048	0.066	0.086
node 64	0.017	0.031	0.047	0.022	0.038	0.056	0.021	0.037	0.054
node 65	0.028	0.044	0.060	0.036	0.053	0.071	0.034	0.052	0.070
node 66	0.011	0.022	0.034	0.015	0.027	0.040	0.014	0.027	0.039
node 67	0.069	0.103	0.137	0.092	0.128	0.165	0.093	0.127	0.164
node 68	0.048	0.077	0.108	0.063	0.095	0.129	0.062	0.094	0.127
node 69	0.020	0.042	0.066	0.025	0.051	0.078	0.023	0.051	0.078
node 70	0.025	0.049	0.073	0.032	0.060	0.088	0.031	0.059	0.088
node 71	0.007	0.017	0.030	0.008	0.021	0.037	0.007	0.021	0.035
node 72	0.018	0.041	0.063	0.026	0.050	0.076	0.027	0.050	0.075
node 73	0.111	0.173	0.252	0.159	0.221	0.312	0.160	0.219	0.309
node 74	0.058	0.105	0.165	0.074	0.134	0.206	0.077	0.133	0.211
node 75	0.012	0.029	0.049	0.016	0.036	0.059	0.016	0.036	0.059

node 76	0.089	0.127	0.170	0.105	0.157	0.212	0.103	0.155	0.206
node 77	0.069	0.092	0.118	0.076	0.103	0.133	0.074	0.102	0.132
node 78	0.016	0.034	0.055	0.019	0.040	0.063	0.019	0.040	0.062
node 79	0.050	0.053	0.060	0.050	0.054	0.062	0.050	0.054	0.062

Table 2.14: Marginal log-likelihoods *P(X | M)* of the data *X* under model *M* (BEAST analysis with CCs as in Table 2 for *SSU* and *LSU* (39 species)).

Model	*P*(*X* | *M*)	SE
4 CCs	-22132.288	0.231
5 CCs	-22132.579	0.208
6 CCs	-22131.978	0.232

Table 2.15: Posterior means of node-specific divergence times (in billion years) with their 5% and 95% bounds; three time constraints are considered with four (4CCs), five (5CCs) or six (6CCs) calibrations. Node identifiers refer to Figure 2.1.

	4 CCs			5 CCs			6 CCs		
	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD
root	0.782	1.146	1.567	0.856	1.243	1.671	0.819	1.173	1.570
node 42	0.476	0.798	1.162	0.506	0.882	1.290	0.506	0.848	1.227
node 43	0.397	0.666	0.943	0.428	0.745	1.073	0.433	0.707	1.013
node 44	0.352	0.621	0.893	0.406	0.691	1.010	0.395	0.649	0.936
node 45	na	na	na	na	na	na	na	na	na
node 46	0.223	0.389	0.576	0.245	0.429	0.639	0.234	0.406	0.590
node 47	0.369	0.596	0.863	0.418	0.661	0.940	0.403	0.592	0.800
node 48	0.133	0.244	0.367	0.145	0.260	0.380	0.147	0.257	0.380
node 49	0.088	0.175	0.275	0.097	0.186	0.281	0.095	0.185	0.284
node 50	0.053	0.124	0.204	0.059	0.130	0.212	0.058	0.130	0.214
node 51	0.183	0.294	0.410	0.245	0.336	0.437	0.245	0.325	0.417
node 52	0.175	0.260	0.357	0.238	0.300	0.374	0.239	0.293	0.361
node 53	0.098	0.169	0.251	0.111	0.190	0.272	0.109	0.187	0.272
node 54	0.049	0.099	0.154	0.056	0.110	0.169	0.051	0.107	0.165
node 55	0.028	0.068	0.113	0.031	0.076	0.127	0.029	0.073	0.122
node 56	0.001	0.006	0.013	0.001	0.007	0.014	0.001	0.006	0.014
node 57	0.143	0.207	0.281	0.220	0.243	0.284	0.220	0.240	0.278
node 58	0.114	0.163	0.218	0.139	0.185	0.228	0.140	0.183	0.232
node 59	0.059	0.106	0.150	0.071	0.116	0.159	0.071	0.115	0.159
node 60	0.002	0.008	0.017	0.002	0.009	0.019	0.002	0.009	0.019
node 61	0.075	0.096	0.123	0.078	0.103	0.132	0.078	0.102	0.130
node 62	0.065	0.070	0.081	0.065	0.072	0.084	0.065	0.071	0.084
node 63	0.025	0.040	0.054	0.026	0.041	0.057	0.025	0.041	0.055
node 64	0.009	0.023	0.038	0.009	0.024	0.039	0.009	0.023	0.039
node 65	0.026	0.042	0.058	0.027	0.044	0.060	0.028	0.044	0.060
node 66	0.011	0.024	0.038	0.012	0.025	0.040	0.012	0.024	0.039
node 67	0.081	0.126	0.176	0.098	0.141	0.190	0.092	0.140	0.187
node 68	0.058	0.096	0.141	0.067	0.108	0.154	0.064	0.106	0.152
node 69	0.019	0.051	0.086	0.023	0.058	0.097	0.020	0.056	0.093
node 70	0.029	0.061	0.099	0.033	0.069	0.108	0.031	0.067	0.108
node 71	0.007	0.024	0.045	0.008	0.026	0.049	0.008	0.026	0.049
node 72	0.011	0.034	0.062	0.012	0.037	0.069	0.012	0.037	0.068
node 73	0.098	0.159	0.226	0.113	0.182	0.245	0.113	0.178	0.240
node 74	0.046	0.087	0.136	0.052	0.099	0.153	0.050	0.097	0.148
node 75	0.011	0.030	0.054	0.012	0.034	0.060	0.012	0.032	0.059
node 76	0.057	0.078	0.103	0.058	0.080	0.106	0.058	0.078	0.104
node 77	0.066	0.089	0.117	0.067	0.092	0.120	0.067	0.090	0.116
node 78	0.012	0.040	0.069	0.014	0.042	0.072	0.014	0.041	0.069
node 79	0.050	0.054	0.061	0.050	0.054	0.062	0.050	0.054	0.061

Table 2.16: Marginal log-likelihoods *P(X | M)* of the data *X* under model *M* (BEAST analysis with CCs as in Table 2 for *rbcL* and *tufA* (23 species)).

Model	*P*(*X* | *M*)	SE
4 CCs	na	na
5 CCs	-17541.472	0.138
6 CCs	-17541.067	0.150

Table 2.17: Posterior means of node-specific divergence times (in billion years) with their 5% and 95% bounds; three time constraints are considered with four (4CCs), five (5CCs) or six (6CCs) calibrations. Node identifiers refer to Figure 2.1.

	4 CCs			5 CCs			6 CCs		
	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD	< 95% HPD	mean	> 95% HPD
root	na	na	na	0.604	0.766	0.949	0.596	0.755	0.929
node 42	na	na	na	0.141	0.354	0.610	0.125	0.376	0.697
node 43	na	na	na	0.219	0.464	0.788	0.233	0.490	0.831
node 44	na	na	na	0.090	0.288	0.509	0.092	0.291	0.497
node 45	na	na	na	na	na	na	na	na	na
node 46	na	na	na	0.025	0.186	0.343	0.036	0.195	0.361
node 47	na	na	na	0.316	0.581	0.863	0.350	0.509	0.743
node 48	na	na	na	na	na	na	na	na	na
node 49	na	na	na	0.055	0.312	0.788	0.033	0.262	0.589
node 50	na	na	na	na	na	na	na	na	na
node 51	na	na	na	na	na	na	na	na	na
node 52	na	na	na	0.236	0.332	0.445	0.239	0.322	0.419
node 53	na	na	na	na	na	na	na	na	na
node 54	na	na	na	na	na	na	na	na	na
node 55	na	na	na	na	na	na	na	na	na
node 56	na	na	na	na	na	na	na	na	na
node 57	na	na	na	0.220	0.243	0.289	0.220	0.241	0.282
node 58	na	na	na	0.140	0.203	0.263	0.138	0.200	0.255
node 59	na	na	na	0.033	0.085	0.144	0.033	0.084	0.142
node 60	na	na	na	na	na	na	na	na	na
node 61	na	na	na	0.066	0.121	0.181	0.066	0.120	0.179
node 62	na	na	na	0.065	0.095	0.128	0.065	0.095	0.128
node 63	na	na	na	0.042	0.069	0.097	0.042	0.069	0.097
node 64	na	na	na	0.012	0.041	0.067	0.012	0.040	0.066
node 65	na	na	na	0.027	0.052	0.079	0.027	0.052	0.078
node 66	na	na	na	0.008	0.026	0.044	0.009	0.026	0.046
node 67	na	na	na	na	na	na	na	na	na
node 68	na	na	na	na	na	na	na	na	na
node 69	na	na	na	na	na	na	na	na	na
node 70	na	na	na	na	na	na	na	na	na
node 71	na	na	na	na	na	na	na	na	na
node 72	na	na	na	0.023	0.071	0.122	0.020	0.071	0.127
node 73	na	na	na	0.232	0.298	0.387	0.232	0.292	0.369
node 74	na	na	na	0.073	0.184	0.294	0.087	0.186	0.286
node 75	na	na	na	na	na	na	na	na	na
node 76	na	na	na	0.220	0.242	0.307	0.220	0.239	0.301
node 77	na	na	na	0.055	0.073	0.107	0.055	0.072	0.105
node 78	na	na	na	na	na	na	na	na	na
node 79	na	na	na	na	na	na	na	na	na

2.8. Figures

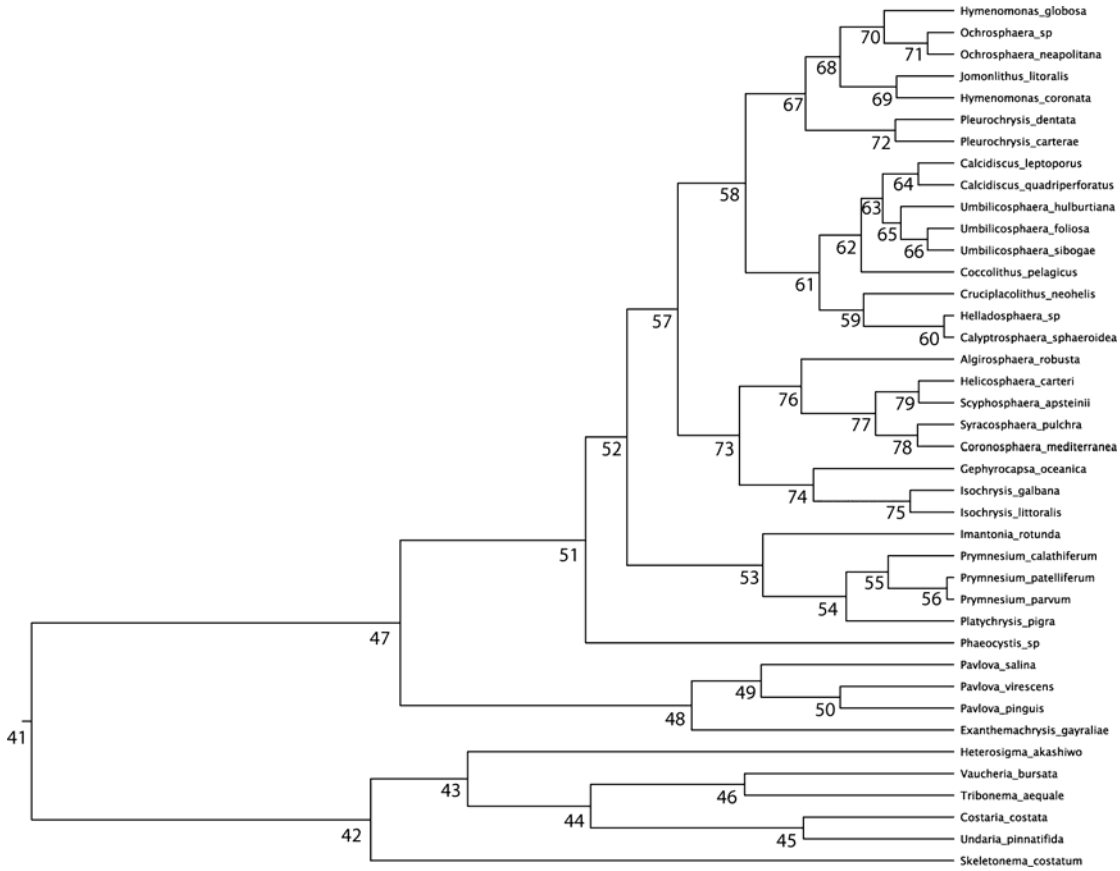


Figure 2.1: Node coding used for estimation of divergence time.

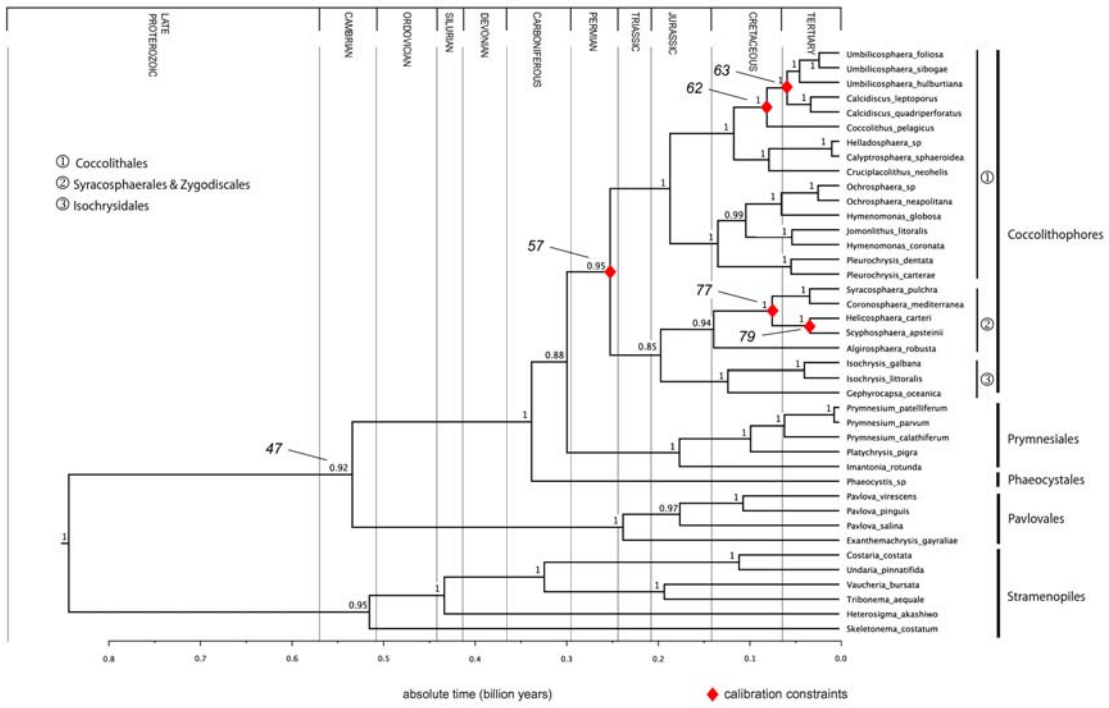


Figure 2.2: Phylogeny and divergence times of the Haptophytes (BEAST analysis). The lognormal model of rate change with five calibration constraints (CCs) was assumed (see text). Placement of CCs on the tree is indicated by red diamonds, labeled as in Figure 2.1. Numbers represent clade posterior probabilities. Times are in billion years.

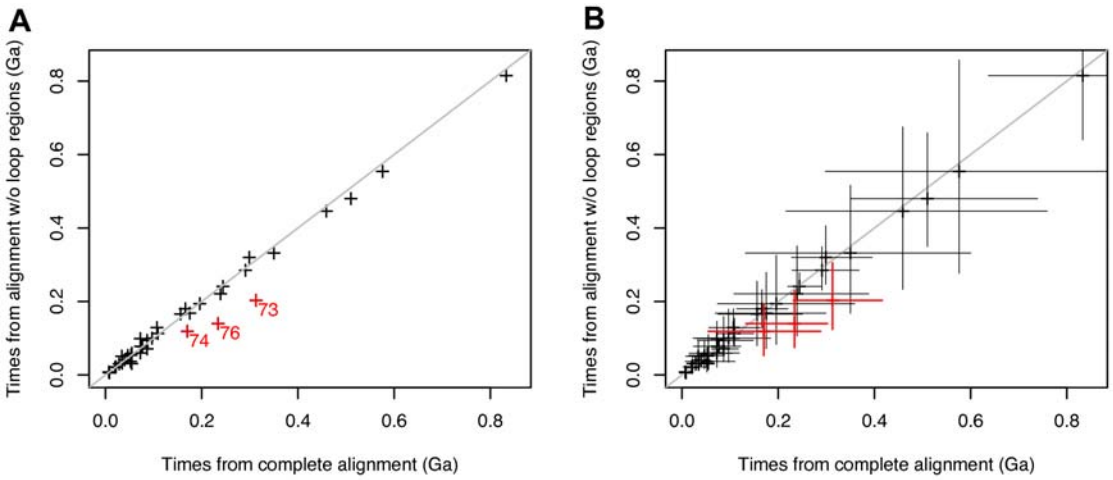


Figure 2.3: Correlation of date estimates obtained with 6 CCs as in Table 2 between the complete alignment (x-axis) and the alignment without loop regions of the *SSU* and *LSU* genes (y-axis). (A) mean posterior estimates; (B) 95% credibility intervals. The three divergences indicated in red are those of the Isochrysidales (see Fig. 2.1 for node identifiers).

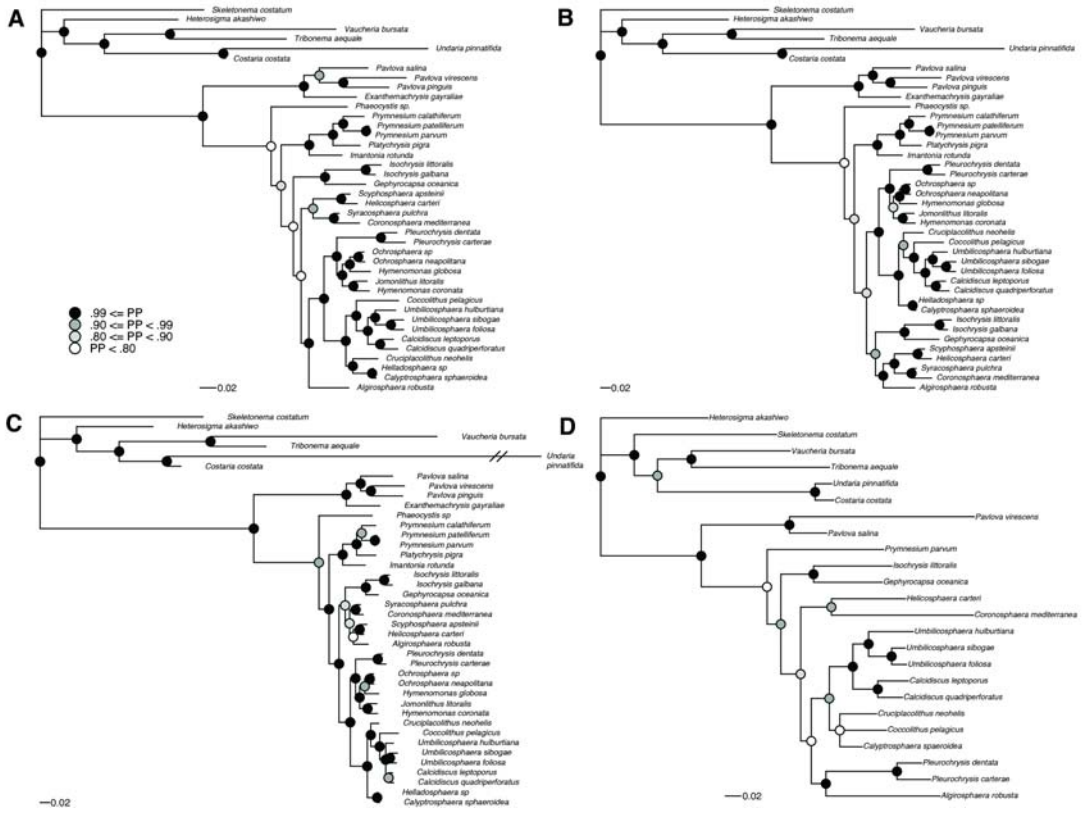


Figure 2.4: Bayesian estimation of the haptophytes phylogeny across different partition of the data: (A) all four genes concatenated; (B) eight partitions: two across the noncoding RNA genes (*SSU* and *LSU*), and one across each of the three codon positions of the two protein-coding genes (*tufA* and *rbcL*); (C) two partitions across each nuclear genes (*SSU* and *LSU*; the long branch leading to *Undaria pinnatifida* is not to scale); (D) two partitions across the protein-coding plastid genes (*tufA* and *rbcL*). PP: posterior probability.

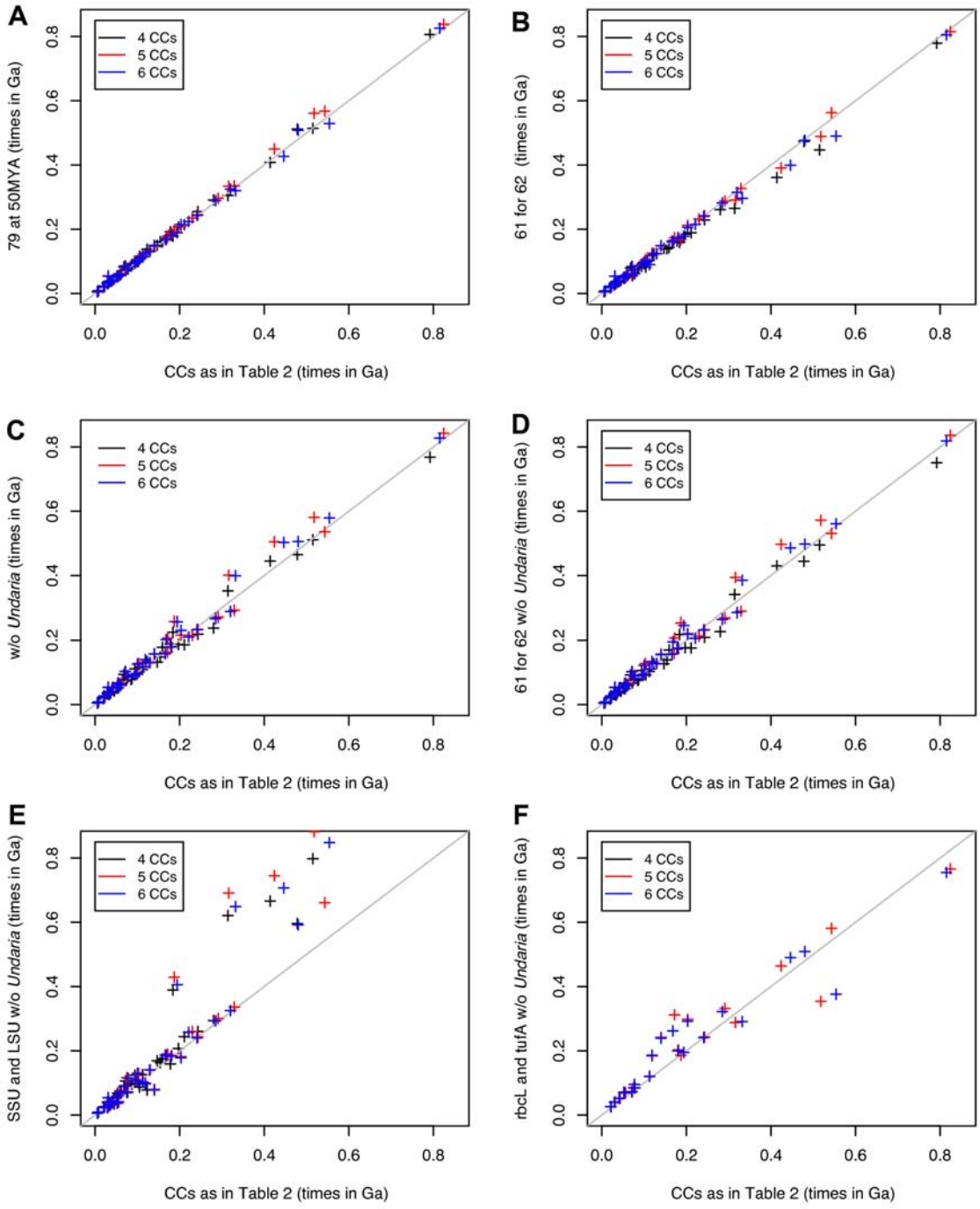


Figure 2.5: Correlation between date estimates obtained with CCs as in Table 2 with the different assumptions examined in the Supplementary Tables: (A) node 79 at 50MYA; (B) CC at node 61 instead of node 62; (C) CCs as in Table 2 without *Undaria pinnatifida*; (D) CC at node 61 instead of node 62 without *Undaria pinnatifida*; (E) CCs as in Table 2 for *SSU* and *LSU* (39 species); (F) CCs as in Table 2 for *rbcL* and *tufA* (23 species).

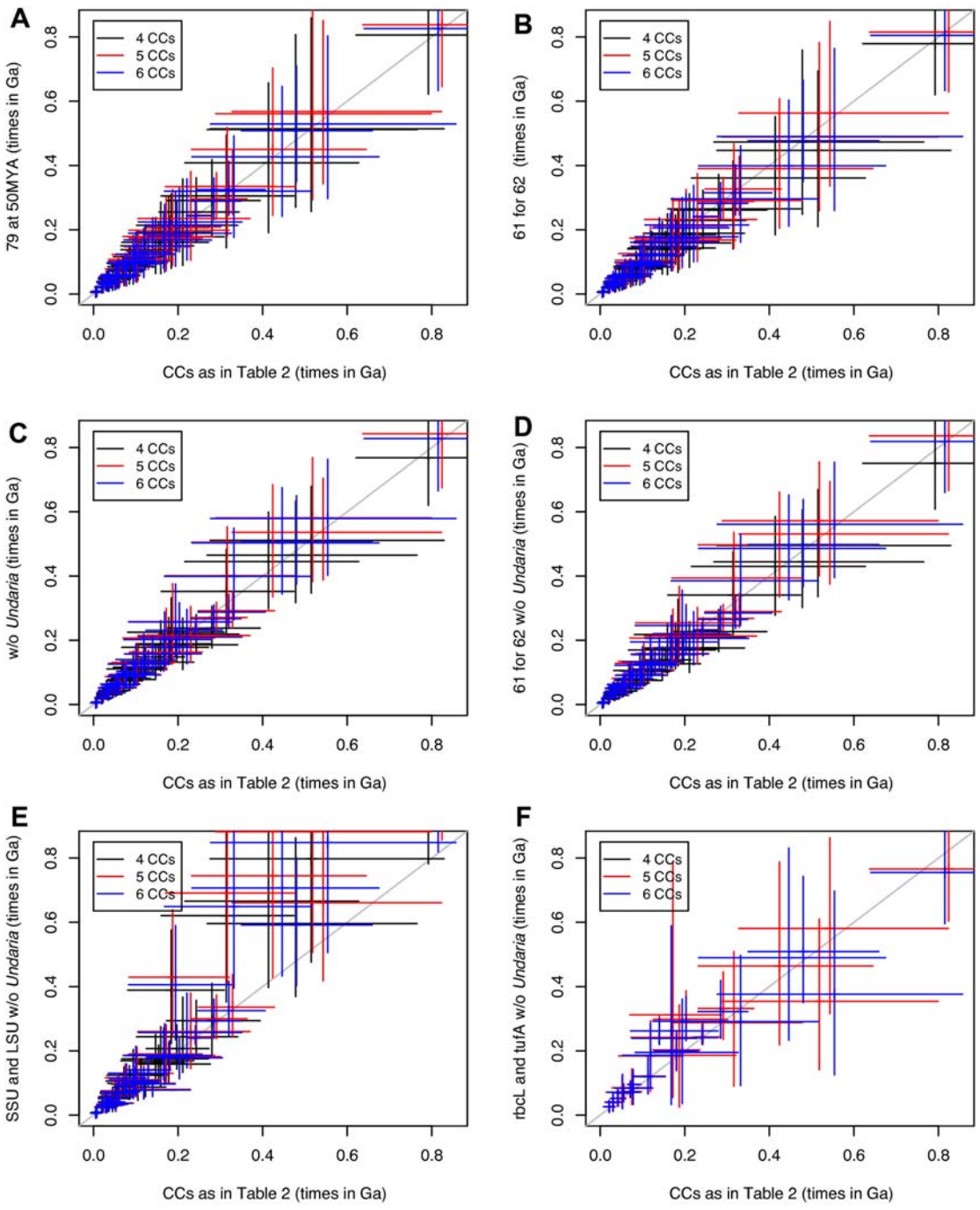


Figure 2.6: Correlation between date estimates obtained with CCs as in Table 2 with the different assumptions examined in the Tables, as in Fig. 2.4, but including the 95% credibility intervals.

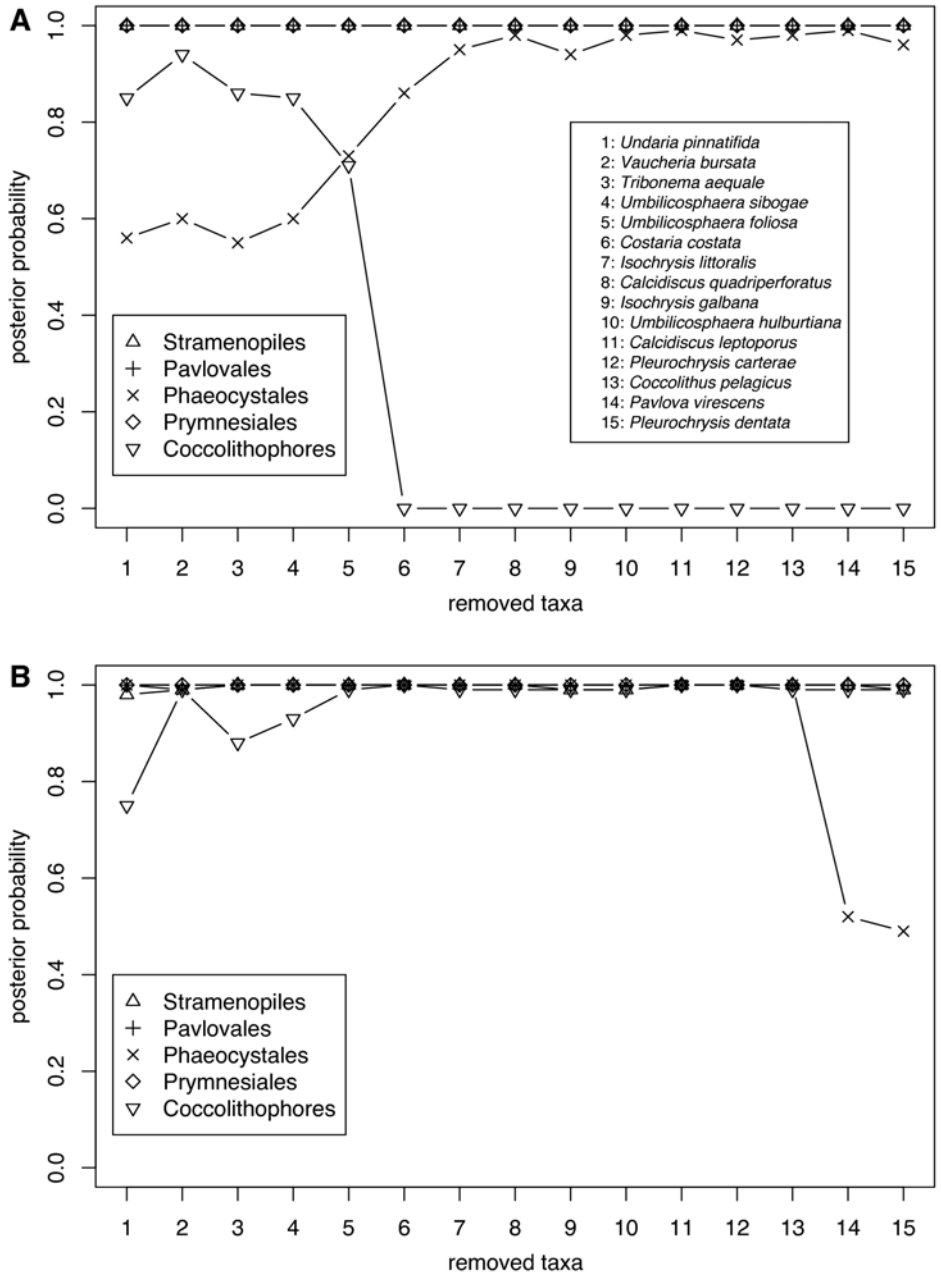


Figure 2.7: Long-branch analysis by taxon removal based on the concatenated alignments: posterior probability of monophyly for (A) time-independent analyses (`MrBayes`); (B) time-dependent analyses (`BEAST`; monophyly not enforced; all six CCs placed and set as in Table 2). The insert in (A) defines the species identifiers used on the x-axis to represent the taxa that were sequentially removed. For the Phaeocystales, the plotted values represent the clade posterior probability of *Phaeocystis sp*.

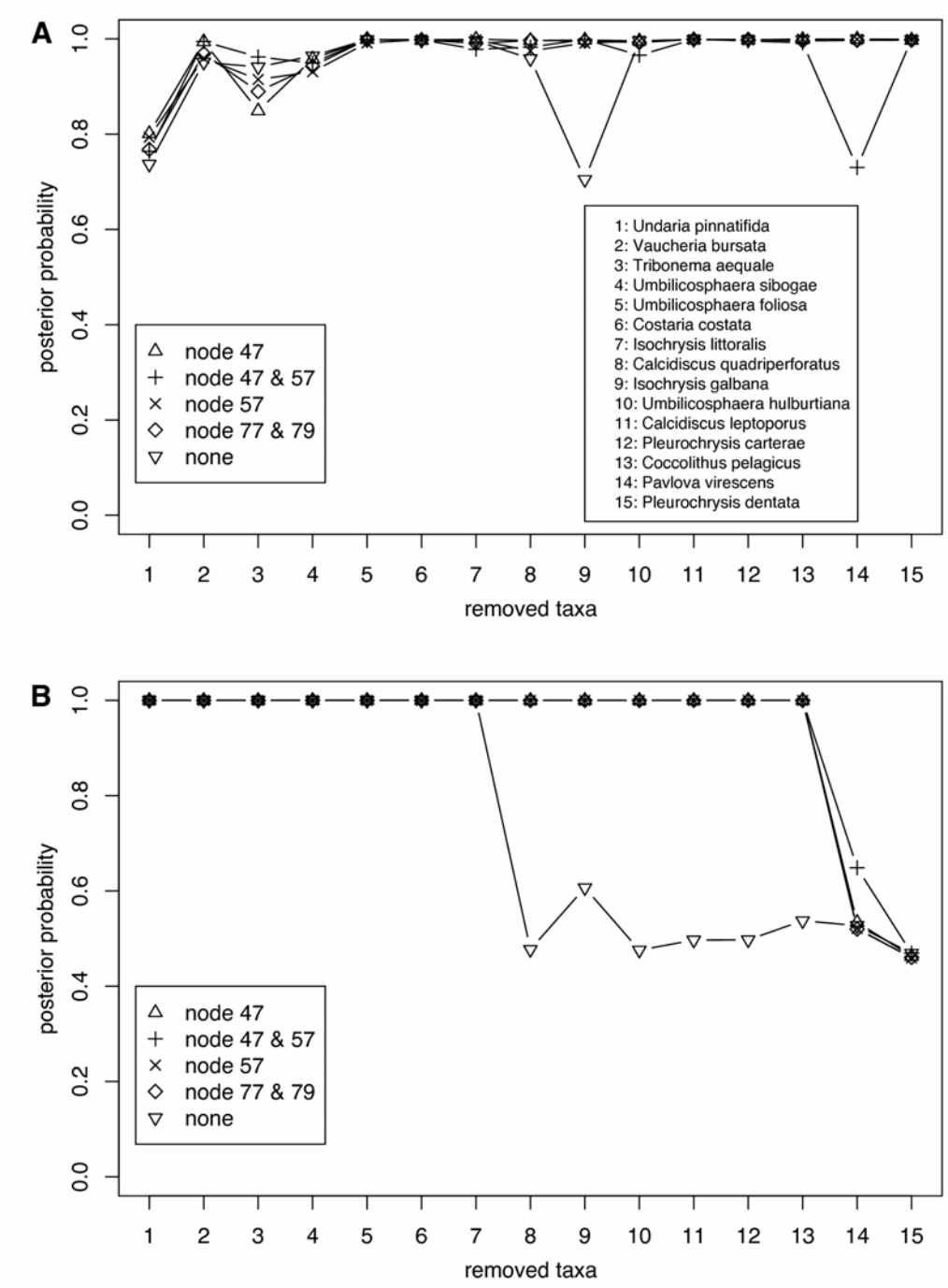


Figure 2.8: Effect of the choice of calibration constraints on the long-branch attraction analysis by CC removal in time-dependent analyses (BEAST): (A) posterior probability of the monophyly of Coccolithophores; (B) posterior probability of the Phaeocytales. The insert in (A) defines the species identifiers used on the x-axis to represent the taxa that were sequentially removed.

