Cryptogamie, Algologie, 2014, 35 (4): 353-377

Morphospecies *versus* phylospecies concepts for evaluating phytoplankton diversity: the case of the Coccolithophores

Jeremy R. YOUNG[a], Hui LIU[c], Ian PROBERT[b], Stéphane ARIS-BROSOU[d] & Colomban de VARGAS[e]*

[a]Earth Sciences, University College London, Gower St, London WC1E 6BT, England

[b]UPMC Université Paris 06 & CNRS, UMR 7144, FR2424 Roscoff Culture Collection, Station Biologique de Roscoff, 29682, Roscoff, France

[c]Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, NJ 08901, USA

[d]Departments of Biology and of Mathematics & Statistics, University of Ottawa, Ontario, Canada K1N 6

[e]UPMC Université Paris 06 & CNRS, UMR 7144, Equipe EPEP – Evolution du Plancton et Ecosystèmes Pélagiques, Station Biologique de Roscoff, 29682, Roscoff, France

Abstract – Genetic approaches to exploring in situ marine phytoplankton assemblages have revealed previously unsuspected diversity at different taxonomic levels. However, the phylogenetic species concept has rarely been compared to classical morphologically based taxonomy, which forms the basis of most current ecological, physiological, and paleontological knowledge of phytoplankton. Here we use the coccolithophores as a case study to test the relationship between these two taxonomic approaches. Analysis of 217 coccolithophore *LSU* rDNA sequences and 729 specimens observed by light and electron microscopy obtained from three water samples (Atlantic Ocean, Pacific Ocean and Mediterranean Sea) demonstrated that parallel analysis of morphological and genetic data highlights limitations inherent to each approach. Combined morpho-genetic analyses increase the scope of description of the composition, richness and structure of natural coccolithophore communities. Overall, genetic determined diversity exceeded morphological determined diversity, which may partly reflect methodological biases, but also probably reflects cryptic speciation and/or the presence of lightly- or non-calcifying species (or life cycle stages) within the coccolithophore clade. Focusing on six coccolithophore family or order level subgroups, we show that the genetic diversity within established morphospecies varied significantly in different environments. Critically, we find that the divergence threshold at which phylospecies corresponded to morphospecies varied between different natural communities, a factor that may have important implications with respect to evaluation of diversity by metagenomics approaches.

Protistan biodiversity / species concepts / morphospecies / phylospecies / Haptophyta / coccolithophores

* Corresponding author: jeremy.young@ucl.ac.uk

doi/10.7872/crya.v35.iss4.2014.353

INTRODUCTION

Planktonic photosynthetic protists (phytoplankton) dominate oceanic primary production and play critical roles in marine biogeochemistry (Field *et al.*, 1998; Liu *et al.*, 2009b), exporting significant amounts of organic and inorganic carbon to the deep sea through the biological pump (Dugdale & Goering, 1967; Eppley *et al.*, 1979), and structuring marine food webs (Ryther, 1969). The taxonomy of phytoplankton is based predominantly on morphological characters and so this forms the basis for much of our current knowledge of their ecology and physiology, as well as virtually all our knowledge of their palaeontology. However, perceptions of biodiversity in marine microbial communities are being radically altered by the discovery of previously unsuspected levels of genetic diversity, initially through the use of PCR-based phylogenetic approaches on prokaryotes (Chisholm *et al.*, 1988; Giovannoni *et al.*, 1990; Rappe *et al.*, 1998), and more recently oceanic protists, mainly from the picoplankton (cell size < 2-3 μm) size fraction (e.g. Diez *et al.*, 2001; Liu *et al.*, 2009b; Moon-Van Der Staay *et al.*, 2000). The recent application of whole-genome shotgun sequencing to marine samples has accelerated the discovery of novel genetic diversity (Biers *et al.*, 2009; DeLong *et al.*, 2006; Venter *et al.*, 2004). These new molecular tools have revealed thousands of previously undocumented rDNA ribotypes, often termed "phylospecies" (Bittner *et al.*, 2013; Huber *et al.*, 2007; Queiroz & Donoghue, 1988).

There are, however, many ambiguities in the new molecular view of marine protistan biodiversity, because of (i) the shortage of links to traditional morphospecies diversity analyses (Finlay, 2004) and (ii) the various biases inherent to PCR-based amplification of rDNA or other genetic markers (Acinas *et al.*, 2005; Hugenholtz & Huber, 2003). In the latter context, firstly, filter samples used for construction of picoplankton clone libraries seem to be prone to contamination by cell debris or gametes of large cells (Not *et al.*, 2009). Therefore, when universal eukaryotic primers are used, many, if not most, of these supposed picoplankton ribotypes may in fact be derived from larger cells. Currently, it is difficult to resolve this problem because of the paucity of clone library surveys of larger cell-size fractions (nano-, micro-, and macro-plankton). Secondly, within any given size-fraction, molecular diversity will be biased by three factors: (i) the nature of the genetic marker, (ii) the techniques used to extract genetic information from total environmental DNA, and (iii) the formation of chimeric sequences during the PCR amplification process. The nature of the genetic marker affects ribotype diversity because ribosomal genes are often present in multiple variants within a given species, as demonstrated in several protistan groups (Alverson & Kolnick, 2005; Darling *et al.*, 2007; Pawlowski *et al.*, 2007). Technique bias occurs since standard eukaryotic rDNA PCR amplification is biased towards short and/or GC poor genes with secondary structures particularly amenable to oligonucleotide priming and polymerase extension. This problem, initially revealed in bacteria (Polz & Cavanaugh, 1998; Suzuki & Giovannoni, 1996) may 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 those of most other eukaryotes currently represented in GenBank, and is thus inaccessible when using standard PCR protocols (Lopez-Garcia *et al.*, 2001; Pawlowski *et al.*, 1997). Despite their importance in both the planktonic and benthic marine realms, foraminifers are virtually absent from all molecular environmental surveys of these environments (Pawlowski, 2000; Stoeck *et al.*, 2006). PCR amplification biases can theoretically be reduced by increasing sequencing depth or by using multiple sets of primers with various levels of specificity (Stoeck *et al.*, 2006). However, the

extent of this type of bias has not been well quantified, and it remains unclear whether these measures would prove effective (Jeon *et al.*, 2008). Chimeric sequences occur because DNA fragments of different origins, even from distantly related organisms, can anneal (*i.e.* join together) during PCR amplification; producing molecular chimaeras which can represent a significant fraction (up to 32% in studies to date) of environmental sequences (Berney *et al.*, 2004; Hugenholtz & Huber, 2003; Robison-Cox *et al.*, 1995, Wang & Wang, 1997). It is relatively easy to detect chimeras that consist of large fragments from widely divergent species using methods such as alignment to reference sequences (Cole *et al.*, 2003) or partial tree building (Wang & Wang, 1997). However, it is much more challenging to detect micro-chimeric patterns produced by annealing of DNA fragments from related species, genera, or families. Together, these amplification biases could significantly and artificially increase the number of ribotypes amplified from natural populations (Speksnijder *et al.*, 2001).

Genetic approaches are therefore revealing novel, but potentially largely artificial chunks of biodiversity at an increasingly fast rate, whereas standard morphological analyses are probably too conservative, lumping together cryptic species and thereby potentially missing an important part of the biological diversity within groups displaying low levels of phenotypic differentiation. The increasing difficulty to attract, educate, and recruit young taxonomists is preventing transmission of expertise between generations, the so called "taxonomic impediment" (Wheeler *et al.*, 2004). This is a factor that will undoubtedly exacerbate the widening gap between the use of genetic and morphological approaches in phytoplankton taxonomy. In view of the drawbacks associated with each approach, it seems likely that combined morpho-genetic surveys will allow better interpretation of diversity patterns in their ecological context compared to the use of either method alone. Morphological analysis provides a means of evaluating the efficiency of coverage of clone libraries, especially when the extent and the potential causes of biases in PCR amplification are poorly constrained. Conversely, metagenetic data should be useful for revealing cryptic speciation and life cycle associations. This type of parallel analysis can also potentially be used to link genotypic and phenotypic data.

Here we present a case study in which we compared morphological and genetic data to assess species-level taxonomic and biogeographic differentiation in an ecologically important group of phytoplankton, the coccolithophores. This group is ideally suited for such morphogenetic inter-calibration for several reasons. First, they are abundant and ecologically significant throughout the world's oceans (Winter *et al.*, 1994). Second, the calcified platelets (coccoliths) produced by these organisms present a rich suite of morphological characters that can be observed by conventional light and electron microscopy techniques. Third, their extant diversity, as described by classical morphology-based taxonomy, is rather limited compared to other important groups of phytoplankton and they have been well-studied (Young *et al.*, 2003), making group-wide analysis feasible. Fourth, there is a reasonable coverage of cultured species (Probert & Houdan, 2004) that have been used for large-scale phylogenetic studies (Bittner *et al.*, 2013; Edvardsen *et al.*, 2011; Liu *et al.*, 2009a; Medlin *et al.*, 2008), thus facilitating the anchoring of environmental diversity to known morphospecies. Nonetheless, the morphological view of coccolithophore biodiversity is limited by several potential problems, including cryptic and pseudo-cryptic speciation (Geisen *et al.*, 2004; Saez *et al.*, 2003), dimorphic haplo-diplontic life cycles (Billard, 1994; Houdan *et al.*, 2004), and the possible existence of non-calcifying species that might be overlooked in conventional morphological studies (de Vargas & Probert, 2004; Young *et al.*, 2005).

To allow an integrated approach we collected samples for parallel morphological and genetic analyses from three geographically distinct locations in the Mediterranean Sea, the Atlantic and the Pacific Oceans. The first extensive clone library data set focusing on coccolithophores is presented, and by assessing the inherent biases in morphological and genetic approaches we demonstrate the advantages of combining these two types of analysis for an accurate assessment of protistan environmental biodiversity.

