Supporting Information

Cuvelier et al. 10.1073/pnas.1001665107

SI Materials and Methods

This supporting text contains a more detailed discussion on all aspects of the materials used, methods, and data analysis.

1. Field Sampling. Eight transect cruises were performed in the Florida Straits between South Florida and the Bahamas. Flow cytometry (FCM), microcopy, DNA, and HPLC samples were collected at the surface and deep chlorophyll maximum (DCM). Most Florida Straits cruises were in 2005, specifically, WS0503, 31 March–April; WS0510, 18–19 May; WS0515, 24–25 June; WS0518, 31 July–1 August; WS0523, 27–28 September; WS0528, 7–8 December, and in 2007 flow sorting was also performed along this transect (WS0705, 27 February) (Table S1). In addition, a transect from coastal New England to the Bermuda Atlantic Time-series Study (BATS) station in the Sargasso Sea (OC413, 23 May–12 June 2005) was performed on which microscopy, DNA, and FCM (12 depths per profile) samples were prepared from standard water column samples as well as dilution experiment samples. Dilution experiments were performed according to the methods of ref. 1 using modifications in ref. 2. Experimental work was conducted at two main stations: BATS, which had early summer conditions with the mixed layer extending to 30 m, and a nearby Northern Sargasso Sea (NSS) station (35° 10′ N, 66° 33′ W), where stratification was less pronounced and the mixed layer extended slightly deeper, to 40 m. On several other cruises only samples for enumerating and sizing picophytoplankton groups were collected and analyzed, specifically, N92S (Equatorial Pacific Ocean, 25 April–5 May 1992), N92F (Equatorial Pacific Ocean, 10–17 September 1992), N93 (Atlantic Ocean, 7 July–28 August 1993), N95 (Indian Ocean, 24 September–23 October 1995), N96 (South West Pacific Ocean/Southern Ocean, 19 January–1 February 1996), S201 (North East Pacific Ocean, 14–25 March 2001). For three additional cruises only DNA samples were analyzed (surface and DCM); these were in the Western Pacific (CN207, October 2007) and the North Atlantic Ocean, from the US east coast to BATS (EN351 and EN360 in April and September 2001, respectively).

For all of the cruises, seawater was collected using either GO-FLO bottles or a Sea-Bird Niskin Rosette equipped with standard conductivity, temperature, and depth (CTD) and photosynthetically active radiation detectors. Vertical profiles of temperature, salinity, irradiance, and fluorometry were recorded in situ at each station. A series of other results and metadata have previously been published for many of these cruises (3–8). Additional metadata for some cruises is available at http://www.mbari.org/bog/roadmap/roadmap.htm#data.

2. Flow Cytometry and Cell Sorting. The flow cytometer used for sorting cells (Influx, Cytopeia, now Becton Dickinson) was equipped with a 488-nm laser (200 mW output) and a 70-μm-diameter nozzle and generally run at a flow rate of 25 μL min^{-1}. Forward angle light scatter (FALS), pulse width, side scatter (90° angle; SSC), red autofluorescence from chlorophyll (692 ± 40-nm band-pass filter), and orange autofluorescence from phycobiliproteins (527 ± 27-nm band-pass filter) were recorded (log-integrated for scatter and fluorescence). The trigger was FALS. This instrument is only ever run using sterile solutions as sheath and is always cleaned extensively and air dried before shutdown. For all cell sorting (hereafter "the sort," "sorts," "flow sorts"), the instrument was cleaned extensively upon setup to eliminate potential sources of contamination. The sheath and sample lines were flushed before running samples using a series of 10% bleach (in 0.2 μm filtered 18.2 MΩ water); 0.2 μm filtered 18.2 MΩ water; and finally 70% ethanol (in 0.2 μm filtered 18.2 MΩ water). Specifically, the sheath tank was filled with the 10% bleach solution, which was run through all of the lines (sheath and sample) for 5 min at a high flow rate. While the lines were filled with the bleach solution, a series of on/off cycles were performed for all pneumatic valves to disrupt any possible particulate buildup. The sheath tank was then removed, emptied, and rinsed five times, filled with 0.2 μm filtered 18.2 MΩ water, and the lines then rinsed for 10 min. Finally, the sheath tank was filled with the 70% ethanol solution and run just long enough to fill all of the sample and sheath lines before stopping the run; all lines were then "blown dry" with filtered compressed air. In all sorts, 1× PBS solution was used as sheath fluid. PBS was prepared from a sterile 20× solution with 0.2 μm filtered 18.2 MΩ water and was subsequently 0.2 μm filter sterilized again and autoclaved before use. Seawater samples were sorted within hours of collection.

Natural populations (and controls) were flow sorted, with capture in two directions, directly into nuclease/pyrogen-free cryovials and frozen at −80 °C. The volume of the sorted droplet was ≈1 μL. Populations were selected according to specific SSC, FALS, pulse width, and chlorophyll autofluorescence criteria, with gates from all of these parameters, as well as an orange fluorescence exclusion gate, used to define each of the populations sorted (Fig. 1, main text), increasing the stringency, although also decreasing yields because only particles meeting all criteria were sorted. A subset of replicate sorts was immediately resuspended and rerun to determine the sort efficiency (32% for sorts used for metagenome material) [i.e., the true number of cells that would then go into the multiple displacement amplification (MDA) reaction used to amplify the whole genome, detailed below]. Approximately 300 cells were actually sorted as determined by this approach. Control sorts (to test for contamination) included sheath fluid collected from the sheath reservoir as well as collected from the test streams that run through the sample lines (see below). The piezo amplitude was 0.56 V, drop frequency 49.3 kHz, and drop delay ≈34.5 droplets.

Flow cytometry was also used to determine picophytoplankton cell counts on the global transects. For "WS" and "OC" cruises FCM samples (1 mL) were collected in triplicate from each depth, preserved with 0.25% glutaraldehyde (final concentration; Tousimis) and stored in the dark for 20 min, a modification of previously published methods (9). Again, data collection was triggered on FALS. Instrument setup and data collection were as above. Samples were flash-frozen in liquid nitrogen and either stored in liquid nitrogen until processing or kept at −80 °C for long-term storage. Samples were thawed just before analysis, and a known volume of yellow-green 0.75-μm beads (Polysciences) was added and used as internal fluorescence and light scatter standards (9). Photo-multiplier tubes were at relatively high voltage settings to enumerate *Prochlorococcus* at the same time as other picophytoplankton.

For all "N" and "S" cruises, samples (1.8 mL) were fixed in 0.2% paraformaldehyde (final concentration) and stored in liquid nitrogen. Samples were analyzed using a Coulter EPICS 753 flow cytometer equipped with two 5-W argon lasers, and data were collected for abundance and fluorescence characteristics according to previously published protocols (5, 10–12).

Listmodes were analyzed either using CYTOPC software (see ref. 12) or WinList (Verity Software House). *Prochlorococcus* and *Synechococcus* were defined according to FALS and fluorescence characteristics (13). Note that for a small number of samples in the Florida Straits the *Prochlorococcus* population

intersected with baseline noise in the red-fluorescence channel. These samples were not used for biomass averages (Fig. 4, main text) or for pico-prymnesiophyte contributions to total phytoplankton biomass. "Nonprymnesiophyte" picoeukaryotes were enumerated using analysis windows as by Buck et al. (5), who showed that for field samples analysis using these windows (encircling the smallest eukaryotes) rendered FCM results that were tightly correlated with the sum of all picoeukaryotes, excluding pico-prymnesiophytes, enumerated by microscopy.

3. Whole-Genome Amplification and Sample Prescreening. Duplicate sorts from two environmental sites underwent Multiple Displacement Amplification (MDA) (Repli-g Midi kit; Qiagen) using methods similar to those in ref. 14 after alkaline lysis (KOH, 10 min, on ice). These Florida Straits sorts were from a right sort population at Station 04, 75 m (used for sort clone libraries and metagenomics) and a population with similar characteristics at Station 08, 141 m (used for sort clone libraries) east of Station 04 (adjacent blue circle, Fig. 2 *Inset*, main text). Station 04 was in the core of the Gulf Stream, as determined by Acoustic Doppler Current Profiler data. Duplicates of each of three different control sorts also underwent MDA, specifically (*i*) sheath fluid run through the entire sheath system, collected after sort test deflection, before introduction of seawater samples to the system on the sort day; (*ii*) sheath run as sample through the sample line collected in tube using sort test deflection; and (*iii*) sheath run through the entire sheath system, collected after sort test deflection (but performed later in the day, after environmental sorts). Hence, the environmental sorts were performed between controls *ii* and *iii*. Finally, a positive DNA control (100 pg gDNA; Qiagen) and a negative control (H_2O) also underwent MDA (both in duplicate). After storage at −80 °C, sort samples and the above controls were transferred to a thin-walled microfuge tube, sample volume determined, and brought to a total volume of 2 μL with Tris-EDTA (TE) buffer; negative controls were performed using the same "template" volumes. The total reaction volume was 10 μL, and handling of all reagents and samples was performed in a PCR workstation with high efficiency particulate filtered air supply. The reactions were incubated at 30 °C for 16 h according the recommended protocol from the manufacturer (Repli-g Handbook, Qiagen). The amplifications were subsequently diluted 5-fold with TE buffer before heat-inactivation at 65 °C for 3 min. The MDA products then served as template for PCR to construct preliminary 16S and 18S rRNA gene libraries (see below) to select a sample to advance for metagenomic sequencing. The purpose of this quality-control step was also to verify potential contamination in the sample handling from collection through whole-genome amplification. Negative controls (H_2O and all sheath sorts) did not render 16S or 18S rRNA gene PCR products. A small number of clones from the flow-sorted phytoplankton population were then sequenced (≈10 clones per replicate) and used to screen different MDA products for target organisms (Fig. 2 in main text and Fig. S1*A*, from sorts at Station 04 and Station 08; see below). We also tested the efficiency of the alkaline lysis used, showing that, at this stage, ≈54% of the cells in the sorts used herein were lysed.

4. Size-Fractionated and Preliminary Flow Sort 18S rRNA Gene Clone Libraries. Standard clone libraries were generated for multiple samples from three Florida Straits cruises, WS0503, WS0518, and WS0528 (Station 14 only) and all of the Sargasso Sea cruises (EN360, EN351, and OC413) from samples collected, processed, and extracted as in refs. 3 and 4. Typically 1 L of seawater was collected, prefiltered by gravity although a 2-μm polycarbonate filter (GE Osmonics) and then vacuum filtered through a 0.2-μm (OC413, WS0503, WS0518, WS0528) or 0.45-μm (EN351, EN360) Supor filter (Pall Gelman). The Supor filter was immediately frozen cryogenically and subsequently moved to −80 °C for long-term storage. In addition, clone libraries were built for CN207, and small preliminary libraries were built from the two MDA products (≈10 clones each) for the 18S rRNA gene, as well as the 16S rRNA gene, which captures both bacteria and eukaryotic chloroplast 16S rDNA sequences (see below). For CN207, seawater from the Niskin rosette was transported to a large reservoir that had been cleaned with a 10% HCl solution. Cells were collected on a 0.8-μm pore size, 293-mm Supor filter (Pall Gelman) after prefiltration through a 3-μm pore size filter (in series, both under vacuum). Before collecting samples the entire filtration system and reservoirs were flushed with a solution composed of 1:9 bleach:18.2 MΩ H_2O to reduce the possibility of contamination. These large filters were flash-frozen by suspension in liquid nitrogen vapor and stored at −40 °C. For CN207, a sucrose extraction protocol (http://www.mbari.org/phyto-genome/resources.html) was used to extract DNA from the 293-mm filters. Environmental conditions for clone library sites are shown in Table S1.

18S rRNA genes were amplified using primers complimentary to conserved regions proximal to the gene termini (forward 5′-ACCTGGTTGATCCTGCCAG-3′ and reverse 5′-TGATCCTTCYGCAGGTTCAC-3′), designed as universal eukaryotic primer set, but likely with some biases (15, 16). Briefly, PCR was performed with an initial "hot start" for 15 min at 95 °C, proceeded by 32 cycles at 94 °C for 30 s, 55 °C for 30 s and 72 °C for 1 min; followed by a final extension at 72 °C for 10 min, as in ref. 4. One microliter of PCR product was ligated into the vector pCR2.1 using the TOPO-TA cloning kit (Invitrogen) and transformed; after colony selection and growth, plasmids were purified. "WS" and "OC" cruises were sequenced with a single primer internal to the PCR product (502F) that rendered a unidirectional product for all 18S rRNA gene clones. "CN" and "EN" cruises were sequenced with a suite of primers, two plasmid targeted primers (M13F and M13R), and two primers internal to the product [1174R and 502F (17)]. Sequencing was performed using Big Dye Terminator chemistry on an AB3730xl sequencers (Applied Biosystems). For "WS" and "OC" cruises, 96 clones were sequenced per library, for "EN" cruises 40 clones were sequenced per library.

BLASTN against the GenBank nonredundant (nr) database was used to make a preliminary taxonomic affiliation for the sequences obtained from the clone libraries. In the Sargasso Sea, prymnesiophyte sequences were retrieved from all four environmental clone libraries (surface and DCM at BATS and NSS) and accounted for 1–12% of the total number (96) of sequences in each library. In the Florida Straits, prymnesiophyte sequences were retrieved from 13 of the 14 clone libraries (1–17% of the total number of sequences in each library). Chromatograms and assemblies of all 18S rRNA gene sequences tentatively assigned to the prymnesiophytes were manually curated. For phylogenetic analyses, only curated sequences were analyzed, alongside prymnesiophyte sequences and out-group sequences retrieved from GenBank (last retrieval February 2009). Manual screening was use to detect chimeras, which were subsequently removed. Sequences were aligned using ClustalW (18). Preliminary neighbor-joining trees were built using PHYLIP modules (19) and 280 sequences (including out-groups). Generally only a single representative of a cultured species, or a strain, as well as a single representative from each clone library found within a single clade, was kept for subsequent phylogenetic analysis. A total of 139 environmental sequences were then used in the final tree, including 111 from our samples (72 from the Florida Straits, 27 from the Sargasso Sea, 7 from the Western Pacific, and 5 from the MDA; see below). We also used 28 environmental sequences retrieved from GenBank, 5 from an earlier study of ours in the Sargasso Sea (3), 4 from the Indian Ocean (20), 13 from the Equatorial Pacific (15, 16), 2 from L'Atalante deep-sea basin (21), 1 from coastal subtropical Western Pacific (22), 1 from the coastal North Western Mediterranean Sea (23), 1 from the Southern California Bight (17), and 1 from unknown origins (location is not specified in the GenBank entry). This represented all environmental prymnesiophyte 18S rRNA gene sequences housed at GenBank as of February 2009 that

had sufficient overlap with our sequenced products to be included in alignments. These sequences were then realigned with ClustalW, and the alignment was manually edited. The final 18S rRNA gene phylogenetic analysis (Fig. 2, main text) was performed using maximum likelihood methods in PhyML (24) after prediction of the best evolutionary model (in this case GTR+I+G) using ModelTest (25). Model parameters used were 1.5886, T_iT_v; 0.3044, p_{inv}; 0.6026, γ distribution shape (α). Data were bootstrapped with 100 replicates. Out-group sequences were from red- and green-lineage organisms, specifically: *Chondrus cripus* (Z14140), *Gracilaria lemaneiformi* (M54986), *Compsopogon coeruleus* (AF342748), *Cryptomonas ovata* (EF180057), *Cryptomonas pyrenoidifera* (AJ421147), *Hemiselmis virescens* (AJ007284), *Rhodomonas salina* (EU926158), *Pyrenomonas helgolandii* (AB240964), *Pyramimonas australis* (AJ404886), *Chlamydomonas reinhardtii* (AY665726), *Chlorella vulgaris* (AY591515), *Micromonas* CCMP1723 (AY954997), *Micromonas* CCMP1545 (AY954994), *Symbiodinium microadriaticum* (EF492496), *Prorocentrum micans* (AJ415519), *Karlodinium micrum* (EF492506), *Coscinodiscus radiates* (X77705), *Thalassiosira weissflogii* (AY485445), *Thalassiosira pseudonana* (DQ093367), *Gloeochaete wittrockiana* (X81901), *Cyanophora paradoxa* (AY823716), and *Glaucocystis nostochinearum* (X70803). For additional phylogenetic analysis of flow sorts see *SI Materials and Methods, Section 5*.

The topology of the overall 18S rRNA gene tree was consistent with previous reports (Fig. 2, main text, and Fig. S1*A*). Bootstrap values at interior nodes were low as commonly seen for 18S rRNA gene trees. Node support was especially low for the Prymnesiales, and most deep branches were unresolved. 18S rRNA gene phylogenies are known to have limited resolution and hence although the trees clearly demonstrate the extensive diversity of uncultured taxa within the prymnesiophytes, evolutionary relationships are difficult to discern. The Pavlovales formed a supported clade distinct from, and basal to, the prymnesiophytes, and several of our clone library sequences were placed basal to cultured prymnesiophytes, but inside the Pavlovales, as seen elsewhere (26). In the 25 size-fractionated libraries, only one sequence, from a single date, was unambiguously assigned to a cultured species (100% *Phaeocystis globosa*). Several broad "clades" identified previously within the prymnesiophytes were represented (Table S2), specifically clades A to E as per refs. 27–29. Few clade C sequences were recovered, likely owing to (*i*) our size fractionation step, which would exclude many of these cells; and (*ii*) the fact that few cells within an appropriate size range for this clade were observed by FISH. In the Florida Straits, of the 36 FISH samples analyzed only 8 had >15% of the prymnesiophyte cells falling within the >3-μm size fraction (*SI Materials and Methods, Section 10*, and Fig. S1*A*); the majority were <3 μm in size. In these eight samples, 28% ± 10% were in the 3–10-μm size fraction, with the rest being smaller.

Note that here the term "prymnesiophytes" is used to refer to the class Prymnesiophyceae Hibberd, which seems to be the most consistent usage in previous oceanographic literature. In general, this group is alternatively referred to as the division Haptophyta (division Haptophyta Hibberd ex Edvardsen et Eikrem), including both the Prymnesiophyceae and the Pavlovophyceae (Cavalier-Smith) Green et Medlin. However, classification of haptophytes has differed noticeably between authors (e.g., refs. 30–33). We adopted the nomenclature of Edvardsen et al., 2000, wherein "coccolithophores" incorporates all haptophytes with calcified scales (Coccolithophores) during some stages of their life cycle (34) and includes two orders: Coccolithales (E. Schwarz) Edvardsen et Eikrem and Isochrysidales (Pascher) Edvardsen et Eikrem.

5. Targeted Metagenome Sequencing and Additional Small Subunit Characterization. The Florida Straits Station 04 MDA-flow sort product was advanced for shotgun cloning and pyrosequencing, as well as being phylogenetically characterized at greater depth. The MDA product was debranched using an S1 nuclease (Fermentas) digestion with 10 U μL^{-1} at 37 °C for 1 h (60 μL reaction). The enzyme was heat-inactivated in the presence of EDTA, and the DNA was phenol-chloroform-isoamyl alcohol extracted and ethanol-precipitated. For Sanger sequencing, 3-kbp shotgun libraries were constructed using the debranched MDA products. For shotgun library construction, MDA products were randomly sheared to 2–4-kbp fragments using a HydroShear (GeneMachines). The sheared DNA was separated on an agarose gel, gel-purified using the QIAquick Gel Extraction Kit (Qiagen), and blunt-ended using T4 DNA polymerase (Roche Applied Science) and Klenow Fragment (New England Biolabs) in the presence of dNTPs and NEB2 buffer. The 2–4-kbp DNA fragments were ligated in pUC19 vector (Fermentas) O/N at 16 °C using T4 DNA ligase (Roche Applied Science) and 4.5% polyethylene glycol (Sigma). The ligation products were phenol-chloroform extracted and ethanol precipitated. According to the manufacturer's instructions, ligations were electroporated into ElectroMAX DH10B Cells (Invitrogen) and clones prepared and sequenced on an ABI PRISM 3730 capillary DNA sequencer (Applied Biosystems) according to standard Joint Genome Institute (JGI) protocols (www.jgi.doe.gov). Pyrosequencing was performed on debranched MDA products using the Genome Sequencer FLX System (454 Life Sciences) (35), according to the manufacturer's protocol.

