doi:10.1038/nature12221

Table of Contents

1.0 Sequencing and Annotation Methods
- 1.1 *Emiliania huxleyi* strain CCMP1516
- 1.2 DNA isolation and library construction
- 1.3 Genome sequencing, filtering and assembly
- 1.4 Estimating genome completeness
- 1.5 cDNA library construction and sequencing
- 1.6 Genome annotation
  - 1.6.1 Filtered models
  - 1.6.2 Refined models
- 1.7 Protein domain analysis

2.0 Genome Structure and Composition
- 2.1 Whole genome tiling array analysis of gene expression
- 2.2 Tag digital gene expression
- 2.3 Gene family analysis
- 2.4 Phylogenomic analysis of *E. huxleyi*
- 2.5 Estimating genome sizes by flow cytometry
- 2.6 Illumina sequencing of genomic DNA from 13 *E. huxleyi* strains
  - 2.6.1 Analysis of Illumina sequenced genomes of *E. huxleyi* by de novo assembly
  - 2.6.2 Analysis of Illumina sequenced genomes of *E. huxleyi* by direct mapping
- 2.7 Comparative genomic hybridization
- 2.8 Concatenated gene phylogenies
- 2.9 Analysis of horizontal gene transfer events
  - 2.9.1 Identification of prokaryotic genes
  - 2.9.2 Identification of putative HGT events between *E. huxleyi* and viruses
- 2.10 Detection of repeats and transposable elements
  - 2.10.1 Analysis of the tandem repeat content in the genome of *E. huxleyi* CCMP1516 in comparison to related organisms

2.10.2 Analysis of transposable elements and other repeats
2.10.3 Analysis of the genomic coverage by tandem repeats and low complexity regions
2.10.4 Analysis of the coverage by tandem repeats and low complexity regions of different genomic regions
2.10.5 GC content in genomic regions and repeats
2.11 Composition of the core and variable genomes
3.0 Ecophysiology
3.1 Phylogenetic and sequence analysis of light harvesting complex genes
3.2 Spider plots
4.0 Supplementary Information Tables
Table SI 1| Genomic libraries included in the *E. huxleyi* genome assembly and their respective assembled sequence coverage levels in the final release
Table SI 2| Summary statistics of the final genome release v1.0.
Table SI 3| GC ranges for partitioning filtered scaffolds
Table SI 4| Partitioning of filtered scaffolds
Table SI 5| Genes in the filtered versus refined models set
Table SI 6| Refined model set classified by gene prediction method
Table SI 7| Properties of refined model set
Table SI 8| Protist and cyanobacterial genomes used for protein domain comparisons with *E. huxleyi*
Table SI 9| *E. huxleyi* protein domains that were absent in 11 other protist and cyanobacterial genomes
Table SI 10| Species included in the phylogenomic analyses
Table SI 11| BLASTn homology statistics between the genomes of the reference strain *E. huxleyi* CCMP1516 and three other deeply sequenced strains using a >90% identity threshold over regions > 100 bp
Table SI 12| BLASTn homology statistics between the genomes of the reference strain *E. huxleyi* CCMP1516 and three other deeply

sequenced strains using a >80% identity threshold over regions > 100 bp

Table SI 13| Sequence from the CCMP1516 reference genome missing from the assemblies of the three deeply sequenced strains

Table SI 14| A Comparison of the amount of strain specific sequence in the *E. huxleyi* pan genome

Table SI 15| BLAST hits between the reduced protein set of the reference genome (CCMP1516) and proteins from 13 other strains of *E. huxleyi*

Table SI 16| Summary of the BLAST results when comparing the reduced protein set of the "type strain" CCMP1516 against the assemblies of genomes from different strains

Table SI 17| The two sets of tandem repeat search parameters used for the TRF program

Table SI 18| List of genomes for which the TR content has been compared to the *E. huxleyi* genome

5.0 Supplementary Information Figures

Figure SI 1| Hierarchical clustering of the *E. huxleyi* gene family expansions.

Figure SI 2| Average gene family size in each species.

Figure SI 3| The distribution of species-specific gene families and orphan genes in a range of genomes.

Figure SI 4 Concatenated phylogeny of 228 *E. huxleyi* genes.

Figure SI 5| The accuracy of flow cytometry for estimating genome sizes.

Figure SI 6| Representation of a 200 kb scaffold (AKBS149348.g2.0 in the reference genome assembly) with information on common regions between other *E. huxleyi* genomes.

Figure SI 7| Comparing the GC and repeat content of the mapped and unmapped Illumina reads from strains B11 and 92D.

Figure SI 8| Genomic coverage of tandem repeats in *E. huxleyi* and other genomes.

Figure SI 9| Genomic densities and mean lengths of individual tandem repeat classes in the *E. huxleyi* genome.

Figure SI 10| Repeat coverage [%] for two different sets of search parameters.

Figure SI 11| Densities and mean lengths of individual tandem repeat classes in intergenic regions and introns.

Figure SI 12| Tandem repeat coverage in different genomic regions.

Figure SI 13| Strandedness of TR patterns and C/G base usage in introns and CDS regions.

Figure SI 14| GC content in different genomic regions and in TRs in different genomic regions.

Figure SI 15| Sequence logo plot of *E. huxleyi* LI818-like promoter cis-acting elements.

Supplementary Information

1.0 Sequencing and Annotation Methods

1.1 *Emiliania huxleyi* strain CCMP1516

The Joint Genome Institute (JGI) used genomic DNA and cDNA from *E. huxleyi* strain CCMP1516 for library construction and sequencing. The diploid strain CCMP1516 was isolated in 1991 from the South Pacific (02.6667S 82.7167W) and was obtained from the National Center for Marine Algae and Microbiota (NCMA, formerly Provasoli-Guillard National Centre for Culture of Marine Phytoplankton CCMP). The strain has been maintained at California State University San Marcos since 1995 where batch cultures are grown photoautotrophically at 17-18°C on a 12 hour light/dark cycle under cool white fluorescent light (660 μmol m^{-2} s^{-1}) in f/50 or f/2 artificial seawater media[1]. To minimize bacterial contamination artificial seawater was supplemented with kanamycin at 100 μg/ml.

1.2 DNA isolation and library construction

Nuclear DNA was isolated using the cetyl-trimethylammonium bromide (CTAB) protocol followed by a cesium chloride density gradient[2]. Cells cultured in f/2 artificial seawater media[1] were harvested by centrifugation and the resulting pellet was resuspended in 9.5 ml of TE Buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). To lyse the cells, 0.5 ml of 10% sodium dodecyl sulfate (SDS) and 50 μl of 20 mg/ml of proteinase K was added, and the sample was incubated at 37°C for one hour. To remove proteins, lipids and polysaccharides, 1.8 ml of 5 M NaCl and 1.5 ml of a CTAB/NaCl (10% CTAB, 0.7 M NaCl) solution was added to the cell lysate, which was mixed thoroughly and incubated at 65°C for 20 minutes. After extraction with an equal volume of chloroform/isoamyl alcohol (24:1), nucleic acids were precipitated by adding 2/3 volume isopropanol. The DNA was then purified on a cesium chloride gradient. Three libraries were constructed; one with 3 kb inserts in a pUC vector, one with 8 kb inserts in the pMCL200 vector, and the third fosmid library was constructed with 20-40 kb inserts in the CopyControl pCC1FOS Vector (Epicentre).

1.3 Genome sequencing, filtering, and assembly

The majority of the sequencing reads were collected with standard Sanger sequencing protocols on ABI 3730XL capillary sequencing at the Department of Energy Joint Genome Institute in Walnut Creek, California (http://www.jgi.doe.gov/sequencing/protocols/prots_production.html) from multiple sized insert libraries (see Table SI1). The assembly of 3,910,095 whole genome shotgun reads was accomplished with Arachne[3] v.20071016 (maxcliq1=150 and BINGE_AND_PURGE=False parameters) using paired-end Sanger sequence reads. After assembly, bacterial filtering was performed using scaffolds and a megablast search for homology against NCBI nt database with the following command-line option:

-D 2 –p 90 –e 1e-50 –z 1e9 –F "m D" –v 100 –b 100

Scaffolds containing mitochondrial and chloroplast sequences were directly screened using the previously sequenced genomes from *E. huxleyi*[4,5]. Otherwise, hits from the megablast run were classified as 1) Eukaryotic-Only, 2) Prokaryotic-Only, 3) Eukaryotic/Prokaryotic Overlap, and 4) Non-cellular; and superimposed on scatter plots of net scaffold length, net scaffold depth, and scaffold GC content. Results revealed sets of low- and high-depth prokaryotic-only scaffolds with the dividing line between them at a mean no-gap scaffold sequence depth greater than 5, along with sets of short and long eukaryotic-only scaffolds, with the dividing line between them at a net length of 20 kb (Table SI 4).

The hits to all of the prokaryotic-only scaffolds with total lengths greater than 100 kb were analyzed. The largest scaffold of the set (3.6 Mb, Scaffold 1) shows significant BLAST similarity to *Erythrobacter*, a common contaminant in *E. huxleyi* cultures. The second scaffold in the set consisted of a single 145 kb contig, and was clearly circular, suggesting that it was a distinct plasmid. The BLAST hits to it mostly involved various *Pseudomonas* species, at % IDs ranging from 92.8%-99.6%, and lengths from 457–467 bases. Little else in the data set hit *Pseudomonas*, hence it was unclear what else might be associated with this plasmid. The low-depth scaffold set had hits almost exclusively to various *Agrobacterium* species, but mainly to *Agrobacterium tumefaciens*. The % IDs of the hits tended to be in the low 90s, and the distribution of hits on the larger scaffolds indicated that these scaffolds did not correspond to an organism already in NCBI. Filtered scaffolds were also classified based on their mean GC contents (Table SI 3). Ranges were chosen based on the distributions of scaffolds with each category of BLAST hit. The results of the classification are provided in Table SI 4.

The high GC content of the *E. huxleyi* genome (65%) and the large amount of repetitive

sequence complicated the sequencing and assembly. As a consequence, 161,432 reads, contigs or scaffolds of less than 3 Kb in length were excluded from the final assembly and treated as a separate database. BLASTn searches indicated that 131,024 of these smaller fragments had significant similarity to portions of the *E. huxleyi* final genome assembly and were considered divergent alleles or multicopy genes.

The final nuclear genome is distributed over 6,995 scaffolds (after removal of alleles, See Refine Models Section 1.6.2 below), and is surprisingly coherent, with 321 large scaffolds harboring 70% of the total sequence (Table SI 2, Supplementary Table 1). The *E. huxleyi* genome moreover, was sequenced to a depth of 10X coverage, and is ~97% complete based on conserved eukaryotic single copy genes[6], and the core eukaryotic genes mapping approach[7] (Supplementary Table 1). Additional assemblies aimed at collapsing haplotypes were performed but did little for the assembly as a whole due to the high heterozygosity rate of the genome. Strategies included: 1) a collapsed assembly similar to the original assembly with slightly different parameter settings, 2) a straight inbred assembly with a binge/purge clean up of repeat content, and 3) a binged/purge assembly. While up to 1 Mb of sequence was collapsed, this occurred at the cost of the collapse of a significant amount of low copy number repeat content.

1.4 Estimating genome completeness

Several criteria were used to examine the completeness of the Sanger sequenced CCMP1516 reference genome including the Conserved Eukaryotic Genes Mapping Approach (CEGMA), the Conserved Eukaryotic Single Copy Genes, Illumina sequencing, and assembly size versus missing sequence. CEGMA[7] was used to identify orthologs in the CCMP1516 genome of the 468 core eukaryotic proteins, and 96.9%, or all but 14 of the gene sequences, were detected using a BlastP threshold of $e < 10^{-5}$ (Supplementary Data File 7). Using a Blast E-value threshold and an alignment coverage >50%, 434 (94.8%) of the CEGMA gene homologs are present in the genome, and of these 94.9% are complete in terms of possessing identifiable start and stop codons. When the larger set of 716 highly conserved single-copy eukaryotic genes[6] were queried, 97% were identified (21 missing), again indicating the genome or the catalog of genes is nearly complete. Scaffold completeness is estimated at 96%, while the ratio of assembly size to genome is 1.02. Together these data indicate although the genome is not finished, it is near completion.

1.5 cDNA library construction and sequencing

Four cDNA libraries corresponding to different developmental stages and growth conditions were used to identify the transcribed regions of the *E. huxleyi* genome. Batch cultures of *E. huxleyi* were grown photoautotrophically at 17-18°C on a 12 hour light/dark cycle under cool white fluorescent light (660 μmol m^{-2} s^{-1}) in f/50 artificial seawater media, and libraries were prepared from RNA extracted from cells 4, 7, 10, and 14 days post-inoculation.

Total RNA for cDNA library construction was isolated as previously described[8] and poly(A) RNA was purified using the Absolutely mRNA Purification Kit (Stratagene, La Jolla, CA). Libraries were prepared using the SuperScript plasmid system with Gateway Technology for cDNA synthesis and cloning (Life Technologies, Carlsbad, CA) whereby 1-2 μg of polyA+ RNA was reverse transcribed with SuperScript II (Life Technologies, Carlsbad, CA) and the oligo dT-*Not*I primer (5'-GACTAGTTCTAGATCGCGAGCGGCCGCCCTTTTTTTTTTTTTTT-3') for first strand cDNA synthesis. RNase H, Polymerase I, T4 DNA Polymerase, and *E. coli* DNA Ligase were used for second strand synthesis. Following ligation of *Sal*I adaptors (5'-TCGACCCACGCGTCCG and 5'- CGGACGCGTGGG) to the cDNA, the ligation products were digested with *Not*I (New England Biolabs, Ipswich, MA) and size selected by agarose gel electrophoresis. After gel purification, cDNA fragments between 0.6-2 kb were ligated into the *Sal*I and *Not*I digested pCMVsport6 vector (Life Technologies, Carlsbad, CA). Ligation products were transformed into ElectroMAX T1DH10B cells.

cDNA libraries were plated at high density (1,000 colonies) onto agar plates and incubated for 18 hours at 37°C. Resulting transformants were picked into 384-well plates containing LB medium supplemented with ampicillin. After 18 hours at 37°C, rolling circle amplification (Templiphi, GE Healthcare, Piscataway, NJ) was used to produce plasmid DNA for sequencing. cDNA inserts were then sequenced on the ABI 3730 (ABI, Foster City, CA) using Big Dye Terminator chemistry with primers complementary to the flanking vector sequence (Fwd: 5'- GTAAAACGACGGCCAGT, Rev: 5' - AGGAAACAGCTATGACCAT).

1.6 Genome annotation

1.6.1 Filtered models

The genome assembly was annotated using the JGI Annotation Pipeline. First the 6,995 *E. huxleyi* CCMP1516 v1.0 scaffolds were masked using RepeatMasker (http://www.repeatmasker.org/) and a custom repeat library of 1,226 putative transposable element-like sequences. Next, 72,513 ESTs from 4 experimental conditions or timepoints were clustered into 21,625 consensus sequences and aligned to the scaffolds with BLAT[9], along with 19,461 unigenes from previous strain CCMP1516 EST sequence analyses[10]. Proteins from the non-redundant (nr) database at the National Center for Biotechnology Information (NCBI; Genbank[11]) were aligned to the genome using BLAST[12].

Gene models were predicted using the following methods: i) *ab initio* methods (FGENESH[13] and Genemark[14]); ii) protein homology-based methods, with models seeded by nr protein alignments (FGENESH+[11] and Genewise[15]; iii) cDNA-based methods, with direct mapping of the EST cluster consensi and unigenes to the genome (Est_map[16]) and iv) a hybrid approach (EuGene[17]). Truncated Genewise models were extended where possible to start and stop codons in the surrounding genome sequence. EST, EST cluster consensus, and unigene alignments were used to extend, verify, and complete the predicted gene models. The multiple sets of models were then filtered based on a scoring scheme which maximizes EST support, sequence similarity to proteins in the nr database, and sequence similarity to other predicted proteins in *E. huxleyi*, to produce a single gene model at each locus. Models with no support were eliminated, as were potential transposable elements, leaving a 'filtered models' (FM) set of 30,569 genes. The genes have a density of 233 per Mb scaffold, and they cover 40% of the genome.

Protein function predictions were made for all predicted gene models using a suite of bioinformatic tools: SignalP[18], TMHMM[19], InterProScan[20], and hardware-accelerated double-affine Smith-Waterman alignments (http://www.timelogic.com/decypher_sw.html) with SwissProt[21], KEGG[22], and KOG[23]. Finally, KEGG hits were used to map the proteins to EC numbers and KEGG pathways, while InterPro, KEGG, and SwissProt hits were used to map proteins to GO terms[24]. For access to each of the KEGG pathways see the *E. huxleyi* JGI portal (http://genome.jgi.doe.gov/cgi-bin/metapathways?db=Emihu1).

1.6.2 Refined models

The *E. huxleyi* CCMP1516 genome is diploid and polymorphic, and so the assembly procedure separated some allelic copies of chromosomal segments and genes from each other. To assess the effect of separated alleles on annotation, we employed the Markov clustering algorithm (MCL) to cluster the FM set, using BLAST alignment scores between proteins as a similarity metric. The FM set clustered into 20,875 'families', including 11,484 singlets (29% of genes) and 7,427 doublets (38% of genes) (Table SI 5). The latter unusually high value is consistent with the separation of some alleles.

To identify separated alleles on the gene level, we aligned the proteins against each other using BLAST, selected best bidirectional hits based on very stringent criteria (identical length and exon number, aa identity > 95%, $e < 10^{-25}$), and looked for at least one neighboring allelic pair as defined by the synteny tool DAGchainer[25]. To identify separate alleles on the chromosomal nucleotide level, we aligned the scaffolds against each other using BLAT and selected stringent alignments (nt identity > 90%, length of alignment either > 90% of smaller scaffold or > 10 kb). These very conservative criteria allowed us to detect 8,557 putatively separated alleles (less conservative criteria captured gene families in addition to separated alleles). Removal from the FM set of one gene model from each allelic pair (always keeping the model on the longer scaffold) produced a 'refined models' (RM) set of 30,569 genes. The RM set clustered into 12,004 'families', including 7,012 singlets (23% of genes) and 4,592 doublets (15% of genes). The latter two values are similar and substantially lower than their corresponding FM values, indicating that we had successfully identified a considerable number of separated alleles.