MATERIALS AND METHODS

Sample 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 (Fig. 1). Sampling took place during the Atlantic Meridional Transect cruise 16 in May 2005 (sample AMT16_4.1), the Hawaii Ocean Time-Series 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; see Fig. 1 and Table 1). At each station, water samples were collected using Niskin bottles from several depths through the photic zone. For molecular analysis, a single depth was selected and up to 100 liters of water were concentrated by filtration through a nominal 5 μm pore size nylon mesh net at the HOT169_S2 and MedEx-6 stations, whereas no such pre-filtration step was undertaken for the AMT16_4.1 sample. The water was then filtered gently using a peristaltic pump (pressure < 150 mm Hg) through poly-ether sulphone filters (0.45 μm pore size) for total DNA extraction. DNA filters were kept dry frozen at – 80°C until genomic DNA was extracted. In parallel, 0.5 to 2 liters of water (not pre-filtered) from the same samples were gently filtered onto both polycarbonate filters (0.4 μm pore size) for scanning electronic microscopy (SEM) and cellulose nitrate filters (0.45 μm pore size) for light microscopy (LM).

The pre-filtration of molecular samples from the HOT169_S2 and MedEx-6 stations was carried out in order to concentrate coccolithophores relative to non-calcifying pico-haptophytes. The nominal 5 μm 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 10 μm than 5 μm, so that there was a significant sampling bias toward larger coccosphere sizes.

Table 1. Summary of hydrographic conditions at study sites

Library	*Cruise*	*Station*	*Long.*	*Lat.*	*Depth (m)*	*Temperature (°C)*	*Salinity (psu)*	*Chlorophyll (μg/L)*
HOT169_S2	HOT169	2	– 158	22.75	79	23.8	35.1	16.31
AMT16_4.1	AMT16	4	9.33	– 30.58	2	19.6	35.7	0.04
MedEx-6	MedEx	D	7.32	43.69	20	22.1	38.1	0.12

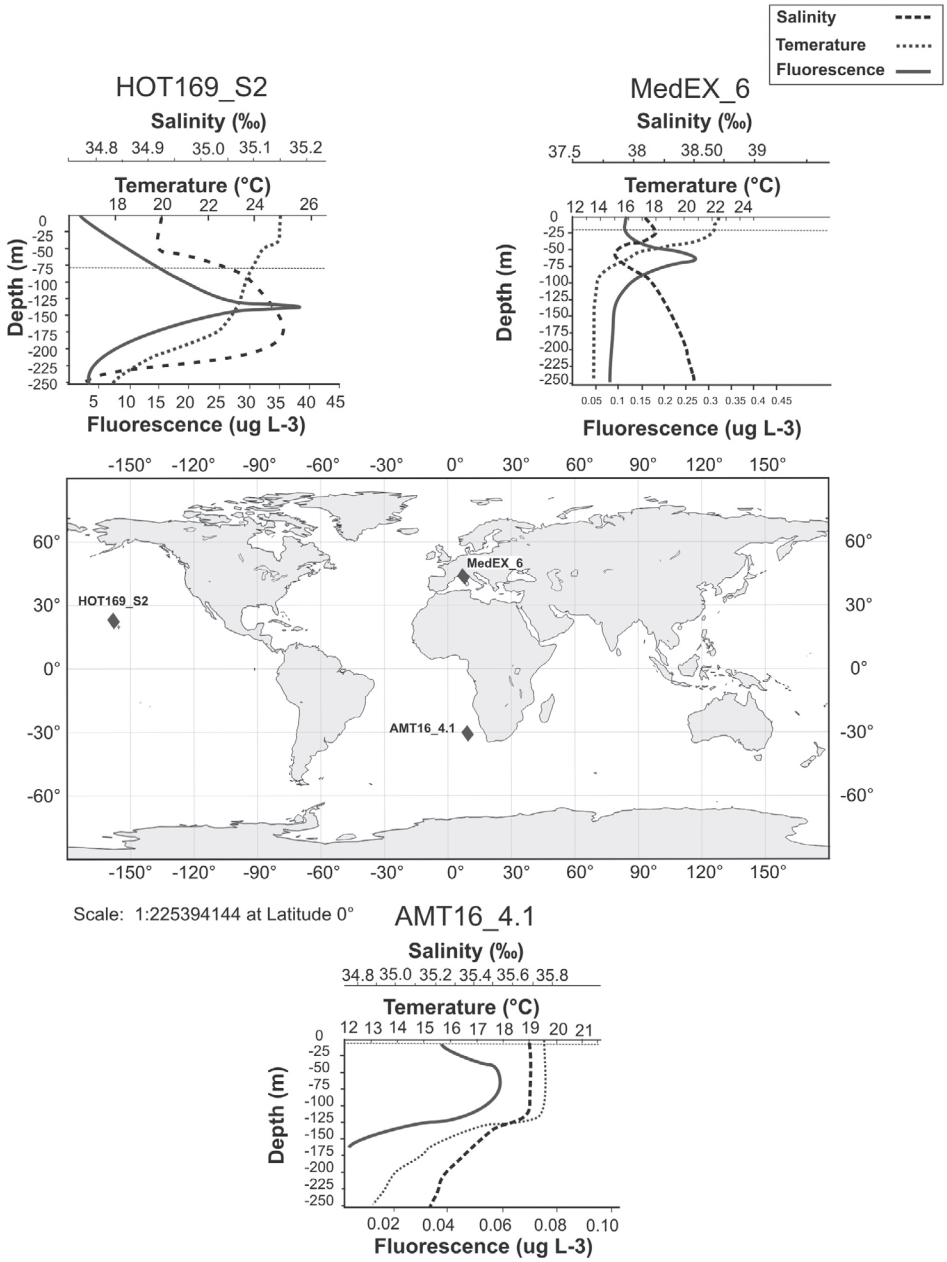


Fig. 1. Map of sampling sites (red diamonds) and plots of hydrographic conditions. Temperature, salinity, and fluorescence profiles down to 250 m depth are given for each station. Dotted lines indicate the depths at which water was sampled for this study.

Morphological and genetic data acquisition

Cross-polarized LM was used to establish the relative abundance of the major components of the assemblages. SEM was subsequently used to confirm LM observations and for identification of smaller and rarer taxa. 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, filter fragments were mounted on aluminum SEM stubs, coated with a 20 nm gold-palladium layer and examined with a Phillips XL30 field emission electron microscope. Morpho-taxonomic concepts applied followed the 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'). In order to limit formation of chimeric sequences, PCR reactions (98°C for 30s, 50°C for 30 s, and 72°C for 30 s, with initial denaturation and final extension steps) were performed over a maximum of 30 cycles using Phusion high-fidelity PCR DNA Polymerase (New England BioLabs), 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 0.2 mM 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 with 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 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 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.

DNA sequence analysis

LSU rDNA sequences were checked for potential chimeras with the Check_Chimera program (Cole *et al.,* 2003). All novel sequences passing this first screening were re-checked manually in multiple sequence alignment and in partial Neighbor-Joining (NJ) phylogenetic trees to remove putative micro-chimeras, that is, sequences containing segments from two or more closely related species. Despite the methodological precautions taken to avoid the formation of chimeric PCR products described above, the Check_Chimera approach identified ~ 15-20% putative chimeric sequences. These problematic sequences were discarded. All remaining sequences were manually aligned with 32 taxonomically defined sequences from cultures representing all major lineages of Haptophyta lineages from the Roscoff Culture Collection (http://www.roscoff-culture-

collection.org) using the *Genetic Data Environment* (GDE) version 2.2 software (Larsen *et al.,* 1993). Note that the sequence of *Gephyrocapsa oceanica* was included as the only representative of the Noelaerhabdaceae since the *LSU* rDNA sequences of *G. oceanica* and *Emiliania huxleyi* are extremely similar (0.1% genetic distance, see Liu *et al.*, 2009b). Phylogenetic analyses were performed using NJ (Saitou, 1987) with MEGA (Tamura *et al.*, 2007) and Phylo_win (Galtier *et al.*, 1996), maximum likelihood (ML) with PhyML (Guindon *et al.*, 2005) and Bayesian approaches with BEAST (Drummond & Rambaut, 2007). For ML and Bayesian analyses, the Akaike Information Criterion (AIC) implemented in ModelTest (Posada & 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% distance levels with DOTUR (Schloss & Handelsman, 2005). Pairwise distances were estimated by ML under the substitution model selected by ModelTest. This clustering resulted in two additional data sets: one that contained all 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).

Estimation of morpho- and phylospecies richness

Rarefaction curves along with diversity indices and richness estimators were calculated in order to assess the diversity found in both the morphological and genetic data sets which respectively included 729 observed individual coccolithophore specimens and all (N = 216) coccolithophore sequences retrieved from the three sample sites. For the HOT169_S2 and MedEx-6 morphological data sets, species smaller than 5 µm (e.g. *Gephyrocapsa ericsonii*, *Syracosphaera nana*, *Ophiaster*, unidentified small *Syracosphera*, *Corisphaera* cf. *gracilis*, *Anthosphaera* sp.) were excluded so that the assemblages more closely corresponded to the 5 µm pre-filtered genetic samples. Rarefaction curves are a means of predicting the total number of species in a randomly sub-sampled population. 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 ver. 9.1 (SAS Institute, Inc.; script available upon request).

For genetic data, 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 ML distance matrices. Each distance matrix was then analyzed with DOTUR assuming the furthest neighbour algorithm to cluster sequences, construct rarefaction curves and calculate the Chao1 estimators (Chao, 1984) and Shannon diversity index (Shannon, 1948). Clustering levels ranged from 0 to 5% differences.

