Supplemental Information: Newly discovered deep-branching marine plastid lineages are numerically rare but globally distributed

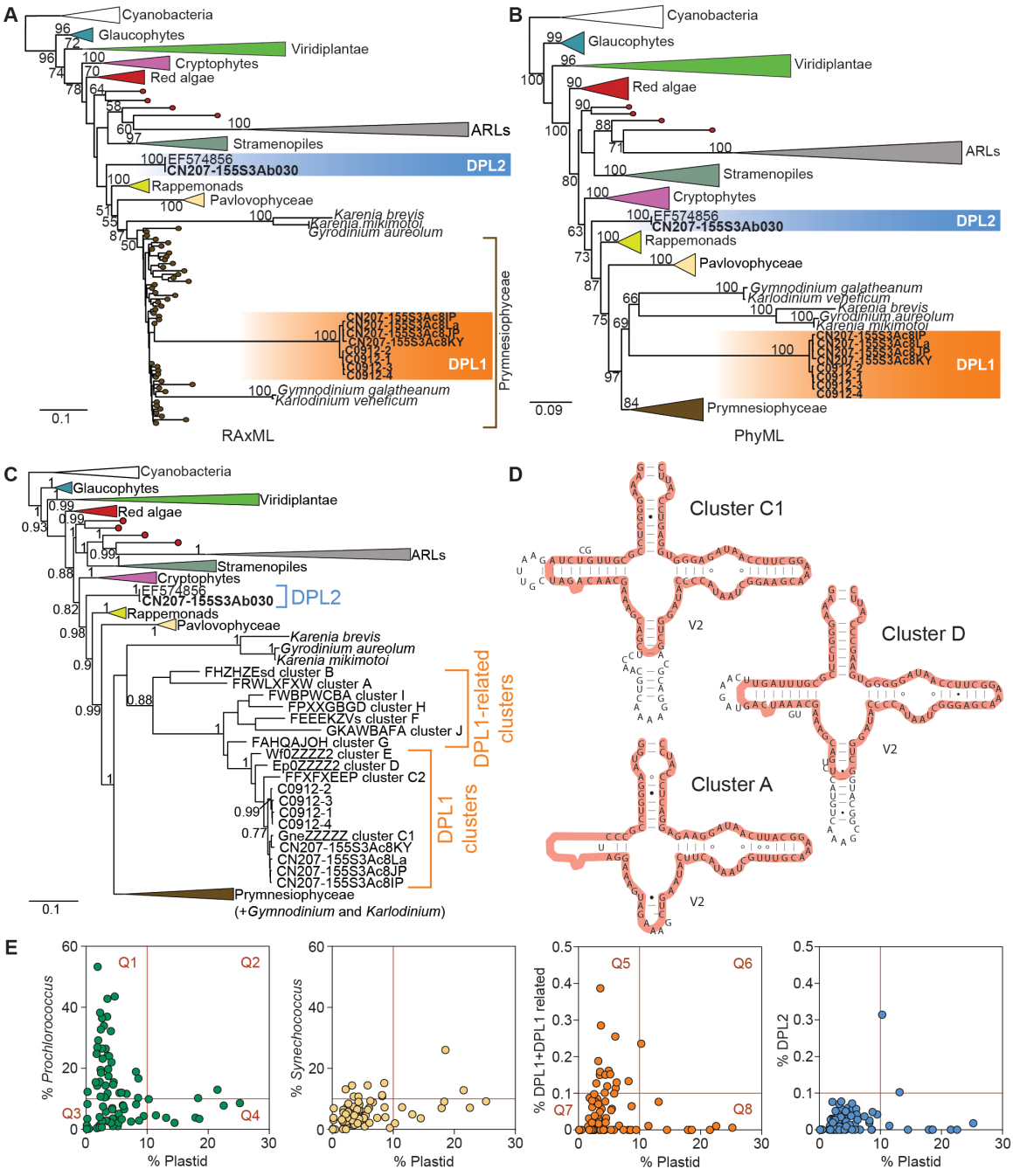
Chang Jae Choi, Charles Bachy, Gualtiero Spiro Jaeger, Camille Poirier, Lisa Sudek, V.V.S.S. Sarma, Amala Mahadevan, Stephen J. Giovannoni, Alexandra Z. Worden

Acknowledgements

We thank the Captains and crews of the *R/Vs Western Flyer*, *Weatherbird II* and *R. Revelle*, S. Sudek for laboratory support, M. Vermeij (CARMABI Marine Research Station), P. Keeling, F. Rohwer, B. Knowles and J. Janouskovec for organizing Curaçao field work. This research was supported by ONR N000141310451 (A.M.), MBARI, GBMF 1668 and GBMF 3788 (A.Z.W.).