A high proportion of the RM set is cDNA-based (20%), due to the relatively large number of available ESTs. In contrast, few models are based on protein similarity (12%), likely due to the large phylogenetic distance between *E. huxleyi* and other taxa with sequenced genomes. The remaining predictions are *ab initio* (68%) (Table SI 6). Complete models with start and stop codons comprise 83% of the predicted genes, 51% of the models are consistent with ESTs, 69% align with proteins in the nr database, 77% are members of multi-gene families, and 34% contain Pfam domains[26] (Table SI 7).

The average gene length is 1.8 kb (Table SI 7). A majority of the genes (75%) possess one or more introns, and this group of multi-exonic genes averages 4.5 exons per gene (Table SI 7). Of the many introns, nearly half exhibit a non-canonical GC donor site. Introns also lack a

clear branch point motif and frequently feature 10-11 bp tandem repeat sequences (see Section 2.10.4). In addition, 1,633 alternatively spliced transcripts, 121 tRNAs, and 371,016 small RNAs have been identified including 40 miRNAs.

The average protein length is 346 amino acids. We predicted that 32% of the proteins possess a leader peptide, 16% possess at least one transmembrane domain, and 10% possess both. The RM set was used for most whole-genome expression and comparative analyses (Table SI 7).

Web-based interactive editing tools available through the JGI genome portal (http://jgi.doe.gov/Ehux/) were used to manually curate the automated annotations in three ways: i) to assess, and if necessary correct, predicted gene structures. ii) to assign gene functions and report supporting evidence, and iii) to create, if necessary, new gene structures. As of 6 July 2012, 3,584 genes models were manually curated (~11% of the RM).

1.7 Protein domain analysis

To determine if the *E. huxleyi* gene complement exhibits functionality not found in potentially competing taxa, we compared the protein domains of the 30,569 proteins predicted from the genome assembly and 8,564 proteins predicted from the nonaligned consensi to the domains of 144,123 proteins of 11 other protistan and cyanobacterial genomes (Table SI 8). Pfam domains were found by subjecting the proteins to a hardware-accelerated Hidden Markov Model tool (http://www.timelogic.com/decypher_hmm.html). Out of 2,307 distinct Pfam domains found in 14,486 *E. huxleyi* proteins, 58 domains were absent from all of the 11 other genomes (Table SI 9). These include the Alliinase domain (PF04864) and the Ecotin domain (PF03974). Alliinase is an exclusively plant thiolyase that synthesizes allicin (a thiosulfinate with anti-microbial properties) in garlic. Ecotin is a bacterial inhibitor of serine proteases such as neutrophil elastase; the domain has been found in bacteria and in kinetoplastida (*Leishmania* and *Trypanosoma* sp.) but not in other eukaryotes until now.

2.0 Genome Structure and Composition

2.1 Whole genome tiling array analysis of gene expression

To identify transcriptional units in the *E. huxleyi* genome, high-density oligonucleotide arrays that tile the entire genome were employed and hybridized to RNA prepared from turbidostat grown cells. The Tiling Array consisted of ~2.1 million oligonucleotide probes that were ~50

bases long and spanned the Sanger sequenced strain CCMP1516 genome. RNA samples were obtained from three unialgal turbiostat cultures of *E. huxleyi* strain CCMP1516 run in parallel at the University of Essex. Cultures were grown at 15°C in growth rooms using 0.2 μm filtered artificial seawater supplemented with metals and vitamins to achieve f/2 medium concentrations. Media was prepared with nitrate added to 200 μM and phosphate added to 40 μM. Turbidostats were operated manually whereby dilution flow rates were adjusted daily to achieve a steady biomass with a target cell density of 1,350 cells μl^{-1}. Cultures were incubated under a 14/10 h light/dark cycle at a PDF of 300 μmol photons $m^{-2}s^{-1}$ in 3 L cylindrical culture vessels, illuminated from the side and incubated at 360 p.p.m.v. CO_2.

Sampling was performed on the 10^{th} and 19^{th} day post-inoculation. On these two days samples for RNA extraction were collected every two hours between 11:00 AM-8:00 PM when cells were in the G1 phase (as determined by flow cytometry). RNA from each day was pooled, converted to cDNA, and labeled with Cy5 and Cy3 fluorochromes. Paired analyses performed in triplicate, with dye swapping, were made across days and chemostats. The experimental array data included 60 individual Tiling Array pairs organized into twelve categories based on: 1) Chemostat (I, II, and III), 2) day (S1, S3: day 10, day 19) and 3) Fluorochrome (Cy3, Cy5). Based on pairwise correlations between replicates within each of the twelve categories we found the data to be of high quality with pairwise correlations ranging from 0.88-0.96. To define transcriptional units across the genome, we examined the signal in the promoter region of JGI predicted gene sequences and established a threshold based on the signal intensity below which 95% of the promoter signals fell[27]. When combining all Tiling Array data, the signal intensity of 840,935 of the ~2.1 million probes was found to be greater than that of the threshold value. This suggests ~40% of the genome is expressed. Out of the 30,569 predicted gene models 23,911 or the equivalent of 78.2%, were detected by the Tiling Array. In terms of accuracy, signal was detected in the intron and/or boundary sequence of 15,536 or 50.8% of the predicted gene models. This may be a reflection of a combination of factors including some of the inherent limitations of Tiling Array data, incorrectly annotated sequences, and/or alternative splicing events.

2.2 Tag digital gene expression

Illumina short sequence reads were used to further examine the sensitivity and accuracy of the automated annotation. The same turbidostat cultures of strain CCMP1516 described above were used for this purpose. RNA was extracted (TRIzol) and pooled from a single turbidostat culture, sampled every two hours between 11:00 AM-8:30 PM (when most cells were likely to be G1 phase), after 10 days of turbidostat acclimation. Sequencing libraries were constructed using Illumina's DGE Tag Profiling *Dpn*II Sample Prep Kit (Illumina, Cat. No. FC-102-1007). PolyA RNA was isolated from total RNA using magnetic oligo-dT beads, first strand cDNA was produced from the immobilized PolyA RNA with SuperScript II (Life Technologies, Carlsbad, CA) Reverse Transcriptase, and second-strand cDNA synthesis was accomplished by using DNA Polymerase I. The immobilized cDNA was digested with *Dpn*II and all fragments other than the 3' fragment attached to the bead were washed away. Fragments retained on the beads were ligated to the GEX *Dpn*II Adapter 1. After digesting with *Mme*I 20 bp tags were released, dephosphorylated with CIAP, and ligated to GPX Adapter 2. PCR was then used to amplify the adapter-ligated tags, which were subsequently gel purified and sequenced on the Illumina Genome Analyser using the Genome DNA Sequencing Sample Prep Kit.

After filtering for primers and short or low quality-scores a total of 13,270,000 short sequence reads between 16 and 23 nt in length were obtained and mapped to the CCMP1516 genome using Mapping and Alignment with Quality (MAQ version 0.6.8). Methods described in the online documentation (http://maq.sourceforge.net) were adopted for the alignment, using the default parameter values. Of the filtered reads, 82.5% successfully mapped to the genome. These reads covered 9,003,395 bp of the genome or 6.3% of the non-gapped region of the genome. When used to examine the sensitivity and accuracy of the automated annotation, 96.8% of the 30,569 predicted genes matched at least one of the Illumina short sequence reads. Analysis also identified 14,247 potentially mis-annotated and/or alternatively spliced genes based on reads mapping to introns, intron-exon junctions, and exon boundaries. To identify transcriptional units across the genome, reads within a sliding window of 2,000 bp were computed and compared to the number of reads within annotated gene regions. A high number of mapped reads were detected in 1,700 previously un-annotated intervals within intergenic regions. The number of mapped reads in these regions was greater than the number of reads mapping to 90% of the annotated genes, indicating these may code for novel structural or catalytic RNA transcripts or un-annotated protein coding sequences.

2.3 Gene family analysis

Gene family analysis was performed by comparing the proteins predicted from the genome of *E. huxleyi* with those predicted from the fully sequenced genomes of 20 other species representing a broad phylogenetic distribution. The predicted proteins from the genome sequences of *E. huxleyi* (JGI, v1 reduced set); eight other chromalveolates including *Cryptosporidium hominis* (VCU), *Plasmodium falciparum* (Plasmodb, v5.5), *Theileria annulata* (Sanger), *Paramecium tetraurelia* (Genoscope), *Phaeodactylum tricornutum* (JGI, v2), *Thalassiosira pseudonana* (JGI, v3), *Phytophthora sojae* (JGI, v1.1) and *Aureococcus anophagefferens* (JGI, v1); one red algae, *Cyanidioschyzon merolae* (University of Tokyo, Release Jan 2008); one plant, *Arabidopsis thaliana* (TAIR8); three green algae, *Chlamydomonas reinhardtii* (JGI, v3), *Micromonas* sp. RCC299 (JGI, v2) and *Ostreococcus tauri* (JGI, v2); one yeast, *Schizosaccharomyces pombe* (GeneDB); two metazoans, *Caenorhabditis elegans* (WormBase, v190) and *Drosophila melanogaster* (Ensembl, Release50); one free living amoeba flagellate, *Naegleria gruberi* (JGI, v1); and two cyanobacteria, *Synechococcus* WH8102 (JGI) and *Prochlorococcus marinus* SS120 (CCMP1375) (RefSeq) were downloaded. All-against-all BLASTp (e-value 1E-5) homology searches with all proteins were performed to delineate gene families, which were then clustered using the TribeMCL (inflation value: 2) algorithm[28]. Mean gene family size and the standard deviation for all gene families (orphans and species-specific gene families excluded) with at least one member in *E. huxleyi* and one other photosynthetic species (i.e. the two diatoms, the red alga, the "brown tide" alga, the three green algae, the plant and the two cyanobacteria), were calculated, and the protein phylogenetic profile was transformed into a z-score matrix. The top 100 families with the highest z-score in *E. huxleyi* were extracted and hierarchically clustered with a Pearson correlation as a distance measure using MeV (Figure SI 1).

We identified 1,527 gene families in *E. huxleyi* containing two or more members, and 123 gene families with ten or more members. Of the genomes analyzed, *E. huxleyi* was found to have the third largest average gene family size (Figure SI 2) next to *A. thaliana* and *P. tetraurelia,* both of which are known to have undergone whole-genome duplications and have a history of gene family expansion. Expanded gene families in *E. huxleyi* contain some of the expected protein kinases including tyrosine and serine/threonine kinases, but also genes involved in iron/macromolecular transport, post-translational modification, cytoskeletal development, and

nucleic acid metabolism. Large families of ammonium transporters, ATPase components of ABC transporters, UDP-N-acetylglucosamine, nucleotide-sugar transporters, triose-phosphate transporter, and nitrate/nitrite transporters are present, in addition to large families of genes coding for proteins involved in the development of the cytoskeleton such as alpha-tubulin suppressors, actin-binding proteins, myosin heavy chain, actin, tubulin, and dynein. Proteins with diverse functions in the ubiquitin system or post-translational modification make up some of the largest expanded gene families and include the ATP-dependent 26S proteasome regulatory subunit, FKBP-type peptidyl-prolyl cis-trans isomerase, ubiquitin, ubiquitin conjugating enzyme E2, RING zinc finger domain proteins, and both the N- terminal BTB (Broad-Complex/Tramtrack/Bric-a-brac) substrate recognition and the C-terminal catalytic HECT domains involved in E3-ubiquitin ligase interactions. The overrepresentation of these gene families suggests proteasome-mediated degradation is a more important mode of regulation in *E. huxleyi* than it is in other unicellular algae. The RING finger and the BTB gene families are the second and third largest gene families in *E. huxleyi*, each with over 150 members.

Functional specialization in *E. huxleyi* is also implied by the large number of species-specific genes (Figure SI 3), and the expanded gene families for a variety of different proteins associated with nucleic acid metabolism and protein-protein interactions. Numerous paralogs of RAP domain containing proteins, ribonuclease inhibitors with leucine rich repeats, and the Superfamily II DNA/RNA helicases are present. While proteins with RING zinc finger or BTB domains have been implicated in binding ubiquitination enzymes, they are also capable of binding to DNA and RNA and represent a large class of transcription factors. Both have functional roles in gene transcription, translation, mRNA trafficking, chromatin remodeling, cytoskeletal organization, and protein folding. The large number (89) of RAP domain containing paralogs in *E. huxleyi* is also noteworthy as these proteins are particularly abundant in apicomplexans and are involved in diverse RNA-binding activities. While numerous families of proteins are shared across all kingdoms including structural proteins, transcription factors, and enzymes and proteins of unknown function, members of particular families of genes have undergone substantial increases and decreases in size in *E. huxleyi.* In fact the five largest gene families, which include the RING zinc finger proteins, BTB and RAP domain containing proteins, the leucine-rich repeat ribonucelases, and a protein of unknown function, distinguish *E. huxleyi* from other phytoplankton species.

2.4 Phylogenomic analysis of *E. huxleyi*

The maximum-likelihood tree of main eukaryotic lineages (Figure 1b) was reconstructed based on a concatenated alignment of 15 nuclear-encoded proteins originally published by Parfrey *et al*.[29] and modified by Nozaki *et al*.[30] to exclude intracellular endoparasite apicomplexans (available at TreeBase, identifier: 13511; http://www.treebase.org). Homologous protein sequences from *E. huxleyi* CCMP1516 genome project and the picoprymnesiophyte targeted metagenome[31] were identified through BLASTp searches[32] and aligned to the concatenated alignment with MAFFT v7[33] (without altering the original alignment). The most likely tree was identified from 100 maximum-likelihood tree reconstructions performed with RAxML v7.2.8[34]. The model of evolution best-fitting our data, LG+G+I, was selected with ModelGenerator v0.85[35] based on the Akaike information criterion. Robustness of phylogenetic relationships was tested with 100 standard bootstrap trees generated with RAxML v7.2.8. Circular tree figure was edited from a scalar vector graphic file generated via iTOL[36]. In this tree, *E. huxleyi* appears as the sister taxon to plastid genes from red algae, indicating a predominantly plastid signal present within this gene set. The clade comprising *E. huxleyi* and the red algal plastid genes is itself sister to a clade consisting of photosynthetic and non-photosynthetic plastids.

Concatenated phylogenetic analysis can provide increased resolution over individual gene datasets but may also mask the presence of multiple valid, but discordant, evolutionary histories that may result from genomic mosaicism (as found in organisms with complex plastids) or from a complex history of gene duplication and loss. To address this, we undertook an analysis that aggregated data from multiple single-gene trees. Predicted proteins from complete genomes and EST datasets of 62 genomes (nuclear and organellar; note, some genes have undergone transfer from oganellar genomes to the nucleus in certain taxa) representing 49 organisms (Table SI 10) were downloaded from JGI, NCBI, and EuPathDB (http://eupathdb.org/eupathdb/). ESTs were translated in all six frames using Transeq from the EMBOSS package (http://emboss.sourceforge.net/). Best pairwise reciprocal BLAST[12] hits were collected from predicted proteins of complete genomes and translated ESTs using an e-value cutoff of 10^{-10} and used to assemble COGs[37]. COGs containing at least 10 taxa were collected and their sequences aligned using MUSCLE with default settings[38]. The selection of protein sequences depended only on the quality of the blast hits and their incorporation into COGs. A

total of 1,563 proteins satisfied these criteria; information on these sequences, including the JGI protein ID, KEGG annotations, bootstrap support for the smallest bipartition into which *E. huxleyi* fell, and the organisms in the bipartition can be found in Supplementary Data File 4. Because these datasets were not filtered for specific taxon representation or gene copy number, the number of taxa varies from alignment to alignment. Phylogenetic trees were found from these alignments using RaxML with an LG amino acid replacement matrix and gamma distributed rates of change for 100 bootstrap replicates per dataset[29]. This method permits the analysis of an unusually heterogeneous dataset, albeit at the expense of the high bootstrap values that can be obtained from very long concatenated alignments. To summarize the varying phylogenetic history of genes within the *E. huxleyi* genome, bipartitions exceeding a given bootstrap value were counted from each tree in which *E. huxleyi* appeared. This approach allows taxon bipartitions to be counted appropriately even when the sequence representation within the clade varies.

Of the 1563 individual protein trees from which counts were made, 500 resolved the relationship between *E. huxleyi* and a sister taxon with bootstrap support of 50% or better, and 240 resolved an *E. huxleyi* relationship with sister group with 70% or better support (the remaining trees with insufficient bootstrap support included proteins that were either too short, too highly conserved, or excessively variable). Regardless of the level of bootstrap support, the frequency with which *E. huxleyi* fell with specific groups remained similar (Supplementary Figure 1). The three most frequent sister groups to *E. huxleyi* were (in declining order) the heterokonts, the green lineage, and the alveolates. The affinity with the heterokonts (and alveolates) presumably reflects the morphological and biochemical similarities shared by these groups, quite possibly because of a common shared ancestor.

The frequency with which *E. huxleyi* appears sister to the green lineage may reflect a prior endosymbiotic event or a green algal ancestor as proposed by Moustafa and coworkers[39], but this has been questioned by Burki et al.[40], who found that the putative green signal in diatom genomes was more likely attributable to a general plastid signal and the relative paucity of red lineage data when compared to the green lineage, as also inferred in an analysis of a wild picoprymnesiophyte targeted metagenome[31]. In this context, the strong green signal in the present data likely represents plastid-associated genes. Although taxon bipartitions specifically grouping *E. huxleyi* with red algae and cryptomonads were less frequent than those placing them

with greens, these were both relatively common placements, and likely represents bias introduced by the much larger green-lineage database. It is noteworthy that among those trees with 70% or higher bootstrap support for the placement of *E. huxleyi*, the frequency of trees supporting a placement with heterokonts increases, while that for green and red algae (i.e., primary plastid lineages) decreases.

To further interrogate the phylogenetic signal in our dataset, a second, large-scale concatenated analysis was performed that made use of the single-gene datasets. Among the 1563 single-gene alignments, there were 228 that shared at least 70% of taxa with at least one other alignment in the group. These were concatenated with GeniousPro v.6.0.5. The resulting alignment consisted of 271694 aligned characters, with 57% missing data (total gaps/total characters for all taxa in all alignments). Trees were inferred using RaxML-HPC v.7.2.8, with the LG model and Prot-CAT for protein evolution. Bootstrap values obtained after 100 runs of the rapid bootstrap algorithm (-f a) were mapped on the best-known likelihood tree found after 10 rapid hill-climbing algorithm searches (-f d) (Figure SI 4). In contrast to the robust support observed for various well-established eukaryotic groupings, features of the topology relating to the larger placement of *E. huxleyi* within the tree were not well resolved, beyond support for a relationship with the cryptophytes, consistent with the multiple statistically supported relationships in the individual gene analyses above. Taken together these data reinforce the chimeric nature of the genome, with contributions from the host lineage and the eukaryotic endosymbiont as well as from the plastid and the mitochondrion.