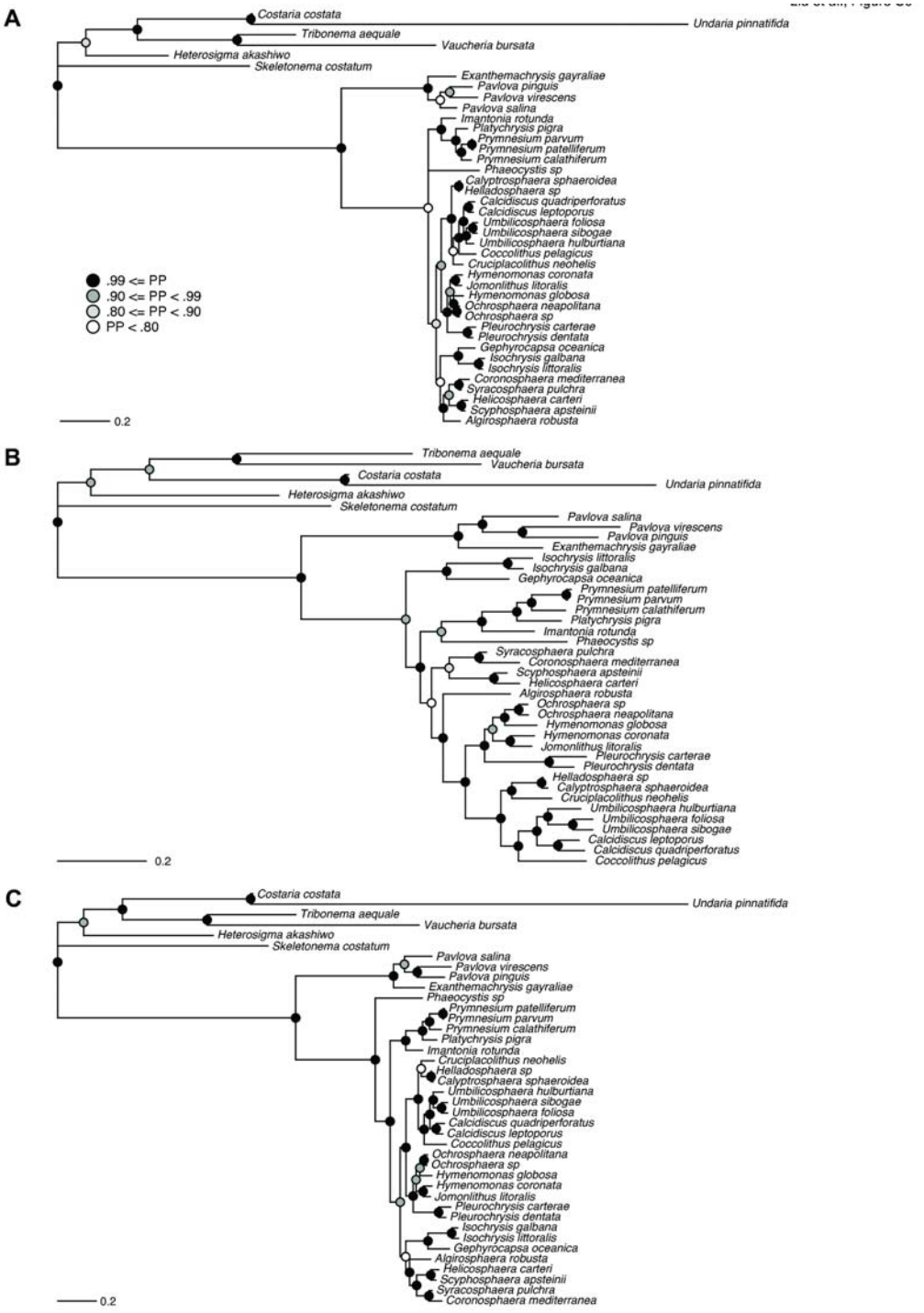


Figure 2.9: Bayesian estimation of the haptophytes phylogeny under different models: (A) CAT; (B) BP; (C) CAT-BP.

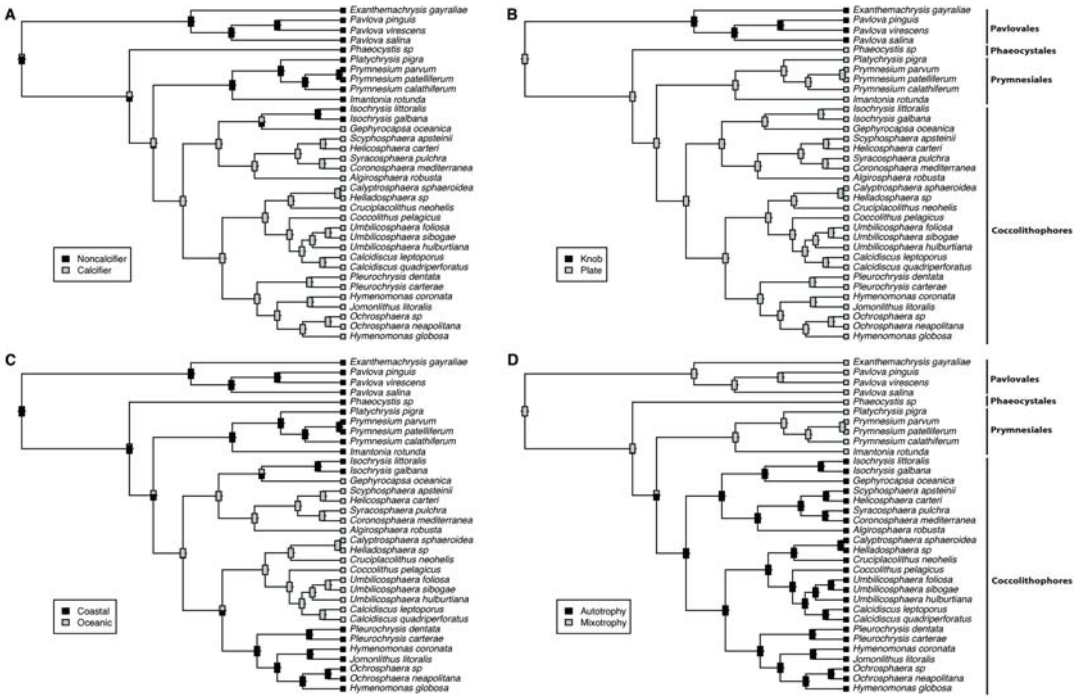


Figure 2.10: Maximum likelihood reconstructions of the paleoecology of the haptophytes: (A) calcification; (B) type of organic plates; (C) marine environment; (D) trophic mode. Box shading at leaf nodes indicates the observed states. Box shading at internal nodes represents the relative probabilities of the different pairs of states.

3.0. Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans

3.1. Abstract

The current paradigm holds that cyanobacteria, which evolved oxygenic photosynthesis more than 2 billion years ago, are still the major light harvesters driving primary productivity in open oceans. Here we show that tiny unicellular eukaryotes belonging to the photosynthetic lineage of the Haptophyta are dramatically diverse and ecologically dominant in the planktonic photic realm. The use of Haptophyta-specific primers and PCR conditions adapted for GC-rich genomes circumvented biases inherent in classical genetic approaches to exploring environmental eukaryotic biodiversity and led to the discovery of hundreds of unique haptophyte taxa in 5 clone libraries from sub-polar and sub-tropical oceanic waters. Phylogenetic analyses suggest that this diversity emerged in Paleozoic oceans, thrived and diversified in the permanently oxygenated Mesozoic Panthalassa, and currently comprises thousands of ribotypic species, belonging primarily to low-abundance and ancient lineages of the 'rare biosphere'. This extreme biodiversity coincides with the pervasive presence in the photic zone of the world ocean of 19'-hexanoyloxyfucoxanthin (19-Hex), an accessory photosynthetic pigment found exclusively in chloroplasts of haptophyte origin. Our new estimates of depth-integrated relative abundance of 19-Hex indicate that haptophytes dominate the chlorophylla-normalized phytoplankton standing stock in modern oceans. Their ecologic and evolutionary success, arguably based on mixotrophy, may have significantly impacted the oceanic carbon pump. These results add to the growing evidence that the evolution of complex eukaryotic cells is a critical force in the functioning of the biosphere.

3.2. Introduction

Oxygenic photosynthesis, the most complex and energetically powerful molecular process in biology, originated in cyanobacteria more than 2 billion years ago in Archean oceans (Brocks et al. 1999). Marine photosynthesis still contributes ~50% of total primary production on Earth (Field et al. 1998). This revolutionary process was integrated, at least once, into an ancestral phagotrophic eukaryotic lineage through the evolution of chloroplasts, which themselves were redistributed to a large variety of aquatic eukaryote lineages via permanent secondary and tertiary endosymbioses (Falkowski and Knoll 2007). Despite this evolutionary trend from photosynthetic prokaryotes to eukaryotes, particularly visible in today's coastal oceans where microalgae such as diatoms and dinoflagellates are omnipresent, cyanobacteria have been repeatedly claimed as the champions of photosynthesis in open ocean waters (Goericke and Welschmeyer 1993). This hypothesis followed the introduction of flow cytometry and molecular genetic approaches to biological oceanography in the 1980s, which revealed astonishing concentrations of minute cyanobacterial cells of the genera *Procholorococcus* and *Synechococcus* in marine waters (Chisholm et al. 1988). The physiology, ecology, and functional and environmental genomics of these prokaryotes are subjects of ongoing intensive study (Kettler et al. 2007).

Several lines of evidence in fact argue for eukaryotic supremacy over marine oxygenic photosynthesis. Flow cytometric cell counts (Li 1995) show that picophototrophic protists (0.2-3μm cell size) are indeed 1-2 orders of magnitude less abundant than cyanobacteria. However, biophysical and group-specific ^{14}C-uptake measurements suggest that tiny eukaryotes can, through equivalent or higher growth rates of relatively larger cells, dominate carbon biomass and net production in both coastal (Worden et al. 2004) and

oceanic (Li 1995) settings. High Performance Liquid Chromatography (HPLC) analyses of group-specific accessory pigments have further stressed the ecologic prevalence of phototrophic protist taxa. In particular, 19'-hexanoyloxyfucoxanthin (19-Hex) was originally estimated to account for 20-50% of total chlorophyll *a* (Chl*a*) biomass in tropical Atlantic and Pacific sites (Andersen et al. 1996) and has since been consistently reported in open ocean photic-zone waters, (e.g., Lovejoy et al. 2006; Not et al. 2008) , suggesting a ubiquitous occurrence of haptophytes in upper layers of the water column. Surveys of genetic diversity based on environmental ribosomal DNA libraries over the last decade have unveiled an unexpected diversity of tiny eukaryotes in all oceans (Moon-van der Staay et al. 2001). Paradoxically, most picoeukaryotic sequence diversity from photic layers represented novel *heterotrophic* (Massana et al. 2004b) and *parasitic* (Guillou et al. 2008) protists within phyla traditionally thought to be dominated by photoautotrophs. This evoked that marine protist diversity might be significantly skewed towards heterotrophic taxa (Vaulot et al. 2002), as appears to be the case for prokaryotes. However, the paucity of haptophyte nuclear rDNA sequences in these surveys contrasts strikingly with the abundance of 19-Hex in marine waters.

Here we use a combination of previously undescribed genetic, pigment, and microscopy data to unveil a dramatic and ancient diversity of unique photosynthetic picoplanktonic protists within the Haptophyta. This diversity could account for the mysteriously high concentration of 19-Hex in the photic layer of the world oceans, our calculations indicating that haptophytes contribute ~2 fold more than either cyanobacteria or diatoms to global oceanic Chl*a* standing stock. The phylogenetic position of these tiny haptophytes implies that they are photophagotrophic, matching the recent discovery of dominant bacterivory by

small eukaryotic phytoplankton in the oceans (Zubkov and Tarran 2008). Mixotrophy may provide a competitive advantage over both purely phototrophic microalgae (including cyanobacteria) and aplastidial protists, and the extreme genetic diversity of tiny haptophytes matches the cellular and behavioral complexity inherent in this mixed mode of nutrition.

3.3. Results and Discussion

3.3.1. A massive unique diversity of oceanic picohaptophytes.

We first show that previous nuclear rDNA PCR-based studies of eukaryotic communities were subject to severe selective amplification biases. Several groups of protists known to have long and/or GC rich rDNA are virtually missing from environmental clone libraries produced by classical PCR amplification protocols using 'general eukaryote' SSU rDNA primers (Massana et al. 2004a; Moon-van de Staay et al. 2001). This is the case for the haptophytes, the rDNA of which has a mean GC content of ~57%. We therefore used haptophyte-specific primers and a PCR protocol designed for GC-rich genomes to amplify LSU rDNA D1-D2 fragments from bulk DNA extracted from the 0.2- to 3- μm fraction of seawater collected at 4 offshore stations in the Arctic and Indian oceans (Fig. 3.1, Table 3.1). Standard eukaryotic rDNA analyses of these samples yielded ~0.4-0.7% haptophyte sequences (Lovejoy et al. 2006; Not et al. 2008). In contrast, our data reveal hundreds of previously undescribed rDNA sequences from tiny haptophytes. Rarefaction curves for individual clone libraries (Fig. 3.2) indicate that current sequencing effort is far from exhaustive, notably in sub-tropical waters where genetic diversity is particularly dramatic. Estimates of the number of unique ribotypes using the Chao1 estimator were 1098-1147 and 325-509, respectively, for the Indian and Arctic ocean samples (with rather large

confidence intervals, see Table 3.2). The frequency distribution of unique ribotypes (Fig. 3.2) indicates higher species richness in subtropical waters, with a substantial number of orphan and deep-branching genotypes (see below) in both warm and cold waters. This parallels recent observations for marine prokaryotes of a 'seed bank' of ancient and rare taxa, termed the 'rare biosphere' (Sogin et al. 2006).

3.3.2. Taxonomy and evolutionary history of the previously undescribed diversity.

The 674 novel environmental LSU rDNA sequences were aligned with 64 orthologous gene sequences from clonal culture strains representing a cross-section of known haptophyte biodiversity. Phylogenetic analyses indicate that all novel environmental sequences belong to the Haptophyta (Fig. 3.3), a eukaryotic phytoplankton division classically considered as nanoplankton (3-20µm) and including the calcifying coccolithophores (de Vargas et al. 2007). However, not a single environmental sequence was strictly identical to any of the taxonomically defined sequences. The vast majority of environmental sequences form new clusters branching deep in the haptophyte phylogeny, most being related to *Chrysochromulina* species from clade B2 within the order Prymnesiales (Edvardsen et al. 2000). The described representatives of this clade are nanoplanktonic and known almost exclusively from coastal and shelf environments (Edvardsen and Paasche 10998); our data show they are in fact derived from open ocean picoplanktonic taxa (Fig. 3.3). Note that the 2 other major prymnesiophyte lineages, the Phaeocystales and the Calcihaptophycidae (de Vargas et al. 2007), also appear to emerge from clusters of picohaptophyte sequences.

Calibration of our tree with the stratigraphic record of coccolithophore taxa (Fig. 3.3) suggests that tiny haptophyte biodiversity emerged more than 250 Ma, in Paleozoic oceans

(Fig.3.4), before the evolution of intracellular biomineralization in the Calcihaptophycidae, which according to both fossil and molecular clock data occurred ~220 Ma (de Vargas et al. 2007). The phylogenetic depth of most picohaptophyte clades argues for a Mesozoic diversification of the group, which may have thrived in the newly permanently oxygenated and largely oligotrophic Panthalassic Ocean, conditions which served as a selection matrix for a wide range of chlorophyll a+c containing protists (Falkowski et al. 2004). Many genotypes or genotype clusters were found exclusively in either sub-arctic or sub-tropical oceans, supporting significant lineage partitioning between cold mixed and warmer stratified waters (Fig. 3.5). The phylogeographic distribution of ribotypes suggests that tropical waters were the original center of diversification, with biodiversity spreading secondarily into higher latitudes, a scenario that fits the putative early radiation of the group in the warm Panthalassa.

3.3.3. Ecologic relevance of the picohaptophytes.

Our genetic survey positions the haptophytes as the most diverse group of picophototrophs in modern open oceans. Recent exploration of chloroplastic SSU rDNA in pelagic (Fuller et al. 2006) and coastal (McDonald et al. 2007) environments supports this conclusion. Haptophytes dominate the emerging chloroplastic view [3] of marine tiny eukaryotic phytoplankton in terms of both diversity and abundance. In a year-round data set from the Gulf of Naples (McDonald et al. 2007), >45% total- and >70% unqiue eukaryote chloroplastic rDNA sequences were of haptophyte origin, 55% of them belonging to the Prymnesiales clade-B2 (Fig. 3.6). This extreme diversity coincides with a numerical significance. Group-specific fluorescent in situ hybridization data from various oceanic

settings indicate that haptophytes represent up to 35% of total picoeukaryotic cell numbers (Worden and Not 2008). Dot blot hybridizations using group-specific chloroplastic rDNA probes indicated a mean dominance of ~45% of haptophytes among other eukaryotic divisions during a 2-year survey of ultraphytoplankton (<5µm cell size) in Mediterranean waters (McDonald et al. 2007). To assess whether these localized observations are representative of a global trend, we evaluated the contribution of haptophytes to oceanic phototrophic biomass using an empirical model based on >2400 worldwide vertical profiles of HPLC pigment data integrated through monthly ocean-color composites of surface Chl*a* concentrations measured by the *SeaWiFS* satellite in the year 2000 (Fig. 3.7). This analysis revealed that 19-Hex is the dominant accessory pigment in the oceans over this period, representing about twice the standing stocks of either fucoxanthin (diatoms) or zeaxanthin (prokaryotes) when normalized to Chl*a*. Haptophytes appear thus to represent the background oceanic light harvesters, contributing from 30 to 50% of total photosynthetic standing stock across the world ocean.

3.3.4. Mixotrophy, the key to the success of tiny haptophytes in open oceans?

The phylogenetic position of the majority of the picohaptophytes in the Prymnesiales strongly suggests that they are mixotrophic, i.e., able to supplement their phototrophic regime with uptake and assimilation of organic nutrients. Laboratory experiments have shown that members of the Prymnesiales are typically capable of ingesting organic particles and preys (e.g., Legrand et al. 2001; Nygaard and Tobiesen 1993; Tillmann 1998). Their third flagellum-like appendix, the haptonema, which is particularly long relative to cell size in all described members of Prymnesiales clade B2, can be used to catch preys and

[3] Note that chloroplast genomes are typically not GC-biased, meaning they are amenable to standard PCR

transfer them to the cell membrane for active phagocytosis (Kawachi et al. 1991). Even in calcihaptophytes, highly modified coccoliths may be involved in harvesting preys (Aubry 2009). Field studies on bacterivory by plastid-containing protists have demonstrated the dominance of tiny haptophyte-like cells in oceanic mixotrophy (e.g., Unrein et al. 2007; Zubkov and Tarran 2008). Recent quantitative evidence (Zubkov and Tarran 2008) revealed that eukaryotic algae with cell size ≤5μm, expected to be mostly haptophytes, carry out *most* of the bacterivory in the euphotic layer of both temperate and tropical Atlantic oceans. Furthermore, significant 19-Hex concentrations were recorded in 200- to 300- m-deep layers of the clearest waters on Earth in the South Pacific gyre (Ras et al. 2008), depths where irradiance at noon is not even sufficient for photosynthesis to cover basic cellular metabolic requirements. The complex combination of phagocytotic and photosynthetic modes of nutrition can be postulated to have allowed haptophytes to attain relatively large size and morphological complexity while maintaining prokaryote-like growth rates, and thus to have radiated into a wide diversity of ecogenotypes. The nutritional flexibility offered by mixotrophy is likely to have equipped the tiny haptophytes with a significant competitive advantage over both purely phototrophic and aplastidial cells under different light (depth) and nutrient regimes[4].

3.4. Conclusion

Besides their unanticipated diversity and abundance, the unveiled haptophytes display morphological features that suggest they play critical roles in organic carbon fluxes on a global scale. Size analyses of cells identified by haptophyte-specific fluorescent probes

---

protocols and may provide a more accurate view of the real phytoplanktonic diversity.

[4] Complex mixotrophic regimes may also explain why none of the open ocean haptophytes are currently available in culture collections.

revealed a mode of ~4μm, with largest sizes of 8-9μm (Fig. 3.8 and Table 3.4). In terms of volume, haptophytes are thus typically 300-3,000 times larger than *Procholorococcus*, the most abundant marine cyanobacteria. The few available electron microscopy images of these open ocean tiny haptophytes indicate that they do produce organic plate scales (Fig. 3.9), a plesiomorphic character common to the overwhelming majority of prymnesiophytes. Interestingly, abundant and diverse *Chrysochromulina* spp. scales were recently observed in Atlantic surface sediments collected at 4,850 m (Gooday et al. 2006). The taxonomic origin and pristine preservation of these scales, previously overlooked in deep-sea sediments due to their minute size (≤1μm), suggest that they were rapidly transported to the sea floor. Eukaryotic scales made of proteins embedded into cellulose and other polysaccharides potentially provide abundant resistant and sticky matter to enhance aggregation and flux of marine snow particles to the deep ocean, contributing to the largely underestimated role of coagulation of small phytoplankters in the biological pump (Richardson and Jackson 2007). Thus, the tiny haptophytes may have been essential mediators of carbon fluxes from the atmosphere to the deep oceans and the lithosphere throughout much of the Phanerozoic Eon.

3.5. Methods

3.5.1. Sampling, DNA extraction, and construction of LSU rDNA clone libraries.

At each sampling station (Figure 3.1 and Table 3.1) 5-15L of seawater was immediately prefiltered through a 200-μm nylon mesh and collected in an acid-washed carboy. The water was then filtered, using peristaltic pumping, through a 3-μm pore-size Nucleopore polycarbonate filter (Millipore), before recovery of picoplanktonic cells in 0.2-μm pore-size Sterivex filter units (Millipore). Filters were preserved in lysis buffer (40 mM

EDTA, 50 mM Tris-HCl, 0.75 M sucrose) and stored at -80°C until genomic DNA extraction was performed as in Diez et al. 2001). Approximately 1,000 bp nuclear LSU rDNA fragments including the D1-D2 domains were PCR amplified using the forward haptophyte-specific primer *Hapto_4* (5'-atggcgaatgaagcgggc-3'), and the reverse general eukaryote primer *Euk_34r* (5'-gcatcgccagttctgcttacc-3'). PCR reactions (98°C for 30s, 50°C for 30s, and 72°C for 60s, with initial denaturation and final extension steps) were performed over a maximum of 30 cycles to limit formation of chimeric sequences (Acinas et al. 2005) using the *Phusion* high-fidelity PCR DNA Polymerase (New England BioLabs) which is specifically suited for amplification of GC-rich DNA. PCR products were purified using the MinElute gel extraction kit (Qiagen) and 3'-A-overhangs were bound to DNA fragments by adding 0.2mM dATP, 1 unit of Taq DNA polymerase and 1X Taq DNA polymerase buffer to the purified PCR product, and incubating for 20 min at 72°C. Classical TA-cloning into OneShot DH5α-T1 competent bacteria using the TOPO TA kit (Invitrogen) was then performed according to the manufacturer's instructions. Clone libraries were checked by PCR using the M13 forward and reverse primers and sequencing of ~25-35 random clones in both directions. The entire process of library construction was repeated until >85% of white colonies yielded high-quality sequences. Libraries were then sent to High-Throughput Sequencing Solutions (www.htseq.org) for random automatic picking of 200 clones, plasmid minipreps, and automatic sequencing of both strands of 150-200 LSU rDNA fragments per library.

3.5.2. Molecular biodiversity, phylogenetic, molecular clock, and biogeographic analyses.

The choice of the *LSU* rDNA D1-D2 fragment over the classically used SSU rDNA to

assess haptophyte diversity was motivated by the inability of the latter marker to distinguish closely related species (Table 3.5). D1-D2 LSU rDNA fragments show virtually no intraspecific variations, while discriminating morphospecies which split in the Pleistocene. Unambiguous LSU rDNA sequences were first screened for chimeras using Check-Chimera (Cole et al. 2003), then double-checked by thorough visual inspection of all sequences producing abnormally long branches in neighbor-joining trees (Saitou and Nei 1987) as described in Hugenholtz and Huber (2003). This conservative approach led to the removal of ~13% of putative chimeric sequences from subsequent analyses. The remaining 674 environmental sequences were added to 64 nuclear LSU rDNA sequences obtained from taxonomically identified clonal haptophyte cultures from the Roscoff Culture Collection (http://www.sb-roscoff.fr/Phyto/RCC), and 3 sequences from GenBank. LSU rDNA sequences were aligned using Muscle (Edgar 2004) and the resulting alignment was manually inspected in Genetic Data Environment 2.2 (Larsen et al. 1993). The Akaike Information Criterion (Posada and Crandall 1998) was used to select the most appropriate model of nucleotide substitution: The general time-reversible model plus Γ_4 and invariable sites. For each of the libraries, PAUP* 4.0b10 (Swofford 2002) was used to build pairwise maximum likelihood distance matrices under the model selected above for estimation of rarefaction curves and rDNA richness based on the average neighbor algorithm implemented in DOTUR (Schloss and Handelsman 2005). Phylogenies including both environmental and culture sequences were reconstructed using MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003), with 2 independent samplers, 10^7 steps, tempering with 1 cold and 3 heated chains, and burn-in of 10^4 steps. A Bayesian analysis implemented in BEAST v1.4.6 (Drummond et al. 2006) was performed to construct the

phylogeny while estimating divergence times (Figure 3.4). This relaxed clock analysis included the 184 sequences from the total alignment that were >95% divergent. Absolute time calibration was based on the earliest geological record for the evolution of calcification (i.e., ~220 Mya Bown et al. 2004) and 4 minimum divergence dates derived from stratigraphic data: in the Order Coccolithales, the ~65-million-year-old first appearance of the genus *Coccolithus* and the ~24-million-year-old divergence between the genera *Umbilicosphaera* and *Calcidiscus*; in the Syracosphaerales: the ~55-million-year-old split between the genera *Coronosphaera* and *Scyphosphaera/Helicosphaera,* and the ~32 million-year-old-divergence between the genera *Scyphosphaera* and *Helicosphaera sensu stricto*. Note that the relative branching pattern between the monophyletic groups (Syracosphaera; Coronosphaera), (Helicosphaera; Scyphosphaera), and (Algirosphaera) is not statistically supported and varies depending on reconstructions methods. Minimum ages were constrained with a diffuse prior $\Gamma(1, .15)$ distribution for the onset of calcification and prior $\Gamma(1, .005)$ distributions for the fossil ages. Convergence was checked by running four independent samplers with 10^8 steps.

3.5.3. Assessment of the contribution of haptophytes to global oceanic photosynthetic biomass.

The model in (Uitz et al. 2006) was modified and adapted to 19'-hexanoyloxyfucoxanthin (19-Hex), fucoxanthin (Fuco), and zeaxanthin (Zea), the major carotenoids of haptophytes, diatoms, and photosynthetic prokaryotes, respectively. We identified a set of parameters to infer vertical profiles of 19-Hex, Fuco, and Zea for stratified or mixed water conditions, and for any given concentration of surface chlorophyll *a* or $[\mathrm{Chl}a]_{\mathrm{surf}}$, based on the

following method. Among the >2400 worldwide HPLC-derived pigment profiles, the ones from stratified waters were discriminated from those from well-mixed waters, based on the ratio of the euphotic layer depth Z_{eu} (the depth at which photosynthetically available radiation is reduced to 1% of its surface value) to the mixed layer depth Z_m. Z_{eu} was computed from the vertical profiles of [Chl*a*] using bio-optical models (Morel and Berthon 1989; Morel and Maritorena 2001), while Z_m was extracted from the *Levitus* global monthly-mean climatology. For both stratified and mixed waters, the vertical profiles of 19-*Hex*, Fuco, and Zea were sorted into "trophic categories" defined by successive intervals of $[Chl a]_{surf}$ values. Average profiles were first computed independently for each trophic category and each pigment. Because the average pigment profiles display a deterministic behavior in terms of vertical shape and magnitude along the trophic gradient, they could be modeled and parameterized as a function of $[Chl a]_{surf}$. The predictive skill of the parameters were successfully tested using an independent dataset (47). Our empirical model was then applied to monthly composites of *SeaWiFS*-derived $[Chl a]_{surf}$ values for year 2000, on a pixel-by-pixel basis. *Z*eu was first computed from $[Chl a]_{surf}$ by using the log-log linear relationship linking $[Chl a]_{surf}$ to the euphotic layer-integrated Chl*a* content (Eq. 8 in (Uitz et al. 2006) and the relationship linking this last parameter to *Z*eu (Morel and Maritorena 2001). The euphotic depth was then compared to the mixed layer depth to determine whether the water column was stratified (i.e., *Z*eu ≥ *Z*m) or mixed (i.e., *Z*eu < *Z*m). For stratified waters, $[Chl a]_{surf}$ was used to produce dimensionless profiles (with respect to depth and biomass) of 19-Hex, Fuco, and Zea, which were then restored to physical units by multiplying depths by *Z*eu and concentrations by the average Chl*a* concentration within the euphotic layer. For mixed water conditions, the surface

concentration of each pigment was inferred from $[Chl\textit{a}]_{surf}$ and extrapolated within the euphotic layer to generate uniform vertical profiles. This procedure yielded monthly depth-resolved fields of 19-Hex, Fuco, Zea, and Chl*a* for the world ocean, which were then integrated over the euphotic zone. For each pixel, the resulting monthly 19-Hex, Fuco, and Zea integrated contents were converted into Chl*a* equivalents using the appropriate pigment to Chl*a* ratios determined by multiple regression analysis performed on the global pigment database (Uitz et al. 2006). The obtained monthly Chl*a* biomasses attributed to each group were averaged over the year to estimate annual mean values. These values were normalized to the annual mean euphotic layer-integrated Chl*a* content to determine the relative contribution (%) of each phytoplankton group to the total phytoplankton chlorophyll-based biomass (Fig. 3.7). Finally, for each of the three phytoplankton groups, an annual mean Chl*a* standing stock was calculated as the sum of the annually-averaged value of each pixel multiplied by the corresponding pixel surface area. Coastal areas (bathymetry <200 m), large lakes and inland seas were not considered in this analysis.

3.6. Tables

Table 3.1: Basic features of the samples analysed in this paper

Library name	Cruise	Station	Lat	Long	Depth (m)	Date	Vol seawater (l)	Temperature (°C)	Salinity (‰)	Fluorescence (mg/m^3)	# sequences retrieved
z11_11	ARCTIC	z11	72.5	19.6	5	8/25/2002	5	11.05	34.33	1.21	153
z61_43	ARCTIC	z61	76.32	7.98	25	8/29/2002	5	8.17	34.91	1.98	95
mv5_19	VANC 10MV	mv5	34.35	37.68	7	5/18/2003	15	20.81	35.75	0.22	144
mv5_21	VANC 10MV	mv5	34.35	37.68	85	5/18/2003	15	20.55	35.73	0.50	160
mv18_59	VANC 10MV	mv18	17.17	83.67	50	6/1/2003	15	25.03	35.04	0.24	122

Note that sea water was pre-filtered through 3µm membrane filters and collected onto 0.2 µm membrane filters.

Table 3.2: LSU rDNA total diversity estimates for each library using the Chao1 and ACE statistics and 2 divergence cutoffs

Sample	rDNA sequence diversity estimate			
	Unique rDNA		3% divergence cutoff	
	Chao1	ACE	Chao1	ACE
mv5_19	1099 (537 - 2401)	1412 (682 - 3076)	169 (114 - 288)	287 (158 - 595)
mv5_21	1098 (587 - 2183)	1408 (790 - 2606)	249 (158 - 446)	347 (210 - 622)
mv18_59	1147 (527 - 2664)	1084 (554 - 2230)	250 (154 - 463)	362 (220 - 642)
z11_11	509 (268 - 1062)	756 (388 - 1585)	88 (52 - 200)	90 (56 - 179)
z61_43	325 (157 - 769)	414 (198 - 956)	26 (19 - 62)	30 (20 - 66)

95% confidence intervals are given in parentheses.The Chao 1 statistics give a lower bound of species richness estimation, while ACE scores indicate point estimation of species richness (Chao and Shen 2003–2005).

Table3.3: Identification and origin of the haptophyte strains isolated, cultured, and characterised by electron microscopy and LSU rDNA D1-D2 sequencing, used in our study to anchor environmental genetic diversity.

Accession Number	Species	Culture Strain	Culture Collection	Isolation_Source
EU729452	*Platychrysis sp.*	RCC1385	Roscoff Culture Collection (RCC), France	Mediterranean - Spain
EU729451	*Platychrysis pienaarii*	RCC1392	RCC, France	unknown
EU729449	*Prymnesium sp.*	RCC1440	RCC, France	Mediterranean - Tunisia
EU729447	*Prymnesium zebrinum*	RCC1432	RCC, France	NAtlantic - France
EU729448	*Prymnesium zebrinum*	RCC1438	RCC, France	NAtlantic - France
EU729446	*Prymnesium sp.*	RCC1446	RCC, France	unknown
EU729445	*Prymnesium sp.*	RCC1443	RCC, France	NAtlantic - Spain
EU729444	*Prymnesium calathiferum*	CCMP707	Center for Culture of Marine Phytoplankton (CCMP), USA	North Island, New Zealand

EU729450	*Prymnesium sp.*	RCC1450	RCC, France	Mediterranean - Tunisia
AF289038	*Prymnesium patelliferum*	unknown	unknown	unknown
EU729443	*Prymnesium parvum*	RCC1434	RCC, France	NAtlantic - English Channel
EU729442	*Prymnesium sp.*	CCMP711	CCMP, USA	NAtlantic – Maine, USA
EU729441	*Chrysochromulina sp.*	RCC1184	RCC, France	NAtlantic - France
EU729458	*Platychrysis pigra*	RCC1390	RCC, France	Mediterranean - France
EU729457	*Imantonia rotunda*	RCC1343	RCC, France	NAtlantic - France
EU729456	*Chrysochromulina brevifilum*	S-3	Algobank Culture Collection, France	NAtlantic - Spain
EU729455	*Chrysochromulina ericina*	CCMP283	CCMP, USA	NAltantic, Gulf of Maine, USA
EU729454	*Chrysochromulina hirta*	S-17	Algobank Culture Collection, France	NAtlantic - Spain
EU729453	*Chrysochromulina cf herdlensis*	CCMP284	CCMP, USA	49.87N 142.67W
EU729440	*Chrysochromulina camella*	CCMP289	CCMP, USA	29.97N 63.86W
EU729439	*Chrysochromulina camella*	RCC1185	RCC, France	NAtlantic - France
EU729438	*Chrysochromulina sp.*	S-14	Algobank Culture Collection, France	NAtlantic - Spain
EU729437	*Chrysochromulina acantha*	S-6	Algobank Culture Collection, France	NAtlantic - Spain
EU729436	*Chrysochromulina throndsenii*	S-5	Algobank Culture Collection, France	NAtlantic - Spain
EU729435	*Chrysochromulina sp.*	No code available	RCC, France	unknown
EU729434	*Chrysochromulina simplex*	RCC1193	RCC, France	NAtlantic - Spain
DQ980469	*Chrysochromulina sp.*	NIES 1333	NIES Collection, Japan	Pacific - Japan
EU729460	*Calcidiscus sp.*	RCC1157	RCC, France	Mediterranean - Spain
EU502878	*Calcidiscus sp.*	RCC1147	RCC, France	SAtlantic, Namibia
EU729463	*Umbilicosphaera hulburtiana*	RCC1474	RCC, France	SAtlantic, South Africa
EU729461	*Umbilicosphaera sibogae*	RCC1468	RCC, France	Mediterranean - Spain
EU729462	*Umbilicosphaera foliosa*	RCC1470	RCC, France	NAtlantic - Puerto Rico
EU729464	*Coccolithus braarudii*	AC613	Algobank Culture Collection, France	NAtlantic - English Channel
EU502875	*Jomonlithus litoralis*	RCC1354	RCC, France	Mediterranean - Spain
EU502872	*Hymenomonas globosa*	RCC1338	RCC, France	NAtlantic - English Channel
EU729469	*Ochrosphaera neapolitana*	RCC1359	RCC, France	NAtlantic - English Channel
EU729468	*Pleurochrysis dentata*	RCC1400	RCC, France	New Mexico - USA
EU729467	*Stauronertha neohelis*	RCC1206	RCC, France	NAtlantic - Guadeloupe
EU729466	*Calyptrosphaera sphaeroidea*	RCC1178	RCC, France	North Sea - Norway
EU729465	*Helladosphaera sp.*	RCC1182	RCC, France	Pacific - Japan
EU502879	*Syracosphaera pulchra*	RCC1460	RCC, France	Mediterranean - Spain
EU729471	*Coronosphaera mediterranea*	RCC1204	RCC, France	SAtlantic – South Africa
EU729473	*Helicosphaera carteri*	RCC1333	RCC, France	SAtlantic – South Africa
EU729472	*Scyphosphaera apsteinii*	RCC1455	RCC, France	Mediterranean - Spain

EU729470	*Algirosphaera robusta*	RCC1128	RCC, France	Mediterranean - Spain
EU729476	*Gephyrocapsa oceanica*	RCC1289	RCC, France	Mediterranean - Spain
EU729475	*Dicrateria sp.*	RCC1207	RCC, France	Mediterranean - Morocco
EU729474	*Isochrysis galbana*	RCC1348	RCC, France	NAtlantic - Irish Sea
EU729459	*Phaeocystis cordata*	CCMP 2495	CCMP, USA	Mediterranean - Italy
AF289040	*Phaeocystis antarctica*	unknown	unknown	unknown
EU502882	*Phaeocystis sp.*	AC618	Algobank Culture Collection, France	NAtlantic - English Channel
EU729479	*Exanthemachrysis gayraliae*	RCC1523	RCC, France	NAtlantic - English Channel
EU729478	*Rebecca salina*	RCC1545	RCC, France	NAtlantic - English Channel
EU729477	*Pavlova virescens*	RCC1535	RCC, France	NAtlantic - France
EU502883	*Pavlova pinguis*	RCC1538	RCC, France	Mediterranean - France

The strains are listed following the branching pattern of the tree in Figure 3.3A (from top to bottom external black branches). All sequences except DQ980469, AF289038, and AF289040 were generated during this study. Note *Stauronertha* is the new genera name for *Cruciplacolithus* (Aubry and Bord 2009).

