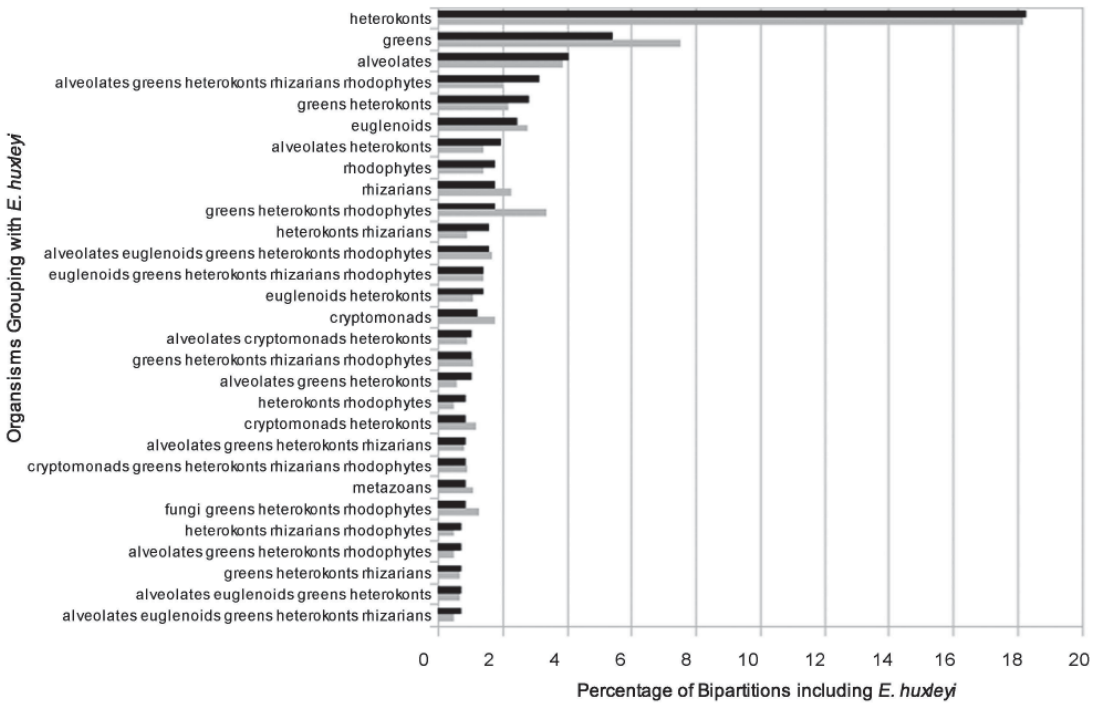
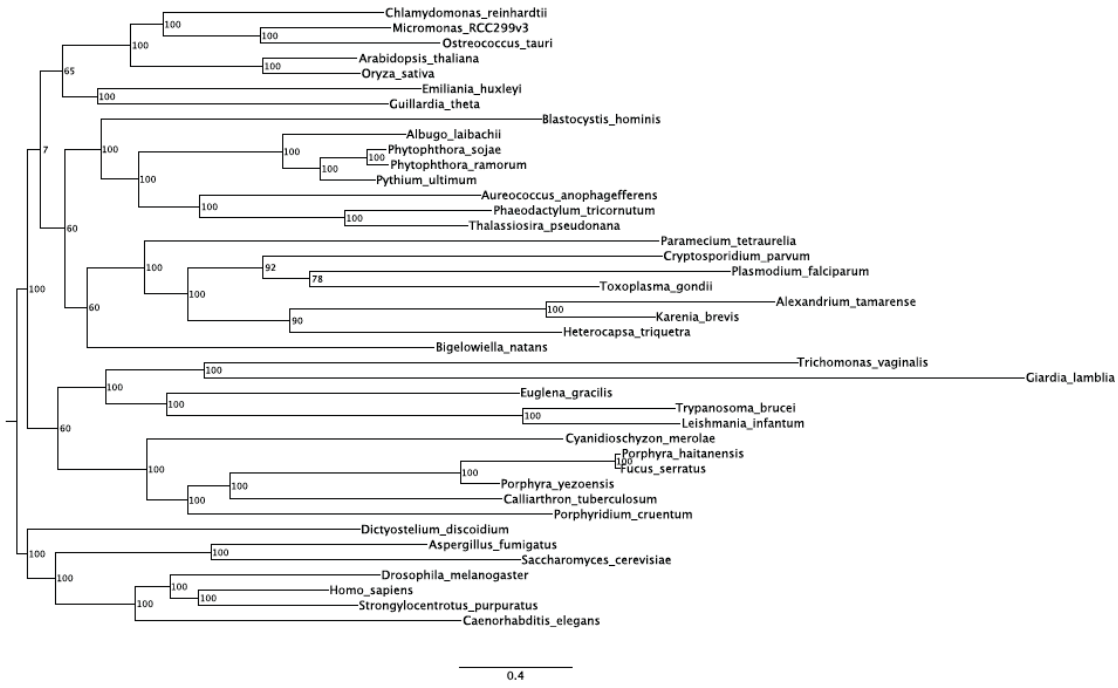
doi:10.1038/nature12221