2.5 Estimating genome size by flow cytometry

Flow cytometry of the nuclear DNA content of the different strains was performed at the Flow Cytometry Core Laboratory at the Benorya Research Institute (Seattle, WA) using a modified method from Arumuganathan and Earle[41]. Briefly, the procedure consists of preparing suspensions of intact nuclei in $MgSO_4$ buffer mixed with DNA standards and stained with propidium iodide (PI) in a solution containing DNase-free RNase. Fluorescence intensities of the stained nuclei are measured by a flow cytometer. Values for nuclear DNA content are estimated by comparing fluorescence intensities of the nuclei of the test population with those of an appropriate internal DNA standard that is included with the sample being tested. To establish linearity of the flow cytometer chicken erythrocyte nuclei (CEN, Biosure, Grass Valley, CA)

with single nuclei and doublet and triplet peaks were employed. Nuclei from chicken red blood cells (2.5 pg/2C), *Glycine max* (2.45 pg/2C), *Oryza sativa* cv. *Nipponbare* (0.96 pg/2C), *Arabidopsis thaliana* (0.36 pg/2C) were used as internal standards and subjected to the same staining protocol as *E. huxleyi* cultures.

Specifically for flow cytometric analysis, 1 mL of fresh algal culture was placed in microfuge tubes and centrifuged for 5 s at 5,000 xg at room temperature. The pellet was suspended by vortexing vigorously in 0.5 mL solution containing 10 mM $MgSO_4 \bullet 7H_2O$, 50 mM KCl, 5 mM Hepes, pH 8.0, 3 mM dithiothreitol, 0.1 mg/mL PI, 1.5 mg/mL DNase-free RNase (Roche, Branchburg, NJ) and 0.25% Triton X-100. The suspended nuclei were withdrawn using a pipettor, filtered through 30-µm nylon mesh, and incubated at 37 °C for 30 min before flow cytometric analysis. Suspensions of sample nuclei were spiked with a suspension of standard nuclei (prepared in above solution) and analyzed with a FACScalibur flow cytometer (Becton-Dickinson, San Jose, CA). For each measurement, the PI fluorescence area signals (FL2-A) from 10,000 nuclei were collected and analyzed by CellQuest software (Becton-Dickinson, San Jose, CA) on a Macintosh computer. The mean position of the G0/G1 (nuclei) peak of the sample and the appropriate internal standard were determined by CellQuest software. The mean nuclear DNA content of each algal sample, measured in picograms, was based on 10,000 scanned nuclei. Samples were run in triplicate (at the very least), and the average DNA content was used to measure genome size. The within strain variation in genome size estimates ranged from 0-8% with an average of 4%, and is shown in Supplementary Table 6.

To gauge the accuracy of the methods used, cultures of three marine alga with sequenced genomes (*Guiardia theta*, *Bigelowiella natans,* and *Thalassiosira pseudonana*), and one cilated paramecium whose genome size had previously been estimated by pulse gel electrophoresis were sent to the Benorya Research Institute for independent genome size estimations. Flow cytometry-based estimates positively correlated with genome size determined by direct sequencing (R^2= 0.86), and with sizes determine by direct sequencing and/or pulse gel electrophoresis genome size estimates R^2= 0.81, Figure SI 5). These results suggest genome sizes estimated by flow cytometry are reliable, and hence the observed differences in the genome size across *E. huxleyi* strains are genuine differences and cannot be attributed to technical error.

2.6 Illumina sequencing of genomic DNA from 13 *E. huxleyi* strains

Genomic DNA was isolated from 13 strains of *E. huxleyi* collected from diverse biogeographical locations using Qiagen DNeasy Plant Mini Kit (Qiagen, Florence, CA). Illumina sequencing of the genomic DNA was performed using established protocols. Briefly, genomic DNA was randomly sheared to 150-200 bp fragments, end-repaired, and ligated to Illumina oligonucleotide adapters. After gel purification fragments were amplified by PCR with primers complementary to the ends of the adaptor. Primers were attached to the surface of the flow cell and clusters were formed by bridge amplification. Sequencing-by-synthesis was then performed using standard procedures to sequence clusters[42]. An average of 36 x 10^9 reads were produced for each of the three deeply sequenced strains (EH2, 92A, and Van556), and an average of 27 x 10^6 reads were produced for each of 10 additional strains.

2.6.1 Analysis of Illumina sequenced genomes of *E. huxleyi* by *de novo* assembly

Quality filtered sequence reads were pooled and assembled *de novo* using CLC Genomics Workbench Performer software (http://www.clcbio.com/products/clc-genomics-workbench) with standard default settings. The estimated coverage of the three deeply sequenced strains ranged from 265-352X while that of the other 10 strains ranged from 14-29X coverage (Supplementary Table 5 and 6). Scaffold N50 values ranged from 2620-3378 bp and from 645-1018 bp for the deeply and moderately covered strains respectively, with total scaffold lengths of 98-117 and 49-76.5 Mb respectively (Supplementary Table 7). In all cases, assembly sizes are smaller than the genome sizes estimated by flow cytometry for individual strains. Short sequence reads prohibit assembly of redundant (repetitive) parts of the genomes, thus the assemblies represents the unique parts of the genomes plus contigs representing the consensus of the repeats.

BLASTn was used to compare the three deeply sequenced genomes (EH2, 92A, and Van556) to that of CCMP1516. Of the sequences from the three strains, only 52-74% of the individual assemblies show significant homology to the CCMP1516 reference genome at >90% identity over regions > 100 bp, and 54-77% at > 80% identity over regions of > 100 bp (Table SI 11,12). Using the identity threshold > 80% this amount of homolgous sequence represents ~ 45% of the reference genome. The amount of non-gapped sequence from the three strains that is absent (< 80 identity) from CCMP1516 varies from 19-54 Mb per strain (Table SI 12). Reciprocal BLASTn analysis shows 48-54 Mb of CCMP1516 does not map to the individual

strains (Table SI 13), with 27.5 Mb exclusive to the genome of CCMP1516. A total of ~8.8, ~21.4, and ~40.7 Mb is exclusive to each of 92A, EH2, and Van556, respectively (Table SI 14).

The coding portion (CDS) of a genome can be defined with gene prediction programs. When predicting genes with a high sensitivity, the false positive rate increases. This is particularly the case if CDSs are being defined in genomes with high GC content. Thus, due to the relative high GC content in the *E. huxleyi* genome the false positive rate might be high. To gain a better understanding of the coding capacity of *E. huxleyi* a comparison of the assemblies was used to estimate the number of shared genes between different strains. The refined haploid set of predicted proteins from the *E. huxleyi* "type strain" CCMP1516 was used to compare the coding capacities between the strains. As can be seen in Table SI 15, in each case more than 16,000 proteins had matches with more than 95% identity at the protein level to a contig of the individual assemblies. The probability of having assembled a particular location in all of the sequenced genomes is roughly 88% (99% for the individual genome multiplied with each other). More than 14,000 proteins had matches with a Bit score higher than 200 to all assemblies (Table SI 16). Thus, the shared gene complement of all strains should be in the range of at least 16,000 genes. This value could be an underestimation if gene predictions are incorrect in the "type strain" or if there is an excess of introns, which would render the gene detection incomplete due to short identical gene fragments.

A BLAST analysis using protein sets can yield only a rough overview of the similarity between the coding capacities of genomes. This analysis does not differentiate between true matches and spurious, false positives, nor does it allow for comparisons at the DNA level. Thus, we aimed to obtain a more detailed view of the genome by analyzing a randomly chosen 200 kb portion of the "type" genome as a test case. A BLAST search at the nucleotide level (BLASTn using a minimum threshold length of 300 nt and a minimum threshold identity of 95%) of this test region against the other assembled genomes yielded matches to contigs, which could then be readily aligned to this region. We constructed a multiple alignment using all matching segments. Since we required the segments to be at least 300 nt long we excluded smaller matching fragments and longer fragments separated by indels. Thus, the final alignment contained only the best matching contigs from each genome.

Interestingly, contigs aligned in clusters at certain positions in the test scaffold (Figure SI 6). This indicates that the genome is divided into high similarity regions and a variable portion.

The high similarity regions therefore constitute the core genome of *E. huxleyi*. Exactly matching EST sequences were added to get an indication of which portions of the scaffold are transcribed. ESTs are commonly enriched in UTRs, which can differ considerably from strain to strain, and therefore might not be represented in the other genomes. EST sequences tended to match to the core genome or directly adjacent to it indicating a high degree of overlap between the core genome and the transcribed genome. The predicted genes in this region are also often associated with the transcribed part and/or the core genome. Gene predictions often, but not always, coincide with the core genome region. Some predicted genes are intron rich with only short coding exons. These genes are not covered by DNA alignments from other strains indicating sequence variability in intron sequences. Furthermore, our previous estimates of the fraction of the core genome was 50% of the total predicted genes, the other 50% being either incorrectly predicted or strain specific genes. A closer look at the annotations connected to the predicted genes reveals it is mainly the core genome genes that contain identifiable domains and have counterparts in databases (Figure SI 6). One gene has no counterparts in databases and is not present in the other *E. huxleyi* genomes (indicated as "unique"). The gene adjacent to this "unique" gene was previously defined as encoding a surface antigen of *E. huxleyi* CCMP1516. Since this type of gene is highly variable, it is not surprising that the other *E. huxleyi* strains do not possess a similar genomic region. From the analysis of the 200 kb segment we concluded: i) the coding portion of the core genome genes is not only identifiable on the protein level, but is represented by highly identical DNA segments in all *E. huxleyi* strains. It comprises roughly 50% of all predicted genes ii) gene predictions in the variable portion of the genome encode strain specific proteins like the surface antigens or may be incorrectly annotated.

2.6.2 Analysis of Illumina sequenced genomes of *E. huxleyi* by direct mapping

In order to analyze the core and variable genome among the 13 Illumina sequenced strains, Illumina reads were mapped to the reference genome using the Mapping and Assembly with Qualities (Maq, http://maq.sourceforge.net/) program. The reads were trimmed to 30 nt with low-quality bases removed, and mapped to the reference genome using the default Maq parameters that allow a maximum of 2 mismatches per alignment. The average depth of mapping coverage across non-gap regions for the deeply sequenced strains was ~250 while that of the strains sequenced to lower coverage ranged from 3.127 to 4.612. To characterize reads not aligning to

the reference genome, which ranged from 5-35% across strains, we compared the GC and repeat content of the mapped and unmapped reads and found while the mapped reads had a slightly higher GC content there was little difference in the repeat content (Figure SI 7), suggesting the unmapped reads are either contaminants or strain specific. Then we examined hits to the refined set of predicted genes of the reference genome, and considered a gene as present in a strain if the coverage by mapped reads was > 50% of the gene length. Using this criteria a total of 5,218 or 17% of the genes are missing from the three deeply sequenced strains (92A, Van556 and EH2) with 1373-2012 different genes missing from each of the individual strains, and 364 appeared to be missing from all three strains. These findings cannot be explained by poor coverage or sequencing bias alone. Based on the alignment coverage >50% of the gene length, 94.8% of the 458 highly conserved single-copy eukaryotic genes from the CEMGA set[7] were identified in the CCMP1516 reference genome and 95% were present in the three deeply Illumina sequenced strains (Supplementary Data File 7). Using the same > 50% gene coverage criteria, a core genome comprised of 20,055 genes that are shared across the 13 studied strains and a variable genome of 10,514 genes that are missing from one or more strains, were identified. CEMGA estimates of completeness for the lower coverage strains range from 91-95%.

2.7 Comparative genomic hybridization

The comparative analysis based on hybridization of genomic DNA of 15 *E. huxleyi* strains (NZEH, EH2, Van556, 12-1 (CCMP 371), CCMP373, Ch24/90, Ch25/90, CCMP374, CCMP379, L, 92, 92A, 92D, 92E, 92F, hybridization with 92A is redundant to CCMP379) against the reference strain CCMP1516 on an *E. huxleyi*-DNA microarray (Agilent, Santa Clara, CA) reveals first insight into the genomic plasticity of *E. huxleyi*. We analyzed a set of 31,940 reproducibly hybridizing probes, each specific for a gene from the sequenced reference strain (sequences from best protein models). Across all strains, using GACK, we identify a core genome from 14,628 probes that hybridize with intensity comparable to or higher than the reference with estimated probability of presence of 95%[43]. Two hundred and twenty-four probes are specific for the reference Protein-IDs (Supplementary Data, File 6).

A GO-enrichment analysis of the core genome was performed with Fisher's exact test (two-tailed) to identify significantly over- and under-represented annotations within the core genome against full annotations from the reference models. This analysis filters for significant

differences in GO-term counts between a selected subset of genes (core genome, or strain-specific gene list) against the annotations of the full genome. In good correspondence with 39% GO-annotated genes of the full genome (12,135 of 30,569 best protein models), 33% core genes have annotations (4,839 of 14,628).

A subset of 104 GO-terms from the annotated core-genome is found significantly overrepresented (FDR < 0.05). Thirteen GO-terms are significantly overexpressed, containing three cellular component terms (mitochondrial part, mitochondrial envelope, organelle envelope).

Under-represented molecular functions include channel activity, ion channel activity, passive transmembrane activity, substrate-specific channel activity, subtilase activity, ion binding, metal ion binding and cation ion binding. The fact, that no molecular function terms from biosynthesis are found supports the view that a high degree of completeness of the core-genome in terms of its functional annotation has been achieved. It is noteworthy that none of the identified 224 reference-strain (CCMP1516) specific gene models have GO-annotations.

An asymptotic core genome size estimate of ~13,000 genes was obtained from resampling the CGH data[43] and fitting non-linearly with 500 repetitions, excluding asymptotic size estimates below 8,000. Results are shown in Supplementary Figure 2. Error bars are from 200 resamplings, permuting the order of hybridization data. The asymptotic size of 13,000 extends to 20,000 when genes of false-negative probability above 5% are considered as present (i.e., extending the conservative core-genome estimate presented above to a set of 22,363 gene models).

For comparisons with the Illumina sequence data, the predicted refined set of *E. huxleyi* CCMP1516 proteins (30,569) was searched (using BLASTx) against assemblies constructed from the raw Illumina sequences of the single strains. Using a lower score threshold of 200 without applying a length threshold we defined the core protein set for all sequenced *E. huxleyi* strains. This approach yielded a potential core set of ~14,000 proteins (Supplementary Figure 2).

2.8 Concatenated gene phylogenies

Nucleotide sequences for 32 genes (gene IDs listed below) were collected for 13 *E. huxleyi* strains for a concatenated phylogenetic analysis. Reference sequences were included for CCMP1516, as well as nucleotide sequences from Illumina sequencing. Genes were selected on

the basis of coverage and variability. The most variable gene sequences, exhibiting identity scores between 92-96%, were selected from amongst the top most highly covered gene sequences where reads covered at least 85% of selected gene regions in every strain, were concatenated.

Sequences were aligned using MUSCLE version 3.6[38]. All ambiguously homologous and invariant sites were removed. The phylogenetic analysis was carried out using two separate programs. Bayesian analysis was performed by the program MrBayes version 3.1.2[44]. The analysis incorporated a gamma correction for rate among sites. The analysis was run for 10^6 MCMC generations with burnin (250) determined by removing all trees before a graphically defined plateau. All calculations were checked for convergence with a splits frequency of <0.01. Maximum likelihood analysis was performed by the program RAxML-VI-HPC v. 2.2.3[34]. The analysis incorporated a GTRMIX model of sequence evolution. Prior to the concatenated phylogenetic analysis, individual phylogenies were constructed for each gene using the programs and settings described above in order to ensure that tree topologies were not strongly inconsistent.

List of genes included in this phylogeny:

61344, 63451, 63753, 68375, 78156, 97861, 103111, 114864, 119769, 194249, 198215, 199511, 201849, 208061, 217143, 217464, 217596, 220352, 222665, 227094, 227111, 230973, 232428, 235187, 235621, 238644, 309298, 433711, 443039, 447697, 457317, 460090

2.9 Analysis of horizontal gene transfer events

2.9.1 Identification of prokaryotic genes

We built a reference database of cellular and viral proteome sequences from the KEGG database[22] and the viral section of the NCBI/RefSeq data set (as of February 2011). The reference database contains proteomes from 156 eukaryotes, 1,168 bacteria, 94 archaea and 3,773 viruses. Regarding *E. huxleyi* protein sequences, we used only protein sequences (JGI reduced gene models) from contigs that encode at least one protein best matching to a eukaryotic protein in the reference database (BLASTp E-value < 10^{-10}). To make an initial list of prokaryotic genes of *E. huxleyi*, we performed the following two-way BLAST searches. First,

BLASTp searches were performed from *E. huxleyi* protein sequences against the reference database (E-value < 10^{-10}). For each *E. huxleyi* sequence that best matched to a prokaryotic sequence, we recorded the BLAST score, which was denoted as X. The prokaryotic sequence was then searched against the reference database to collect its close homologs with a threshold of score ≥ X. If these close homologs were all belonging to prokaryotes (bacteria or archaea), the corresponding *E. huxleyi* query sequence was retained for further analysis. With this two-way BLAST method, we identified 819 *E. huxleyi* protein sequences. Homologs of these *E. huxleyi* sequences were then collected using BLASTp against the reference database (E-value < 10^{-5}) and BLASTCLUST[32], and aligned using MUSCLE[38]. All the gap-containing sites in the alignments were excluded before phylogenetic analysis. Bootstrapped neighbor-joining trees were produced using QuickTree with Kimura's correction[45]. Maximum likelihood trees were produced using PhyML with LG substitution model and a gamma distributed site rates (four rate categories). The generated phylogenetic trees were mid-point rooted by Phylip/Retree[46] to facilitate the identification of sequence groups. After these phylogenetic reconstructions, we discarded cases where *E. huxleyi* sequences form a clade with other eukaryotic sequences. We also discarded cases likely corresponding to eukaryote-to-bacteria gene transfer, where *E. huxleyi* sequences and other closely related eukaryotic sequences were grouped with prokaryotic sequences. Branches were considered supported when the minimum value of the approximate likelihood ratio test (aLRT), parametric (Chi2-based) branch support and the aLRT non-parametric branch support based on a Shimodaira-Hasegawa-like test was greater than or equal to 95%[47].