Table 3.4: Time, space, and depth information for the 28 worldwide samples used to measure total haptophyte cell size (Fig. 3.8).

Station	Location	Date	Depth (m)
Roscoff (France), SOMLIT-Astan.	48°46′ N, 3°57′W	May to July 2006 & January, April, June, October 2007	Sub-surface (n=6)
Japan, Station A	40°N, 143°E	May 2006	10, 30
Japan, Station B	40°N, 145°E	May 2006	5, 20, 30, 50, 90
Japan, Station E	34°04′N, 140°E	May 2006	10, 25, 40
Japan, Station F	34°26′N, 139°E	May 2006	subsurface, 30
Japan, Station G	33°21′N, 140°E	May 2006	subsurface, 20, 70
Villefranche sur mer (France), SOMLIT Point B.	43°41’N, 7°19′E	September 2007	subsurface, 20, 40, 50, 70, 150, 200

For each depth, water was prefiltered through a 60µm sieve, and planktonic cells were recovered onto 0.2 µm membranes as in (Frada et al. 2006).

Table 3.5: *SSU* versus *LSU* (D1-D2 fragment) rDNA Tajima & Nei genetic distances for several couples of haptophytes species.

Compared cultured strains	SSU rDNA distance	LSU rDNA distance
Emiliania huxleyi / Gephyrocapsa oceanica	0,0%	0,1%
Calcidiscus quadriperforatus / Calcidiscus leptoporus	0,2%	3,1%
Umbillcosphaera foliosa / Umbilicosphera sibogae	0,4%	4,1%
Helicosphera carteri / Scyphosphaera apsteinii	1,2%	1,7%
Syracosphera pulchra / Coronosphaera mediterrancea	1,2%	2,2%

Pleurochrysis carterae / Pleurochrysis carterae var. dentata	1,3%	2,4%
Jomonlithus litoralis / Hymenomonas coronata	1,8%	3,9%
Chrysochromulina acantha / Chrysochromulina throndsenii	0,1%	0,9%
Chrysochromulina ericina / Chrysochromulina simplex	5,7%	10,9%
Chrysochromulina hirta / Chrysochromulina brevifilum	1,1%	4,2%
Prymnesium zebrinum / Prymnesium parvum	1,9%	5,5%
Pavlova pinguis / Pavlova virescens	5,4%	11,7%

Pairs of sequences were automatically aligned using ClustalX and the program Mega 4.0 (Tamura et al. 2007) was used to compute genetic distances. As a mean value, LSU rDNA evolves ~5 times faster than SSU rDNA. Note that closely related species which split in the Pleistocene, such as *E. huxleyi* and *G. oceanica*, cannot be separated using SSU rDNA sequences. In addition, no intraspecific variability was detected between the several LSU rDNA clones we sequenced from cultured strains.

3.7. Figures

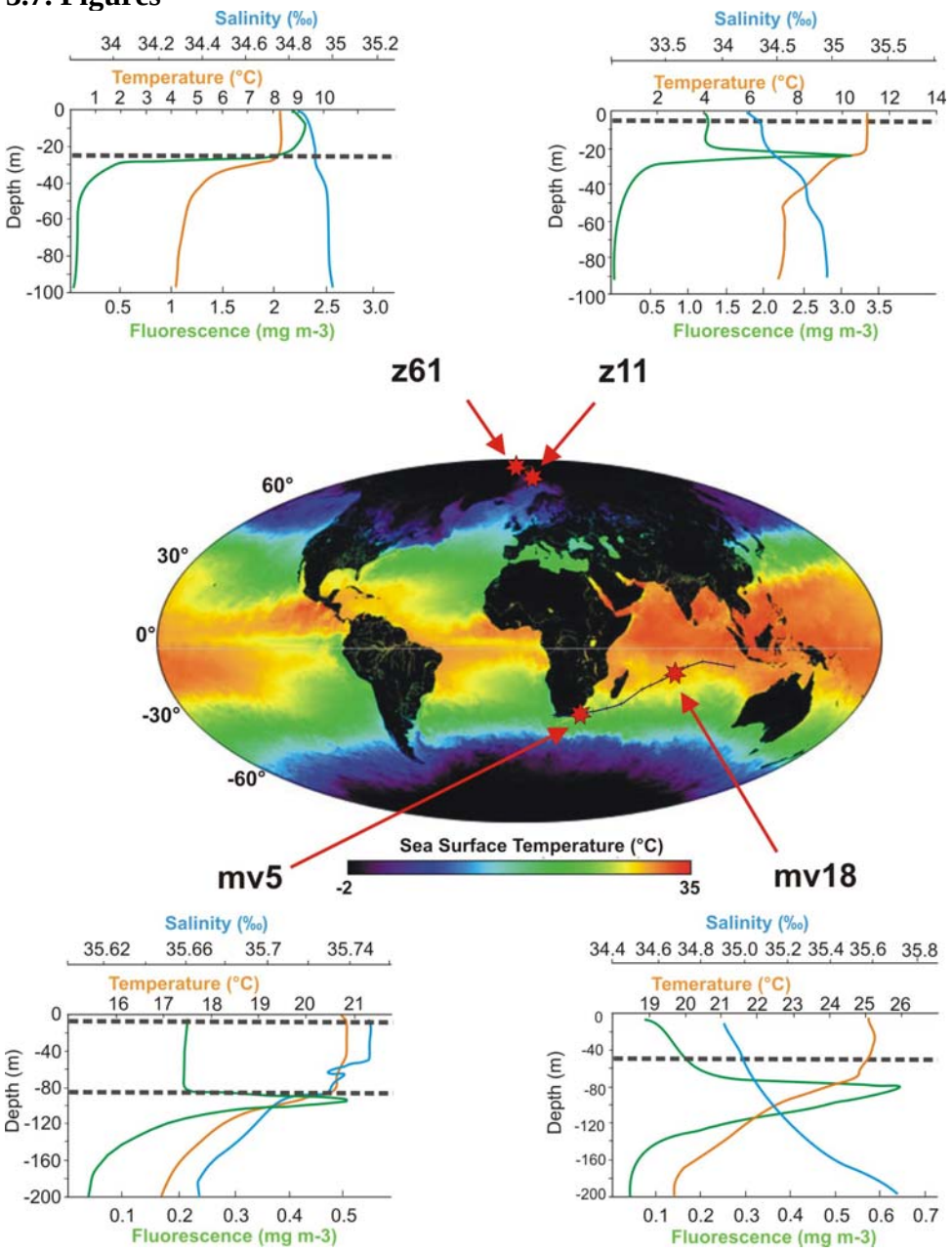


Figure 3.1: Cruise tracks and sampling locations. The 4 stations are marked with red stars. Temperature, salinity, and fluorescence profiles down to 100 (z) or 200m (mv) depth are given for each station. Dotted lines indicate the depths at which water used for DNA extraction and rDNA sequencing was sampled with Niskin bottles. Global sea surface temperature corresponds to a monthly (May 2001) composite of data captured by the satellite *MODIS* (**http://modis.gsfc.nasa.gov/**). Further details showed in Table 3.1 above.

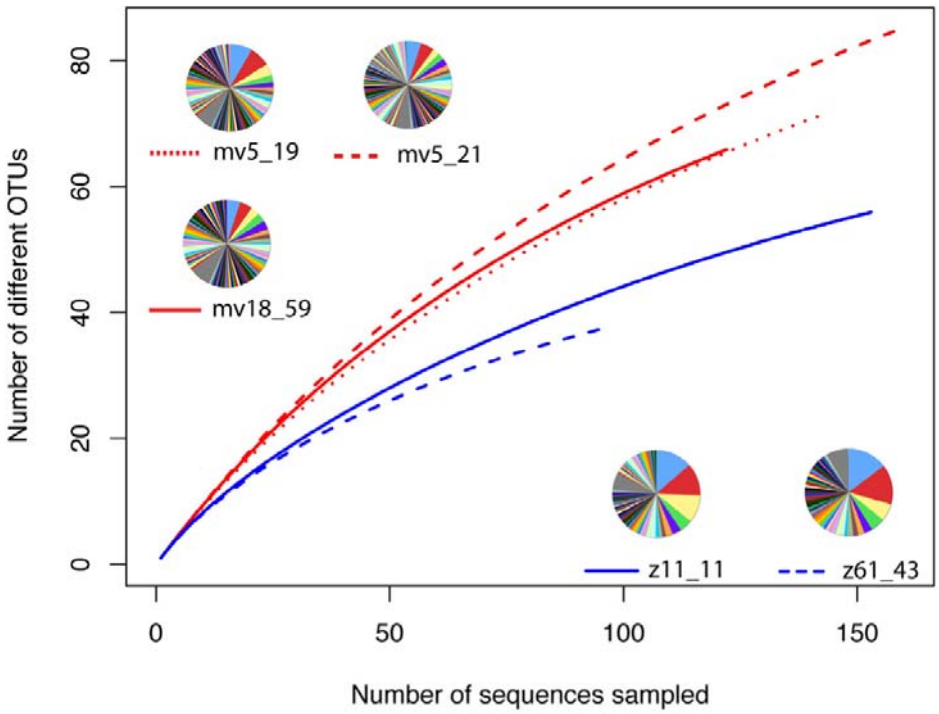


Figure3.2: Rarefaction analysis for each environmental clone library based on unique 28S rDNA sequences (OTUs). 72, 85, 65, and 56, 37 OTUs were respectively obtained from the Indian ocean (*Mv* 19, 21, 18) and subarctic (*z* 11, 43) clone libraries (Fig. 3.1) The pie charts show, for each library, the amount of identical sequences in each retrieved OTU.

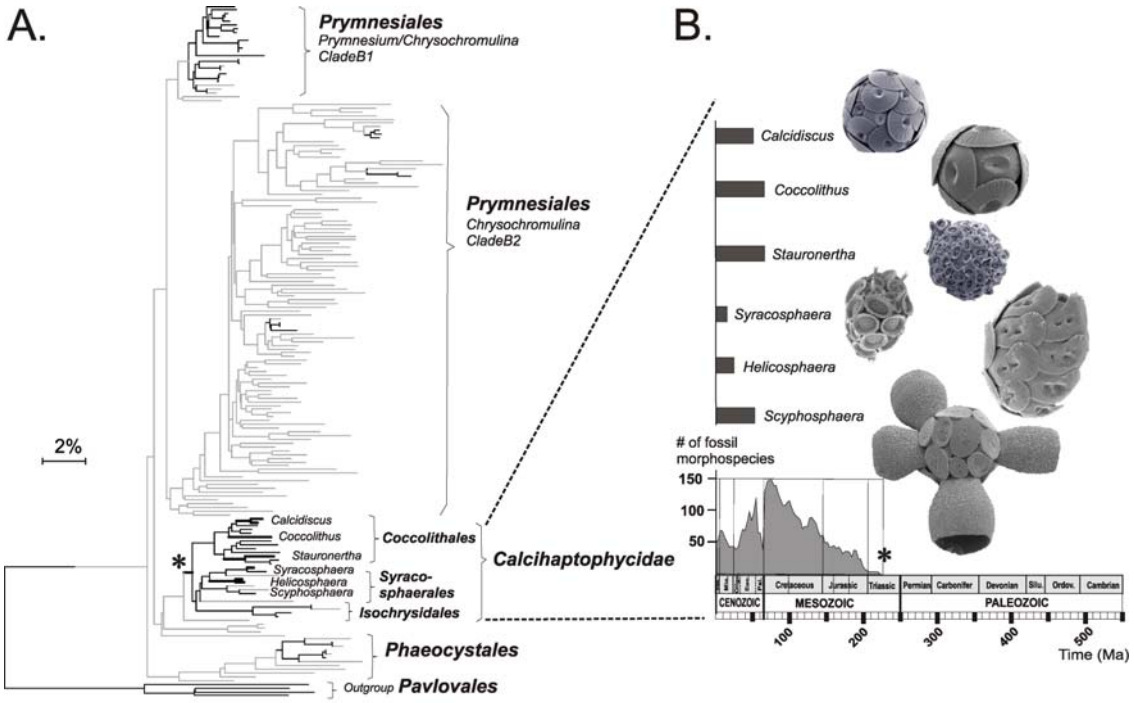


Figure3.3: Phylogenetic assessment of the previously undescribed haptophyte environmental diversity. (A). *LSU* rDNA tree (5% divergence cutoff) including environmental sequences (grey branches) and a taxonomic cross-section of cultured haptophyte taxa (black branches, see Table 3.3 for species identification). (B). Focus on the stratigraphic ranges (black rectangles) of key genera within the calcifying haptophytes (de Vargas et al. 2007) (thick black branches in the tree in A, and SEM images in B). The coccolithophore fossil record (Bown 2005b) (lower right) represents number of fossil morpho-species along time in million years. Black clover symbols indicate the origin of haptophyte calcification ~220 Ma. Note that *Stauronertha* is the new genus name for *Cruciplacolithus* (Aubry and Bord, 2009).

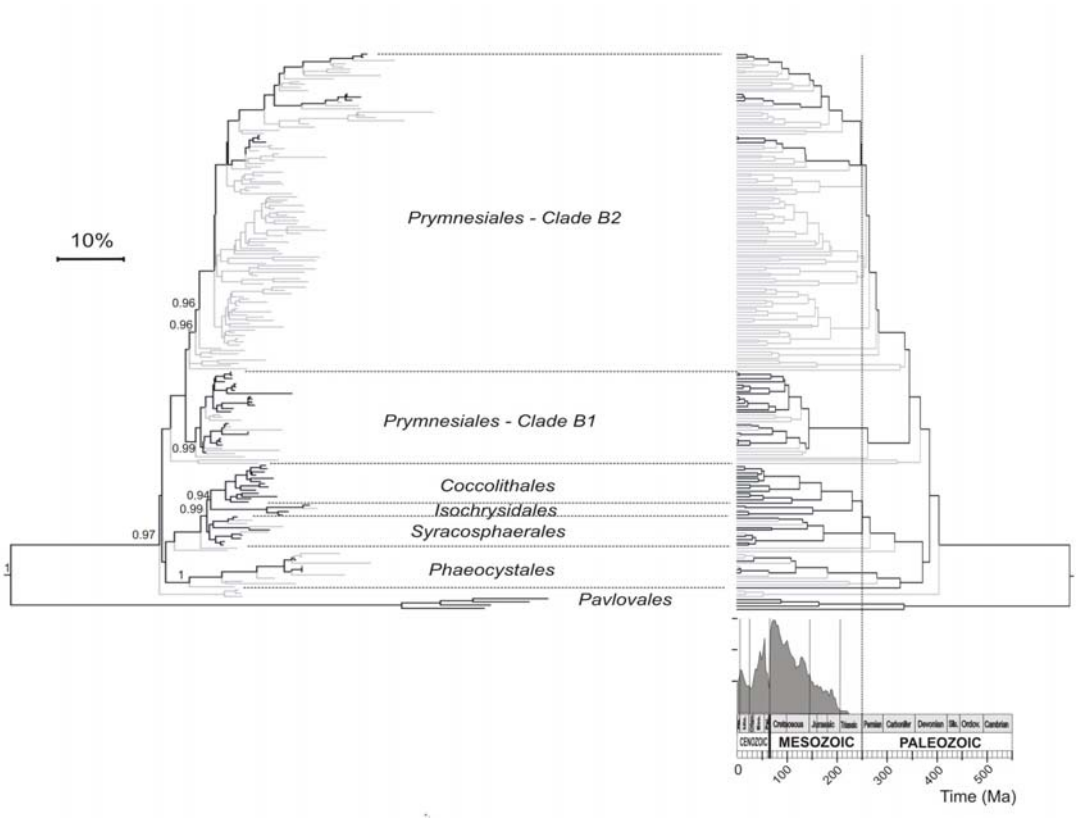


Figure 3.4: Diversification of the haptophytes along geological time. Relaxed clock analysis calibrated on the coccolithophore fossil record. The tree on the left shows the pattern of diversification; numbers represent Bayesian posterior probabilities of key divergences only. The tree on the right shows the corresponding divergence times. Branches are colour coded: black for sequences obtained from cultured and previously described haptophytes; grey for environmental and previously unknown sequences.

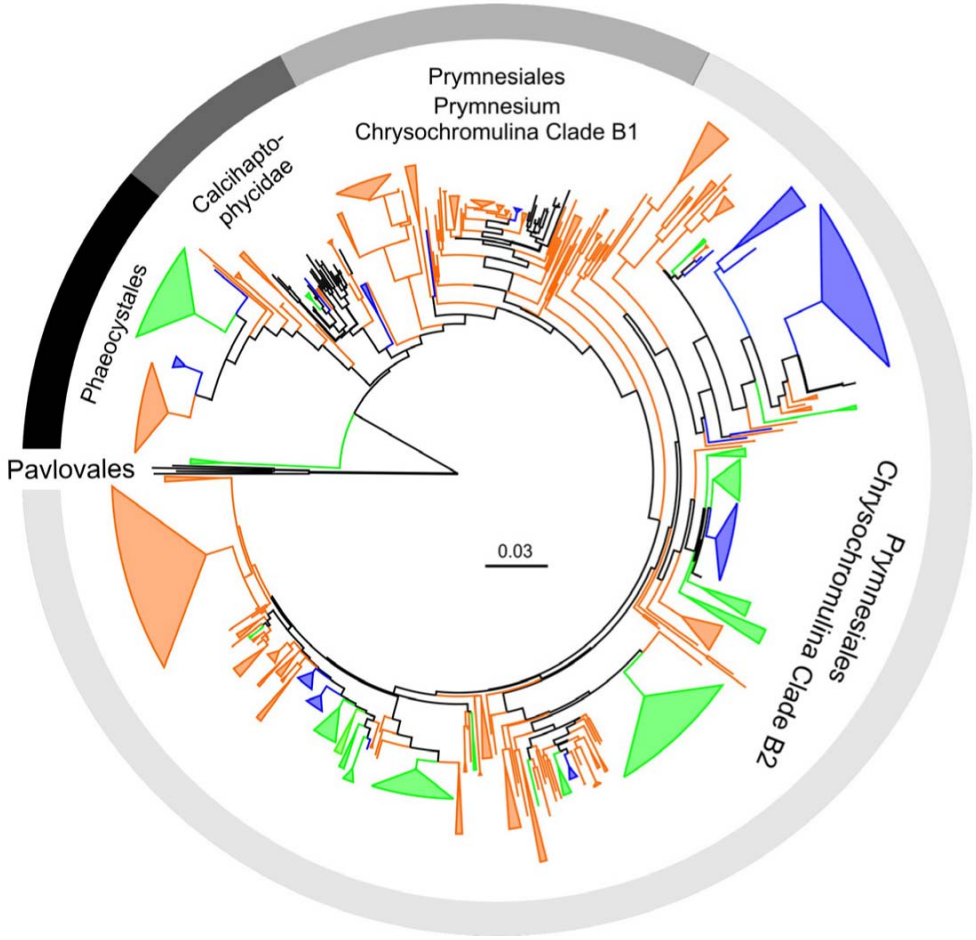


Figure 3.5: Biogeographic partitioning of environmental tiny haptophyte diversity. This Maximum Likelihood tree contains all 674 environmental LSU rDNA sequences, with clustering above 97% similarity. Colour code: orange, subtropical; blue, subpolar; green, both subpolar and subtropical; black external branches, taxonomically-defined sequences from cultured haptophyte strains. Colour code applies to internal branches when they are part of strictly subtropical or subpolar monophyletic groups.

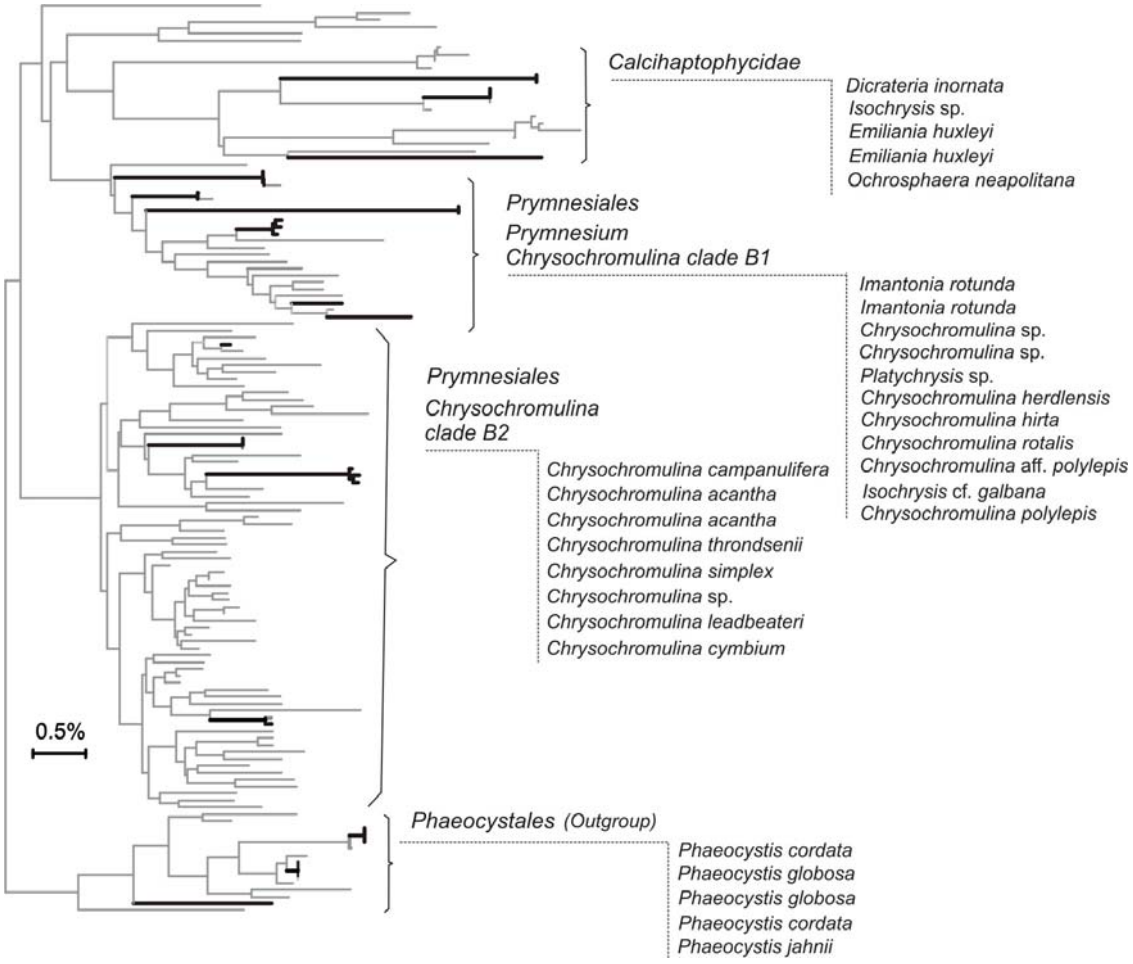


Figure3.6: Chloroplastic 'view' of eukaryotic and haptophyte diversity from the Gulf of Naples, Mediterranean sea. McDonald and collaborators (McDonald et al. 2007) used chloroplastic-biased 16S rDNA primers to explore 6 environmental clone libraries over an annual cycle. 46% of the retrieved unique eukaryotic sequences and 73% of the total eukaryotic OTUs belonged to the Haptophyta. This overwhelming haptophyte biodiversity is reanalyzed here using the Neighbor-joining phylogenetic method based on a Tajima-Nei distance matrix. Grey branches represent the unveiled environmental diversity, integrated into taxonomically-known 16S rDNA data (black branches). Note that several taxonomic inconsistencies were removed as compared to the original dataset presented in (McDonald et al. 2007).

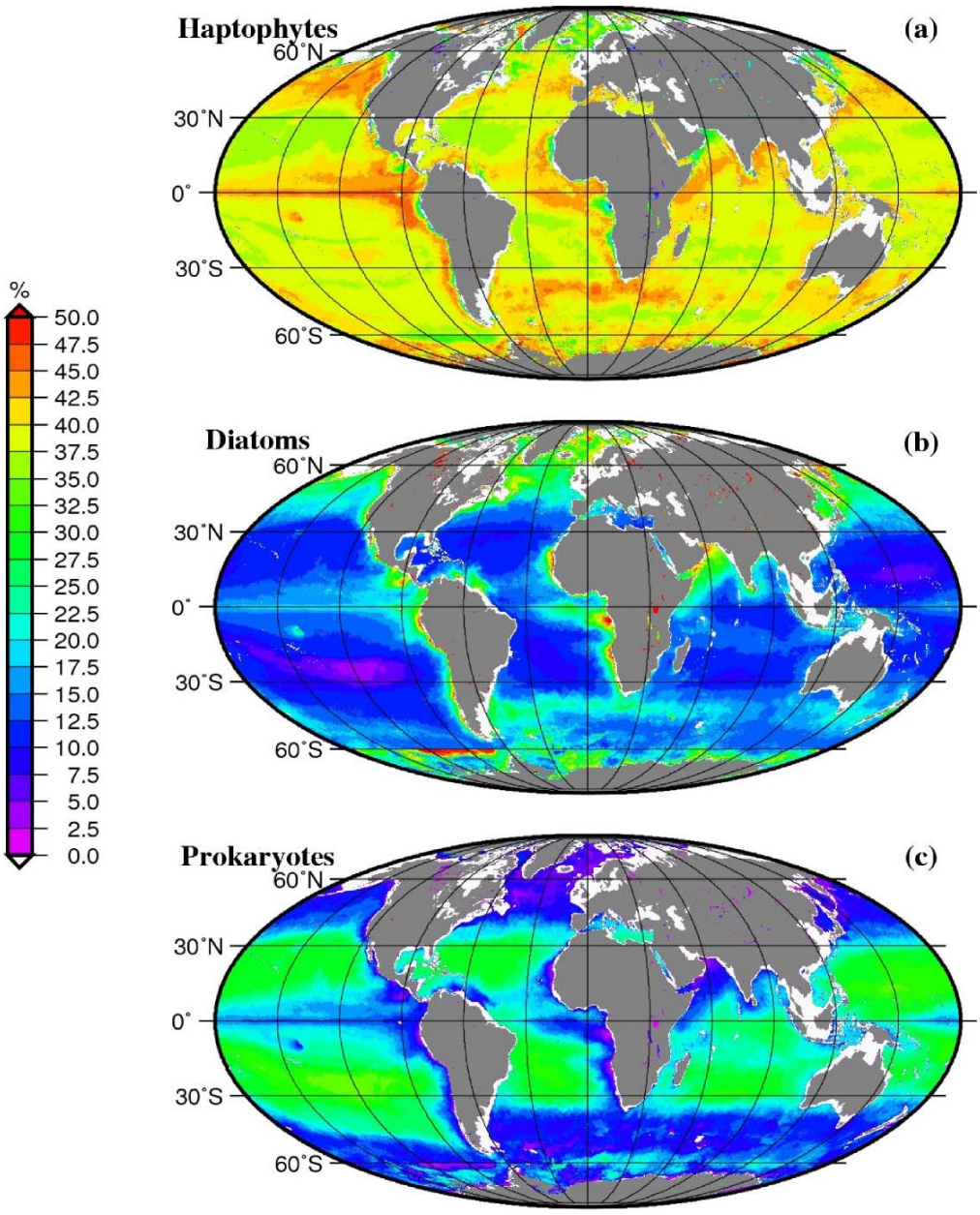


Figure3.7: Accessory pigments based relative contribution of (A) haptophytes, (B) diatoms, and (C) photosynthetic prokaryotes to total chlorophyll-*a* biomass in the photic layer of the world ocean over the year 2000. The average yearly standing stocks associated with these three groups are respectively 2.5×10^9, 1.3×10^9, and 1.1×10^9 kg Chl*a*. See methods sections for details of the calculation.

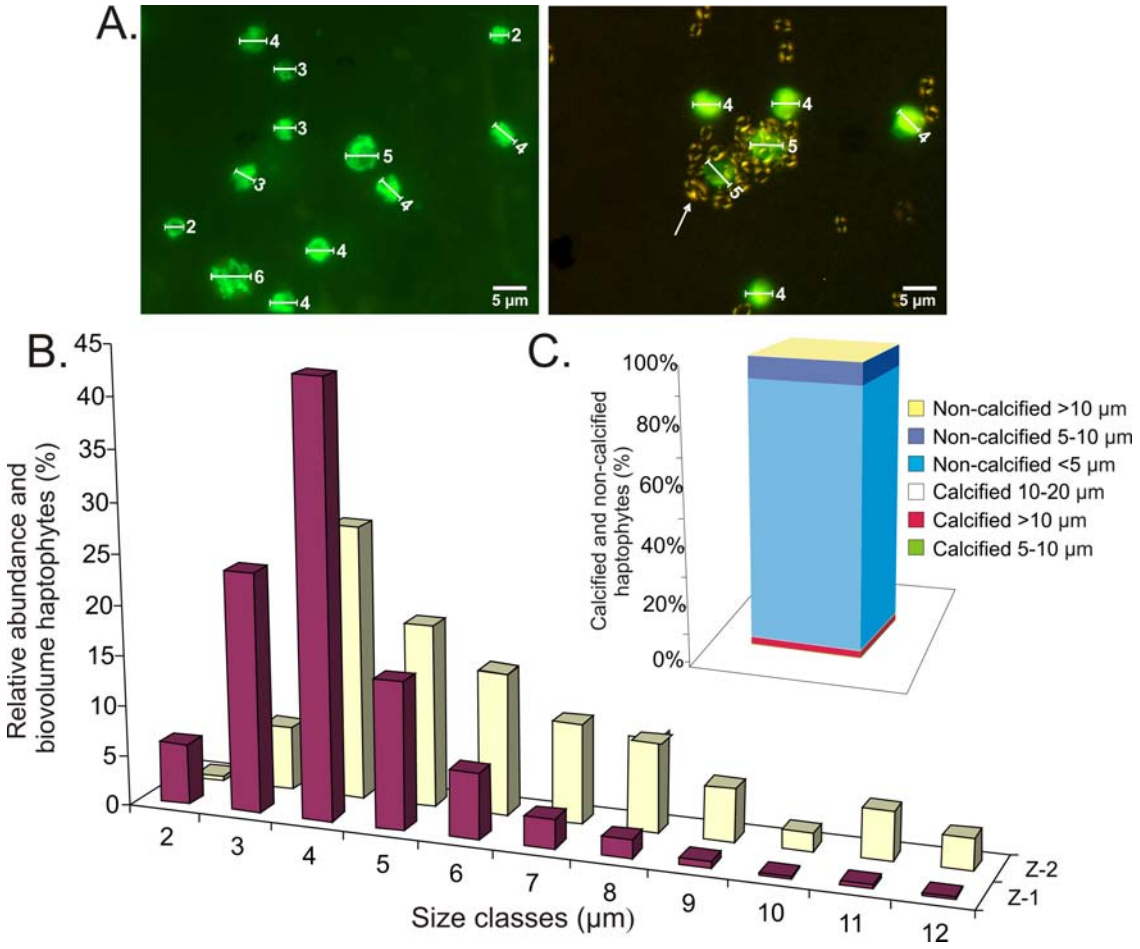


Figure3.8: Abundance, size, and biovolume in non-calcifying and calcifying haptophytes. (A). Haptophyte cells from 28 plankton samples from various depths in North Pacific, Mediterranean Sea, and North Atlantic waters (see Table 3.4) were CODFISHed (CaCO3 optical detection with haptophyte-specific fluorescent in situ hybridization) (Frada et al. 2006) and cell diameters were measured from 548 individuals. The white arrow in the right panel points to a single CaCO3 coccolith displaying typical light polarization pattern and allowing the detection of calcifying versus non-calcifying cells. The microscopy field shown in the left panel displays 12 non-calcifying cells. (B). Relative abundance (Z-1) and relative biovolume (Z-2, estimated as a sphere [4/3*π*r3]) of *non-calcifying* haptophytes in various size-classes. **C**. Relative abundance of different size fractions of non-calcifying and calcifying haptophytes. Note that the extensive diversity of LSU rDNA types reported herein and recovered from <3µm filtrates, may in fact partly come from larger, nanoplanktonic cells disrupted by the vacuum pumping filtration process.

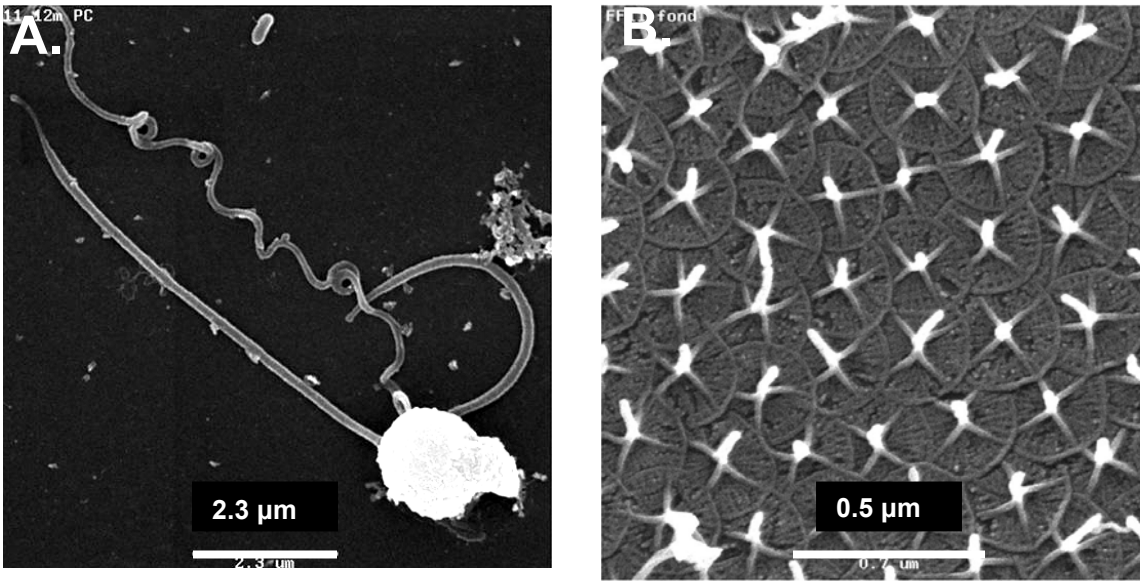


Figure 3.9: Tiny Chrysochromulina: haptonema and scales. (A) Whole-mounts transmission electron microscopy picture of a tiny *Chrysochromulina* sp. collected from 24m depth in the Bay of Banyuls/Mer, France, on May 15th, 2001 (personal communication, M-J Chrétiennot-Dinet). Such unidentified species from the genus *Chrysochromulina* are very common and diverse in oligotrophic water but do not grow in current culture media. White arrow indicates the *haptonema*, which can be used to capture bacteria; bacterial ingestion is commonly observed in tiny haptophytes from both culture and field samples (refs. 16, 25, 26, 29 from the core paper). (**B)** Typical organic scales covering the surface of cells within the genus *Chrysochromulina* (here the species *C. ephippium)*. Both images were graciously contributed from M-J Chrétiennot-Dinet.