Supplemental Figure

Figure S1. Related to Figure 1 and Experimental Procedures. Phylogenetic testing of the robustness of the phylogeny in Figure 1A, 16S V2 secondary structures, and comparison of different phytoplankton groups at BATS. (A, B) The alignment of near full-length 16S rRNA genes was modified by adding five fast-evolving dinoflagellate plastid sequences from taxa that acquired their plastid from a haptophyte alga so that in total 151 sequences from a representative suite of taxa were included and resulting topologies are similar to [S1]. Maximum Likelihood trees with (A) RAxML rapid bootstrap analysis of 1,000 replicates [S2] using the GTR+Γ model of evolution (-m GTRGAMMA -f a -# 1000) and (B) PhyML [S3] using the substitution model of GTR+Γ with 100 bootstrap replicates (-m GTR -f e -v e -c 8 -a e -b 100 -s BEST --n_rand_starts 10) are shown. (C) 16S rRNA gene phylogeny including representative amplicons from DPL1 and DPL1-related clusters. Bayesian tree with MrBayes using lset nst=6 rates=invgamma ncat=6, and ngenval =10,000,000 samplefreqval =1,000 and tempval =0.200 parameters is shown. Only posterior probabilities >0.70 are shown. (D) Example secondary structures for the 16S rRNA gene V2 region for DPL1 and DPL1-related sequences. For comparison, the pale red background shows the structure for *Escherichia coli* as in [S4]. (E) BATS surface water amplicon contributions by different photosynthetic groups. Relative abundance of the cyanobacteria *Prochlorococcus* and *Synechococcus*, as well as DPL1/DPL1-related clusters and DPL2 as a percent of total 16S amplicons (including all bacteria) plotted against the percent plastid amplicons (again as a percent of total 16S amplicons). For sample inclusion we required that >2,000 total amplicons passed quality control from the sample. Quadrants of interest are labeled. (F) Means (± standard deviation) for various environmental parameters for identified quadrants in (E) are provided and the number of samples (n) used to compute the mean is indicated. Note that while numerical values for temperature, salinity and silicate were available for all BATS samples, for many of these samples phosphate and nitrate were below detection limits, resulting in the large standard deviations in (F); interpretation of these data must be treated with caution. Additional information on seasonal dynamics at BATS based on the BATS program cruises can be found in [S5-9]. N.A., not available.



F

	Temp (°C)	Sal (ppt)	PO_4^{3-} (µM)	$NO_3^-+NO_2^-$ (µM)	SiO_4^{4-} (µM)
Q1	23.921 ± 2.680 (n=38)	36.568 ± 0.174 (n=38)	0.004 ± 0.013 (n=37)	0.027 ± 0.112 (n=34)	0.653 ± 0.262 (n=38)
Q2	19.473 ± 0.158 (n=2)	36.636 ± 0.014 (n=2)	0.000 ± 0.000 (n=2)	0.055 ± 0.078 (n=2)	0.685 ± 0.049 (n=2)
Q3	24.250 ± 3.163 (n=36)	36.575 ± 0.162 (n=36)	0.007 ± 0.021 (n=35)	0.013 ± 0.044 (n=36)	0.720 ± 0.322 (n=33)
Q4	20.382 ± 0.886 (n=8)	36.593 ± 0.141 (n=8)	0.015 ± 0.030 (n=8)	0.043 ± 0.098 (n=8)	0.690 ± 0.297 (n=6)

	Temp (°C)	Sal (ppt)	PO_4^{3-} (µM)	$NO_3^-+NO_2^-$ (µM)	SiO_4^{4-} (µM)
Q5	23.646 ± 2.772 (n=15)	36.574 ± 0.163 (n=15)	0.003 ± 0.010 (n=15)	0.005 ± 0.017 (n=13)	0.662 ± 0.246 (n=13)
Q6	21.719 (n=1)	36.518 (n=1)	0.000 (n=1)	0.000 (n=1)	N.A.
Q7	24.192 ± 2.955 (n=59)	36.571 ± 0.170 (n=59)	0.006 ± 0.019 (n=57)	0.023 ± 0.092 (n=57)	0.689 ± 0.302 (n=58)
Q8	20.031 ± 0.731 (n=9)	36.611 ± 0.130 (n=9)	0.013 ± 0.028 (n=9)	0.050 ± 0.095 (n=9)	0.689 ± 0.252 (n=8)

Supplemental Table

Table S1. Samples Sequenced or Analyzed and Presence of Plastid, DPL1 and DPL2

The table shows sample name, number of total amplicons after QC (postQC) and plastid-encoded 16S rRNA gene V1-V2 amplicons, as well as the DPL1, DPL1-related and DPL2 amplicons as percent of all plastid-derived amplicons in the respective sample. BATS data are provided as means ± standard deviation from 53 samples in which >200 plastid-derived 16S amplicons were recovered from the sample (32 samples did not meet this criterion). For BATS, samples are described as T/N where T indicates time adjusted to the month of deep mixing (DM) and N is the number of samples used to compute the means provided. Additional sample information (metadata) is available in the SRA depositions (see Supplemental Experimental Procedures for accessions) and can be downloaded from the Baselines Initiative website, at http://www.mbari.org/resources-worden-lab. Results from analyzing Tara Oceans data [S10] is shown for total 16S miTAGs (from the photic-zone) and the number of miTAGs fragments representing DPL1 and DPL2 are denoted. Note that the Tara Oceans samples were prefiltered through either a 3 or 1.6 μm pore-sized filter (as indicated below), whereas our (Baselines) samples were not prefiltered (with the exception of the first three samples listed for CN207 in Table S1). -, not detected.

Samples	Amplicons (postQC)	Plastid amplicons	%DPL1 clusters	%DPL1-related clusters	%DPL2
Eastern North Pacific Line-67					
CN207_155surf08A	12979	1368	-	-	0.66
CN207_155surf08B	10478	1431	-	-	0.14
CN207_155surf30	4072	1642	1.77	2.68	1.10
WFAD09_H3	211029	25758	-	0.02	0.05
WFAD09_70N10	162268	16671	0.62	0.25	0.29
WFAD09_155N10	174626	3383	0.71	0.59	0.27
Indian Ocean					
INDIOCE_S1	126880	6191	-	-	-
INDIOCE_S2	127622	3589	0.06	-	0.11
INDIOCE_S3	91890	2121	0.14	0.14	0.28
INDIOCE_S4	89713	3868	-	-	0.10
INDIOCE_S5	98316	2100	0.05	0.14	0.43
INDIOCE_S6	105523	3051	0.46	0.10	0.43
INDIOCE_S7	146981	3545	0.51	0.03	0.31
INDIOCE_S8	178588	4583	0.31	0.20	0.24
INDIOCE_S9	103341	2725	0.18	0.44	0.37
INDIOCE_S10	246524	2228	1.80	0.49	1.17
INDIOCE_S11	196217	2470	0.08	0.12	0.40
INDIOCE_S12	125041	2825	-	0.07	0.28
INDIOCE_S13	179843	1880	0.37	0.21	0.27
INDIOCE_S14	39954	539	-	0.19	0.19
INDIOCE_S15A	233372	2168	0.05	0.05	0.09
INDIOCE_S15B	344900	3104	0.06	0.19	0.45
INDIOCE_S16	165953	2871	0.73	0.07	0.14
INDIOCE_S17	233719	2871	0.59	0.03	0.66
INDIOCE_S18	202448	2228	0.63	0.13	0.72
INDIOCE_B1	189754	2536	0.35	0.16	0.67
INDIOCE_B2	84724	802	0.25	-	0.12
INDIOCE_B3	181522	1921	0.10	-	0.36
INDIOCE_B4	220377	3782	0.13	0.03	0.29
INDIOCE_B5	237490	2234	1.48	0.18	0.81
INDIOCE_B6	279177	4006	0.10	0.17	0.52
INDIOCE_B7	121052	1603	0.19	0.25	0.31
INDIOCE_B8	230744	5264	0.27	0.57	0.59
INDIOCE_B9	91022	1221	-	0.41	0.16