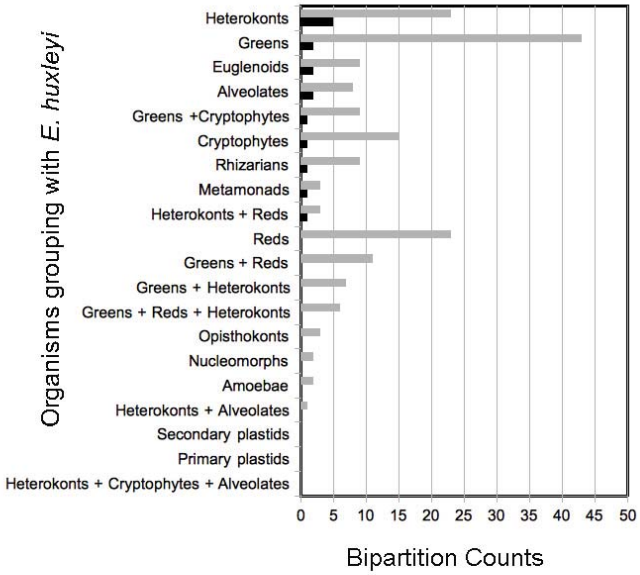
a



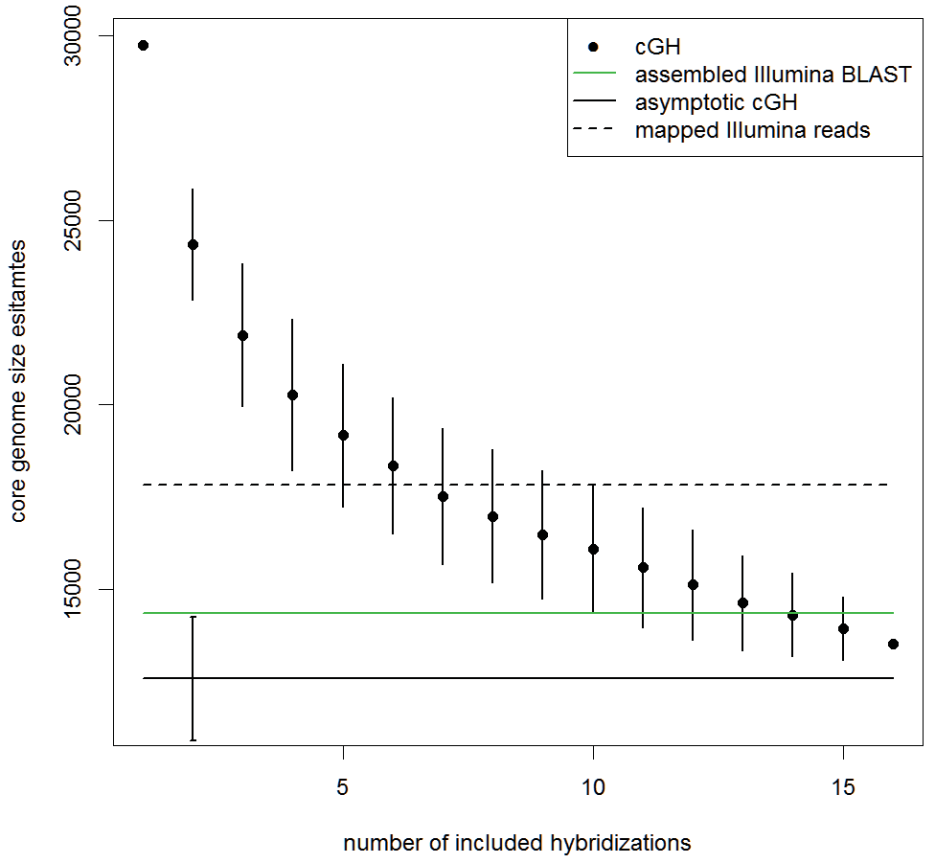
b



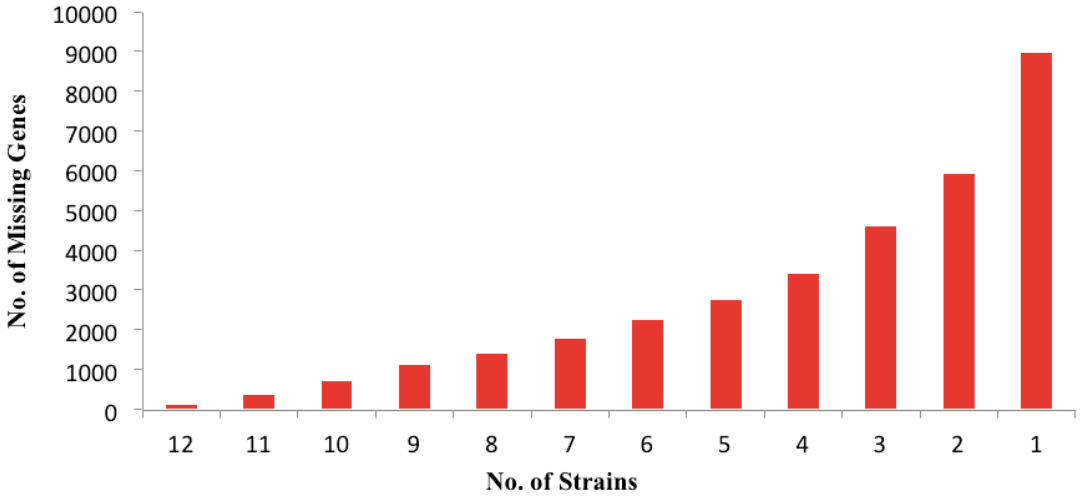
c



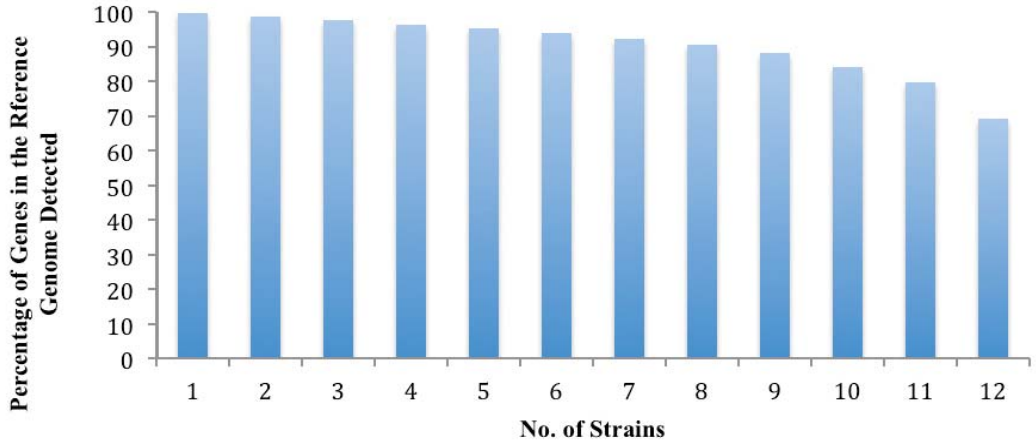
Supplementary Figure 1| Phylogenomic analysis using single gene trees and a concatenated phylogeny. a) Counts of otherwise monophyletic bipartitions containing *E. huxleyi* from 1563 single protein trees. Black bars represent counts from only those trees in which the branch leading to *E. huxleyi* and its sister taxon is represented by a bootstrap value of 70 or above. Grey bars represent counts from all trees. **b)** Concatenated phylogeny of 228 *E. huxleyi* genes showing **t**he best 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. **c)** 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.



Supplementary Figure 2| Asymptotic analysis of the core genome size estimates obtained from comparative genomic hybridization (cGH). The core-genome size from cGH is 12,757 +/- 1,673 genes. From an exponential fit the estimated asymptotic size is ~13,000 (horizontal black line). The unassembled Illunmina comparative approach yielded a core-genome of 17,826 genes (dotted black line) while the assembled reads yielded a core of 14,344 genes (green), see text. Vertical bars indicate ± one standard deviation from 200 random re-samples. The asymptotic estimate is the average from repeated non-linear exponential fits under 500 random permutations of the order of hybridization data and dropping estimates lower than 8,000.

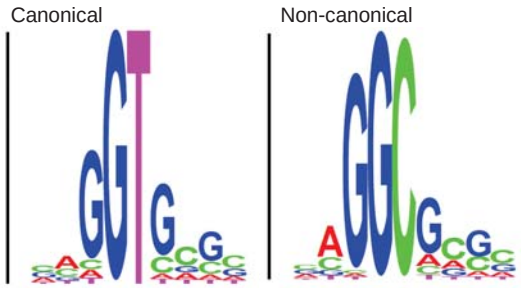


Supplementary Figure 3| The cumulative number of genes in the reference genome not detected in the indicated number of strains (~9000 variable genes are not detected in at least one strain, but all of these genes are detected in the genetic repository of all strains). Illumina reads generated from genomic DNA extracted from 13 different strains, including AWI1516, were mapped to the refined gene model set assembled from the reference genome. Genes were detected using a threshold value equivalent to the coverage of ~50% of the gene length), whereby genes with < 50% coverage with mapped reads were regarded as missing. The total number of missing genes is represented on the Y-axis and the number of strains is indicated along the X-axis.

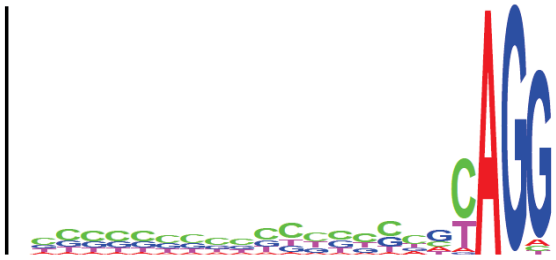


Supplementary Figure 4| The percentage of reference genes detected across multiple strains of *E. huxleyi*. Illumina reads generated from genomic DNA extracted from13different strains, including AWI1516, were mapped to the reduced gene model set assembled from the reference genome. Genes with $\geq$ 50% coverage based on mapped reads were called present, and the total number of genes found in each strain was then compared to the 30,569 predicted gene models in the reference genome.

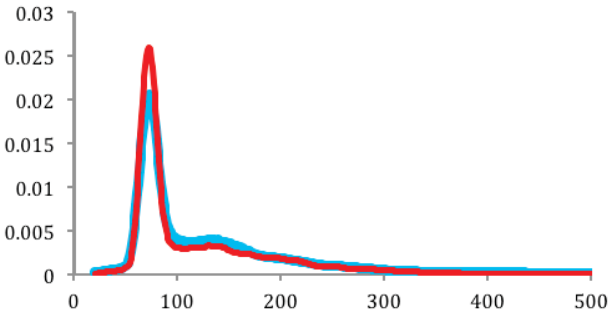
a. Intron donor site logo



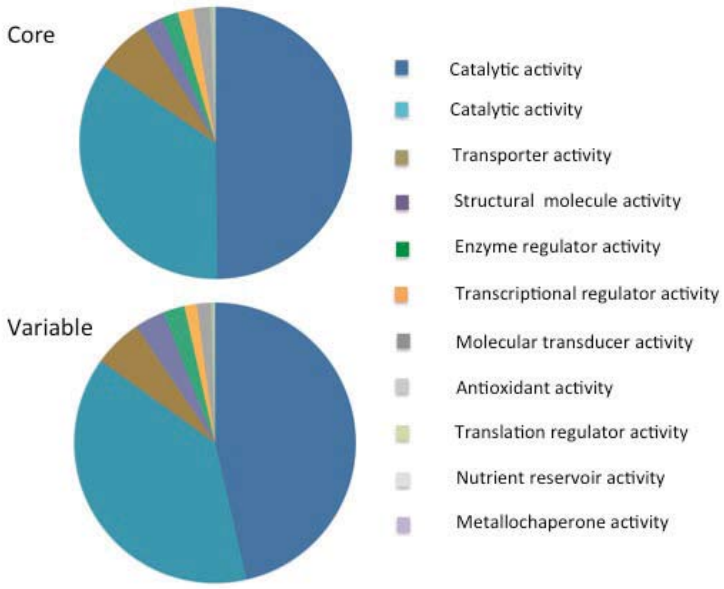
b. Intron acceptor site logo



c. Intron length distribution



Supplementary Figure 5| Logo images of intron donor and acceptor sites. **a,** The sequence logo for the *E. huxleyi* canonical and non-canonical donor sites in the data set with the most frequent nucleotides being CAGGT|GCGC and CAGGC|GCGC. **b,** The sequence logo for the acceptor sites in the *E. hxuleyi* data set with the most frequent nucleotides being GT/GC|AGG. **c,** Intron length distribution where GC donor introns are shown in red and the GC donor introns are in blue. The vertical axis shows the percentage of introns and the horizontal axis, the base pair length of the introns.