Finally, our analysis revealed 388 *E. huxleyi* protein sequences forming a clade only with prokaryotic sequences in both neighbor joining and maximum likelihood trees. We classified these *E. huxleyi* sequences into three categories: 47 *E. huxleyi* sequences grouped with prokaryotic sequences from a few genera from a single class or phylum of prokaryotes with significant branch supports; 110 *E. huxleyi* sequences grouped with prokaryotic sequences from a wide range of prokaryotic genera with significant branch supports; and 231 *E. huxleyi* sequences grouped with prokaryotic sequences with no branch support. We considered the 47 cases of the first category as likely candidates for horizontal gene transfers (HGTs) from prokaryotes to *E. huxleyi* (Supplementary Table 8).

2.9.2. Identification of putative HGT events between *E. huxleyi* and viruses

Homologous sequences of *E. huxleyi* protein models (JGI reduced set) were gathered from UniProt[48] using BLASTp, PSI-BLAST and BLASTCLUST[32]. Multiple sequence alignments were generated using MUSCLE[38]. All the gap-containing sites in the alignment were excluded in the phylogenetic analysis. Phylogenetic analyses were performed using the neighbor-joining (NJ) method implemented in ClustalW[49] and the maximum likelihood (ML) method implemented in PhyML[50,51]. NJ analysis was performed based on the distances with Kimura's correction. ML analysis was performed with WAG substitution model and a gamma distributed site rates (four rate categories). We considered *E. huxleyi* and viral proteins as potentially originating from horizontal gene transfer if their closest homologs were only found in *E. huxleyi* and viruses, or if the *E. huxleyi* and viral protein sequences formed a monophyletic group in both NJ- and ML-analyses (Supplementary Table 8).

2.10 Detection of repeats and transposable elements

All densities and coverage values of repetitive elements given in this section are computed with respect to the number of A,C,G,T bases in the haploid genome (~131 Mb), thus correcting densities for the number of Ns in the reference assembly (~141 Mb). Repeat densities are given in base pairs per mega base pairs (bp/Mbp). They can be converted to a coverage value measured in percent by dividing the repeat density in bp/Mbp by 10,000. The total number of A,C,G,T sites in the haploid genome assembly as well as in introns, intergenic regions, CDS regions, 5'UTR and 3'UTR are respectively: ~131 Mb; ~17 Mb; ~78 Mb; ~29 Mb; ~0.6 Mb and ~1.7 Mb. A small amount of genomic bases (3.7%) has not been attributed to any of the haploid genomic regions if ambiguous annotations have been found in a single gene model.

2.10.1. Analysis of the tandem repeat content in the genome of *E. huxleyi* CCMP1516 in comparison to related organisms

For this comparative genomic study, tandem repeats (TR) have been detected with the Phobos program (v3.3.1, http://www.rub.de/spezzoo/cm/cm_phobos.htm). We searched for imperfect tandem repeats with search parameters chosen such that only repeats with a considerably conserved repeat structure are allowed. The search parameters were: match score: 1, mismatch and indel score: -5, minimum score without counting the first unit: 12 (or the unit length, whichever is higher), recursion depth: 5, unit size range: 1-51 bp, maximum number of

successive Ns in repeat: 4. TR characteristics in the unit size range 1-50 bp, have been determined with Sat-stat (v1.3.0, Christoph Mayer, unpublished) which computes TR statistics as well as the TR coverage in the genome for different unit size ranges.

We found that in the *E. huxleyi* genome the density of TRs with conserved repeat structure is high compared to most other genomes (Figure SI 8, Table SI 18). In particular, minisatellites in the size range 7-50 bp contribute significantly to the TR content. The genomic density of individual TR classes is shown in Figure SI 9. Specifically, the TRs with a pattern size of 10 and 11 bp constitute the dominant repeat classes in the *E. huxleyi* genome.

2.10.2 Analysis of transposable elements and other repeats

Dispersed repeats present in the *E. huxleyi* CCMP1516 genome were identified and annotated using the REPET pipeline (v1.3.13; http://urgi.versailles.inra.fr/index.php/urgi/Tools/REPET) that integrates a combination of *de novo* and similarity-based approaches[52,53]. Initially, high-scoring segment pairs (HSPs) were identified by comparing the whole *E. huxleyi* genome to itself using the program BLASTER[52]. HSPs were clustered using the GROUPER, RECON, and PILER programs[52,54,55], and groups comprising at least three HSPs (n=10,354) were retained for further analysis. Clusters of HSPs were then aligned using the MAP algorithm[56] and multiple sequence alignments were used to derive a consensus sequence for each cluster.

Each consensus was classified using an in-house tool called PASTEC. PASTEC combines three complementary approaches to detect a variety of features in the consensus sequences: i) screen for structural features characteristic of transposable elements (TEs) such as long terminal repeats (LTRs), terminal inverted repeats (TIRs), and polyA tails, as well as for the presence of TRs using Tandem Repeats Finder (TRF version 4.0.0[57]); ii) search for similarity with known nucleic and amino acid TE sequences deposited in Repbase (version 15.11; http://www.girinst.org/[58]) using BLASTx, tBLASTx, and BLASTn; iii) probe for virtually all hidden Markov models (HMMs) from Pfam annotation database using HMMER[59]. The bank of HMMs was adapted to distinguish between two classes of Pfam annotations: TE-specific or not (host gene-specific). According to the features detected, PASTEC proposes an automated classification of the input sequences. In an effort to improve TE classification, we attempted to manually construct a library of *E. huxleyi*-specific TEs. Indeed, because most TEs are fast evolving sequences, they display only weak conservation across eukaryotic super-groups,

essentially at the level of core catalytic domains. For that reason, manual TE identification can help address the TE content from species that are distantly related to other eukaryotes for which TEs are referenced in public databases[60-62]. Thereby, LTR FINDER[63] was used with the whole genome as input in order to identify full length LTR-retrotransposon sequences in the genome. In addition, consensus sequences were compared to TEs referenced in Repbase and two in-house databases using BLASTx and tBLASTx. The results were manually curated to compile a library of *E. huxleyi* reference TEs that was appended to the Repbase library to launch PASTEC. In addition, transfer and ribosomal RNA genes were searched in the *E. huxleyi* CCMP1516 genome using the tRNAscan-SE and RNAmmer programs, respectively[64,65], and compared to the consensus sequences using BLASTn. The features collected from each consensus sequence were subsequently examined and used as a support for the manual curation of the results obtained from automated classification with PASTEC.

Of the consensus 1,815 were annotated as host genes (consensus with only significant host gene Pfam annotation). These include main categories such as ankyrin repeat (91 consensus), GCC2 protein repeat (n=76), protein kinase (n=47) and kelch repeat protein (n=44). 160 sequences were classified as autonomous TEs (136 Class 1 and 24 Class 2) because they fit at least one of four criteria: i) to display > 80% BLASTn coverage by manually identified TEs; ii) to display > 70% BLASTn and > 60% tBLASTx coverage by manually identified TEs; iii) to display TE-specific Pfam HMM profile and to be validated manually if best hit (BLASTx) against GenBank nr protein database is a cognate TE sequence; iv) to display BLASTx or tBLASTx hit with a sequence from Repbase and to be validated manually if best hit (BLASTx) against GenBank is a cognate TE sequence. We also annotated 347 consensus sequences as putative non-autonomous Class 2 elements because they display TIRs, including two subcategories: TIRs (n=288) ranging 501-5,000 bp, and MITEs (n=59) ranging 101-500 bp. In addition, 20 consensus sequences were annotated as putative rDNA genes because they display > 80% rDNA coverage, and 50-80% rDNA coverage was also detected in another 114 consensus sequences including 6 putative non-autonomous Class 2 elements (TIRs and MITEs), 1 LINE and 2 consensus classified as TRs. Furthermore, 526 consensus sequences present 20 to 50% rDNA coverage, including 13 TEs, 11 potential genes, and 16 consensus classified as TRs. In addition, 17 consensus sequences were found to contain predicted tRNAs.

Finally, 7,425 consensus sequences were annotated as 'NoCat' (no category) because no

significant feature could be detected or because a classification could not be unambiguously established. These include 1,863 consensus sequences with over 50% TR coverage. Masking of the *E. huxleyi* genome was accomplished by aligning the set of consensus sequences to the *E. huxleyi* CCMP1516 genome using the RepeatMasker (Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-3.0*.1996-2010) and CENSOR[66] programs. MATCHER was used to handle overlapping HSPs and to make connections (also called defragmentation) and locally co-linear annotations of the same consensus were recovered and joined using the 'long join' procedure if the fragments were of similar age and interrupted by younger TE insertions. In addition, the whole genome was screened for TRs using the TRF program. Results were split into two categories: TRs which overlap with dispersed repeats (consensus sequences) and those mapping outside of dispersed repeats. The densities (corrected for the occurrence of Ns) for the different classes of repeated sequences found in the haploid *E. huxleyi* CCMP1516 genomic sequences are summarized in Supplementary Table 9.

2.10.3 Analysis of the genomic coverage by tandem repeats and low complexity regions

When comparing the results described in 2.10.1 and 2.10.2 above, we observed that different search parameters and algorithms led to strongly different TR densities. The dependence of TR characteristics with respect to the choice of parameters and algorithms has been described before[67,68] but, in *E. huxleyi*, this effect was considerably stronger than expected. In contrast to the Phobos program, which was designed to detect TRs that show a mostly clear repeat structure, the TRF program can be parameterized to report more highly degraded TRs, i.e. so-called low complexity regions. Therefore, TRF (v4.0.0) was used with two different sets of search parameters to conduct a comparative analysis. The two sets of TRF parameters are given in Table SI 17.

The parameters of run 2 have high mismatch and indel penalties, minimizing the detection of highly degraded repeats. These parameters are closest to those used in the analysis conducted with the Phobos program in Section 2.10.1. Parameters of run 1 have low mismatch and indel penalties and will tend to detect not only repeats with a well conserved pattern, but also highly degraded repeats and low complexity regions.

The results of the TRF program have been imported into the Sat-stat software version 1.3.10.

This analysis revealed enormous differences in the TR density found with different search parameters (Table SI 17, Figure SI 10). For low mismatch and indel penalties the coverage is as high as 34.4% of the genome (run 1), whereas for stringent penalties the density still has a high value of 12.3% (run 2). These results show that the *E. huxleyi* genome has not only a high proportion of TRs with a clear repeat structure, but that as much as 35% of the genome consists of low complexity regions which are formed by more or less degraded TRs. Interestingly, the dominant pattern in TRs in low complexity regions is the trinucleotide pattern CCG and variations of this motif with longer pattern sizes. Such low complexity regions can serve as recombination hot spots for a quick reshuffling of the genome and thus likely play an important role in the evolution of this genome. The high proportion of degraded (imperfect) repetitive structures testifies that low complexity regions have been impacting the genome over evolutionary times. This in turn may explain to some extent the many peculiarities found in the *E. huxleyi* pan genome, where the variable genome has apparently undergone vast changes on short time scales.

2.10.4. Analysis of the coverage by tandem repeats and low complexity regions of different genomic regions

We analyzed the distribution of TRs detected with Phobos and TRF in different genomic regions. For this task, introns, CDS, untranslated regions (UTRs), and intergenic regions have been excised according to the "Filtered models" annotation file. Redundancy in the gene annotation resulting for example from alternative splicing, alternative sources, or uncertainty, has been removed according to the following rule: If two gene models overlapped by more than 50% of the length of the shorter one, the longer gene model has been removed. We found that the TRs identified by Phobos in 2.10.1 (the set of TRs with most conserved structure) are most abundant in introns (Figs. SI 11, SI 12). Strikingly, repeats with a pattern size of 10 and 11 bp are not overrepresented in intergenic regions in contrast to them being the dominant repeat classes in the whole genome. Instead, their contribution stems almost completely from their high density in introns. Furthermore, intronic 10 and 11 bp repeats show a strong strandedness (Figure SI 13), meaning that on the sense and antisense strand either the motif or its reverse complement is highly favored. Together, this suggests that intronic 10 and 11 bp repeats might have a functional relevance.

Consistently, low complexity regions identified using relaxed search parameters with TRF were also found to be most abundant in introns. The coverage of introns by low complexity regions is as high as ~50% (Figure SI 12). Interestingly, for TRs found with stringent search parameters, the repeat density in introns is more than a factor of 2 higher than in other genomic regions, whereas for low complexity regions, this factor is smaller than 2. Densities of TRs and low complexity regions in introns of this magnitude as well as a dominance of certain repeat patterns have not been reported for any other genome to our knowledge.

2.10.5 GC content in genomic regions and repeats

E. huxleyi has a GC content of 65%, but high GC contents are not rare in algae as we see in Figure SI 14. Interestingly, the GC content in TRs (found with stringent search parameters with Phobos) is even higher than the average value in the genome for all genomes investigated. In *E. huxleyi*, the GC content in TRs is highest in CDS regions, which can also be observed in several other taxa. A preference for TRs with high GC content could in part explain the high GC content in the genomes as a whole.

2.11 Composition of the core and variable genomes

Genes in the core and variable genomes (based on direct mapping criteria where a gene is called present if hits cover > 50% of the gene length) were functionally classified using gene ontology (GO) annotation. GO annotation was performed with BLAST2GO [69] using second level molecular function terms. The general gene composition of the core and variable genomes is similar, both being dominated by genes of unknown function, and genes encoding housekeeping proteins involved in metabolic processes, transcription, membrane transport, and ion and protein binding (Supplementary Figure 6). Although approximately 70% of the genes in the core and variable genomes were of unknown function, genes in the variable genome appeared shorter and contained fewer introns (51% have one or no introns, compared to 44% in the core genome).

3.0 Ecophysiology

3.1 Phylogenetic and sequence analysis of light harvesting complex genes

Amino acid and nucleotide BLAST searches were performed using NCBI BLAST software (http://www.ncbi.nlm.nih.gov/BLAST/BLAST.chi) and Open Reading Frames (ORF) were determined using ORF finder software (http://www.ncbi.nlm.nih.gov/projects/gorf/). Sequences alignments and phylogeny programs were run from the "Phylogeny.fr: Robust Phylogenetic Analysis For The Non-Specialist" platform accessible at http://phylogeny.lirmm.fr/phylo_cgi/index.cgi[50]. Sequence alignments were performed using MUSCLE 3.7[38] run in full mode with the number of iterations equal to 16. Maximum-likelihood analysis was performed using PhyML 3.0 program[51,70] with the WAG substitution model. The number of substitution rate categories was four and the number of bootstrap samples was 100**.** The branch lengths of the tree were optimized and the tree topology improved with Nearest Neighbor Interchange (NNI) branch swapping operations.

Amongst photosynthetic organisms sequenced to date, *E. huxleyi* has the highest number of predicted full length light harvesting complex proteins (LHCs), totaling 68. Phylogenetic analysis shows that *E. huxleyi*'s LHCs could be classified in three previously described clades[71,72]: the Chlorophyll *a*/*c* LHC group I and II, the red algal LHC, and the LI818-like LHC and a related clade, LHCZ-like. LI818 and LI818-like genes (LHCSR, LHCX) of microalgae species have been shown to be up-regulated under various stresses, including high light[73-76], iron, phosphorous and sulphur deprivation[77,78], while the other members of the light harvesting family exhibit opposite response. These different lines of evidence have prompted the classification of LI818-like LHCs as stress response genes with possible role in photoprotection in other microalgae species. Analysis of LHC promoters (2 kb) using *de novo* motif discovery (MEME SUITE[79]) revealed a canonical motif (E-value 8.2e-024) from position -55 to -28 in the 5'-upstream region relative to the translational start site (Figure SI 15). This motif was detected solely in *E. huxleyi* LI818-like LHC promoters and was absent in the promoters of the other *E. huxleyi* LHCF clades and LHCX of the diatoms *T. pseudonana*, *P. tricornutum* and *F. cylindrus*, suggesting that this motif is specific to *E. huxleyi* LI818-like promoters. It is possible that this motif is acting as a cis-element, regulating the transcription of *E. huxleyi* LI818-like proteins.

3.2 Spider plots

Spider plots (Figure 4) detailing of the distribution of selected genes encoding proteins potentially important to the ecophysiology of *E. huxleyi* were made by cataloging the presence of

these genes in each of the non-reference strains based on data generated by direct mapping of Illumina reads. For this purpose, manually curated genes from the variable genome of the reference strain CCMP1516 were used and as previously described, a threshold of 50% gene coverage by short sequence reads was used for calling genes in other strains present. Genes which included ammonium transporters (ATMs), urea transporters (UT), nitrite reductase (NII), nitrilase (NIT), phosphate transporters (PTA), alkaline phosphatase (PHOA), ferredoxin (FDX), flavodoxin (FldA), nitrate reductase (NAR), Ca^{2+} binding EF hand proteins (CaEF), and genes coding for proteins that bind metals such as copper (Cu), zinc (Zn), iron (Fe) showed non-uniform distribution across strains. Linking these patterns with niche specificities awaits further experimental work.

Table SI 1| Genomic libraries included in the *E. huxleyi* genome assembly and their respective assembled sequence coverage levels in the final release.

Library	Sequencing Platform	Average Read/Insert Size	Read Number	Assembled Sequence Coverage (x)
FIOP	Sanger	3,284±438	364,984	0.40
AKBS	Sanger	3,749±617	1,098,155	2.79
ACCS	Sanger	3,762±618	50,282	0.20
AONE	Sanger	3,823±622	85,803	0.09
AWCG	Sanger	6,294±884	450,511	1.16
ACCT	Sanger	6,321±872	1,8915,52	3.57
ACCU	Sanger	36,026±4,690	333,792	0.45
Total		N/A	3,910,095	8.26

Table SI 2| Summary statistics of the final genome release v1.0. The table shows total contigs and total assembled basepairs for each set of scaffolds greater than the size listed in the left hand column.