Chapter 4

4.0. Morphospecies versus phylospecies concepts in marine phytoplankton: the case of the coccolithophores

4.1. Abstract

Genetic approaches to exploring *in situ* marine phytoplankton assemblages have unveiled, over the last decade, previously unknown biodiversity at different taxonomic levels. However, these new data typically measure biodiversity in terms of the phylogenetic concept of species, without reference to other species concepts. In particular, what makes a phylospecies has never been assessed against parallel morphological analyses, upon which most of the current ecological, physiological, and paleontological knowledge on oceanic phytoplankton relies. Here we use the coccolithophores as a case study to test the relationship between these two species concepts and evaluate the diversity level at which phylospecies and morphospecies can be considered equivalent concepts. By analyzing 217 coccolithophore *LSU* rDNA sequences and 729 specimens for morphological (light and electron microscopy) characters obtained from three water samples from the Atlantic Ocean, Pacific Ocean and the Mediterranean Sea, we show that a parallel analysis of morphology and genetic data overcome several limitations inherent to both methods as such a comparison more precisely describes the composition, richness and structure of natural coccolithophore communities. We compared the genetic and morphological diversity within six coccolithophore sub-groups (family or order level) and within each sampling location. We show that the genetic variability within established morphospecies varies significantly between different environments. Critically, we find that the threshold at which phylospecies and morphospecies are defined varies across different natural

communities, which has severe implications with respect to our evaluation of diversity as estimated from metagenomics approaches.

4.2. Introduction

Oceanic photosynthetic protists (phytoplankton) play critical roles in marine biogeochemistry by dominating primary production (Field et al. 1998; Liu et al. 2009), exporting significant amounts of organic and inorganic carbon to the deep sea through the biological pump (Dugdale and Goering 1967; Eppley et al. 1979), and structuring marine food webs (Ryther 1969). However, the taxonomy of phytoplankton, on which most of the current knowledge on the ecology, physiology, and paleontology of the different groups relies, is based on morphological characters. Over the last decade, the emerging field of metagenomics has benefited from the revolution in molecular ecology used the tools of PCR-based phylogenetics and whole-genome shotgun sequencing to analyze and unveil previously unsuspected levels of environmental diversity in marine microbes (Biers et al. 2009; DeLong et al. 2006; Venter et al. 2004). The initial studies, limited to prokaryotes (Chisholm et al. 1988; Giovannoni et al. 1990; Rappe et al. 1998), were soon extended to oceanic protists, mainly those from the pico-planktonic size fraction (cells <2-3 μm e.g., Diesz et al. 2001; Moon-Van Der Staay et al, 2000). The new tools revealed hundreds of previously undocumented rDNA phylotypes or ribotypes that can eventually be characterized as phylospecies (Huber et al. 2007; Queiroz and Donoghue 1988).

However, this promising molecular view of oceanic protistan biodiversity remains largely ambiguous because of both the lack of links to traditional diversity analyses that rely on the concept of morphospecies (Finlay 2004), and the important biases inherent to

PCR-based amplification of rDNA or other genetic markers (Acinas et al. 2005; Hugenholtz and Huber 2003). First, filter samples used for construction of clone libraries are often contaminated by larger, morphological known cells that release their genetic materials through gametes/swarmers or cell debris to the water column. Therefore, many if not most of these picoplantonic ribotypes may correspond to these larger cells when a set of universal eukaryotic primers are used. Currently, this problem cannot be resolved because of the relative absence of clone library surveys of larger cell-size fractions (nano-, micro-, and macro-plankton). Second, within any given size-fraction, the molecular diversity will be biased by three confounding factors: (i) the nature of the genetic marker, (ii) the techniques used to extract genetic material from total environmental DNA and (iii) the formation of artificial chimeras. The nature of the genetic marker affects ribotype diversity because ribosomal genes are present in multiple copies within a given species, as demonstrated in several protistan groups (Alverson and Kolnick 2005; Darling et al. 2007; Pawlowski et al. 2007). In addition to this first factor of bias, standard eukaryotic rDNA PCR amplification is biased toward short and/or GC poor genes with secondary structures especially amenable to oligonucleotide priming and polymerase extension. This problem, initially revealed in bacteria (Polz and Cavanaugh 1998; Suzuki and Giovannoni 1996), may in fact be much worse in eukaryotes, whose rDNA varies greatly in length and GC content. For example, in foraminifers, the *SSU* rDNA gene can be 3–5 times longer than in most other eukaryotes currently available in GenBank, and is thus inaccessible when using standard PCR protocols. Despite their importance in both the planktonic and benthic marine realms, foraminifers are virtually absent from all environmental surveys of these environments (Pawlowski 2000; Stoeck et al. 2006). Such PCR primer bias can

theoretically be reduced by increasing sequencing depth or by using multiple sets of primers with various levels of specificity (Stoeck et al. 2006), but the extent of this bias has not been well quantified, and it remains unclear whether these measures would prove effective (Jeon et al. 2008). On top of two sources of potential biased evaluation of ribotype diversity comes the formation of artificial chimeras, which are genes made of fragments from the genomes of different species, during PCR amplification. Highly conserved regions within rDNA sequences can anneal even between sequences from distantly related organisms and chimeras can therefore represent up to 32% of environmental sequences (Berney et al. 2004; Hugenholtz and Huber 2003; Robison-Cox et al. 1995; Wang and Wang 1997). It is relatively easy to detect chimeras consisting of large fragments from widely divergent species using methods such as alignment to reference sequences (Cole et al. 2003) or partial tree building (Wang and Wang 1997). However, it is much more challenging to detect micro-chimeric patterns between related species, genera, or families, and such amplification biases could significantly and artificially increase the amount of phylospecies amplified from natural populations (Speksnijder et al. 2001).

As a consequence of all these issues, a major gap is growing between genetic and morphological surveys of marine protistan biodiversity. The genetic approach is revealing novel but potentially largely artificial biodiversity at an increasingly fast rate, while standard morphological analyses are probably too conservative, lumping together cryptic species into single categories and potentially missing important part of the diversity within groups displaying poor phenotypic differentiation. The difficulty to educate and recruit young taxonomists prevents transmission of expertise between generations (the so called

'taxonomic impediment', (Wheeler et al. 2004) and will dangerously increase the gap between the genetic and morphological species concepts. On the other hand, we posit that combined morpho-genetic surveys will allow better interpretation of diversity patterns in their ecological (and potentially functional) context compared to the use of either method alone. Parallel morphological analysis provides a means to evaluate the efficiency of coverage of clone libraries, especially when the extent and the potential causes of bias in PCR amplification are poorly understood, and potentially to link genotypic and phenotypic data. The reverse is also true, clone libraries providing data with which to assess issues related to morphological analyses, such as cryptic speciation, poor preservation of the material, or the presence of various life cycle stages of unknown taxonomic status.

Here we present a case study where we compare morphological and genetic approaches to assessing species-level genotypic, phenotypic, and ecological differentiation in an ecologically important group of phytoplankton, the coccolithophores. Coccolithophores represent an ideal group for such morphogenetic inter-calibration. They are abundant and ecologically relevant throughout the world's oceans. The calcified platelets (coccoliths) produced by these organisms exhibit a rich suite of morphological characteristics that can be observed by conventional electron and light microscopy techniques. Their extant diversity, as described by classical morphology-based taxonomy, is rather limited compared to other important groups of phytoplankton, making group-wide analysis feasible. In addition, there is a reasonable coverage of cultured species (Probert & Houdan 2004) that have been used for large-scale phylogenetic studies (e.g., Liu et al. 2009; Medlin et al. 2008) thus facilitating the anchoring of environmental diversity to known morphospecies. However, the morphological view of coccolithophore biodiversity is

limited by several potential problems that include cryptic and pseudo-cryptic speciation (Geisen et al. 2004; Saez et al. 2003), the dissolution of small or delicate coccoliths during sample preparation, dimorphic haplo-diplontic life cycles (Billard 1994; Houdan et al. 2004), and the hidden 'naked' (i.e. non-calcifying) or poorly calcified species that are practically inaccessible to observation-based identification (see also (de Vargas and Probert 2004; Young et al. 2005). We collected samples for a parallel morphological and genetic analysis from three geographically distinct locations in the western Mediterranean Sea, the Atlantic and the Pacific oceans. We report the first extensive clone library dataset focusing on coccolithophores and, by assessing the inherent biases in morphological and genetic approaches, we demonstrate the advantages of combining morphology and genetics for an accurate assessment of protistan environmental biodiversity.

4.3. Materials and Methods

4.3.1. Samples locations and collection

Samples were collected from three geographically distinct mid-latitude oceanographic regions: the southeast Atlantic Ocean, the North Pacific gyre, and the Mediterranean Sea. Sampling took place during the Atlantic Meridional Transect (AMT) cruise 16 in May 2005 (sample AMT16_4.1), the Hawaii Ocean Time-Series (HOTS) cruise 169 also in May 2005 (sample HOT169_S2), and the *BOOM*-project survey of living coccolithophores conducted in the bay of Villefranche-sur-Mer (France) in September 2007 (sample MedEx-6), respectively (Fig. 4.1 and Table 4.1). At each station, water samples were collected using Niskin bottles from the three depths of the water-column: subsurface, intermediate mixed layer and deep chlorophyll maximum (DCM) waters. For

molecular analysis, up to 100 l water samples were concentrated by filtration through a nominal 5 μm nylon mesh net at the HOTS169_S2 and MedEx-6 stations; no such prefiltration step was undertaken for the AMT16_4.1 sample. The water then was filtered gently by a peristaltic pump (pressure <150mm Hg) through poly-ether sulphone filters (0.45 μm) for total DNA extraction. In parallel additional filter samples were prepared from the same water samples for morphological work, including for most samples both polycarbonate filters (0.4 μm) for scanning electronic microscopy (SEM); and cellulose nitrate filters (0.45 μm) for light microscopy (LM). DNA filters were kept dry frozen at –70°C until genomic DNA was extracted.

The pre-filtration of molecular samples from the HOTS169_S2 and MedEx-6 stations was carried out in order to concentrate coccolithophores relative to non-calcifying pico-haptophytes. The nominal 5μm mesh should have retained >90% of coccospheres, however subsequent LM measurement of cells from unfiltered and pre-filtered samples indicated that the effective filtration diameter was nearer 10μm than 5μm and so that there was significant biasing of the sampling toward larger coccosphere sizes.

4.3.2. Morphological and DNA data acquisition

Cross-polarized LM was used to establish the relative abundance of the major components of the assemblage and SEM was used to confirm LM identifications and for determination of smaller and rarer species, or ultrastructural details in morphologically distinct species. For LM filter segments were mounted on glass slides using Norland Optical Adhesive NOA74. This is a low-viscosity mounting medium that yields excellent optical results. Samples were enumerated on a Zeiss Axioplan photomicroscope at 1600X magnification. For SEM analysis, samples were mounted on aluminum SEM stubs, coated with a 20nm

gold-palladium layer and examined with a Phillips XL30 field emission electron microscope. Morpho-taxonomic concepts applied follow the recent synthesis of Young et al. (2003).

Total DNA was extracted from frozen filters using the DNeasy Plant Mini Kit (Qiagen) according to the manufacturer's instructions. Nuclear *LSU* rDNA fragments of ~950 bp containing the D1–D2 domains were PCR amplified using a forward *Haptophyta-specific* primer, Hapto_4 (5'-atggcgaatgaagcgggc-3'), and a reverse general eukaryote primer, Euk_34r (5'-gcatcgccagttctgcttacc-3'). PCR reactions (98°C for 30s, 50°C for 30s, and 72°C for 30s, with initial denaturation and final extension steps) were performed over a maximum of 30 cycles and using the Phusion high-fidelity PCR DNA Polymerase (New England BioLabs) specifically suited for amplification of GC-rich DNA, in order to limit formation of chimeric sequences. PCR products were purified using the MinElute gel extraction kit (Qiagen), and 3'-A-overhangs were bound to DNA fragments by adding 0.2mM dATP, 1 unit of Taq DNA polymerase, and 1X Taq DNA polymerase buffer to the purified PCR product and incubating the mixture for 20 min at 72°C. Classical TA-cloning into OneShot DH5α-T1 competent bacteria using the TOPO TA kit (Invitrogen) was then performed according to the manufacturer's instructions. Clone libraries were checked by PCR using the M13 forward and reverse primers included in the kit and sequencing of ~25–35 random clones in both directions. The entire process of library construction was repeated until >85% of white colonies yielded high-quality sequences. Libraries then were sent to High-Throughput Sequencing Solutions (www.htseq.org) for random automatic picking of 200 clones, plasmid minipreps, and automatic sequencing of both strands of ~150 *LSU* rDNA fragments per library. All sequences obtained in this study were

deposited in *GenBank* under accession numbers EU729435-EU729479, EU502872–EU502882 and FJ696920–FU696921 for culture sequences and FJ787731–FJ788096 for environmental sequences.

4.3.3. DNA sequences analysis

LSU rDNA sequences were checked for potential chimeras with the Check_Chimera program (Cole et al. 2003). All novel sequences passing this first screen were re-checked manually in multiple sequence alignment and in partial phylogenetic trees to remove putative micro-chimeras, that is, sequences containing segments from two or more closely related species. Despite the experimental cautiousness to avoid the formation of chimeric PCR products described above, our stringent approach to detecting chimera identified ~15-20% of the sequences as such. These putative chimeric sequences were removed from calculation of genetic biodiversity. All remaining sequences were manually aligned with 33 taxonomically known sequences using the *Genetic Data Environmen*t (GDE) 2.2 software (Larsen et al. 1993). Phylogenetic analyses were subsequently performed using neighbor-joining (NJ; Saitou and Nei 1987) with MEGA (Tamura et al. 2007) and Phylo_win (Galtier et al. 1996), maximum likelihood with PhyML (Guindon et al. 2005) and Bayesian approaches with BEAST (Drummond and Rambaut 2007). For maximum likelihood and Bayesian analyses, the Akaike Information Criterion (AIC) implemented in ModelTest (Posada and Crandall 1998) was used to determine the most appropriate nucleotide substitution model. For NJ and ML analyses 1000 bootstrap replicates were generated to assess clade support values. For Bayesian analyses, two Markov chain Monte Carlo samplers were run, each of 100 million steps with thinning of 1000; convergence was checked with Tracer (http://tree.bio.ed.ac.uk/ software/ tracer/), which was also used to

determine that a burn-in period of 2 million steps was generally appropriate; a Perl in-house script was then used to combine post-burn-in tree files. Sequences were also clustered into operational taxonomic units (OTUs) at both unique and 3% nucleotide divergence levels with DOTUR (Schloss and Handelsman 2005), resulting in two additional datasets: one that contained only the 266 unique OTUs and one that contained only the 87 OTUs at the 3% difference level. Morpho-genetic diversity was compared in detail within six phylogenetic subgroups corresponding to classical order or family-level taxonomic divisions (Jordan et al. 2004; Young et al. 2003).

4.3.4. Estimation of morpho- and phylospecies richness

Rarefaction curves along with diversity indices and richness estimators were then calculated in order to assess the diversity found in both the morphological and genetic datasets. Rarefaction curves allow for the calculation of the expected number of species from a collection of random samples, so that they represent what is statistically expected from the accumulation curve. For the morphological data, rarefaction curves were produced by repeated random sampling of all identified morphospecies. Rarefaction curves and species richness estimators from morphological analyses were obtained using Proc IML in SAS software version 9.1 (SAS Institute, Inc.; script available upon request). For genetic sampling, AIC was used as above when constructing phylogenies to select the most appropriate model of nucleotide substitution. For each library, `PAUP*` 4.0b10 (Swofford 2002) was used to build pairwise maximum likelihood distance matrices. Each distance matrix was then analyzed with `DOTUR` assuming the furthest neighbor algorithm to cluster sequences, construct rarefaction curves and calculate the Chao1 (Chao 1984)

estimators and Shannon diversity index (Shannon 1948). Clustering levels ranged from 0 to 5% differences.

4.3.5. Comparison of the proportion of morphological and genetic diversity in each defined taxonomic sub-group

Both morphological and genetic surveys of the environmental diversity of coccolithophores have known limitations and biases and may not be quantitative. However this does not preclude the possibility that the proportion of DNA sequences and morphotypes observed by SEM in each defined taxonomic sub-group are homogeneous. To test this hypothesis, the Cochran-Mantel-Haenszel test was performed for each sampling location using two sets of data: 1) including all six defined taxonomic sub-groups, and 2) a fewer number of subgroups, where any group presenting apparent discrepancies between morphological and genetic samplings were excluded.

4.4. Results

4.4.1. Species richness

Morphological and genetic diversity, assessed here with the concepts of morphospecies and phylospecies, respectively, were estimated from each sample at each oceanic site. Morphological analyses recorded a total number of 22, 28, and 35 morphospecies out of 238, 191, and 300 observed individuals in the Mediterranean, Atlantic, and Pacific samples, respectively. Genetic analyses identified 45, 26, and 74 phylospecies, defined here as unique OTUs, out of a total of 75, 62, and 80 coccolithophore sequences retrieved from the Mediterranean, Atlantic, and Pacific samples. Table 4.3 lists the number of morphospecies and phylospecies obtained and their respective Shannon diversity indexes. Rarefaction curves were built for both morphological and genetic data (Fig. 4.2). At the

level of unique OTUs, the phylospecies rarefaction curves did not seem to reach a plateau with our current sequencing efforts, whereas all three morphospecies rarefaction curves showed a tendency towards saturation. According to Chao1/ACE estimators, the Mediterranean, Atlantic, and Pacific sites contained respectively 221/274, 121/112, and 399/493 total estimated unique phylospecies. The highest genetic and morphological diversities were observed in the Pacific ocean based on both rarefaction curves and the Shannon diversity index. The Atlantic site showed the least genetic diversity of the three locations, but a higher morphological diversity was observed in this sample compared to MedEx-6. Morphospecies rarefaction curves were then compared to phylospecies rarefaction curves based on different sequence similarity levels for each sample (Fig. 4.2). The morphological rarefaction curve for HOT169_S2 was closest to the genetic rarefaction curve at the 3% divergence cut-off. The rarefaction curve for the MedEx-6 morphological data was closest to the curve constructed from the genetic analysis at the >1% divergence cut-off level and the rarefaction curve for the AMT16_4.1 morphological analysis was found to be in between the unique and the 0-1% difference level.

4.4.2. Global coccolithophore phylospecies diversity

Of the 366 environmental haptophyta rDNA sequences retrieved, 266 unique OTUs were identified using DOTUR, of which 130 are coccolithophore OTUs. 33 taxonomically-defined sequences from cultured strains were aligned to the environmental coccolithophore sequences, allowing assessment of their phylogenetic position (Fig. 4.3). Although none of environmental phylospecies were strictly identical to any sequences from cultured coccolithophores, the majority could be classified into six clusters of calcified haptophytes corresponding to family or order levels: Noelaerhabdaceae (N=9),

Rhabdosphaeraceae (N=2), Coccolithales (N=6), Zygodiscales (N=15), Syracosphaeraceae (N=109), Umbellosphaeraceae (N=47). Only two groups of phylotypes could not be identified using reference sequences.

4.4.3. Comparative interpretation of morpho-genetic data by sub-group

We present our results and interpretations within six sub-groups, each corresponding to a family or an order according to current taxonomy. The vast majority of coccolithophore species recorded in this study fall into these six sub-groups, with a few exceptions (rare species) that were classified as 'others' and not included in the comparative interpretations. The number of OTUs at unique and 3% levels and the number of species identified from morphological sampling by sub-group are listed in table 4.4. The difference in frequencies was only found to be not significant between morphological and genetic samplings in MedEx-6 sample when problematic groups (i.e. the Neolaerhabdaceae and putative Umbellosphaeraceae, for which almost no sequences were retrieved but abundant individuals were observed in SEM) were excluded from construction of the contingency table (CMH test p=0.1686). This indicates the difficulty of correlating the frequency of retrieved DNA sequences and species abundance in the sample unless biases in methods could be reduced to a substantial level. We present detailed interpretation of morpho-genetic data for each sub-group below.

Group 1. Noelaerhabdaceae

The Noelaerhabdaceae, comprising the extant genera *Emiliania, Gephyrocapsa* and *Reticulofenestra* is the most abundant family of coccolithophores. They are distinguished from other coccolithophores by many characters, e.g., production of alkenones and having a motile non-calcifying haploid stage (de Vargas et al. 2006). Reflecting this, in almost all

molecular phylogenies they show a basal divergence from the other coccolithophores. Morphologically they show rapid species turnover in the Quaternary (e.g., Perch-Nielen 1985) and genetically they show very low divergences in both *SSU* and *LSU* rDNA genes.

The Noelaerhabdaceae clade can be unambiguously identified in the combined phylogeny (Figs. 4.3, 4.4, 4.7) since two sequences from HOT169_S2 were found to be genetically close to *Gephyrocapsa oceanica* or *Emiliania huxleyi* (note that the LSU rDNA sequences of these two closely related species differ by only one out of >900 base pairs), one of which presented 1% and the other 3% genetic distance. No Noelaerhabdaceae sequences were retrieved from the MedEx-6 sample. However, 25 *E.huxleyi* cells from HOT169_S2 and 71 cells from MedEx-6 were observed by SEM. Six sequences from AMT16_4.1 were identical or very close to *G.oceanica* or *E.huxleyi* with genetic distance smaller than 1% and 117 *E.huxleyi* cells were recorded by SEM. The anomalously low frequency of Noelaerhabdaceae sequences may in part reflect the use of >5µm pre-filtration on the HOT169_S2 and MedEx-6 samples which would have allowed virtually all of the *E.huxleyi (*~5µm cell size) and all of the *Gephyrocapsa ericsonii* (<5µm cell size) cells to pass through. Pre-filtration was not carried out on the AMT16_4.1 sample and this sample yielded significantly more Noelaerhabdaceae sequences. However, even in this sample the group is seriously under-represented in the clone library relative to the morphological analysis, suggesting that an additional factor is involved. This apparent discrepancy between morphological and genetic sampling in all three samples may be due to the high G+C content of the rDNA of Neolaerhabdaceae (~60%), which would be expected to reduce the efficiency of PCR amplification compared to other coccolithophore species with lower rDNA G+C content (~55%–59%).

Group 2. Rhabdosphaeraceae

The Rhabdosphaeraceae are a morphologically distinctive, moderately diverse (ca 22 extant species) family of coccolithophores showing highest abundances in oligotrophic waters (Young et al. 2003). To date, *Algirosphaera robusta* is the only species of the family that has been isolated into clonal laboratory culture (Probert et al. 2007), thus the identity of the unknown environmental sequences can only be established by their phylogenetic affinity with respect to this species. One sequence from MedEx-6 was similar to *A. robusta* at the 3% difference level. A more distant clade of four close sequences from the same sample was also observed at the 7% difference threshold from *A. robusta*. *A. robusta*, *Rhabdosphaera clavigera* and *R. stylifera* were observed by parallel morphological sampling. The clade of four very similar sequences (≤1%) from MedEx-6 (Fig. 4.5, 4.7, group2) is more likely be *Rhabdosphaera* species given the 7% difference threshold from *A. robusta*. Rhabosphaeraceae were quite abundant in the HOT169_S2 morphological sample. One sequence from the HOT169_S2 library exhibited a 2% difference from *A. robusta*. The more distant clade of two sequences from HOT169_S2 could be *R. clavigera* or *Discopshaera tubifera*, both of which were common in the sample; *R. clavigera* is perhaps more likely because 79 m is below the typical depth range for *D. tubifera* and the observed specimens were probably mainly sinking dead cells. No Rhabdosphaeraceae sequences were retrieved from the AMT16_4.1 clone library, but eight individuals were observed in the parallel morphological examination. The apparent discrepancy may reflect genetic under-sampling or PCR preferential bias.

Group 3. Coccolithales

The Coccolithales comprises the oceanic families Coccolithaceae, and Calcidiscaceae and neritic families Pleurochrysidaceae and Hymenomonadacaea. The group is predominantly mesotrophic. As a result, Coccolithales are relatively rare in the oligotrophic samples studied here, but they are well represented in culture collections and molecular phylogenies based on them.

Calcidiscus sequences were found in the MedEx-6 library and *Calcidiscus* coccospheres were observed in the corresponding morphological sample, which contained both *C. quadriperforatus* HOL and *C. leptoporus* HET. In the phylogenetic analysis, the two *Calcidiscus* sequences from MedEx-6 were identical to *C. leptoporus*. No *Umbilicosphaera sp.* specimens were observed in the MedEx-6 morphological sample, but *U. sibogae* and *U. hulburtiana* sequences occurred in the clone library, perhaps indicating that the (unknown) haploid stage was present in the water column. One sequence retrieved from the MedEx-6 library exhibited a 3% difference from *Calyptrosphaera sphaeroidea* or *Helladosphaera sp*.; besides, one *Helladosphaera pienarii* coccosphere was observed in the morphological sample, which suggests that this sequence might be *Helladosphaera pienarii*. The clade of four MedEx-6 sequences nesting within the Coccolithales but outside the Coccolithaceae and Calcidiscaceae is intriguing (bootstrap<0.70), since no obvious candidate species conventionally assigned to the Coccolithales were observed in morphological analysis. One possibility is that these are *Ceratolithus* since this enigmatic genus was common in the MedEx-6 sample and could conceivably fall almost anywhere in the coccolithophore tree. Almost no Coccolithales were found in the HOT169_S2 sample, and very few were found in the other upper water column samples from Hawaii; thus, the absence of any Coccolithales sequences in the corresponding clone library is to be

expected. Note, however, that parallel culture isolation from this sample resulted in the initiation of several cultures of *Calcidiscus* spp. and *Umbilicosphaera* spp...This result coincides with study carried out in exploring cyanobacterial mat communities (Jungblut, et al, 2005), where phylogenetic diversities retrieved by clone libraries from three ponds are not similar, yet, known culture strain sequences clustered together with clones were obtained from all the three ponds. In the AMT16_4.1 sample, *C.leptoporus* was quite common and a few *Umbilicosphaera* specimens occurred, but no Cocolithales sequences were found in the corresponding clone library; this discrepancy could be due to selective PCR amplification.

Group 4. Zygodiscales

The Zygodiscales is a rather low diversity group, which is well-supported both morphologically and paleontologically (Aubry 1989; Frada et al. 2009; Young et al. 2003). It includes two extant families the Pontosphareceae and Helicosphaerceae, members of which have been cultured and sequenced. Recent molecular phylogenetic studies confirm the monophyly of the group (Saez et al. 2004, de Vargas et al. 2007, liu et al, 2009).

Heterococcoliths and holococcoliths of *Helicosphaera* were common in the MedEx-6 morphological sample and the clone library contained 15 *Helicosphaera* sequences. In contrast, *Helicosphaera* was very rare in the AMT16_4.1 and HOT169_S2 morphological samples and no sequences occurred in the clone libraries. Six *Helicosphaera* sequences from MedEx-6 were identical to the *H. wallichii* culture sequence, another five sequences were similar to *H. carteri* and *H. wallichii* at 0% divergence, and the remaining four sequences were more distant from *H. carteri* and *H. wallichii* (3 sequences at 1% divergence and 1 sequence at 4% divergence). However, virtually all of the observed

heterococcospheres from the morphological sample were *H. carteri*. The holococcosphere *Syracolithus ponticuliferus,* which is suspected to be the holococcolith-bearing stage of a *Helicosphaera* species (Geisen et al. 2004) was common in the MedEx-6 morphological sample and one combination coccosphere of *H. wallichii* and *S. ponticuliferus* was observed (Couapel et al. 2009)). The comparison of morphological and genetic sampling is not straightforward because (a) we have observed *S. ponticuliferus* holococcoliths co-occurring with typical *H. carteri* type holococcoliths on single coccospheres, and (b) an *H. wallichii* strain that forms holococcoliths more like the typical (solid) type has been observed for *H. carteri* holococcoliths (Kyoko Hagino pers. comm.).

Group 5. Syracosphaeraceae

The Syracosphaeraceae are the most morphologically complex and species-rich group of coccolithophores, including ca 50 described species, many of which being possible pseudo-cryptic species (Cros & Fortuño 2002, Young et al. 2003). However, only two species have been isolated from laboratory cultures, *Syracosphaera puchra* and *Coronopshaera mediteranea*. As a result, their genetic diversity is essentially unknown.

The putative Syracosphaeraceae form a large and very diverse clade of sequences. The identification of this clade as corresponding to the Syracosphaeraceae is based on the presence of culture sequences from *Coronosphaera mediterranea* in a basal position and of *Syracosphaera pulchra* deep within the clade. The clade can itself be sub-divided into three well-separated, diverse, sub-clades. The sub-clade containing *S. pulchra* is almost certainly a *Syracosphaera* clade. The other two sub-clades could contain other genera such as *Calciosolenia*, *Ophiaster*, and *Michaelsarsia*, but most likely they are dominated by *Syracosphaera* species. Heterococcolith and holococcolith phases of *S. pulchra* were

common in both the MedEx-6 and HOT169_S2 samples, but rare in the AMT16_4.1 sample. Numerous sequences were found in the MedEx-6 clone library and some from the HOT169library, and they clustered close to the known *S. pulchra* sequence. *Syracosphaera histrica*, which we would predict to be the sister species of *S. pulchra on* morphological grounds, was also common in the MedEx-6 morphological sample. Some of the sequences closed to *S. pulchra* are probably *S. histrica*. Beyond this it is difficult even to speculate, all three morphological samples contained diverse low-abundance assemblages of Syracosphaeraceae and yielded numerous clones within the Syracosphaeraceae clade. Large-scale divisions of *Syracosphaera* have been discussed (e.g., Young et al. 2003) and it is conceivable that the three sub-clades seen here correspond roughly to the *S. pulchra, S. nodosa* and *S. molischii* groups. However there is not enough data here to test this assertion, and those groupings are tentative (Bootstrap<0.70; Fig 4.8). Overall, the molecular tree is more complex than we would predict from the observed morphological diversity, however more data are required confirm these conclusions.

Group 6. Umbellosphaeraceae

Umbellosphaera is a very common oligotrophic coccolithophore genus of uncertain affinity and no cultures (and hence no reference sequences) of this genus exist, it contains two well-established species *U. tenuis* and *U. irregularis*, but it has been suggested that *U. tenuis* is a cluster of at least five pseudo-cryptic species, informally termed *U. tenuis* types I, II, IIIa, IIIb and IV (Boeckel and Baumann 2008; Kleijne 1993; Young et al. 2003). Because the genus shows no obvious morphological affinities to other coccolithophores so a new family *incertae sedis* was established for it by Young et al. (2003).

Umbellosphaera was abundant in all the morphological samples. The AMT_4.1 and HOT169_S2 samples contained *U. irregularis* and *U. tenuis*, whereas the MedEX-6 sample contained *U. tenuis* but not *U. irregularis*. One large clade of sequences fell well outside all of the clades containing known coccolithophore sequences. This clade contains numerous sequences from both the AMT16_4.1 and the HOT169_S2 samples. Therefore, a plausible hypothesis is that this clade represents *Umbellosphaera*. This hypothesis is strongly supported by the data from the HOT169_S2 sample because (i) *Umbellosphaera* coccospheres represents ~ 70% of the observed assemblage in the morphological sample at this site, (ii) the clade contains 26 out of ~76 coccolithophore sequences in the library, and (iii)*Umbellosphaera* coccospheres represented ~ 70% of the observed assemblage in the morphological sample at this site.

The AMT and Hawaii morphological samples each contain both *U. tenuis* and *U. irregularis*, but the sequences from the two sites form discrete groups within the overall *Umbellosphaera* clade (Fig. 4.5, 4.7). For *U. tenuis* this arguably supports the previous morphological work suggesting that *U. tenuis* is a complex of several cryptic species. The Hawaii sample contains *U. tenuis* type IV, whilst the AMT sample contains primarily *U. tenuis* type IIIa, so it is quite possible that the two main clades, with respectively 25 and 21 clones represent these two *U. tenuis* types. The more basal sequences within the putative *Umbellosphaera* clade may represent additional *U. tenuis* types and/or *U. irregularis*.

This set of hypotheses appears rather convincing, even if there are two unresolved anomalies. First, *Umbellosphaera* was abundant in the morphological samples from MedEx but there are no putative *Umbellosphaera* sequences in the MedEx-6 library. Second, the AMT16_4 morphological samples contained both *U. tenuis* and *U. irregularis*

but all clones fall in a single low diversity clade, tentatively identified as *U. tenis* type III. Cortes et al. (2001) in a detailed study of coccolithophres from the HOTS station showed that *U. tenuis* occurred deeper in the water column than *U. irregularis* so it is possible that the observed coccospheres of *U. irregularis* in this relatively deep sample (79m) were settling dead cells.

4.5. Discussion

4.5.1. Methodologies

Despite our limited sampling effort, we were able to document 126 unique previously undescribed coccolithophores OTUs out of 130 unique coccolithophore OTUs observed. SEM examination of samples from all three locations revealed ~70 morphotypes (table 4.5), however, at least two potential biases are likely to affect our results. First, the three clone libraries were clearly skewed towards certain groups, such as the Syracosphaeraceae and the putative Umbellosphaeraceae. We argue here that such skewed sample distributions do not necessarily represent skewed species abundances. One of the primary challenges in environmental diversity studies is to overcome the PCR bias and retrieve the actual diversity from community samples (Acinas et al. 2005). Our parallel morphological examination addressed this problem by serving as a reference to screen for the lineages that could be missing in the clone libraries. For example, the Neolaerhabdaceae were very abundant in the morphological samples from all three sampling locations, but only a very small number of Neolaerhabdaceae sequences were obtained. In the HOT169_S2 sample, a large number of Rhabdosphaeraceae coccospheres was observed by SEM analysis, but again only three sequences were retrieved, while in the MedEx-6 sample, 24 *Umbellosphaera* coccospheres were observed by SEM without retrieving any sequence.

All of these discrepancies could be due to PCR bias or limited sequencing efforts. Alternatively, in several cases, sequences were retrieved, but no individuals were observed using SEM. For example, no *Umbilicosphaera* specimens were observed in the MedEx-6 morphological sample, but *U. sibogae* and *U. hulburtiana* sequences occurred in the clone library. These cases could represent biases or limitations of morphological methods.

The pre-filtertion step used in our study represented the second major bias potentially affecting our results. This step is an excellent technique for cleaning up the coccolithophore samples, but here it almost certainly skewed the assemblage composition to an undue degree. The reason we used the pre-filtration step has to do with the use of Haptophyta-specific primers in our study. Currently, it has proven impossible to design efficient coccolithophore-specific primers with a desirable fragment size based on our Haptophyta *LSU* rDNA sequence database. Previous studies (Liu et al, 2009) showed that the non-calcifying pico-haptophytes (< 3 μm) are very abundant and diverse, thus without pre-filtration a majority of the Haptophyta sequences retrieved by clone libraries will fall into this size <3 μm size category. We conducted analyses with (HOT169_S2 and MedEx-6) and without (AMT16_4.1) the 5 μm pre-filtration step. The two libraries with pre-filtration resulted in 75.5-87.2% coccolithophore sequences, however only 35.6% coccolithopohre sequences were retrieved from the AMT16_4.1 library.

4.5.2. Comparison of species richness

Prior to this study, knowledge of coccolithophore species richness relied mainly on morphological identifications made by SEM or LM. This study illustrates that even a limited sequencing effort reveals a far greater diversity than that discovered solely through a corresponding morphological analysis. No apparent plateau appeared in the genetic

rarefaction curves at the unique level (Fig.4. 2), suggesting that the libraries did not encompass the full extent of OTU richness in any of the three sampling locations. However, the rarefaction curves based on morphological data showed a tendency to plateau.