Supplementary Figure 6| Pie charts of the function categories of genes in the pan genome (core and variable sets based on the reference strain CCMP 1516). Colors are coded according to the insert legend, with percentages given in parentheses for the core versus variable genomes. Note that the major difference is the percentage of unknown/orphan genes.

Supplementary Table 1| *Emiliania huxleyi* CCMP 1516 reference genome assembly statistics.

General assembly	
Total number of reads	2991780
Genome size (Mb)	141.7
Read depth coverage (fold)	10
Scaffold sequence gap (%)	6.98
Scaffold number	6995
Scaffold N50/L50	109/404.8 Kb
Contig number	26075
Contig N50/L50	1315/29.7 Kb
Average Properties	
Gene length (bp)	1718
Transcript length (bp)	1129
Protein length (aa)	346
Exon frequency number per gene	3.65
Exon length (bp)	308
Intron length (bp)	242
Gene density per Mbp	233
Estimating genome completeness:	
Conserved eukaryotic single copy genes[a]	97%
Core eukaryotic genes mapping approach (CEGMA)[b]	97%
Genes were validated by the following evidence:	
Total number of gene models	30,569 (100%)
Start + stop codons	25,315 (83%)
EST support (# of genes / %)	15,642 (51%)
RNASeq (# of genes / %)	29,637 (96%)
Tiling array (# of genes / %)	23,911 (72%)
NR hit (# of genes / %)	21,143 (69%)
SwissProt hit (# of genes / %)	19,722 (65%)
Pfam hit (# of genes / %)	10496 (34%)

[a]Based on the 716 highly conserved single-copy eukaryotic genes detailed in the *Daphnia pulex* genome paper.
[b]Based on the 458 highly conserved single-copy eukaryotic genes from the CEMGA set.

Supplementary Table 2| Genes putatively identified as encoding enzymes involved in the catabolism of DMSP. tBlastn homology searches were performed against the genome using corresponding protein sequences from *Halomonas* sp. HTNK1.

Gene Description	Gene Name	No. of Orthologs	Protein ID	Allelic Variant	EST evidence	e-value	Homolog
Choline dehydrogenase	DddA	6	441720		Y	1.32E-42	*Halomonas sp. HTNK1*
			249893		Y	2.94E-41	
			115992		Y	3.29E-35	
			467417	463443	Y	1.18E-34	
			419045	426209	Y	5.53E-25	
			108462		Y	4.89E-10	
DMSP lyase	DddD	2	558015		Y	1.56E-63	*Halomonas sp. HTNK1*
			558016		Y	2.75E-55	
Enoyl-CoA hydratase	AcuK	5	436143		Y	1.51E-31	*Halomonas sp. HTNK1*
			355649		Y	1.51E-31	
			367461	360949	Y	8.00E-12	
			558212		N	1.80E-09	
			440242		Y	6.28E-08	
Iron containing alcohol dehydrogenase	DddB	2	437150	357847	Y	1.79E-89	*Pseudomonas sp.* J465
			451417	244398	Y	9.28E-25	
Aldehyde dehydrogenase	DddC	4	445304	434256	Y	2.10E-25	*Halomonas sp. HTNK1*
			358314		Y	3.97E-24	
			558209		Y	1.16E-20	

			558211		Y	6.58E-20	
MMPA-CoA ligase	DmdB	7	429472	427245	Y	1.30E-20	*Ruegeria pomeroyi*
			416064	432437	Y	3.06E-15	
			432438		Y	3.06E-15	
			59160		N	1.75E-13	
			216570		N	1.75E-13	
			194364		Y	4.27E-12	
			71614	61069	Y	1.62E-11	
MMPA-CoA dehydrogenase	DmdC	3	452720	459883/458546	Y	1.38E-37	*Ruegeria pomeroyi*
			453920		Y	9.47E-42	
			441379		Y	8.18E-10	
MTA-CoA hydratase	DmdD		105067	442456	N	2.16E-06	*Ruegeria pomeroyi*

Supplementary Table 3| Proteins involved in sphingolipid and omega-3- polyunsaturated fatty acid synthesis.

Protein Description	Functionally Characterized	No. of Paralogs	Protein ID	Core vs. Variable Genome
Genes involved in sphingolipid synthesis				
Serine Palmitoyl Transferase (SPT)	yes	1	432901	core
Ketosphinganine reductase	no	2	437991	core
	no		434961	core
Ceramide synthase (LAG1)	yes	1	200862	nd
Ceramide glucosyltransferase	no	1	436316	core
Sphingolipid LCB Δ4-desaturase	yes	1	461341	core
Sphingolipid LCB Δ8-desaturase	yes	1	72565	core
Sphingolipid LCB Δ9-methyl transferase	yes	1	365614	core
Genes involved in the biosynthesis of omega-3 LC-PUFAs				
Δ12-desaturase	no	2	432965	nd
	no		424263	core
Δ15-desaturase	no		468588	core
Δ6-desaturase	no		417285	core
Δ9-elongase	no	2	433098	core
	no		434059	core
Δ8-desaturase	no		216445	core
Δ5-desaturase	no		443389	core
Δ5-elongase	no	3	369636	core
	no		414244	core
Δ4-desaturase	no		433467	core

Supplementary Table 4| Genes encoding proteins potentially involved in biomineralization processes.

Protein Description	Family member/ Subfamily/ Complex subunit	Number of Paralogs	Protein ID	Core vs Variable
Proton/calcium exchanger	CAX	7	72273	core
			66584	variable
			416800	core
			415715	core
			103021	core
			223499	nd
Cation/calcium exchanger	CCX	2	203920	core
			449053	core
Potassium dependent sodium/calcium exchangers	NCKX	2	205210	core
			447939	variable
Potassium independent sodium/calcium exchangers	NCX	1	454623	core
Calcium pumps	ECA	6	522052	nd
			522053	nd
			62350	core
			426283	core
			101130	variable
			466567	core
$Ca^{2+/}Mg^{2+}$-permeable cation channels Transient Receptor Potential	TRP	4	449985	core
			468587	core
			455760	core
			107737	core
4 Domain voltage-gated Ca^{2+} channels	CAV	2	448907	core
			448526	core

	VDC	2	97870	core
			196523	nd
IP3R	IPR	2	209909	nd
			250663	variable
Eukaryotic Na^+/H^+ exchanger	NAH1	5	465194	core
			434034	core
			467182	core
			196090	core
			464223	core
Bacterial Na^+/H^+ exchanger	NHA	5	219535	variable
			447659	core
			105293	core
			521936	nd
Voltage gated proton channel	HVCN	1	631975	nd
V-ATPase	ATPVA	2	464767	core
	ATPVA		439538	core
	ATPVB	1	435128	core
	ATPVC	3	359783	core
	ATPVC		451883	core
	ATPVC		558234	nd
	ATPVD	2	413949	core
	ATPVD		420005	core
	ATPVE	1	433060	core
	ATPVF	1	352209	core

	ATPVG	1	355949	core
P type ATPase (PM H^+ ATPase)		2	67081	core
			76123	variable
H^+ PPase		2	415047	core
			439740	core
Putative Ci Transporters				
SLC4 Na^+ independent Cl^-/HCO_3^- exchangers	SLC4	3	436956	core
			450694	core
			466232	core
	AE	3	99943	core
			198643	core
			200137	core
Carbonic anhydrases	Alpha CA		62679	core
			456048	core
			558406/437352	nd
	Beta CA		469462	nd
			239690	nd
			115240	variable
			233823	variable
	Delta CA		436031	core
			469783	core
			222602	core
			229394	core
			227360	variable
			195575	variable

	Gamma CA	432493	variable
		373149	variable
		63173	core
		214705	core

Table 5| A comparison of the sanger sequenced CCMP 1516 reference genome with the CLC assemblies of three other *E. huxleyi* strains deeply sequenced using the Illumina platform.