Comparison of the proportion of morphological and genetic diversity in taxonomic subgroups

Both morphological and genetic surveys of environmental diversity of coccolithophores have inherent limitations and may not be fully quantitative. This does not, however, preclude the possibility that the proportion of DNA sequences and morphotypes observed by SEM in a particular taxonomic group are homogeneous (*i.e.* no significant differences exists in observed taxonomic frequencies between morphological and genetic analyses). To test this hypothesis, the Cochran-Mantel-Haenszel (CMH) test was performed for each sampling location using two sets of data: (1) including all six defined taxonomic subgroups (*i.e.* Noelaerhabdaceae, Rhabdosphaeraceae, Coccolithales, Zygodiscales, Syracosphaeraceae, Umbellosphaeraceae), each corresponding to family or order level clades based on the phylogeny constructed with all environmental and culture coccolithophores, and (2) a limited selection of subgroups, excluding groups for which apparent discrepancies were found between morphological and genetic analyses.

RESULTS

Species richness

Morphological and genetic diversities, assessed with the corresponding concepts of morphospecies and phylospecies, respectively, were estimated for each sample from the three sampling locations. Morphological analyses recorded 22, 28, and 35 morphospecies out of a total 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 2 lists the number of morphospecies and phylospecies obtained and their respective Shannon diversity indices. Rarefaction curves were calculated for both morphological and genetic data (Fig. 2). At the level of unique OTUs, the phylospecies rarefaction curves did not reach a plateau with our current sequencing effort, whereas all three morphospecies rarefaction curves showed a tendency towards saturation.

Table 2. Diversity and richness estimations from morphological and genetic sampling

	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

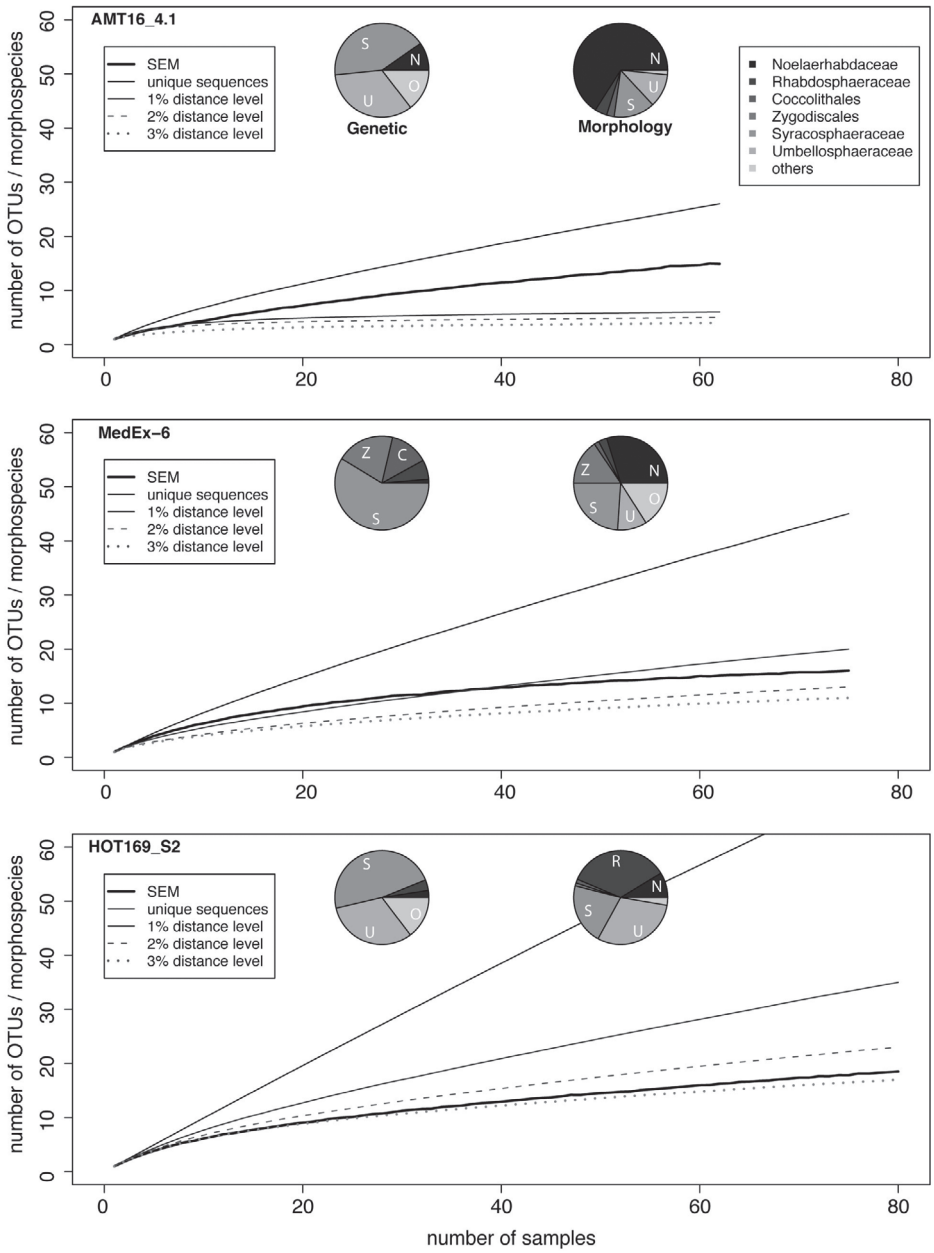


Fig. 2. Rarefaction curves for coccolithophore morpho- and phylospecies samplings at each sampling site. Three distance levels (unique, 0%, 1%, 2% and 3%) were used for phylospecies sampling. Pie charts indicate the relative frequency of sequences or individuals identified in each subgroup for both the morphological and the genetic sampling.

According to Chao1/ACE estimators, the Mediterranean, Atlantic, and Pacific samples contained an estimated 221/274, 121/112, and 399/493 coccolithophore phylospecies, respectively. The highest genetic and morphological diversities were observed in the Pacific Ocean based on both rarefaction curves and the Shannon diversity index. The Atlantic sample had the least genetic diversity of all three locations, but a higher morphological species richness was observed at this site than in the Mediterranean sample. Morphospecies rarefaction curves were also compared to phylospecies rarefaction curves based on different sequence similarity levels for each sample site (Fig. 2). The morphological rarefaction curve for HOT169_S2 (Pacific) was closest to the genetic rarefaction curve at the 3% divergence cut-off. The rarefaction curve for the MedEx-6 (Mediterranean) 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 (Atlantic) morphological analysis was found to be in between the unique and the 1% divergence level.

Global coccolithophore phylospecies diversity

Of the 366 environmental Haptophyta rDNA sequences retrieved, 266 unique OTUs were identified using DOTUR, of which 130 were coccolithophore OTUs. Thirty-two taxonomically-defined sequences from cultured strains were aligned to the environmental coccolithophore sequences, allowing assessment of the phylogenetic position of the latter (Fig. 3). None of the environmental sequences were strictly identical to any sequences from cultured coccolithophores. However, in the phylogenetic tree containing mixed environmental and culture sequences (Fig. 4), most of the major clades contained two or more culture sequences all belonging to a given family or order level taxonomic group. Hence we infer that the major genetic clades are likely to correspond to these morphology-based taxa. The five clades identified in this way and the numbers of sequences within them were: Noelaerhabdaceae (N = 9), Rhabdosphaeraceae (N = 2), Coccolithales (N = 6), Zygodiscales (N = 15), Syracosphaeraceae (N = 109). A sixth clade, containing 47 environmental sequences but no culture sequences, was inferred to represent the Umbellosphaeraceae, as explained below.

Comparative interpretation of morpho-genetic data by subgroup

The vast majority of coccolithophore OTUs recorded in this study fell into the six subgroups defined above, with a few rare exceptions that were classified as "others" and not included in comparative interpretations. The number of OTUs at the unique and 3% levels, as well as the number of morphospecies identified are listed by subgroup in Table 3. There was a significant difference in frequencies between the morphological and genetic analyses in the MedEx-6 sample unless the problematic groups (the Noelaerhabdaceae and putative Umbellosphaeraceae, for which almost no sequences were retrieved but abundant specimens observed in SEM) were excluded from construction of the contingency table (CMH test $P = 0.169$). As a result, it is difficult to correlate the frequency of retrieved DNA sequences with morphospecies abundance in a given sample in the present study. A breakdown of the morpho-genetic data by subgroup is presented below.

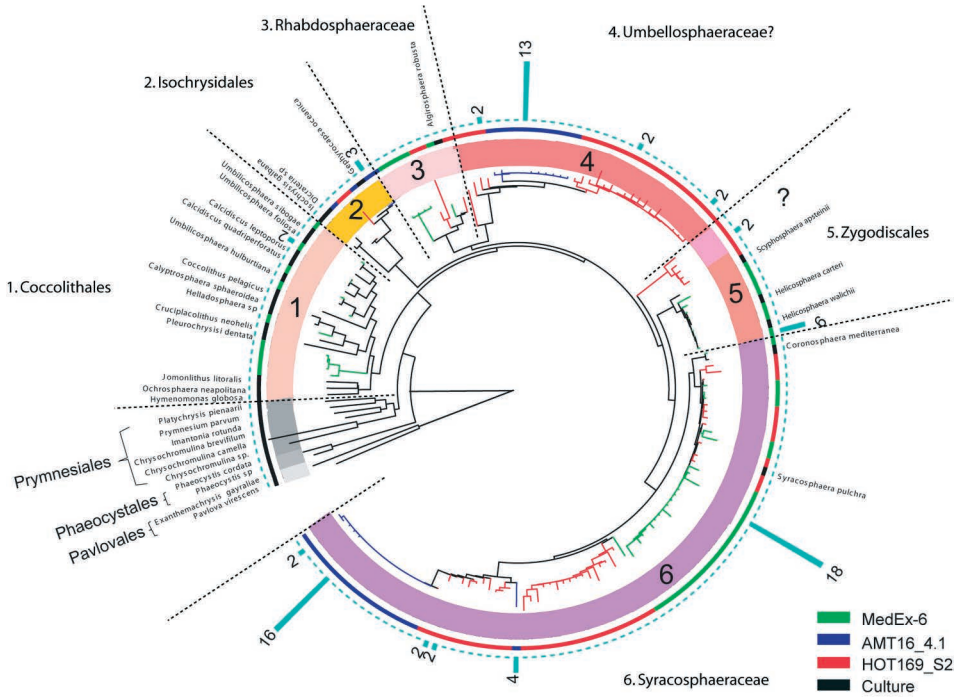


Fig. 3. *LSU* rDNA Neighbor Joining tree, including all unique coccolithophore environmental sequences and a full taxonomic cross-section of 32 known, cultured, haptophyte species. The color code used in the tree topology and outer circle highlights the origin of the phylospecies: black = culture collection, red = Pacific ocean, blue = Atlantic ocean, and green = Mediterranean sea. Names are provided for each identified cultured strains, which give a taxonomic status to the 10 clades containing environmental sequences. The number of sequences included in each unique OTU is indicated by light blue bars and the actual species count is given if > 1. The tree was formatted using the interactive tree of life (iTOL, http://itol.embl.de/itol.cgi).