Significantly, the highest genetic and morphological diversity levels occurred in the HOT169_S2 sample (Table 4.2). This sample was collected from Hawaiian tropical waters at 79 m depth, where the standing stock of coccolithophores was highest based on our CTD (conductivity, temperature, density) data (Fig. 4.1) and previous studies conducted at the same sampling location (HOT station ALOHA, Hawaii (Cortés et al. 2001). The dramatic diversity at this site possibly correlates with optimum growth conditions at this station. Estimates of the number of unique ribotypes using the Chao1 and ACE estimator were 399 (95% CI: 219–804) and 493 (95% CI: 136–2922), respectively, which exceed the total number of all currently identified morphological species (~280) (Young et al, 2003, 2005). Conversely, the AMT16_4.1 sample was collected from surface waters in the oligotrophic southeastern Atlantic Ocean, where the cell density of coccolithophores is rather low. The total genetic richness estimated by the Chao1/ACE and Shannon diversity index was lowest in the AMT16_4.1 sample. The MedEx-6 sample presented an intermediate genetic diversity between that of HOT169_S2 and AMT16_4.1. However the AMT16_4.1 sample has the lowest morphological diversity as indicated by Chao1 estimator. MedEx-6 sample was collected in the Bay of Villefranche-sur-Mer, where the water is considered to be oligo-mesotrophic according to our CTD data and classifications of water systems (Sorokin 1981). The differences in genetic and morphological diversity of coccolithphore

community may correlate with the different the trophic states and / or more complex multiple environmental drivers of the water column.

While we found that some lineages are present in more than one sampling sites, it is notable that more lineages seem to be specific to one site (Fig. 4.5). For example, we found both cases in Syracosphaeraceae, however in Umbellsphaeraceae, all clades formed seem to be specific to only one site (either AMT16_4.1 or HOT169_S2).

We also compared the morphological and genetic rarefaction curves at different similarity levels to determine the genetic variability within and between morphological species in each sample. The three samples yielded different results. The morphological rarefaction curve for HOT169_S2 was closest to the genetic rarefaction curve at the 2% difference cut-off. The rarefaction curve for the MedEx-6 morphological sampling was closest to the curve constructed from genetic sampling at the 1% difference cut-off and the rarefaction curve for the AMT16_4.1 morphological sampling fell in between the curves constructed from genetic sampling at the unique and 0% difference levels. These results suggests that the genetic variability within established morphospecies varies significantly across different environments, and that the genetic distance threshold at which a morphological species is defined varies when different natural communities are sampled. Here we show that for coccolithophore *LSU* rDNA this threshold stands at almost 3% in species-rich environments, and drops below 0%, in species-poor environments, that is, the phylospecies threshold is underestimated.

The question of whether molecular taxonomy should be used to supplement or even to replace traditional morphological taxonomy is the subject of considerable debate. Many researchers have suggested that molecular taxonomy can be used to identify and classify

species (e.g., Blaxter 2004; Hebert et al. 2003; Tautz et al. 2003), whereas others have argued that morphology must continue to play the central role in defining species boundaries (Dunn 2003; Will and Rubinoff 2004). The similarity threshold at which morphological and phylogenetic species should be defined is still not properly addressed. In bacteria, sequences with < 3% difference are typically assigned to the same species and those with 5% differences are typically assigned to the same genus (Bond et al. 1995; McCaig et al. 1999; Michelle Sait 2002). However, these numbers are sometimes controversial and subject to debate (Pedrós-Alió 2006). Identity cut-offs ranging from 1 to 3 % are often are used to define OTUs for 16S rDNA (Munson et al. 2004).

4.5.3. Comparison of molecular and morphological analyses

The morphological and molecular data were first compared qualitatively and quantitatively within each phylogenetic group (family or genera level) as defined above. In summary, the Neolaerhabdaceae were abundant in all three morphological analyses but were absent or very rare in all three genetic analyses; this discrepancy most likely is due to preferential PCR amplification. The Rhabdosphaeraceae showed discrepancy in the HOT169-S2 sample, in which many coccospheres were observed, but only three sequences were found in the clone library. As this group was rare in both genetic and morphological analyses for the other two samples, the possible explanation for the discrepancy in the HOT169-S2 data could also be due to PCR preferential amplification. Very few Coccolithales sequences or coccospheres were found in the three samples and the morphological and genetic data were broadly consistent at each sampling location. Morphological and genetic data were also correlated for all three sampling locations for the Zygodiscales. Many sequences and coccospheres were found in the MedEx-6 sample, whereas almost none were found in the

HOT16_S2 and AMT16_4.1 samples. The Syracosphaeraceae was the most diverse and abundant group in this study. Many sequences/coccospheres were found in all three sampling locations. The putative Umbellosphaeraceae exhibited a discrepancy in the MedEx-6 sample: many individuals were observed in SEM, but no sequences were retrieved, which may be a sampling bias artifact due to the prefiltration.

For this highly diverse group our genetic data (Fig. 4.5, 4.7) provides a first insight into the likely phylogenetic structure of the group and suggests that the genus *Syracosphaera* is probably paraphyletic. However, more intensive comparisons of morphological and genetic data are required to test for a correlation between molecular and morphological data. For instance, the Identification of the putative Umbellosphaeraceae clade is an interesting premise yielded from our approach, although it requires further testing. However, even within this group there was a major discrepancy in the MedEx-6 sample, where many individuals were observed by SEM, but no sequences were retrieved.

4.5.4. Complexity of combing molecular and morphological data

Numerous anomalies were found in attempting to make connections between sequences and species and the organisms identified by SEM. Intriguingly, a simple one-to-one mapping does not exist. A good example of this is the discrepancies observed with respect to our putative Umbellosphaeraceae samples. Firstly, *U. tenuis* was common in the MedEX-6 morphological sample, but there were no MedEx-6 sequences in the putative *Umbellosphaera* clade. A possible explanation for this result is that the *U. tenuis* specimens in the MedEX-6 sample were mostly rather small (~7-8µm) and compact coccospheres (type IIIb) (Young et al. 2003). As a result they may have passed through the 5µm mesh that was used for pre-filtration. Secondly, the AMT16_4.1 and HOT169_S2

samples both contained *U. tenuis* and *U. irregularis,* but the sequences from the two sites formed discrete clades within the overall *Umbellosphaera* clade. For *U. tenuis* this arguably supports previous morphological work suggesting that *U. tenuis* is a complex of several species (Kleijne 1993, Young et al. 2003, Boeckel & Baumann 2008). The Hawaii sample contained *U. tenuis* type IV (which is a large form and so would probably be retained on the >5µm filter), while the AMT16_4.1 sample contained primarily *U. tenuis* type IIIa. The *U. irregularis* coccospheres from the two areas appeared very similar, so the absence of any overlap is surprising. Thirdly, the AMT16_4.1 sample contained *U. irregularis* and *U. tenuis* in similar abundances and with no morphological overlap, so we would expect two well-separated clades from this sample. However, all of the identified sequences fell into one clade.

4.6. Concluding remarks

This study represents the first detailed attempt to reconcile the operation species definitions at the morphological and genetic levels in the context of the metagenetics analysis of a group of marine planktonic eukaryotes. The study clearly shows that working definitions of phylospecies and of morphospecies depend on the environment sampled, where genetic studies overestimate morphospecies in species-rich (equatorial) environments, and seriously underestimate them in species-poorer (temperate)environments. As a result, the frequency of retrieved DNA sequences that are retrieved is unlikely to be a simple function of species abundance in the sample without taking latitude and therefore species-richness into account. Reciprocally, morphotaxonomy is unlikely to be an accurate reflection of species-level diversity. Integration of molecular and morphological techniques is not straightforward. Yet, our study demonstrates the

possibility to match the majority of the 126 previously undescribed unique OTUs to 70 morphotypes observed at species or genus level. With more data it should be possible both to develop a comprehensive phylogeny of coccolithophores linked to morphological taxonomy, and to calibrate the amplification biases so that environmental DNA techniques can be used for automated analyses of populations. Such techniques can both provide us with revolutionary new insights into marine eukaryote diversity and ecology, and allow us to integrate these insights gained from large-scale metagenetic or metagenomic studies with more traditional metamorphological methods

4.7. Tables.

Table 4.1: Summary of hydrographic conditions at study sites.

Library	Cruise	Station	Long.	Lat.	Depth (m)	Temperature (^{o}C)	Salinity (psu)	Chlorophyll (ug/L)
HOT169_S2	HOT169	S2	-158	22.75	79	23.8	35.1	16.31
AMT16_4.1	AMT16	4	9.33	-30.58	2	19.6	35.4	0.04
MedEx_6	MedEx	D	7.32	43.69	20	22.1	38.1	0.12

Table 4.2. Diversity and richness estimations from morphological and genetic sampling.

Sampling method	Sample name	No. of sampled individuals or sequences	No of species or unique OTUs observed	Shannon Diversity Index
Genetic sampling	HOT169-S2	80	74	4.278
	AMT16_4.1	62	26	2.615
	MedEx_6	75	45	3.307
Morphological sampling	HOT169-S2	300	35	2.467
	AMT16_4.1	191	28	1.777
	MedEx_6	238	22	2.406

Table 4.3: List of number of OTUs at unique and 3% levels retrieved from genetic sampling and number of species identified from morphological sampling by sub-group.

Taxonomic Group (Fmaily or Order)	MedEx_6			HOT169_S2			AMT16_4.1		
	Genetic OTUs		Morphological # of species	Genetic OTUs		Morphological # of species	Genetic OTUs		Morphological # of species
	unique	-3%		unique	-3%		unique	-3%	
Noelaerhabdaceae	1	1	1	2	1	2	4	1	2
Rhabdosphaeraceae	5	2	3	3	3	6	0	0	6
Coccolithales	9	5	3	0	0	3	0	0	2
Zygodiscales	6	1	2	0	0	3	0	0	0
Syracosphaeraceae	26	3	9	37	4	14	18	2	13
Umbellosphaeraceae	0	0	1	22	4	2	11	1	2
Others	1	1	3	6	2	5	0	0	2

Table 4.4: Species counts based on LM and SEM for the three morphological samples within each defined subgroup.

Subgroups	MedEx-6		AMT 16.4.1		Hawaii HOT169-S2	
	LM	SEM	LM	SEM	LM	SEM
NOELAERHABDACEAE						
Emiliania huxleyi	78	71	117	117	18	29
Gephyrocapsa oceanica	1				1	1
G. ericsonii	7	10		9		
Reticulofenestra sessilis						1
COCCOLITHALES						
Calcidiscus leptoporus HET		2	6	3		
Calcidiscus leptoporus I				1		
Calcidiscus quadriperforatus I	1	1				1
Hayaster perplexa						2
Oolithotus fragilis			?			
Umbilicosphaera sibogae			1		1	
Umbilicosphaera foliosa			1			
Umbilicosphaera hulburtiana				1		1
ZYGODISCALES						
Helicosphaera carteri HET	9	5				
Helicosphaera wallichii HET	5					1
Helicosphaera hyalina HET						1
H. carteri HOL solid	7	7				
H. carteri HOL perforate	12	12				
H. carteri HOL *ponticuliferus*	34	13				
H. pavimentum					1	
Scyphosphaera apsteinii						1
Pontosphaera discopora						1
Pontosphaera japonica						1
SYRACOSPHAERALES						
Syracosphaeraceae						
Syracosphaera pulchra HET	19	14	3		8	14
S. pulchra HOL *oblonga* type	18	2			1	11
S. pulchra HOL *pirus* type						6
Syracosphaera anthos HET	8	1	3	6	1	
Syracosphaera anthos HOL	2		1	1	1	
Sy bannockii				1		1
Sy cf. *bannockii*				2		
Sy corolla				3		1
Sy dilatata						9
Sy histrica		9		1		
Sy marginoporata				1		
Sy molischii		1		3		2
Sy nana	50		8	1	5	
Sy nodosa		6				1
Sy ossa		5				1
Sy protrudens		11				
Sy rotula						5
Sy sp cf. *nana*						2
Sy sp. type G						1
Sy sp. type L		2				2
Sy. cf. *orbiculus*				2		
Ophiaster				2		3
Michaelsarsia elegans				1		
Coronosphaera binodata					1	4
Coronosphaera mediterranea	1	4				
Coronosphaera mediterranea HOL		2				
Calciosoleniaceae						
Calciosolenia brasiliensis			1	3		1
Calciosolenia murrayi						3

Subgroups	MedEx-6		AMT 16.4.1		Hawaii HOT169-S2	
	LM	SEM	LM	SEM	LM	SEM
Rhabdosphaeraceae						
Rhabdosphaera stylifer	32	5	3	1	4	5
Rh. clavigera		1	4	1	6	12
Discosphaera tubifer			7	3	6	80
Rhabdosphaera xiphos						4
Palusphaera vandellii				1		
Algirosphaera robusta		1		1	1	1
Acanthoica quattrospina			1	1	3	2
Cyrtosphaera aculeata						2
INCERTAE SEDIS						
Holococcoliths of uncertain affinities						
Corisphaera cf. gracilis		1				
Calyptrolithophora papilifera						1
Anthosphaera sp	8			1		
Syracolithus schilleri						3
Poricalyptra aurisinae						1
Helladosphaera pienaarii		1				
Umbellosphaeracea						
Umbellosphaera tenuis	22	24	5	8	57	58
Umbellosphaera irregularis			6	14	51	58
Heterococcolith genera inc. sedis & nannoliths						
Ceratolithus cristatus CER	30	28	1			2
Ceratolithus cristatus HET coccolithomorpha						1
Ceratolithus cristatus HET nishidae	2	7	1		1	
Alisphaera sp.		3				3
Florisphaera profunda			3	2		
TOTAL	**346**	**249**	**172**	**191**	**167**	**340**

Table 4.5: Species counts based on LM and SEM for the three morphological samples within each defined subgroup

Subgroups	MedEx-6		AMT 16.4.1		Hawaii HOT169-S2	
	LM	SEM	LM	SEM	LM	SEM
NOELAERHABDACEAE						
Emiliania huxleyi	78	71	117	117	18	29
Gephyrocapsa oceanica	1				1	1
G. ericsonii	7	10		9		
Reticulofenestra sessilis						1
COCCOLITHALES						
Calcidiscus leptoporus HET		2	6	3		
Calcidiscus leptoporus I				1		

Subgroups	MedEx-6		AMT 16.4.1		Hawaii HOT169-S2	
	LM	SEM	LM	SEM	LM	SEM
Calcidiscus quadriperforatus I	1	1				1
Hayaster perplexa						2
Oolithotus fragilis			?			
Umbilicosphaera sibogae			1		1	
Umbilicosphaera foliosa			1			
Umbilicosphaera hulburtiana				1		1
ZYGODISCALES						
Helicosphaera carteri HET	9	5				
Helicosphaera wallichii HET	5					1
Helicosphaera hyalina HET						1
H. carteri HOL solid	7	7				
H. carteri HOL perforate	12	12				
H. carteri HOL *ponticuliferus*	34	13				
H. pavimentum					1	
Scyphosphaera apsteinii						1
Pontosphaera discopora						1
Pontosphaera japonica						1
SYRACOSPHAERALES						
Syracosphaeraceae						
Syracosphaera pulchra HET	19	14	3		8	14
S. pulchra HOL *oblonga* type	18	2			1	11
S. pulchra HOL *pirus* type						6
Syracosphaera anthos HET	8	1	3	6	1	
Syracosphaera anthos HOL	2		1	1	1	
Sy bannockii				1		1
Sy cf. *bannockii*				2		
Sy corolla				3		1
Sy dilatata						9
Sy histrica		9		1		
Sy marginoporata				1		
Sy molischii		1		3		2
Sy nana	50		8	1	5	
Sy nodosa		6				1
Sy ossa		5				1
Sy protrudens		11				
Sy rotula						5
Sy sp cf. *nana*						2
Sy sp. type G						1
Sy sp. type L		2				2
Sy. cf. *orbiculus*				2		
Ophiaster				2		3
Michaelsarsia elegans				1		
Coronosphaera binodata					1	4
Coronosphaera mediterranea	1	4				
Coronosphaera mediterranea HOL		2				
Calciosoleniaceae						
Calciosolenia brasiliensis			1	3		1
Calciosolenia murrayi						3

Subgroups	MedEx-6		AMT 16.4.1		Hawaii HOT169-S2	
	LM	SEM	LM	SEM	LM	SEM
Rhabdosphaeraceae						
Rhabdosphaera stylifer	32	5	3	1	4	5
Rh. clavigera		1	4	1	6	12
Discosphaera tubifer			7	3	6	80
Rhabdosphaera xiphos						4
Palusphaera vandellii				1		
Algirosphaera robusta		1		1	1	1
Acanthoica quattrospina			1	1	3	2
Cyrtosphaera aculeata						2
INCERTAE SEDIS						
Holococcoliths of uncertain affinities						
Corisphaera cf. gracilis		1				
Calyptrolithophora papilifera						1
Anthosphaera sp	8			1		
Syracolithus schilleri						3
Poricalyptra aurisinae						1
Helladosphaera pienaarii		1				
Umbellosphaeracea						
Umbellosphaera tenuis	22	24	5	8	57	58
Umbellosphaera irregularis			6	14	51	58
Heterococcolith genera inc. sedis & nannoliths						
Ceratolithus cristatus CER	30	28	1			2
Ceratolithus cristatus HET coccolithomorpha						1
Ceratolithus cristatus HET nishidae	2	7	1		1	
Alisphaera sp.		3				3
Florisphaera profunda			3	2		
TOTAL	**346**	**249**	**172**	**191**	**167**	**340**

4.8. Figures

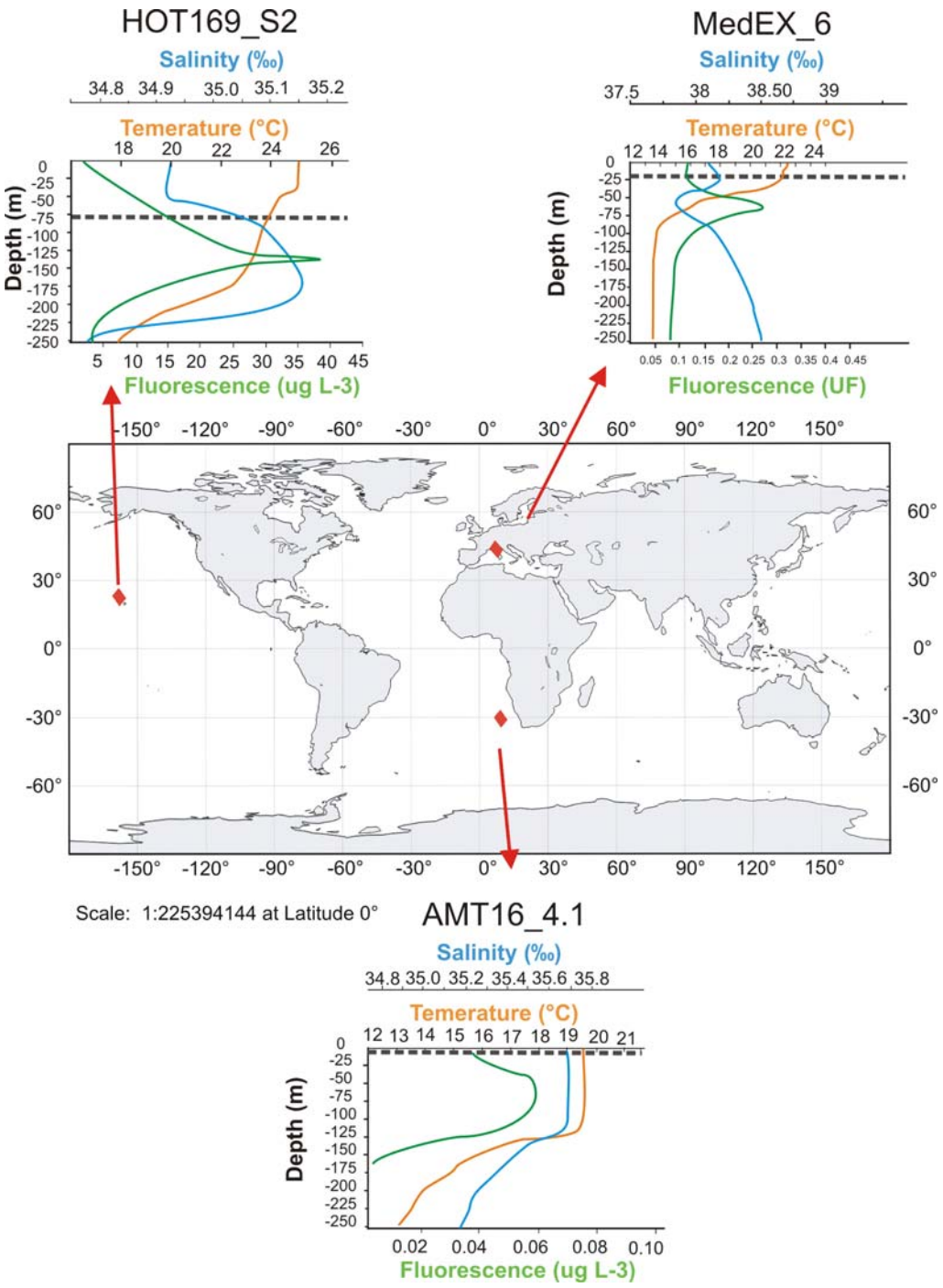


Figure 4.1: Map of sampling sites (red squares) and hydrographic conditions at each site. Temperature, salinity, and fluorescence profiles down to 250m depth are given for each station. Dotted lines indicate the depths at which water was sampled for comparative genetic and morphological analysis.

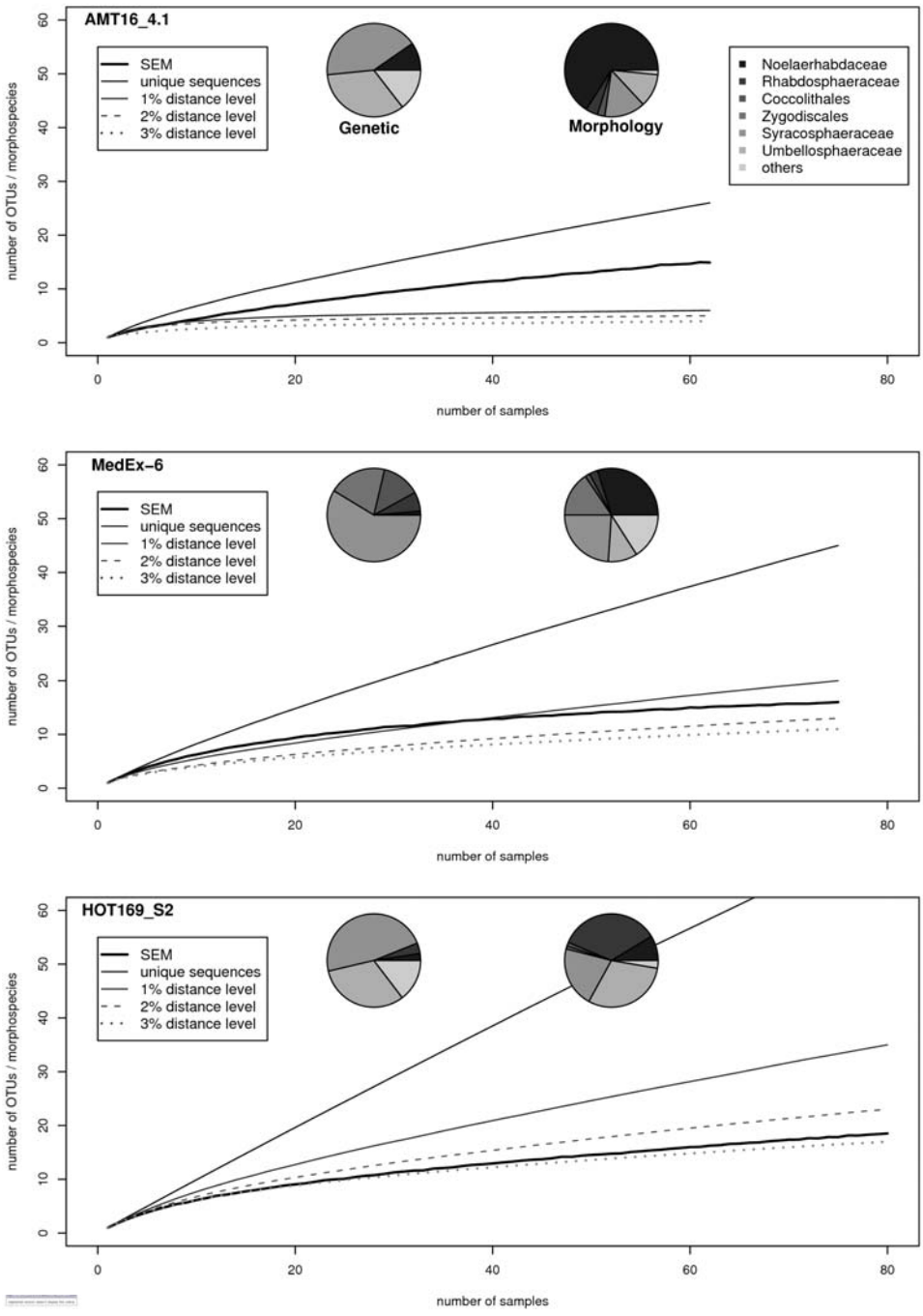


Figure 4.2: Rarefaction curves for both morpho- and phylospecies samplings at each site. Different levels (unique, 0%, 1%, 2% and 3%) of genetic divergence were used for phylospecies sampling. Pie chart indicating the relative frequency of sequences or individuals identified in each subgroup from each morpho-genetic sampling

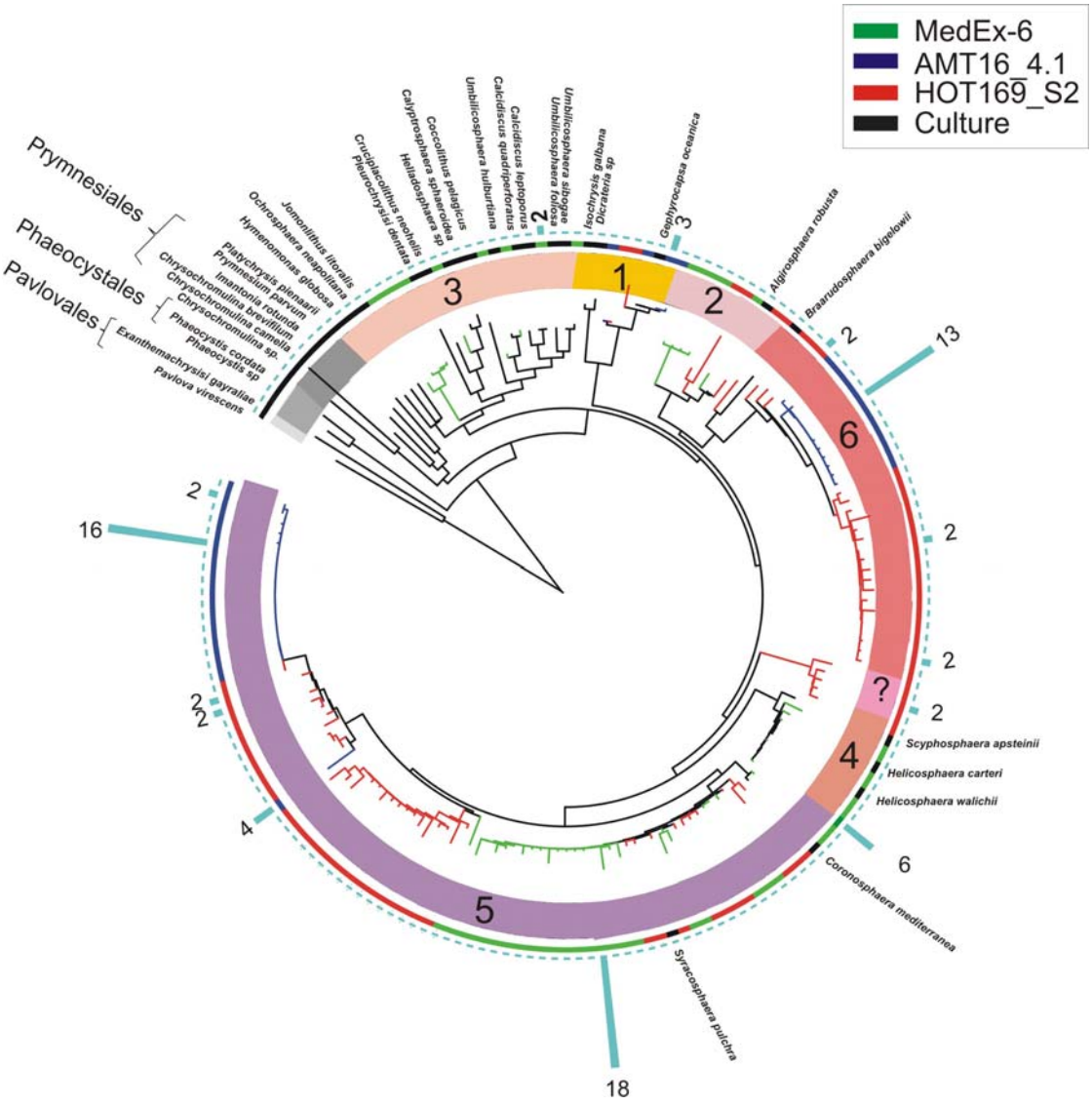


Figure 4.3: LSU rDNA based Neighbor Joining tree, including all unique coccolithophore environmental sequences and a full taxonomic cross-section of known, cultured, haptophyte species. The color code used in the tree topology and outer circle highlights the origin of the phylospecies: black = culture collection, red = Pacific ocean, blue = Atlantic ocean, and green = Mediterranean sea. Names are given for each identified, cultured strains allowing to give a taxonomic status to the 10 detected environmental clusters. The number of sequences included in each unique OTU are indicated by light blue bars and associated number if >1. The tree was formatted using the interactive tree of life (iTOL, http://itol.embl.de/itol.cgi).

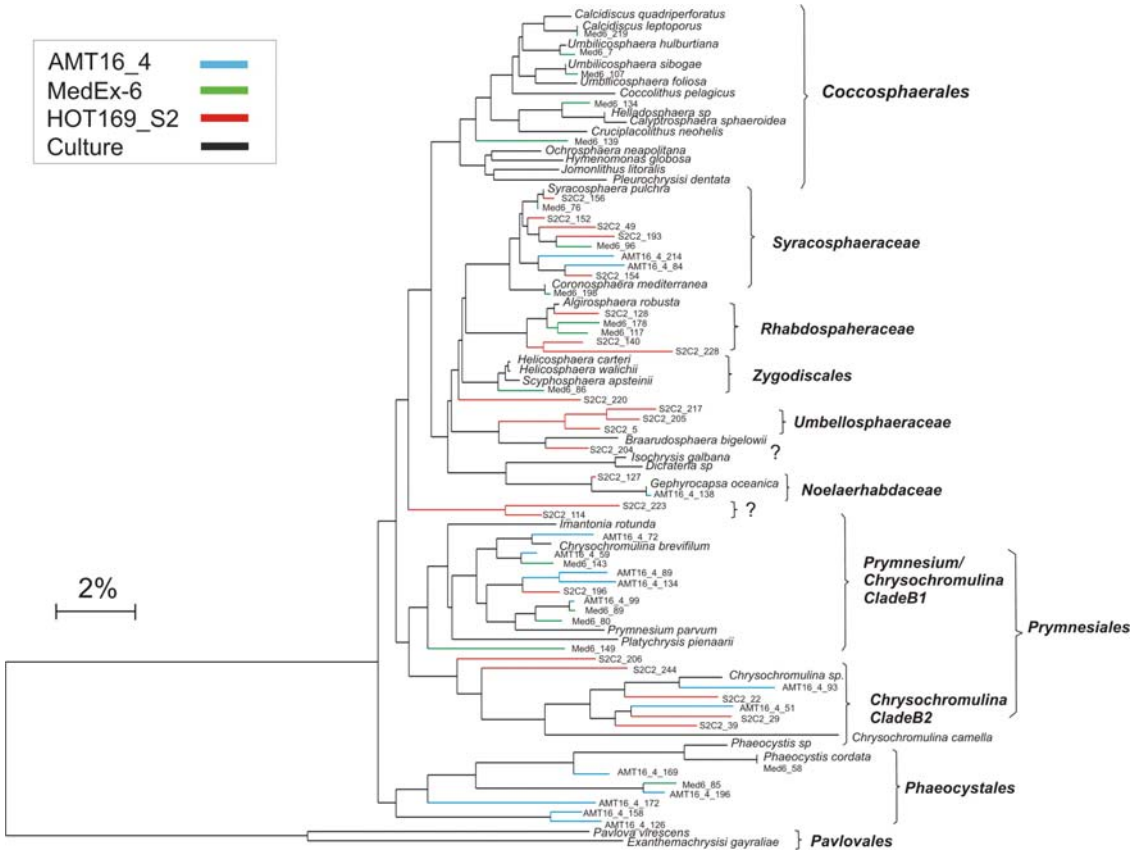


Figure 4.4: *LSU* rDNA Bayesian tree including all environmental haptophyte sequences at the 3% divergence cut-off.

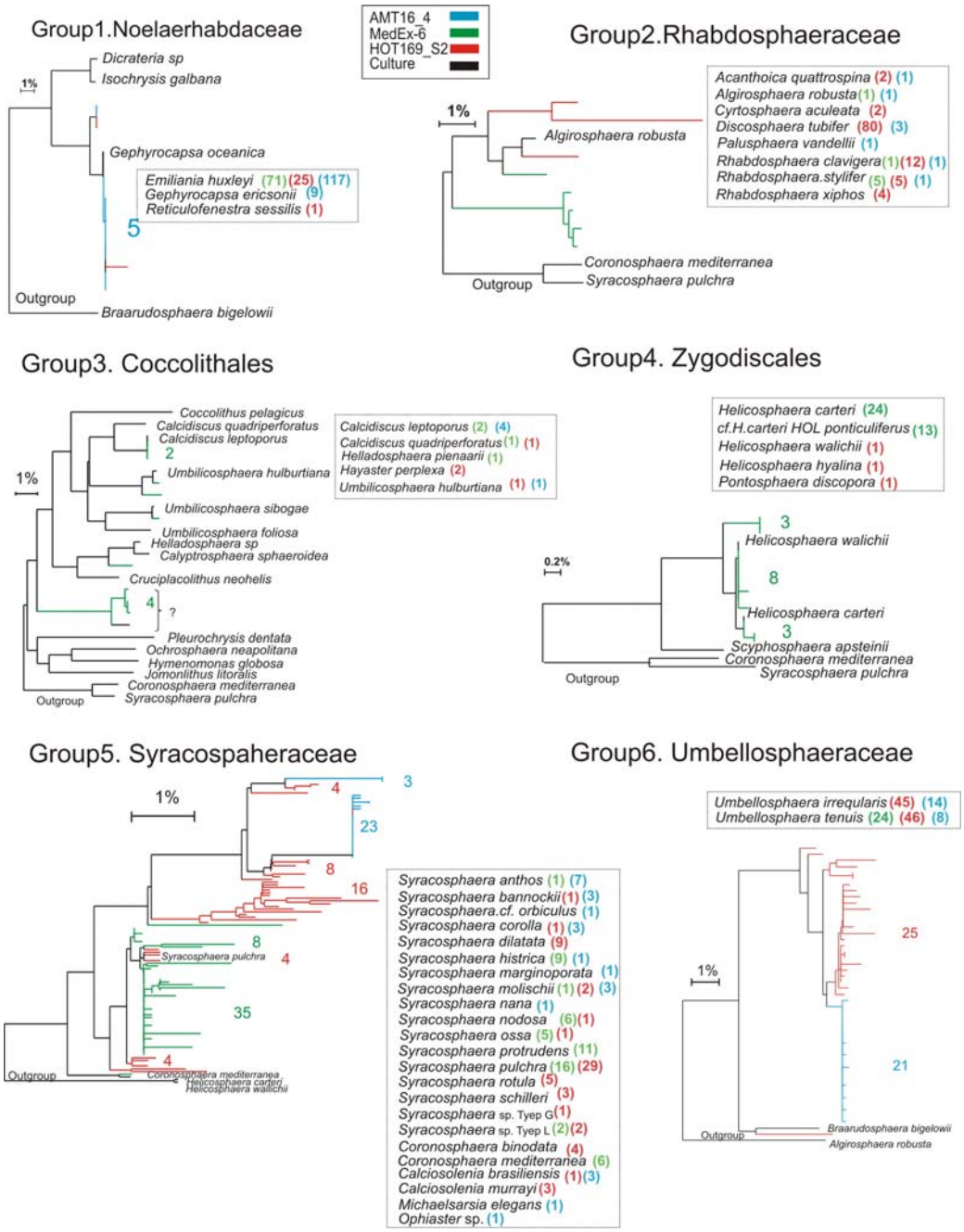


Figure 4.5: Sub-trees for groups of interest, the number of individuals identified in each morphological sampling are listed in parallel (i.e., those labeled A, B, C, D, E, and F).