Strain	Flow Cytometry Haploid Genome Size (Mb)	Genome Assembly Size (Mb)	Illumina Read Depth Coverage	CEGMA % Completeness	Ref Genome Coverage (%)	Non-aligning sequence[a] (Mb)	Strain Specific Sequence[b] (Mb)
M202 (EH2)	120 +/- 0.01	117.7	322	95	44.8	54.9 (47%)	40.7 (35%)
M222 (VAN556)	133 +/- 0.02	109.7	265	95	47.8	42.8 (39%)	21.4 (20%)
92A (CCMP379)	99 +/- 0.00	85.6	352	95	46.9	19.9 (23%)	8.8 (10%)

[a]Non-aligning sequence, the amount of assembled non-gapped sequence with blast identity scores to the CCMP 1516 reference genome < 80% over segments > 100 bp; with percentage of the total strain specific assembly specified in parenthesis.
[b]Strain Specific sequence, the amount of assembled non-gapped sequence with blast identity scores to other strains < 80% over segments > 100 bp; with percentage of the total strain specific assembly specified in parenthesis.
[c]Missing genes are defined as genes with <50% length coverage by mapped reads.

Supplementary Table 6| Characteristics of *E. huxleyi* strains used for Illumina sequencing.

Strain	Location	Length in Culture	Flow Cytometry Genome Size (Mb)	Genome Coverage	% Ref Genome Covered[b]	%Unmapped Reads	Missing Genes[c]	SNPs
AWI1516[a]	South Pacific (02.6667S 82.7167W)	1991	137+/- 0.05	14	91.31	6.8	598	62,889
92D	English Channel (50,02N 04,22W)	1975	106 +/- 0.02	19	68.02	30.4	4953	187,925
92E	English Channel (49, 52N 6,12W)	1992	100 +/- 0.07	19	77.85	26.1	2366	175,800
92F	English Channel (49,52N 6,12W)	1992	nd	nd	68.00	30.8	4832	184,053
B11	Bergen Sea (60, 16N 5,14E)	1992	168 +/- 0.08	14	77.52	30.2	2268	133,854
B39	Bergen Sea (60° 25' 58" N, 5° 38' 13" E)	1992	136 +/- 0.00	17	76.51	29.3	2426	138,100
L	Oslo fjord (60°N 11°E)	1959/68/80	nd	nd	78.21	25.3	2296	145,623
M217	Bergen Sea (60,16N 5,14E)	1992	152 +/- 0.08	16	91.98	17.5	238	46,897
12_1 (CCMP 371)	Sargasso Sea (32N 62W)	1987	nd	nd	82.64	24.2	1454	144,487
M219 (NZEH)	Big Glory Bay, New Zealand	1992	154 +/- 0.03	29	77.12	37.7	2061	123,536
M202 (EH2)	Gt. Barrier Reef, Australia	1990	120 +/- 0.01	322	76.79	22.5	3119	132,166
92A	English Channel	1957	99 +/- 0.00	352	74.66	29.5	3348	155,785
M222 (VAN556)	Vancouver, BC (49,5N 144,40W)	1984	133 +/- 0.02	265	70.19	32.9	3929	204,852

[a]AWI1516[*] was nominally considered CCMP1516 but maintained at the Alfred Wegener Institute and is not the CCMP culture collection strain used for JGI Sanger sequencing.
[b]Coverage was estimated by direct mapping of reads to the reference genome using the Mapping and Assembly with Quality software package.
[c]Missing genes are operationally defined as genes with less than 50% of their length covered by mapped reads.

Supplementary Table 7| Assembly statistics of genomic reads from 12 different *E. huxleyi* strains using the CLC Genomics Workbench.

Strain	Total Reads	Total bp	% GC	N[a]	n:N50[b]	N50[c] (bp)	Max[d] (bp)	Assembly Size (Mb)	Completeness CEGMA (%)[e]
AWI1516	25,260,678	1919811528	65.1	80774	15305	878	84510	49.1	95.0
92D	26,153,077	1,987,633,852	64.3	89837	17878	759	18908	52.8	91.7
92E	27,029,626	2,054,251,576	64.9	87542	15706	1018	47013	59.4	94.5
92F	26,501,216	2,676,622,816	64.2	95147	18511	821	44419	59	91.0
B11	23,074,156	2,330,489,756	64.4	132288	26692	695	21343	75.1	93.9
B39	22,759,741	2,298,733,841	64.4	112729	19512	938	27923	73,2	94.1
L	22,905,701	1,740,833,276	64.8	87386	15697	939	37123	54.2	93.9
M217	24,012,306	2,425,242,906	62.8	117904	20463	938	208111	76.5	95.2
12-1 (CCMP371)	26,604,262	2,021,923,912	64.8	94172	17490	837	84508	56.4	95.0
M219 (NZEH)	50,355,481	4,462,328,406	61.5	143724	29874	645	29910	101.3	95.0
M202 (EH2)	386,571,098	38,657,109,800	61.2	77783	5129	3378	2043190	117.7	94.5
92A	344,534,050	34,453,405,000	62.2	56794	8753	2620	84520	85.6	94.8
M222 (Van556)	352,924,546	35,292,454,600	62.4	75716	9927	2665	183618	109.7	95.0

[a] total number of contigs
[b] contig number of the median contig length
[c] median contig length
[d] maximum contig length
[ee] percent completeness based on conserved eukaryote genes mapping approach based on criteria using an alignment coverage >50% of the gene length

Supplementary Table 8| *E. huxlei* genes likely transferred from bacteria or viruses.

Protein ID	Gene Description	Bacterial/Viral homologs	e-value
57193	HisI Phosphoribosyl-AMP cyclohydrolase	Verrucomicrobia (*Coraliomargarita akajimensis*)	8.00E-32
61543	TrxB Thioredoxin reductase	Cyanobacteria (*Synechococcus* sp., *Prochlorococcus marinus*)	2.00E-101
63173	PaaY Carbonic anhydrases/acetyltransferases, isoleucine patch superfamily	Verrucomicrobia (*Coraliomargarita akajimensis*, *Akkermansia muciniphila*)	9.00E-45
66010	Uncharacterized conserved protein	Cyanobacteria (*Synechococcus* sp., *Prochlorococcus marinus*)	6.00E-137
99196	CaiD Enoyl-CoA hydratase/carnithine racemase	Actinobacteria (*Frankia* sp., *Acidimicrobium ferrooxidans*)	5.00E-28
113427	DdpA ABC-type dipeptide transport system, periplasmic component	Fusobacteria (*Fusobacterium nucleatum*)	2.00E-57
116424	HisC Histidinol-phosphate/aromatic aminotransferase and cobyric acid decarboxylase	Gammaproteobacteria (*Saccharophagus degradans*)	4.00E-34
118625	DraG ADP-ribosylglycohydrolase	Epsilonproteobacteria (*Arcobacter nitrofigilis*)	2.00E-104
195029	RacX Aspartate racemase	Betaproteobacteria (*Burkholderia cepacia*, *Burkholderia ambifaria*)	6.00E-15
199862	unknown	Betaproteobacteria (*Bordetella avium*, *Bordetella bronchiseptica*, *Bordetella petrii*, *Achromobacter xylosoxidans*)	4.00E-28
205111	AslA Arylsulfatase A and related enzymes	Verrucomicrobia (*Coraliomargarita akajimensis*)	4.00E-34
212130	unknown	Actinobacteria (*Mycobacterium smegmatis*)	3.00E-30
214575	unknown	Alphaproteobacteria (*Jannaschia* sp.)	4.00E-26
216096	Predicted CoA-binding protein	Firmicutes (*Clostridium perfringens*, *Clostridium botulinum*, *Clostridium cellulovorans*)	3.00E-19
221901	Uncharacterized conserved protein	Actinobacteria (*Geodermatophilus obscurus*, *Beutenbergia cavernae*)	2.00E-13
223179	Predicted integral membrane protein linked to a cation pump	Alphaproteobacteria (*Agrobacterium tumefaciens*)	1.00E-37