To characterize phylogenetic diversity at greater depth than the initial prescreening, 18S (515F 5′-GTGCCAAGCAGCCGCGGTAA-3′, 1209R 5′-GGGCATCACAGACCTG-3′) and 16S (27F 5′-AGAGTTTGATCCTGGCTCAG-3′, 1391R 5′-GACGGGCRGTGWGTRCA-3′) rRNA gene (PCR) clone libraries were created from debranched MDA products using the above universal primers. PCR amplicons of five replicate reactions were combined and ligated into the pCR4-TOPO vector using the TOPO TA Cloning Kit (Invitrogen). Ligations were then electroporated into One Shot TOP10 Electrocomp *Escherichia coli* cells and plated on selective media agar plates. The bidirectional rDNA sequence reads were end-paired and trimmed for PCR primer sequence and quality. Because we used flow sort parameters to target a discrete population of photosynthetic eukaryotes that contained both pico-prymnesiophytes and *Pelagomonas*, we characterized overall diversity in the targeted metagenome from Florida Straits Station 04 by a comprehensive phylogenetic analysis of sequences within the specific sample. A greater number of clones were sequenced for these libraries than for initial libraries used to select the flow sort to be advanced for metagenomic sequencing. A clustering approach was first applied to the 663 16S and 326 18S rDNA successful sequences, by clustering at the 98.9% identity level with 90% overlap using BLASTCLUST (36). Only clusters with three or more representatives were further analyzed for 18S rRNA gene sequences to exclude those with assembly errors, because manual curation of assemblies was not possible. Maximum likelihood phylogenetic reconstructions were performed using PhyML (24). The TrN+G model of evolution was selected for the 18S rDNA phylogenetic reconstruction, which used 749 sites and AICc criterion; the likelihoods of different substitution models were computed by PhyML using MrAIC [Nylander JAA (2004) MrAIC.pl. Program distributed by the author. Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden]. In the case of the 18S rRNA gene phylogenetic reconstruction, 176 of the 326 sequences formed pico-prymnesiophyte clusters (Fig. S1*B*). Use of more stringent criteria than used in BLASTCLUST showed that those clustering with group 8 all had 99–100% identity to representatives from this group across the entire sequence. The same was found for those clustering with group 3. Fifty-eight clones formed a third cluster by the BLASTCLUST criteria and showed 99–100% identity to a range of sequences from the tip of the tree (i.e., the region of the tree starting around group 14 and ending with sequences near group 22), for example with 100% identity to group 15. Sequences in this region were all >97% identical over the entire sequence, which

also included some *Chrysochromulina* species. The extent to which these sequences represent different taxa, or possibly microdiversity within multiple copies of the 18S rRNA gene in an individual taxon, is unknown. Two other groups were seen as well, one branching with *Micromonas* (comprising 6 sequences) and two with *Pelagomonas calceolata* (one with 115 and one with 3 sequences). Phylogenetic analysis was also performed on 16S rRNA gene sequences derived from the metagenomic scaffolds using 678 positions and the TrN+I+G model of evolution determined as above (Fig. S3*A*). The three pico-prymnesiophyte chloroplast scaffolds recovered each contained a 16S rRNA gene sequence with 99% identity to one of each of three 16S rRNA gene clusters from clone library sequences, supporting the fidelity of the metagenomic assembly process. Clustering of the 663 16S rDNA sequences from the MDA-flow sort clone libraries showed the same major pico-prymnesiophyte groups, in addition to 148 *Pelagomonas* sequences. 16S rDNA clone library sequence assemblies were not manually curated, and a large number were mitochondrially derived (322, with 315 falling within one 99% identity cluster), whereas 12 sequences (forming two 99% identity clusters, one with 11 sequences and one singlet) branched with bacteria, although not with high identity to known sequences. Because of the paucity of relevant mitochondrial 16S rDNA sequences in GenBank, mitochondrial sequences were not analyzed further. *Micromonas* and *Pelagomonas* 16S rDNA clades also attracted three plastid scaffolds from the complete metagenome, before any screening and removal, with sizes 1,844 nt, 5,672 nt, and 92,558 nt, the latter two being from *Pelagomonas*. Three primary groups were identified in a single sort from a recent study in the North East Atlantic, in which a clone library was constructed at a single station among the 20 stations where primary production was measured (37). Of the clone groups recovered, one corresponded to ours (99% to C5574; Fig. S3*A*), and the other two bore 95% and 93% highest identity to the 664 16S rDNA clones we sequenced. However, the 16S rRNA primer sets used were different from ours and may not recover the same phylotypes with the same efficiency, or these results possibly reflect true compositional differences.

In both the 18S rDNA and chloroplast genome–derived 16S rDNA trees, Pavlovales sequences served as an out-group along with a number of more distant taxa (e.g., prasinophytes and pelagophytes). Because these PCR-based analyses of the MDA product confirmed that more than one eukaryotic taxon was present in the sort (as expected given the cell number and size of the sort population), we applied a stringent criterion to select scaffolds that would undergo further analyses, and all data failing these criteria were removed from further analyses (see below).

6. Assembly, Gene Modeling, and Filtering of the Pico-prymnesiophyte Metagenome. A Newbler assembly was run with all 454 data (1,031,617 reads totaling 209,914,830 bases). Fifty-two percent of the reads assembled in some way, forming a total of 89,375 contigs containing 27,027,311 bases; of these, 20,118 were "large" (>500 bp) contigs containing 10,824,818 bases, the largest of which was 21 kb. For assembly into a scaffold a minimum of 40 nt overlap at 90% nt identity was required. The large contigs only were "shredded" into 1,000-bp pieces with 100-bp overlaps and used for the subsequent assembly. From this process there was a remainder of 41,450 "small" (<499 bp) 454 contigs (11,514,939 bases) and 294,515 454 singlets (≈64 Mb). Sanger data (125,925 reads) was trimmed with Lucy (version 1.19p) and assembled along with the 454 "shreds" from the above assembly using the Paracel Genome Assembler (PGA version 2.6.2). This produced 19,905 contigs totaling 32,538,830 contiguous bases and 52,057 singlets containing 43,568,548 bases. Singlets were not considered further. Approximately 64% of the Sanger reads and 76% of the 454 shreds went into the assembly. This two-phase assembly process was found to produce larger and more accurate contigs than a single Newbler assembly with all of the data. Notably, neither type of sequencing platform reached saturation; that is, assembly continued to improve with each incremental increase in sequencing.

A phylogenomic approach was used to narrow the metagenomic data for analysis of pico-prymnesiophytes only. First, an initial set of nuclear genes was predicted from the 19,905 PGA metagenome scaffolds using the Program to Assemble Spliced Alignments (38). This approach used a combination of a protein-to-genome alignment method using GeneWise (39) and an expressed sequence tag (EST)-to-genome alignment method with the most up-to-date *Emiliania huxleyi* EST dataset (from GenBank). Initially, GeneWise alignments were performed to generate the preliminary pico-prymnesiophyte protein set using filtered model proteins from 12 annotated protistan genomes, from which gene predictions were generated at JGI. The following 12 species were used as references for GeneWise alignments: *Micromonas* CCMP1545 (JGI; v2.0), *Micromonas* RCC299 (JGI; v2.0), *Aureococcus anophagefferens* (JGI; v1.0), *Ostreococcus lucimarinus* CCE9901 (JGI; v2.0), *Ostreococcus tauri* (JGI; v2.0), *Phaeodactylum tricornutum* CCAP 1055/1 (JGI; v2.0), *T. pseudonana* CCMP1335 (JGI; v3.0), *C. reinhardtii* (JGI; v3.0), *Phytophthora ramorum* (JGI; v1.0), *Phytophthora sojae* (JGI; v1.0), *Volvox carteri f. nagariensis* (JGI; v1.0), and *E. huxleyi* CCMP1516 (JGI; v1.0). Regions of the selected scaffolds that did not contain protein alignments but contained EST alignments were evaluated and genes created (or updated) according to EST data. As a result, 25,230 genes were predicted in the targeted metagenome assemblies, representing a total of 32.5 Mb of scaffold sequence. Singletons, representing 44 Mb of sequence, were not considered.

ORFs were predicted (Table S3) on the largest chloroplast genome scaffold (C19847) in addition to the genome modeling performed for nuclear scaffolds (above). This scaffold contained a pico-prymnesiophyte plastid 16S rRNA gene (Fig. S3*A*) and was annotated using National Center for Biotechnology Information and BLASTX (36) searches to the nonredundant protein sequence database and a custom database containing all complete chloroplast genomes. rRNAs were identified by BLASTN against GenBank-NT. tRNA sequences (Table S4) were identified using tRNASCAN-SE (40). This chloroplast metagenome assembly had 9X coverage, and the 16S rDNA sequence found on this scaffold had 99% identity to the largest cluster detected in the 16S rDNA clone library. Manual observation of this chloroplast scaffold indicated a few polymorphisms in the small subunit (SSU) sequence and fewer over the rest of the scaffold.

All predicted nuclear-encoded genes were analyzed phylogenetically using APIS (Automated Phylogenetic Inference System; see, e.g., ref. 41), an automated system for creation and summarizing of phylogenetic trees for each protein encoded by a genome. The homologs used by APIS for each phylogenetic tree were obtained using BLASTP (42) to compare query proteins against an extended version of ComboDB that contained taxonomic, genomic, protein, and coding DNA information for 46 eukaryotic genomes, including 10 phytoplankton genomes and two additional protistan stramenopile genomes, as well as 52 archaeal, 687 bacterial, and 1,928 viral complete (or nearly complete) genomes (as of June 1, 2008). The full-length sequences of these homologs were retrieved from the database and aligned using MUSCLE (43). Bootstrapped neighbor-joining trees were produced using QuickTree (44). The inferred tree was then midpoint-rooted before analysis, allowing automated determination of taxonomic classification of the query sequence based on placement within known taxonomic groups. The bootstrap value of the node connecting metagenomic sequences from Station 04 to the only available prymnesiophyte genome (*E. huxleyi*) was noted to identify particularly robust groupings. Sequence scaffolds were filtered to remove those that did not contain a gene set of which at least half of the APIS classifiable genes were classified as a prymnesiophyte (i.e., directly sistered by *E. huxleyi* to the exclusion of all other taxa). Predicted genes from this reduced set of scaffolds, repre-

senting a 2-Mb assembly (≈10% of the total *Micromonas* genome size), were further filtered by SEG (45), and genes were eliminated from further analysis that contained fewer than 30 residues outside of low-complexity regions. A final set of 1,624 pico-prymnesiophyte genes was used for nuclear genomic analysis. GC% of the reduced assembly was 60%, with the average of scaffolds composing the assembly being 61% (SD 8%). Average coverage for these scaffolds was 1.9X.

7. Functional and Evolutionary Analysis of the Targeted Pico-prymnesiophyte Metagenome. Predicted nuclear-encoded proteins underwent functional annotation using several different approaches. BLASTP searches were performed against an internal database of predicted proteins from phytoplankton genomes (downloaded from JGI), Panther, and Swissprot, as well as domain (Pfam and TIGRfam) and motif searches (Prosite and InterPro), and programs to predict membrane localization, such as SignalP, TMHMM, and TargetP, as well as CDD (see, e.g., refs. 46–49).

Comparisons of gene density were done on a sliding window of 100 kb sampled at 10-kb increments to estimate the distribution of local gene density. The comparison of gene density on a partial set of relatively short scaffolds with that of much larger scaffolds or chromosomes of the selected references necessitates use of a sliding window to visualize local variations in gene density. Gene density was calculated as the number of nucleotides within the bounds of a predicted gene per total nucleotides in the window of 100 kb or smaller. Smaller windows result from contigs or regions smaller than 100 kb. This measure was used, as opposed to the number of genes per kilobase, to account for variation in gene sizes. The gene boundaries used were those predicted by GeneWise models.

Sequence coverage and pico-prymnesiophyte assembly genome size were estimated from core marker genes from our metagenomic assemblies, as well as from a predictive model of gene abundance from functional diversity. Near single-copy core marker genes were identified for reference green algal and chromalveolate genomes: *A. anophagefferens*, *C. reinhardtii*, *Cryptosporidium parvum* Iowa II, *Ectocarpus siliculosus*, *Micromonas* CCMP1545, *Micromonas* RCC299, *O. lucimarinus* CCE9901, *O. tauri*, *P. tricornutum*, *P. ramorum*, *P. sojae*, *Plasmodium falciparum* 3D7, *Tetrahymena thermophila*, *T. pseudonana*, and *V. carteri*. Clustering of all peptide sequences was based on reciprocal best BLASTP hits, and fragment hidden Markov models (HMMs) were built using HMMER (49) from clusters containing at least one peptide from each of the reference genomes. The reference genomes were then searched using the initial set of HMMs, and a final set of 132 HMMs was chosen that resulted in near single-copy matches (between one and three) to each of the reference genomes at e-value $\leq 1.0e^{-5}$, with an SD of the number of matches to all reference genomes at or below 0.66. The set of 1,624 pico-prymnesiophyte genes was searched using the 132 core HMMs, and 19 (14%) core genes were detected at e-value $\leq 1.0e^{-5}$.

At the most basic level, in small genomes, large protein families make up less of the overall genome content, whereas in larger genomes much of the gene content is contained in the largest protein families. A global statistic was developed to describe the spread of functions over the space of KOGs (euKaryotic Orthologous Groups of proteins). Models were trained on a range of subsets of genes from complete genomes and were found to be predictive of the total number of genes in such genomes. The proportion of BLASTP hits at e-value $\leq 1.0e^{-9}$ to all KOGs that account for the largest 20% of KOGs (KOG20) is a statistic summarizing the distribution of genes across functional categories. The KOG20 statistic was chosen for ease of interpretability in general, and specifically the value 20% was determined to maximize the ratio of between- to within-sample variation among reference genomes. The KOG20 measure of functional diversity differs significantly from current measures of the spread of genetic diversity, such as counting lineage-specific expansions of gene families (50, 51). Calculation of KOG20 does not require de novo clustering of gene families, relying instead on preclustered KOGs, and is not lineage specific. Significant expansions of gene families with limited homology to known KOGs would not be detected by measuring the spread of gene families across KOGs. However, KOG20 does represent a measure that is easy to calculate, is comparable among reference chromalveolate and green algal species, and is predictive of a large range of eukaryotic genome sizes.

A relationship exists between the KOG20 value and the total number of genes in reference Chromalveolata and green algal genomes, which range between 5,000 and 25,000 in haploid gene number. The number of genes contained within a "complete" genome for the pico-prymnesiophyte was predicted by calculating KOG20 for the 1,624 predicted genes and comparing this with KOG20 of samples of genes from the reference genomes. The KOG20 values were transformed by –log (1-KOG20) and a linear fit performed by ordinary least squares on reference gene samples using the statistical package R. The reference genomes were sampled randomly to 13% of the total number of genes in each genome and 100 replicates performed. The average KOG20 was calculated for each genome over the 100 replicate samples and a prediction for pico-prymnesiophyte total genes made based on the KOG20 value of the 1,624 pico-prymnesiophyte genes. Predictions were made by sampling reference genomes at fractions ranging from 5% to 30%, and a convergence of predicted pico-prymnesiophyte sequence coverage and reference genome sample fraction used in the prediction was found at 13% (Fig. S2). Sampling reference genomes at exactly 1,624 genes was also performed, and a similar prediction of pico-prymnesiophyte sequence coverage was found.

The pico-prymnesiophyte chloroplast metagenome assembly was analyzed in relation to published chloroplast genomes, including *E. huxleyi* (52). GC content (37.2%) was similar to that of *E. huxleyi* (36.8%), and many other chloroplast genomes, which, according to other literature, tend to have low %GC. Semicircular representation of the pico-prymnesiophyte chloroplast metagenome assembly and the *E. huxleyi* chloroplast genome (52) were obtained using the GenomeVx Web server (53). In addition, we verified presence of other *E. huxleyi* chloroplast gene sequences using these protein sequences as the query and TBLASTN, for those not contained on the largest assembled scaffold. Using this approach, representatives of all genes encoded on the *E. huxleyi* chloroplast genome were identified. For phylogenetic analyses, however, only those on the largest contig, C19847, were included. We used a previously published alignment composed of 44 concatenated chloroplast and cyanobacterial protein sequences originating from 20 different species as an initial alignment template (54). The original alignment was trimmed to 22 protein sequences (a subset within Table S3) to add sequences from the environmental scaffold and the green alga *Micromonas* RCC299. In addition to the latter, sequences from the moss *Physcomitrella patens* were added to increase the number of representatives of the green lineage in the analysis. These concatenated sequences were aligned to the trimmed template alignment using the T-Coffee profile alignment mode (55). The resulting alignment was manually curated, and ambiguously aligned sites were removed along with all gap-containing sites. The final alignment consisted of 4,425 sites. Maximum likelihood reconstruction was performed using PhyML under Jones-Taylor-Thornton amino acid substitution matrix (56). Statistical support was computed using 100 bootstrap replicates. Global nucleotide conservation between chloroplast genomes was performed using MUMMER (with forward and reverse complement matches) and visualized with MUMMERplot (57). The diatom genomes were reported previously (58), as was the *E. huxleyi* chloroplast genome (52).

As mentioned above, phylogenomic analyses were conducted for all predicted nuclear-encoded genes using APIS. Among many other genomes, APIS included stramenopile reference sequences from the following genomes: *A. anophagefferens*, *P. tricornutum*, *P. ramorum*, *P. sojae*, and *T. pseudonana*, as well as the following Archaeplastida: *Arabidopsis thaliana*, *C. reinhardtii*, *Micromonas* CCMP1545, *Micromonas* RCC299, *O. lucimarinus* (CCE9901), *O. tauri*, *P. patens*, *Populus trichocarpa*, *Selaginella moellendorffii*, and *V. carteri*. Venn diagrams were produced from BLASTP hits at e-value $\leq 1.0e^{-9}$ of the 1,624 pico-prymnesiophyte genes to three groups of reference genomes: prasinophyte (*Micromonas* CCMP1545, *Micromonas* RCC299, *O. lucimarinus* CCE9901, *O. tauri*), *Phytophthora* (*P. ramorum*, *P. sojae*), and diatom (*P. tricornutum*, *T. pseudonana*). These groups captured all published marine algal genomes. *Phytophthora* was also included because, although not marine or photosynthetic, it broadened representation of the stramenopiles. The 1,624 pico-prymnesiophyte genes were divided into Venn groups according to their overlapping hits to each of the reference groups, and functional profiles were produced for each Venn group. Gene products were classified using Gene Ontology (GO) (59), a set of organism-independent controlled vocabularies for describing molecular function, biological process, and cellular component. GO assignments were made using Pfam HMM searches below trusted cutoffs and the Pfam2GO tool, which maps Pfam hits to GO terms (46). The relative distributions across GO molecular functions were calculated from BLASTP hits at e-value $\leq 1.0e^{-5}$ and normalized as a proportion of all hits within each Venn group. GO functions are inclusive of all lower-level GO terms. Enzyme Commission (EC) numbers were assigned using PSI-BLAST hits against PRIAM profiles, built using protein sequences from the ENZYME database (60). The EC assignments were then made using a modified algorithm implemented in metaSHARK (61) that searches for the best match to the PRIAM profile.