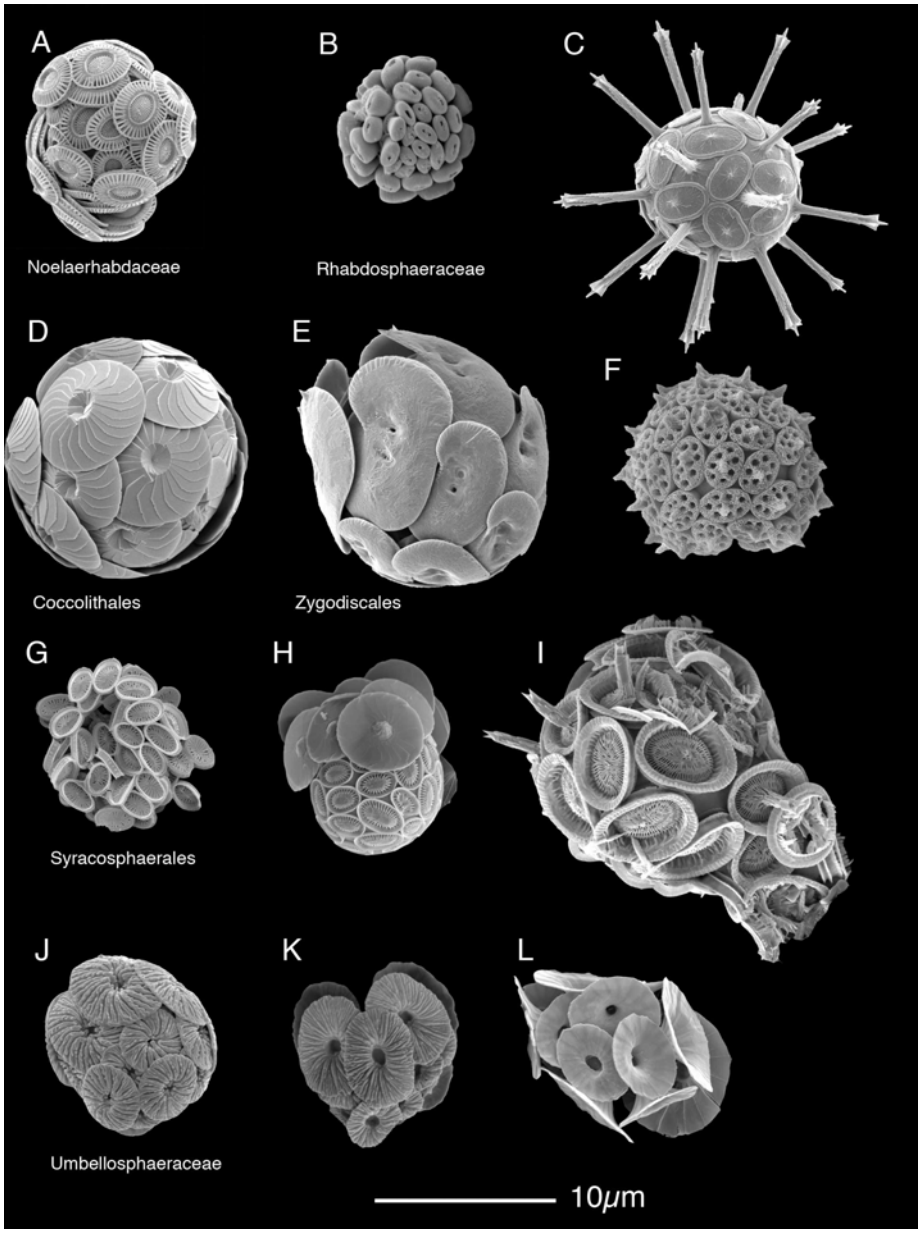


Figure 4.6: Scanning electron micrographs of representatives of the different coccolithophore clades discussed. All images at the same scale. From the upper left are Noelaerhabdaceae A. *Emiliania huxleyi,* from AMT14; Rhabdosphaerceae B *Algirosphaera robusta,* from HOTS169**;** C *Rhabdosphaera stylifera*, from Alboran Sea - to be replaced by MEDEX specimen Coccolithales; D *Calcidiscus leptoporus*, from culture- to be replaced by MEDEX specimen; Zygodicales E *Helicosphaera carteri* HET, from Alboran Sea - to be replaced by MEDEX specimen; F *Helicosphaera carteri* HOL confusus type, from MEDEX; Syracosphaerales; G *Syracosphaera dilatata*, from HOTS 169; H. *Syracosphaera anthos*, from AMT16; I. *Syracosphaera pulchra*, from MEDEX; Umbellosphaeraceae J. *Umbellosphaera tenuis* type IIIa from AMT16 ; J. *Umbellosphaera tenuis* type IV from HOTS169; J. *Umbellosphaera irreglaris* type IIIa from HOTS169

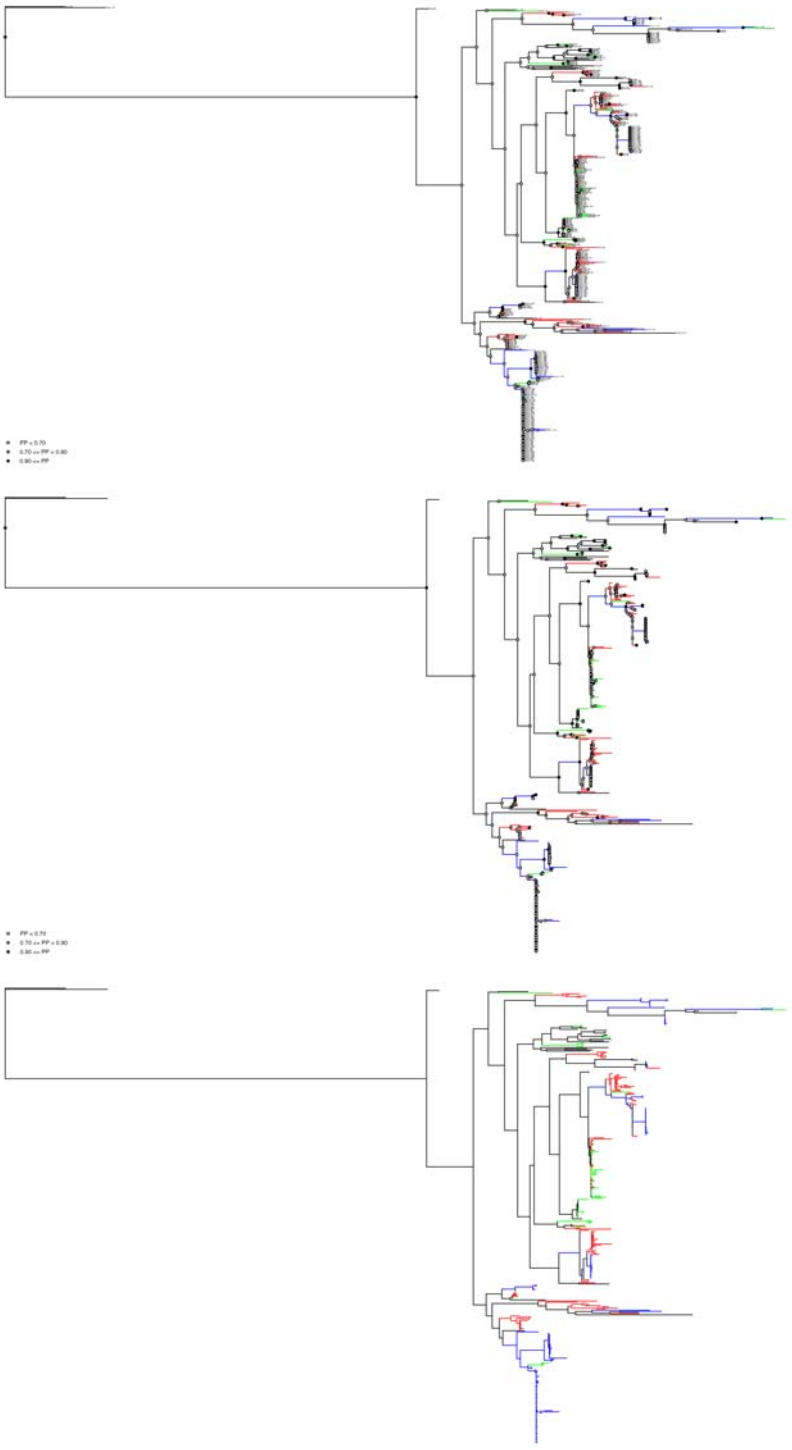


Figure 4.7. Phylogeny of all all environmental haptophyte diversity reconstructed using Neighbor-Joining, maximum likelihood and Bayesian statistics.

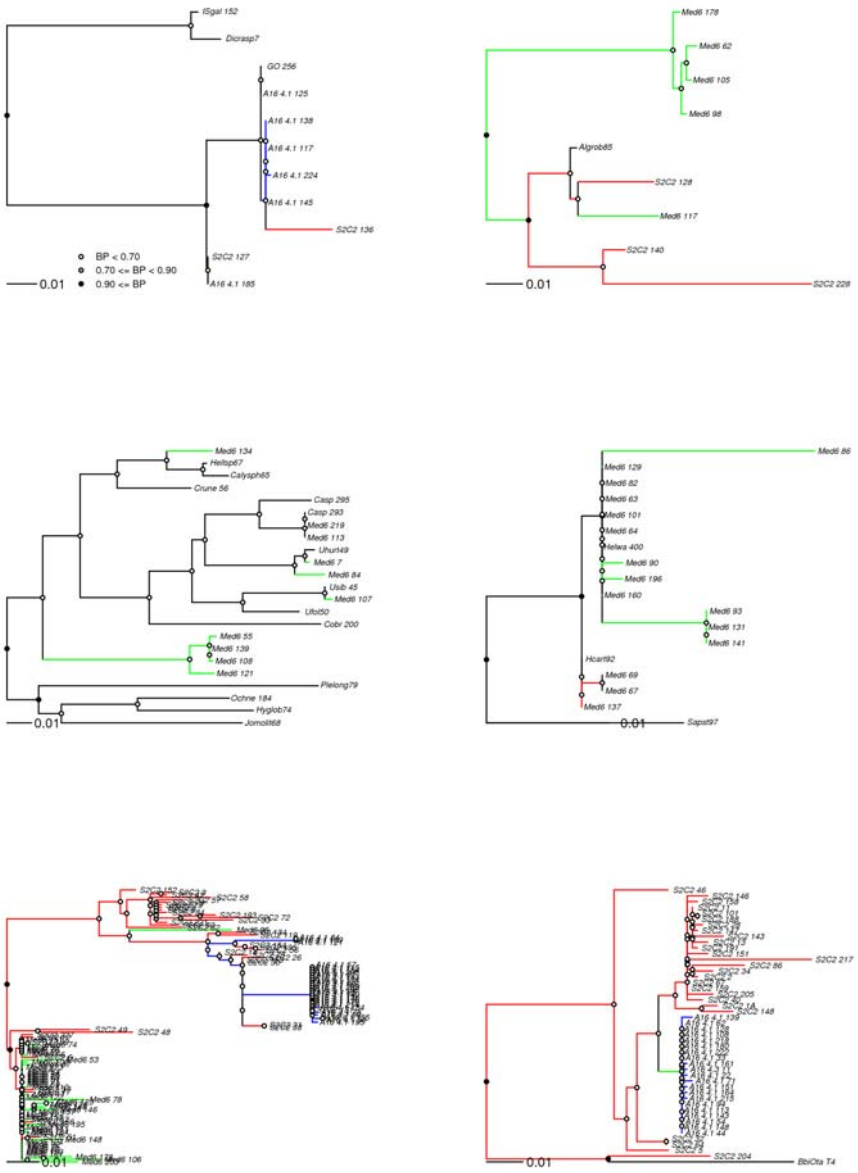


Figure 4.8. Six subgroup trees extracted from the ML tree. Support values are included for internal nodes (1000 bootstrap replicates) (See Figure 4. 5 for group name).

Chapter 5

5.0. Diversity, biogeography, and evolution in non-spinose planktonic foraminifera (*Neogloboquadrinids*)

5.1. Abstract

Recent studies of genetic sequences from the oceans unveiled an astounding level of cryptic diversity in all organisms, from viruses to metazoans. Newly discovered genetic types are particularly abundant among unicellular organisms. How did this extensive diversity arise, evolve, and colonize the immense fields of oceanic waters? Here we present a case study of pelagic microbial evolution and biogeography based on rDNA analyses of the Neogloboquadrinids, a family of non-spinose planktonic foraminifera that left one of the best fossil records on Earth. The Neogloboquadrinids first appeared ~22 Ma and radiated ~11 Ma into eight morphological species that successfully colonized the global ocean from the Equator to both poles. Based on a worldwide sampling of the three modern morphospecies within the family, we show that at least 10 distinct genetic types can be defined. Each type corresponds to a monophyletic group with slightly divergent ribotypes that vary as much between as within individuals. We reinterpreted the phylogeny of the family based on a comparison between genetic and geological data, and we discuss the biogeography of each genetic type in terms of adaptation and dispersal. While the genetic types inhabiting the equatorial to temperate waters have transbasin and transhemispheric and likely continuous distributions, the types collected in subpolar and polar waters have circumglobal but monopolar distributions.

5.2. Introduction

Recent molecular studies have revealed the widespread presence of cryptic species among many groups of marine plankton, such as copepods (Goetze 2005), ciliates (Katz et al. 2005), diatoms (Amato et al. 2007), coccolithophores (Saez et al. 2003), and planktonic foraminifera (Darling et al. 2004; Darling et al. 2007; de Vargas et al. 1999). These studies have led to the discovery of a huge hidden genetic diversity. Furthermore, exploration of the biogeography and evolutionary history of these cryptic species has provided insight into the pace and mechanisms of speciation and diversification in the pelagic ecosystem. The cryptic species concept is particularly important in the study of planktonic foraminifera. The remains of these organisms constitute one of the most complete fossil records on Earth, dating to ~130 million years ago. Planktonic foraminifera have long been used as markers in the study of paleoenvironments and paleoecology. For example, planktonic foraminifera tests are used to reconstruct ancient sea surface temperature, ice volume, and salinity (CLIMAP 1981; CLIMAP 1984); Some foraminifera species such as *Neogloboquadrina pachyderma, Globigerina bulloides,* and *G. inflata* have well-defined preferences for certain conditions that can be used as climate indicators (Bandy 1972). However, with increasing molecular data showing that most of the morphospecies actually are composed of several cryptic species with particular habitats (de Vargas et al. 1999), the universal assumption made in these paleoapplications is seriously challenged. Study of the speciation and biogeography of planktonic foraminifera will not only help us to better understand the evolution of this organism, but also to achieve better precision for their use as paleo-oceanographic proxies.

In this study, we choose the Neogloboquadrinids, a family of non-spinose planktonic foraminifera, to address questions about pelagic microbial evolution and biogeography. This family first appeared about 22 Ma; they radiated much later (~11 Ma) into eight morphological species that successfully colonized the global ocean from the Equator to both poles (Figure 5.1) (Kennett and Srinivasan 1983). The three extant morphological species—*Neogloboquadrina pachyderma*, *N. dutertrei*, and *Pulleniatina obliquiloculata*—are among the most important foraminiferal species in terms of ecological success, evolutionary complexity, and paleo-oceanographic tracers. *Neogloboquadrina pachyderma* evolved from *N. continuosa* about 11.2 Ma ago (Bandy 1972) and dominates planktonic foraminifera assemblages throughout the high-latitude marine provinces of both hemispheres (Darling et al. 2007). Today, this species exhibits two coiling directions. The left-coiling *N. pachyderma* dominates polar regions with temperatures < 9 ºC (Bandy 1972), especially north of the Arctic Front in the Greenland, Iceland, and Norwegian Seas (Pflaumann 1996); the right-coiling *N. pachyderma* occurs in waters with temperatures ranging from 9 to 18 ºC, extending from polar to warm subtropical areas. *Neogloboquadrina dutertrei,* which arose from *N. humerosa* about 5.5 Ma ago, is dominant in tropical to warm subtropical waters and thrives in eutrophic areas. *Pulleniatina obliquiloculata*, which evolved from *P. praecursor* about 4 Ma ago, inhabits tropical to warm subtropical regions. In this lineage, *P. praecursor* arose from *P. primalis* about 5.2 Ma, which itself evolved from *N. acostaensia* about 6.2 Ma ago.

Members of the Neogloboquadrinid family exhibit considerable morphological and genetic variability (Darling et al. 2000; Srinivasan and Kennett 1976) thus

providing an ideal model for the study of cryptic speciation in open-ocean plankton. To date, nine cryptic species have been defined within the family; in particular, *N. pachyderma* has been used extensively as a model to study global biogeography and cryptic speciation (Darling et al. 2004; Darling et al. 2007; Darling et al. 2000). However, the geographic scaling and the intraindividual genetic variability remain the two major impediments to the study of this family. First, all three extant morphospecies occupy worldwide, bi-hemispheric biogeographic ranges. It is reasonable to assume that each morphospecies could be split into 5 to 10 biological species with more restricted ranges. However, the range of genetic types, even if more restricted, might also be widespread among the oceans and might be seasonal. Thus, *worldwide* and *multi-seasonal* sampling is a prerequisite to assess the geographic and ecological range of the family. Such an approach was partially undertaken for *N. pachyderma* (Darling et al. 2006; Darling et al. 2004; Darling et al. 2007), but the spatial coverage was seriously biased toward higher latitudes areas, preventing a thorough assessment of the genetic diversity of *P. obliquiloculata* and *N. dutertrei*. Another major limitation to understanding the evolutionary history of this family is that different copies of rDNA clusters are genetically variable within a single individual; even the 18S rDNA, which is a marker often used as a good proxy for biological species(Darling et al. 2006; Darling et al. 2004), might be subject to intraindividual variation. Thus, assessment of the extent of this intraindividual genetic variability is a fundamental prerequisite to defining a (cryptic) species.

In this study, we revisited these studies in the context of a more comprehensive sampling of the World Ocean and a detailed examination of both intraindividual and

intraspecific genetic variations. Our goals were to (1) examine the SSU rRNA genotypic diversity and spatial distribution of each morphospecies and correctly define genotypes within each morphospecies and (2) assess the macroevolutionary history of the Neogloboquadrinids family. To achieve these objectives, we cloned and sequenced 206 individuals from unsampled areas. Together with all of the data available in GenBank, we redefined 10 genotypes within the three morphospecies. We tested different evolutionary scenarios using Neighbor-Joining, maximum-likelihood, and Bayesian analyses. We reinterpreted the phylogeny of the family based on a comparison between genetic and geological data. We also analyzed the degree of intra- and interindividual genetic diversity that exists to better define the genetic boundary within a given morphospecies. Finally, we discuss the biogeography of each genetic type in terms of adaptation and dispersal with regards to paleo-oceanographic changes over the last million years.

5.3. Material and methods

5.3.1. Organism collection, DNA extraction, amplification, and sequencing

Samples were collected using plankton nets (100 μm mesh size) and vertical or subvertical tows to filter water from 200 m depth to the sea surface. In total, 130 stations were visited during the *AMT-5* (September–October, 1997), *AMT-8* (May–June, 1999), *OISO-4* (January–February, 2000), *Revelle-2001* (January, 2001), *Melville-2003* (May–June, 2003), and *BJ8-2003* (July, 2003) cruises, as well as cruises off the coasts of Bermuda (April–May, 1996), Santa Barbara (February–March, 1998), Puerto-Rico (March–April, 1997), and Guam (December, 1999) (Figure 5.2). The three Neogloboquadrinid morphospecies (*Neogloboquadrina dutertrei*, *N. pachyderma*, and

Pulleniatina obliquiloculata) were isolated from the samples using a dissecting microscope and transferred to Petri dishes containing filtered sea water. Specimens were then individually cleaned with a brush to remove the detritus and microorganisms from their surface. The single foraminifers were then transferred in a special buffer, GITC* (de Vargas et al. unpublished), which allows DNA extraction while preserving the calcareous test. A ~750 bp fragment localized at the 3' end of the SSU rDNA was amplified by PCR using the foraminiferal specific primers S15rf (5'-GTGCATGGCCGTTCTTAGTTC-3' coupled with S19f (5'CCCGTACTAGGCATTCCTAG-3'). The amplified PCR products were ligated into the pGEM-T Vector System (Promega Co., Ltd., Madison, WI, USA) following the manufacturer's instructions. Positive clones were submitted to colony PCR, purified, and sequenced on a 3100-Avant Genetic Analyser**.** Two to twelve rDNA clones were sequenced for thirteen individuals to examine the species boundaries for each genotype.

5.3.2. DNA sequences analyses

Partial SSU rDNA sequences were manually aligned using the Genetic Data Environment (GDE) 2.2 software (Larsen et al. 1993). The Neighbor-Joining method (NJ Saitou and Nei 1987), the maximum likelihood method (ML) (Felsenstein 1981), and Bayesian statistics were used to reconstruct the family phylogeny. NJ analyses were performed using the Phylo_win program (Galtier et al. 1996), and distances were corrected for multiple hits according to the Tajima and Nei substitution model (Tajima and Nei 1984). ML analyses were performed using PAUP* version 4.0b10 (Swofford 2002) and PhyML (Guindon et al. 2005). The `ModelTest` program (Posada and

Crandall 1998) was used to choose the DNA substitution model that was most appropriate to analyze our data. Nonparametric bootstrapping (Felsenstein 1985) was performed with 1000 replicates for both ML and distance analyses. Bayesian phylogenetic analyses were conducted for all species trees with MrBayes 3.0 (Huelsenbeck and Ronquist 2001) under the same model selected for ML. Each Markow chain was started from a random tree and run for 10^6 generations; the chains were sampled every 100^{th} cycle. All sample points prior to reaching stationary were discarded as burn-in samples. Three to five representatives from each genotype were used to reconstruct the family phylogeny, both including and excluding the highly divergent right-coiling *N. pachyderma*. Two sets of data were explored in parallel: One included all available unambiguously aligned nucleotide sites (N = 774) and the other contained 550 sites, from which the most highly variable sites had been removed**.** Three samples of *Gloiborotalia inflata* were used as outgroups. NJ and MP were used to build the phylogenetic trees within each morphospecies. Intraspecies and colonial pairwise genetic distance (Tajima and Nei 1984) were calculated using Mega3 (Kumar et al. 2004). Relative-rate tests (Robinson-Rechavi and Huchon 2000) were performed both including and excluding *N. pachyderma* right coiling to compare substitution rates between each lineage of the species complex.

5.3.3. Restriction fragment length polymorphism (RFLP) analysis

PCR product digestions were performed using the endonuclease *PsiI*, which cuts the nucleotide sequence at specific sites. The restriction enzyme was selected to rapidly discriminate between the different *P. obliquiloculata* genotypes. The protocol used is as follows: 12.5 µl of the ~1000 bp *SSU* rDNA PCR-products were directly digested

for 5 hours at 37 °C in a total volume of 25 µl containing 2.5 µl of the diluted enzyme (1.25 units), 2.5 µl of 10X-buffer (Roche), and 7.5 µl of distilled H_2O. Distinct patterns for each genotype were UV detected after migration of the digested PCR-products on 2% agarose gel and ethidium bromide coloration.

5.4. Results

5.4.1. Macroevolution of the Neogloboquadrinid species complex

We distinguished 10 distinct genetic types within the 206 specimens that we examined from the worldwide ocean. Specifically, we redefined four genotypes within the left-coiling *N. pachyderma* and identified two genotypes within the right-coiling *N. pachyderma*; we discovered for the first time two discrete genotypes within *P. obliquiloculata*; and we found two major genotypes within *N. dutertrei*. Further details on each type will be provided later.

The Neogloboquadrinids are characterized by significant variations in the rates of rDNA substitution. To test the significance of these variations, we performed relative rate tests that included and excluded the highly divergent right-coiling *N. pachyderma*. Results clearly demonstrate that the left-coiling *N. pachyderma* has a significantly different rate of evolution compared to the rest of the lineages of the family, in addition to the well-known fast-evolving right-coiling *N. pachyderma* (Table 5.1).

The resulting SSU rDNA phylogeny of the species complex was evaluated using two different methods (Figure 5.3). Tree A was reconstructed by the PhyML method and Tree B was the best supported tree reconstructed by Bayesian analyses (excluding the fast-evolving right-coiling *N. pachyderma*). Both topologies contradict

the monophyly of *P. obliquiloculata* and *N. dutertrei* that were derived from fossil records. To better resolve this contradiction, we tested two hypotheses: Hypothesis A states that the two coiling *N. pachyderma* form a monophyletic group and that *P. obliquiloculata* and *N. dutertrei* form a monophyletic group, whereas hypothesis B rejects one or both of the premises stated in A (see Figure 5.4). Out of the 16 trees we reconstructed using the different methods, 11 trees supported hypothesis B by rejecting the monophyly of *P. obliquiloculata* and *N. dutertrei*. Most of the analyses supported the monophyly of the two coiling types of *N. pachyerma*; only two trees rejected it (Figure 5.4).

5.4.2. Species definition in non-spinose foraminifera

To illustrate the evolutionary relationship within each morphospecies, we reconstructed unrooted NJ trees within each morphospecies (Figures 5.5 and 5.7). We compared the intraindividual variation to the genetic distance that defined the genotypic boundary using 45 clones retrieved from 17 individuals that were randomly selected from all genotypes (Table 5.2). The average intraindividual distance ranged from 0.0015 to 0.008; values were as small as 2.1% in left-coiling *N. pachyderma* or as large as 46% in *N. dutertrei* when compared to the distance that defined the genotypic boundary. This result strengthens our premise that we need to evaluate intraindividual variability before defining any cryptic species.

We identified two genotypes (RI and RII) within the right-coiling *N. pachyderma*. The NJ tree is shown in Figure 5.5. RI and RII are separated by a 7.2% distance and with 100% bootstrap support. RI was first identified by Darling (2000) as a bipolar genotype that inhabits both the subpolar Arctic and the subpolar Antarctic;

RII was defined from the samples collected in the Santa Barbara Channel (Darling et al. 2003). We sequenced 24 specimens of right-coiling *N. pachyderma*, which were all type RI**,** and the four specimens from Japan's Tsugaru Strait all were type RII**.**

We next re-evaluated the genetic diversity within the left-coiling *N. pachyderma*. Four distinct genotypes were clearly presented in the tree, each forming a monophyletic group with 98–100% bootstrap support (Figure 5.5). The principal division occurred between Type IV and Types I–III; this result presents a major disagreement with previous data (Darling et al. 2004; Darling et al. 2007). Therefore, we tested the two alternative hypotheses by reconstructing the phylogeny using different methods and two sets of alignments including and excluding the most variable sites. Of our 24 trees, 22 supported our conclusion that the first early divergence separated Type IV from the remaining genotypes (Figure 5.6). The sequences from Japan's Tsugaru Strait form a new genotype (III) that has not been previously reported. We further found that the previously defined genetic types II, III, and V are in fact a single genetic type with bootstrapping support of > 99% from both MP and NJ methods when more comprehensive samples are included in the analysis. Therefore, we named this single genetic type Type II, which includes the previously defined genotypes II, III, and V (Darling et al. 2004; Darling et al. 2007). Type IV was defined previously by Darling (2004) as being located in Antarctic cold water, south of the polar front, and in the Bellingshausen Sea. Twenty-two of the samples of left-coiling *N. pachyderma* we analyzed were Type IV and two were Type III.

Two new genotypes were identified within *P. obliquiloculata* for the first time (see Figure 5.7). At least two genotypes were defined within *N. dutertrei*. Compared

with previous data from GenBank, our new genotype type II is identical to the *N. dutertrei* genotype Ic (Darling et al. 2003) and type I includes the *N. dutertrei* genotype Ib (Darling et al, 2003), but with a much higher genetic diversity.

5.4.3. Biogeographic distribution across hemispheres and basins

The right-coiling *N. pachyderma* type RI is distributed globally between 65°N and 52.5°S instead of being a bipolar species, as reported previously (Darling et al. 2004; Darling et al. 2000). This discrepancy may be because previous studies did not sample any non-polar regions. The right-coiling *N. pachyderma* type RII was found exclusively in the Pacific Ocean in the Santa Barbara Channel and in Japan's Tsugaru Strait. Both sites are channels in which different water masses converge, which results in relatively large-scale differences in hydrological parameters.

We found that each genotype of the left-coiling *N. pachyderma* seems to be adapted to a specific hydrographic or trophic environment. Type L I occurs in the Arctic, whereas Type L II occurs in the Southern Ocean in the sub-Antarctic area. We found Type III in the Tsugaru Strait, Japan and Type L IV in the Antarctic's cold water. L II and L IV were both Antarctic genotypes, although L IV seemed to be more successful in colder water.

N. dutertrei is composed of two genotypes: Type I seems to be cosmopolitan whereas Type II is present only in the Eastern Pacific (Figure 5.8).

Due to sampling limitations, we do not have a comprehensive understanding of the distribution of the two genotypes of *P. obliquiloculata*; however, both types inhabit the Indo-Pacific Ocean in the Indonesian Straits, and Type II was found alone in the subtropical Atlantic Ocean (Figure 5.9).

5.5. Discussion

5.5.1. Species definition in planktonic foraminifera

SSU rDNA is considered to be highly conservative at the species or higher taxonomic level; thus, it has been used widely as a marker to reconstruct the phylogeny of planktonic foraminifera (Darling et al. 2004; de Vargas et al. 1997; Pawlowski et al. 1996). Direct sequencing from PCR or purified PCR products was used in most of these studies, with the assumption that the variation among the multiple gene copies of the *SSU* rDNA is minor and does not introduce substantial noise to the topology at the species or higher taxonomic levels (Darling et al. 2004; Darling et al. 2007; Darling et al. 2003; Darling et al. 2000; de Vargas et al. 2002). Therefore, the extent to which intraindividual rDNA polymorphism influences our definition of cryptic species in planktonic foraminifera has not yet been adequately addressed, although Darling (2007) did recently sequence multiple clones from a single individual.

The intraindividual genetic distance calculated in our study shows that the values are large enough to bias our definition of the genotypes within a morphospecies. For example, within one *N. dutertrei* specimen, AMT8_538, the intraindividual genetic distance (0.011) was almost equal to the genetic distance between the two genotypes of this species (0.013). The intraindividual genetic distance seems species independent, and the genetic plasticity of a species may be essential for us to understand its process of evolution.

5.5.2. Bipolarity and biogeography

Bipolar distributions appear to be common in many planktonic marine species (Lazarus 1983). As one of the major disjoint distribution patterns on Earth, bipolarity has been frequently studied and many hypotheses have been put forward to uncover the evolutionary processes behind this phenomenon. With recent molecular data becoming available, this issue can be directly investigated. Studies of dinoflagellates (Montresor et al. 2003) and planktonic forminifera (Darling et al. 2007; Darling et al. 2000; de Vargas et al. 2001) have provided evidence of genetic divergence between different populations. For example, using a calibrated molecular clock, Darling (2004) theorized a mechanism for the diversification of foraminifera species starting with the allopatric isolation of Arctic and Antarctic *N. pachyderma* populations after the onset of the Northern Hemisphere's glaciations. A more recent study further discussed the evolutionary processes that drove the divergence of *N. pachyderma* on a global scale (Darling et al. 2007) However, when global-scale sampling was performed we found that the right-coiling *N. pachyderma* type I that was previously described as a bipolar species actually is rather cosmopolitan. We found 22 individuals of right-coiling *N. pachyderma* type RI in warm water masses; specifically, we identified 9 individuals between 37.5°N and 52°N, 2 individuals at 15°N off the Mauritanian coast, and 11 individuals between 37.5°S and 52.5°S. On the other hand, we did not find bipolarity in left-coiling *N. pachyderma*, in that no genetic affinity was shared between populations living in the Arctic and Antarctic regions. What is also striking is the difference between the Pacific population compared to Indo-Atlantic ones: Types RII and LIII are endemic to the Pacific

5.5.3. Diversification in left-coiling *N. pachyderma*

Instead of isolation between northern and southern hemisphere populations since the early Pleistocene, as suggested by Darling et al. (2004), we propose a different evolutionary scenario to explain the diversification of left-coiling *N. pachyderma*. Our results suggest that genetic mixing may have occurred more than once since the early Pleistocene. Our detailed phylogenetic analyses within left-coiling *N. pachyderma* show that Antarctic polar type IV was the ancestral genotype, and it evolved first. This ancestor then gave rise to type I in the Arctic Ocean, and later in geological time the divergence between type II (subAntarctic) and type III (Tsugaru Strait, Japan) occurred. A possible explanation for this pattern is that significant oceanic cooling associated with the Pleistocene glacial stages might have enabled the latitudinal range of *N. pachyderma* to be extended closer to the Equator in some areas, thereby enabling reasonably free interchange of the northern and southern hemisphere populations. Morphological data show that Arctic populations have more in common with subantarctic populations than with Antarctic populations (Kennett 1968), which also supports the evolutionary scenario that mixing occurred between the northern and southern hemispheres.

5.6. Conclusion

Overall, we identified 10 discrete genotypes in the family Neogloboquadrinids. We redefined the cryptic species, as each type corresponds to a monophyletic group of slightly divergent ribotypes that might vary as much between as within individuals. The reconstructed molecular phylogeny of this family indicates that the left-coiling and right-coiling *N. pachyderma* cluster together, which is consistent with current

interpretations of the fossil record. *Neogloboquadrina dutertrei* and *P. obliquiloculata*, however, might represent a divergence that occurred earlier than the 7 Ma derived from the fossil record. We have demonstrated that cryptic speciation in planktonic foraminifera could be more complicated than previously believed, given the large global scale of their habitats and their as yet poorly known genetic complexity (as illustrated by the intra-individual genetic variations reported herein).

5.7. Tables

Table 5.1: Relative rate tests between the four main Neogloboquadrinid species complexes.

Including right-coiling *N. pachyderma*		
Lineage 1	**Lineage 2**	**P-Value**
N. dutertrei	*P. obliquiloculata*	*0.806359*
N. dutertrei	*N. pachyderma* Left	0.00017219 *
N. dutertrei	*N. pachyderma* Right	1.00E-07 **
P. obliquiloculata	*N. pachyderma* Left	0.000414967 *
P. obliquiloculata	*N. pachyderma* Right	1.00E-07 **
N. pachyderma Left	*N. pachyderma* Right	1.00E-07 **
Excluding right-coiling *N. pachyderma*		
Lineage 1	**Lineage 2**	**P-Value**
N. dutertrei	*P. obliquiloculata*	*0.806082*
N. dutertrei	*N. pachyderma* Left	0.000304786 *
P. obliquiloculata	*N. pachyderma* Left	0.000703864 *

Table 5.2: Interindividual versus intraindividual genetic diversity (genetic distance and number of genetic substitutions) within the Neogloboquadrinid species complex.