Minimum Scaffold Length	Number of Scaffolds	Number of Contigs	Scaffold Size	Basepairs	% Non-gap Basepairs
5 Mb	0	0	0	0	0.00%
2.5 Mb	2	240	5,580,372	5,425,492	97.22%
1 Mb	23	1,304	32,463,260	31,162,139	95.99%
500 Kb	81	3,237	71,299,831	68,107,657	95.52%
250 Kb	178	5,341	105,972,389	100,308,677	94.66%
100 Kb	323	7,163	129,130,382	121,141,357	93.81%
50 Kb	487	8,209	140,644,545	131,004,310	93.15%
25 Kb	616	8,622	145,105,830	134,827,580	92.92%
10 Kb	1,165	9,783	153,561,179	142,743,837	92.96%
5 Kb	1,689	10,741	157,496,082	145,883,981	92.63%
2.5 Kb	2,246	11,379	159,433,041	147,732,059	92.66%
1 Kb	7,809	16,942	167,727,055	156,026,073	93.02%
0 bp	7,809	16,942	167,727,055	156,026,073	93.02%

Table SI 3| GC ranges for classification filtered scaffolds

Classification	GC Range
Mitochondrion	≤ 0.305
Chloroplast	$0.305 < GC \leq 0.38$
Eukaryotic	$0.38 < GC \leq 0.53$
Prokaryotic	$0.53 < GC \leq 0.63$
Eukaryotic	$GC > 0.63$

Table SI 4| Classification of Filtered Scaffolds

Category	No. Scaffolds	% Scaffolds	No. of Reads	% of Reads
Mitochondrion	5	0.046	761	0.065
Chloroplast	23	0.21	3,967	0.34
Eukaryotic	7,809	69.1	1,054,392	90.0
Prokaryotic	3,314	30.5	111,771	9.5
Non-Cellular	5	0.046	24	0.002

Table SI 5| Genes in the filtered versus refined models set. Many doublets in Filtered Models set are separated alleles and were removed from the Filtered Models set to create the Refined Models set.

Number	Filtered Models	Refined Models
# genes	39126	30569 (⇓22%)
# clusters	20875	12004 (⇓43%)
# clusters w/ 1 gene (singlets)	11484	7012 (⇓37%)
# clusters w/ 2 genes (doublets)	7427	4492 (⇓40%)
# genes in doublets	14854	9684 (⇓40%)

Table SI 6| Refined model set classified by gene prediction method

Prediction method	# models
total	30569
protein-based	3668 (12%)
cDNA-based	6114 (20%)
ab initio	20787 (68%)

Table SI 7| Properties of refined model set

Property or number	Value
Avg. gene length	1,718 nt
Avg. transcript length	1,129 nt
Avg. protein length	346 aa
Avg. exon length	365 nt
Avg. intron length	242 nt
Avg. exon frequency	3.65 exons per gene
# multiexon genes	22,927 (75%)
# genes with similarity to NR protein	21,143 (69%)
# genes with *E. huxleyi* gene family	23,538 (77%)
# genes with EST support	15,642 (51%)
# genes with Pfam domain	10,496 (34%)
# genes with signal peptide	9,782 (32%)
# genes with transmembrane domain	4,891 (16%)
# genes with EC number	3,363 (11%)
# genes with GO term	11,005 (36%)

Table SI 8| Protist and cyanobacterial genomes used for protein domain comparisons with *E. huxleyi*. All proteomes were downloaded on August 8, 2008.

Species	# Proteins	Source
Aureococcus anophagefferens[80]	11501	http://genome.jgi-psf.org/Auran1/
Chlamydomonas reinhardtii[81]	15256	http://genome.jgi-psf.org/Chlre3/
Cyanidioschyzon merolae[82]	5014	http://merolae.biol.s.u-tokyo.ac.jp/
Naegleria gruberi[83]	15753	http://genome.jgi-psf.org/Naegr1/
Ostreococcus lucimarinus[84]	7805	http://genome.jgi-psf.org/Ost9901_3/
Paramecium tetraurelia[85]	39604	http://paramecium.cgm.cnrs-gif.fr/
Phaeodactylum tricornutum[61]	10025	http://genome.jgi-psf.org/Phatr2/
Phytophthora sojae[86]	19027	http://genome.jgi-psf.org/Physo1_1/
Plasmodium falciparum[87]	5484	http://plasmodb.org/
Synechocystis sp. PCC6803[88]	3264	http://www.kazusa.or.jp/e/
Thalassiosira pseudonana[89]	11390	http://genome.jgi-psf.org/Thaps3/

Table SI 9| *E. huxleyi* protein domains that were absent in 11 other protist and cyanobacterial genomes.

Pfam domain	# Proteins with the domain	Pfam description
PF00777.9	11	Glycosyltransferase family 29 (sialyltransferase)
PF04013.3	3	Protein of unknown function (DUF358)
PF07173.3	3	Protein of unknown function (DUF1399)
PF00040.10	2	Fibronectin type II domain
PF00066.8	2	Notch (DSL) domain
PF01773.11	2	Na+ dependent nucleoside transporter N-terminus
PF01784.9	2	NIF3 (NGG1p interacting factor 3)
PF03360.7	2	Glycosyltransferase family 43
PF04148.4	2	Transmembrane adaptor Erv26
PF06977.2	2	SdiA-regulated
PF08939.1	2	Domain of unknown function (DUF1917)
PF09296.2	2	NADH pyrophosphatase-like rudimentary NUDIX domain
PF00404.9	1	Dockerin type I repeat
PF00934.11	1	PE family
PF01033.8	1	Somatomedin B domain
PF01129.9	1	NAD:arginine ADP-ribosyltransferase
PF01270.8	1	Glycosyl hydrolases family 8
PF03313.6	1	Serine dehydratase alpha chain
PF03315.6	1	Serine dehydratase beta chain
PF03974.4	1	Ecotin
PF04041.4	1	Domain of unknown function (DUF377)
PF04181.4	1	Domain of Unknown Function (DUF408)
PF04303.4	1	Protein of unknown function (DUF453)
PF04864.4	1	Alliinase
PF05090.5	1	Vitamin K-dependent gamma-carboxylase
PF05493.4	1	ATP synthase subunit H
PF05569.2	1	BlaR1 peptidase M56
PF05751.2	1	FixH
PF05840.4	1	Bacteriophage replication gene A protein (GPA)
PF05962.2	1	Bacterial protein of unknown function (DUF886)

PF06439.2	1	Domain of Unknown Function (DUF1080)
PF06800.3	1	Sugar transport protein
PF07035.3	1	Colon cancer-associated protein Mic1-like
PF07382.2	1	Histone H1-like nucleoprotein HC2
PF07565.4	1	Band 3 cytoplasmic domain
PF07593.3	1	ASPIC and UnbV
PF07799.3	1	Protein of unknown function (DUF1643)
PF08097.2	1	Conotoxin T-superfamily
PF08409.2	1	Domain of unknown function (DUF1736)
PF08612.2	1	TATA-binding related factor (TRF)
PF08666.3	1	SAF domain
PF08885.2	1	GSCFA family
PF09133.1	1	SANTA (SANT Associated)
PF09332.2	1	Mcm10 replication factor
PF09663.1	1	Amidohydrolase ring-opening protein (Amido_AtzD_TrzD)

Table SI 10| Species included in the phylogenomics analyses

Species Name
Ostreococcus tauri
Drosophila melanogaster
Ostreococcus tauri (PL)
Porphyra haitanensis
Saccharomyces cerevisiae
Giardia lamblia
Alexandrium tamarense
Trypanosoma brucei
Bigelowiella natans (PL)
Cyanidium caldarium (PL)
Aureococcus anophagefferens (PL)
Pythium ultimum
Aspergillus fumigatus
Chlamydomonas reinhardtii
Thalassiosira pseudonana (PL)
Strongylocentrotus purpuratus
Euglena gracilis
Blastocystis hominis
Calliarthron tuberculosum
Cryptosporidium parvum
Phaeodactylum tricornutum (PL)
Phytophthora ramorum
Arabidopsis thaliana
Cyanidioschyzon merolae
Albugo laibachii
Plasmodium falciparum

Porphyridium cruentum.fasta
Porphyra yezoensis
Fucus serratus
Dictyostelium discoidium
Karenia brevis
Heterocapsa triquetra
Bigelowiella natansNM
Emiliania huxleyi
Paramecium tetraurelia
Micromonas RCC299v3
Aureococcus anophagefferens
Porphyra purpurea (PL)
Bigelowiella natans
Phaeodactylum trichornutum
Oryza sativa
Cryptomonas paramecium (PL)
Chlamydomonas reinhardtii (PL)
Caenorhabditis elegans
Guillardia thetaNM
Oryza sativa (PL)
Gracilaria tenuistipitata (PL)
Guillardia theta (PL)
Schizosaccharomyces pombe
Toxoplasma gondii
Leishmania infantum
Thalassiosira pseudonana
Cyanidioschyzon merolae (PL)
Guillardia theta
Homo sapiens
Porphyridium cruentum
Rhodomonas salina (PL)
Fucus_vesiculosus (PL)
Ectocarpus_siliculosus (PL)
Arabidopsis_thaliana (PL)
Phytophthora_sojae
Trichomonas_vaginalis

(PL) indicates plasmid sequence

Table SI 11| BLASTn homology statistics between the genomes of the reference strain *E. huxleyi* CCMP1516 and three other deeply sequenced strains using a >90% identity threshold over regions > 100 bp.

	92A-paired	EH2-Paired	Van-Paired
Number of contigs	56,794	77,783	75,716
Assembly size (bp)	85,612,925	117,731,447	109,723,373
Sequence aligning to CCMP1516 (bp); Gapped Non-gapped	60,105,021 63,050,273	57,239,795 60,740,518	58,443,441 62,130,927
Absent sequences (bp)	25,507,904	60,491,652	51,279,932
Non-gap absent sequences (bp)[1]	22,562,652	56,990,929	47,592,446
Absent sequences GC content (%)	61.0%	56.6%	62.1%
Number of absent sequences	46,482	55,692	75,856
Number of contigs containing absent sequences	34,088	46,037	55,982

Table SI 12| BLASTn homology statistics between the genomes of the reference strain *E. huxleyi* CCMP1516 and three other deeply sequenced strains using a >80% identity threshold over regions > 100 bp.

	92A-paired	EH2-Paired	Van-Paired
Num of contigs	56,794	77,783	75,716
Assembly size (bp)	85,612,925	117,731,447	109,723,373
Sequence aligning to CCMP1516 (bp); Gapped Non-gapped	62,848,354 65,685,833	59,360,887 62,773,345	63,446,431 66,934,211
Absent sequences (bp)	22,764,571	58,370,560	46,276,942
Non-gap absent sequences (bp)	19,926,092	54,958,102	42,789,162
Absent sequences GC content (%)	60.5%	56.3%	61.8%
Num of absent sequences	44,232	52,375	70,802
Num of contigs containing absent sequences	32,012	43,337	51,810

Table SI 13| Sequence from the CCMP1516 reference genome missing from the assemblies of the three deeply sequence strains. Missing sequence was determined by BLASTn using a threshold of < 80% identity over regions of > 100 bp.

	92A	EH2	Van	All
Absent non-gap reference sequences (bp)	48,428,154	53,668,496	52,521,826	27,491,107
GC content of the absent reference sequences	64.6%	64.3%	63.5%	63.6%

Table SI 14| A comparison of the amount of strain specific sequence in the *E. huxleyi* pan genome.

Strain	Novel Sequence (bp)	% of the sequenced genome
92A	8,884,876	10.3
EH2	40,773,754	34.5
Van556	21,442,533	19.5
CCMP1516	27,491,107	19.5

Table SI 15| BLAST hits between the reduced protein set of the reference genome (CCMP1516) and proteins from 13 other strains of *E. huxleyi*. We used a threshold score of 200 in order to reduce the false positive rate and avoid a too high false negative rate. The identity values were taken from the BLAST output, thus representing only uninterrupted parts of the alignments with the query proteins.

% Iden	12_1	92A	92D	92E	92F	B11	B39	AWI1516	EH2	L	M217	M219	Van556	total
100	6839	4282	3808	5607	3376	5689	5598	9515	5399	6213	8944	8857	2603	179
95-100	10836	11445	11074	11632	11218	11058	11401	9525	10394	10504	10002	11071	11121	4131
90-95	3069	4045	3710	3334	4074	3610	3589	2765	3706	3267	2939	2921	4423	3271
85-90	1415	2235	1646	1512	1876	1596	1545	1282	1880	1439	1460	1295	2415	2029
80-85	943	1394	1023	1070	1232	1100	1109	845	1386	957	919	823	1577	1427
75-80	754	1217	725	812	821	736	760	615	1106	795	695	630	1291	1128
70-75	624	1035	758	739	732	593	721	573	925	699	593	466	1112	1154
65-70	618	734	712	652	688	597	584	641	751	671	589	380	921	1155
60-65	546	574	693	564	653	609	567	550	580	632	497	356	731	1149
55-60	472	407	627	488	524	534	466	467	493	534	418	331	565	1246
50-55	511	353	607	508	569	512	461	370	513	528	400	326	485	1427
45-50	415	265	605	384	550	466	482	348	396	449	324	329	367	1464
40-45	561	245	588	439	559	456	485	464	377	490	381	391	395	1722
35-40	420	250	583	382	590	469	377	301	359	495	310	292	321	1898
30-35	275	155	459	233	392	281	254	215	221	373	177	173	225	1446
25-30	84	56	158	100	129	95	83	63	95	114	45	32	66	552
<25	5	9	12	13	13	9	5	0	6	13	1	337	3	49
no	635	321	1234	553	1026	612	535	483	435	849	328	12	401	9
total	29022	29022	29022	29022	29022	29022	29022	29022	29022	29022	29022	29022	29022	25436

Table SI 16| Summary of the BLAST results when comparing the reduced protein set of the "type strain" CCMP1516 against the assemblies of genomes from different strains. Two different measures for protein similarity (identity and bit score) were used.

	Genes Present in ALL Strains	Genes Present in Some Strains	Genes Present Only in CCMP1516	Genes present in only one additional strain
ID>=95%	4310	20647	2295	1170
Score>=200	20084	8811	71	56

Table SI 17| The two sets of tandem repeat search parameters used for the TRF program

Run	Match score	Mismatch score	Indel score	Mismatch probability	Indel probability	Minimum score	Unit size range (bp)
1	2	3	5	80	10	20	1-2000
2	2	10	10	80	10	24	1-2000

Table SI 18| List of genomes for which the TR content has been compared to the *Emiliania huxleyi* genome

Species name	Assembly version	Source
Arabidobsis thaliana[90]	6.0	ftp://ftp.ncbi.nih.gov
Aureococcus anophagefferens[80]	1.0	http://genome.jgi-psf.org/Auran1/
Chlamydomonas reinhardtii[81]	4.0	http://genome.jgi-psf.org/Chlre4/
Chlorella sp.[91]	1.0	http://genome.jgi-psf.org/ChlNC64A_1/
Coccomyxa sp.C-169[92]	2.0	http://genome.jgi-psf.org/Coc_C169_1/
Cyanidioschyzon merolae[82]	1.0	http://merolae.biol.s.u-tokyo.ac.jp/download/
Emiliania huxleyi CCMP1516	1.0	http://genome.jgi-psf.org/Emihu1/
Micromonas pusilla CCMP1545[93]	2.0	http://genome.jgi-psf.org/MicpuC2/
Micromonas sp. RCC299[93]	3.0	http://genome.jgi-psf.org/MicpuN3/
Neurospora crassa[94]	7.0	http://www.broad.mit.edu/annotation/genome/neurospora/
Ostreococcus lucimarinus[84]	2.0	http://genome.jgi-psf.org/Ost9901_3/
Ostreococcus RCC809	1.0	http://genome.jgi-psf.org/OstRCC809_1
Ostreococcus tauri[84]	2.0	http://genome.jgi-psf.org/Ostta4
Phaeodactylum tricornutum[61]	2.0	http://genome.jgi-psf.org/Phatr2/
Phytophthora capsici[95]	1.0	http://genome.jgi-psf.org/PhycaF7/
Physcomitrella patens patens[96]	1.1	http://genome.jgi-psf.org/Phypa1_1/
Phytophthora ramorum[86]	1.1	http://genome.jgi-psf.org/Phyra1_1/
Phytophthora sojae[86]	1.1	http://genome.jgi-psf.org/Physo1_1/
Saccharomyces cervesiae[97]	2.1	ftp.ncbi.nih.gov
Selaginella moellendorffii[98]	1.0	http://genome.jgi.doe.gov/selaginella/
Sorghum bicolor[99]	1.0	http://genome.jgi.doe.gov/Sorbi1/
Thalassiosira pseudonana[89]	3.0	http://genome.jgi-psf.org/Thaps3/
Volvox carteri[100]	1.0	http://genome.jgi-psf.org/Volca1/

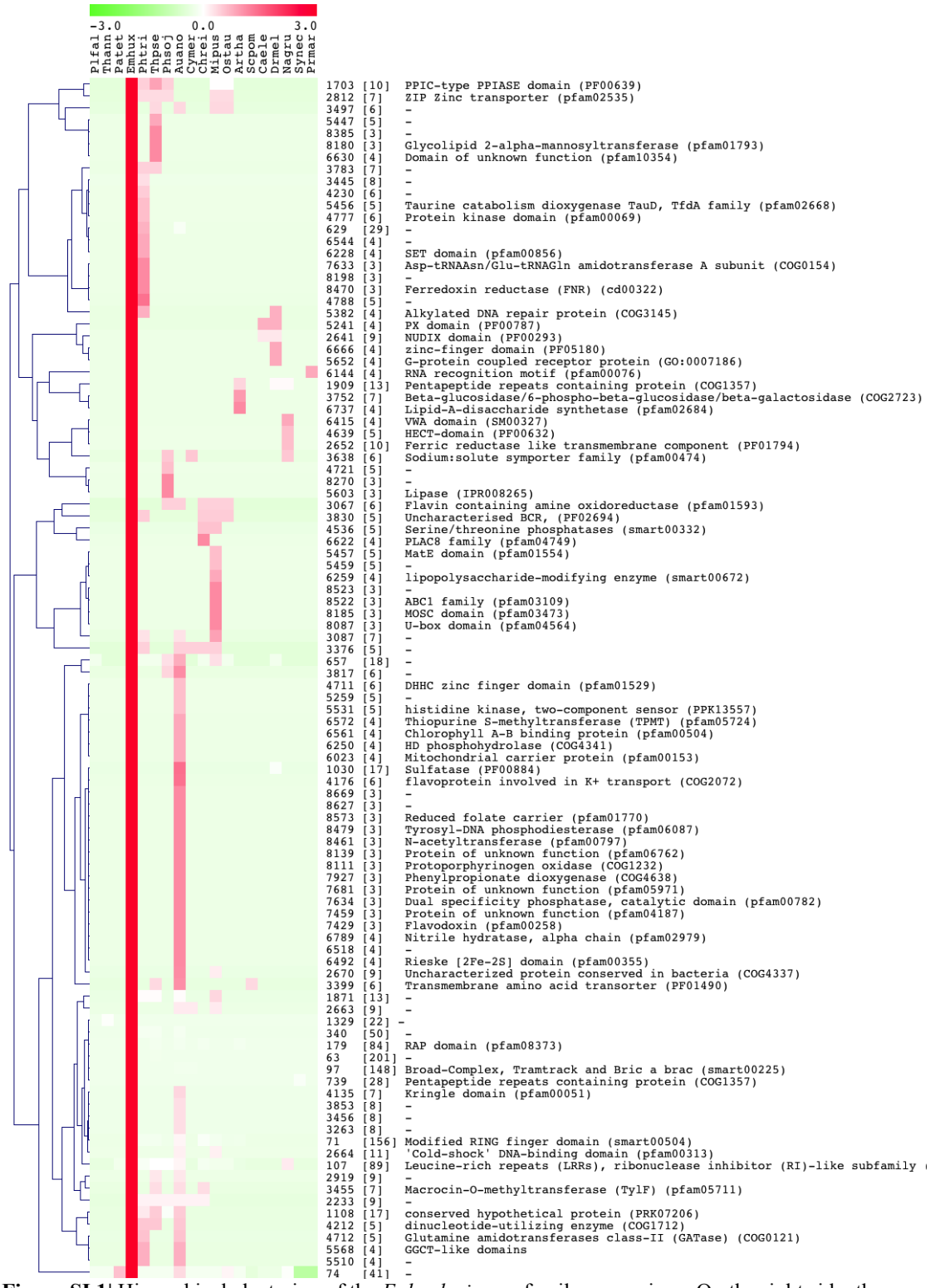


Figure SI 1| Hierarchical clustering of the *E. huxleyi* gene family expansions. On the right side, the gene family IDs are shown followed by the number of *E. huxleyi* genes in that family. The functional description of the gene and their corresponding conserved domain database IDs are displayed. The green and red scale (based on z-scores) show where gene family sizes are substantially smaller or larger than the mean gene family size. Red blocks represent gene family expansions.