Table 3. Numbers of OTUs at unique and 3% levels retrieved from genetic sampling compared to number of species identified from morphological sampling by sub-group

	HOT 169_S2			*AMT16_4.1*			*MedEx-6*		
	Genetic diversity		*Morphologic Species*	*Genetic diversity*		*Morphologic species*	*Genetic diversity*		*Morphologic species*
Taxonomic group	*OTUs (unique)*	*OTUs - 3%*		*OTUs (unique)*	*OTUs - 3%*		*OTUs (unique)*	*OTUs - 3%*	
Noelaerhabdaceae	2	1	2	3	1	2	0	0	1
Rhabdosphaeraceae	3	3	6	0	0	6	5	2	3
Coccolithales	0	0	3	0	0	2	9	5	3
Zygodiscales	0	0	3	0	0	0	6	1	2
Syracosphaeraceae	41	4	14	13	2	13	24	3	9
Umbellosphaeraceae	22	4	2	10	1	2	0	0	1
Others	6	2	5	0	0	3	1	1	3

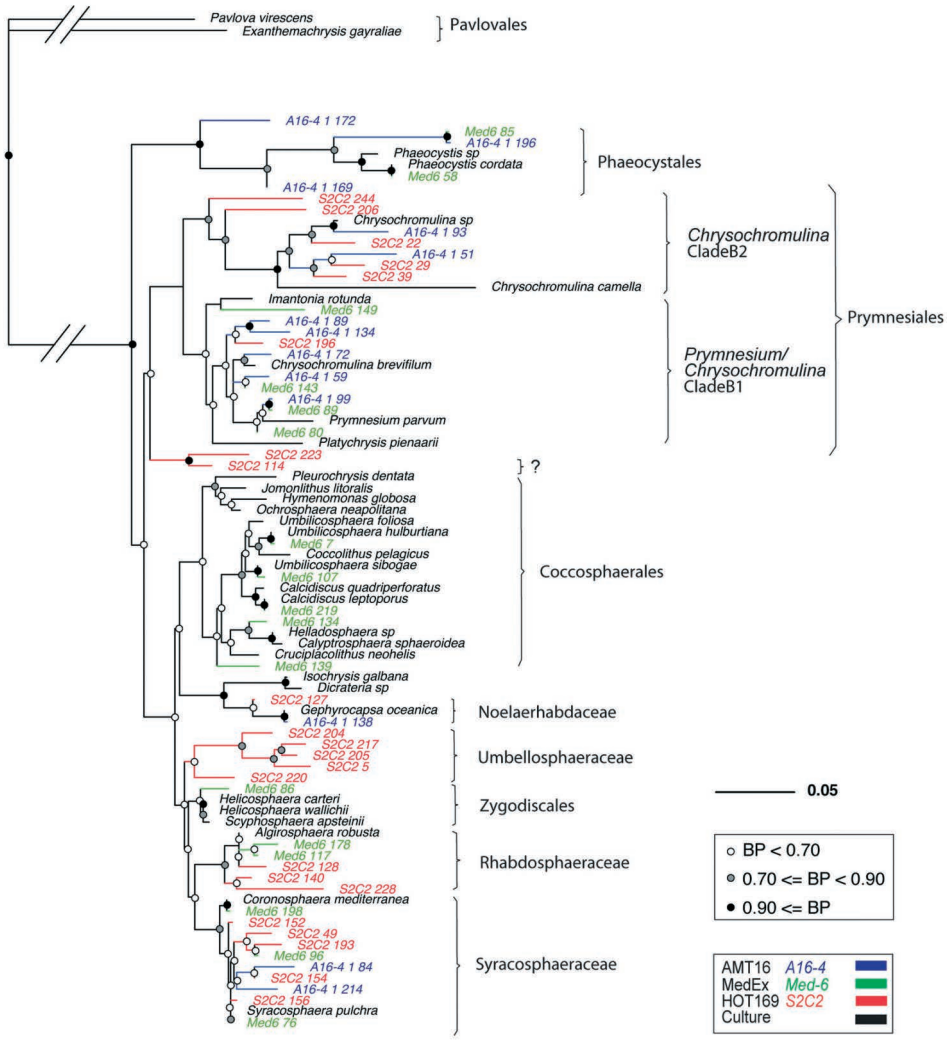


Fig. 4. *LSU* rDNA maximum likelihood tree obtained with PhyML, including only environmental haptophyte sequences at the 3% divergence cut-off level. BP: Bootstrap Probability. A16: sequences from AMT16_4.1. Med6: sequences from MedEx-6, S2C2: sequences from HOT169_S2.

Group 1. Noelaerhabdaceae

The Noelaerhabdaceae, comprising the extant genera *Emiliania, Gephyrocapsa* and *Reticulofenestra*, is the most abundant family of coccolithophores in contemporary oceans. They are distinguished from other coccolithophores by many characters, such as the production of alkenones and the presence of a motile non-calcifying haploid stage (de Vargas *et al.*, 2007). Reflecting this morphological distinction, the Noelaerhabdaceae show a basal divergence from other coccolithophores in most molecular phylogenies (Fig. 4). Morphologically, the Noelaerhab-

daceae have undergone rapid species turnover in the Quaternary (e.g. Perch-Nielsen, 1985) and exhibit very low genetic divergences in their *SSU* rDNA (Medlin *et al.*, 1996), *LSU* rDNA (de Vargas *et al.*, 2007) and *rbcL* genes (Fujiwara *et al.*, 2001).

The Noelaerhabdaceae clade can be unambiguously identified in the combined phylogeny (Figs 3-4) since two sequences from HOT169_S2 were found to be genetically close (one sequence with 1% and the other with 3% genetic divergence) to cultured *G. oceanica* or *E. huxleyi* (note that the *LSU* rDNA sequences of these two closely-related species differ by only 1 out of > 900 base pairs). 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 (genetic distance < 1%) to *G. oceanica/E. huxley*i, and 117 *E. huxleyi* cells were recorded by SEM. The anomalously low frequency of Noelaerhabdaceae sequences may in part reflect the use of the nominal > 5 µm pre -filtration on the HOT169_S2 and MedEx-6 samples which would have allowed virtually all *E. huxleyi* (~ 5 µm cell size) and all *G. ericsonii* (< 5 µm 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 (Fig. 5). However, even in this sample the group was significantly 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 results in all three samples may be due to the high GC content of the rDNA of Noelaerhabdaceae (~ 60%), which might be expected to reduce the efficiency of PCR amplification compared to other coccolithophore species with lower rDNA GC content (~ 55%-59%).

Group 2. Rhabdosphaeraceae

The Rhabdosphaeraceae are a morphologically distinct, moderately diverse family of coccolithophores (*ca* 22 extant morphospecies) 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 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 closely related 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 in the parallel morphological analysis. The clade of four very similar sequences ≤ 1% difference between them) from MedEx-6 (Fig. 5, Group 2) is likely constituted of species of the genus *Rhabdosphaera* given the 7% difference threshold from *A. robusta*. Rhabdosphaeraceae 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 is likely to be *R. clavigera* or *Discosphaera tubifera*, both of which were common in the sample. It is perhaps more likely to be *R. clavigera* because 79 m is below the typical depth range of *D. tubifera* (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.