I. P. obliquiloculata	**% genetic distance (Tajima & Nei's, 1984) / # of substitutions (Standard Errors)**
Overall Mean Value	0.012 / 7.008 (SE: 0.003 / 1.582)
Within Type I	0.004 / 2.218 (SE: 0.001 / 0.566)
Within Type II	0.006 / 3.550 (SE: 0.002 / 0.993)

Between Type I & Type II	0.02 / 11.397	(SE: 0.005 / 2.951)
Mean value within colonies	0.003/2.000	(SE: 0.002 / 0.713)
II. *N. pachyderma* right coiling		
Overall Mean Value	0.018/11.361	(SE: 0.002/1.364)
Within Type R_I	0.003/2.209	(SE: 0.001/0.436)
Within Type R_II	0.006/3.800	(SE: 0.002/1.130)
Between Type R_I &Type R_II	0.072/45.368	(SE: 0.011/6.218)
Mean value within colonies	0.0015/1.15	(SE: 0.001/0.631)
III. *N. pachyderma* left coiling		
Overall Mean Value	0.044/25.605	(SE: 0.005/2.636)
Within Type II	0.028/17.333	(SE: 0.005/2.8000)
Within Type III	0.009/5.667	(SE: 0.003/1.634)
Within Type IV	0.008/4.709	(SE: 0.002/1.190)
Between Types (mean value)	0.067/38.45	(SE: 0.009/5.276
Mean value within colonies	0.008/5.000	(SE: 0.0015/1.38)
IV. *N. dutertrei*		
Overall Mean Value	0.008/4.855	(SE: 0.002/1.163)
Within Type I (Type Ib)	0.007/4.008	(SE: 0.002/1.012)
Within Type II (Type Ic)	0.001/0.756	(SE: 0.001/0.519)
Between Type I &Type II	0.013/7.624	(SE: 0.004/2.084)
Mean value within colonies	0.006/2.708	(SE: 0.002/1.226)

5.8. Figures

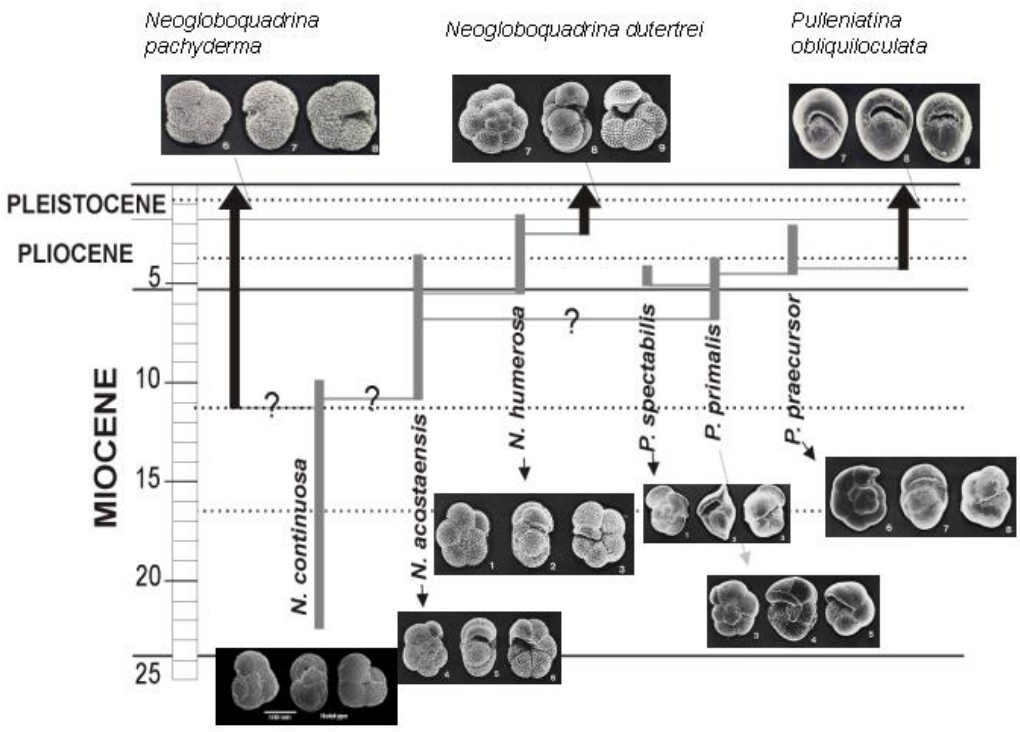


Figure 5.1: Fossil record of the Neogene non-spinose planktonic foraminifera, with a focus on the Neogloboquadrinids. A. The Neogloboquadrinids form a natural assemblage containing two genera–*Neogloboquadrina* and *Pulleniatina*-divided into nine morphological species, whose stratigraphic range (modified from (Kennett and Srinivasan 1983) and SEM pictures are depicted here. B. Stratigraphic record of the 62 morphospecies of Neogene globorotaliid-like, non-spinose, planktonic foraminifera, the morphological group to which the Neogloboquadrinids belong. This major group can be split into eight morphological subgroups (names below parentheses), whose origin and phylogenetic relationships are mostly uncertain (question marks).

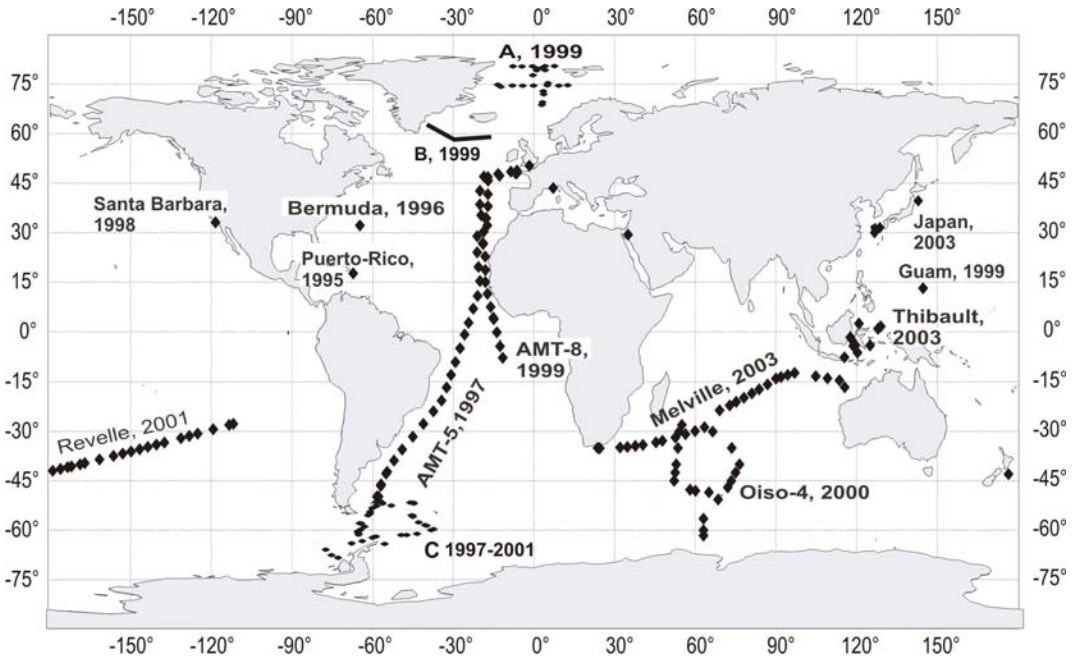


Figure 5.2: Location of sampling sites where living planktonic foraminifera were collected. The black diamonds depict the 138 stations studied in this paper. Stations and cruise tracks by Darling et al. (2000, 2004) are shown under A, B, and C. Data from Japan are from Kimoto and Tsuchiya (2003 submitted). Cruise names and dates are also indicated. For detailed assessment of where each species was collected, refer to the *CalcOBIS* web-site, http://marine.rutgers.edu/MEEOOP/CalcOBIS/.

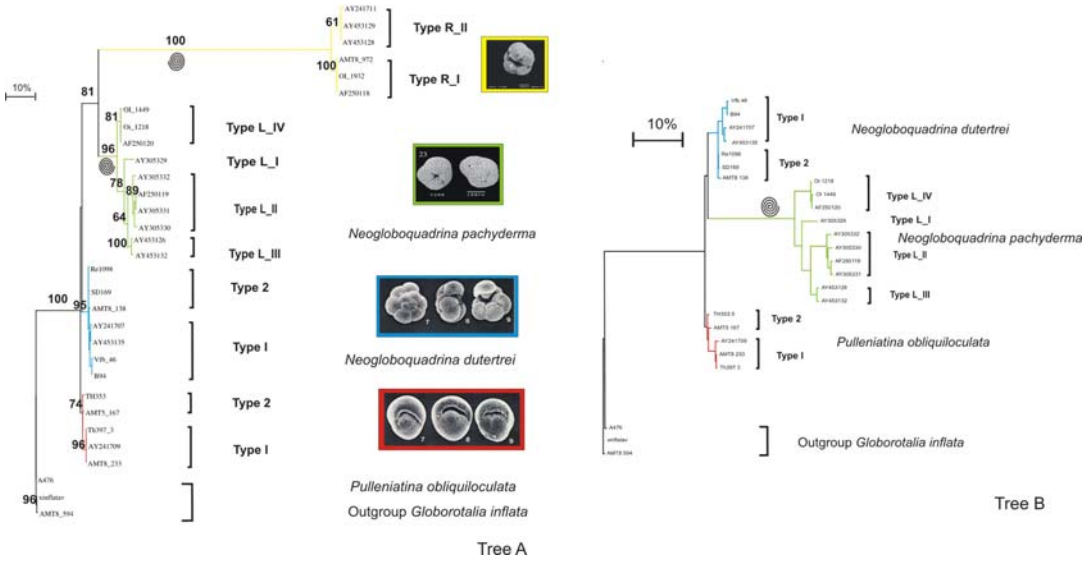


Figure 5.3: Phylogenetic trees of the different SSU rDNA genotypes we detected among 206 samples (left-coiling *N. pachyderma*, n = 16; right-coiling *N. pachyderma*, n = 33; *N. dutertrei*, n = 67; *P. obliquiloculata*, n = 90). Trees were constructed both including (tree A) (PhyML) and excluding (tree B) (Bayesian analyses) the highly divergent right-coiling *N. pachyderma*. Three samples of *G. inflata* were used as outgroups. Bootstrap values (PhyML) indicate support for branches in tree A.

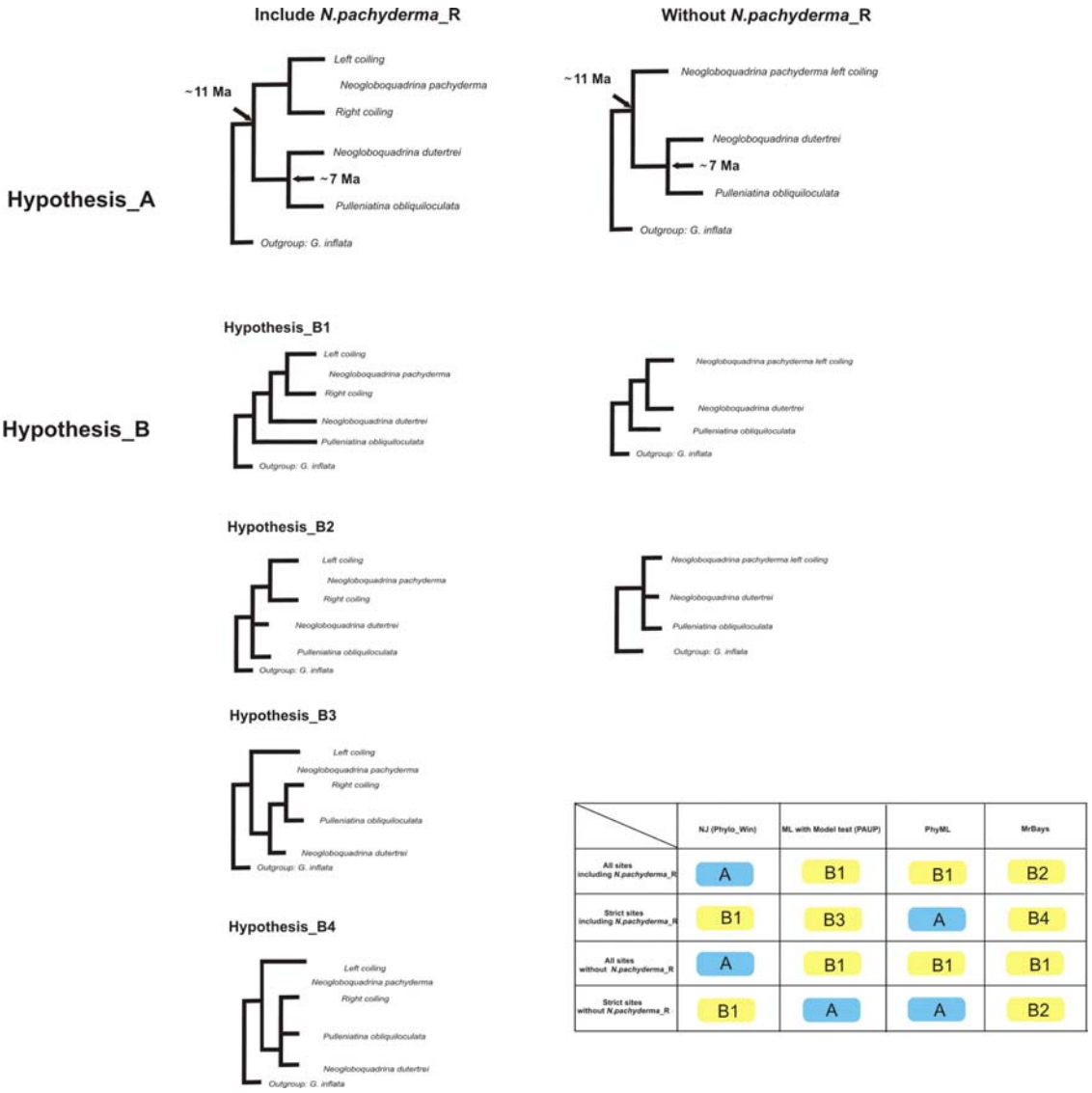


Figure 5.4: Tests of the monophyly of left- and right-coiling *N. pachyderma* and the monophyly of *P. obliquiloculata* and *N. dutertrei* using 16 phylogenetic trees (methods and different datasets are shown in the table). Hypothesis A = monophyly present throughout the evolutionary history of the Neogloboquadrinids; hypothesis B = either or neither of the monophyly premises is true.

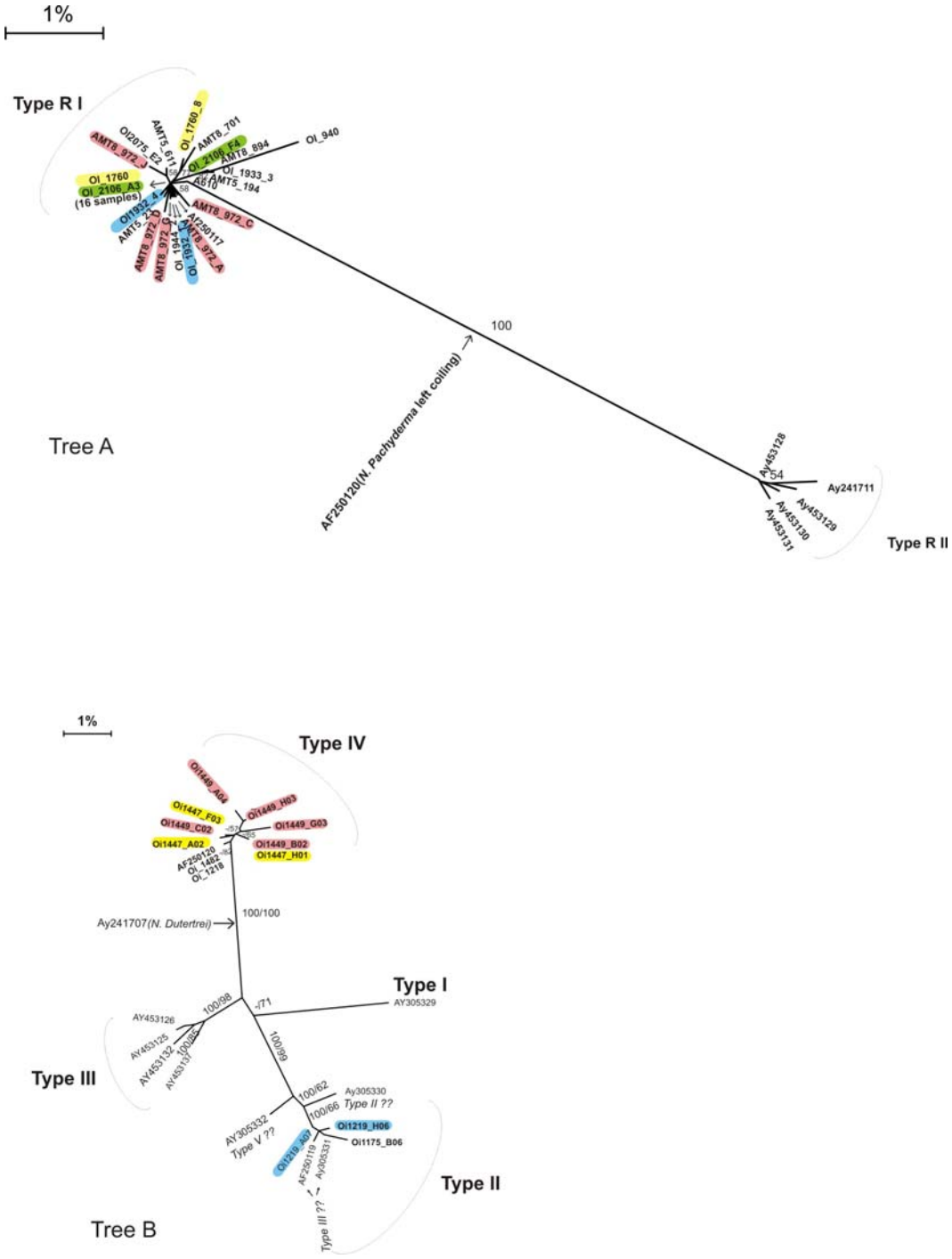
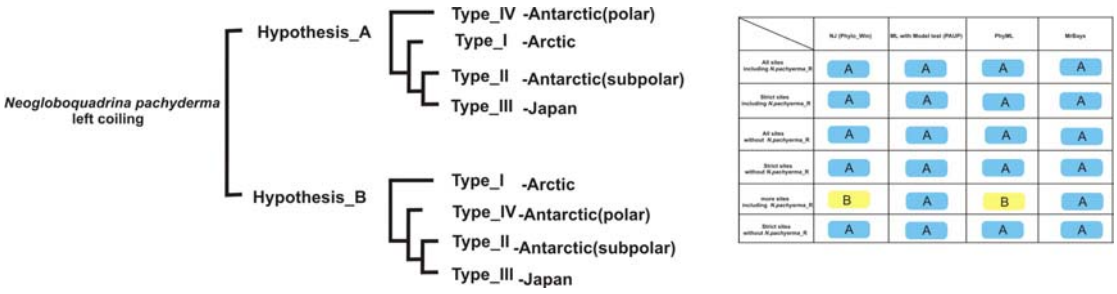


Figure 5.5: Unrooted NJ trees within right-coiling (tree A) and left-coiling (tree B) *N. pachyderma*. Bootstrap percentages (1000 replicates) for MP and NJ methods are given next to each internal branch of the tree (only > 50%). The root for each tree is indicated by an arrow with the outgroup species name. Each color shows the positions for each colony of one individual**.**



	A	A	A	A
	A	A	A	A
	A	A	A	A
	A	A	A	A
	B	A	B	A
	A	A	A	A

Figure 5.6: Tests of the first division of the different genotypes within the left-coiling *N. pachyderma* using 24 phylogenetic trees (methods and different datasets are shown in the table). Hypothesis A = the first division is between Type IV and Type I–III; hypothesis B = it is between Type I and Type II–IV.

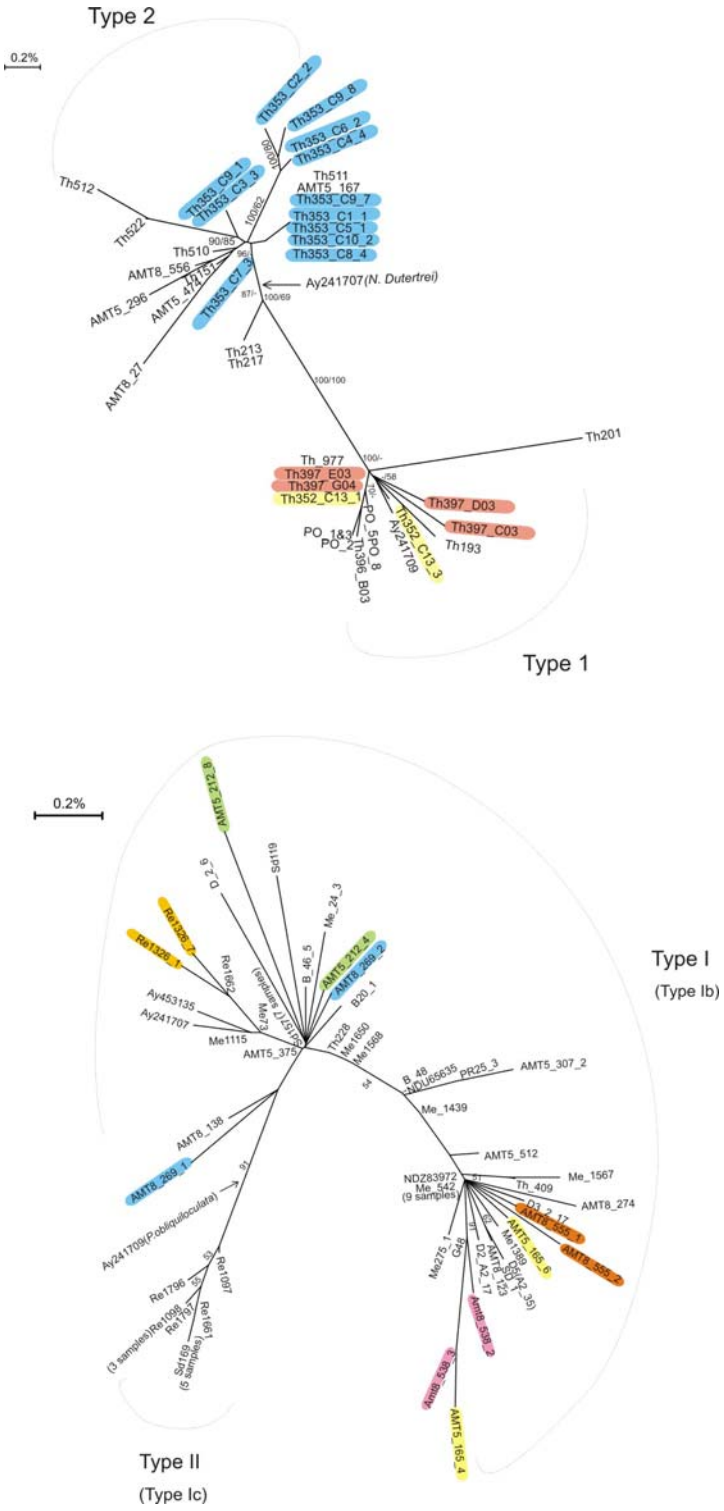


Figure 5.7: Unrooted NJ trees within *P. obliquiloculata* (tree A) and *N. dutertrei* (tree B). Bootstrap percentages (1000 replicates) for MP and NJ methods are given next to each internal branch of the tree (only > 50%). The root for each tree is indicated by an arrow with the outgroup species name. Each one color shows the positions for each colony of one individual.

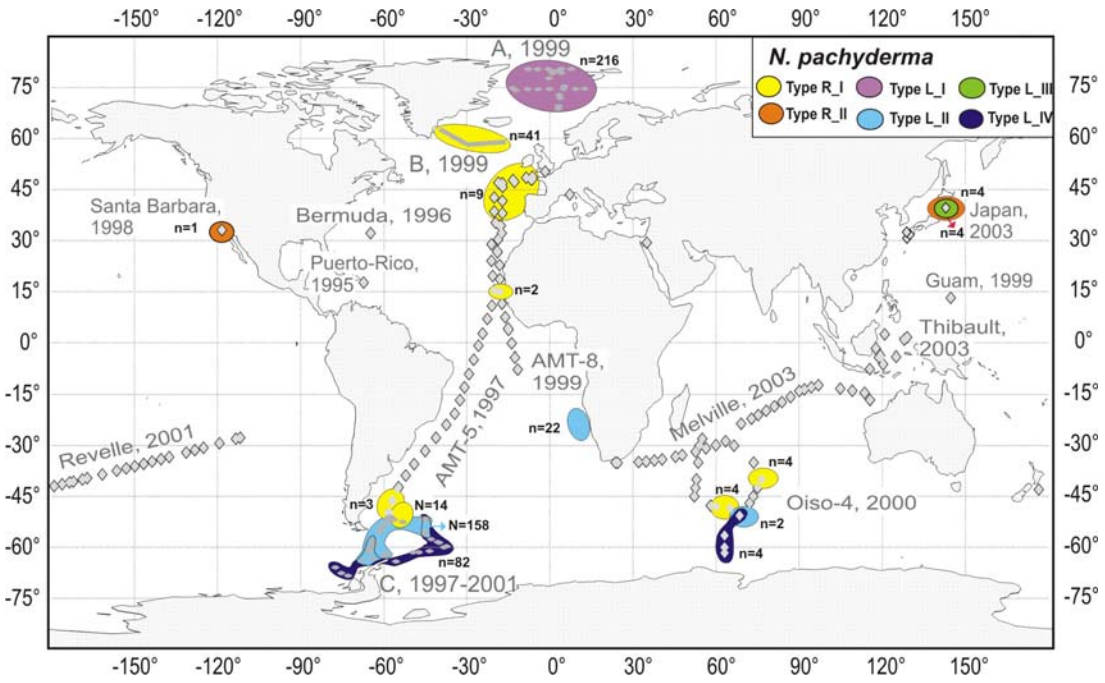


Figure 5.8: Distribution map of the different genotypes of left- and right-coiling *N. pachyderma*. Each genotype is indicated by a different color. The number of individuals for each genotype found in a specific area is indicated.

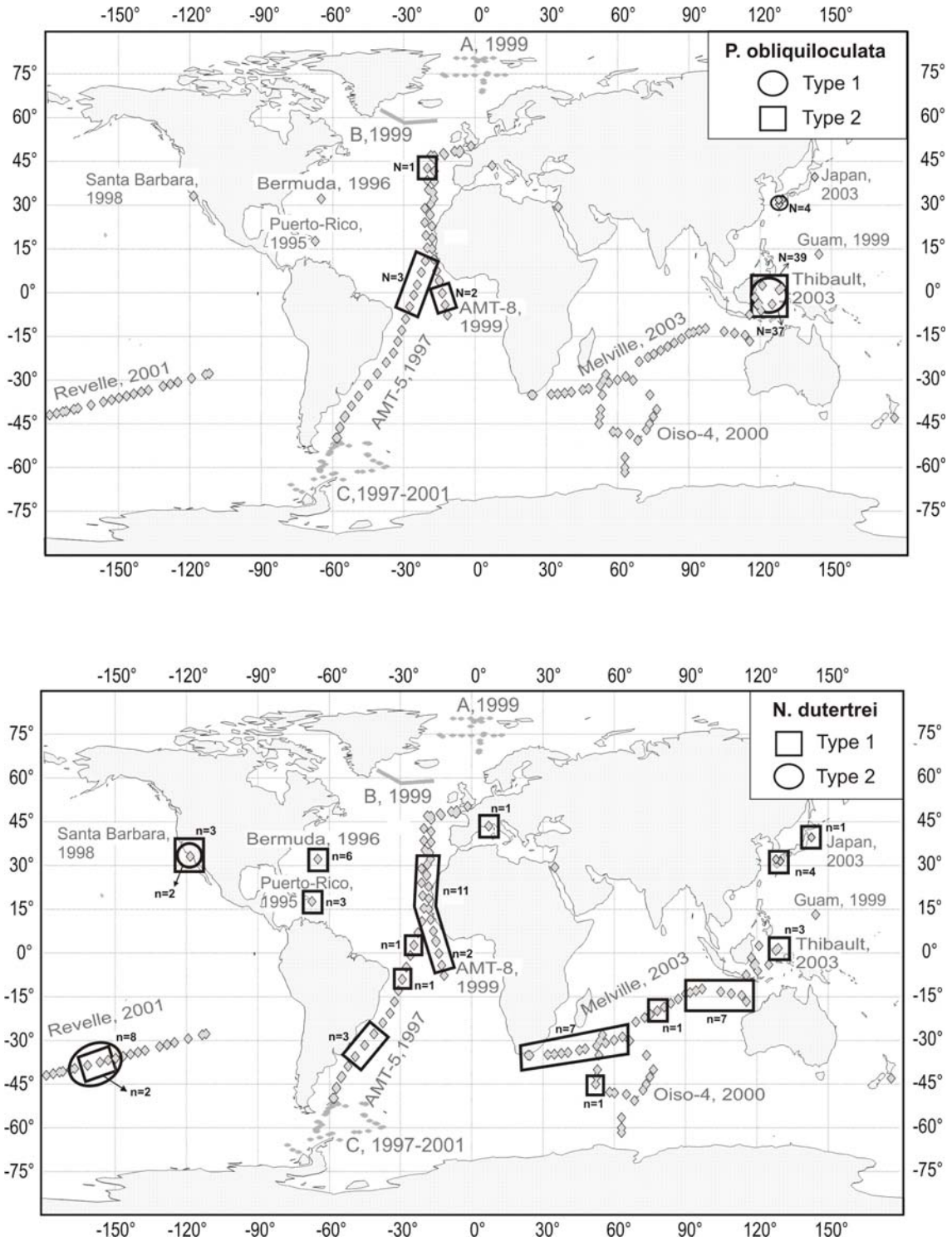


Figure 5.9: Distribution of the different genotypes of *P. obliquiloculata* (map A) and *N. dutertrei* (map B). A circle or square represents one genotype. The number of individuals for each genotype found in a specific area is indicated.

Chapter 6

6.0. Conclusion and Perspectives

6.1. Conclusions and contributions

The research described herein yielded new and exciting insights about marine protistan phylogeny, diversity, and biogeography. The major contributions of this dissertation include establishment a molecular timeline for haptophyte evolution and the discovery and description of the dramatic diversity of novel haptophyte lineages, which are responsible for the majority of global photosynthesis in modern oceans. The first part of the thesis focused on understanding the macroevolution of the haptophytes. Based on extensive analysis of a multi-gene dataset using several Bayesian models, I presented a robust and extensive phylogeny of the haptophytes and dated the origin of this group back to 824 MYA (95% highest probability density 1031–637 MYA). The use of maximum likelihood reconstruction of ancestral characters provided more precise dating of the key diversification and transition events of haptophyte evolution. For example, the ability to calcify and the transition from mixotrophy to autotrophy evolved between 329–291 MYA in the Carboniferous period. I also presented several scenarios about the origin of algal plastids, the transition of trophic modes, and the emergence of organic scales in planktonic protists; these scenarios can be more rigorously tested in future studies using dated phylogenies of multiple groups of marine protists that have such characteristics. The interpretation of the timing of these key evolutionary transitions in both ecological and geological contexts provides a better understanding of how the diversification of haptophytes may have been correlated with past environmental changes. In turn, this understanding may help us predict how these species will react to the rapid climate change

and ocean acidification that currently is taking place. The dataset and phylogeny generated in this part of the study also served as a taxonomically constrained framework to anchor the second part of the study: the analysis of global environmental diversity of haptophytes.

The second part of the thesis focused on understanding the extent of haptophyte biodiversity, including the tiny, naked, or poorly calcifying taxa as well as the calcifying taxa. A growing number of studies have reported the presence of unanticipated levels of diversity and many undescribed protistan taxa through cloning and sequencing of natural assemblages from numerous marine ecosystems (Countway et al. 2007; Lovejoy et al. 2007; Massana et al. 2004a; Romari and Vaulot 2004). These studies altered our comprehension of the overall diversity, composition, and function of protistan assemblages. However, prior to this thesis, the diversity of the Haptophyta has been largely unexplored, primarily due to the GC-rich rDNAs they possess, which cause them to be virtually missing from environmental clone libraries produced by classical PCR amplification protocols using universal eukaryotic primers. To address this issue, I used a novel genetic protocol adapted for GC-rich genomes and Haptophyta-specific primers and discovered dramatic diversity among the non-calcifying haptophytes. An estimation of depth-integrated relative abundance of 19-Hex during the year 2000 suggested that the biomass of these organisms might be as much as two times greater than that of cyanobacteria or diatoms. This finding helped explain an important oceanographic paradox: the omnipresence in seawater of a photosynthetic pigment (19-hexanoyloxyfucoxanthin) borne by an unsuspected diversity of organisms.

I continued my exploration of haptophyte diversity by studying larger calcifying cells, the coccolithophores. The use of a combined molecular and morphological approach

allowed me to decipher the genetic, morphological, and ecological variations that define species and to assess more precisely the diversity of coccolithophores in natural communities. As the first detailed attempt to reconcile the working definition of species at the morphological and genetic levels, this part of the thesis showed that the threshold at which phylospecies and morphospecies are defined varies across different natural communities. This finding has severe implications with respect to our evaluation of diversity as estimated from metagenomic approaches. This research also revealed high levels of genetic diversity in the Syracosphaeraceae and Umbellosphaeraceae. Future genetic surveys using a set of more specific primers at the family or genus levels will be very helpful for investigating the diversity and biogeographic distributions of these groups. The combination of morphological and genetic analyses improved our understanding of the spatio-temporal dynamics of morphological versus genetic species-level differentiations, and the detailed phylogenetic analysis within each family or genus helped to uncover possible occurrences of cryptic speciation in a few ecologically important morphospecies. The latter groups will require further single-cell or culture-based genetic studies.

In the last part of the thesis, I used the Neogloboquadrinids, a family of non-spinose planktonic foraminifera, as a model to study cryptic speciation and biogeography in marine protists. I reinterpreted the phylogeny of the family based on a comparison of genetic and geological data and identified the biogeographic patterns of each genetic type. The genetic types inhabiting the equatorial to temperate waters have transbasin and transhemispheric and likely continuous distributions, whereas the types collected in subpolar and polar waters have circumglobal but monopolar distributions. The study demonstrated that

cryptic speciation in planktonic foraminifera could be more complicated than previously believed, given the large global scale of their habitats and their as yet poorly known genetic complexity. The illustration of intra-species vs. intra-individual genetic variations also makes this group a useful case study for understanding common problems in the use of sequence data.

6.2. Perspectives and Future Challenges

The results of this study opened up several new avenues of study that require follow-up. First, the study unveiled a dramatic and ancient diversity of unique photosynthetic picoplanktonic protists within the Haptophyta. Size analysis of these cells identified by haptophyte-specific fluorescent probes indicated that they are ~4 μm (maximum of 8–9 μm), which may explain their critical roles in organic carbon fluxes on a global scale. The phylogenetic analysis indicated that the majority of the diversity lies in the Chrysochromulina-B2 clade (Fig. 3.3), but almost nothing is known about the morphology and ecophysiology of members of this clade, and this hampered our understanding of how they function in the ocean's dynamics. Thus, a critical next step is to obtain culture representatives of these novel species. However, it is difficult to isolate such tiny organisms from a field sample, and the sorting process often is biased towards fast-growing types. Moreover, it is challenging to maintain these organisms without prior knowledge of their ecophysiology. Nevertheless, successful culture of a few representatives of these organisms will provide opportunities to intensively investigate their morphology, ecophysiology, and genetics in the lab and will help us to better interpret the metaphylogenomic data collected from the global ocean. Second, the clone libraries reconstructed in this study from eight water samples collected from the worldwide ocean

revealed a dramatic diversity among both calcifying and non-calcifying haptophytes. However, the rarefaction curves from all samples indicated that the current sequencing effort (~200 sequences analyzed per library) is far from exhaustive. Much more extensive large-scale environmental sequencing is required to fill in the protistan gap in the tree of life. Recent techniques, such as 454 tag sequencing, which allows massive parallel sequencing at a relatively low cost, are powerful tools that can generate large datasets about the genetic composition, species diversity, and interactions between different species in the marine environment. The 454 tag sequencing method first applied to bacteria (Kysela et al. 2005; Neufeld et al. 2004), then to archaea (Huber et al. 2007), estimated the species richness at approximately 30,000 (for bacteria) and 3,000 (for archaea) per liter of seawater; moreover, it indicated the presence of a "seed bank" of ancient and rare taxa in marine prokaryotes, termed the "rare biosphere " (Sogin et al. 2006). The application of this technique to protists has lagged primarily due to extreme variations in eukaryotic *SSU* rRNA gene copy number and to the limitation of read length in the early 454-pyrosequencing system. With recent advances in pyrosequencing technology with read lengths of up to 240 nucleotides and a newly designed protocol suitable for exploring protistan diversity (Amaral-Zettle et al. 2009), several exciting projects (e.g., ICoMM, MIRADA-LTERs) have begun to explore the "rare protistan biosphere" using massive parallel tag sequencing. Another project (POSEIDON, as part of Tara-OceanNs) proposes to generate more than half a million 454 sequence reads from 10 proposed open oceanic stations along a circumglobal cruise. The ability to sequence "deeply" into the huge diversity that characterizes microbial eukaryote assemblages is a critical step in understanding why such high-level environmental genetic diversity exists and how the

many species function in complex ecological systems.

Research on the form and function of protistan diversity continues to be scientifically fascinating. Novel hypotheses about the evolution of protists will continue to emerge and be tested. Which protistan taxa are contributing to community function in a given space and at a given time will be identified, and how changes in assemblage structure relate to the overall protistan diversity and community function will be explained.

References:

Acinas SG, Sarma-Rupavtarm R, Klepac-Ceraj V, Polz MF (2005) PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. Appl. Environ. Microbiol. 71:8966-8969

Alverson AJ, Kolnick L (2005) Intragenomic nucleotide polymorphism among small subunit (18S) rDNA paralogs in the diatom genus within the diatom Skeletonema costatum (Bacillariophyta). Journal of Phycology 41:1248-1257

Amato AK, Levialdi Ghiron JH, Mann DG, Pröschold T, M M (2007) Reproductive Isolation among Sympatric Cryptic Species in Marine Diatoms. Protist 158:193-207.