INDIOCE_B10	230270	5565	0.14	0.47	0.20
INDIOCE_B11	223090	6122	0.33	0.05	0.11
Curacao					
CUR_SP1_1	217435	3349	-	-	0.03
CUR_SP1_2	207949	4549	-	-	-
CUR_MG1	193999	4380	-	-	-
CUR_MG2_2	205398	21084	-	-	0.06
CUR_MG3_1	274347	14989	-	-	-
CUR_MG3_2	218578	25533	-	-	-
CUR_MG3_3	229384	17378	-	-	-
CUR_WF2_C0	168462	3412	0.38	0.15	0.06
CUR_WF2_C1	166470	5058	0.08	-	0.02
CUR_WF2_C2	330924	4228	0.19	0.05	0.09
CUR_EP1_coral	165012	8507	0.09	0.01	0.06
CUR_EP1_surf	213654	12468	-	0.01	0.10
CUR_EP2_C0	218089	4303	0.05	-	0.63
CUR_EP2_C4	187197	9408	-	-	0.03
BATS					
-1/6	8083 ± 4939	1186 ± 1527	-	0.05 ± 0.07	0.11 ± 0.11
DM/7	4400 ± 2704	783 ± 687	0.01 ± 0.02	-	0.07 ± 0.19
+1/7	11242 ± 4320	616 ± 360	0.18 ± 0.31	1.08 ± 0.83	0.66 ± 0.53
+2/4	11296 ± 7000	875 ± 384	0.82 ± 0.49	0.50 ± 0.46	0.62 ± 0.15
+3/3	10298 ± 9287	460 ± 392	0.75 ± 0.25	1.64 ± 0.73	1.74 ± 1.43
+4/4	9016 ± 2631	309 ± 85	1.31 ± 1.18	1.07 ± 0.87	0.60 ± 0.72
+5/4	8054 ± 4299	331 ± 89	0.69 ± 0.80	1.44 ± 1.38	0.67 ± 0.37
+6/3	11495 ± 5498	371 ± 141	1.18 ± 1.15	0.66 ± 0.37	0.63 ± 0.62
+7/5	9677 ± 5387	272 ± 100	0.43 ± 0.61	0.51 ± 0.69	1.05 ± 0.21
+8/3	8725 ± 1081	239 ± 62	0.21 ± 0.37	2.36 ± 4.08	0.93 ± 0.81
+9/4	13205 ± 5225	435 ± 92	0.36 ± 0.71	0.95 ± 1.30	0.38 ± 0.26
+10/3	10912 ± 7547	456 ± 363	0.08 ± 0.13	0.21 ± 0.20	0.44 ± 0.33

Tara Oceans

Samples	16S miTAGs	DPL1	DPL2
TARA_056_SRF_0.22-3	101902	-	4
TARA_057_SRF_0.22-3	100892	-	-
TARA_058_DCM_0.22-3	95349	-	-
TARA_062_SRF_0.22-3	69585	1	1
TARA_064_DCM_0.22-3	161337	1	4
TARA_064_SRF_0.22-3	186898	1	7
TARA_065_DCM_0.22-3	155299	-	3
TARA_065_SRF_0.22-3	61426	-	1
TARA_066_DCM_0.22-3	39410	-	-
TARA_066_SRF_0.22-3	70388	1	-
TARA_067_SRF_0.22-3	40590	-	-
TARA_068_DCM_0.22-3	56160	-	1
TARA_068_SRF_0.22-3	68154	3	6
TARA_070_SRF_0.22-3	40905	2	11
TARA_072_DCM_0.22-3	49909	-	-
TARA_072_SRF_0.22-3	78627	-	4
TARA_076_DCM_0.22-3	68102	-	3
TARA_076_SRF_0.22-3	78051	3	69
TARA_078_DCM_0.22-3	87909	-	1
TARA_078_SRF_0.22-3	78486	-	18
TARA_082_DCM_0.22-3	161515	1	-
TARA_082_SRF_0.22-3	72336	1	-
TARA_084_SRF_0.22-3	134591	-	-
TARA_085_DCM_0.22-3	128279	-	-