Transcription-related domains (Tables S5 and S6) were analyzed using HMMsearch from the HMMer 3.0 package with a manually curated HMM file of 374 transcription-related domain alignments, which are collected from the plant transcription factor library (http://plntfdb.bio.uni-potsdam.de/v3.0/) and DBD: transcription factor prediction database (www.transcriptionfactor.org; version 2.0). The inclusion cutoff was e-values <0.001. Searches were done against the 1,624 predicted gene sequences. If more than two domains overlapped in the same sequence, the domain with a lower e-value was selected. Identified domains were then searched against other published genomes for comparison, these genomes included *T. pseudonana* (JGI; v3.0), *P. tricornutum* (JGI; v2.0), *P. sojae* (JGI; v1.1), *T. thermophila* (www.ciliate.org/; August 2004), *P. falciparum* (plasmodb.org/plasmo/; v5.4), *Micromonas* RCC299 (JGI; v3.0), *Micromonas* CCMP1545 (JGI; v2.0), *O. tauri* (JGI; v2.0), *O. lucimarinus* CCE9901 (JGI; v2.0), *C. reinhardtii* (JGI; v4.0), and *A. thaliana* (www.arabidopsis.org/; v8.0).

Because of the relatively high number of SET-domain protein family sequences identified we performed phylogenetic analysis of these genes (Fig. S5*A*). Representative SET-domain sequences were collected from animals, fungi, and plants on the basis of previously defined SET-domain subfamilies (62, 63). Additional SET-domain protein sequences were obtained from the published genomes of green algae and chromalveolates by BLASTP against each representative of all subfamily sequences. Sixty-three SET-domain sequences were aligned in MUSCLE and edited manually. Searches for the best maximum-likelihood tree were performed using RAxML 7.2.3 (64) with a parameter of "-# 100 -f a -m PROTCATIBLOSUM62F." Bootstrap support values are shown only for those greater than 50%, except branches of major subfamilies, for which bootstrap support values were calculated by RAxML and MultiPHYL (65).

8. High-Performance Liquid Chromatography. Samples for pigment analysis were obtained by filtering 1 L to 5 L of seawater, depending on the depth and location, through a 25-mm glass fiber filter (Whatman). The filter was placed in a cryovial and frozen in liquid nitrogen. For analysis, filters were thoroughly dried, placed in 3 mL of 90% acetone, and vortexed for 45 s before placing them at −20 °C. After 24 h, filters were sonicated for 30 s and vortexed again for 45 s. The extract was then cleared through 0.8-μm filters. One milliliter of extract was mixed with 0.2 mL of 0.2-μm-filtered autoclaved 18.2 MΩ water and placed in an Autosampler tray at 4 °C. The HPLC hardware and analysis was performed as previously described (66). Chl *a*, as the sum of monovinyl (MVChl *a*) and divinyl (DVChl *a*) Chl *a*, was used as a measurement of total phytoplankton pigment biomass. *Prochlorococcus* contribution to Chl *a* was estimated directly as DVChl *a*. The contribution of the rest of major groups to MVChl *a* was quantified using Chemtax (67) with a newer version that was provided to us before publication (68). Samples were initially separated in two subgroups: DCM and surface samples. The pigment dataset was carefully checked to distinguish the potential presence of a total of seven phytoplankton groups that could contribute to MVChl *a*: Prymnesiophyceae, Pelagophyceae, Prasinophytae, *Synechococcus*, Cryptophyceae, Dinophyceae, and diatoms; with the following pigments: Chl c_2, peridinin, 19'-butanoyloxyfucoxanthin, fucoxanthin, prasinoxanthin, violaxanthin, 19'-hexanoyloxyfucoxanthin, alloxanthin, and zeaxanthin. Among the distinguished groups zeaxanthin, the pigment marker of *Synechococcus*, also occurs in prasinophytes and *Prochlorococcus*. According to the abundance of the pigment marker prasinoxanthin, Prasinophytes were a minor group compared with *Synechococcus* and *Prochlorococcus*. This result, together with the low concentration of zeaxanthin in prasinophytes (69), made the contribution of prasinophytes to the zeaxanthin pool practically negligible. Therefore, it was considered that only *Synechococcus* and *Prochlorococcus* contributed significantly to the zeaxanthin pool. Because only the former group contributes to MVChl *a*, it is necessary to distinguish between $Zeax_{Syn}$ and $Zeax_{Pro}$ to apply Chemtax. We partitioned Zeax as $Zeax_{FCM} = Zeax\ Syn_{cell}^{-1} \times [Syn]_{FCM} + Zeax\ Pro_{cell}^{-1} \times [Pro]_{FCM}$, where Zeax Syn_{cell}^{-1} and Zeax Pro_{cell}^{-1} are the Zeax content per cell of *Synechococcus* and *Prochlorococcus*, respectively, and $[Syn]_{FCM}$ and $[Pro]_{FCM}$ were the *Synechococcus* and *Prochlorococcus* cell concentrations obtained from FCM for the same sample. Initial values for Zeax Syn_{cell}^{-1} and Zeax Pro_{cell}^{-1} were estimated by minimizing the $\sum (Zeax_{HPLC} - Zeax_{FCM})^2$ using the function Solver of Microsoft Excel in default mode (time = 100 s, iterations = 100, precision = 0.000001, tolerance = 5, convergence = 0.0001, lineal estimation, progressive derivative, Newton's method). We used a single, common Excel cell for all Zeax Syn_{cell}^{-1} with a seed value of 1.8 fg Zeax Syn_{cell}^{-1} as per ref. 70. The same procedure was applied for all Zeax Pro_{cell} samples but with a seed value of 1 fg Zeax Pro_{cell}^{-1} from ref. 71. This procedure provides a single value of Zeax Syn_{cell}^{-1} and Zeax Pro_{cell}^{-1} for all of the samples. A further refinement consisted of applying Solver a second time allowing the change of all of the individual values of Zeax Syn_{cell}^{-1} and Zeax Pro_{cell}^{-1}. Prymnesiophytes have previously been categorized as falling into eight major pigment groups (72). Some have pigment characteristics of diatoms (type 1), others of diatoms with some additional minor pigments (types 2–5), and still others that are a mixture of more typical prymnesiophytes (types 6 and 7), as well as one that has characteristics of pelagophytes and prymnesiophytes. Here we used types 6 and 7 to represent prymnesiophytes, and type 8 did not exist in the matrix, but rather was divided to pelagophytes or to prymnesiophyte types 6 and 7. It should be noted that dinoflagellates can be abundant in the tropics (73), and some dinoflagellates contain the prymnesiophyte-indicative marker pigment 19'-hexanoyloxyfucoxanthin (67, 74) and therefore can contribute to HPLC overestimation of prymnesiophytes.

Chemtax was applied according to the procedures described in ref. 75, using version 1.95 of Chemtax (68). Random pigment to

Chl *a* ratios between 0.1 and 1 were used as seed values of 16 input matrices. Chemtax was run using the following parameters: ratio limits = 1,000, initial step size = 50, step ratio = 2, ε limit = 0.0001, cutoff step = 30,000, iterations limit = 1,000, elements varied = 10 (number of pigments), subiterations = 1, weighting = bound relative (50). The output of each run was used as input for the following run and this procedure repeated eight times. The median of each pigment ratio was incorporated to the final pigment ratio matrix. This matrix was then used to estimate the contribution of the different groups to MVChl *a* stock.

9. Prymnesiophyte Cell Counts. Prymnesiophytes were enumerated by FISH using a prymnesiophyte-specific probe or using a characteristics-based method based on their chloroplast arrangements, flagellar characteristics, and occasionally the presence of a haptonema (76). No significant difference (*t* test, $P = 0.428$) was detected between these two microscopy methods. Comparison of data from between 25° to 35° N in the Atlantic showed comparably abundances, with the prymnesiophyte characteristics-based average being 593 ± 108 mL^{-1} (SE, $n = 12$), whereas the FISH average was 500 ± 61 mL^{-1} (SE, $n = 26$) for different sample sets, and cruises.

FISH was performed on "WS" and "OC" cruise samples, using a prymnesiophyte-specific probe [PRYM02, 5′ GGA ATA CGA GTG CCC CTG AC 3′ (77)] and hybridized cells enumerated by epifluorescence microscopy. To prepare and store samples for hybridization, seawater (180 mL for OC413 CTD profiles and 100% raw seawater treatments in the dilution experiments; 405 mL of seawater for the 40% and 20% raw seawater; 90 mL for all "WS" 2005 cruises) was preserved with paraformaldehyde (1%, final concentration) for a minimum of 1 h at 4 °C in the dark. The seawater was filtered onto a 0.2-μm Anodisc (25 mm; Whatman), and the filters were dried with an ethanol series (50%, 80%, and 100% ethanol diluted in autoclaved 18.2 MΩ water for 3 min each) and stored at −80 °C before hybridization. FISH was performed on replicate filter pieces in conjunction with tyramide signal amplification using a modification of a previously reported method (4, 78). The PRYM02 probe had no mismatches with the prymnesiophyte 18S rRNA sequences from clone libraries (Fig. 2, main text, and Fig. S1*A*), with the following exceptions: OLI16029, one mismatch; OLI51033, two mismatches; OLI51059, two mismatches; OC413BATS_O071_75m, two mismatches; FS01AA-77_01Aug05_5m, one mismatch; *Chrysochromulina leadbeateri*, three mismatches; *Chrysoculter rhomboideus*, one mismatch. Note that several of the OLI sequences had gaps, or apparent nucleotide substitutions, in several highly conserved positions for other eukaryotes. Hybridization efficiency of PRYM02 was tested on a culture of a larger cultured prymnesiophyte species, *Isochrysis sp.* CCMP1244; out of the 1,492 cells detected using the DNA-specific dye DAPI, 1,480 cells (or 99.3% ± 3.3% of the cells) were positively hybridized. After hybridization, FISH filters were counterstained with DAPI. This was performed by counterstaining with 2.5 μg mL^{-1} for 5 min, rinsing for 5 min at room temperature in autoclaved 18.2 MΩ water, briefly dipping in 80% ethanol, and then air drying for approximately 10 min, and finally applying 7 μL of mounting solution [1:5 antifading solution AF1 (Citifluor) and Vectashield mounting medium (Vector Laboratories)] for "OC" samples. For "WS" samples, filters were air dried for approximately 15 min and Vectashield mounting medium, containing DAPI, was applied to each piece. In either case, the coverslip was sealed to the slide with nail polish and filters counted within 24 h.

Thirty ("OC" samples) and 50 ("WS" samples) 100 μm × 100 μm fields were enumerated per filter piece using a ×100 oil-immersion objective on an Olympus BX61 epifluorescence microscope. Probe signal was detected in the FITC channel and associated DAPI fluorescence (showing the cell nucleus) verified during enumeration. The volume filtered and area of the filter were considered and cell concentrations calculated accordingly. Cells were placed into three size categories (using the largest cell dimension): <3 μm, 3–10 μm, and >10 μm, by measurement against grid markings (1-μm increments). More specific size measurements were performed as below. The number of cells in the >10-μm size fraction was statistically unreliable (0.8% ± 1.8%) and therefore not considered further.

Several controls were performed alongside PRYM02 hybridization of field samples. The bacterial antisense NON338 probe (5′ ACTCCTACGGGAGGCAGC 3′) (79) was used as a negative control for all hybridizations. In addition, filters of *Micromonas* RCC299 and *Isochrysis sp.* CCMP1244 cultures were used as negative and positive controls for the PRYM02 probe, respectively. These were hybridized alongside all field samples, including those from dilution experiments. A no-probe control was added for each environmental sample at least once, but not necessarily for each hybridization.

For all of the "N" and "S" cruises a known volume of surface water was added to a filter funnel fitted with a polycarbonate filter (Nucleopore, 25-mm diameter, 0.2-μm pore size) and a diffuser filter underneath, preserved with a small volume of 50% glutaraldehyde (1–2%, final concentration) and vacuumed onto the filter. Filters were mounted on glass slides in subdued light to preserve phytoplankton pigment fluorescence and counted aboard the ship on the day of collection using a Zeiss Axioskop equipped with epifluorescence and a ×100 oil-immersion objective. The excitation filter was a Zeiss 48.77.09, under which phycoerythrin fluoresces orange and chlorophyll fluoresces red. Glutaraldehyde-induced green fluorescence revealed cell membranes and, in combination with pigment fluorescence and the unique chloroplast and flagellar configurations of prymnesiophytes (76), pico-prymnesiophytes were counted and sized. Pico-prymnesiophytes were binned to four size categories (Table S7). A 10 × 10 grid of 4,624 μm^2 was used to count abundant picophytoplankton. For lower abundances of picophytoplankton the iris diaphragm was closed to give a 120-μm diameter field and a portion of this field counted until >500 picophytoplankton had been routinely counted.

A recent study exploring label uptake in eukaryotes from labeled *Prochlorococcus* prey (presumably uptake was direct) showed that sequences close to group 14 (Fig. 2, main text) were present at Station ALOHA, in the north Pacific Gyre (e.g., hotxp4g5) (80). Thus, a consideration regarding some uncultured prymnesiophytes lies in emerging evidence that some may be capable of consuming *Prochlorococcus* (80). In our study, the pico-prymnesiophytes evaluated contained chlorophyll and showed no evidence of captured prey. Some potential prey, like *Synechococcus*, would be difficult to overlook, owing to its intense phycobiliprotein fluorescence.

10. Picophytoplankton Cell Size and Biomass. Cell size of prymnesiophytes was determined as above using epifluorescence microscopy. For each pico-prymnesiophyte size category (Table S7), biovolume was calculated using the formula $V = 4/3\,\pi \times L/2 \times W/2 \times W/2$, where L, length = the longest visible cell dimension, and W, width = the shortest visible cell dimension. Because only two dimensions could be measured on the microscope, the third dimension for volume was assumed to be the shortest of the two dimensions measured (W), thus potentially biasing the data in a way that could underestimate pico-prymnesiophyte biovolume values. For a small portion of the data ("OC" and "WS" cruises), pico-prymnesiophytes were enumerated for three bins only: <3 μm, 3–10 μm, and >10 μm. Because the midsize category (3–10 μm) contained many cells in the smaller end of this range, we more precisely sized cells at two sites and two depths in the Sargasso Sea. L and W were precisely measured for PRYM02-hybridized cells using a calibrated sizing grid for the NSS station (15 m, $n = 60$, and 70 m, $n = 60$) and BATS station (15 m, $n = 60$, and 75 m, $n = 60$). To determine averages the data were placed into two size categories (those with L <3 μm and those with L between 3 and <5 μm). Ten

percent of the cells were ≥5 μm (largest dimension) and were not included for further analyses to avoid overestimation of pico-prymnesiophyte biomass. This resulted in an average cell length of 3.4 ± 0.5 μm (instead of 3.8 ± 1.1 μm, when including all cells >5 μm) and average width of 2.8 ± 0.6 μm (instead of 3.1 ± 1.0 μm, when including all cells >5 μm). The vast majority of pico-prymnesiophytes counted for "WS" cruises were composed of <3-μm cells (Fig. S6*A*). Pico-prymnesiophytes biovolumes for these samples ranged from 4.0 ± 1.0 μm^3 to 14.0 ± 3.6 μm^3. Biovolume for the four pico-prymnesiophyte size categories identified in all other cruises ranged from 4.2 to 11.5 μm^3 (Table S7).

Biomass of various size groups was then estimated using the product of abundance and mean cellular carbon content. The latter was taken as the product of cell biovolume and a single carbon conversion factor used for all groups, 237 fg C μm^{-3}, previously reported for *Prochlorococcus*, *Synechococcus*, and several "non-prymnesiophyte" picoeukaryote groups (81). For pico-prymnesiophytes, mean cellular carbon content was determined using this biovolume-based carbon conversion factor for each size category (Table S7). *Prochlorococcus* and *Synechococcus* cellular carbon conversion values were 39 fg C $cell^{-1}$ and 82 fg C $cell^{-1}$, respectively, as determined previously on discrete populations enumerated by FCM and analyzed by carbon, hydrogen, and nitrogen (CHN) (81). As noted above, counts from our eukaryotic FCM analysis window for the smallest eukaryotes showed a tight correlation with the sum of all nonprymnesiophyte red-fluorescing picoeukaryotes counted by microcopy (5). Given that nonprymnesiophyte picoeukaryotes tend to be smaller than the pico-prymnesiophytes (e.g., pico-prasinophytes; for instance, *Ostreococcus* is ≈1 μm diameter and *Micromonas* ≈1.4–1.6 μm), the biomass conversion factor 530 fg C $cell^{-1}$ was used for the FCM-enumerated nonprymnesiophyte picoeukaryotes, as determined previously for field populations with few prymnesiophytes in the eastern North Pacific (81). Unlike some studies that have used larger cellular conversion factors for eukaryotes that likely overestimate their contributions, the cellular carbon conversion factors used here for the four picophytoplankton groups (*Prochlorococcus*, *Synechococcus*, nonprymnesiophyte picoeukaryotes, and pico-prymnesiophytes) were derived using the same carbon per unit volume, from ref. 81. This value is similar to that of Booth et al. (82), 220 fg C μm^{-3}. However, Grob et al. (83) reported in situ cellular carbon of *Prochlorococcus* being 29 ± 11 fg C $cell^{-1}$ and for combined picophytoeukaryotes (all lineages) being 730 ± 226 fg C $cell^{-1}$, on the basis of a combination of culture-based work, environmental Coulter Counter data, and CHN measurements. The latter value likely reflects an average between our non-prymnesiophyte picophytoeukaryote and pico-prymnesiophyte cellular carbon values. Our *Prochlorococcus* cellular carbon value is higher than estimated by Grob et al., but similar to that of ref. 84. Total picophytoplankton carbon was taken as the sum of the population biomasses for *Prochlorococcus*, *Synechococcus*, and nonprymnesiophyte picoeukaryotes and the various contributions of the different cells within the different pico-prymnesiophyte size ranges.

Picophytoplankton biomass at the NSS station was dominated by *Prochlorococcus* at most depths above 80 m (between 1.8 μg C L^{-1} and 3.0 μg C L^{-1}). At this site, average pico-prymnesiophyte biomass was higher at the surface than at the DCM (1.7 ± 0.6 μg C L^{-1} and 0.8 ± 0.4 μg C L^{-1}, respectively). Nonprymnesiophyte picoeukaryotes and *Synechococcus* also contributed significantly to the picophytoplankton biomass at the same depths (between 0.7 μg C L^{-1} and 1.5 μg C L^{-1} for the former and 1.4 μg C L^{-1} and 2.6 μg C L^{-1} for the latter). Nonprymnesiophyte eukaryote biomass peaked at the DCM, reaching a maximum of 2.6 μg C L^{-1} for CTD081. At BATS, surface picophytoplankton biomass was lower than in the NSS. Generally (date dependent), maximum biomass was reached at the DCM (85 m, 3.0 μg C L^{-1}) or just above (65 m, 4.8 μg C L^{-1}) at BATS. Overall, pico-prymnesiophyte cell concentrations range from 177 ± 116 cells mL^{-1} (CTD056, DCM) to 872 ± 45 cells mL^{-1} (CTD056, surface) in the Sargasso Sea. Differences between mean cell concentrations were not significant between the two sites (P = 1.0). At BATS, average surface abundance of pico-prymnesiophytes were almost identical (536 ± 231 cells mL^{-1}) to those at DCM (539 ± 224 cells mL^{-1}). Nevertheless, these trends were not observed for individual CTD casts; numbers of cells were significantly higher at the surface than at the DCM for CTD004 ($P < 0.05$), and the opposite was seen for CTD029 ($P < 0.01$; Fig S6*A*). At the NSS station, pico-prymnesiophytes were more abundant at the surface (768 ± 129 cells mL^{-1}) than at the DCM. In the Florida Straits, pico-prymnesiophytes ranged up to 1.2 × 10^3 cells mL^{-1} (December, Station 01) at the surface and 6.0 × 10^2 cells mL^{-1} (September, Station 14) in the DCM but at times were below 100 cells mL^{-1} (e.g., May, Station 04 surface, July/August, Station 14 DCM). Pico-prymnesiophyte biomass ranged from 0.10 to 1.9 μg C L^{-1} in the surface and 0.10 ± 0.63 μg C L^{-1} at the DCM over the six Florida Straits transects in 2005.