Andersen RA, Bidigare RR, Keller MD, Latasa M (1996) A comparison of HPLC pigment signatures and electron microscopic observations for oligotrophic waters of the North Atlantic and Pacific Oceans Deep Sea Research Part II: Topical Studies in Oceanography 43:517-537

Aris-Brosou S (2007) Dating phylogenies with hybrid local molecular clocks. PLoS ONE 2:e879

Aubry M-P (1989) Phylogenetically based calcareous nannofossil taxonomy: Implications for the interpretation of geological events. . In: E. VHS, Crux JA (eds) Nannofossils and their Applications - Proceedings of the INA Conference, London , Ellis Horwood, Chichester: 21-40

Aubry M-P (2009) A Sea of Lilliputians. In: Twitchett R, Wade BS (eds) Extinction, Dwarfing and the Lilliput Effect. Palaeogeography, Palaeoclimatology, Palaeoecology, Special Publication

Aubry M-P, Bord D (2009) Reshuffling the Cards in the Photic Zone at the Eocene/Oligocene Boundary. In: Koeberl C, Montanari A (eds) The Late Eocene Earth - Hot House, Ice House, and impacts. Geological Society of America Bull

Baldauf SL (2003) The Deep Roots of Eukaryotes. Science 300:1703-1706

Bandy OL (1972) Origin and development of Globorotalia (Turborotalia) pachyderma (Ehrenberg). Micropaleontology 18:294-318

Berggren WA, Kent, D.V., Swisher, C.C., III, and Aubry, M. P., (1995) A revised Cenozoic geochronology and chronostratigraphy. In Geochronology, Time Scales, and Global Stratigraphic Correlation, SEPM (Society for Sedimentary Geology) Special Publication 54:129-212

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

Berney C, Pawlowski J (2006) A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. Proc Biol Sci 273:1867-72

Biegala IC, Not F, Vaulot D, Simon N (2003) Quantitative Assessment of Picoeukaryotes in the Natural Environment by Using Taxon-Specific Oligonucleotide Probes in Association with Tyramide Signal Amplification-Fluorescence In Situ Hybridization and Flow Cytometry. Appl. Environ. Microbiol. 69:5519-5529

Biers EJ, Sun S, Howard EC (2009) Prokaryotic Genomes and Diversity in Surface Ocean Waters: Interrogating the Global Ocean Sampling Metagenome. Appl. Environ. Microbiol. 75:2221-2229

Billard C (1994) Life cycles. The Haptophyte Alage 51:167-186

Blanquart S, Lartillot N (2008) A site- and time-heterogeneous model of amino acid replacement. Mol Biol Evol 25:842-858

Blaxter ML (2004) The promise of a DNA taxonomy. Philos Trans R Soc Lond B Biol Sci 359:669-79

Boeckel B, Baumann K-H (2008) Vertical and lateral variations in coccolithophore community structure across the subtropical frontal zone in the South Atlantic Ocean. Marine Micropaleontology 67:255-273

Bond PL, Hugenholtz P, Keller J, Blackall LL (1995) Bacterial community structures of phosphate-removing and non-phosphate- removing activated sludges from sequencing batch reactors. Appl. Environ. Microbiol. 61:1910-1916

Bown P (2005a) Selective calcareous nannoplankton survivorship at the Cretaceous-Tertiary boundary. Geology 33:653-656

Bown PR (1987) Taxonomy, evolution, and biostratigraphy of Late Triassic–Early Jurassic calcareous nannofossils. Spec. Pap. Paleontol. 38:1–118

Bown PR (1998) Calcareous nannofossil biostratigraphy, Chapman & Hall.

Bown PR (2005b) Calcareous nannoplankton evolution: a tale of two oceans. Micropaleontology 51:299-308

Bown PR, Lees JA, Young JR (2004) Calcareous nannoplankton evolution and diversity through time. In: Thierstein HR, Young JR (eds) Coccolithophores - from molecular processes to global impact. Springer Verlag, Berlin, p 427-554

Brassell SC, Dumitrescu M (2004) Recognition of alkenones in a lower Aptian porcellanite from the west-central Pacific. Org Geochem 35:181-188

Brassell SC, Eglinton G, Marlowe IT, Pflaumann U, Sarnthein M (1986) Molecular stratigraphy: a new tool for climatic assessment. Nature 320:129–133

Brinkmann H, van der Giezen M, Zhou Y, Poncelin de Raucourt G, Philippe H (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst Biol 54:743 - 757

Brocks JJ, Logan GA, Buick R, Summons RE (1999) Archean Molecular Fossils and the Early Rise of Eukaryotes. Science 285:1033-1036

Brown CW, Yodar JA (1994) Coccolithophorid blooms in the global ocean. J Geophys Res 99:7467-7482

Brownlee C, Taylor A (2004) Calcification in coccolithophores: a cellular perspective. In: Thierstein. HR, Young JR (eds) Coccolithophores. from molecular processes to global Impact. Springer-Verlag, Berlin, p 31-49

Buckley TR (2002) Model Misspecification and Probabilistic Tests of Topology: Evidence from Empirical Data Sets. Syst Biol 51:509-523

Bukry D (1978) Biostratigraphy of Cenozoic Marine Sediment by Calcareous Nannofossils. Micropaleontology 24:44-60

Castle DM, Montgomery MT, Kirchman DL (2006) Effects of naphthalene on microbial community composition in the Delaware estuary. FEMS Microbiology Ecology 56:55-63

Cavalier-Smith T (2004) Chromalveolate diversity and cell megaevolution: interplay of membranes, genomes and cytoskeleton. In: Hirt R, P., Horner DS (eds) Organelles, genomes and eukaryotic phylogeny. Taylor and Francis, London, p 75-108

Cavalier-Smith T (2006) Cell evolution and Earth history: stasis and revolution Philos Trans R Soc Lond B Biol Sci 361:969–1006

Chandler DP, Fredrickson JK, Brockman FJ (1997) Effect of PCR template concentration on the composition and distribution of total community 16S rDNA clone libraries. Molecular Ecology 6:475

Chao A (1984) Non-parametric estimation of the number of classes in a population. Scand. J. Stat., 11:265-270

Chao A, Shen T (2003–2005) Program SPADE (species prediction and diversity estimation)

Chisholm SW, Olson RJ, Zettler ER, Goericke R, Waterbury JB, Welschmeyer NA (1988) A novel free-living prochlorophyte abundant in the oceanic euphotic zone. Nature 334:340-343

CLIMAP (1976) The surface of the ice-age Earth. Science 191:131-1144

CLIMAP (1981) Seasonal reconstruction of the Earth's surface at the last glacial maximum. . Geological Society of America Map and Chart Series MC-36: :1-18.

CLIMAP (1984) The last interglacial ocean. Quat. Res 21:123-224

Cole JR, Chai B, Marsh TL, Farris RJ, Wang Q, Kulam SA, Chandra S, McGarrell DM, Schmidt TM, Garrity GM, Tiedje JM (2003) The Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. Nucl. Acids Res. 31:442-443

Conte MH, Thompson A, Lesley D, Harris RP (1998) Genetic and physiological influences on the alkenone/alkenoate versus growth temperature relationship in *Emiliania huxleyi* and *Gephyrocapsa* oceanica. Geochim Cosmochim Ac 62 51-68

Cortés MY, Bollmann J, Thierstein HR (2001) Coccolithophore ecology at the HOT station ALOHA, Hawaii. Deep Sea Research Part II: Topical Studies in Oceanography 48:1957-1981

Couapel MJJ, Beaufort L, Young JR (2009)) A new Helicosphaera-Syracolithus combination coccosphere (Haptophyta) dfrom the western Mediterranean sea. Journal of Phycology 45

Countway P, Gast R, Dennett M, Savai P, Rose J, Caron D (2007) Distinct protistan assemblages characterize the euphotic zone and deep sea (2500 m) of the western North Atlantic (Sargasso Sea and Gulf Stream). Environmental Microbiology 9:1219-1232

Darling KF, Kucera M, Kroon D, Wade CM (2006) A resolution for the coiling direction paradox in Neogloboquadrina pachyderma. Paleoceanography 21:PA2011

Darling KF, Kucera M, Pudsey CJ, Wade CM (2004) Molecular evidence links cryptic diversification in polar planktonic protists to Quaternary climate dynamics. PNAS 101:7657-7662

Darling KF, Kucera M, Wade CM (2007) Global molecular phylogeography reveals persistent Arctic circumpolar isolation in a marine planktonic protist. PNAS 104:5002-5007

Darling KF, Kucera M, Wade CM, von Langen P, Pak D (2003) Seasonal occurrence of genetic types of planktonic foraminiferal morphospecies in the Santa Barbara Channel Paleoceanography 18:1032

Darling KF, Wade CM, Stewart IA, Kroon D, Dingle R, Brown AJL (2000) Molecular evidence for genetic mixing of Arctic and Antarctic subpolar populations of planktonic foraminifers. Nature 405:43-47

Darling MKaKF (2002) Cryptic species of planktonic foraminifera: their effect on palaeoceanographic reconstructions. Philosophical Transactions: Mathematical, Physical and Engineering Sciences 360:695 - 718

Daugbjerg N, Andersen RA (1997) Phylogenetic analyses of the *rbcL* sequences from haptophytes and heterokont algae suggest their chloroplasts are unrelated. Mol Biol Evol 14:1242-1251

de Vargas C, Aubry MP, Probert I, Young J (2007) The origin and evolution of coccolithophores: form coastal hunters to oceanic farmers. In: Falkowski PG, Knoll AH (eds) Evolution of primary producers in the sea. Elsevier Academic Press, New York, p 251-286

de Vargas C, Norris R, Zaninetti L, Gibb SW, Pawlowski J (1999) Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces. PNAS 96:2864-2868

de Vargas C, Probert I (2004) New keys to the Past: Current and future DNA studies in Coccolithophores. Micropaleontology 50:45-54

de Vargas C, Renaud S, Hilbrecht H, Pawlowski J (2001) Pleistocene adaptive radiation in Globorotalia truncatulinoides: genetic, morphologic, and environmental evidence. Paleobiology 27:104-125

de Vargas C, Sáez A, Medlin, L., Thierstein H (2004) Super-Species in the Calcareous Nannoplankton. In: (eds) ITHaYJ (ed) Coccolithophores: From the molecular processes to global impact. Springer Verlag, New York, Berlin, Heidelberg, London, Paris, Tokyo

de Vargas C, Sáez, A.G., Medlin, L.K. & Thierstein, H.R. (2004) Super-species in the calcareous plankton. Coccolithophores - From molecular processes to global impact. Springer.

de Vargas C, Zaninetti L, Pawlowski J, Hilbrecht H (1997) Phylogeny and rates of molecular evolution of planktonic foraminifera: SSU rDNA sequences compared to the fossil record. Journal of Molecular Evolution 45:285-294

DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM (2006) Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. Science 311:496-503

Diez B, Pedros-Alio C, Massana R (2001) Study of Genetic Diversity of Eukaryotic Picoplankton in Different Oceanic Regions by Small-Subunit rRNA Gene Cloning and Sequencing. Appl. Environ. Microbiol. 67:2932-2941

Douzery EJP, Snell EA, Bapteste E, Delsuc F, Philippe H (2004) The timing of eukaryotic evolution: Does a relaxed molecular clock reconcile proteins and fossils? Proc Natl Acad Sci USA 101:15386-15391

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4:e88

Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7:214

Dugdale RC, Goering JJ (1967) Uptake of new and regenerated forms of nitrogen in primary productivity. Limnol. Oceanogr., 12:196-206

Dunn CP (2003) Keeping taxonomy based in morphology. Trends Ecol. Evol 18:270–271

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792-1797

Edvardsen B, Eikrem W, Green JC, Andersen RA, Moon-van der Staay SY, Medlin L (2000) Phylogenetic reconstructions of the Haptophyta inferred from 18S ribosomal DNA sequences and available morphological data. Phycologia 39:19-35

Edvardsen B, Paasche E (1998) Bloom dynamics and physiology of Prymnesium and Chrysochromulina. In: Anderson DM, Cembella, A.D., Hallegraeff, G.M. (ed) Physiological ecology of harmful algal blooms. NATO ASI Series G. Springer-Verlag, Heidelberg, p 193-208.

Eppley RE, Swift E, Redalje DG, Landry MR (1979) Particulate organic matter flux and planktonic new production in the deep ocean. Nature 282:677-680

Fabry VJ (2008) Ocean science: marine calcifiers in a high-CO2 ocean. Science 320:1020-1022

Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJR (2004) The Evolution of Modern Eukaryotic Phytoplankton. Science 305:354-360

Falkowski PG, Knoll AH (2007) Evolution of primary producers in the sea. Elsevier Academic Press

Farrimond P, Eglinton G, Brassell SC (1986) Alkenones in Cretaceous black shales, Blake-Bahama Basin, western North Atlantic. Org. Geochem 10:897–903

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401 - 410

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17:368-376

Felsenstein J (1985) Confidence Limits on Phylogenies: An Approach Using the Bootstrap. Evolution 39:783-791

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281:237-240

Finlay BJ (2004) Protist taxonomy: an ecological perspective. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 359:599-610

Fogg G (1995) Some comments on picoplankton and its importance in the pelagic ecosystem. Aquatic Microbial Ecology 9:33-39

Frada M, Not F, Probert I, de Vargas C (2006) CaCO3 optical detection with fluorescent in situ hybridization: A new method to identify and quantify calcifying microorganisms from the oceans. J Phycol 42:1162-1169

Frada M, Percopo I, Young J, Zingone A, de Vargas C, Probert I (2009) First observations of heterococcolithophore-holococcolithophore life cycle combinations in the family Pontosphaeraceae (Calcihaptophycideae, Haptophyta). Marine Micropaleontology 71:20-27

Fujiwara S, Sawada M, Someya J, Minaka N, Kawachi M, Inouye I (1994) Molecular phylogenetic analysis of *rbcL* in the Prymnesiophyta. J Phycol 30:863-871

Fujiwara S, Tsuzuki M, Kawachi M, Minaka N, Inouye I (2001) Molecular phylogeny of the haptophyta based on the *rbcL* gene and sequence variation in the spacer region of the RUBISCO operon. J Phycol 37:121-129

Fuller NJ, Tarran GA, Cummins DG, Woodward EMS, Orcutt KM, Yallop M, LeGall F, Scanlan DJ (2006) Molecular analysis of photosynthetic picoeukaryote community structure along an Arabian Sea transect. Limnology and Oceanography 51:2502-2514

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput. Appl. Biosci. 12:543-548

Geisen M, Young JR, Probert I, Sáez AG, Baumann KH, Bollmann J, Cros L, de Vargas C, Medlin LK, Sprengel C (2004) Species level variation in coccolithophores. In: Thierstein HR, Young JR (eds) Coccolithophores: From Molecular Processes to Global Impact. Springer, Berlin, p 327-366

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG (1990) Genetic diversity in Sargasso Sea bacterioplankton. Nature 345:60-63

Goericke R, Welschmeyer NA (1993) The marine prochlorophyte Prochlorococcus contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. Deep Sea Research Part I: Oceanographic Research Papers 40:2283-2294

Goetze E (2005) Global population genetic structure and biogeography of the oceanic copepods eucalanus hyalinus and E.spinifer. Evolution 59:2378–2398

Gooday AJ, Esteban GF, Clarke KJ (2006) Organic and siliceous protistan scales in north-east Atlantic abyssal sediments. Journal of the Marine Biological Association of the United Kingdom 86:679-688

Guillou L, Eikrem W, Chretiennot-Dinet M-J, Le Gall F, Massana R, Romari K, Pedros-Alio C, Vaulot D (2004) Diversity of Picoplanktonic Prasinophytes Assessed by Direct Nuclear SSU rDNA Sequencing of Environmental Samples and Novel Isolates Retrieved from Oceanic and Coastal Marine Ecosystems. Protist 155:193-214

Guillou L, Viprey M, Chambouvet A, Welsh RM, Kirkham AR, Massana R, Scanlan DJ, Worden AZ (2008) Widespread occurrence and genetic diversity of marine parasitoids belonging to Syndiniales (Alveolata). Environmental Microbiology 10:3349-3365

Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucl. Acids Res. 33:W557-559

Hackett JD, Yoon HS, Li S, Reyes-Prieto A, Rummele SE, Bhattacharya D (2007) Phylogenomic analysis supports the monophyly of Cryptophytes and Haptophytes and the association of Rhizaria with Chromalveolates. Mol Biol Evol 24:1702-1713

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB, Roger AJ (2009) Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". Proc Natl Acad Sci U S A 106:3859-3864

Harper JT, Waanders E, Keeling PJ (2005) On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. Int J Syst Evol Microbiol 55:487-496

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London B:313–322

Houdan A, Billard C, Marie D, Not F, Saez A, Young G, Probert I (2004) Holococcolithophores-heterococcolithophores (Haptophyta) life cycles: flow cytometry analysis of relative ploidy levels. Systematics and Biodiversity 1:453-465

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML (2007) Microbial Population Structures in the Deep Marine Biosphere. Science 318:97-100

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-5

Hugenholtz P, Huber T (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. Int J Syst Evol Microbiol 53:289-293

Iglesias-Rodriguez D, Brown CW, Doney SC, Kleypas J, Kolber D, Kolber Z, Hayes PK, Falkowski PG (2002) Representing key phytoplankton functional groups in ocean carbon cycle models: Coccolithophorids. Global Biogeochemical Cycles 16(4):47-1-47-20

Inouye I (1997) Systematics of haptophyte algae in Asia-Pacific waters. Algae (Kor. J. Phycol) 12:247-261

Inouye I, Kawachi M (1994) The haptonema. The Systematics Association Special Volume 51:77-89

Jeon S, Bunge J, Leslin C, Stoeck T, Hong S, Epstein S (2008) Environmental rRNA inventories miss over half of protistan diversity. BMC Microbiology 8:222

Jordan RW, Cros L, Young JR (2004) A revised classification scheme for living Haptophytes. Micropaleontology 50:55-79

Katz La, McManus GB, Snoeyenbos-West OLO, Griffin A, Pirog K, Costas B, Foissner W (2005) Reframing the 'Everything is everywhere' debate: evidence for high gene flow and diversity in ciliate morphospecies. Aquatic Microbial Ecology 41:55-65

Kawachi M, Inouye I, Maeda O, Chihara M (1991) The haptonema as a food-capturing device: observations on *Chrysochromulina hirta* (Prymnesiophyceae). Phycologia 30:563-573

Kennett JP (1968) Latitudinal variation in Globigerina pachyderma (Ehrenberg) in surface sediments of the southwest Pacific Ocean. Micropaleontology 14:305-318

Kennett JP, Srinivasan MS (1983) Neogene Planktonic Foraminifera: A Phylogenetic Atlas. Hutchinson Ross Publishing Company, Stroudsburg, PA

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S, Chen F, Lapidus A, Ferriera S, Johnson J, Steglich C, Church GM, Richardson P, Chisholm SW (2007) Patterns and Implications of Gene Gain and Loss in the Evolution of Prochlorococcus. PLoS Genetics 3:e231

Kleijne A (1993) Morphology, taxonomy and distribution of extant coccolithophorids (calcareous nannoplankton). Vrije Universiteit, Amsterdam, p 321

Kumar S, Tamura K, Nei M (2004) MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform %R 10.1093/bib/5.2.150 5:150-163

Lancelot C, Keller MD, Rousseau V, Smith WO, S M (1998) Autecology of the marine haptophyte Phaeocystis sp. In: Anderson DM, Cembella AD, Hallegraeff GM (eds) Physiological ecology of harmful algal blooms, NATO-ASI series 41. Springer, Berlin, p 209-224

Landry MR (2002) Integrating classical and microbial food web concepts: evolving views from the open-ocean tropical Pacific. Hydrobiologia V480:29-39

Lane CE, Archibald JM (2008) The eukaryotic tree of life: endosymbiosis takes its TOL. Trends Ecol Evol 23:268-275

Larsen N, Olsen GJ, Maidak BL, McCaughey MJ, Overbeek R, Macke TJ, Marsh TL, Woese CR (1993) The ribosomal database project. Nucl. Acids Res. 21:3021-3023

Lartillot N, Brinkmann H, Philippe H (2007) Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol Biol 7:S4

Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095 - 1109

Lartillot N, Philippe H (2006) Computing Bayes factors using thermodynamic integration. Syst Biol 55:195 - 207

Lazarus D (1983) Speciation in Pelagic Protista and Its Study in the Planktonic Microfossil Record: A Review. Paleobiology 9:327-340

Legrand C, Johansson N, Johnsen G, Børsheim KY, Granéli E (2001) Phagotrophy and toxicity variation in the mixotrophic Prymnesium patelliferum (Haptophyceae) Limnol. Oceanogr 46:1208-1214

Lemmon AR, Moriarty EC (2004) The importance of proper model assumption in bayesian phylogenetics. Syst Biol 53:265-277

Li WKW (1995) Composition of ultraphytoplankton in the central North Atlantic. Marine Ecology - Progress Series 122:1-8

Liu H, Aris-Brosou S, Probert I, de Vargas C (2009) A timeline of the environmental genetics of the haptophytes. Molecular Biology and Evolution: :msp222.

Liu H, Probert I, Uitz J, Claustre H, Aris-Brosou Sp, Frada M, Not F, de Vargas C (2009) Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. Proceedings of the National Academy of Sciences 106:12803-12808

Lopez-Garcia P, Rodriguez-Valera F, Pedros-Alio C, Moreira D (2001) Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. 409:603-607

Lovejoy C, Massana R, Pedros-Alio C (2006) Diversity and Distribution of Marine Microbial Eukaryotes in the Arctic Ocean and Adjacent Seas. Appl. Environ. Microbiol. 72:3085-3095

Lovejoy C, Vincent W, Bonilla S, Roy S, Martineau M, Terrado R, Potvin M, Massana R, Pedros-Alio C (2007) Distribution, phylogeny, and growth of cold-adapted picoprasinophytes in arctic seas. Journal of Phycology 43:78-89

Malin G, Steinke, M., (2004) Dimethyl sulfide production: what is the contribution of the coccolithphores.127-164

Massana R, Balague V, Guillou L, Pedros-Alio C (2004a) Picoeukaryotic diversity in an oligotrophic coastal site studied by molecular and culturing approaches. FEMS Microbiology Ecology 50:231-243

Massana R, Castresana J, Balague V, Guillou L, Romari K, Groisillier A, Valentin K, Pedros-Alio C (2004b) Phylogenetic and Ecological Analysis of Novel Marine Stramenopiles. Appl. Environ. Microbiol 70:3528-3534

McCaig AE, Glover LA, Prosser JI (1999) Molecular Analysis of Bacterial Community Structure and Diversity in Unimproved and Improved Upland Grass Pastures. Appl. Environ. Microbiol. 65:1721-1730

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

McIntyre A (1967) Coccoliths as paleoclimatic indicators of Pleistocene glaciation. Science 158:1314–1317

Medlin LK, Kooistra WHCF, Potter D, Saanders G, Wandersen RA (1997) Phylogenetic relationships of the 'golden algae' (haptophytes, heterokont chromophytes) and their plastids, The origin of the algae and their plastids. In: Bhattacharya D (ed) Plant Syst Evol, p 187-219

Medlin LK, Sáez AG, Young JR (2008) A molecular clock for coccolithophores and implications for selectivity of phytoplankton extinctions across the K/T boundary. Mar Micropaleontol 67:69-86

Michelle Sait PHPHJ (2002) Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. Environmental Microbiology 4:654-666

Milliman JD (1993) Production and accumulation of calcium carbonate in the ocean: budget of a nonsteady state. Global Biogeochem Cy 7:927-957

Montresor M, Lovejoy C, Orsini L, Procaccini G, Roy S (2003) Bipolar distribution of the cyst-forming dinoflagellate Polarella glacialis. Polar Biology 26:186-194

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

Morel A, Berthon JF (1989) Surface Pigments, Algal Biomass Profiles, and Potential Production of the Euphotic Layer: Relationships Reinvestigated in View of Remote-Sensing Applications. Limnol Oceanogr 34:1545-1562

Morel A, Maritorena S (2001) Bio-optical properties of oceanic waters : a reappraisal. Limnol Oceanogr 106 7163-7180

Munson MA, Banerjee A, Watson TF, Wade WG (2004) Molecular Analysis of the Microflora Associated with Dental Caries. J. Clin. Microbiol. 42:3023-3029

Norris RD, de Vargas C (2000) Evolution all at sea. 405:23-24

Not F, Latasa M, Scharek R, Viprey M, Karleskind P, Balague V, Ontaria I, Cumino A, Goetze E, Vaulot D, Massana R (2008) Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. Deep Sea Research Part I: Oceanographic Research Papers 55:1456-1473

Not F, Ramon Massana, Mikel Latasa, Dominique Marie, Céline Colson, Wenche Eikrem, Carlos Pedrós-Alió, Daniel Vaulot, and Nathalie Simon (2005) Late summer community composition and abundance of photosynthetic picoeukaryotes in Norwegian and Barents seas. Limnol. Oceanogr 50:1677–1686

Nygaard K, Tobiesen A (1993) Bacterivory in Algae: A Survival Strategy During Nutrient Limitation. Limnology and Oceanography 38:273-279

Pagel M, Meade A, Barker D (2004) Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53:673 - 684

Paradis E (2006) Analysis of phylogenetics and evolution with R. Springer, New York

Patron NJ, Inagaki Y, Keeling Patrick J (2007) Multiple Gene Phylogenies Support the Monophyly of Cryptomonad and Haptophyte Host Lineages. Curr Biol 17:887-891

Pawlowski J (2000) Introduction to the Molecular Systematics of Foraminifera Micropaleontology 46:1-12

Pawlowski J, Bolivar I, Fahrni JF, Cavalier-Smith T, Gouy M (1996) Early origin of foraminifera suggested by SSU rRNA gene sequences. Molecular Biology and Evolution 13:445-450

Pawlowski J, Fahrni J, Lecroq B, Longet D, Cornelius N, Excoffier L, Cedhagen T, Gooday AJ (2007) Bipolar gene flow in deep-sea benthic foraminifera. Molecular Ecology 16:4089-4096

Pawlowski J, Holzmann M (2002) Molecular phylogeny of Foraminifera a review. European Journal of Protistology 38:1-10

Pedrós-Alió C (2006) Marine microbial diversity: can it be determined? Trends in Microbiology 14:257-263

Perch-Nielsen K (1985) Cenozoic calcareous nannofossils. In: H.M. Bolli, Saunders. JB, Perch-Nielsen. K (eds) Plankton Stratigraphy. Cambridge University Press, Cambridge, p 427-555

Pflaumann UD, Josette; Pujol, Claude; Labeyrie, Laurent D. (1996) SIMMAX: A modern analog technique to deduce Atlantic sea surface temperatures from planktonic foraminifera in deep-sea sediments. Paleoceanography 11:15-36

Polz MF, Cavanaugh CM (1998) Bias in Template-to-Product Ratios in Multitemplate PCR. Appl. Environ. Microbiol. 64:3724-3730

Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. Bioinformatics 14:817-818

Prahl FG, Wakeham SG (1987) Calibration of unsaturation patterns in long-chain ketone compositions for palaeotemperature assessment. Nature 330: 367-369

Probert I, Houdan A (2004) The laboratory culture of coccolithophores In: Thierstein HR, Young JR (eds) Coccolithophores: from molecular processes to global impact. Springer Verlag, Berlin

Queiroz K, Donoghue MJ (1988) Phylogenetic systematics and the species problem. Cladistics 4:317-338

Rappe MS, Suzuki MT, Vergin KL, Giovannoni SJ (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. 64:294-303

Ras J, Claustre H, Uitz J (2008) Spatial variability of phytoplankton pigment distributions in the Subtropical South Pacific Ocean: comparison between in situ and predicted data. Biogeosciences 5:353-369

Rice DW, Palmer JD (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 Biol 4:31

Richardson TL, Jackson GA (2007) Small Phytoplankton and Carbon Export from the Surface Ocean. Science 315:838-840

Ridgwell A, Zeebe RE (2005) The role of the global carbonate cycle in the regulation and evolution of the Earth system. Earth Planet Sci Lett 234:299-315

Robertson JE, Robinson C, Turner DR, Holligan P, Watson AJ, Boyd P, Fernandez E, Finch M (1994) The impact of a coccolithophore bloom on oceanic carbon uptake in the northeast Atlantic during summer 1991. Deep-Sea Res Pt I 41:297-314

Robinson-Rechavi M, Huchon D (2000) RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. Bioinformatics 16:296-7

Robison-Cox J, Bateson M, Ward D (1995) Evaluation of nearest-neighbor methods for detection of chimeric small- subunit rRNA sequences. Appl. Environ. Microbiol. 61:1240-1245

Romari K, Vaulot D (2004) Composition and temporal variability of picoeukaryote communities at a coastal site of the English Channel from 18S rDNA sequences. Limnology and Oceanography 49:784-798

Ronquist F, Huelsenbeck J (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers Y-H, Falc, oacute, n LI, Souza V, Bonilla-Rosso G, aacute, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC (2007) The Sorcerer II Global Ocean Sampling Expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biology 5:e77

Ryther JH (1969) Photosynthesis and Fish Production in the Sea. Science 166:72-76

Saez AG, Lozano E (2005) Body doubles. 433:111

Saez AG, Probert I, Geisen M, Quinn P, Young JR, Medlin LK (2003) Pseudo-cryptic speciation in coccolithophores. PNAS 100:7163-7168

Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol Biol Evol 19:101-109

Sanderson MJ (2003) r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19:301-302

Schloss PD, Handelsman J (2005) Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. Appl Environ Microbiol 71:1501-1506

Seo T-K, Kishino H (2008) Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst Biol 57:367-377

Shannon CE (1948) A mathematical theory of communications. Bell Syst.Techn. J. 27:379-423 and 623-656.

Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16:1114-1116

Simon N, Brenner J, Edvardsen B, Medlin LK (1997) The identification of Chrysochromulina and Prymnesiumspecies (Haptophyta, Prymnesiophyceae) using fluorescent or chemiluminescent oligonucleotide probes: a means for improving studies on toxic algae. Eur J Phycol 32:393 - 401

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, J. HG (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". PNAS 103:12115-12120

Soltis PS, Soltis DE, Savolainen V, Crane PR, Barraclough TG (2002) Rate heterogeneity among lineages of tracheophytes: integration of molecular and fossil data and evidence for molecular living fossils. Proc Natl Acad Sci USA 99:4430-4435

Sorokin YI (1981) Microheterotrophic organisms in marine ecosystems. In: Longhurst AR (ed) Analysis of marine ecosystems. Academic Press, New York, p 293-342

Speksnijder AGCL, Kowalchuk GA, De Jong S, Kline E, Stephen JR, Laanbroek HJ (2001) Microvariation Artifacts Introduced by PCR and Cloning of Closely Related 16S rRNA Gene Sequences. Appl. Environ. Microbiol. 67:469-472

Srinivasan MS, Kennett JP (1976) Evolution and phenotypic variation in the Late Cenozoic Neogloboquadrina dutertrei plexus. progress in micropaleontology

Stoeck T, Hayward B, Taylor GT, Varela R, Epstein SS (2006) multiple PCR-primer approach to access the microeukaryotic diversity in environmental samples. . Protist 157:31-43

Suchard MA, Weiss RE, Sinsheimer JS (2001) Bayesian selection of continuous-time Markov chain evolutionary models. Mol Biol Evol 18:1001-1013

Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl. Environ. Microbiol. 62:625-630

Swofford DL (2002) PAUP*: Phylogenetic analysis using parsimony (*and other methods). Version 4.0b10. Sinauer and Associates, Sunderland, Massachusetts

Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1:269-285

Takano Y, Hagino K, Tanaka Y, Horiguchi T, Okada H (2006) Phylogenetic affinities of an enigmatic nannoplankton, Braarudosphaera bigelowii based on the *SSU* rDNA sequences. Mar Micropaleontol 60:145-156

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. Mol Biol Evol 24:1596-1599

Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003) A plea for DNA taxonomy. Trends in Ecology and Evolution 18:70–74

Thompson J, Gibson TJ, Plewniak F, Jeanmougin F, DG. H (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876-4882

Thomsen HA, Buck KR, Chavez FP (1994) Haptophytes as components of marine phytoplankton. In: Green JC, Leadbeater BSC (eds) The haptophyte algae. Clarendon Press, Oxford, p 187-208

Tillmann U (1998) Phagotrophy of a plastidic haptophyte, Prymnesium patelliferum. . Aquat Microb Ecol 14:155-160

Tyrrel T, Merico A (2004) Emiliania huxleyi: bloom observations and the conditions that induce them. In: Thierstein HR, Young JR (eds) Coccolithophores: from the molecular processes to global impact. Springer Verlag, Berlin

Uitz J, Claustre H, Morel A, Hooker SB (2006) Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. Limnol Oceanogr 111: C08005

Unrein F, Massana R, Alonso-Saez L, Gasol JM (2007) Significant year-round effect of small mixotrophic flagellates on bacterioplankton in an oligotrophic coastal system. Limnol. Oceanogr. 52:456-469

Van Lenning. K, Latasa. M, Estrada. M, Sáez. A. G, Medlin. L, Probert. I, Véron. B, Young. J.R (2003) Pigment signatures and phylogenetic relationships of the Pavlovophyceae (Haptophyta). J Phycol 39:379-389

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

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-H, Smith HO (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. Science 304:66-74

Wade C, Darling K (2002) Fossilized records of past seas. Microbiology Today 29:183-185

Wang GC, Wang Y (1997) Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. Appl. Environ. Microbiol. 63:4645-4650

Webster G, Newberry C, Fry J, Weightman A (2003) Assessment of bacterial community structure in the deep sub-seafloor biosphere by 16S rDNA-based techniques: A cautionary tale. J Microbiol Methods 55:155–164

Westbroek P, Brown CW, Van Bleijswijk J, Brownlee C, Brummer GJ, Conte M, Egge J, Fernández E, Jordan R, Knappertsbusch M, Stefels J, Veldhuis MJW, van der Wall P, Young JR (1993) A model system approach to biological climate forcing. The example of Emiliania huxleyi Global Planetary Change 8:27–46

Wheeler QD, Raven PH, Wilson EO (2004) Taxonomy: impediment or expedient. Science 303:285

Will KW, Rubinoff D (2004) Myth of the molecule: DNA barcodes for species cannot replace morphology for identification and classification. Cladistics 20:47–55

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

Worden AZ, Not F (2008) Ecology and Diversity of Picoeukaryotes In: Kirchman DL (ed) Microbial Ecology of the Oceans. Wiley, Hoboken, p 159-205

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591

Yang Z, Rannala B (2005) Branch-length prior influences Bayesian posterior probability of phylogeny. Syst Biol 54:455-470

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D (2004) A Molecular Timeline for the Origin of Photosynthetic Eukaryotes. Mol Biol Evol 21:809-818

Yoon HS, Hackett JD, Pinto G, Bhattacharya D (2002) The single, ancient origin of chromist plastids. Proc Natl Acad Sci USA 99:15507-15512

Yoshida M, Noel MH, Nakayama T, Naganuma T, Inouye I (2006) A haptophyte bearing siliceous scales: ultrastructure and phylogenetic position of Hyalolithus neolepis gen. et sp. nov. (Prymnesiophyceae, Haptophyta). Protist 157:213-34

Young JR (1998) Neogene. In: P.R. B (ed) Calcareous nannofossil biostratigraphy. British micropalaeontology society series., p 225-265

Young JR, Davis SA, Bown PR, Mann S (1999) Coccolith ultrastructure and biomineralisation. Journal of Structural Biology 126:195-215

Young JR, Geisen M, Cros L, Kleijne A, Sprengel C, I. P, Ostergaard J (2003) A guide to extant coccolithophore taxonomy. J. Nannoplankton Res Special Issue 1:125

Young JR, Geisen M, Probert I (2005) A review of selected aspects of coccolithophore biology with implications for paleobiodiversity estimation. Micropaleontology 51:267-288

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

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

CURRICULUM VITAE

Hui Liu

EDUCATION

Rutgers, the State University of New Jersey, New Brunswick, NJ
- M.S., Statistics, October 2008
- Ph.D., Oceanography, October 2009

Xiamen University, Xiamen, P. R. China
- BS, Biological Oceanography, July 2002

EMPLOYMENT

10/2008 - Present	Clinical Research Laboratories, Inc., Piscataway, NJ Biostatistician
6/2008 – 8/2008	Professional Service Solutions, Inc., Princeton, NJ Statistical Analyst
9/2003 – 8/2008	Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ Graduate Assistant
6/2002 – 5/2003	Institute of Oceanology, Chinese Academy of Sciences, Qingdao, P. R. China Research Assistant

PUBLICATIONS

Liu H., Probert I., Uitz J., Claustre H., Aris-Brosou S., Frada M., Not F., de Vargas C. 2009. Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. Proceedings of the National Academy of Sciences USA. 106:12803-12808.

Hui Liu, Stephane Aris-brosou, Ian Probert, Colomban de Vargas. A timeline of the environmental genetics of the haptophytes. 2009 Molecular Biology and Evolution:msp222.