TARA_085_SRF_0.22-3	163215	-	-
TARA_093_DCM_0.22-3	186478	-	-
TARA_093_SRF_0.22-3	89154	-	-
TARA_094_SRF_0.22-3	127765	-	-
TARA_096_SRF_0.22-3	136798	1	6
TARA_098_DCM_0.22-3	73002	-	-
TARA_098_SRF_0.22-3	72312	1	5
TARA_099_SRF_0.22-3	96066	-	5
TARA_100_DCM_0.22-3	103040	-	1
TARA_100_SRF_0.22-3	122482	-	-
TARA_102_DCM_0.22-3	112336	-	-
TARA_102_SRF_0.22-3	74287	-	-
TARA_109_DCM_0.22-3	77265	3	4
TARA_109_SRF_0.22-3	128823	2	4
TARA_110_DCM_0.22-3	128677	1	2
TARA_110_SRF_0.22-3	106731	6	2
TARA_111_DCM_0.22-3	112059	-	-
TARA_111_SRF_0.22-3	115161	3	4
TARA_112_DCM_0.22-3	106315	-	1
TARA_112_SRF_0.22-3	119900	2	4
TARA_122_DCM_0.22-3	93757	-	-
TARA_122_SRF_0.22-3	108073	3	-
TARA_123_MIX_0.22-3	125791	-	1
TARA_123_SRF_0.22-3	86324	3	2
TARA_124_MIX_0.22-3	141432	-	1
TARA_124_SRF_0.22-3	144234	1	5
TARA_125_MIX_0.22-3	79206	-	-
TARA_125_SRF_0.22-3	129135	3	1
TARA_128_DCM_0.22-3	76317	-	2
TARA_128_SRF_0.22-3	102780	1	5
TARA_132_DCM_0.22-3	104381	1	-
TARA_132_SRF_0.22-3	106220	3	8
TARA_133_DCM_0.22-3	92255	-	2
TARA_133_SRF_0.22-3	149404	9	1
TARA_137_DCM_0.22-3	118179	-	-
TARA_137_SRF_0.22-3	119711	5	8
TARA_138_DCM_0.22-3	68607	-	1
TARA_138_SRF_0.22-3	111087	6	3
TARA_140_SRF_0.22-3	115773	1	2
TARA_141_SRF_0.22-3	77211	2	-
TARA_142_DCM_0.22-3	100452	1	1
TARA_142_SRF_0.22-3	109445	-	4
TARA_145_SRF_0.22-3	84116	1	-
TARA_146_SRF_0.22-3	118221	-	3
TARA_148_SRF_0.22-3	118053	-	2
TARA_149_SRF_0.22-3	120367	-	8
TARA_150_DCM_0.22-3	118668	2	3
TARA_150_SRF_0.22-3	123658	-	5
TARA_151_DCM_0.22-3	104170	-	3
TARA_151_SRF_0.22-3	122930	-	11
TARA_152_MIX_0.22-3	113269	-	3
TARA_152_SRF_0.22-3	92320	-	4
TARA_004_DCM_0.22-1.6	143169	-	-
TARA_004_SRF_0.22-1.6	91124	-	-
TARA_007_DCM_0.22-1.6	77642	-	-
TARA_007_SRF_0.22-1.6	75826	-	-

TARA_009_DCM_0.22-1.6	96015	-	-
TARA_009_SRF_0.22-1.6	147018	2	2
TARA_018_DCM_0.22-1.6	115928	-	-
TARA_018_SRF_0.22-1.6	133643	-	-
TARA_023_DCM_0.22-1.6	53084	-	-
TARA_023_SRF_0.22-1.6	81189	-	-
TARA_025_DCM_0.22-1.6	96926	-	-
TARA_025_SRF_0.22-1.6	133238	-	1
TARA_030_DCM_0.22-1.6	157853	-	2
TARA_030_SRF_0.22-1.6	95690	-	1
TARA_031_SRF_0.22-1.6	124569	-	-
TARA_032_DCM_0.22-1.6	127264	-	-
TARA_032_SRF_0.22-1.6	86601	-	-
TARA_033_SRF_0.22-1.6	69314	-	-
TARA_034_DCM_0.22-1.6	57414	-	-
TARA_034_SRF_0.22-1.6	64924	-	-
TARA_036_DCM_0.22-1.6	100170	2	-
TARA_036_SRF_0.22-1.6	62530	-	-
TARA_038_DCM_0.22-1.6	100385	2	-
TARA_038_SRF_0.22-1.6	68985	-	-
TARA_039_DCM_0.22-1.6	71132	-	-
TARA_041_DCM_0.22-1.6	76328	-	-
TARA_041_SRF_0.22-1.6	101365	-	-
TARA_042_DCM_0.22-1.6	56690	-	-
TARA_042_SRF_0.22-1.6	97737	-	-
TARA_045_SRF_0.22-1.6	124540	-	-
TARA_048_SRF_0.22-1.6	139279	-	-
TARA_052_DCM_0.22-1.6	96919	-	-
TARA_052_SRF_0.22-1.6	108588	-	-

Supplemental Experimental Procedures and Information

Oceanographic Sampling and Sequence Generation:

Baselines Initiative samples were collected using Niskin bottles mounted on a rosette along with a conductivity, temperature, and depth (CTD) sensor in the eastern North Pacific and Sargasso Sea, as well as most Indian Ocean sites. Eleven of the 29 Indian Ocean samples were collected using a clean bucket due to a cable failure and loss of the rosette in the ocean. Curaçao samples were collected using Niskins (open sea/coral adjacent waters) or Amber bottles (mangroves and saline ponds). Unless otherwise specified, DNA samples were collected by filtering 500 to 2,000 ml seawater through a 0.2 μm pore size membrane filter (Supor 200, Pall Gelman, East Hills, NY, USA); filters were flash-frozen in liquid nitrogen, and transferred to -80°C until DNA extraction according to [S11].