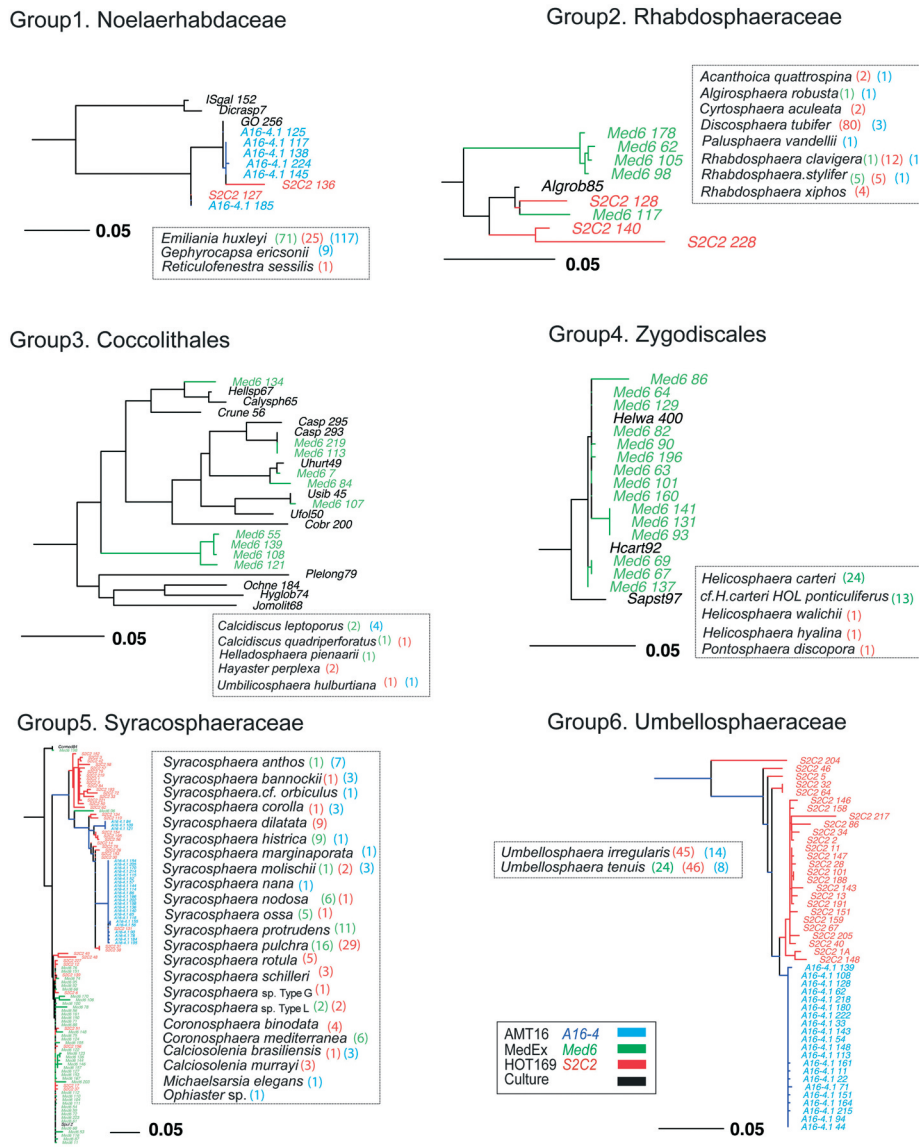


Fig. 5. Trees of subgroups identified in this study. Each tree was extracted from the tree that includes all 399 sequences (with environmental sequences). For comparison, the number of specimens of each species identified from morphology in the samples is given in the six boxes adjacent to each tree. BP: Bootstrap Probability. A16: sequences from AMT16_4.1. Med6: sequences from MedEx-6, S2C2: sequences from HOT169_S2.

Group 3. Coccolithales

The Coccolithales comprises the oceanic families Coccolithaceae and Calcidiscaceae, as well as the neritic families Pleurochrysidaceae and Hymenomonadaceae. The group is predominantly mesotrophic. As a result,

Coccolithales were relatively rare in the oligotrophic samples studied here, although 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* in the holococcolithophore (HOL) life cycle stage and *C. leptoporus* in the heterococcolithophore (HET) life cycle stage. In the phylogenetic analysis, the two *Calcidiscus* sequences from MedEx-6 were identical to *C. leptoporus*. No *Umbilicosphaera* 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% divergence from *Calyptrosphaera sphaeroidea* or *Helladosphaera* sp. culture sequences, and one *Helladosphaera pienaarii* coccosphere was observed in the morphological sample, which suggests that this unknown sequence might be correspond to this morphospecies. The clade of four MedEx-6 sequences nesting within the Coccolithales but outside of the Coccolithaceae and Calcidiscaceae is intriguing (bootstrap < 0.70; Fig. 5), since no obvious candidate species conventionally assigned to the Coccolithales were observed in morphological analyses. One possibility is that these are *Ceratolithus* since this enigmatic genus was common in the MedEx-6 sample and could conceivably be placed almost anywhere in the coccolithophore tree. Almost no Coccolithales were found in the HOT169_S2 sample, and very few were found in other upper water column samples from Hawaii. Therefore, the absence of any Coccolithales sequences in the corresponding clone library is to be expected. Nonetheless, parallel culture isolation from this sample resulted in the initiation of several cultures of *Calcidiscus* spp. and *Umbilicosphaera* spp. This mirrors a study carried out on cyanobacterial mat communities (Jungblut *et al.*, 2005), where phylogenetic diversities retrieved by clone libraries from three ponds were not similar, yet known culture strain sequences clustered together with clones obtained from all three ponds. In the AMT16_4.1 sample, *C. leptoporus* was quite common and a few *Umbilicosphaera* specimens occurred, but no Coccolithales sequences were found in the corresponding clone library.

Group 4. Zygodiscales

The Zygodiscales is a rather low diversity group that is well supported both morphologically and paleontologically (Aubry, 1989; Frada *et al.*, 2009; Young *et al.*, 2003). It includes two extant families, the Pontosphaeraceae and Helicosphaeraceae, certain members of which have been cultured and sequenced. Recent molecular phylogenetic studies confirm the monophyly of the group (Liu *et al.*, 2009a; Medlin *et al.*, 2008; de Vargas *et al.*, 2007).

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 (n.b. < 0.5% was rounded down to 0%), 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 holococcolithophore *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 with coccoliths of both *H. wallichii* and *S. ponticuliferus* was observed, suggesting that these two morphospecies are alternate life cycle stages of a single taxon (Couapel *et al.*, 2009). However, in LM we also observed *S. ponticuliferus* holococcoliths co-occurring with typical *H. carteri* type holococcoliths on single combination coccospheres so the life-cycle pairings may not be straightforward.

Group 5. Syracosphaeraceae

The Syracosphaeraceae is the most morphologically complex and morphospecies-rich group of coccolithophores, including *ca* 50 described species, many of which include possible pseudo-cryptic species (Cros & Fortuño, 2002; Young *et al.*, 2003). However, only two species, *Syracosphaera pulchra* and *Coronosphaera mediterranea*, have been isolated into laboratory culture. 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 *C. mediterranea* in a basal position and of *S. pulchra* deep within the clade. The clade can itself be subdivided into three 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 they are most likely 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 HOT169 library, and they clustered close to the known *S. pulchra* sequence. *S. 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 closely related to *S. pulchra* probably correspond to *S. histrica*. Beyond this, it is difficult even to speculate, as 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 hypothesis, since the morphological groupings are tentative and the clades were not well supported (bootstrap < 0.70; Fig. 5).

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. This group contains two well-established species, *U. tenuis* and *U. irregularis*, but it has been suggested that *U. tenuis* is a cluster of at least six pseudo-cryptic species, informally termed *U. tenuis* types O, I, II, IIIa, IIIb and IV (Boeckel & Baumann, 2008; Kleijne, 1993; Young *et al.*, 2003) – see below.

Because the genus shows no obvious morphological affinities to other coccolithophores, a new family *incertae sedis* was established for it by Young *et al.* (2003).

Umbellosphaera was abundant in all morphological samples. The 16_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 and well-supported (BP ≥ 0.90; Fig. 5) 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 represented ~ 70% of the observed assemblage in the morphological sample at this site, and (ii) the clade contains 26 out of ~ 76 coccolithophore sequences in the library.

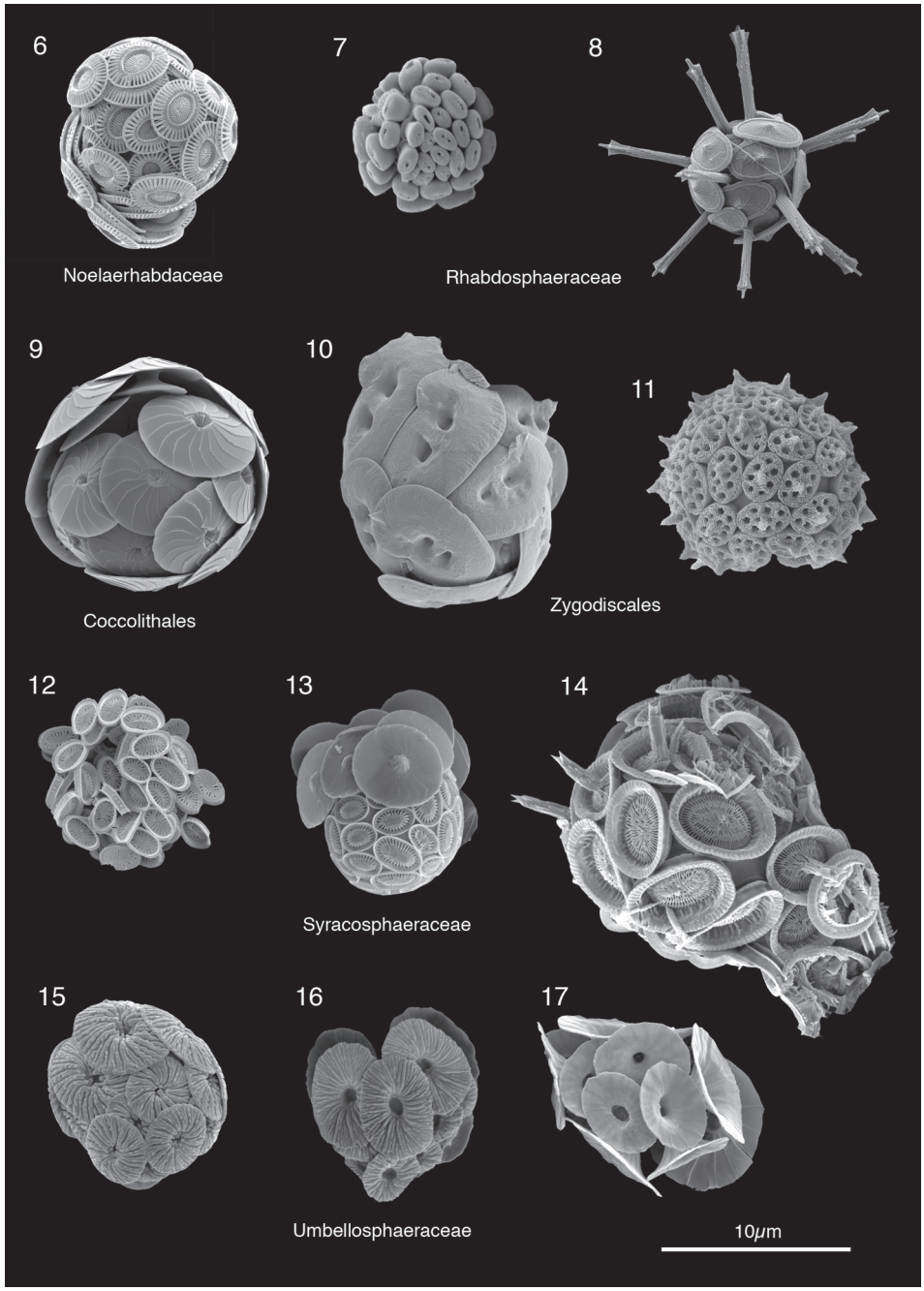
The AMT16_4.1 and HOT169_S2 morphological samples each contained both *U. tenuis* and *U. irregularis*, but sequences from the two sites form discrete sub-clades within the overall *Umbellosphaera* clade (Fig. 5). For *U. tenuis*, this arguably supports the previous morphological work suggesting that *U. tenuis* is a complex of several cryptic species (Boeckel & Baumann, 2008; Kleijne, 1993; Young *et al.*, 2003). The HOT169_S2 sample contained *U. tenuis* type IV (which is a large form and hence would probably be retained on the > 5 μm filter), whilst 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 similar sequences is surprising. Cortes *et al.* (2001) in a detailed study of coccolithophores from the HOT station showed that *U. tenuis* occurred deeper in the water column than *U. irregularis* so one possibility is that the observed coccospheres of *U. irregularis* in this relatively deep sample (79 m) were settling dead cells and so all sequences came from *U. tenuis*. The most basal sequences within the putative *Umbellosphaera* clade may represent additional *U. tenuis* types and/or *U. irregularis*.