223483	unknown	Bacteroidetes (*Paludibacter propionicigenes*)	7.00E-55
223586	Predicted esterase	Deltaproteobacteria (*Haliangium ochraceum*)	2.00E-29
227462	HemL Glutamate-1-semialdehyde aminotransferase	Alphaproteobacteria (*Methylobacterium radiotolerans*, *Methylobacterium nodulans*, *Methylobacterium* sp.)	7.00E-36
227959	MhpC Predicted hydrolases or acyltransferases	Betaproteobacteria (*Burkholderia xenovoran*)	3.00E-34
232290	TrkA Predicted flavoprotein involved in K+ transport	Actinobacteria (*Mycobacterium smegmatis*)	6.00E-43
237718	Hfi Hydroxypyruvate isomerase	Bacteroidetes (*Spirosoma linguale*, *Zunongwangia profunda*, *Maribacter* sp.)	4.00E-80
238888	CysC Adenylylsulfate kinase and related kinases	Cyanobacteria (*Synechococcus* sp.)	6.00E-90
249598	Predicted amino acid aldolase or racemase	Acidobacteria (*Solibacter usitatus*, Candidatus 'Koribacter versatilis')	1.00E-55
317236	PyrF Orotidine-5'-phosphate decarboxylase	Alphaproteobacteria (*Azospirillum* sp.)	8.00E-53
368782	Icc Predicted phosphohydrolases	Gammaproteobacteria (*Legionella pneumophila*)	9.00E-33
414413	CysK Cysteine synthase	Firmicutes (*Clostridium difficile*)	4.00E-88
422943	Era GTPase	Bacteroidetes (*Dyadobacter fermentans*, *Leadbetterella byssophila*)	3.00E-72
423817	PurC Phosphoribosylaminoimidazolesuccinocarboxamide synthase	Gammaproteobacteria (*Ferrimonas balearica*, *Vibrio cholerae*)	8.00E-149
429133	GlmS Glucosamine 6-phosphate synthetase,	Bacteroidetes (*Salinibacter ruber*, *Rhodothermus marinus*)	4.00E-164
430491	unknown	Betaproteobacteria (*Variovorax paradoxus*)	5.00E-50
434256	PutA NAD-dependent aldehyde dehydrogenases	Actinobacteria (*Streptomyces scabiei*, *Mycobacterium abscessus*)	4.00E-157
438874	MntH Mn^{2+} and Fe^{2+} transporters of the NRAMP family	Bacteroidetes (*Rhodothermus marinus*, *Spirosoma linguale*)	2.00E-22
440614	Protein involved in biosynthesis of mitomycin antibiotics/polyketide fumonisin	Deltaproteobacteria (*Sorangium cellulosum*, *Myxococcus xanthus*)	8.00E-16
441379	CaiA Acyl-CoA dehydrogenases	Alphaproteobacteria (*Parvularcula bermudensis*)	8.00E-94
441750	Rpe Pentose-5-phosphate-3-epimerase	Cyanobacteria (*Prochlorococcus marinus*)	5.00E-104
441988	Short-chain alcohol dehydrogenase of unknown specificity	Alphaproteobacteria (*Roseobacter denitrificans*, *Maricaulis maris*, *Hirschia baltica*)	2.00E-28

442445	SpsE Sialic acid synthase	Bacteroidetes (*Croceibacter atlanticus*)	6.00E-69
448150	PdxK Pyridoxal/pyridoxine/pyridoxamine kinase	Gammaproteobacteria (*Dickeya dadantii*)	8.00E-81
450347	AceE Pyruvate dehydrogenase complex, dehydrogenase (E1) component	Alphaproteobacteria (*Sphingomonas wittichii*, *Caulobacter* sp., *Bradyrhizobium* sp., *Acidiphilium cryptum*)	0
452720	CaiA Acyl-CoA dehydrogenases	Gammaproteobacteria (*Pseudomonas syringae*, *Pseudomonas putida*)	9.00E-93
453623	unknown	Bacteroidetes (*Pedobacter heparinus*)	5.00E-50
453687	Predicted acetamidase/formamidase	Alphaproteobacteria (*Sphingomonas wittichii*, *Rhizobium leguminosarum*)	3.00E-46
458546	CaiA Acyl-CoA dehydrogenases	Gammaproteobacteria (*Pseudomonas syringae*, *Pseudomonas entomophila*)	2.00E-102
459783	Acd 1-aminocyclopropane-1-carboxylate deaminase	Deltaproteobacteria (*Desulfovibrio desulfuricans*)	2.00E-85
461437	WcaA Glycosyltransferases involved in cell wall biogenesis	Deltaproteobacteria (*Desulfovibrio desulfuricans*)	3.00E-28
469992	Predicted acetamidase/formamidase	Alphaproteobacteria (*Rhizobium leguminosarum*, *Sphingomonas wittichii*)	2.00E-41
43654	ELO, GNS1/SUR4 family	Emiliania huxleyi virus 86	1.00E-63
54601	Dihydroceramide desaturase (Dsd1, delta-4)	Emiliania huxleyi virus 86	3.00E-36
61414	Phosphate permease	Emiliania huxleyi virus 86	3.00E-57
97888	Lipocalin-like	Emiliania huxleyi virus 86	5.00E-19
102590	ERG3, Sterol desaturase	Emiliania huxleyi virus 86	7.00E-34
111551	Hypothetical protein	Emiliania huxleyi virus 86	1.00E-11
122629	Putative DNA cytosine methylase	Enterobacteria phages	5.00E-26
193896	Hypothetical protein	Emiliania huxleyi virus 86	8.00E-10
193908	Lipid phosphate phosphatase (PAP2 superfamily)	Emiliania huxleyi virus 86	4.00E-15
196284	Fatty acid desaturase (Aco1, delta-9)	Emiliania huxleyi virus 86	3.00E-22
197639	Hypothetical protein	Emiliania huxleyi virus 86	7.00E-09
200323	Hypothetical protein	Emiliania huxleyi virus 86	5.00E-11
200862	Dihydroceramide synthase	Emiliania huxleyi virus 86	1.00E-44

205088	Hypothetical protein	Emiliania huxleyi virus 86	5.00E-08
208320	DNA repair/recombination protein pif1-like with HRDC domain	Emiliania huxleyi virus 86	2.00E-39
210457	ERG3, Sterol desaturase	Emiliania huxleyi virus 86	1.00E-54
212478	MFS_1, Major Facilitator Superfamily	Emiliania huxleyi virus 86	3.00E-95
215136	Tmk, Thymidylate kinase	Emiliania huxleyi virus 86	5.00E-48
235604	Glycosyl transferase family 8-like protein	Emiliania huxleyi virus 86	3.00E-17
242737	YqaJ viral recombinase family	PBCV-1, OtV5	5.00E-11
243604	Hypothetical protein	NCLDVs (Mimivirus, PBCVs, ASCV1)	4.00E-11
420219	ATP-dependent DNA ligase	Emiliania huxleyi virus 86	1.00E-104
432191	Sec14p-like lipid-binding domain	Emiliania huxleyi virus 86	1.00E-12
432205	Hypothetical protein	Emiliania huxleyi virus 86	1.00E-54
432901	Serine palmitoyltransferase	Emiliania huxleyi virus 86	1.00E-145
432978	Hypothetical protein	Emiliania huxleyi virus 86	3.00E-60
434519	Hypothetical protein	Emiliania huxleyi virus 86	2.00E-32
439872	2OG-FeII_Oxy domain-containing protein	Synechococcus phage Syn9	6.00E-29
440222	Hypothetical protein	Emiliania huxleyi virus 86	2.00E-07
443105	Hypothetical protein	Emiliania huxleyi virus 86	1.00E-78
446612	Methyltransferase	Emiliania huxleyi virus 86	2.00E-22
454190	SSL2, DNA or RNA helicases of superfamily II	NCLDVs (Mimivirus, EsV1, OtV5, PBCVs, ASCV1)	2.00E-23
461715	Hypothetical protein	Emiliania huxleyi virus 86	2.00E-21
508420	Formamidopyrimidine-DNA glycosylase	Mimivirus	4.00E-24

Supplementary Table 9| Classification of the repeat content of the *E. huxleyi* genome.

Class	Category	Family	coverage (bp)	% of Genome
Class 1			**1456459**	**1.115**
	LTR RT		992754	0.760
		Ty1/Copia	573139	0.439
		Ty3/Gypsy	5969	0.005
		DIRS	41333	0.032
		other*	372313	0.285
	Non-LTR RT			
		LINE	463705	0.355
Class 2			**3888700**	**2.976**
	TIR		3848046	2.945
		Harbinger	46255	0.035
		hAT	11260	0.009
		PiggyBac	39231	0.030
		Sola	32087	0.025
		Tc1/Mariner	70312	0.054
		MITE**	412678	0.316
		TIR**	3236223	2.477
	Helitron			
		Helitron	40654	0.031
rDNA-related* **			**3751552**	**2.871**
	>80% rDNA coverage		65245	0.050
	50-80% rDNA coverage		293121	0.224
	20-50% rDNA coverage		3393186	2.597
NoCat* **			**40417639**	**30.935**
	>50% TR coverage		11855716	9.074
	<50% TR coverage		28561923	21.861
Host gene* **			**13226696**	**10.123**
Tandem repeats and low complexity regions			44946107	34.401
	In dispersed repeats†		23582094	18.049
	Not in dispersed repeats†		21364013	16.351
Total ††			**~84Mb**	**~64%**

* Family not clearly determined.