Several 16S rRNA gene datasets were generated as part of the Baselines Initiative (http://www.mbari.org/resources-worden-lab) and analyzed here. The Baselines Foundational Set (near full-length 16S rRNA gene sequences) was generated by PCR from DNA extracted from eastern North Pacific Line-67 samples collected in 2007 (cruise CN207) as described in [S12]. The samples came from 5 to 10 m depth at Station 155 (33.286 Lat, -129.428 Lon; 7 Oct. 2007), Station 70 (36.129 Lat, -123.490 Lon; 9 Oct. 2007) and Station H3 near Monterey Bay (36.740 Lat, -122.020 Lon; 10 Oct. 2007), which represent oligotrophic, mesotrophic/upwelling and coastal environments, respectively. Additionally a sample from the deep chlorophyll maximum at Station 155 (86 m; 6 Oct. 2007) was sequenced. The 16S rRNA gene was amplified from three sequentially-filtered size fractions (3 to <20 μm, 0.8 to <3 μm, 0.1 to <0.8 μm) from each sample using two sets of universal 16S rRNA gene primers: 27F (5′-AGAGTTTGATCMTGGCTCAG-3′) with 1492R (5′-GGTTACCTTGTTACGACTT-3′) and 27F (5′-AGAGTTTGATCMTGGCTCAG-3′) with 1391R (5′-GACGGGCGGTGWGTRCA-3′). In addition for

Station H3 the 0.1 to 20 μm size fraction was sequenced. The resulting products were cloned and a total of 8,823 clones were bidirectionally Sanger sequenced; chimeric sequences were checked with UCHIME against the ChimeraSlayer gold reference database with additional manual inspection [S13]. Mitochondrial sequences were removed (1,149 in total) from the data set prior to deposition, so that a total of 3,223 clones from the 27F/1492R primer set and 2,954 from the 27F/1391R primer set form the Baselines Foundational Set and have been deposited under GenBank accessions KX932494 to KX938325.

The other Baselines Initiative datasets (http://www.mbari.org/resources-worden-lab, Baselines Amplicon Release 1 (R1)) involved 16S rRNA V1-V2 amplicon sequencing from products generated using the primers 27FB (5′-AGRGTTYGATYMTGGCTCAG-3′) and 338RPL (5′-GCWGCCWCCCGTAGGWGT-3′) as in [S14]. These were sequenced on two high-throughput platforms. Two sets were sequenced using the 454-platform: Amplicons from the 2007 Line-67 cruise CN207 were sequenced from 5 m at Station 155 (latitude and longitude as above) from the 0.8 to <3 μm (two replicates) and 3 to <20 μm (1 sample) size fractions. Quality control (QC) of 454-pyrosequenced reads was checked using published methods [S15]. This amounted to 27,529 amplicons in total after QC (47,748 amplicons prior to QC) [S16] that have been deposited in the SRA (SRX707412, SRX707413, SRX708079) as well as 662,211 16S V1-V2 amplicons after QC from 85 BATS samples [S14] (latitude ranges from 31.164 to 31.906 and longitude from -64.679 to -63.773) deposited at iMicrobe (http://data.imicrobe.us/project/view/101). Note that BATS collection and extraction methods are described in [S7]). Other Baselines datasets were generated using the Illumina-platform (MiSeq PE). Illumina reads were assessed with FastQC 0.11.4 for the overall read quality [S17], followed by quality processing including score-based trimming using Sickle with default parameters [S18] and then paired-end reads were assembled with PANDAseq using a minimum overlap of 50 bp [S19]. This reduced the number of Illumina sequenced raw reads from 11,414,650 to 8,670,868 amplicon sequences, with the latter representing the final number after quality control and assembly of paired-end reads. The samples came from: a 2009 Line-67 cruise [S16] with sampling at the three above stations (547,923 total amplicons after QC); 29 Indian Ocean sites sampled in 2015 (5,126,047 total amplicons after QC) for which temperature on average was very similar (28.97 ± 0.25^{o}C); and from near/in Curaçao sampled in 2015 (2,996,898 total amplicons after QC) deposited in the SRA (SRP090781 to SRP090783). Latitude and longitude are provided in the deposition files for each sample analyzed here and at http://ocean-microbiome.embl.de/companion.html (Table W1) for the Tara Oceans samples we reanalyzed (see below).

Nine DPL1 (accessions KX938311 to KX938319) and one DPL2 (accession KX935025) near full-length 16S rRNA gene sequences were recovered from Baselines Foundational Set samples along with 27 sequences from DPL1-related clusters A, B, H and I (accessions KX934896, KX937741 to KX937744, KX938218 to KX938235 and KX938320 to KX938323). Additional DPL1 near full-length 16S plastid sequences (accessions KX938326 to KX938329) were generated by primer walking after preliminary analysis of short read datasets (which identified a cluster of sequences, now DPL1, without clear taxonomical attribution). A forward primer (N99F1:5′-CCYCGACRAAAGCARMAGATCG-3′) was designed in the V1 region to match with this cluster of sequences placed at a deep node. Sequences were generated by PCR from a DNA sample collected on 14 Sept. 2012 at 10 m (sample C0912: 36.3033 Lat, -122.3828 Lon; 15.32^{o}C, 33.34 ppt, 2.025 μM NO_3^-, 0.669 μM PO_4^{3-}, cruise C0912) using the N99F1 specific primer and 1492R universal 16S primer (as above), cloned into pCRII-TOPO and Sanger sequenced bidirectionally using plasmid primers M13F and M13R. Then, a reverse primer was designed near the end of the 16S (N99R1:5′-TTCTCCGAATCACGAAGACGC-3′), 100 bp downstream of the 1492R primer (above). Products were then generated by PCR using 27F (as above) and the N99R1 specific primer, cloned into pCRII-TOPO and Sanger sequenced bidirectionally using plasmid primers M13F and M13R. In total four high-quality full length sequences were generated (C0912-1, C0912-2, C0912-3 and C0912-4) and these were 97% identical to the four DPL1 16S rRNA gene sequences from the Line-67 Station 155 data in the Baselines Foundational Set (sequences CN207-155S3Ac8La, CN207-155S3Ac8KY, CN207-155S3Ac8IP and CN207-155S3Ac8JP) confirming these divergent plastid sequences in samples

collected at different geographical locations and in different years. Similar sequences were not found in GenBank, and best blastn hits (~85% nucleotide identity) were to environmental clones termed "uncultured cyanobacteria", "uncultured bacteria" or "environmental eukaryotic plastids".