DISCUSSION

This was an exploratory study designed to investigate how readily the results from environmental DNA analysis correlate with those from morphological analysis and to explore the potential of combined morphological and environmental DNA analyses as a tool for exploring the biodiversity of plankton. Several useful points emerge from this study.

Broad consistency between results from clone library and morphological samples

Despite many inconsistencies there is, as outlined in the results section, a broadly logical correlation between the clone library phylogeny, with clade identification using culture sequences and the assemblages observed in the morphological samples. So, since the clone library also includes numerous sequences and sub-clades not previously known from cultures it does appear to provide an alternative way to assemble a phylogeny of haptophytes, avoiding the restrictions posed by needing to isolate cultures or undertake single cell DNA


6
7
8
Noelaerhabdaceae
Rhabdosphaeraceae
9
10
11
Coccolithales
Zygodiscales
12
13
14
Syracosphaeraceae
15
16
17
Umbellosphaeraceae
10μm

extractions. Indeed the new tree (Figs 3-5) provides many new insights into relationships of taxa, and allows the development of testable hypotheses. In particular the following are noteworthy results.

i. Identification of a putative Umbellosphaeraceae clade (Fig. 3). Although the Umbellosphaeraceae are the dominant coccolithophores in oligotrophic waters their affinities were previously unknown. This result suggests that they are a distinct group from any of the better-known families.

ii. The phylogeny further suggests that the Umbellosphaeraceae is a sister taxon of the Rhabdosphaeraceae. This is a surprising result since they do not show obvious morphological or structural similarities to the Rhabdosphaeraceae. Also, since the Rhabdosphaeraceae have an unambiguous fossil record back into the Early Palaeocene (Perch-Nielsen, 1985) this suggests that the Umbellosphaeraceae should have a similar geological record, although their known fossil record is much more recent (Young, 1998).

iii. Identification of novel sub-clades inside the Coccolithales and sister to the Zygodiscales. It is not possible to identify them yet but there is a range of possible groups that these might correspond to, such as the Ceratolithaceae or Alisphaeraceae.

iv. Confirmation that the Syracosphaeraceae dominate modern coccolithophore biodiversity and suggestion that there are three major sub-clades within the group as predicted from knowledge of their morphological diversity. To date there has been little success in culturing the numerous known species of *Syracosphaera* so this type of approach seems much more promising for resolving their phylogenetic relationships and so constraining their evolutionary history.

So with additional environmental sequencing work we can envisage a comprehensive molecular phylogeny of the coccolithophores being assembled in a relatively short time. Development and testing of hypotheses on the identity of clades or groups of OTUs can be undertaken through continued comparison with conventional morphospecies counts. In addition, the genetic data can be used to design molecular probes for particular clades to allow the use of in-situ hybridisation to reveal the morphological identity of the clades. In the particular case of coccolithophores, the COD-FISH protocol (Frada *et al.*, 2006) can be used to allow combined morphological identification using cross-polarised light microscopy and fluorescence from the probe.

◀ Figs 6-17. Scanning electron micrographs of representatives of the different coccolithophore subgroups identified in this study. All images are at the same scale. **6.** Noelaerhabdaceae (*Emiliania huxleyi,* from AMT16_4.1); **7.** Rhabdosphaeraceae (*Algirosphaera robusta,* from HOT169_S2); **8.** Rhabdosphaeraceae (*Rhabdosphaera stylifera*, from MedEx-6; **9.** Coccolithales (*Calcidiscus leptoporus*, from HOT169_S2); **10.** Zygodicales (*Helicosphaera carteri*, from MedEx-6, this is the typical heterococcolith-bearing diploid life-cycle stage); **11.** Zygodicales (*Helicosphaera carteri* HOL confusus type, from MedEx-6, this is the alternate holococcolith-bearing haploid life-cycle stage); **12.** Syracosphaeraceae (*Syracosphaera dilatata*, from HOT169_S2); **13.** Syracosphaeraceae (*Syracosphaera anthos*, from AMT16_4.1); **14.** Syracosphaeraceae (*Syracosphaera pulchra*, from MedEx-6); **15.** Umbellosphaeraceae (*Umbellosphaera tenuis* type IIIa from AMT16_4.1); **16.** Umbellosphaeraceae (*Umbellosphaera tenuis* type IV from HOT169_S2); **17.** Umbellosphaeraceae (*Umbellosphaera irregularis* type IIIa from HOT169_S2).

There is only a weak quantitative correlation between the frequency of clades in morphological samples and in the clone libraries

As summarised in Figure 2, the molecular and morphological analyses give very different impressions of the relative abundance of different taxonomic groups, and the picture is no better at other taxonomic levels. For example, the Noelaerhabdaceae were very abundant in morphological samples from all sampling locations, but only a very limited number of sequences of this group were obtained. Similarly, in the HOT169_S2 sample, many Rhabdosphaeraceae coccospheres were observed by SEM analysis whilst only three sequences were retrieved, and in the MedEx-6 sample 24 *Umbellosphaera* coccospheres were observed by SEM but no sequences were retrieved.

There may be some special factors which caused abnormally poor correlation of molecular and morphological results in this particular study, and which hopefully can be avoided or compensated for in future studies. Specifically:

i. The Noelaerhabdaceae are known to have anomalously high GC frequencies in their DNA sequences and this is likely to have resulted in amplification bias during PCR (Suzuki & Giovannoni, 1996; Polz & Cavanaugh, 1998).

ii. As explained in the methods section, there was a problem with the pre-filtration step we employed in the Hawaii and MEDEX samples, and as a result the clone library assemblages, but not the morphological samples, were biased toward larger coccolithophores.

iii. Coccospheres may remain intact long after the cell has died. During morphological counts such empty dead coccospheres are not usually distinguished from live coccospheres and we did not attempt to do so in this study.

iv. In some coccolithophore species there may be naked life-cycle stages that are not recognised in morphological counts, for example the haploid stage of *Emiliania huxleyi* is non-calcifying and is unidentifiable in routine morphological counts.

All of these factors doubtless contributed to the low correlation of the molecular and morphological counts, nonetheless it is hard to avoid the conclusion that frequency in clone library is a rather poor proxy for frequency in environmental samples and that biasing factors are indeed important. These may include: (i) variation in the amount of ribosomal DNA present in different species. One of the advantages of ribosomal DNA for environmental DNA work is the fact that it is present in numerous copies per cell but the degree of variability between species or physiological states is poorly known; (ii) variable ease of extraction and amplification of rDNA between species. Again little is known about this but it has the potential to be a major biasing factor; (iii) free DNA occurring in environmental samples. There have been many instances of sequences of relatively large eukaryotes such as dinoflagellates and copepods occurring in environmental DNA samples which had been filtered to remove such organisms. One interpretation of these anomalies is that DNA from dead cells can be preserved in the marine environment. If this is so for larger organisms then clearly it might also be the case for coccolithophores.

More work is therefore needed on inter-calibration of morphological and environmental DNA results, and identification of which biasing factors are most important. This is becoming especially relevant as environmental DNA count frequency is starting to be used as a measure of taxon abundance in ecological surveys, at least for organisms that are traditionally hard to count. In this respect, study of groups such as coccolithophores which are accessible both for traditional counts and environmental DNA study may play a key role in developing the science.

There appears to be implausibly high variability between clones in the clone libraries

An especially striking statistic is that of the 366 environmental Haptophyta rDNA sequences retrieved, there were 266 unique OTUs, *i.e.* most OTUs were only recovered once or twice, and differences between sister OTUs is frequently > 1%, even > 3%.