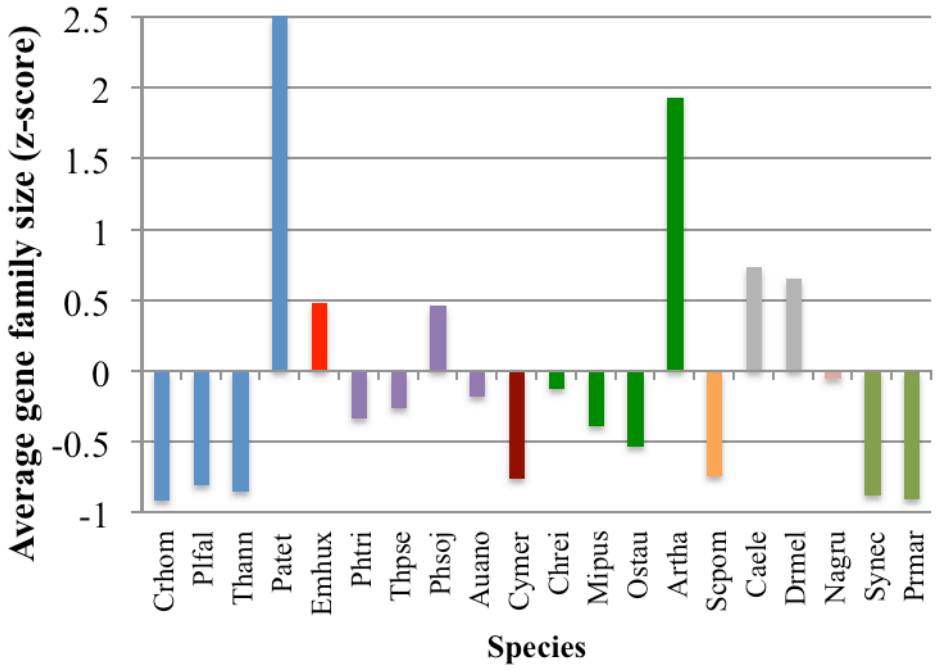


Figure SI 2| Average gene family size in each species. The average gene family size was calculated based on the protein phylogeny profile excluding the orphan genes and converted into z-score. The abbrevation of each species: Crhom: *Cryptosporidium hominis*[101], Plfal: *Plasmodium falciparum*[87], Thann: *Theileria annulata*[102], Patet: *Paramecium tetraurelia*[85], Phtri: *Phaeodactylum tricornutum*[61], Thpse: *Thalassiosira pseudonana*[89], Phsoj: *Phytophthora sojae*[86], Auano: *Aureococcus anophagefferens*[80], Cymer: *Cyanidioschyzon merolae*[82], Artha: *Arabidopsis thaliana*[90], Chrei: *Chlamydomonas reinhardtii*[81], Mipus: *Micromonas* strain RCC299[93], Ostau: *Ostreococcus tauri*[84], Scpom: *Schizosaccharomyces pombe*[103], Caele: *Caenorhabditis elegans*[104], Drmel: *Drosophila melanogaster*[105,106], Nagru: *Naegleria gruberi*[83], *Sycom: Synechococcus s*train WH8102[107] and Prmar: *Prochlorococcus marinus* strain SS120[108].

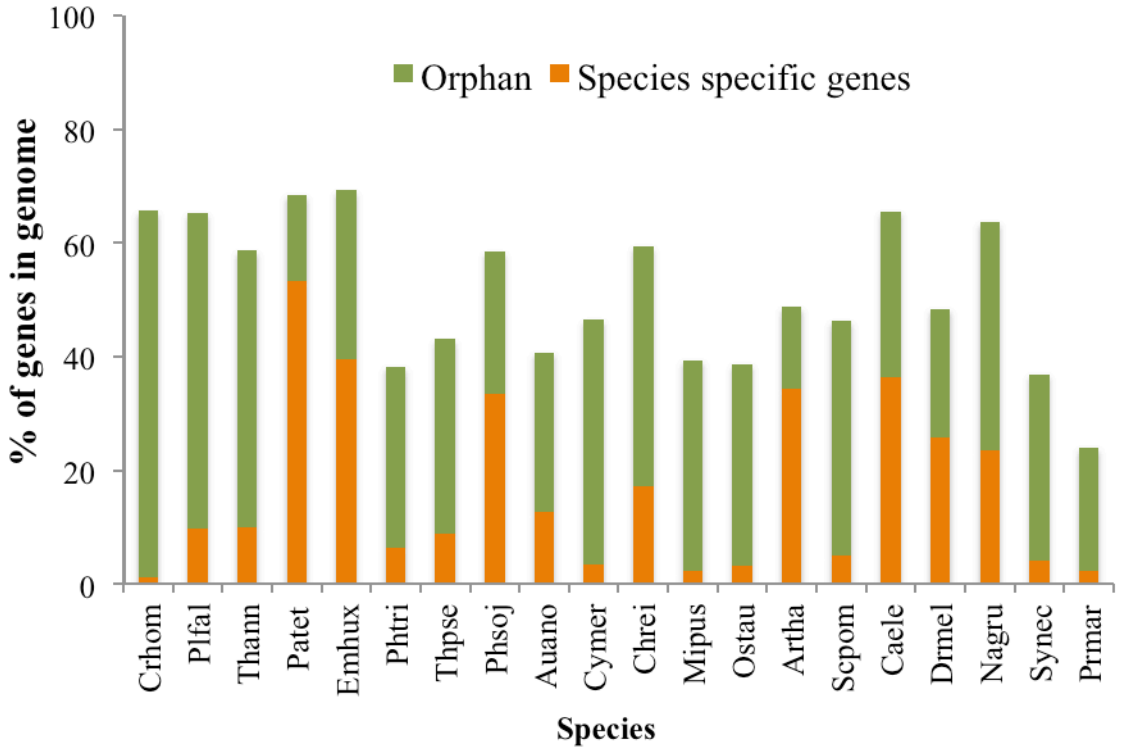
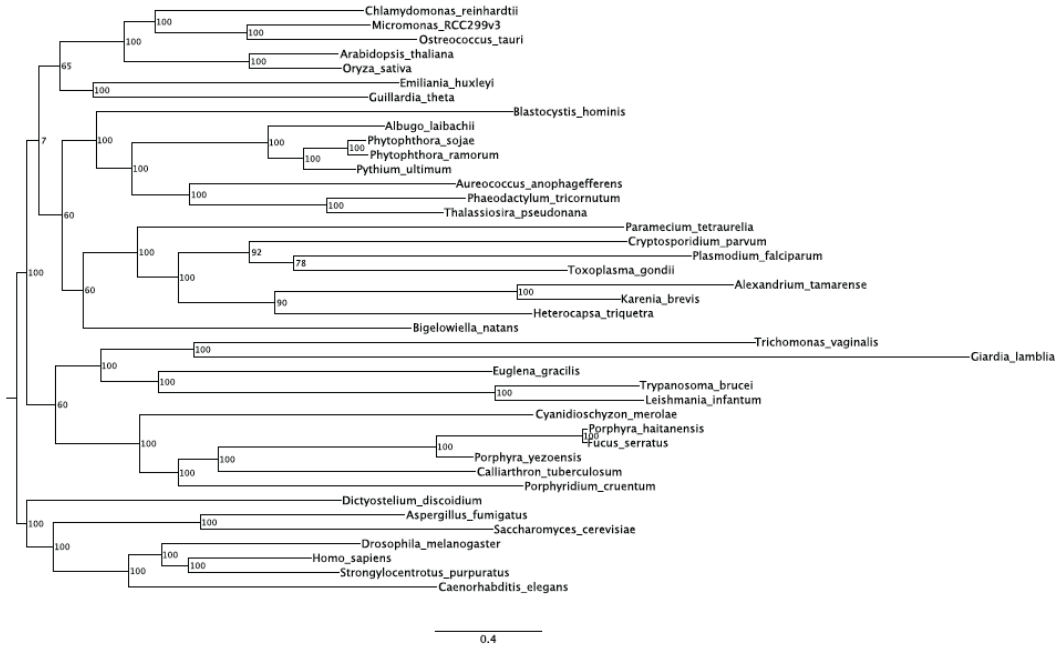


Figure SI 3| The distribution of species-specific gene families and orphan genes in a range of genomes. Abbreviations are listed in the caption of Figure SI 2.

a



b

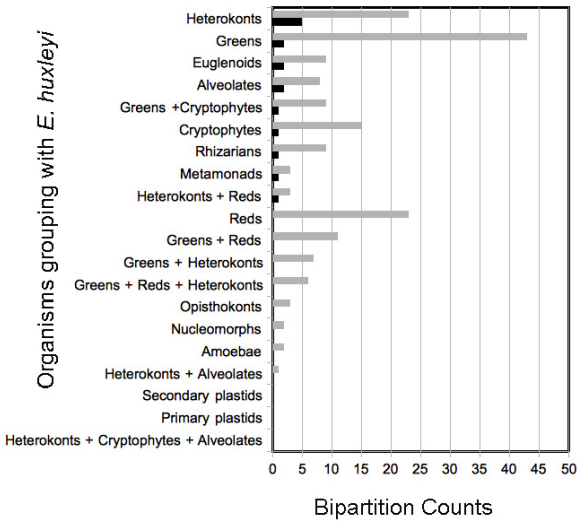


Figure SI 4| Concatenated phylogeny of 228 *E. huxleyi* genes. a) RAxML topology with support values from 100 bootstrap replicates. Note the poorly resolved placement of the *E. huxleyi+Guillardia theta* clade with respect to the other plastid-containing lineages. b) Graph of trees from concatenation with counts of otherwise monophyletic bipartitions containing *E. huxleyi* from single protein trees corresponding to the alignments used in the concatenated tree. Black bars represent counts from only those trees in which the branch leading to *E. huxleyi* and its sister taxon is represented by a bootstrap value of 70 or above. Grey bars represent counts from all trees.

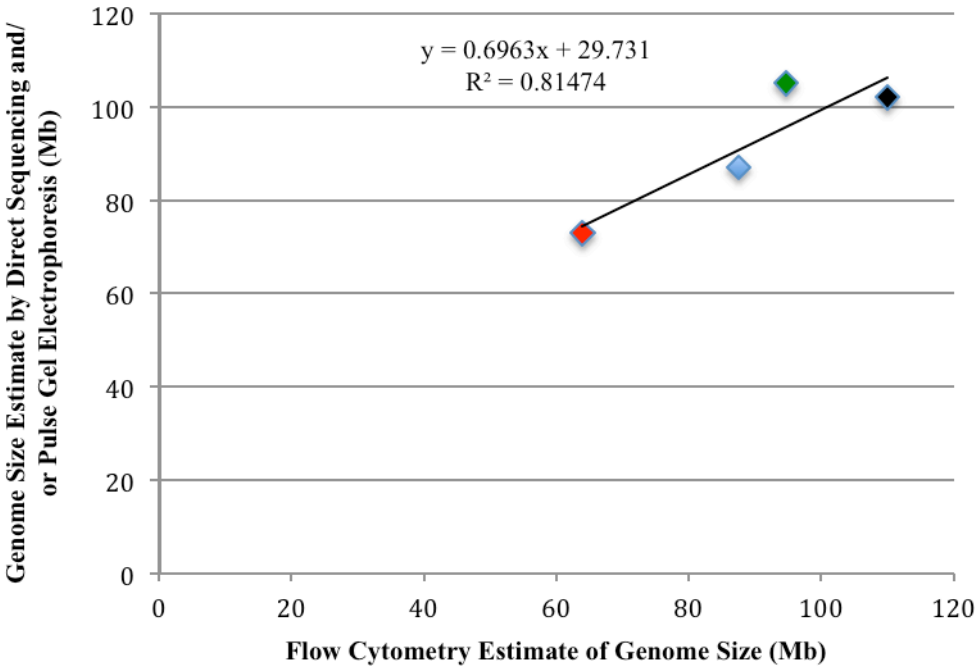


Figure SI 5| The accuracy of flow cytometry for estimating genome sizes. Genome size estimates made by flow cytometry of three marine alga with sequenced genomes (*Guiardia theta*-blue, *Bigelowiella natans*-green, and *Thalassiosira pseudonana*-red), and one ciliated paramecium (black) whose genome size had previously been estimated by pulse gel electrophoresis, positively correlated with estimates made by flow cytometry.

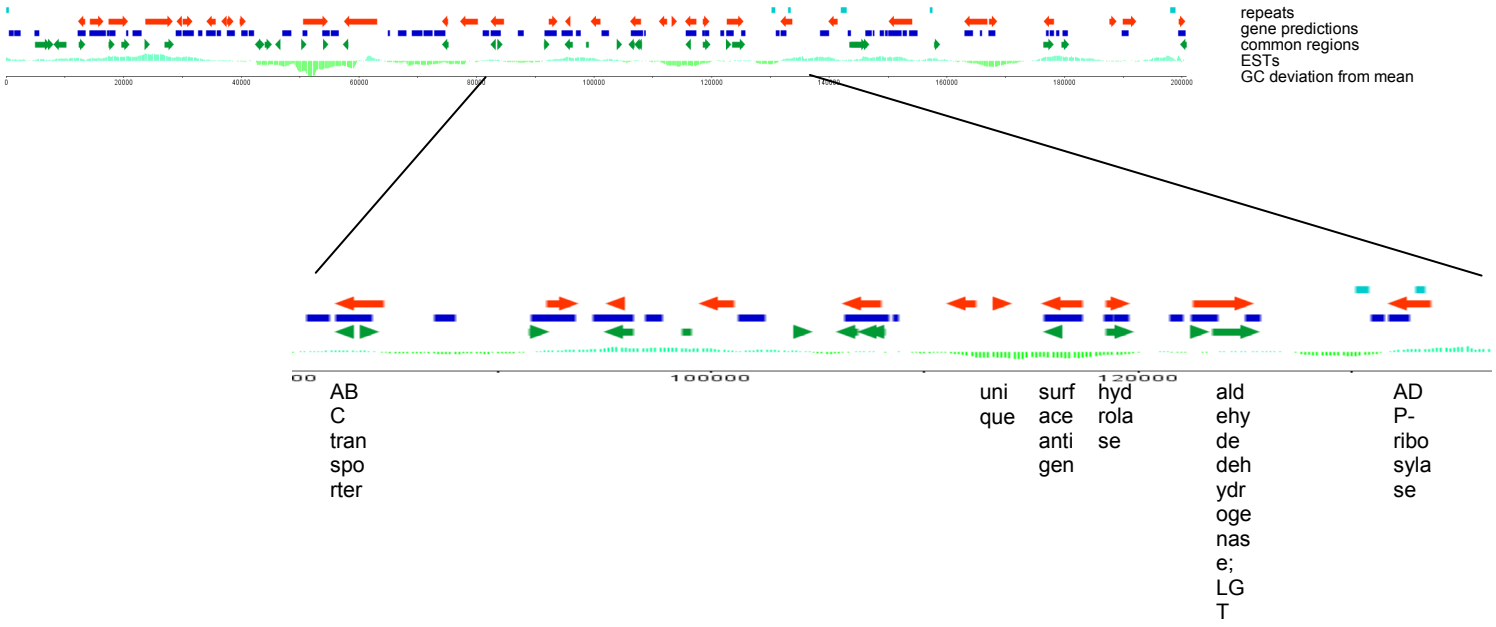


Figure SI 6| Representation of a 200 kb scaffold (AKBS149348.g2..0 in the reference genome assembly) with information on common regions between other *E. huxleyi* genomes (blue), matching EST sequences (green), gene predictions (red), and long regions repeated in this scaffold (light blue). For the enlarged portion the annotations for the predicted genes are indicated. Matches to hypothetical proteins were omitted. The gene prediction with the label "unique" has no counterpart in the databases presumably due to a false positive prediction.