Nutrient Measurements:
For Line-67 samples nutrients were quantified as in [S20], while for BATS the methods are described in [S5] and the data for BATS has been published in [S6]. Samples from the Bay of Bengal in the northern Indian Ocean were measured following [S21] using an Autoanalyzer (San Plus, SKALAR Instrumentation, The Netherlands).

Phyloassigner 16S rRNA Alignments, Amplicon Placement and Additional 16S V1-V2 Amplicon Analyses:
16S rRNA V1-V2 region amplicons were initially parsed phylogenetically using PhyloAssigner version089 [S14], which performs a profile alignment of amplicons to a multiple sequence alignment using HMMER [S22] and assigns phylogenetic positions in an unmasked reference tree based on maximum likelihood methods using pplacer [S23]. The 9,360,608 amplicons that passed quality control were run against a reference alignment and tree constructed from near full-length bacterial 16S rRNA sequences as well as a subset of eukaryotic plastid rRNA gene sequences [S14]. The 3,296,582 amplicons assigned to the cyanobacterial/plastid region of the tree were then retrieved for further analysis as below.

For accurate placement of these amplicons, we first constructed a specialized alignment optimized for plastid sequence placement with the most up-to-date full-length reference data possible by retrieving all 16S rRNA gene sequences from the SILVA database (release 119, July 14, 2014, http://www.arb-silva.de/documentation/release-119/) [S24]. This database contained 4,346,367 entries (chimera pre-screened) and highly identical sequences (99% identity criterion) had been removed using UCLUST, resulting in a total of 534,968 representative rRNA gene sequences termed the “non-redundant (NR) SSU Ref dataset”. Cyanobacteria and plastid clusters were selected from the NR SSU Ref dataset based using the ARB software package (http://www.arb-home.de), resulting in a total of 9,945 sequences to which we added 197 unique sequences from the Baselines Foundational Set (near full-length 16S rRNA gene sequences). Due to the small number of dinoflagellate 16S rRNA gene sequences in SILVA we also added dinoflagellate 16S rRNA retrieved from MMETSP datasets [S25]. These dinoflagellate 16S rRNA gene sequences were initially aligned to the reference alignment with the SINA aligner, available through the SILVA website (http://www.arb-silva.de/aligner/), followed by further screening for chimeric or short sequences by viewing placement on the total 16S rRNA tree and subsequent modification of the alignment as needed. The remaining high-quality near full-length 16S rRNA sequences (691 in total) were used to construct our global cyanobacterial and plastid phylogenetic tree using Maximum Likelihood inference under the gamma corrected GTR model of evolution as implemented in FastTree 2.1 [S26] from masked alignments. The mask was made in ARB software by removing non-homologous and ambiguously aligned positions, particularly those at the ends of hairpin turns. FastTree used the SH test with 1,000 bootstrap replicates to estimate the confidence in branching and accuracy of the phylogenetic inferences estimated by FastTree were assessed using a Maximum Likelihood algorithm implemented in RAxML assuming the GTR+Γ substitution model with 1,000 bootstrap replicates (-m GTRGAMMA -f a -# 1,000 parameters). Both approaches yielded consistent results with the same topologies. The accuracy of amplicon placement was verified for our global cyanobacterial and plastid reference tree and alignment by running a test set of 50 16S rRNA gene sequences derived from characterized plastid sequences (not present in our alignment) that were trimmed to the size of the V1-V2 amplicons. Finally, PhyloAssigner was run using the global cyanobacterial and plastid reference tree to bin environmental amplicons into either plastids or cyanobacteria, and then retrieving those assigned to DPL1 and DPL2 clusters for further analyses. The samples with highest overall DPL contributions to plastid-derived amplicons were collected at BATS on 15 June 1993 (sample ID s57d0; 9.7% DPL1) and 7 July 1998 (s118d0; 3.6%).

We applied a number of criteria to the data before computing averages for DPL amplicon data from various samples. We performed UCLUST clustering on all amplicons assigned by PhyloAssigner to the DPL1 and DPL2 nodes and further examined them using additional phylogenetic placement tests. Representative amplicons from these UCLUST clusters (99% nt identity cutoff) were added into the alignment used for Figure 1A and aligned using MAFFT with default parameters. Positions with gaps were masked using Gblocks and phylogenetic inferences were done in RAxML [S2], PhyML [S3] and MrBayes [S27], and the positions of amplicons were examined. Only if amplicons were placed and grouped either in the DPL1 or DPL2 branches with bootstrap support and had >87% nt identity were the amplicons considered DPL1 or DPL2. Note that for Figure 1D bar graphs only samples with >200 plastid-derived amplicons were included in the analysis and DPL1 represents the sum of DPL1 clusters C, D and E. For Supplemental Figure S1E and S1F, only BATS samples with >2,000 total amplicons (to avoid potential detection limit issues for the four groups analyzed) were included in the analysis.