11. At-Sea Growth Rate Experiments. Dilution experiments were performed according to the methods of ref. 1 with modifications similar to those in ref. 2. These experiments allowed us to estimate the growth and grazing mortality rates of natural phytoplankton populations (1, 85). Briefly, a series of bottles containing different ratios of raw seawater to filtered seawater were incubated for 24 h. For each experiment, triplicate bottles were prepared for the following dilution factors: 1.0, 0.6, 0.4, 0.3, 0.2, and 0.1 (the factors represent the fraction of raw seawater diluted with 0.2-μm-filtered seawater for a final volume of 1 L). Bottles were incubated in on-deck water baths for 24 h (from sunrise to sunrise). In situ light and spectral conditions were simulated using a combination of blue and/or neutral-density gel filters (Lee Filters). Water temperature was maintained using a flow-although system that constantly pumped surface seawater through the on-deck water baths. Two experiments were performed at BATS: experiment Exp. 1 (75 m) and Exp. 2 (15 m); and two at the NSS station: Exp. 3 (15 m) and Exp. 4 (70 m). Surface (15 m) experiments were conducted at in situ temperatures, whereas those from deeper [e.g., the 70 m experiment (DCM at 93 m)] were incubated at <2 °C higher than at 70 m. Pico-prymnesiophyte abundances were 324 ± 136 cells mL^{-1} and 238 ± 94 cells mL^{-1} in Exp. 1 and Exp. 2, respectively. These concentrations were slightly lower than values at the surface and DCM of the day previous to each experiment, in part because the DCM was deeper in the water column (i.e., Exp. 1 was conducted in a region of the water column with lower abundance than the DCM). For Exp. 3 and Exp. 4, pico-prymnesiophyte cell concentrations were 448 ± 144 cells mL^{-1} and 651 ± 282 cells mL^{-1}, respectively. Environmental groups 13, 15, and 16 were detected in these samples.

FCM samples were collected from each of the triplicate bottles at t = 0 h and t = 24 h, as were FISH samples. For FISH, 180 mL was filtered, with replication, for each of the replicate 1.0 dilution treatment bottles, whereas 405 mL was filtered for the more dilute bottles. The net (apparent) growth rate in each bottle was calculated as the natural logarithm of the ratio of the final cell concentration to the initial concentration. Linear regression of the net growth rates against dilution factors was used to estimate the grazing mortality rates (g, slope) and growth rates (μ, y-intercept). Growth rates and grazing mortality rates of *Prochlorococcus*, *Synechococcus*, and picophytoeukaryotes were estimated using FCM samples collected from each bottle. Prymnesiophytes were enumerated and sized using FISH on samples collected from 1, 0.4, and 0.2 dilution factors. An ANOVA was used to test the significance of the regression, and only rates from statistically significant data (P = 0.06, r^2 = 0.73, P = 0.06, r^2 = 0.87) are reported. Low abundance, which was magnified in dilution treatments, made it difficult to enumerate sufficient numbers of cells in diluted bottles, particularly at BATS. The ANOVA results led to BATS experimental data being discarded

for pico-prymnesiophytes and small eukaryotes. At the NSS station, pico-prymnesiophyte growth and grazing mortality rates were higher at the surface (1.12 d^{-1} and 1.41 d^{-1} for μ and g, respectively, $r^2 = 0.87$, $P = 0.06$) than deeper in the water column (70 m; 0.29 d^{-1} and 0.70 d^{-1} for μ and g, respectively, $r^2 = 0.73$, $P = 0.06$). For small eukaryotes, the opposite trend was observed, with higher growth and grazing mortality rates at depth (0.51 d^{-1} and 0.74 d^{-1} for μ and g, respectively, $r^2 = 078$, $P < 0.0001$) than at the surface (0.22 d^{-1} and 0.29 d^{-1} for μ and g, respectively, $r^2 = 0.41$, $P = 0.06$). Although no previous literature values exist for pico-prymnesiophyte–specific growth rates, unadjusted total (all size fractions) prymnesiophyte HPLC ratio-based growth estimates in the Equatorial Pacific, a region dominated by picophytoplankton (73), are similar to those herein.

Phytoplankton primary production and grazing losses were calculated for the NSS station using the dilution experiment results. Abundance and biomass of picophytoplankton groups along with growth rates (μ, d^{-1}) and grazing mortality rates (g, d^{-1}) of *Prochlorococcus*, *Synechococcus*, nonprymnesiophyte picoeukaryotes, and pico-prymnesiophytes were used to calculate primary production (i.e., PP, μg C L^{-1} d^{-1}) of each group using the following equations derived from refs. 86 and 87. At any instant, t :

$$PP_t = \mu * B_t$$

and where Bt (μg C L^{-1}) is the phytoplankton biomass at time t calculated with the following equation:

$$B_t = B_0 * \left[e^{(\mu - g)*t}\right]$$

so that PP integrated over the length of the experiment T (here 1 d), can be calculated with the following equations:

$$PP = \mu * B_0 * \left[e^{(\mu - g)*T} - 1\right] \Big/ \left[(\mu - g) * T\right]$$

where B_0 (μg C L^{-1}) is the initial biomass.

We also calculated biomass production without grazing mortality, which would represent the maximum potential PP (PP_{max}) of each group using the following equation:

$$PP_{max} = B_0 * \left[e^{\mu * T} - 1\right] / T.$$

Pico-prymnesiophyte primary production at the NSS site was 1.1 μg C L^{-1} d^{-1} at the surface, or 2.4 μg C L^{-1} d^{-1} if production is considered without the effect of grazing, almost 4-fold more than for other picoeukaryotes (0.27 μg C L^{-1} d^{-1}, or 0.30 μg C L^{-1} d^{-1} if production is considered without the effect of grazing). These roles were reversed at 70 m, where pico-prymnesiophyte primary productions was 0.3 μg C L^{-1} d^{-1} (or 0.5 μg C L^{-1} d^{-1} if production is considered without the effect of grazing) and nonprymnesiophyte picoeukaryotes were estimated to produce 1.8 μg C L^{-1} d^{-1} (2.6 μg C L^{-1} d^{-1} if production is considered without the effect of grazing).

12. Author Contributions. Cruise sampling was designed by M.L.C., R.M.W., C.G., K.R.B., F.P.C. and A.Z.W. Sampling was performed by the same, as well as J.A.H, and other cruise participants (not necessarily co-authors). Flow sorting was performed by R.M.W and A.Z.W. The overall 18S rRNA gene tree was constructed by M.L.C. and A.Z.W. with significant input from A.M. The MDA-flow sort approach was conceived by A.Z.W., T.I., R.M.W., and R.S.L. T.I. performed MDA. A.Z.W. and T.I. performed preliminary sort characterizations. T.I. and T.W. constructed/sequenced MDA SSU clone libraries. S.G.T. and T.W. led metagenomic sequencing and assembly. A.M. performed final MDA-SSU phylogenetics. M.T. and E.C. performed gene modeling. Final metagenome analyses including phylogenomics were performed by A.E.A., A.M., J.P.M., J.H.L., C.L.D., and A.Z.W. K.R.B. performed characteristics-based microscopy. M.L.C. and A.Z.W. ran and analyzed FCM for FS and SS samples, while F.P.C. was responsible for other FCM analyses; M.L. ran and analyzed HPLC samples. M.L.C. and J.A.H. performed FISH counts. B.J.B. designed and implemented dilution experiments. M.L.C. analyzed dilutions experiments by FCM and FISH. M.M. and F.P.C. developed global biomass analyses with input from M.L.C. and A.Z.W. The manuscript was written by M.L.C., A.E.A. and A.Z.W., with significant input from A.M., J.P.M., S.G.T., T.W., as well as input from J.-H. L., C.L.D., F.P.C. and M.L. The project was conceived by A.Z.W.

1. Landry MR, Hassett RP (1982) Estimating the grazing impact of marine microzooplankton. *Mar Biol* 67:283–288.
2. Worden AZ, Binder BJ (2003) Application of dilution experiments for measuring growth and mortality rates among *Prochlorococcus* and *Synechococcus* populations in oligotrophic environments. *Aquat Microb Ecol* 30:159–174.
3. Not F, Gausling R, Azam F, Heidelberg JF, Worden AZ (2007) Vertical distribution of picoeukaryotic diversity in the Sargasso Sea. *Environ Microbiol* 9:1233–1252.
4. Cuvelier ML, et al. (2008) Widespread distribution of a unique marine protistan lineage. *Environ Microbiol* 10:1621–1634.
5. Buck KR, Chavez FP, Campbell L (1996) Basin-wide distributions of living carbon components and the inverted trophic pyramid of the central gyre of the North Atlantic Ocean, summer 1993. *Aquat Microb Ecol* 10:283–298.
6. Chavez FP, Buck KR, Service SK, Newton J, Barber RT (1996) Phytoplankton variability in the eastern and central tropical Pacific. *Deep-Sea Res Pt II* 43:835–870.
7. Zinser ER, et al. (2007) Influence of light and temperature on *Prochlorococcus* ecotype distributions in the Atlantic Ocean. *Limnol Oceanogr* 52:2205–2220.
8. Sullivan MB, et al. Genomic analysis of oceanic cyanobacterial myoviruses compared to T4-like myroviruses from diverse hosts and environments. *Environ Microbiol*, in press.
9. Olson RL, Zettler ER, DuRand MD (1993) *Handbook of Methods in Aquatic Ecology*, eds Kemp PF, Sherr BF, Sherr EB, Cole JJ (Lewis Publishers, Boca Raton, FL), pp 175–186.
10. Monger BC, Landry MR (1993) Flow cytometric analysis of marine bacteria with Hoechst 33342. *Appl Environ Microbiol* 59:905–911.
11. Campbell L, Nolla HA, Vaulot D (1994) The importance of *Prochlorococcus* to community structure in the central North Pacific Ocean. *Limnol Oceanogr* 39:954–961.
12. Campbell L, Vaulot D (1993) Photosynthetic picoplankton community structure in the subtropical North Pacific Ocean near Hawaii (station ALOHA). *Deep-Sea Res* 40:2043–2060.
13. Olson RJ, Chisholm SW, Zettler ER, Armbrust EV (1990) Spatial and temporal distributions of prochlorophyte picoplankton in the North Atlantic Ocean. *Deep-Sea Res* 37:1033–1051.
14. Dean FB, et al. (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA* 99:5261–5266.
15. Moon-van der Staay S, et al. (2000) Abundance and diversity of prymnesiophytes in the picoplankton community from the equatorial Pacific Ocean inferred from 18S rDNA sequences. *Limnol Oceanogr* 45:98–109.
16. Moon-van der Staay SY, De Wachter R, Vaulot D (2001) Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature* 409:607–610.
17. Worden AZ (2006) Picoeukaryote diversity in coastal waters of the Pacific Ocean. *Aquat Microb Ecol* 43:165–175.
18. Chenna R, et al. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* 31:3497–3500.
19. Felsenstien J (2005) PHYLIP (Phylogeny Inference Package) version 3.6 (Department of Genome Sciences, University of Washington, Seattle, WA).
20. Not F, et al. (2008) Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. *Deep-Sea Res Pt I* 55:1456–1473.
21. Alexander E, et al. (2009) Microbial eukaryotes in the hypersaline anoxic L'Atalante deep-sea basin. *Environ Microbiol* 11:360–381.
22. Cheung M, Chu K, Li C, Kwan H, Wong C (2008) Genetic diversity of picoeukaryotes in a semi-enclosed harbour in the subtropical western Pacific Ocean. *Aquat Microb Ecol* 53:295–305.
23. Massana R, Balagué V, Guillou L, Pedrós-Alió C (2004) Picoeukaryotic diversity in an oligotrophic coastal site studied by molecular and culturing approaches. *FEMS Microbiol Ecol* 50:231–243.
24. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
25. Posada D, Crandall KA (1998) MODELTEST: Testing the model of DNA substitution. *Bioinformatics* 14:817–818.
26. 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.
27. Edvardsen B, et al. (2000) Phylogenetic reconstructions of the Haptophyta inferred from 18S ribosomal DNA sequences and available morphological data. *Phycologia* 39:19–35.
28. Edvardsen B, Medlin L (2007) *Molecular Systematics of Haptophyta. Unravelling the Algae—The Past, Present and Future of Algal Molecular Systematics, The Systematics*

Association Special Volume Series, eds Brodie J, Lewis J (Taylor and Francis, Oxford), Vol Vol 75, pp 183–196.
29. Saez A, et al. (2004) *Coccolithophores—From Molecular Processes to Global Impact*, eds Thierstein H, Young J (Springer, Berlin), pp 251–269.
30. Parke M, Green JC (1976) Haptophyta. Check-list of British Marine algae, 3rd revision. *J Mar Biol Assoc* 56:527–594.
31. Chretiennot-Dinet M, Sournia M, Ricard M, Billard C (1993) A classification of the marine phytoplankton of the world from class to genus. *Phycologia* 32:159–179.
32. Green JC, Jordan RW (1994) *Systematic History and Taxonomy. The Haptophyte Algae, Systematics Association Special*, eds Green JC, Leadbeater BSC (Oxford Univ Press, Oxford), Vol 51, pp 1–21.
33. Jordan RW, Cros L, Young JR (2004) A revised classification scheme for living haptophytes. *Micropaleo* 50 Suppl 155–79.
34. 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.
35. Margulies M, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380.
36. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
37. Jardillier L, Zubkov MV, Pearman J, Scanlan DJ (2010) Significant CO_2 fixation by small prymnesiophytes in the subtropical and tropical northeast Atlantic Ocean. *ISME J*, doi: 10.1038/ismej.2010.36.
38. Haas BJ, et al. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* 31:5654–5666.
39. Birney E, Clamp M, Durbin R (2004) GeneWise and Genomewise. *Genome Res* 14: 988–995.
40. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955–964.
41. Worden AZ, et al. (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* 324:268–272.
42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410.
43. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.
44. Howe K, Bateman A, Durbin R (2002) QuickTree: Building huge Neighbour-Joining trees of protein sequences. *Bioinformatics* 18:1546–1547.
45. Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 17:149–163.
46. Finn RD, et al. (2008) The Pfam protein families database. *Nucleic Acids Res* 36 (Database issue):D281–D288.
47. Zdobnov EM, Apweiler R (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847–848.
48. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J Mol Biol* 300:1005–1016.
49. Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* 14:755–763.
50. Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res* 11:555–565.
51. Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* 12:1048–1059.
52. Sánchez Puerta MV, Bachvaroff TR, Delwiche CF (2005) The complete plastid genome sequence of the haptophyte *Emiliania huxleyi*: A comparison to other plastid genomes. *DNA Res* 12:151–156.
53. Conant GC, Wolfe KH (2008) GenomeVx: Simple web-based creation of editable circular chromosome maps. *Bioinformatics* 24:861–862.
54. Le Corguillé G, et al. (2009) Plastid genomes of two brown algae, *Ectocarpus siliculosus* and *Fucus vesiculosus*: Further insights on the evolution of red-algal derived plastids. *BMC Evol Biol* 9:253.
55. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302:205–217.
56. Jones DT, Taylor WR, Thornton JM (1994) A mutation data matrix for transmembrane proteins. *FEBS Lett* 339:269–275.
57. Kurtz S, et al. (2004) Versatile and open software for comparing large genomes. *Genome Biol* 5:R12.
58. Oudot-Le Secq MP, et al. (2007) Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: Comparison with other plastid genomes of the red lineage. *Mol Genet Genomics* 277:427–439.
59. Ashburner M, et al.; The Gene Ontology Consortium (2000) Gene ontology: Tool for the unification of biology. *Nat Genet* 25:25–29.
60. Claudel-Renard C, Chevalet C, Faraut T, Kahn D (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* 31:6633–6639.
61. Pinney JW, Shirley MW, McConkey GA, Westhead DR (2005) metaSHARK: Software for automated metabolic network prediction from DNA sequence and its application to the genomes of *Plasmodium falciparum* and *Eimeria tenella*. *Nucleic Acids Res* 33: 1399–1409.
62. Dillon SC, Zhang X, Trievel RC, Cheng XD (2005) The SET-domain protein superfamily: Protein lysine methyltransferases. *Genome Biol* 6:227.
63. Veerappan CS, Avramova Z, Moriyama EN (2008) Evolution of SET-domain protein families in the unicellular and multicellular Ascomycota fungi. *BMC Evol Biol* 8:190.
64. Stamatakis A (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
65. Keane TM, Naughton TJ, McInerney JO (2007) MultiPhyl: A high-throughput phylogenomics webserver using distributed computing. *Nucleic Acids Res* 35 (Web Server issue):W33–W37.
66. Latasa M, et al. (2001) Losses of chlorophylls and carotenoids in aqueous acetone and methanol extracts prepared for RPHPLC analysis of pigments. *Chromatographia* 53: 385–391.
67. Mackey MD, Mackey DJ, Higgins HW, Wright SW (1996) CHEMTAX—a program for estimating class abundances from chemical markers: Application to HPLC measurements of phytoplankton. *Mar Ecol Prog Ser* 144:265–283.
68. Wright SW, et al. (2009) Composition and significance of picophytoplankton in Antarctic waters. *Polar Biol* 32:797–808.
69. Latasa M, Scharek R, Le Gall F, Guillou L (2004) Pigment suites and taxonomic groups in Prasinophyceae. *J Phycol* 40:1149–1155.
70. Kana TM, Glibert PM (1987) Effect of irradiances up to 2000 $\mu E\ m^{-2}\ s^{-1}$ on marine *Synechococcus* WH7803. *Deep-Sea Res* 34:479–495.
71. Cailliau C, Claustre H, Vidussi F, Marie D, Vaulot D (1996) Carbon biomass, and gross growth rates as estimated from ^{14}C pigment labelling, during photoacclimation in *Prochlorococcus* CCMP 1378. *Mar Ecol Prog Ser* 145:209–221.
72. Zapata M, et al. (2004) Photosynthetic pigments in 37 species (65 strains) of Haptophyta: Implications for oceanography and chemotaxonomy. *Mar Ecol Prog Ser* 270:83–102.
73. Landry MR, et al. (2003) Phytoplankton growth and microzooplankton grazing in high-nutrient, low-chlorophyll waters of the equatorial Pacific: Community and taxon-specific rate assessments from pigment and flow cytometric analyses. *J Geophys Res-Oceans* 108:8142–8155.
74. Carreto JI, Seguel M, Montoya NG, Clement A, Carignan MO (2001) Pigment of the ichthyotoxic dinoflagellate *Gymnodinium sp.* from a massive bloom in southern Chile. *J Plankton Res* 23:1171–1175.
75. Latasa M (2007) Improving estimations of phytoplankton class abundances using Chemtax. *Mar Ecol Prog Ser* 329:13–21.
76. Andersen RA (2004) Biology and systematics of heterokont and haptophyte algae. *Am J Bot* 91:1508–1522.
77. Simon N, et al. (2000) Oligonucleotide probes for the identification of three algal groups by dot blot and fluorescent whole-cell hybridization. *J Eukaryot Microbiol* 47: 76–84.
78. Not F, Simon N, Biegala IC, Vaulot D (2002) Application of fluorescent *in situ* hybridization coupled with tyramide signal amplification (FISH-TSA) to assess eukaryotic picoplankton composition. *Aquat Microb Ecol* 28:157–166.
79. Worden AZ, Chisholm SW, Binder BJ (2000) In situ hybridization of *Prochlorococcus* and *Synechococcus* (marine cyanobacteria) spp. with rRNA-targeted peptide nucleic acid probes. *Appl Environ Microbiol* 66:284–289.
80. Frias-Lopez J, Thompson A, Waldbauer J, Chisholm SW (2009) Use of stable isotope-labelled cells to identify active grazers of picocyanobacteria in ocean surface waters. *Environ Microbiol* 11:512–525.
81. Worden AZ, Nolan JK, Palenik B (2004) Assessing the dynamics and ecology of marine picophytoplankton: The importance of the eukaryotic component. *Limnol Oceanogr* 49:168–179.
82. Booth BC, Lewin J, Lorenzen CJ (1988) Spring and summer growth rates of subarctic Pacific phytoplankton assemblages determined from carbon uptake and cell volumes estimated using epifluorescence microscopy. *Mar Biol* 98:287–298.
83. Grob C, et al. (2007) Picoplankton abundance and biomass across the eastern South Pacific Ocean along latitude 32.5 degrees S. *Mar Ecol Prog Ser* 332:53–62.
84. Bertilsson S, Berglund O, Karl DM, Chisholm SW (2003) Elemental composition of marine *Prochlorococcus* and *Synechococcus*: Implications for the ecological stoichiometry of the sea. *Limnol Oceanogr* 48:1721–1731.
85. Landry MR, Kirshtein J, Constantou J (1995) A refined dilution technique for measuring the community grazing impact of microzooplankton, with experimental tests in the central equatorial Pacific. *Mar Ecol Prog Ser* 120:53–63.
86. Landry MR, Calbet A (2004) Microzooplankton production in the oceans. *ICES J Mar Sci* 61:501–507.
87. Landry MR, et al. (2000) Biological response to iron fertilization in the eastern equatorial Pacific (IronEx II). III. Dyanamics of phytoplankton growth and microzooplankton grazing. *Mar Ecol Prog Ser* 201:57–72.