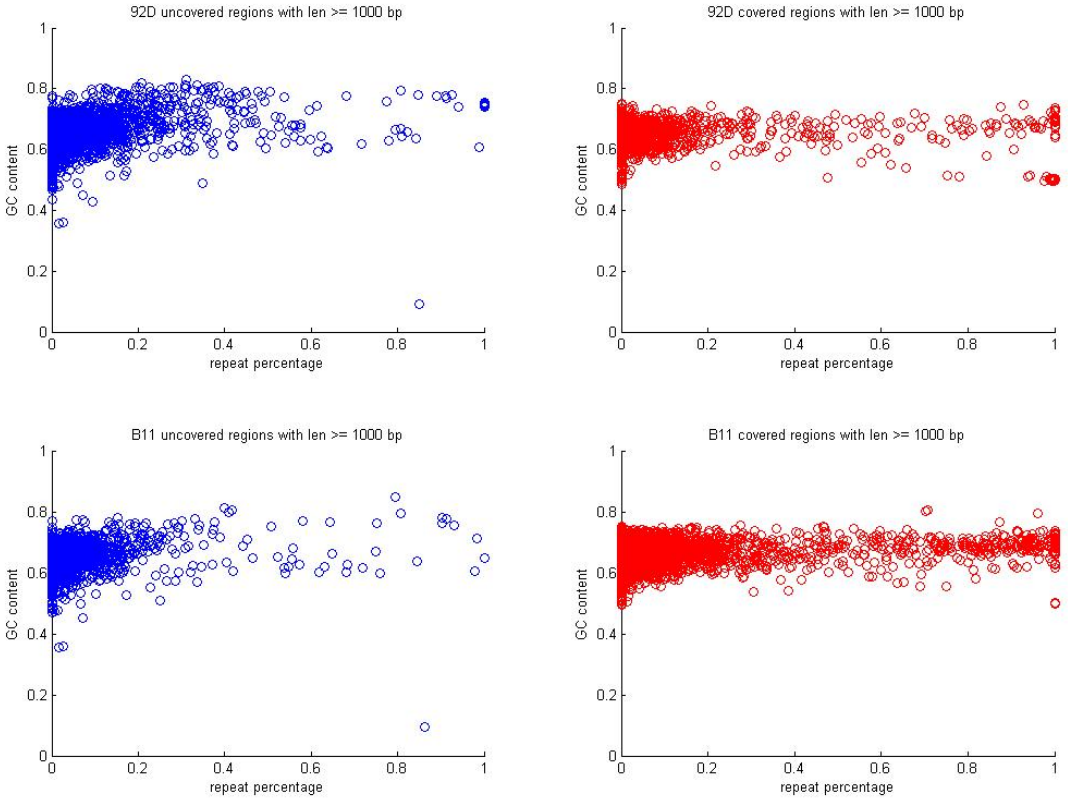


Figure SI 7| Comparing the GC and repeat content of the mapped and unmapped Illumina reads from strains B11 and 92D. Strains B11 and 92D were sequenced to 13 and 32X coverage, respectively. When comparing the reads that mapped to the reference genome to those that did not, in both instances there appears to be little difference in the repeat content and only a small difference in the GC content where the mapped reads having a slightly higher GC content.

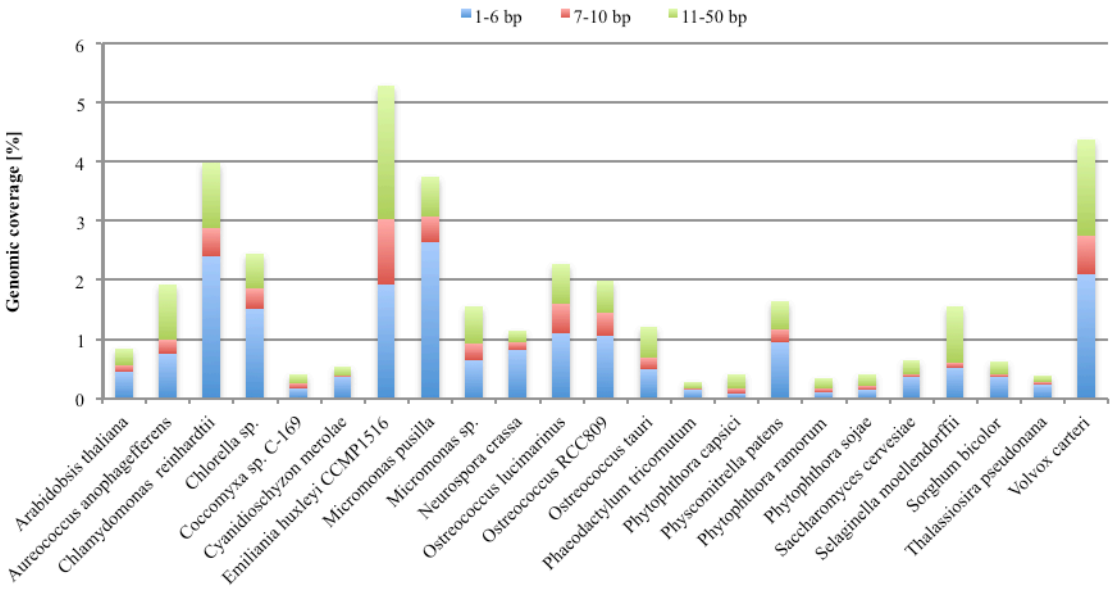


Figure SI 8| Genomic coverage of tandem repeats in *E. huxleyi* and other genomes (see Table SI 18). A mismatch and indel penalty of -5 allows the detection of only slightly imperfect repeats that still have a good repeat structure. Among this diverse group of species *E. huxleyi* stands out as the species with the highest TR coverage. If highly degenerate repeats are included, the repeat coverage is significantly higher for *E. huxleyi* (see Section 2.10.3 and Figure SI 10).

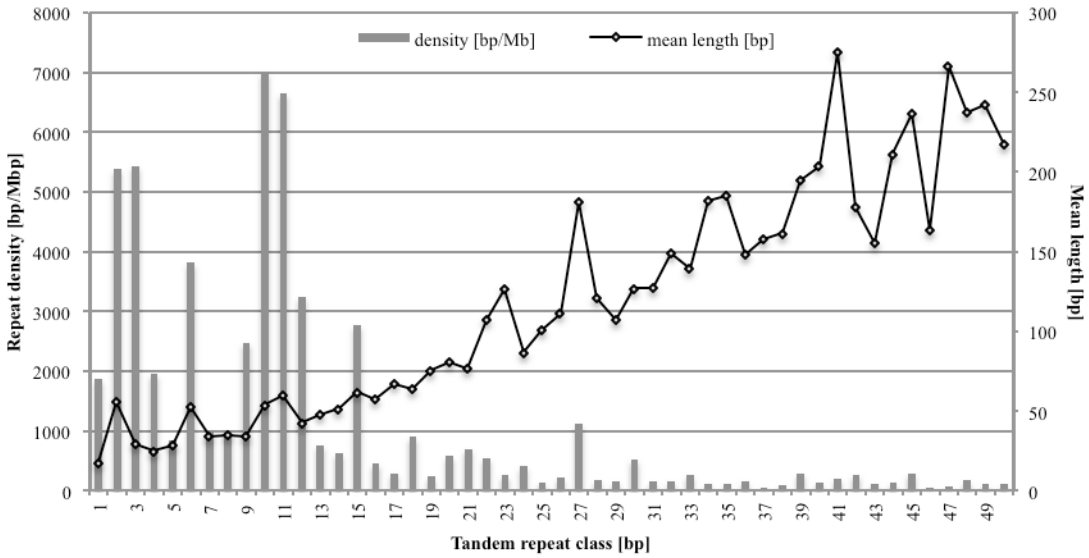


Figure SI 9| Genomic densities (columns) and mean lengths (black dots) of individual tandem repeat classes in the *E. huxleyi* genome.

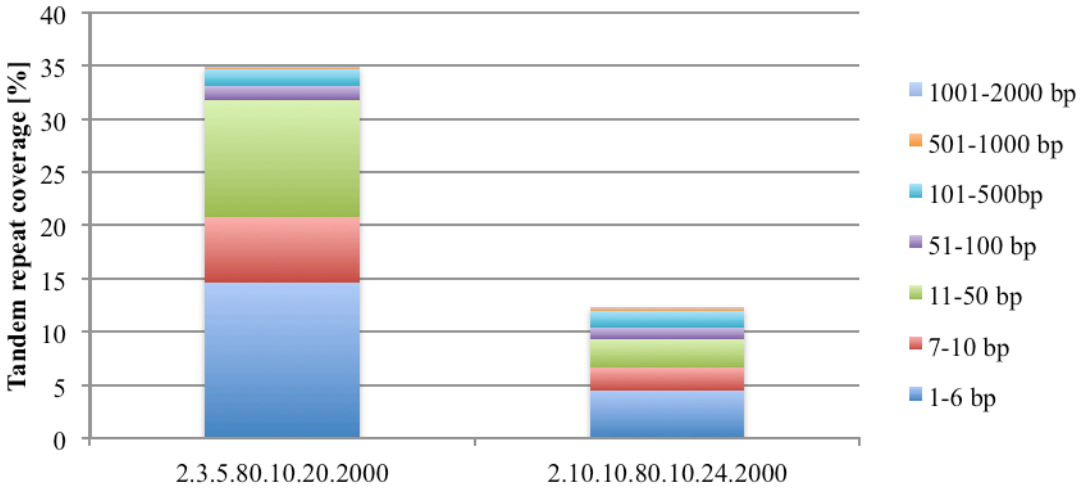
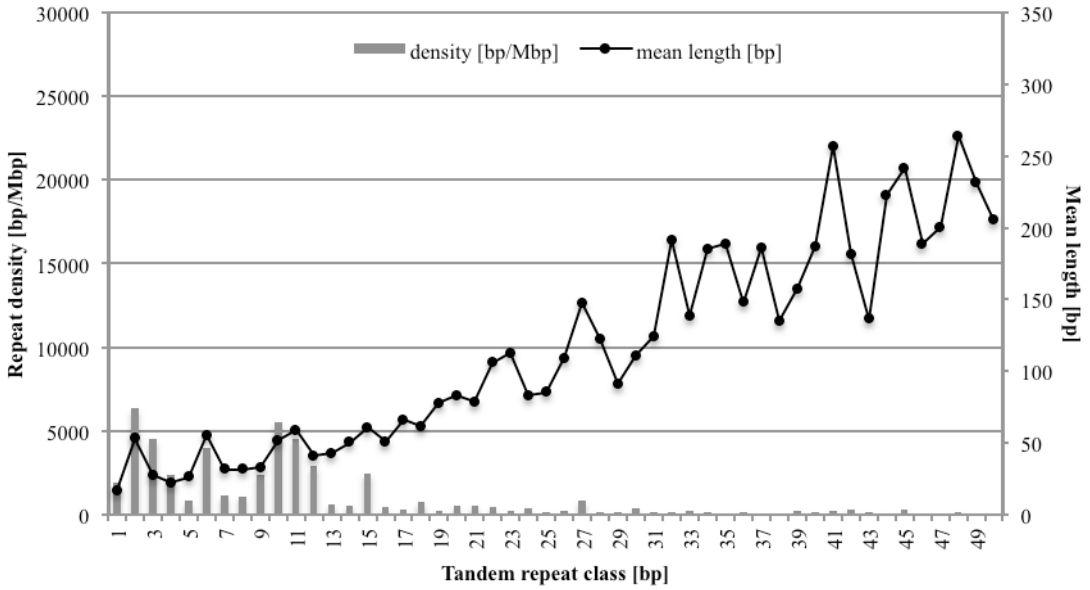


Figure SI 10| Repeat coverage [%] (y-axis) for two different sets of search parameters (x-axis) using the TRF program and for different pattern size ranges (in bp) as shown in the legend. Description on the x-axis is composed of the TRF search parameters used in each run. The more relaxed parameters specified on the left allow for the detection of low complexity regions.

(a)



(b)

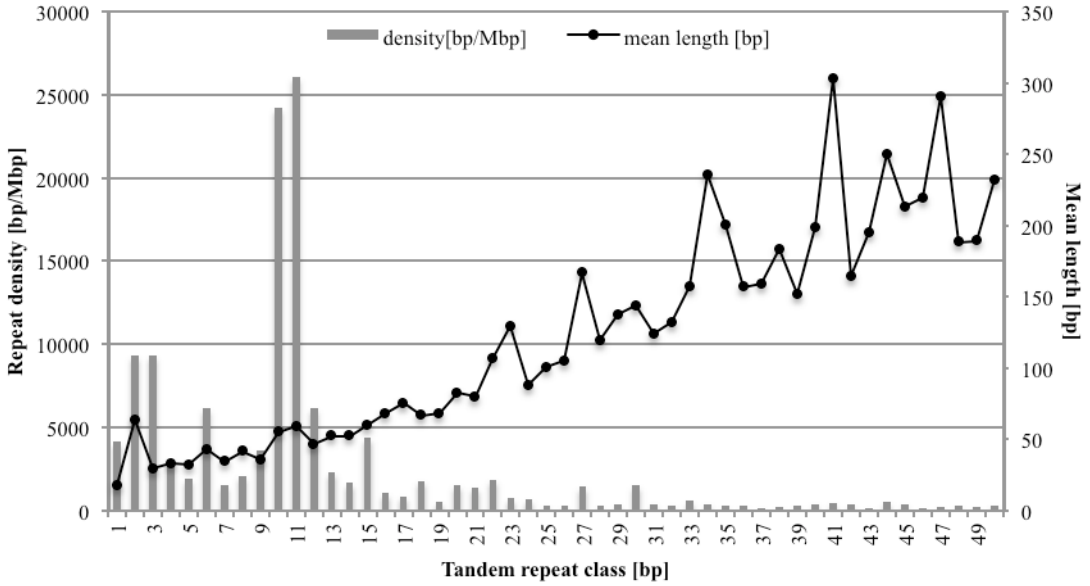


Figure SI 11| Densities (columns) and mean lengths (black dots) of individual tandem repeat classes (a) in intergenic regions and (b) in introns of the *E. huxleyi* genome.

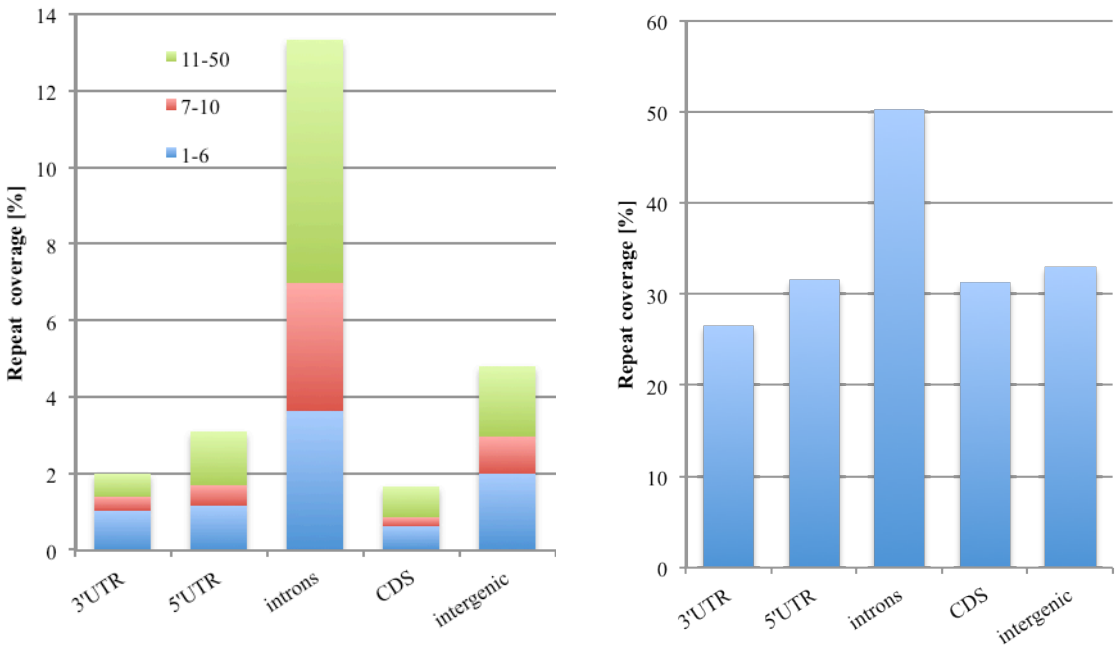


Figure SI 12| Tandem repeat coverage in different genomic regions (i) for high mismatch and indel penalties using Phobos (left plot) and (ii) for very relaxed search parameters allowing to detect low complexity regions with TRF (right plot).

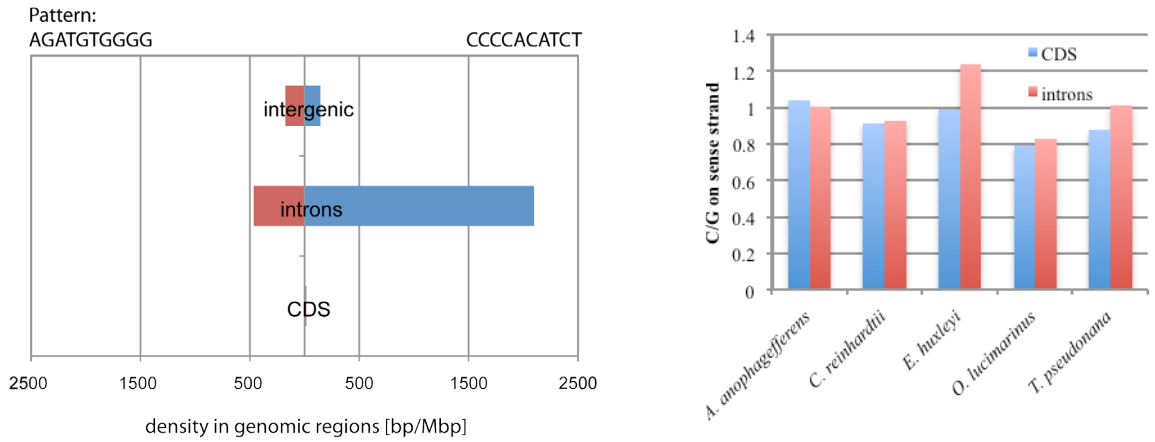
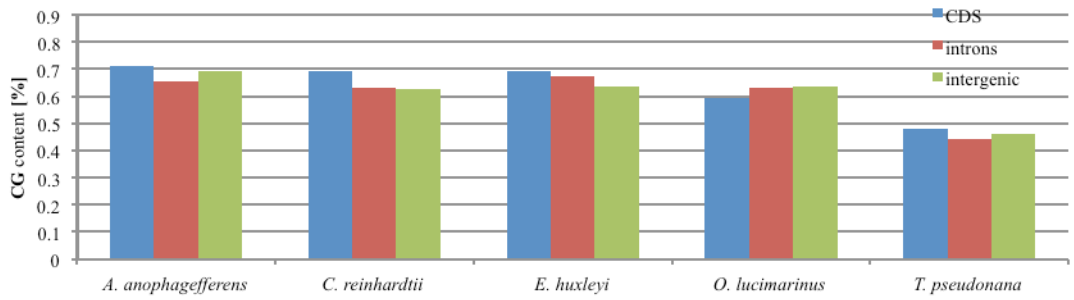


Figure SI 13| Strandedness of TR patterns and C/G base usage in introns and CDS regions. The left panel shows the repeat densities of two 10 bp repeat patterns in different genomic regions. For introns and CDS regions, the density is given for the sense strand of the corresponding gene. For intergenic regions the density is computed for the strand found in the genome assembly. The two patterns are reverse complements to each other. For neutral selection, identical repeat densities of both patterns would be expected on the sense strand in genes. The deviation from identical characteristics on the sense and antisense strands is called a strandedness. The right panel shows the C/G base usage on the sense strand in CDS regions and introns. The high C/G on the sense strand in introns of *E. huxleyi* is correlated with repeat patterns that favor C versus G on this strand.