** Only terminal inverted repeats were detected.

*** Only NoCat and rDNA-related with at least 10 copies and host genes with at least 5 copies are counted.

† Dispersed repeats comprise all above classes.

†† The total repeat content is computed by adding the subtotal of each class (first line of each class) except for "Tandem repeats and low complexity regions" where only the exclusive contribution "Not in dispersed repeats" has to be added.

Supplementary Table 10| Genes involved in the protection and acclimation to high light.

Protein Description	Protein ID	Core/Variable
Light harvesting machinery and photoreceptors		
UVR8 Photoreceptor	858012	nd
Blue light cryptochrome	51761	variable
Blue light cryptochrome	450581	core
Blue light cryptochrome	42748	core
Blue light cryptochrome	52551	core
Blue light cryptochrome	209270	core
Related to blue light cryptochrome	118405	variable
Related to blue light cryptochrome	103064	variable
Putative PAS domain sensor hybrid histidine kinase; phytochrome domain, homology to PHYA red or far red photoreceptor	464698	core
Similarity to blue light receptors but only has the PAS domain	53974	variable
	313274	variable
	557981	nd
	557999	nd
	455463	nd
	557997	bd
	436693	core
	207698	core
	439376	core
	31374	nd
	75737	nd
Phototropin	529487	nd
Phytochrome-like protein similar to phytochrome	447461	core
Antioxidants and proteins involved in non-photochemical quenching		
Peroxiredoxin	235642	core
	432312	core
Peroxiredoxin-2E, chlorophastic	470080	nd
	98076	core
Thiol peroxidase	221257	core
Violaxanthin de-epoxidase	437347	core
Violaxanthin de-epoxidase	455406	core
Zeaxanthin epoxidase	457120	core
	456674	core
11 kDa protein of photosystem II, psbZ, critical role in NPQ	415669	core
Alternative Oxidase	43821	variable
	43799	variable

Ascorbate peroxidase	72999	variable
	69501	core
	434150	core
	451669	variable
Glutathione peroxidase	558203	core
	558390	core
	632133	core
	558200	core
	558201	core
Catalase	464198	core
	350790	core
Superoxide dismutase	96386	variable
	212678	core
	354736	core
	364889	core
Degradation, assembly and repair of PSII Components		
Photosystem II stability/assembly factor HCF136	461307	core
Photosystem II stability/assembly factor HCF137	425813	core
Photosystem II stability/assembly factor HCF138	253334	variable
Photosystem II D1 protease	214344	core
Carboxy-terminal processing protease	315400	core
	461027	core
	97800	core
	243334	nd
	436416	core
FtsH cell division protein	416950	core
FtsH	427625	core
	466116	core
High light response proteins		
High light inducible protein	450258	core
	250739	core

Supplementary Table 11| Table of 68 Light harvesting complex proteins in the *E. huxleyi* genome

Protein ID	Allelic Variant	Annotation name	Core or Variable Genome	Group on Tree	Proteomics Detection
413829		Lhcf 2	core	Red	Y
439022		Lhcf 3	core	Red	Y
438393	445064	Lhcf 4	core	Not resolved	N
441453		Lhcf 5	core	Red	Y
211477	317615	Lhcf 6	core	Chlorophyll a/c group II	Y
419663		Lhcf 7	variable	Chlorophyll a/c group I	Y
46861		Lhcf 8	variable	Chlorophyll a/c group I	N
231654	260243	Lhcf 9	variable	Chlorophyll a/c group I	N
433058	434108	Lhcf 10	core	Chlorophyll a/c group I	Y
432202		Lhcf 11	core	Chlorophyll a/c group I	Y
200525	249526	Lhcf 12	core	Chlorophyll a/c group I	N
66220		Lhcf 13	core	Chlorophyll a/c group I	N
432006		Lhcf 14	core	Chlorophyll a/c group I	Y
415729	446702	Lhcf 15	core	Chlorophyll a/c group I	Y
435472	372663	Lhcf 16	core	Chlorophyll a/c group I	Y
363095		Lhcf 17	variable	Chlorophyll a/c group I	N
443305		Lhcf 18	nd	Chlorophyll a/c group I	Y
422150	77826	Lhcf 19	variable	Chlorophyll a/c group I	Y
356694	362770	Lhcf 20	variable	Chlorophyll a/c group I	N
67603	74777	Lhcf 21	core	Chlorophyll a/c group I	N
65742		Lhcf 22	core	Chlorophyll a/c group I	Y
64682		Lhcf 23	core	Chlorophyll a/c group I	N
438462		Lhcf 24	core	Chlorophyll a/c group I	Y
441864	442651	Lhcf 25	core	Chlorophyll a/c group II	Y
360421	431505	Lhcf 26	variable	Chlorophyll a/c group II	Y
435222		Lhcf 27	core	Chlorophyll a/c group II	Y
70030		Lhcf 28	core	Chlorophyll a/c group II	Y
417035		Lhcf 29	variable	Chlorophyll a/c group II	N
437015	420179	Lhcf 30	variable	Chlorophyll a/c group II	Y
434417		Lhcf 31	core	Chlorophyll a/c group II	N
443878		Lhcf 32	core	Chlorophyll a/c group II	Y
358662		Lhcf 33	core	Chlorophyll a/c group II	Y
45662	467343	Lhcf 34	core	LI818	N
422697		Lhcf 35	core	LI818	Y
69651	73966	Lhcf 36	variable	LI818	N
443721		Lhcf 36	nd	LI818	N
217340*		Lhcf 37	core	LI818	Y
430242		Lhcf 38	variable	LI818	N
45605	45591	Lhcf 39	variable	LI818	N
436450*		Lhcf 41	core	LI818	N

460117*		Lhcf 42	core	LI818	N
416733*		Lhcf 43	core	LI818	N
442232		Lhcf 44	core	LhcZ	N
203047	224646	Lhcf 45	core	LhcZ	N
434226	419743	Lhcf 46	core	LhcZ	Y
451739		Lhcf 47	core	LhcZ	Y
356951		Lhcf 48	core	LhcZ	Y
358243		Lhcf 49	core	LhcZ	N
414772		Lhcf 50	core	Red	Y
417267		Lhcf 51	core	LI818	Y
428685*		Lhcf 52	core	LI818	Y
364696		Lhcf 53	core	LhcZ	N
76288		Lhcf 54	core	Lhcz	N
461003		Lhcf 55	core	LI818	Y
353537		Lhcf 56	core	Chlorophyll a/c group I	N
240836		Lhcf 57	core	Chlorophyll a/c group I	N
362550		Lhcf 58	core	Chlorophyll a/c group I	Y
75517		Lhcf 60	variable	Chlorophyll a/c group I	Y
433847		Lhcf 62	variable	Chlorophyll a/c group II	Y
312801		Lhcf 63	variable	Chlorophyll a/c group II	Y
446310		Lhcf 64	core	Chlorophyll a/c group II	Y
366130		Lhcf 65	core	Chlorophyll a/c group II	Y
355940		Lhcf 66	core	Chlorophyll a/c group II	Y
353073	370423	Lhcf 67	core	Chlorophyll a/c group II	N
312310		Lhcf 68	variable	Chlorophyll a/c group II	N
455747		Lhcf 69	core	Red	Y
358647		Lhcf 70	core	Red	Y
632166		Lhcf 71	nd	Not resolved	N

* denotes the detection of a CO_2 enhancer element

Supplementary Table 12| Genes encoding proteins putatively involved in phosphate metabolism.

Protein Description	No. of Paralogs	Protein ID	Core vs. Variable Genome
Transporters			
General phosphate transporter	14	466430	nd
		68046	variable
		218713	core
		433548	core
		200606	variable
		438712	variable
		438750	variable
		456911	variable
		463311	core
		95576	core
High affinity phosphate transporter		67455	core
		450989	variable
		61414	core
Phosphate metabolism			
Alkaline phosphatse	3	369509	variable
		433041	core
		229451	core
Alkaline phosphatase domain containing protein	2	230310	core
		457924	nd
5' nucleotidase	5	237438	core
5' nucleotidase		237172	core
5' nucleotidase		251982	core
5' nucleotidase		111381	variable
5' nucleotidase		111972	core
Cyclic GMP phosphodiesterase	1	108787	variable
Purple acid phosphatase-like protein/Phytase	4	62631	core
		462501	nd
		62875/78521	core
		451170/67476	nd
Inorganic pyrophosphatase	3	44801	core
		437335	core
		439540/447001	core
Acidocalcisomal pyrophosphatase	2	415968	core
		436192	core
Nudix hydrolase	4	45305	nd

		45054	core
		45000	core
		50111	core
Evidence of Phosphonate Metabolism			
HAD superfamily hydrolase	1	65222	core
Haloacid dehalogenase-like hydrolase	2	428568	nd
		428161	core
Carboxyvinyl-carboxyphosphonate phosphorylmutase	3	47077	core
		254202	variable
		238585	core
ABC transporter with homology to phosphonate transporter	4	427008	core
		427164	core
		110965	core
		417116	core
Regulation			
Two component response regulator	2	69109	nd
		67954	core
Homeobox transcription factor	2	434760	core
		445626	nd
Cyclin dependent kinase type C	1	436308	core
Protein kinase	3	444655	nd
		452712	core
		313453	variable
Phospholipid Replacement			
Sulfolipid biosynthesis protein	1	466230	core
Sulfolipid synthetase	1	97678	core
Phospholipid/glycerol acyltransferase	2	211471	core
		75696	nd
N-acetyltransferase 12 isoform	2	105120	core
		240186	nd

Supplementary Table 13| Genes encoding proteins putatively involved in nitrogen metabolism.