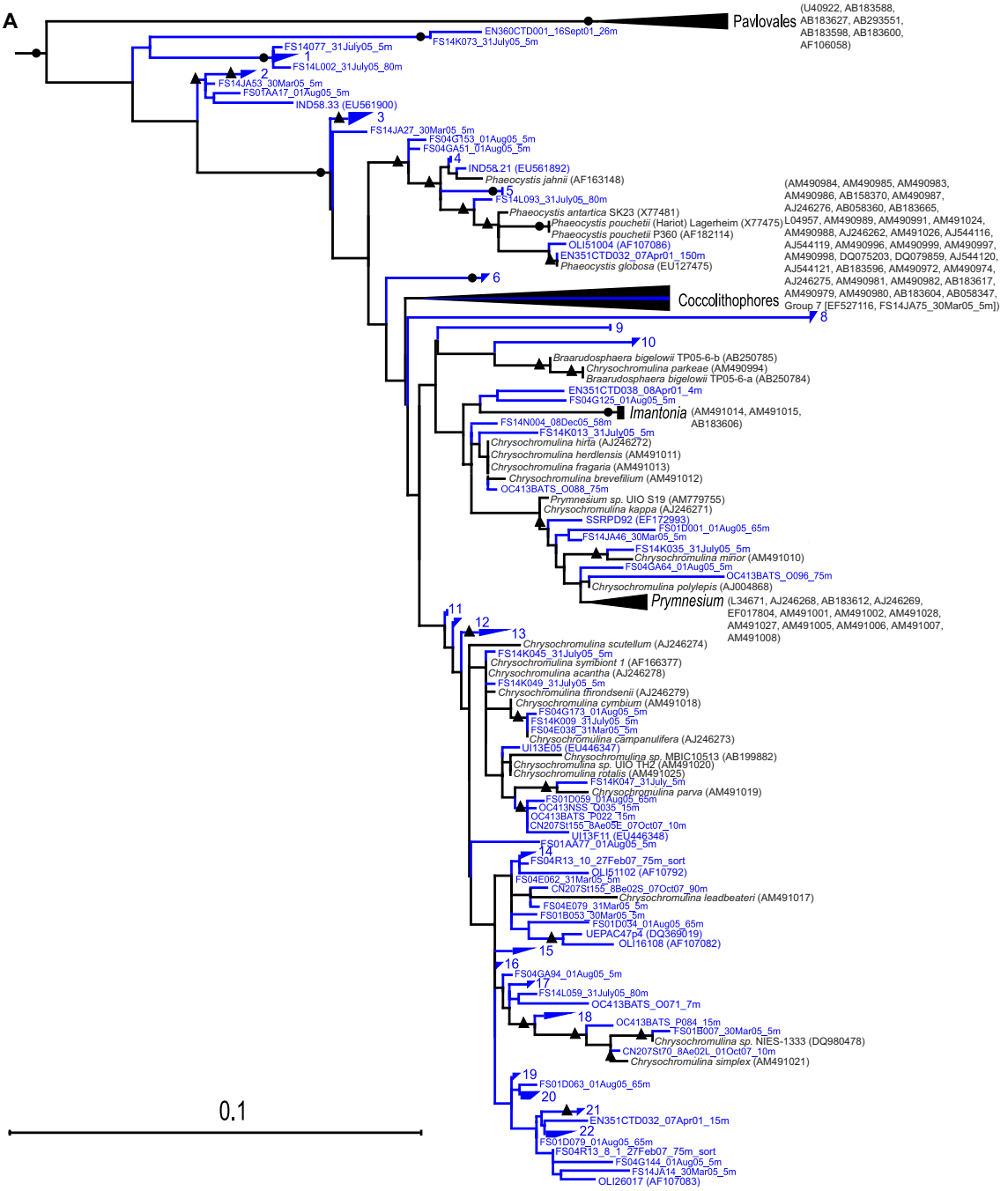

Fig. S1. (Continued)

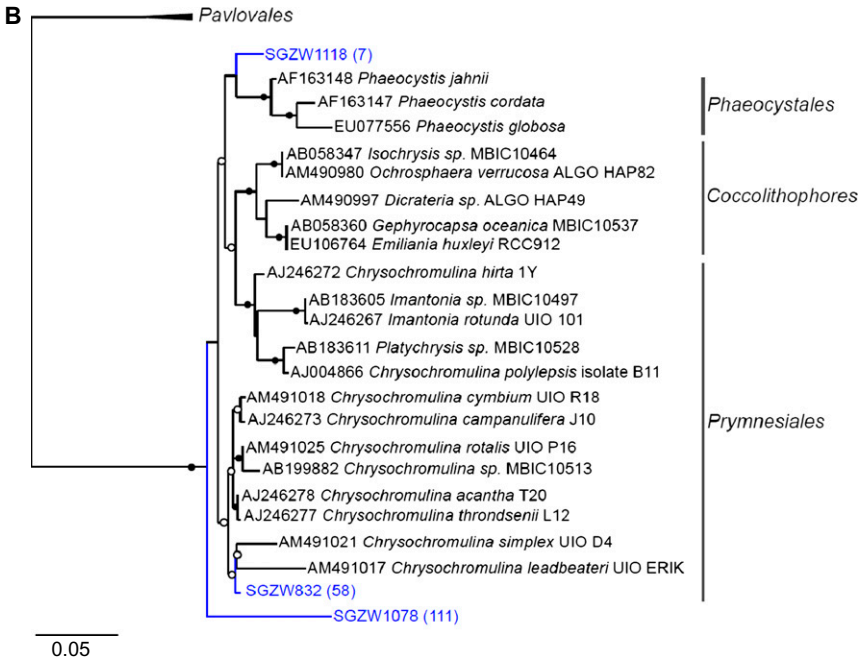

Fig. S1. Maximum likelihood reconstructions of 18S rRNA gene sequences. (*A*) Reproduction of Fig. 2 (main text) showing all environmental clone names, except for 99% identity groups 1–22, for which clone IDs are given in Table S2. Note that all sequences within each 99% identity group (Table S2) were used in the alignment (although the alignment only included a single representative of redundant sequences from within each clone library or previously published data from an individual site), and clades were collapsed for visualization purposes after the tree was constructed. Identity levels were determined using the original sequence (not from a masked alignment). GenBank accession numbers for previously published sequences are also provided. More than half (58%) of our environmental sequences fell within the previously identified clade B (*SI Materials and Methods, Section 4*), especially, clade B2, as highlighted by the number of environmental groups within this clade. Clade B contains the nonmineralized order Prymnesiales, including *Chrysochromulina*, *Prymnesium* and *Imantonia*. Some sequences were not definitively placed in previously existing clades (e.g., group 8), and some clades were represented by only a few sequences from our libraries. For example, one sequence, from the NSS (150 m) had 100% identity to *P. globosa* and fell within clade A, which contains all *Phaeocystis*. Five environmental sequences fell within the coccolithophores (clade C), but none were closely affiliated with sequenced taxa, several falling within group 7 (Table S2) and others being singletons (FS14JA16_30Mar05_5m, FS14JA75_30Mar05_5m, and FS14M081_08Dec05_58m). Isochrysidales were also within the collapsed coccolithophore group. (*B*) Pico-prymnesiophyte 18S rRNA gene sequences in the environmental flow sort advanced for metagenomic sequencing. The clone library was built from the MDA-flow sort product. Before phylogenetic analysis sequences were clustered at the 98.9% identity level. 18S rDNA sequences (blue) from the MDA-flow sort clone library and GenBank reference sequences (black) taken only from cultured strains that have both 16S and 18S rRNA gene sequences available (to allow comparison with 16S rRNA gene-bearing chloroplast scaffolds), with the exception of coccolithophores, for which the criteria were at the genus level for two genera owing to limited sequence availability. Group 8 is represented by SGZW1078 and group 3 by SGZW1118. SGZW832 represents clade B2 sequences (e.g., groups 14–22), at the tip of the tree with high identity to one another and clustering under the 98.9% criteria (*SI Materials and Methods, Section 5*). The number of sequences within each cluster is provided in parentheses beside sequence names. Bootstrap support was computed using maximum likelihood (100 replicates) and neighbor-joining (10,000 replicates) methods. Nodes retaining bootstrap support are indicated for those above 70% by both methods (black circles) and above 70% with only one method (white circles).

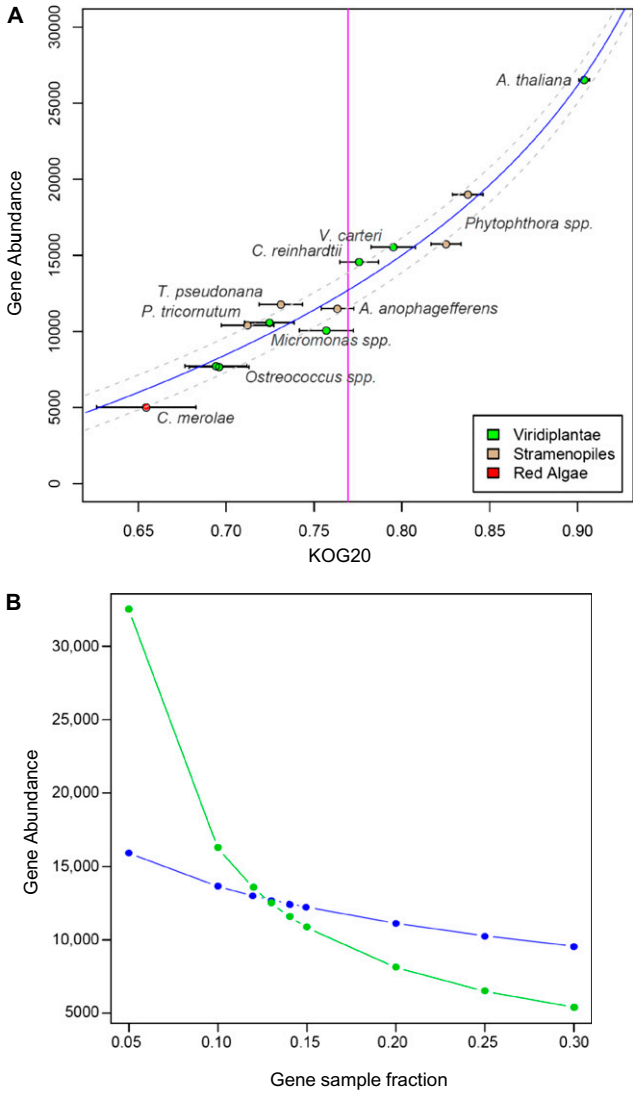

Fig. S2. Gene abundance and genome size. (*A*) Total gene content predicted for a representative pico-prymnesiophyte in the MDA-flow sort based on the targeted metagenome placed relative to gene subsamples of the metagenome KOG20 data (magenta line) and selected reference species. Functional diversity is measured as the proportion of BLASTP hits to KOGs ($e \leq 1.0e^{-9}$) in the largest 20% of KOGs, labeled KOG20 (*SI Materials and Methods, Section 7*). Smaller KOG20 values represent a larger functional spread of genes relative to total KOGs and thus greater functional diversity. The positive relationship between KOG20 and gene abundance allows for a prediction of gene total based on functional composition that is generally independent of the number of genes recovered for the pico-prymnesiophyte. The genome size prediction is based on fitting the model $y = a + b[-\log(1 - x)]$ using ordinary least squares, where y = genome size, and x = kog20. The parameters are estimated to be a = −10,928, b = 16,112, (blue line), with 1 SD in residual error (gray dashed lines). The model was chosen to simplify the parameter space as an exponential relationship between KOG20 and gene abundance and to enforce a theoretical upper bound at x = 1. An extensive range of reference eukaryotic genome sizes was used for fitting the model. SDs of 100 subsamples are shown as horizontal error bars for each reference species. (*B*) Model used to determine the appropriate level of KOG20 subsampling for reference genomes in *A*. The expected total number of pico-prymnesiophyte genes (green), treating the metagenomic assembly as representing a partial pico-prymnesiophyte genome, is plotted such that 1,624 genes represent a range of complete genome coverage estimates. The KOG20 predicted pico-prymnesiophyte genome size (blue) based on subsampling reference genomes to specific sample fractions. These values converge near 13% (0.13 gene sample fraction), where the number of genes in the pico-prymnesiophyte metagenomic assembly estimated by KOG20 equals the expected number based on the same genome sample size. Although the KOG20 estimate of genome size varies with the sample fraction, it is more stable than an estimate based on extrapolating from sample fraction directly, as seen by the wide range of genome size estimates of the expected compared with the KOG20 lines.

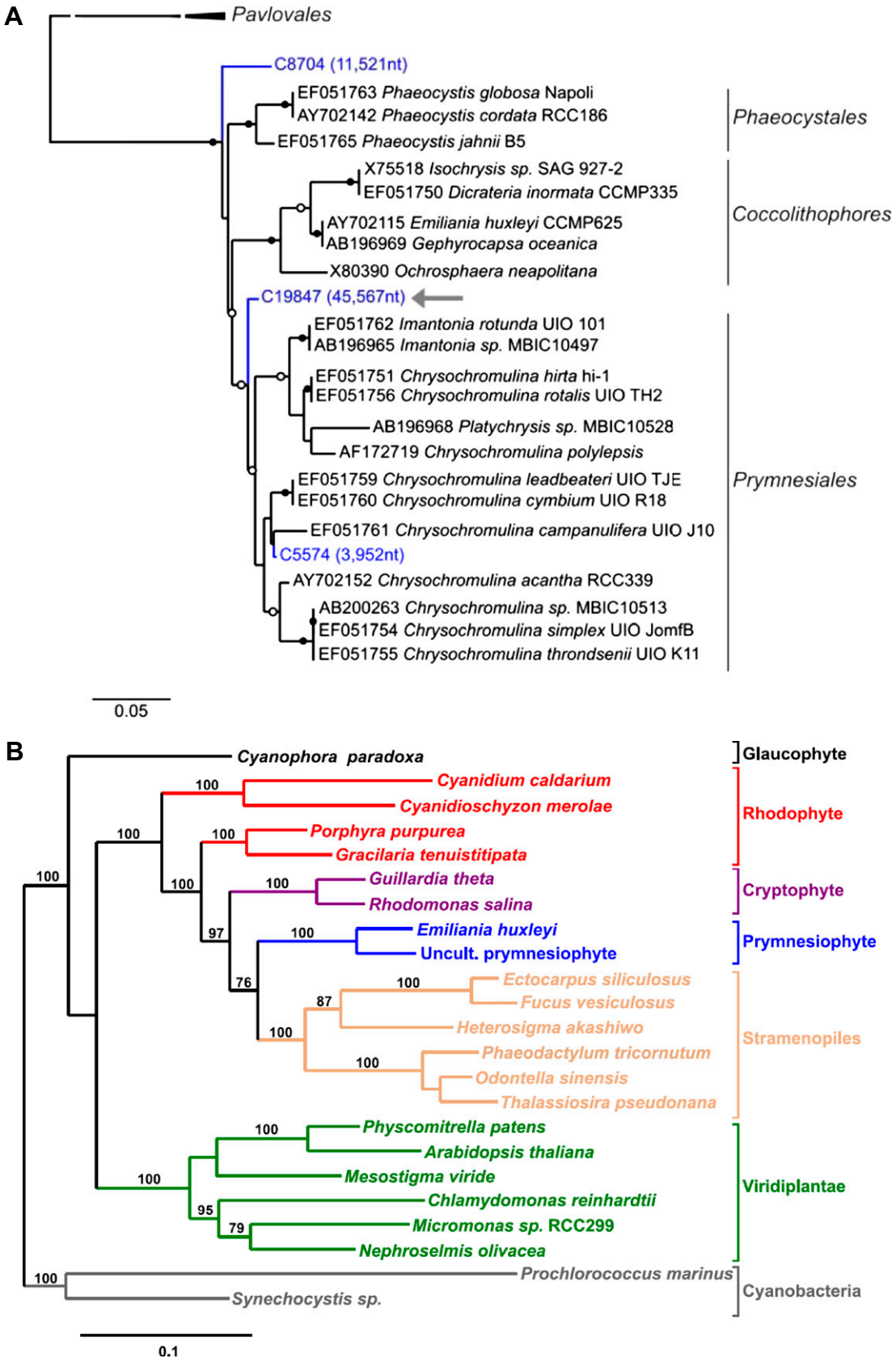

Fig. S3. (Continued)