(**a**)



(**b**)

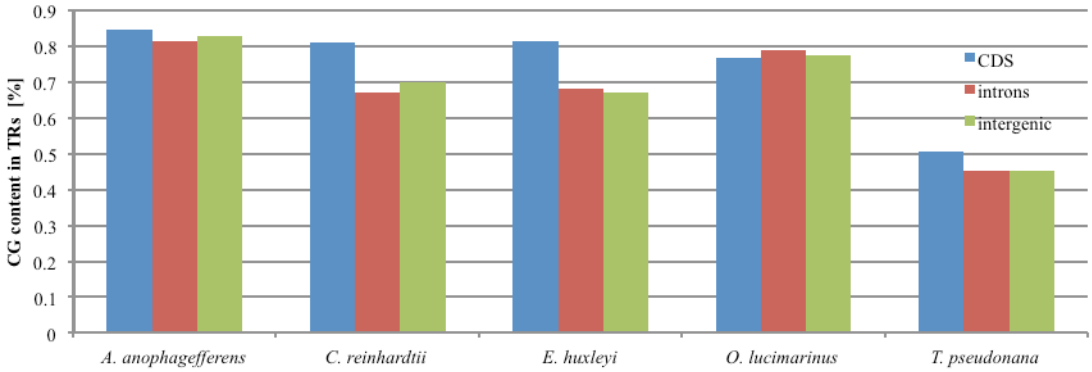


Figure SI 14| (**a**) GC content in percent of different genomic regions in the genome of *E. huxleyi* and other genomes and (**b**) GC content in the tandem repeats detected within these genomes using Phobos. In all genomes and regions, the GC content is higher in TRs than in the corresponding region.

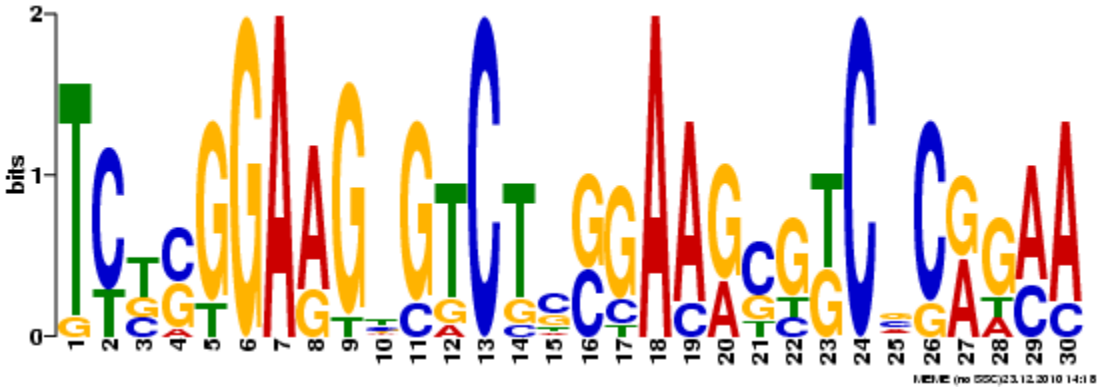


Figure SI 15| Sequence logo plot of *E. huxleyi* LI818-like promoter cis-acting elements.

References

1 Guillard, R. R. & Ryther, J. H. Studies on marine planktonic diatoms. I. *Cyclotella nana Hustedt* and *Detonula confervacaea* (Cleve) Gran. *Can. J. Microbiol.* **8**, 229-239 (1962).

2 Wilson, K. *Preparation of genomic DNA from bacteria.*, (John Wiley & Sons, Inc., 2001).

3 Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91-96. (2003).

4 Sánchez Puerta, M. V., Bachvaroff, T. R. & Delwiche, C. F. The complete plastid genome sequence of the haptophyte *Emiliania huxleyi*: a comparison to other plastid genomes. *DNA Res* **12**, 151-156, doi:10.1093/dnares/12.2.151 (2005).

5 Sánchez Puerta, M. V., Bachvaroff, T. R. & Delwiche, C. F. The complete mictochondrial genome sequences of the haptophyte *Emiliania huxleyi* and its relation to heterokonts. *DNA Res* **11**, 67-68 (2004).

6 Colbourne, J. K. *et al.* The Ecoresponsive Genome of *Daphnia pulex*. *Science* **331**, 555-561 (2011).

7 Para, G., Bradnam, K., Ning, Z., Keane, T. & Korf, I. Assessing the gene space in draft genomes. *Nucleic Acid Res.* **37**, 289-297 (2009).

8 Wahlund, T. M. *et al.* Analysis of expressed sequence tags from calcifying cells of the marine coccolithophorid, *Emiliania huxleyi*. *Marine Biotechnology* **6**, 278-290 (2004).

9 Kent, W. J. BLAT-The BLAST-like alignment tool. *Genome Res.* **12**, 656-664 (2002).

10 Wahlund, T. M., Zhang, X. & Read, B. A. in *Proceedings of the INA Workshop on extant Coccolithophorid research* Vol. 50 *Advances in the biology, ecology and taxonomy of extant calcareous nanoplankton* (ed M. Triantaphyllou) 145-155 (Micropaleontology, Crete, 2005).

11 Benton, D. Recent changes in the GenBank on-line service. *Nucleic Acids Res.* **18**, 1517-1520 (1990).

12 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment searach tool. *J. Mol. Biol.* **215**, 403-410 (1990).

13 Salamoc, A. A. & Solovyev, V. V. *Ab initio* gene finding in *Drosophilia* genomic DNA. *Genome Res.* **10**, 516-522 (2000).

14 Borodovsky, M. *et al.* Detection of new genes in a bacterial genome using Markov models for three gene classes. *Nucleic Acids Res* **23**, 3554-3562 (1995).

15 Birney, E. & Durbin, R. Using Genewise in the *Drosophila* annotation experiment. *Genome Res.* **10**, 547-548 (2000).

16 Solovyev, V., Korsarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7**, 1-12 (2006).

17 Foissac, S., Bardou, P., Moisan, A., Cros, M. J. & Schiex, T. EUGENE'HOM: A genetic similarity-based gene finder using multiple homologous sequences. *Nucleic Acids Res.* **31**, 3742-3745 (2003).

18 Nielsen, H., Brunak, S. & von Heijne, G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.* **12**, 3-9 (1999).

19 Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567-580 (2001).

20 Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116-120 (2005).

21 Schneider, M., Tognolli, M. & Bairoch, A. The Swiss-Prot protein knowledgebase and ExPASy: providing the plant community with high quality proteomic data and tools. *Plant Physiol. Biochem.* **42**, 1013-1021 (2004).

22 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27-30 (2000).

23 Koonin, E. V. *et al.* A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5**, R7 (2004).

24 Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).

25 Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* **20**, 3643-3646 (2004).

26 Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein doimain families based on seed alignments. *Proteins* **28**, 405-420 (1997).

27 Yamada, K. *et al.* Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**, 842-846 (2003).

28 Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575-1584 (2002).

29 Parfrey, L. W. *et al.* Evaluating support for the current classification of eukaryotic diversity. *PLoS Genet.* **2** (2006).

30 Nozaki, H., Yang, Y., Maruyama, S. & Suzaki, T. A case study for effects of operational taxonomic units from intracellular endoparasites and ciliates on the eukaryotic phylogeny: phylogenetic position of the haptophyta in analyses of multiple slowly evolving genes. *PLos One* **7**, doi:10.1371/journal.pone.0050827 (2012).

31 Cuvelier, M. L. *et al.* Targeted metagenomics and ecology of globally important uncultured eukaryotic phytoplankton. *Proc. Natl. Acad. Sci. USA* **107**, 14679-14684 (2010).

32 Altschul *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database serach programs. *Nucleic Acids Res.* **25**, 3389-3402 (1997).

33 Katoh, K. & Frith, M. C. Adding unaligned sequences into an existing alignment using MAFFT and LAST. *Bioinformatics* **28**, 3144-3146 (2012).

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

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

36 Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & Mclnerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6**, 29 (2006).

37 Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A Genomic Perspective on Protein Families. *Science* **278**, 631-637 (1997).

38 Edgar, R. C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5** (2004).

39 Moustafa, A. *et al.* Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**, 1724-1726 (2009).

40 Burki, F. *et al.* Phylogenomics reshuffles the eukaryotic supergroups. *PloS One* **2**, e790 (2007).

41 Arumuganathan, K. & Earle, E. D. Estimates of nuclear DNA amounts of plants by flow cytometry. *Plant Mol. Biol. Rep.* **9**, 229-241 (1991).

42 Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380 (2005).

43 Lefébure T, S. M. Evolution of the core and pan-genome of *Streptococcus*: positive selection, recombination, and genome composition. *Genome Biol.* **8**, R71 (2007).

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

45 Howe, K., Bateman, A. & Durbin, R. QuickTree:building huge neighbor-joining trees of protein sequences. *Bioinformatics* **18**, 1546-1547 (2002).

46 Felsenstein, J. *PHYLIP (Phylogeny Inference Package) version 3.6.* (Department of Genome Sciences, University of Washington, Seattle).

47 Anisimova, M. & Gascuel, O. Approximate likelihood-ratio test for branches: A fast, accurate and powerful alternative. *Syst. Biol.* **55**, 539-552 (2006).

48 Consortium. & UniProt. The universal protein resource (UniProt) in 2010. *Nucleic Acids Res.* **38**, D142-148 (2010).

49 Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acid Res.* **31**, 3497-3500 (2003).

50 Dereeper, A. *et al.* Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acid Res.* **36**, W465-469 (2008).

51 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 534-543 (2003).

52 Quesneville, H. *et al.* Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol.* **1**, 166-175 (2005).

53 Flutre, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, e16526 (2011).

54 Bao, Z. & Eddy, S. R. Automated *do novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269-1276 (2002).

55 Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, i152-i158 (2005).

56 Huang, X. On global sequence alignment. *Comput. Appl. Biosci.* **10**, 227-235 (1994).

57 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acid Res.* **27**, 573-580 (1999).

58 Jurka, J. *et al.* Repbase update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res.* **110**, 462-467 (2005).

59 Bateman, A. *et al.* Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acid Res.* **27**, 260-262 (1999).

60 Maumus, F. *et al.* Potential impact of stress activated retrotransposons on genome evolution in a marine diatom. *BMC Bioinformatics* **10**, 624 (2009).

61 Bowler, C. *et al.* The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* **456**, 239-244, doi:10.1038/nature07410 (2008).

62 Cock, J. M. *et al.* The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* **465**, 617-621 (2010).

63 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acid Res.* **35**, W265-268. (2007).

64 Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acid Res.* **35**, 3100-3108 (2007).

65 Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acid Res.* **25**, 955-964 (1997).

66 Kohany, O., Gentles, A. J., Hankus, L. & Jurka, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**, 474 (2006).

67 Merkel, A. & Gemmell, N. Detecting microsatellites in genome data: variance in definitions and bioinformatic approaches cause systematic bias. *Evol. Bioinform.*, 1-6 (2008).

68 Leclercq, S., Rivals, E. & Jarne, P. Detecting microsatellites within genomes: significant variation among algorithms. *BMC Bioinformatics* **8**, 125 (2008).

69 Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics resaerch. *Bioinformatics* **21**, 3674-3676 (2005).

70 Guindon, S., Lethiec, F., Duroux, P. & Gascuel, O. PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acid Res.* **33**, W557-559 (2005).

71 Koziol, A. G. & Durnford, D. G. Euglena light-harvesting complexes are encoded by multifarious polyprotein mRNAs that evolve in concert. *Mol Biol Evol.* **25**, 92-100 (2008).

72 Green, B. R. in *Evolution of Aquatic Photoautotrophs* (eds P.G. Falkowski & A.H. Knoll) (Academic Press, 2007).

73 Nymark, M. *et al.* An integrated analysis of molecular acclimation to high light in the marine diatom *Phaeodactylum tricornutum*. *PLoS ONE* **4**, e7743 (2009).

74 Lefebvre, S. C. *et al.* Characterization and expression analysis of the Lhcf gene family in *Emiliania huxleyi* (Haptophyta) reveals differential resonses to light and CO2. *J. Phycol.* **46**, 123-134 (2010).

75 Park, S., Jung, G., Hwang, Y. S. & Jin, E. Dynamic response of the transcriptome of a psychrophilic diatom, *Chaetoceros neogracile*, to high irradiance. *Planta* **231**, 349-360 (2010).

76 Zhu, S. H. & Green, B. R. Photoprotection in the diatom *Thalassiosira pseudonana*: Role of LI818-like proteins in response to high light stress. *Biochim. Biophys. Acta.* **1797**, 1449-1457 (2010).

77 Moseley, J. L. *et al.* Adaptation to Fe-deficiency requires remodeling of hte photosynthetic apparatus. *EMBO* **21**, 6709-6720 (2002).

78 Naumann, B. *et al.* Comparative quantitative proteomics to investigate the remodeling of bioenergetic pathays under iron deficiency in *Chlamydomonas reinhardtii*. *Proteomics* **7**, 3964-3979 (2007).

79 Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acid Res.* **37**, W202-208 (2009).

80 Gobler, C. J. *et al.* Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics. *Proc Natl Acad Sci U S A* **108**, 4352-4357, doi:10.1073/pnas.1016106108 (2011).

81 Merchant, S. S. *et al.* The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* **318**, 245-250, doi:10.1126/science.1143609 (2007).

82 Matsuzaki, M. *et al.* Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* **428**, 653-657, doi:10.1038/nature02398 (2004).

83 Fritz-Laylin, L. K. *et al.* The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* **140**, 631-642, doi:10.1016/j.cell.2010.01.032 (2010).

84 Palenik, B. *et al.* The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci U S A* **104**, 7705-7710, doi:10.1073/pnas.0611046104 (2007).

85 Aury, J.-M. *et al.* Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* **444**, 171-178, doi:10.1038/nature05230 (2006).

86 Tyler, B. M. *et al. Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313**, 1261-1266, doi:10.1126/science.1128796 (2006).

87 Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511, doi:10.1038/nature01097 (2002).

88 Kaneko, T. & Tabata, S. Complete genome structure of the unicellular cyanobacterium *Synechocystis sp. PCC6803*. *Plant Cell Physiol* **38**, 1171-1176, doi:9435137 (1997).

89 Armbrust, E. V. *et al.* The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79-86, doi:10.1126/science.1101156 (2004).

90 Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815, doi:10.1038/35048692 (2000).

91 Blanc, G. *et al.* The *Chlorella variabilis NC64A* genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. *Plant Cell* **22**, 2943-2955, doi:10.1105/tpc.110.076406 (2010).

92 Blanc, G. *et al.* The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. *Genome Biol* **13**, R39, doi:10.1186/gb-2012-13-5-r39 (2012).

93 Worden, A. Z. *et al.* Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. *Science* **324**, 268-272, doi:10.1126/science.1167222 (2009).

94 Galagan, J. E. *et al.* The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**, 859-868, doi:10.1038/nature01554 (2003).

95 Lamour, K. H. *et al.* Genome sequencing and mapping reveal loss of heterozygosity as a mechanism for rapid adaptation in the vegetable pathogen *Phytophthora capsici*. *Mol Plant Microbe Interact* **25**, 1350-1360, doi:10.1094/MPMI-02-12-0028-R (2012).

96 Rensing, S. A. *et al.* The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. *Science* **319**, 64-69, doi:10.1126/science.1150646 (2008).

97 Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546, 563-547, doi:8849441 (1996).

98 Banks, J. A. *et al.* The *Selaginella* genome identifies genetic changes associated with the evolution of vascular plants. *Science* **332**, 960-963, doi:10.1126/science.1203810 (2011).

99 Paterson, A. H. *et al.* The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**, 551-556, doi:10.1038/nature07723 (2009).

100 Prochnik, S. E. *et al.* Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. *Science* **329**, 223-226, doi:10.1126/science.1188800 (2010).

101 Xu, P. *et al.* The genome of *Cryptosporidium hominis*. *Nature* **431**, 1107-1112, doi:10.1038/nature02977 (2004).

102 Pain, A. *et al.* Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science* **309**, 131-133, doi:10.1126/science.1110418 (2005).

103 Wood, V. *et al.* The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415**, 871-880, doi:10.1038/nature724 (2002).

104 C. elegans Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012-2018, doi:9851916 (1998).

105 Celniker, S. E. *et al.* Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol* **3**, RESEARCH0079, doi:12537568 (2002).

106 Adams, M. D. *et al.* The Genome Sequence of *Drosophila melanogaster*. *Science* **287**, 2185-2195, doi:10731132 (2000).

107 Palenik, B. *et al.* The genome of a motile marine *Synechococcus*. *Nature* **424**, 1037-1042, doi:10.1038/nature01943 (2003).

108 Dufresne, A. *et al.* Genome sequence of the cyanobacterium *Prochlorococcus marinus SS120*, a nearly minimal oxyphototrophic genome. *Proc Natl Acad Sci U S A* **100**, 10020-10025, doi:10.1073/pnas.1733211100 (2003).