Phylogenetic Reconstructions for Evolutionary Analyses of DPL and Other DPL Sequence Analyses:
For evolutionary analyses complete 16S rRNA gene sequences were aligned using MAFFT [S28] with default values. The alignment incorporated sequences from representative taxa from across the plastid tree (128 sequences, including 5 fast-evolving dinoflagellate sequences) and a subset of cyanobacteria (13 sequences) along with 10 DPL sequences. The dataset used for Figure 1A did not include the fast-evolving dinoflagellate sequences (they were removed prior to performing the reconstruction) and the final phylogenetic reconstruction was based on analysis of 1,164 homologous positions. Phylogenetic reconstructions were performed in MrBayes 3.2.6 [S27] with the parameters lset nst=6 rates=invgamma ncat=6, and ngenval =10,000,000 samplefreqval =1,000 and tempval =0.200. In RAxML 8 [S2] the parameters were -m GTRGAMMA -f a -# 1,000 and for PhyML 3.0.1 [S3] the parameters were -m GTR -f e -v e -c 8 -a e -b 100 -s BEST --n_rand_starts 10. Final trees figures were produced with assistance from FigTree (Rambaut, A., 2012, version 1.4. 0, http://tree.bio.ed.ac.uk/software/figtree) and topologies from the Bayesian reconstruction (Figure 1A). Taxon sampling for known eukaryotic groups was similar to [S1] and red algae were paraphyletic (as in [S1] as well).

To test whether the new plastid-bearing eukaryote sequences were from described dinoflagellates that acquired their plastid from a haptophyte alga, we also included the fast-evolving sequences of dinoflagellate plastids (*Gymnodinium galatheanum* AF172716, *Gyrodinium aureolum* AF172717, *Karenia brevis* EU043083, *Karenia mikimotoi* EU0431111 and *Karlodinium veneficum* JN039300) to the reference alignment (Figure S1A, B). Using this alignment, we also included amplicon sequence representatives from identified DPL1 and DPL1-related clusters (A-J) to test how they related to one another and to taxa for which near full-length 16S rRNA gene sequences were available (Figure S1C). This phylogeny was inferred with the Bayesian methods described above. As the eleven representative amplicon reads were only the V1-V2 region, the other positions (across the full-length alignment) were considered missing characters (i.e. did not contribute phylogenetic information) for these sequences.

Secondary structure predictions for the DPL 16S rRNA gene sequences were examined for all DPL1 and DPL1-related clusters with assistance from Mfold [S27] and from comparisons to published secondary structures. These analyses highlighted within DPL1 diversity, especially in extensions within the 16S rRNA V2 variable region (DPL1 clusters C, D, and E), and phylogenetically-related novel amplicons that formed the DPL1-related clusters (Figure S1C) which either have additional extensions (DPL1-related clusters H, I, and J) or lack them entirely (DPL1-related clusters A, B, F, and G) (e.g., Figure S1D).

Other Data Retrieval and Analyses:
Tara Oceans contextual data and taxonomic profiling based on assembled metagenomic Illumina reads (miTAGs) that had signatures of the 16S rRNA gene were retrieved from the website http://ocean-microbiome.embl.de/companion.html [S10]. Blastn searches against all 16S miTAGs was performed using our DPL2 (e.g., CN207-155S3Ab030) and DPL1 (e.g., C0912-2) reference sequences with cutoff criteria of

e-value $<1 \times 10^{-80}$ and a maximum of three mismatches. This stringency was used to avoid any possible over assignment with potentially misassembled miTAGs. Additionally, we retrieved all DPL2 sequences assigned to DPL2 sequence EF574856 from near Cocos Island (GOS site 25) in the Tara Oceans USEARCH cluster centroids. Note that the Tara Oceans samples were pre-filtered through either a 3 or 1.6 µm pore-sized filter (as indicated on Table S1), whereas our (Baselines) samples were not pre-filtered (with the exception of the specific Station 155 sample above). These differences in procedures likely influence the low numbers of DPL recovered from Tara Oceans miTAG data based on proposed size classes for DPL. We also analyzed Tara samples from deeper in the water column (250 m to 1000 m, 0.22 to <3 µm size-fractionated) and out of 26 such samples found only 1 DPL1 sequence in each of two samples (TARA_110_MES_0.22-3; TARA_056_MES_0.22-3) and 2 DPL2 samples in one sample (TARA_152_MES_0.22-3). This is 7.7% (DPL1) and 3.8% (DPL2) of the deep samples as opposed to 41.6% (DPL1) and 64.9% (DPL2) of 0.22 to <3 µm size-fractionated photic-zone samples. The fact that fewer sequences were found in deep ocean samples is supportive of the idea that DPL sequences come from photosynthetic taxa which would live in the surface ocean. It should be noted however that apochlorotic phytoplankton lineages exist that have lost the ability to perform photosynthesis, e.g. some diatoms and green algae as well as dinoflagellates, and tend to have long 16S rRNA gene branch lengths [S29, S30].

Statistical analyses:
Pairwise comparisons were performed using T-tests in SigmaStat Version 13 (Systat Software House, San Jose, CA 95131 USA). For tests on differences in relative abundance of DPL1 and DPL2 (among plastid amplicons; Figure 1D) in winter versus stratified periods, samples were divided as having a winter/deep mixing signal when temperature was $\leq 21^{o}$C, or stratified when $>21^{o}$C, based on information in [S9]. Thus, the tests were not applied directly by the month designation and only samples with >200 plastid amplicons were analyzed statistically (also the criteria for inclusion in Figure 1D). For the winter period samples came from +9 (1 sample), +10 (1 sample), -1 (5 samples), 0 (DM, 7 samples), +1 (7 samples), +2 (one sample) and thirty other samples belonged to the stratified period, for which there were no samples from the -1, 0 or +1 months. A Mann-Whitney test implemented in GraphPad Prism Version 6.0 (GraphPad Software, Inc., La Jolla, CA, USA) was used to test significance of differences in relative DPL abundances between the two period types.

Supplemental References

S1. Janouskovec, J., Horak, A., Barott, K.L., Rohwer, F.L., and Keeling, P.J. (2012). Global analysis of plastid diversity reveals apicomplexan-related lineages in coral reefs. Curr. Biol. *22*, R518-519.

S2. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312-1313.

S3. Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. *59*, 307-321.