This does not correlate well with our experience of rDNA variability from culture based studies or with the morphological observations. In culture based work we do not observe this level of variability in genetic sequences from repeated sampling of single morphospecies. Within individual morphospecies of coccolithophores only minor differences (< 1%) in 18S rRNa sequences have been observed from sequences of different strains and these differences have typically corresponded to distinct morphotypes (Saez *et al.*, 2004). Indeed it is often the case that clear morphological differences are not reflected in differences in ribosomal DNA. For instance, the diverse *Emiliania huxleyi* strains and those of its morphologically distinct sister taxon *Gephyrocapsa oceanica* have identical 18S rDNA sequences (Sáez *et al.*, 2004). Likewise, *Coccolithus pelagicus* and *C. braarudii* are well differentiated in terms of morphology, ecology and tufa sequences, but have identical 28S sequences (Saez *et al.*, 2004). We would not therefore predict that major differences would occur within single morphospecies. Further, even when the total diversity is high in morphological counts, single morphospecies occur repeatedly with individual assemblages being dominated by a few species. For example, within the AMT16 sample *Umbellosphaera tenuis* formed *ca* 5% of the assemblage and all of the observed specimens were similar (morphotype IIIa of Young *et al.*, 2003). So, it is highly anomalous that there were 9 putative *Umbellosphaera* OTUs differing by 1-3% and one differing by > 3%.

Given this, our apparent evidence of unsuspected molecular diversity within the coccolithophores needs to be questioned. A more likely explanation of this data may be that problems such as micro-chimeric sequences and sequencing errors have introduced artificial diversity into our sequence libraries. It is important for future research to resolve these uncertainties, not least since the problems are liable to be exacerbated when using new generation sequencing since the fragments sequenced are shorter than in traditional clone libraries.

Possible evidence of geographical divergence

Generally speaking, morphological data on coccolithophore biogeography has suggested that they show very low levels of endemicity, with individual species having very broad biogeographic ranges (Winter *et al.*, 1994). Molecular evidence has suggested that species concepts need to be refined (Saez *et al.*, 2003) and this has reduced some biogeographic ranges (Ziveri *et al.*, 2004). Notwithstanding all of the provisos discussed above, there are some tantalising suggestions in our data of unexpected geographical differentiation between sites. The most striking example is provided by *Umbellosphaera*. Morphologically the specimens from the Pacific (Hawaii HOTS169-S2) and Atlantic (AMT16-4.1) were indistinguishable, but the DNA sequences in the putative *Umbellosphaera* clade are well separated.

CONCLUSIONS

This study, whilst in many ways preliminary, has demonstrated that environmental DNA analysis, when combined with clade identification using known sequences and morphological data, has the potential to rapidly improve

our knowledge of the diversity of planktonic species and of their phylogenetic relationships. Conversely, the study suggests that there is very limited correlation between the relative abundance of a sequence in an environmental DNA analysis and the observed abundance of the corresponding taxon in the assemblage count data from the same sample.

Next generation techniques such 454 and Illumina sequencing are rapidly producing immense quantities of sequence data and it is tempting to interpret this data in terms of population abundances. The results of this study highlight the multiple factors that may limit the validity of such interpretations. We suggest that integrated studies, of the type described here, are pursued as a matter of urgency, in order to determine the limitations of quantitative environmental DNA data and to develop approaches for compensating for particular biases in this data.

Acknowledgments. We are grateful to numerous colleagues for assistance with this research. The AMT16_4.1 samples were collected by Miguel Frada with support of the Captain and crew of the *R.S. James Clark Ross* and its science team. The HOT169_S2 sampling was facilitated by Dave Karl and assisted by Sebastian Meier. The MedEx-6 sampling was facilitated by J Dolan and Jean-Pierre Gattuso and supported by many colleagues. Financial support for the research program was provided by the NSF (US), NSERC (Canada), IRND (U. Ottawa, Canada), the French ANR BOOM project, the EU Biodiversa project *BioMarks*, the EU infrastructure project ASSEMBLE (grant number 227799) and the French infrastructure project EMBRC-France.

REFERENCES

ACINAS S.G., SARMA-RUPAVTARM R., KLEPAC-CERAJ V. & POLZ M.F., 2005 — PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample. *Applied and environmental microbiology* 71(12): 8966-8969.

ALVERSON A.J. & KOLNICK L., 2005 — Intragenomic nucleotide polymorphism among small subunit (18S) rdna paralogs in the diatom genus *Skeletonema* (Bacillariophyta). *Journal of phycology* 41(6): 1248-1257.

AUBRY M.-P., 1989. — Phylogenetically based calcareous nannofossil taxonomy: Implications for the interpretation of geological events. *In*: Van Heck S.E. & Crux J.A. (eds), *Nannofossils and their Applications – Proceedings of the INA Conference*. London, Chichester, Ellis Horwood, pp. 21-40.

BERNEY C., FAHRNI J. & PAWLOWSKI J., 2004 — How many novel eukaryotic "kingdoms"? Pitfalls and limitations of environmental DNA surveys. *BMC Biology* 2(1): 13.

BIERS E.J., SUN S. & HOWARD, E.C. 2009 — Prokaryotic Genomes and Diversity in Surface Ocean Waters: Interrogating the Global Ocean Sampling Metagenome. *Applied and environmental microbiology* 75(7): 2221-2229.

BILLARD C., 1994. — Life cycles. *In*: Green J.C. and Leadbeater B.S.C. (eds), *The Haptophyte Alage*. Oxford, Clarendon Press, pp. 167-186.

BITTNER L., GOBET A., AUDIC S., ROMAC S., EGGE E.S., SANTINI S., OGATA H., PROBERT I., EDVARDSEN B. & DE VARGAS C., 2013 — Diversity patterns of uncultured Haptophytes unravelled by pyrosequencing in Naples Bay. *Molecular ecology* 22: 87-101.

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(3-4): 255-273.

CHAO A., 1984 — Non-parametric estimation of the number of classes in a population. *Scandinavian journal of statistics* 11: 265-270.

CHISHOLM S.W., OLSON R.J., ZETTLER E.R., GOERICKE R., WATERBURY J.B. & WELSCHMEYER N.A., 1988 — A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* 334(6180): 340-343.

COLE J.R., CHAI B., MARSH T.L., FARRIS R.J., WANG Q., KULAM S.A., CHANDRA S., MCGARRELL D.M., SCHMIDT T.M., GARRITY G.M. & TIEDJE J.M., 2003 — The

Ribosomal Database Project (RDP-II): previewing a new autoaligner that allows regular updates and the new prokaryotic taxonomy. *Nucleic acids research* 31(1): 442-443.

CORTÉS M.Y., BOLLMANN J. & THIERSTEIN H.R., 2001 — Coccolithophore ecology at the HOT station ALOHA, Hawaii. *Deep sea research Part II: Topical studies in oceanography* 48(8-9): 1957-1981.

COUAPEL M.J.J., BEAUFORT L. & YOUNG J.R., 2009 — A new Helicosphaera-Syracolithus combination coccosphere (Haptophyta) dfrom the western Mediterranean sea. *Journal of phycology* 45(4).

DARLING K.F., KUCERA M. & WADE C.M., 2007 — Global molecular phylogeography reveals persistent Arctic circumpolar isolation in a marine planktonic protist. *Proceedings of the national academy of science* 104(12): 5002-5007.

DELONG E.F., PRESTON C.M., MINCER T., RICH V., HALLAM S.J., FRIGAARD N.-U., MARTINEZ A., SULLIVAN M.B., EDWARDS R., BRITO B.R., CHISHOLM S.W. & KARL D.M., 2006 — Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science* 311(5760): 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. *Applied and environmental microbiology* 67(7): 2932-41.

DRUMMOND A. & RAMBAUT A., 2007 — BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary biology* 7(1): 214.

DUGDALE R.C. & GOERING J.J., 1967 — Uptake of new and regenerated forms of nitrogen in primary productivity. *Limnology and oceanography* 12: 196-206.

EDVARDSEN B., EIKREM W., THRONDSEN J., SÁEZ A.G., PROBERT I. & MEDLIN L.K., 2011 — Ribosomal DNA phylogenies and a morphological revision provide the basis for a revised taxonomy of the Prymnesiales (Haptophyta). *European journal of phycology* 46(3): 202-228.

EPPLEY R.E., SWIFT E., REDALJE D.G. & LANDRY M.R., 1979 — Particulate organic matter flux and planktonic new production in the deep ocean. *Nature* 282: 677-680.

FIELD C.B., BEHRENFELD M.J., RANDERSON J.T. & FALKOWSKI P., 1998 — Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* 281(5374): 237-240.

FINLAY B.J., 2004 — Protist taxonomy: an ecological perspective. *Philosophical Transactions of the royal society of London. Series B: Biological sciences* 359(1444): 599-610.

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. *Journal of phycology* 42(1): 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(1-2): 20-27.

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. *Journal of phycology* 37: 121-129.

GALTIER N., GOUY M. & GAUTIER C., 1996 — SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Computer applications in the biosciences* 12(6): 543-548.

GEISEN M., YOUNG J.R., PROBERT I., SÁEZ A.G., BAUMANN K.H., BOLLMANN J., CROS L., DE VARGAS C., MEDLIN L.K. & SPRENGEL C., 2004. — Species level variation in coccolithophores. *In*: Thierstein H.R. & Young J.R. (eds), *Coccolithophores: From Molecular Processes to Global Impact*. Berlin, Springer, pp. 327-366.

GIOVANNONI S.J., BRITSCHGI T.B., MOYER C.L. & FIELD K.G., 1990 — Genetic diversity in Sargasso Sea bacterioplankton. *Nature* 345(6270): 60-63.

GUINDON S., LETHIEC F., DUROUX P. & GASCUEL O., 2005 — PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic acids research* 33(suppl_2): W557-559.

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(4): 453-465.

HUBER J.A., MARK WELCH D.B., MORRISON H.G., HUSE S.M., NEAL P.R., BUTTERFIELD D.A. & SOGIN M.L., 2007 — Microbial Population Structures in the Deep Marine Biosphere. *Science* 318(5847): 97-100.