Name	Protein ID	Core or Variable Genome
Nitrogen Metabolism		
Glutamate synthase	212450	core
Glutamate synthase	467437	core
Glutamate synthase	449382	nd
Glutamate synthase	225692	core
Glutamate synthase	469498	nd
Glutamine synthetase	470023	core
Glutamine synthetase	437187	core
Glutamine synthetase	110084	nd
Glutamine synthetase	52323	core
Glutamine synthetase	69253	core
Glutamine dehydrogenase	433762	core
Glutamine dehydrogenase	69206	core
Inorganic Nitrogen Metabolism		
Nitrogen fixation protein	254468	core
Nitrogen fixation protein	197151	core
Nitrite reductase	428968	core
Copper nitrite reductase	69470/70548/70565	variable
Nitrite transporter	232402	nd
Nitrite transporter	231096	core
Nitrite transporter	439254	core
Nitrite transporter	64600	core
Nitrite transporter	373579	core
High affinity nitrate transporter	62811	core
High affinity nitrate transporter	440685	core
Nitrate transporter	361445	core
Nitrate transporter	195544	core
Nitrate transporter	460408	variable
Nitrate transporter	73160	core
Nitrate transporter	115184	core
Nitrate transporter	311732	variable
Nitric oxide synthase	550420	nd
flavohemoglobin	522194	nd
Cyanate		
Cyanate lyase	73978	nd
Cyanate lyase	69646	core
Cyante transporter	ACCT9561-J03.xld-F	nd
Cyante transporter	AWCG1205-H03.xld-F	nd
Nitriliase	198908	core
Nitriliase	456340	core
Nitriliase	454646	core
Nitriliase	210750	core
Nitriliase	462146	core

Nitriliase	415168	variable
Urea		
Urea transporter	194219	core
Urea transporter	440179	core
Urea transporter	460978	core
Urea transporter	68140	variable
Urea transporter	225018	variable
Urea transporter	217311	nd
Arginase	573947	nd
Urease	631885	nd
Carbamoyl phosphate synthetase-III	422007	core
Carbamoyl phosphate synthetase-II	463837	core
Ornithine transcarbamyolase	204139	core
Ornithine transcarbamyolase	236460	nd
Argininosuccinic acid synthetase	528272	nd
Argininosuccinic acid synthetase	124858	core
Argininosuccinate lyase	458298	core
Argininosuccinate lyase	465437	core
Argininosuccinate lyase	552739	nd
Ammonium transporter	212609	nd
Ammonium transporter	217537	nd
Ammonium transporter	70117	variable
Ammonium transporter	62984	nd
Ammonium transporter	464017	core
Ammonium transporter	415646	variable
Ammonium transporter	456888	variable
Ammonium transporter	433750	core
Ammonium transporter	74037	core
Ammonium transporter	471321	variable
Ammonium transporter	201096	variable
Ammonium transporter	521917	nd
Ammonium transporter	452779	core
Ammonium transporter	61319	core
Ammonium transporter	433338	core
Ammonium transporter	451136	core
Ammonium transporter	193894	variable
Ammonium transporter	65124	core
Ammonium transporter	463302	variable
Ammonium transporter	122984	variable
Other		
Phenylalanine/Histidine ammonia lyase	69043	variable
Glutaminase	68199	core
Glutaminase	101092	core
Glutaminase	470892	nd
Glutaminase	64305	core
Hypoxanthine phosphoribosyltransferase	428604	core
Polyamine transporter	558131	nd
Polyamine transporter	558095	nd
Polyamine transporter	446672	nd

Polyamine transporter	558152	nd
Polyamine transporter	240775	variable
Polyamine transporter	539234	nd
Polyamine transporter	108831	nd
Polyamine transporter	543743	nd
Polyamine transporter	558060	nd
Polyamine transporter	558138	nd
Polyamine transporter	558140	nd
Polyamine transporter	454057	core
Polyamine transporter	196046	variable
Polyamine transporter	77006	nd
Polyamine transporter	204281	core
Polyamine transporter	457373	variable
Polyamine transporter	70520	core
Polyamine transporter	558156	nd
Polyamine transporter	558130	nd
Polyamine transporter	201292	core
Polyamine transporter	108941	core
Polyamine transporter	438821	core
Polyamine transporter	558157	nd
Polyamine transporter	73488	core
Dipeptidase plasmid membrane bound	247883	variable
Dipeptidase plasmid membrane bound	466525	core
Dipeptidase plasmid membrane bound	467505	nd
Dipeptidase plasmid membrane bound	227593	nd
Amine oxidase	464657	core
Amine oxidase	201294	core
Amine oxidase copper	111175	core
Amine oxidase copper	205572	core
Amine oxidase copper	220023	core
Amine oxidase copper	197391	variable
D-amino acid oxidase	210589	core
L-amino acid oxidase	120029	variable
FAD amine oxidase	244346	variable
FAD amine oxidase	112568	variable
FAD amine oxidase	116825	core
Polyamine oxidase	352602	core
Chinitase	117626	core
Chitinase	65231	core
Chitinase	453682	core
Chitinase	353656	variable
Chitinase	429900	core
Chitinase	469375	nd
Chitinase	417127	core
Chitinase	225160	variable
Amidase	227739	variable
Amidase	451302	variable
Amidase	456188	variable
Amidase	109981	variable

Amidase	102453	variable
Acetamidase/Formamidase	219043	variable
Acetamidase/Formamidase	219044	variable
Acetamidase/Formamidase	453687	core
Acetamidase/Formamidase	444643	core
Acetamidase/Formamidase	245832	variable
Acetamidase/Formamidase	469992	variable
Acetamidase/Formamidase	215750	variable
Acetamidase/Formamidase	452038	variable
Acetamidase/Formamidase	440936	core
Nitrile hydratase beta subunit	56930	variable
Nitrile hydratase alpha subuni	107811	nd
Nitrile hydratase alpha subuni	106673	nd
Nitric oxide reductase	Present in unassembled reads	

Supplementary Table 14| Genes encoding putative selenoproteins.

Selenoprotein	Gene name	Protein ID	Allelic Variant
Dithiodisulfide oxidoreductase	DSBA1	558350	558351
Dithiodisulfide oxidoreductase	DSBA2	558352	
Dithiodisulfide oxidoreductase	DSBA3	632129	632130
Dithiodisulfide oxidoreductase	DSBA4	632131	
Dithiodisulfide oxidoreductase	DSBA5	632132	
Gamma interferon inducible lysosomal thiol reductase	GILT1	632140	
Gamma interferon inducible lysosomal thiol reductase	GILT2	632141	
Glutathione peroxidase	GPX1	558200	
Glutathione peroxidase	GPX2	558201	
Phospholipid hydroperoxide glutathione peroxidase	PHGPX3	558202	558204
Glutathione peroxidase	GPX5	558203	558205
Glutathione peroxidase	GPX6	558390	
Glutathione peroxidase	GPX7	632133	632134
Iodothyronine deiodinase	DIO1	558396	
Iodothyronine deiodinase	DIO2	632124	632125
Iodothyronine deiodinase	DIO3	632126	632127
Iodothyronine deiodinase	DIO4	632128	
Iron-sulfur oxidoreductase	GLPC1	632148	
Membrane selenoprotein	MSP1	632149	
Methylated-DNA-[protein]-cysteine-S-methyltransferase	MGMT1	558389	
Peptide methionine-S-sulfoxide reductase	MSRA1	558199	
Peptide methionine-S-sulfoxide reductase	MSRA9	558301	
Protein disulfide isomerase	PDI1	558386	558387
Protein disulfide isomerase	PDI2	632135	632136
Protein disulfide isomerase	PDI3	632139	
Selenoprotein 15	SEL15	558397	558398
Selenoprotein H	SELH1	558360	
Selenoprotein H	SELH2	632122	632123
Selenoprotein M	SELM1	632113	
Selenoprotein M	SELM2	632114	
Selenoprotein M	SELM3	632115	
Selenoprotein M	SELM4	632117	632116
Selenoprotein M	SELM5	632118	632119
Selenoprotein M	SELM6	632120	632121
Selenoprotein O	SELO1	558338	558339
Selenoprotein O	SELO3	558340	
Selenoprotein O	SELO4	558343	
Selenoprotein O	SELO2	558344	
Selenoprotein T	SELT1	558358	558359
Selenoprotein T	SELT2	632105	

Selenoprotein U	SELU1	558353	
Selenoprotein U	SELU2	632109	632111
Selenoprotein U	SELU3	632112	
Selenoprotein W	SELW1	632107	632108
Thiol protein	THIO1	632145	632144
Thiol protein	THIO2	632146	
Thiol protein	THIO3	632147	
Thiol:disulfide interchange protein	TDIP1	632142	632143
Thioredoxin reductase	TXR1	417208	

Supplementary Table 15| Genes encoding proteins involved in the metabolism or synthesis of vitamins.