S4. Yarza, P., Yilmaz, P., Pruesse, E., Glockner, F.O., Ludwig, W., Schleifer, K.H., Whitman, W.B., Euzeby, J., Amann, R., and Rossello-Mora, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat. Rev. Microbiol. *12*, 635-645.

S5. Carlson, C.A., Morris, R., Parsons, R., Treusch, A.H., Giovannoni, S.J., and Vergin, K. (2009). Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. ISME J. *3*, 283-295.

S6. Treusch, A.H., Demir-Hilton, E., Vergin, K.L., Worden, A.Z., Carlson, C.A., Donatz, M.G., Burton, R.M., and Giovannoni, S.J. (2012). Phytoplankton distribution patterns in the northwestern Sargasso Sea revealed by small subunit rRNA genes from plastids. ISME J. *6*, 481-492.

S7. Treusch, A.H., Vergin, K.L., Finlay, L.A., Donatz, M.G., Burton, R.M., Carlson, C.A., and Giovannoni, S.J. (2009). Seasonality and vertical structure of microbial communities in an ocean gyre. ISME J. *3*, 1148-1163.

S8. Lomas, M.W., Steinberg, D.K., Dickey, T., Carlson, C.A., Nelson, N.B., Condon, R.H., and Bates, N.R. (2010). Increased ocean carbon export in the Sargasso Sea linked to climate variability is countered by its enhanced mesopelagic attenuation. Biogeosciences *7*, 57-70.

S9. Steinberg, D.K., Carlson, C.A., Bates, N.R., Johnson, R.J., Michaels, A.F., and Knap, A.H. (2001). Overview of the US JGOFS Bermuda Atlantic Time-series Study (BATS): a decade-scale look at ocean biology and biogeochemistry. Deep Sea Res. Part II Top. Stud. Oceanogr. *48*, 1405-1447.

S10. Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., et al. (2015). Ocean plankton. Structure and function of the global ocean microbiome. Science *348*, 1261359.

S11. Demir-Hilton, E., Sudek, S., Cuvelier, M.L., Gentemann, C., Zehr, J.P., and Worden, A.Z. (2011). Global distribution patterns of distinct clades of the photosynthetic picoeukaryote *Ostreococcus*. ISME J. *5*, 1095-1107.

S12. Monier, A., Welsh, R.M., Gentemann, C., Weinstock, G., Sodergren, E., Armbrust, E.V., Eisen, J.A., and Worden, A.Z. (2012). Phosphate transporters in marine phytoplankton and their viruses: cross-domain commonalities in viral-host gene exchanges. Environ. Microbiol. *14*, 162-176.

S13. Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., and Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. Bioinformatics *27*, 2194-2200.

S14. Vergin, K.L., Beszteri, B., Monier, A., Cameron Thrash, J., Temperton, B., Treusch, A.H., Kilpert, F., Worden, A.Z., and Giovannoni, S.J. (2013). High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. ISME J. *7*, 1322–1332.

S15. Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J., and Knight, R. (2008). Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. Nat. Methods *5*, 235-237.

S16. Sudek, S., Everroad, R.C., Gehman, A.L., Smith, J.M., Poirier, C.L., Chavez, F.P., and Worden, A.Z. (2015). Cyanobacterial distributions along a physico-chemical gradient in the Northeastern Pacific Ocean. Environ. Microbiol. *17*, 3692-3707.

S17. Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. (http://www.bioinformatics.babraham.ac.uk/projects/fastqc).

S18. Joshi, N.A., and Fass, J. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). (https://github.com/najoshi/sickle).

S19. Masella, A.P., Bartram, A.K., Truszkowski, J.M., Brown, D.G., and Neufeld, J.D. (2012). PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics *13*, 31.

S20. Simmons, M.P., Sudek, S., Monier, A., Limardo, A.J., Jimenez, V., Perle, C.R., Elrod, V.A., Pennington, J.T., and Worden, A.Z. (2016). Abundance and biogeography of picoprasinophyte ecotypes and other phytoplankton in the Eastern North Pacific Ocean Appl. Environ. Microbiol. *82*, 1693–1705.

S21. Grashoff, K., Ehrhardt, M., and Kremling, K. (1992). Methods of Seawater Analysis, (New York, NY: Verlag Chemie).

S22. Eddy, S.R. (2011). Accelerated Profile HMM Searches. PLoS Comput. Biol. *7*, e1002195.

S23. Matsen, F.A., Kodner, R.B., and Armbrust, E.V. (2010). pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics *11*, 538.

S24. Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M., Ludwig, W., Peplies, J., and Glockner, F.O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. *35*, 7188-7196.

S25. Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. *12*, e1001889.

S26. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS ONE *5*, e9490.

S27. Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. *61*, 539-542.

S28. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772-780.

S29. Vernon, D., Gutell, R.R., Cannone, J.J., Rumpf, R.W., and Birky, C.W., Jr. (2001). Accelerated evolution of functional plastid rRNA and elongation factor genes due to reduced protein synthetic load after the loss of photosynthesis in the chlorophyte alga *Polytoma*. Mol. Biol. Evol. *18*, 1810-1822.

S30. Kamikawa, R., Yubuki, N., Yoshida, M., Taira, M., Nakamura, N., Ishida, K., Leander, B.S., Miyashita, H., Hashimoto, T., Mayama, S., et al. (2015). Multiple losses of photosynthesis in *Nitzschia* (Bacillariophyceae). Phycol. Res. *63*, 19-28.

Author Contributions

Conceptualization, A.Z.W., C.J.C. and C.B.; Sampling, A.Z.W., C.P., G.S.J., A.M., and S.J.G.; Methodology, S.J.C., A.Z.W., C.J.C. and C.B.; Sample Preparation: L.S.; Analyses: C.J.C., C.B., C.P. and V.V.S.S.; Writing: A.Z.W., C.B. and C.J.C.; Major Review/Editing, S.J.C. and A.M.