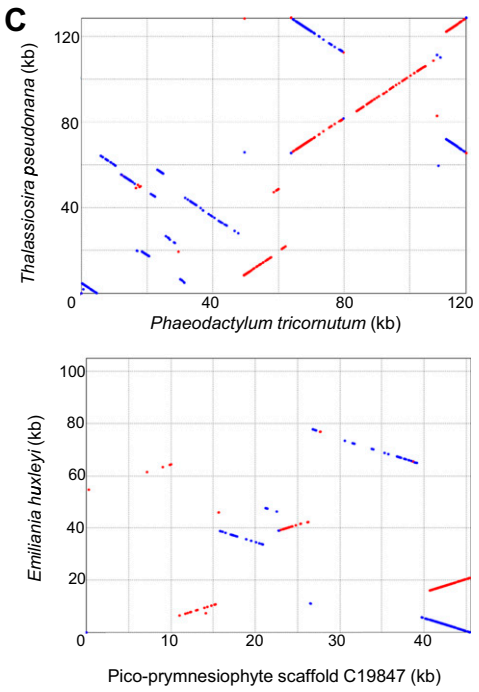

Fig. S3. (Continued)

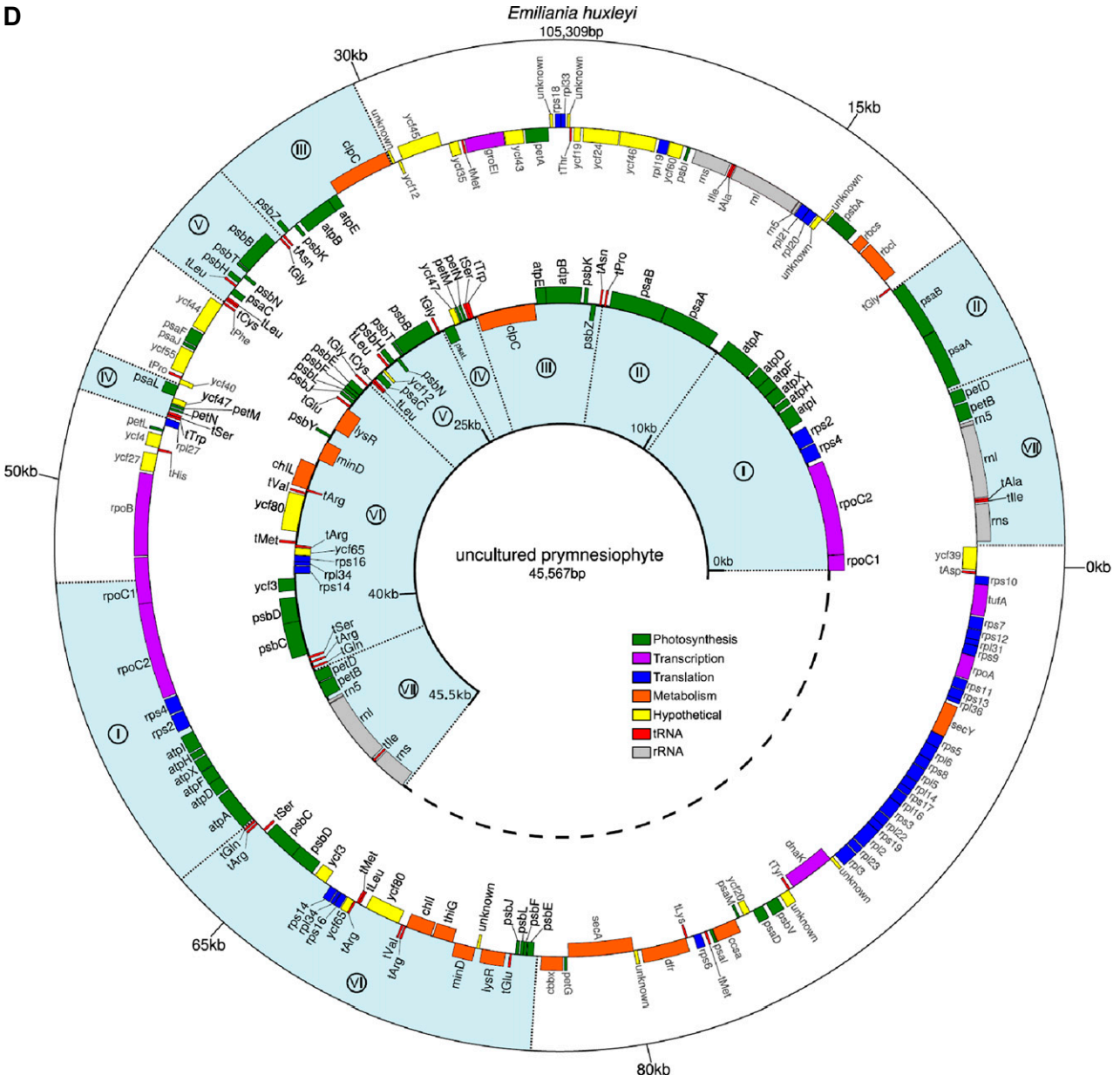

Fig. S3. Pico-prymnesiophyte chloroplast genome scaffolds. (*A*) Maximum likelihood reconstruction of prymnesiophyte chloroplast-derived 16S rRNA gene sequences in the environmental flow sort. 16S rRNA gene sequences (blue) identified on pico-prymnesiophyte chloroplast metagenomic scaffolds, and GenBank reference sequences (black) taken only from cultured strains that have both 16S and 18S rRNA gene sequences available, with the exception of coccolithophores, for which the criteria were at the genus level for two genera owing to limited sequence availability. The scaffold size is provided in parentheses. Bootstrap support was computed using maximum likelihood (100 replicates) and neighbor-joining (10,000 replicates) methods. Nodes retaining bootstrap support are indicated for those above 70% by both methods (black circles) and above 70% with only one method (white circles). Arrow indicates the chloroplast assembly annotated in subsequent figures. (*B*) Maximum likelihood tree of 22 concatenated plastid-encoded and cyanobacterial protein sequences, including genes on the pico-prymnesiophyte metagenomic scaffold C19847. Phylogenetic reconstruction was performed using the JTT matrix and based on a multiple alignment of 4,425 sites. Bootstrap support was computed using maximum likelihood (100 replicates) and values above 75% shown. (*C*) Global nucleotide conservation between chloroplast genomes from the two diatoms, *P. tricornutum* and *T. pseudonana*, and between pico-prymnesiophyte scaffold C19847and *E. huxleyi*. (*D*) Genome maps of scaffold C19847and the *E. huxleyi* chloroplast genome. Genes are color-coded according to functional attributes. Genes depicted on the outside are transcribed clockwise; genes depicted on the inside are transcribed counterclockwise. Blue background indicates gene clusters conserved between the uncultured pico-prymnesiophyte and *E. huxleyi* (Roman numerals I to VII). White background indicates *E. huxleyi* genes not represented in the pico-prymnesiophyte partial chloroplast genome assembly.

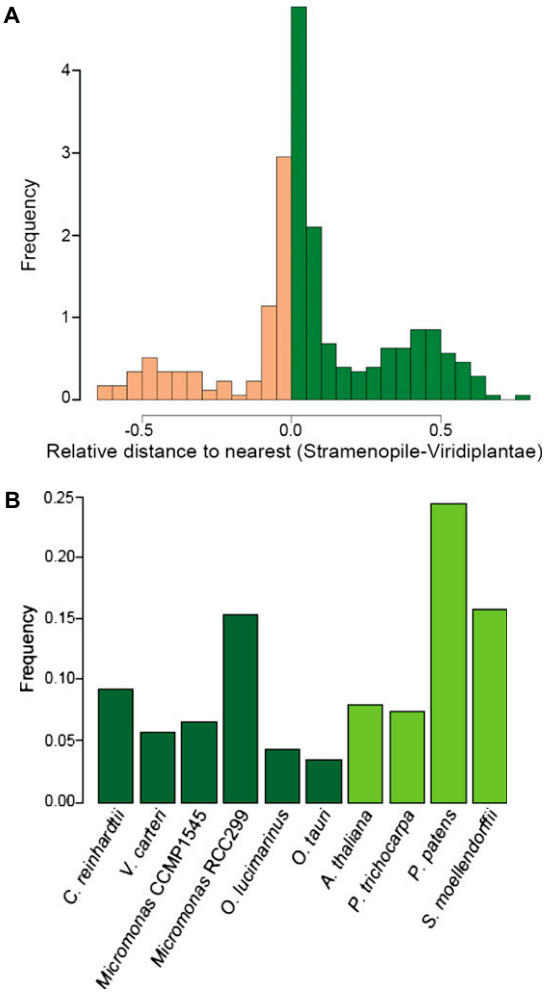

Fig. S4. Comparison of pico-prymnesiophyte tree distances to stramenopile vs. Archaeplastida (specifically Viridiplantae) genes, and distributions. (*A*) For pico-prymnesiophyte gene trees that contained at least one stramenopile representative and one Viridiplantae representative (352 in total), the number closer to the former, or to the latter, was determined first. A total of 229 genes (65%) were closer to a Viridiplantae gene than to a stramenopile gene. The histogram represents proportions of tree distances from pico-prymnesiophyte genes to the nearest reference stramenopile minus the distance to the nearest Viridiplantae, relative to the longer of the two distances. Frequency was calculated such that the area of the entire histogram would sum to 100% and was greater than 1 for some bins because the bin size used (0.05) is less than 1. Distance was measured as the sum of all branch lengths on the path from a pico-prymnesiophyte gene to a reference stramenopile or Viridiplantae in the best tree inferred (*SI Materials and Methods, Section 7*). This measure of distance is used instead of a pairwise edit distance, taking advantage of the phylogenetic context of the other species in the tree, rather than just the pair of species being compared. Tree distance takes advantage of the best tree topology and is robust to changes in topology because branch lengths are estimated during tree construction to minimize globally the difference between this measure of tree distance and pairwise edit distances. For this reason, bootstrap values were not considered, and distances were calculated for the 352 pico-prymnesiophyte genes with trees containing at least one reference from each group. Using this formula, sequences closer to Viridiplantae were represented by positive values (green), because one must go further to get to a stramenopile sequence. Pico-prymnesiophyte sequences closer to a (beige) stramenopile result in a negative *x* axis value. Those pico-prymnesiophyte genes for which the comparison could be made were on average 6.7% closer to the Viridiplantae gene than to the stramenopile gene. (*B*) Distribution within the closest Viridiplantae for pico-prymnesiophyte sequences also found in stramenopiles. Proportions of nearest reference Viridiplantae to the 229 pico-prymnesiophyte genes that are closer to Viridiplantae than to stramenopiles. Reference species are divided into Chlorophyta (dark green) and Streptophyta (light green), accounting for 45% and 55% of genes, respectively. The two nonvascular plant species within Streptophyta (*P. patens* and *S. moellendorffii*) account for 40% of the total.

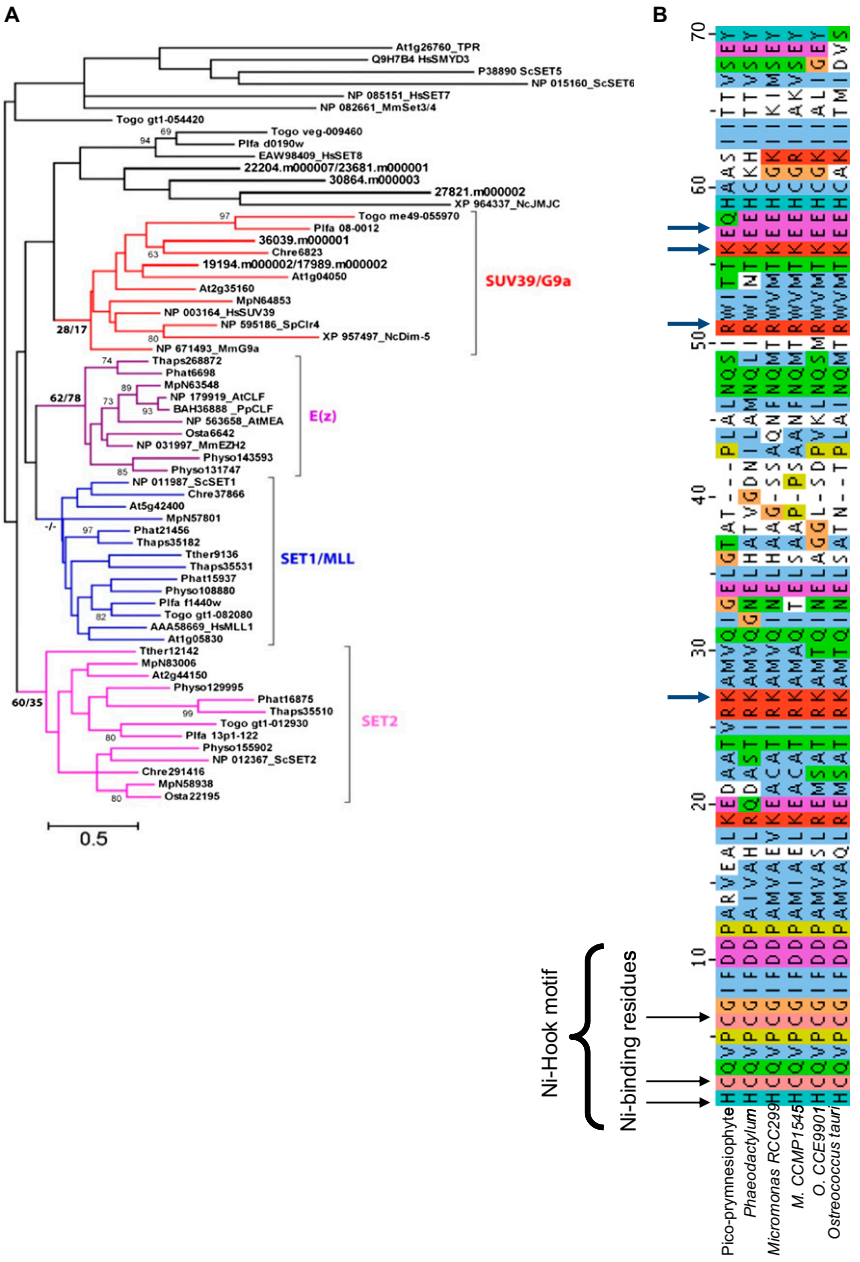

Fig. S5. Phylogenetic analysis and alignments of some protein encoding genes discussed in the main text. (*A*) Genes containing SET domains formed well-conserved major subfamilies including SET1/MLL (blue), E(z) (violet), SET2 (pink), and SUV39/G9a (red). Ten SET domain homologs were identified in the pico-prymnesiophyte metagenome, including three LSMT (RUBISCO large subunit methyltransferase) homologs, not analyzed further. 17989*/19194/36039 belong to the SUV39 subfamily, which was not found in available stramenopile genomes. Four other SET domain genes (22204/23681**, 30864, and 27821) fell among divergent subfamilies such as SET8 and JmjC (fungi-specific). These less-conserved subfamilies (black) include many lineage specific expansions. The specific expansion of SET domain genes contributed to the enlarged SET domain gene family in pico-prymnesiophytes. Pico-prymnesiophyte sequences are indicated by thick branches with enlarged font. Sequences were retrieved either from GenBank (starting with NP and accession, followed by species symbol) or genome catalogs (species symbol plus catalog protein ID). Species symbols: At, *A. thaliana*; Hs, *H. sapiens*; Sc, *S. cerevisiae*; Mm, *M. musculus*; Togo, *T. gondii*; Plfa, *P. falciparum* 3D7; Nc, *N. crassa*; Chre, *C. reinhardtii*; MpN, *Micromonas* RCC299; Sp, *S. pombe*; Thaps, *T. pseudonana*; Phat, *P. tricornutum*; Pp, *P. patens*; Osta, *O. tauri*; Physo, *P. sojae*; Tther, *T. thermophila*. Where retained, node support is shown as RAxML/MultiPHYL percentages. *17989 contained only a partial fragment of the SET domain and was 94% identical to sequence 19194, thus the branch of 17989 and 19194 was collapsed. **22204 and 23681 are distinct (98% identical) but are 100% identical in the alignment (because of masking). (*B*) Putative Ni-SOD proteins found in marine eukaryotic phytoplankton genomes. Encoded proteins were screened for using the published Ni-hook motif identified in *O. lucimarinus*. This 12-aa polypeptide binds Ni and catalyzes SOD activity. Other residues important to protein maturation and electrostatic guidance (blue arrows) are also conserved in eukaryotic Ni-SODs. Signal peptides or other N-terminal extensions were divergent and trimmed before alignment with ClustalX. The complete alignment is not shown because of space limitations. Protein IDs are as follows (JGI Prot. IDs, except for *O. tauri,* Ghent ID, and the pico-prymnesiophyte sequence): 64440, *Micromonas* RCC299; 36384, *Micromonas* CCMP1545; 49037, *P. tricornutum* CCAP1055; 29162, *O. lucimarinus*; Ot01g05280, *O. tauri* ; 26474.m000001, pico-prymnesiophyte.

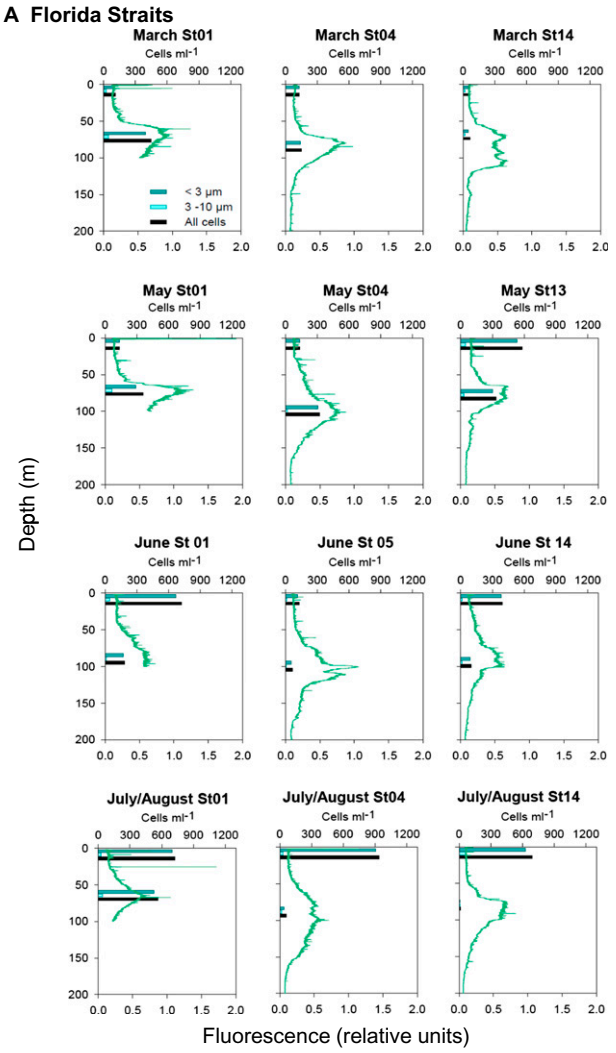

Fig. S6. (Continued)