HUGENHOLTZ P. & HUBER T., 2003 — Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *International journal of systematic and evolutionary microbiology* 53(1): 289-293.

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(1): 222.

JORDAN R.W., CROS L. & YOUNG J.R., 2004 — A revised classification scheme for living Haptophytes. *Micropaleontology* 50(supplement 1): 55-79.

JUNGBLUT A.D., HAWES I., MOUNTFORT D., HITZFELD B., DIETRICH D.R., BURNS B.P. & NEILAN B.A., 2005 — Diversity within cyanobacterial mat communities in variable salinity meltwater ponds of McMurdo Ice Shelf, Antarctica. . *Environmental microbiology* 7(4): 519-529.

KLEIJNE A., 1993. *Morphology, taxonomy and distribution of extant coccolithophorids (calcareous nannoplankton)*. Ph.D. Thesis, Vrije Universiteit, Amsterdam, 321 p.

LARSEN N., OLSEN G.J., MAIDAK B.L., MCCAUGHEY M.J., OVERBEEK R., MACKE T.J., MARSH T.L. & WOESE C.R., 1993 — The ribosomal database project. *Nucleic acids research* 21(13): 3021-3023.

LIU H., ARIS-BROSOU S., PROBERT I. & DE VARGAS C., 2009a — A timeline of the environmental genetics of the haptophytes. *Molecular biology and evolution*: doi:10.1093/molbev/msp222

LIU H., PROBERT I., UITZ J., CLAUSTRE H., ARIS-BROSOU S.P., FRADA M., NOT F. & DE VARGAS C., 2009b — Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. *Proceedings of the national academy of sciences* 106(31): 12803-12808.

LOPEZ-GARCIA P., RODRIGUEZ-VALERA F., PEDROS-ALIO C. & MOREIRA D., 2001 — Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. *Nature* 409(6820): 603-607.

MEDLIN L.K., BARKER G.L.A., CAMPBELL L., GREEN J.C., HAYES P.K., MARIE D., WRIEDEN S. & VAULOT D., 1996 — Genetic characterization of Emiliania huxleyi (Haptophyta). *Journal of marine systematics* 9: 13–31.

MEDLIN L.K., SÁEZ A.G. & YOUNG J.R., 2008 — A molecular clock for coccolithophores and implications for selectivity of phytoplankton extinctions across the K/T boundary. *Marine micropaleontology* 67(1-2): 69-86.

MOON-VAN DER STAAY S.Y., VAN DER STAAY G.W.M., GUILLOU L., VAULOT D., CLAUSTRE H. & MEDLIN L.K., 2000 — Abundance and Diversity of Prymnesiophytes in the Picoplankton Community from the Equatorial Pacific Ocean Inferred from 18S rDNA Sequences. *Limnology and oceanography* 45(1): 98-109.

NOT F., DEL CAMPO J., BALAGUÃE V., DE VARGAS C. & MASSANA R., 2009 — New Insights into the Diversity of Marine Picoeukaryotes. *PLoS ONE* 4(9): e7143.

PAWLOWSKI J., BOLIVAR I., FAHRNI J.F., DE VARGAS C., GOUY M. & ZANINETTI L., 1997 — Extreme differences in rates of molecular evolution of foraminifera revealed by comparison of ribosomal DNA sequences and the fossil record. *Molecular biology and evolution* 14(5): 498-505.

PAWLOWSKI J., 2000 — Introduction to the Molecular Systematics of Foraminifera. *Micropaleontology* 46: 1-12.

PAWLOWSKI J., FAHRNI J., LECROQ B., LONGET D., CORNELIUS N., EXCOFFIER L., CEDHAGEN T. & GOODAY A.J., 2007 — Bipolar gene flow in deep-sea benthic foraminifera. *Molecular ecology* 16(19): 4089-4096.

PERCH-NIELSEN K., 1985. — Cenozoic calcareous nannofossils. *In*: Bolli, H.M., Saunders J.B. & Perch-Nielsen K. (eds), *Plankton Stratigraphy*. Cambridge, Cambridge University Press, pp. 427–554.

POLZ M.F. & CAVANAUGH C.M., 1998 — Bias in Template-to-Product Ratios in Multitemplate PCR. *Applied and environmental microbiology* 64(10): 3724-3730.

POSADA D. & CRANDALL K.A., 1998 — Modeltest: testing the model of DNA substitution. *Bioinformatics* 14(9): 817-818.

PROBERT I. & HOUDAN A., 2004. — The laboratory culture of coccolithophores. *In*: Thierstein H.R. & Young J.R. (eds), *Coccolithophores - From molecular processes to global impact*. Berlin, Springer, pp. 217-250.

PROBERT I., FRESNEL J., BILLARD C., GEISEN M. & YOUNG J.R., 2007 — Light and Electron Microscope Observations of *Algirosphaera robusta* (Prymnesiophyceae). *Journal of phycology* 43(2): 319-332.

QUEIROZ K. & DONOGHUE M.J., 1988 — Phylogenetic systematics and the species problem. *Cladistics* 4(4): 317-338.

RAPPE M.S., SUZUKI M.T., VERGIN K.L. & GIOVANNONI S.J., 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. *Applied and environmental microbiology* 64(1): 294-303.

ROBISON-COX J., BATESON M. & WARD D., 1995 — Evaluation of nearest-neighbor methods for detection of chimeric small- subunit rRNA sequences. *Applied and environmental microbiology* 61(4): 1240-1245.

RYTHER J.H., 1969 — Photosynthesis and Fish Production in the Sea. *Science* 166(3901): 72-76.

SAEZ A.G., PROBERT I., GEISEN M., QUINN P., YOUNG J.R. & MEDLIN L.K., 2003 — Pseudo-cryptic speciation in coccolithophores. *Proceedings of the national academy of science* 100(12): 7163-7168.

SÁEZ A.G., PROBERT I., YOUNG J.R., EDVARDSEN B., WENCHE E. & MEDLIN L.K., 2004. — A review of the phylogeny of the Haptophyta. *In*: Thierstein H.R. & Young J.R. (eds), *Coccolithophores - from molecular processes to global impact*. Berlin, Springer, pp. 251-270.

SAITOU N.N.M., 1987 — The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution* 4(4): 406-425.

SCHLOSS P.D. & HANDELSMAN J., 2005 — Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness. *Applied and environmental microbiology* 71(3):1501-1506.

SHANNON C.E., 1948 — A mathematical theory of communications. *Bell system technical journal* 27: 379-423 and 623-656.

SPEKSNIJDER A.G.C.L., KOWALCHUK G.A., DE JONG S., KLINE E., STEPHEN J.R. & LAANBROEK H.J., 2001 — Microvariation Artifacts Introduced by PCR and Cloning of Closely Related 16S rRNA Gene Sequences. *Applied and environmental microbiology* 67(1): 469-472.

STOECK T., HAYWARD B., TAYLOR G.T., VARELA R. & EPSTEIN S.S., 2006 — multiple PCR-primer approach to access the microeukaryotic diversity in environmental samples. *Protist* 157(1): 31-43.

SUZUKI M.T. & GIOVANNONI S.J., 1996 — Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and environmenta microbiology* 62(2): 625-630.

SWOFFORD D.L., 2002. — PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4b10. Sinauer Associates, Sunderland, MA.

TAMURA K., DUDLEY J., NEI M. & KUMAR S., 2007 — MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Molecular Biology and Evolution* 24(8): 1596-1599.

DE VARGAS C. & PROBERT I., 2004 — New keys to the Past: Current and future DNA studies in Coccolithophores. *Micropaleontology* 50(Suppl_1): 45-54.

DE VARGAS C., AUBRY M., PROBERT I. & YOUNG J., 2007. — The origin and evolution of coccolithophores: form coastal hunters to oceanic farmers. *In*: Falkowski P. & Knoll A. (eds), *Evolution of primary producers in the sea*. New York, Elsevier Academic Press, pp. 251-286.

VENTER J.C., REMINGTON K., HEIDELBERG J.F., HALPERN A.L., RUSCH D., EISEN J.A., WU D., PAULSEN I., NELSON K.E., NELSON W., FOUTS D.E., LEVY S., KNAP A.H., LOMAS M.W., NEALSON K., WHITE O., PETERSON J., HOFFMAN J., PARSONS R., BADEN-TILLSON H., PFANNKOCH C., ROGERS Y.-H. & SMITH H.O., 2004 — Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science* 304(5667): 66-74.

WANG G.C. & WANG Y., 1997 — Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Applied and environmental microbiology* 63(12): 4645-4650.

WHEELER Q.D., RAVEN P.H. & WILSON E.O., 2004 — Taxonomy: impediment or expedient. *Science* 303: 285.

WINTER A., JORDAN R. & ROTH P., 1994 — Biogeography of living coccolithophores in ocean waters. *In*: Winter A. & Siesser W. (eds), *Coccolithophores*. Cambridge, Cambridge University Press, pp. 161-177.

YOUNG J.R., GEISEN M., CROS L., KLEIJNE A., SPRENGEL C., PROBERT I. & OSTERGAARD J., 2003 — A guide to extant coccolithophore taxonomy. *Journal of nannoplankton research* Special Issue 1: 125.

YOUNG J.R., GEISEN M. & PROBERT I., 2005 — A review of selected aspects of coccolithophore biology with implications for paleobiodiversity estimation. *Micropaleontology* 51(4): 267-288.

ZIVERI P., BAUMANN K.-H., BÖCKEL B., BOLLMANN J. & YOUNG J.R., 2004. — Present day coccolithophore biogeography of the Atlantic Ocean. *In*: Thierstein H.R. & Young J.R. (eds), *Coccolithophores – From molecular processes to global impact*. Berlin, Springer, pp. 403-427.