Vitamin Name	PID	Core/Variable
Pro-Vitamin A		
IPP/DMAPP Isomerase	359474	core
GGPP synthase	470946	core
	465220	core
Phytoene synthase	68943	core
Phytoene desaturase	74977	nd
	430592	core
Carotene desaturase	421941	core
Lycopene β and ε-cyclase	42521	core
	448849	core
B_1 Synthesis		
Dinucleotide-utilizing enzyme involved in molybdopterin and thiamine biosynthesis	102319	variable
	68584	core
	102870	core
Thiamine monophosphate synthase	55659	core
	102278	core
	53832	variable
Thiamine pyrophosphokinase	56054	core
	72909	nd
Doxyxylulose-5-phosphate synthase	440786	nd
ABC-Type thiamine transporter	50800	core
	418765	nd
	50947	core
B_1 Related (thiamine)		
Thiamine monophosphate synthase	102278	core
Thiamine triphosphatase	453444	core
	458212	nd
Thiamine pyrophosphate enzyme-like TPP binding protein	199879	core
	312361	variable
Acetohydroxyacid synthase	453895	core
	233263	nd
	118308	nd
	113095	variable
	107052	core
Plastid transketolase	437959	core

Transketolase	420320	core
	450347	core
	423965	core
B_2 (Riboflavin)		
GTP cyclohydrolase II	46070	core
3,4-dihydroxy-2-butanone 4-phosphate synthase	240699	nd
Bifunctional 2,5-diamino-6-ribosylamino-4 (3H)-pyrimidinone 5'-phosphate deaminase/reductase	457738	nd
	457487	core
6,7-dimethyl-8-ribityllumazine synthase	434395	core
Riboflavin synthase	430591	core
Riboflavin kinase	434713	core
	214135	core
	237657	nd
	243687	nd
FAD synthetase	101902	core
B_3 (Niacin)		
Aspartate oxidase	450514	core
Quinolinate synthase		
Quinolinate phosphoribosyltransferase	99818	core
B_5 (Pantothenic acid)		
Bifunctional ketopantoate hydroxymethyltransferase/panthothenate synthase	439738	variable
Ketopantoate reductase	364279	nd
B_6 Salvage Pathway (pyridoxal related compounds)		
Pyridoxal biosynthesis protein	470721	nd
	426056	core
	427499	core
Probable glutamine aminotransferase	222283	core
	123272	variable
Pyrodoxal kinase	448150	core
	44317	core
	45753	variable
Pyridoxal reductase	417225	core
B_7 Metabolism (Biotin)		
8-amino-7-oxononanoate synthase	72721	core
	62552	core
	250705	variable

Bifunctional detiohiotin synthetase/7,8-diamino-pelargonic acid aminotransferase	456336	core
Biotin synthase	423761	variable
DAPA	456336	core
B_9 (folic Acid)		
GTP cyclohydrolase I	73174	core
5-aminoimidazole ribonucleotide carboxylase	421581	core
5-phosphoribosyl-1- pyrophosphate synthase/6-hydroxymethyldihydropterin pyrophosphokinase	457751	nd
	43668	variable
	43644	core
4-amino-4-deoxychorismate synthase	313952	core
	78576	nd
Dihydrofolate reductase/synthase	451981	core
Dihydroneotperin (DHN) aldolase	372344	nd
	351975	core
Aminodeoxychorismate lyase	110960	core
Dihydropteroate synthase	235458	variable
	234285	core
	215177	variable
Folylpolyglutamate synthase	471063	nd
	467825	core
UDP-glucose-p-aminobenzoate glycosyltransferase	226129	core
	233893	variable
γ-glutamylhydrolase	68526	core
	78425	nd
B_{12} Related (Cobalamin)		
S-adenosyl homocysteine hydrolase	441258	core
	440113	core
	440119	core
Glycine hydroxymethyltransferase	433725	core
	422976	core
S-adenosylmethioine synthetase	462645	core
	198673	core
Homocystein S-methyltransferase	423073	core
	463726	core
Methylmalonyl-CoA mutase	417351	core
Methylenetrtrahydrofolate reductase	437840	core
	471108	core
B12 Synthesis/Acquisition		
Uroporphyrin-III C/tetrapyrrole (Corrin/Porphyrin) methyltransferase	213726	core
	45130	core

	55801	variable
Cobyrinic acid a,c-diamide synthase	251565	variable
	437870	core
Cobalamin adensyltransferase	350873	core
	366962	nd
Colalamin biosynthesis protein CobN and Mg=chelatases	309350	core
	317426	core
Cobalamin-5-phosphate synthase	95466	core
Nicotinate-nucleotide-dimethylbenzimidazole	210261	core
	241304	nd
	242675	nd
Cobalamin synthesis protein	417487	core
	75934	nd
	122283	core
	73828	core
	61544	core
	443176	core
	439522	core
	61289	core
	461078	core
	456070	nd
	465373	core
	430429	core
	52778	core
	60847	nd
	74134	nd
	119467	variable
	70806	core
	43607	core
	96729	variable
	438497	core
	456427	core
	41835	core
	102634	core
	43480	variable
	461078	core
	61544	core
Vitamin E Tocopherol Synthesis		
Homogentisate Phytyl transferase (VTE2)	56139	nd

	55059	core
	115547	nd
Solanyl synthase		
Phytol kinase	436338	core
	99736	core
Tocopherol cyclase (VTE1)	54133	core
	98577	core
	237937	nd
Tocopherol O-methyltransferase	60986	core
	98396	core
MPBQ/MSBQ methyltransferase	221641	core
Phytyltransferase	115547	nd
	95285	core
	56139	nd
	55059	variable
C (Ascorbate)		
GTP-mannose pyrophosphorylase	445627	nd
	434761	core
GTP Mannose-3,5-epimerase/phosphmannose isomerase	52642	core
	52327	nd
L-galactose guanyltransferase		
L-galactose-1-P phosphatase	66633	core
	310251	core
L-galactose dehydrogenase	68207	core
	462242	core
L-galactono-1, 4-lactone dehydrogenase	70809	core
	118863	nd
D (calciferol)		
D3 24-hydoxylase	217091	variable
25-hydroxyvitamin D-1 alpha hydrolase	467235	variable
Calcidiol 1- monooxygenase	454070	variable
	458776	core
	452377	core
	72799	variable
	470128	core
	111143	core
	97603	variable
E (tocopherols)		
Homogentisate geranylgeranyl transferase	115547	nd
	95285	core

	56139	nd
	55059	variable
MPBQ/MSBQ methyltransferase	221641	core
	98396	core
Homogentisate solanyltransferase transferase	95285	core
	115547	variable
	56139	nd
	55059	core
Tocopherol cyclase	54133	core
	98577	core
	237937	nd
Chorismate mutase	470315	core
Hydroxyphenylpyruvate (HPP) dioxygenase	452207	core
	470913	nd
Tocopherol methyltransferase	60986	core
K (Phylloquinone)		
Isochorismate synthase	448402	core
PHYLLO	96767	core
1,4-dihydroxy-2-naphtoate phytyl transferase		
O-succinylbenzoate CoA ligase		
1,4-dihydroxy-2-naphthoate CoA synthase		
1,4-dihydroxy-2-naphthoate CoA thioesterase		
1,4-dihydroxy-2-naphthoate phytyptransferase		
Demethylphylloquinone methyltransferase	75975	core
	74671	core
Vitamin K epoxide reductase	361809	core