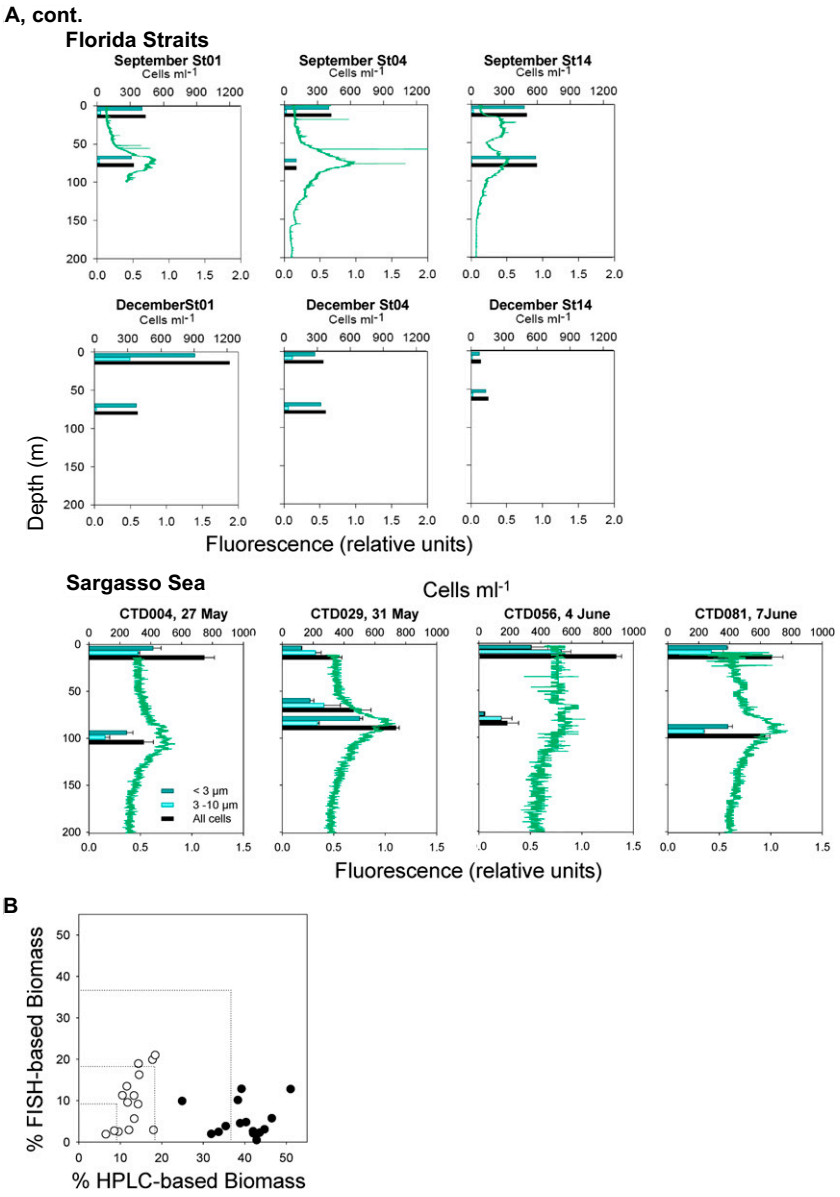

Fig. S6. Depth and size distribution of pico-prymnesiophytes in the subtropical North Atlantic and comparison with HPLC data. (*A*) Seasonal pico-prymnesiophyte cell concentrations in the Florida Straits and Sargasso Sea in 2005. In the Florida Straits time-series pico-prymnesiophytes were enumerated by FISH at three stations. Dark turquoise, <3-μm cells; light turquoise, 3–10-μm cells, measured cells in the latter range averaged 2.8 μm × 3.4 μm (Table S7); black, total numbers of cells including the <3-μm size fraction, 3–10-μm size fraction, and >10-μm size fraction. Data for two profiles at BATS (CTD004 and CTD029) and two at the NSS station (CTD056 and CTD081) are also shown. Error bars represent the SD of two hybridizations. At the NSS station, the average abundance was statistically higher at the surface than at the DCM ($P < 0.02$); no statistical difference was detected at BATS ($P < 0.98$). The light green line represents the in vivo fluorescence signature from the rosette mounted fluorometer (not available for the Florida Straits in December 2005). (*B*) Comparison of HPLC-based and FISH-based prymnesiophyte biomass contributions (%) to picophytoplankton biomass in the Florida Straits. At the DCM (black circles), prymnesiophyte biomass by HPLC seemed to overestimate their average contribution relative to the combined FISH- (pico-prymnesiophytes) and FCM- (other picophytoplankton groups) based estimates from the 2005 time series. Alternatively, the FISH- and FCM-based estimates may have underestimated prymnesiophyte contributions, or overestimated contributions by the other three picophytoplankton groups. At the surface (white circles), the relationship between these two data types was closer and no significant difference was detected. Only surface data were used to generate global biomass contribution data.

Table S1. Coordinates, dates, and environmental characteristics of 18S rRNA gene clone library sites and sort sites

Location	Latitude (N)	Longitude (W)	Date (d/m/y)	Sample depth (m)	Temperature (°C)	Sal (ppt)
BATS	31°39′20″	64°37′21″	29/05/05	75	20.1*	36.8*
BATS	31°39′20″	64°37′21″	01/06/05	15	21.9*	36.7*
NSS	35°09′24″	66°33′46″	05/06/05	15	21.4–21.6*	36.5–36.6*
NSS	35°09′24″	66°33′46″	08/06/05	70	20.1	36.6
NSS	33°14′07″	64°53′19″	07/04/01	150	18.6	36.6
NSS	36°07′30″	67°10′03″	08/04/01	4	21.9	36.5
NSS	33°14′07″	64°53′19″	07/04/01	15	18.7	36.6
CS	40°15′07″	70°25′23″	09/04/01	4	5.6	32.7
CS	40°15′07″	70°25′23″	09/04/01	20	5.3	32.7
CS	39°59′38″	71°48′01″	16/09/01	26	14.4	33.3
FS St01	25°30′07″	80°04′04″	30/03/05	5	24.3	36.4
FS St01	25°30′07″	80°04′04″	30/03/05	70	21.2	36.4
FS St01	25°30′04″	80°03′59″	01/08/05	5	30.1	36.1
FS St01	25°30′04″	80°03′59″	01/08/05	65	23.5	36.3
FS St04	25°30′01″	79°57′20″	31/03/05	5	24.7	36.3
FS St04	25°30′04″	79°57′18″	01/08/05	5	30.3	36.0
FS St04	25°30′04″	79°57′18″	01/08/05	89	24.2	36.5
FS St04[†]	25°30′04″	79°57′18″	27/02/07	75	23.3	36.7
FS St08[†]	25°18′00″	79°34′12″	27/02/07	141	23.25	36.7
FS St14	25°29′59″	79°20′58″	30/03/05	5	25.7	36.2
FS St14	25°29′59″	79°20′58″	30/03/05	70	24.7	36.6
FS St14	25°29′55″	79°20′54″	31/07/05	5	29.9	36.0
FS St14	25°29′55″	79°20′54″	31/07/05	80	25.9	36.5
FS St14[‡]	25°30′01″	79°21′04″	08/12/05	58	26.2	36.3
NEP[§]	36°07′34″	123°29′24″	01/10/07	10	16.1	33.0
NEP	33°17′12″	129°25′41″	07/10/07	10	19.0	33.2
NEP[§]	33°17′12″	129°25′41″	07/10/07	90	13.7	33.1

BATS, Bermuda Atlantic Time-series Study; NSS, Northern Sargasso Sea; FS, Florida Straits; St, station; CS, continental shelf; NEP, North East Pacific Ocean.
*Parameters measured with CTD detector before the sample collection (if single measurement) or before and the following day (if range of measurements), as these samples were collected with a GO-FLO bottle not equipped with a CTD detector.
[†]From MDA-flow sort DNA.
[‡]Two clone libraries, two size fractions: <2 μm and >2 μm.
[§]Two clone libraries.

Table S2. Sequences belonging to collapsed clades of Fig. 2 (main text)

Group no.	Sequence name	Clade
1	FS01AA64_01Aug05_5m, FS01AA86_01Aug05_5m, FS14K020_31July05_5m	ND
2	FS01D004_01Aug05_65m, FS14JA13_30Mar05_5m	ND
3	FS14L039_31July05_80m, FS14I027_30Mar05_70m, FS14M001_08Dec05_58m OC413BATS_O069_75m, OC413BATS_P038_15m, CN207St155_8Be04F_07Oct07_90m, CN207St70_BBe08M_01Oct07_10m, FS04R13_ SGZW1118_27Feb07_75m (sort), *OLI11007, OLI16010, OLI26047, OLI51080*	D
4	OC413BATS_P009_15m, EN351CTD040_40_09Apr01_4m, EN351CTD040_09Apr01_20m	A
5	FS01B073_30Mar05_5m, FS04E091_31Mar05_5m, FS14JA52_30Mar05_5m OC413BATS_P053_15m	A
6	*OLI26041, OLI51050*	E
7	FS04E051_31Mar05_5m, FS14K029_31July05_5m	C
8	FS04R13_7_27Feb07_75m (sort), FS04R14_1_27Feb07_75m (sort), FS04R13_ SGZW1078_27Feb07_75m (sort), FS08L1_2_27Feb07_141m sort (sort), FS14L086_31July05_80m, FS14M077_08Dec05_58m, *F01N5*	ND
9	*OLI51059, OLI51033*	B
10	OC413BATS_P036_15m, *Q2G11N10*	B1
11	FS04GA78_01Aug05_5m, OC413BATS_P070_15m	B2
12	FS14K084_31July05_5, *IND70.03*	B2
13	FS01D092_01Aug05_65m, OC413NSS_Q058_15m	
14	FS01D027_01Aug05_65m, FS04GA50_01Aug05_5m, OC413BATS_P003_15m CN207St155_8Ae02Y_07Oct07_90m, *MB07.32*	B2
15	FS01B058_30Mar05_5m, FS04H103_01Aug05_89m, FS14L014_31July05_80m FS14M055_08Dec05_58m, OC413NSS_Q042_15m, OC413NSS_R062_70m, *OLI16029*	B2
16	FS01D093_01Aug05_65m, FS04E093_31Mar05_5m, FS04G183_01Aug05_5m OC413NSS_Q003_15m, OC413NSS_Q028_15m	B2
17	FS01D024_01Aug05_65m, *B04N10*	B2
18	FS01D021_01Aug05_65m, FS01C021_30Mar05_70m, CN207St155_8Ae02U_07Oct07_10m, EN351CTD040_16_09Apr01_4m	B2
19	OC413BATS_P088_15m, *BL010625.10, IND31.140*	B2
20	FS01D058_01Aug05_65m, *N10E02*	B2
21	FS14L079_31July05_80m, OC413BATS_O028_75m	B2
22	FS01AA62_01Aug05_5m, FS04GA79_01Aug05_5m	B2

Sequences retrieved from size-fractioned (<2 or <3 μm) samples or flow sorts were collapsed after phylogenetic analysis when all sequences within a bootstrap-supported clade had ≥99% identity. Names in Italics denote previously published sequences, whereas other sequences are from this study. Also indicated is assignation to clades A to E as identified in previous publications (*SI Materials and Methods, Section 4*). ND, not determined (too divergent from clades noted in previous publications).

Table S3. Chloroplast protein-encoding genes for pico-prymnesiophyte metagenome assembly scaffold C19847

Gene name	Strand	ORF start	ORF stop	Length	Annotation
rpoC1	+	n/a	615	n/a	RNA polymerase β′ subunit (partial)
rpoc2	+	641	4410	1256	RNA polymerase β′′ subunit*
rps4†	+	4583	5190	202	30S ribosomal protein S4*
rps2†	+	5250	5927	225	30S ribosomal protein S2
atpI	+	6196	6897	233	ATP synthase CF0 A subunit
atpH†	+	7000	7248	82	ATP synthase CF0 C subunit
atpX	+	7335	7826	163	ATP synthase CF0 B′ subunit
atpF	+	7852	8328	158	ATP synthase CF0 B chain
atpD	+	8332	8877	181	ATP synthase CF1 δ subunit
atpA†	+	8972	10474	500	ATP synthase CF1 α subunit
psaA†	+	10982	13240	752	Photosystem I P700 chlorophyll a apoprotein A1
psaB†	+	13270	15474	734	Photosystem I P700 chlorophyll a apoprotein A2
psbZ	—	16247	16059	62	Photosystem II protein Z
psbK†	+	16421	16558	45	Photosystem II protein K
atpB	+	16699	18126	475	ATP synthase CF1 β subunit
atpE†	+	18132	18524	130	ATP synthase CF1 epsilon subunit
clpC	—	21062	18609	786	Clp protease ATP binding subunit
petN	+	21541	21628	29	Cytochrome b6/f complex subunit VIII
petM	+	21693	21788	32	Cytochrome b6/f complex subunit VII
ycf47	+	21832	22053	73	Hypothetical chloroplast protein RF12
psaL	—	22549	22112	145	Photosystem I subunit XI
psbB†	+	22999	24528	509	Photosystem II 47 kDa protein
psbT†	+	24574	24669	32	Photosystem II protein T
psbN†	—	24903	24772	43	Photosystem II protein N
psbH†	+	25015	25215	66	Photosystem II protein H
ycf12	—	25652	25548	34	Hypothetical chloroplast RF12
psaC†	—	26020	25775	81	Photosystem I subunit VII
psbE†	+	26794	27048	84	Photosystem II protein V
psbF†	+	27071	27199	42	Photosystem II protein VI
psbL†	+	27210	27326	38	Photosystem II protein L
psbJ†	+	27385	27504	39	Photosystem II protein J
lysR	—	28778	27810	322	LysR transcriptional regulator
psbY	+	29093	29201	36	Hypothetical protein EmhuCp074
mind	—	30223	29408	271	Septum-site determining protein
chlI	+	30496	31551	352	Mg-protoporyphyrin IX chelatase
ycf80	+	31858	33378	506	Hypothetical chloroplast RF80
ycf65	—	34350	34054	98	Hypothetical chloroplast RF65
rps16	—	34617	34375	80	30S ribosomal protein S16
rpl34	—	34772	34635	45	50S ribosomal protein L34
rps14†	—	35122	34820	100	30S ribosomal protein S14
ycf3†	+	35323	35838	171	Photosystem I assembly protein ycf3
psbD†	+	36117	37169	351	Photosystem II protein D2
psbC†	+	37120	38535	471	Photosystem II 44 kDa protein
petD	—	39816	39334	160	Cytochrome b6/f complex subunit IV
petB†	—	40510	39863	215	Cytochrome b6

*Frameshift.
†Genes used in concatenated phylogeny.

Table S4. tRNAs annotated on chloroplast metagenome scaffold C19847

tRNA type	Anti-codon	Start	Stop	Strand	Cove score
Pro	TGG	15639	15712	+	76.73
Asn	GTT	15838	15909	+	69.62
Trp	CCA	21217	21289	+	68.8
Ser	GCT	21352	21440	+	68.86
Gly	GCC	22748	22819	+	73.52
Leu	TAG	25316	25398	+	59.2
Leu	TAA	26214	26131	–	64.49
Cys	GCA	26296	26226	–	62.91
Gly	TCC	26495	26565	+	73.93
Glu	TTC	27645	27717	+	55.77
Arg	TCT	31624	31552	–	56.62
Val	TAC	31779	31708	–	87.27
Met	CAT	33798	33882	+	54.06
Arg	CCG	34012	33940	–	68.84
Ser	TGA	38741	38656	–	62.44
Arg	ACG	38949	38876	–	64.55
Gln	TTG	39179	39108	–	68.73
Ile	GAT	43860	43787	–	82.71

Table S5. Peptide ID and annotation for nuclear genes discussed in main text

Putative role/domain detection	Protein IDs
Nudix hydrolases	20464.m000001, 20464.m000002, 25045.m000009
Arylsulfatses	18367.m000003, 26116.m000009, 17445.m000009
Polyketide synthase	24556.m000001, 19587.m000001, 34726.m000001, 34726.m000002
acid phosphatase, EC 3.1.3.2	29432.m000003
AP2, PF00847	32547.m000001, 17980.m000002, 25976.m000003, 16791.m000001, 29301.m000002
Cir_N, PF10197	23425.m000001
JmjC, PF02373	27358.m000001, 19723.m000001
MOV34, PF01398	20223.m000002
zf-C2H2, PF00096	16291.m000006
Acetyltransf_1, PF00583	20820.m000001, 30054.m000002, 30012.m000001
mTERF, PF02536	33687.m000001
Myb_DNA-binding, PF00249	16507.m000002, 16824.m000002
Response_reg, PF00072	31813.m000002
SET, PF00856	19194.m000002, 36039.m000001, 17989.m000002, 23681.m000001, 22204.m000007, 30864.m000003, 27821.m000002, 17701.m000003, 24482.m000005, 28775.m000002
Sigma70	17671.m000003
SNF2_N, PF00176	22427.m000006, 22256.m000002, 26579.m000003, 25564.m000001, 26857.m000001, 24084.m000001, 28354.m000003, 28354.m000002
SWIRM, PF04433	16824.m000002
zf-CCCH, PF00642	19661.m000003
SWIRM+MYB, PF04433 +PF00642	16824.m000002

Table S6. Comparison of pico-prymnesiophyte metagenome assembly transcription factors to those in other genomes

	Metagenome	Stramenopiles			Ciliates	Apicom	Prasinophytes				Chloro	Plant	
	Pico-prym	*Tpseu*	*Ptric*	*Psoja*	*Tthe*	*Pfal*	*RCC299*	*CCMP1545*	*Otaur*	*Oluci*	*Crein*	*Athal*	*Pfam*
Transcription regulators													
AP2/EREBP	**0.31**	0.09	0.01	0.01	0.03	0.46	0.15	0.13	0.10	0.12	0.09	0.44	PF00847
Cir_N	0.06	0.02	0.02	0.01	0.00	0.00	0.02	0.02	0.00	0.01	0.01	0.01	PF10197
Zf-C2H2	0.06	0.01	0.01	0.13	0.03	0.02	0.08	0.09	0.03	0.09	0.02	0.17	PF00096
mTERF	0.06	0.05	0.04	0.02	0.00	0.02	0.04	0.05	0.05	0.04	0.06	0.11	PF02536
Myb	0.18	0.30	0.29	0.37	0.12	0.13	0.38	0.31	0.48	0.47	0.25	0.87	PF00249
Response reg	0.06	0.07	0.15	0.03	0.33	0.02	0.11	0.11	0.10	0.14	0.08	0.113	PF00072
Sigma 70	0.06	0.07	0.08	0.00	0.00	0.00	0.02	0.02	0.03	0.01	0.01	0.02	PF04542
SWIRM	0.06	0.02	0.02	0.01	0.00	0.02	0.04	0.03	0.04	0.05	0.01	0.03	PF04433
Transcription regulators-associated													
Mov34	0.06	0.06	0.05	0.04	0.03	0.07	0.10	0.09	0.08	0.11	0.07	0.05	PF01398
Zf-CCCH	0.06	0.13	0.12	0.12	0.10	0.20	0.24	0.19	0.19	0.26	0.11	0.19	PF00642
Chromatin regulators													
Acetyltransf_1	0.25	0.25	0.34	0.16	0.09	0.11	0.38	0.36	0.36	0.38	0.22	0.09	PF00583
JmjC	0.12	0.06	0.09	0.04	0.02	0.02	0.16	0.18	0.16	0.22	0.07	0.06	PF02373
SET	**0.62**	0.47	0.43	0.28	0.07	0.13	0.45	0.46	0.40	0.45	0.31	0.15	PF00856
SNF2_N	**0.43**	0.23	0.25	0.13	0.07	0.20	0.31	0.31	0.36	0.35	0.13	0.15	PF00176

Shown is the percentage of total genes composed of the respective group. Genome versions and methods are as in *SI Materials and Methods, Section 7*, whereas protein IDs for the pico-prymnesiophyte sequences are provided above. The high representation of SET and SNF2 chromatin regulators in the pico-prymnesiophyte metagenome, relative to other sequenced protistan genomes, as well as AP2 domain containing genes, is emphasized with bold text. Pico-prym, pico-prymnesiophyte; *Tpseu*, *T. pseudonana*; *Ptric*, *P. tricornutum*; *Psoja*, *P. sojae*; *Tthe*, *T. thermophila*; *Pfal*, *P. falciparum*; RCC299, *Micromonas* RCC299; CCMP1545, *Micromonas pusilla* CCMP1545; *Otaur*, *O. tauri*; *Oluci*, *O. lucimarinus*; *Crein*, *C. reinhardtii*; *Athal*, *A. thaliania*; *Apicom*, *Apicomplexans*; *Chloro*, *Chlorophytes*.

Table S7. Average size, biovolume, and biomass conversion factor for various picophytoplankton groups

Organism	Size (μm)	Biovolume (μm^{-3})	Conversion factor (fg C $cell^{-1}$)
Prochlorococcus	*	*	39*
Synechococcus	*	*	82*
"Nonprym" picoeukaryotes	*	*	530*
Pico-prym class 1	2.0 × 2.0	4.2	995
Pico-prym class 2	2.0 × 2.5	5.2	1,232
Pico-prym class 3	2.0 × 3.0	6.3	1,493
Pico-prym class 4	2.5 × 3.5	11.5	2,726

Measurements for four size classes of pico-prymnesiophytes binned during counting at all locations, except OC413 and Florida Straits. For the pico-prymnesiophytes classes, biovolume and then a biomass conversion factor was calculated from average size (numbers are rounded after calculation). More precise cell size information was available for the NSS and BATS (i.e., average within size class of 1.9 ± 0.4 × 2.1 ± 0.3 μm, n = 89; 2.8 ± 0.6 × 3.4 ± 0.5 μm, n = 127), resulting in slightly different biovolumes (4.0 and 14.0 μm^{-3}) than for the below class 1 and 4. The former were used to generate 2 of the 121 global data points, averaging all values for each Sargasso site. In the Florida Straits, representing one biogeographical province data point (Fig. 4), cells were binned into two size classes, ≤3 μm and >3 μm; the majority were <3 μm in their largest dimension. Pico-prymnesiophyte biomass values refer to the sum of the individually calculated biomass for each group (i.e., biomass conversion factor multiplied by cell concentration). The same carbon factor per unit volume was used for all organisms (237 fg C μm^{-3}) as previously published (*SI Materials and Methods, Section 10*). Pico-prym, pico-prymnesiophytes.
*From a previous publication (*SI Materials and Methods, Section 10*).

Table S8. Picophytoplankton group abundances and biomass

				Pro		*Syn*		Nonprym		Pico-prym	
Lat	Long	Loc	Temp (°C)	Abund	Bio	Abund	Bio	Abund	Bio	Abund	Bio
65°40′ S	170°01′ W	SO	−0.24	517	0.0	0	0.0	3,468	1.8	590	1.6
63°29′ S	170°01′ W	SO	0.53	1,301	0.1	0	0.0	2,201	1.2	1,196	3.2
61°28′ S	169°59′ W	SO	1.60	1,101	0.0	0	0.0	1,567	0.8	862	2.3
61°00′ S	170°01′ W	SO	2.16	1,767	0.1	0	0.0	1,801	1.0	1,479	4.0
59°00′ S	170°02′ W	SWP	4.97	1,667	0.1	133	0.0	5,036	2.7	1,920	5.2
57°00′ S	170°00′ W	SWP	5.59	3,418	0.1	67	0.0	9,254	4.9	6,122	16.6
55°29′ S	170°02′ W	SWP	6.68	7,187	0.3	550	0.0	13,723	7.3	4,856	13.2
51°59′ S	170°05′ W	SWP	9.90	9,104	0.4	37,701	3.1	15,240	8.1	2,128	5.8
50°00′ S	169°59′ W	SWP	12.92	22,344	0.9	39,252	3.2	8,070	4.3	1,237	3.4
48°29′ S	170°00′ W	SWP	13.54	8,020	0.3	6,286	0.5	6,220	3.3	743	2.0
45°33′ S	172°17′ W	SWP	15.77	36,817	1.4	66,648	5.5	15,190	8.1	1,054	2.9
44°19′ S	173°45′ W	SWP	16.86	26,946	1.1	33,732	2.8	7,920	4.2	1,247	3.4
43°00′ S	095°00′ E	IO	10.53	12,047	0.5	1,441	0.1	5,163	2.7	482	1.1
42°10′ S	171°14′ W	SWP	19.14	243,197	9.5	333	0.0	3,969	2.1	1,188	3.2
39°59′ S	110°00′ E	IO	10.56	51,989	2.0	5,830	0.5	6,956	3.7	913	2.5
39°59′ S	095°00′ E	IO	11.04	11,659	0.5	5,707	0.5	9,050	4.8	497	1.1
36°59′ S	044°59′ E	IO	12.29	83,765	3.3	1,844	0.2	10,180	5.4	987	1.8
35°31′ S	081°58′ E	IO	13.37	52,868	2.1	5,660	0.5	5,821	3.1	1,205	2.2
34°54′ S	081°17′ E	IO	13.8	62,253	2.4	1,324	0.1	8,744	4.6	1,051	1.9
34°20′ S	079°20′ E	IO	14.33	41,270	1.6	4,065	0.3	6,719	3.6	885	1.5
34°10′ S	087°09′ E	IO	14.31	40,496	1.6	3,857	0.3	6,830	3.6	1,501	2.7
34°00′ S	095°00′ E	IO	13.48	113,211	4.4	5,168	0.4	11,386	6.0	1,413	2.5
33°10′ S	090°10′ E	IO	14.95	71,949	2.8	2,211	0.2	8,067	4.3	1,171	2.1
31°59′ S	080°00′ E	IO	14.47	134,779	5.3	7,586	0.6	8,589	4.6	1,213	2.0
31°44′ S	094°59′ E	IO	14.51	103,695	4.0	16,263	1.3	10,949	5.8	2,052	2.9
28°59′ S	079°59′ E	IO	17.27	77,111	3.0	281	0.0	1,645	0.9	380	0.7
25°58′ S	079°59′ E	IO	20.18	79,461	3.1	442	0.0	4,981	2.6	387	0.7
25°00′ S	080°00′ E	IO	20.86	88,349	3.4	241	0.0	6,509	3.4	472	0.9
22°58′ S	079°59′ E	TrIO	20.63	99,815	3.9	805	0.1	7,604	4.0	348	0.6
22°00′ S	080°00′ E	TrIO	21.79	105,832	4.1	1,406	0.1	1,688	0.9	394	0.8
19°58′ S	080°01′ E	TrIO	22.47	144,585	5.6	1,929	0.2	1,527	0.8	474	0.9
16°59′ S	079°59′ E	TrIO	23.64	165,308	6.4	3,500	0.3	2,414	1.3	465	0.8
14°00′ S	080°00′ E	TrIO	25.28	176,459	6.9	5,877	0.5	4,428	2.3	782	1.3
11°59′ S	080°00′ E	TrIO	26.07	225,628	8.8	9,228	0.8	6,730	3.6	549	1.0
10°00′ S	140°01′ W	EPO	29.30	220,429	8.6	18,319	1.5	2,744	1.5	548	1.5
09°59′ S	080°00′ E	EPO	27.58	215,510	8.4	1,408	0.1	1,206	0.6	441	0.8
07°59′ S	080°00′ E	EIO	28.09	243,709	9.5	442	0.0	844	0.4	399	0.8
07°01′ S	140°00′ W	EPO	26.70	224,379	8.8	8,484	0.7	7,747	4.1	884	2.4
06°11′ S	140°20′ W	EPO	25.81	257,336	10.0	15,492	1.3	9,317	4.9	1,074	2.9
05°59′ S	080°00′ E	EIO	28.66	232,766	9.1	1,649	0.1	1,006	0.5	434	0.7
05°00′ S	140°01′ W	EPO	29.14	166,671	6.5	15,854	1.3	3,312	1.8	485	1.3
04°58′ S	140°01′ W	EPO	26.25	270,257	10.5	21,761	1.8	12,665	6.7	874	2.4
04°00′ S	025°00′ W	EAO	26.70	274,711	10.7	1,674	0.1	752	0.4	624	1.3
04°00′ S	080°00′ E	EIO	28.61	266,383	10.4	2,698	0.2	886	0.5	512	0.9
03°01′ S	025°01′ W	EAO	26.72	219,453	8.6	1,674	0.1	540	0.3	443	1.0
03°00′ S	140°00′ W	EPO	26.27	176,015	6.9	18,082	1.5	6,646	3.5	1,064	2.9
02°02′ S	139°53′ W	EPO	26.21	136,368	5.3	14,757	1.2	7,768	4.1	817	2.2
02°00′ S	140°01′ W	EPO	28.70	108,404	4.2	11,631	1.0	3,141	1.7	1,226	3.3
02°00′ S	025°00′ W	EAO	26.17	201,656	7.9	6,420	0.5	2,193	1.2	975	1.9
02°00′ S	080°01′ E	EIO	28.75	280,186	10.9	5,562	0.5	1,370	0.7	632	1.2
01°00′ S	140°00′ W	EPO	25.67	133,634	5.2	12,016	1.0	5,286	2.8	657	1.8
01°00′ S	140°00′ W	EPO	28.8	220,445	8.6	17,710	1.5	6,164	3.3	712	1.9
01°00′ S	025°01′ W	EAO	25.73	316,115	12.3	34,037	2.8	4,688	2.5	982	2.0
00°01′ S	140°03′ W	EPO	28.52	139,269	5.4	7,874	0.6	4,681	2.5	784	2.1
00°00′	025°00′ W	EAO	25.16	185,668	7.2	29,758	2.4	5,783	3.1	1,123	2.3
00°00′	080°00′ E	EIO	29.00	257,406	10.0	4,762	0.4	1,655	0.9	669	1.1
01°00′ N	140°00′ W	EPO	27.39	103,203	4.0	15,806	1.3	9,821	5.2	1,246	3.4
01°01′ N	140°02′ W	EPO	17.75	168,415	6.6	9,713	0.8	5,279	2.8	1,037	2.8

Table S8. Cont.

				Pro		*Syn*		Nonprym		Pico-prym	
Lat	Long	Loc	Temp (°C)	Abund	Bio	Abund	Bio	Abund	Bio	Abund	Bio
01°01′ N	025°01′ W	EAO	n.a.	519,792	20.3	79,899	6.6	3,401	1.8	487	1.1
01°30′ N	080°00′ E	EIO	28.26	193,698	7.6	39,136	3.2	8,369	4.4	695	1.1
02°00′ N	140°00′ W	EPO	28.4	43,828	1.7	8,380	0.7	9,239	4.9	898	2.4
02°00′ N	140°05′ W	EPO	26.97	198,767	7.8	13,178	1.1	6,257	3.3	967	2.6
02°00′ N	025°00′ W	EAO	27.13	360,438	14.1	10,869	0.9	1,769	0.9	436	1.0
03°00′ N	140°07′ W	EPO	27.10	185,590	7.2	16,270	1.3	8,358	4.4	1,415	3.8
03°01′ N	025°31′ W	EAO	27.27	239,751	9.4	12,649	1.0	879	0.5	426	1.0
03°30′ N	080°00′ E	EIO	28.32	242,568	9.5	49,653	4.1	7,515	4.0	851	1.4
03°59′ N	140°00′ W	EPO	27.87	183,326	7.1	17,934	1.5	7,433	3.9	1,050	2.9
04°01′ N	025°49 W	EAO	n.a.	268,058	10.5	12,395	1.0	1,261	0.7	697	1.2
04°57′ N	140°04′ W	EPO	26.64	225,176	8.8	10,935	0.9	5,912	3.1	1,503	4.1
04°58′ N	140°00′ W	EPO	27.01	178,535	7.0	16,530	1.4	5,146	2.7	1,811	2.7
05°00′ N	025°27′ W	EAO	n.a.	239,624	9.3	4,852	0.4	456	0.2	806	1.6
06°00′ N	140°00′ W	EPO	27.66	203,578	7.9	10,338	0.8	3,910	2.1	1,034	2.8
06°01′ N	026°28′ W	EAO	27.59	213,563	8.3	11,336	0.9	328	0.2	941	1.8
07°01′ N	026°50′ W	EAO	27.64	134,756	5.3	14,755	1.2	220	0.1	388	0.8
08°00′ N	027°10′ W	EAO	n.a.	150,143	5.9	15,444	1.3	29	0.0	605	1.2
09°00′ N	027°28′ W	EAO	27.72	230,386	9.0	24,514	2.0	710	0.4	489	1.0
09°56′ N	140°05′ W	EPO	18.78	168,378	6.6	1,358	0.1	895	0.5	416	1.1
09°59′ N	027°50′ W	EAO	27.44	196,782	7.7	11,971	1.0	244	0.1	400	0.8
10°00′ N	140°00′ W	EPO	28.8	249,458	9.7	597	0.0	1,122	0.6	378	1.0
11°00′ N	028°10′ W	NETA	n.a.	230,598	9.0	10,997	0.9	456	0.2	665	1.3
12°00′ N	028°21′ W	NETA	n.a.	248,735	9.7	7,606	0.6	328	0.2	2,063	4.2
12°59′ N	028°49′ W	NETA	27.04	168,744	6.6	3,119	0.3	67	0.0	478	1.0
13°59′ N	029°00′ W	NETA	25.82	295,179	11.5	3,835	0.3	286	0.2	1,241	2.3
15°00′ N	029°01′ W	NETA	n.a.	348,765	13.6	24,628	2.0	912	0.5	749	1.5
15°59′ N	029°00 W	NETA	24.64	213,831	8.3	3,336	0.3	302	0.2	941	2.0
16°59′ N	029°01′ W	NETA	n.a.	218,891	8.5	3,873	0.3	437	0.2	1,075	2.1
17°59′ N	028°59′ W	NETA	24.66	290,324	11.3	4,617	0.4	297	0.2	577	1.1
19°00′ N	028°59′ W	NETA	24.33	174,788	6.8	2,216	0.2	437	0.2	424	0.9
19°59′ N	029°02′ W	NETA	24.38	185,668	7.2	2,933	0.2	257	0.1	463	1.0
21°01′ N	028°28′ W	NETA	23.73	201,294	7.9	2,351	0.2	257	0.1	1,083	2.1
21°59′ N	027°56′ W	NETA	23.98	252,974	9.9	5,030	0.4	142	0.1	856	1.6
22°59′ N	027°26′ W	NETA	23.55	144,565	5.6	2,977	0.2	481	0.3	592	1.2
23°55′ N	154°33′ W	NEP	23.02	101,051	3.9	983	0.1	1,010	0.5	1,433	4.5
24°00′ N	027°00 W	NEA	23.85	328,550	12.8	5,133	0.4	864	0.5	1,168	2.1
24°58′ N	026°21′ W	NEA	23.28	131,984	5.1	2,754	0.2	123	0.1	861	1.9
25°30′ N	079°57′ W	FS	27.28!	86,518*	3.4	13,048†	1.1	1,394‡	0.7	433§	0.5
25°59′ N	025°47′ W	NEA	22.83	58,341	2.3	2,733	0.2	201	0.1	464	1.0
27°00′ N	025°14′ W	NEA	22.85	66,138	2.6	4,767	0.4	159	0.1	587	1.2
28°00′ N	024°41′ W	NEA	n.a.	41,602	1.6	2,267	0.2	328	0.2	297	0.7
28°11′ N	143°31′ W	NEP	18.67	83,484	3.3	2,872	0.2	1,388	0.7	233	0.4
28°59′ N	024°08′ W	NEA	22.79	58,256	2.3	3,411	0.3	286	0.2	545	1.1
30°00′ N	023°32′ W	NEA	23.16	26,131	1.0	3,244	0.3	102	0.1	410	0.9
30°24′ N	137°37′ W	NEP	17.18	98,120	3.8	2,027	0.2	1,471	0.8	300	0.4
31°39′ N	064°37′ W	BATS	22.01!	16,000	0.6	15,964	1.3	1,061	0.6	536¶	1.1
32°59′ N	021°49′ W	NEA	23.27	32,324	1.3	9,377	0.8	91	0.0	443	1.0
33°58′ N	021°13′ W	NEA	n.a.	60,874	2.4	3,986	0.3	139	0.1	826	2.0
34°04′ N	128°26′ W	NEP	13.03	8,969	0.3	2,379	0.2	5,720	3.0	1,616	4.6
35°00′ N	020°34′ W	NEA	22.71	62,180	2.4	8,037	0.7	481	0.3	537	1.2
35°09′ N	066°33′ W	NSS	21.87!	74,124	2.9	31,394	2.6	2,500	1.3	768‖	1.7
36°35′ N	122°31′ W	NEP	12.02	2,434	0.1	10,847	0.9	24,925	13.2	526	5.3
39°59′ N	019°59′ W	NEA	20.10	77,575	3.0	1,212	0.1	1,225	0.6	1,678	3.8
41°59′ N	019°59′ W	NEA	19.19	158,574	6.2	3,328	0.3	1,375	0.7	1,554	3.6
44°59′ N	020°00 W	NEA	19.07	101,111	3.9	8,022	0.7	770	0.4	1,801	3.9
52°01′ N	020°00 W	NEA	15.08	9,286	0.4	31,842	2.6	8,560	4.5	3,024	7.5
55°01′ N	019°59′ W	NEA	14.21	4,492	0.2	18,863	1.5	10,449	5.5	7,144	16.4

Table S8. Cont.

				Pro		*Syn*		Nonprym		Pico-prym	
Lat	Long	Loc	Temp (°C)	Abund	Bio	Abund	Bio	Abund	Bio	Abund	Bio
56°00′ N	019°59′ W	NEA	13.15	3,087	0.1	28,210	2.3	8,608	4.6	4,420	9.8
57°01′ N	020°00 W	NEA	13.01	660	0.0	17,977	1.5	3,974	2.1	2,074	4.8
58°01′ N	019°59′ W	NEA	12.31	278	0.0	22,592	1.9	3,911	2.1	757	1.8
59°00′ N	020°01 W	NEA	11.10	2,317	0.1	40,319	3.3	3,347	1.8	3,126	8.1
60°01′ N	020°00 W	NEA	11.04	1,053	0.0	56,299	4.6	1,392	0.7	1,012	2.5
61°00′ N	019°55′ W	NEA	10.18	3,330	0.1	17,410	1.4	5,073	2.7	1,907	4.7

Abundance (Abund; cell mL^{-1}) and biomass (Bio; μg C L^{-1}) of *Prochlorococcus* (*Pro*), *Synechococcus* (*Syn*), nonprymnesiophyte picoeukaryotes (Nonprym), and pico-prymnesiophytes (Pico-prym) for surface samples used to generate the biogeographical province biomass averages (Fig. 4). Values represent individual surface samples with the exception of those marked, which represent averages at sites with more intensive or seasonal sampling. Lat, latitude; Long, longitude; Loc, location; n.a., not available; SO, Southern Ocean; SWP, South West Pacific Ocean; IO, Indian Ocean; TrIO, tropical Indian Ocean; EP, Equatorial Pacific Ocean; EIO, Equatorial Indian Ocean; EAO, Equatorial Atlantic Ocean; NETA, North East tropical Atlantic Ocean; NEP, North East Pacific Ocean; NEA, North East Atlantic Ocean; FS, Florida Straits; BATS, Bermuda Atlantic Time-series Study; NSS, Northern Sargasso Sea; !, temperature data averaged over multiple dates. Averaged count data as below.

*SD = 23,017, range = 53,048–139,675 (n = 14).

†SD = 8,103, range = 4,554–31,865 (n = 18).

‡SD = 1,219, range = 647–6,054 (n = 18).

§SD = 331, range = 66–1,225, (n = 18).

¶SD = 231, range = 310–777 (n = 4).

‖SD = 289, range = 614–903 (n = 4).