Supplementary Information

Draft genome sequence and genetic transformation of the oleaginous alga *Nannochloropsis gaditana*

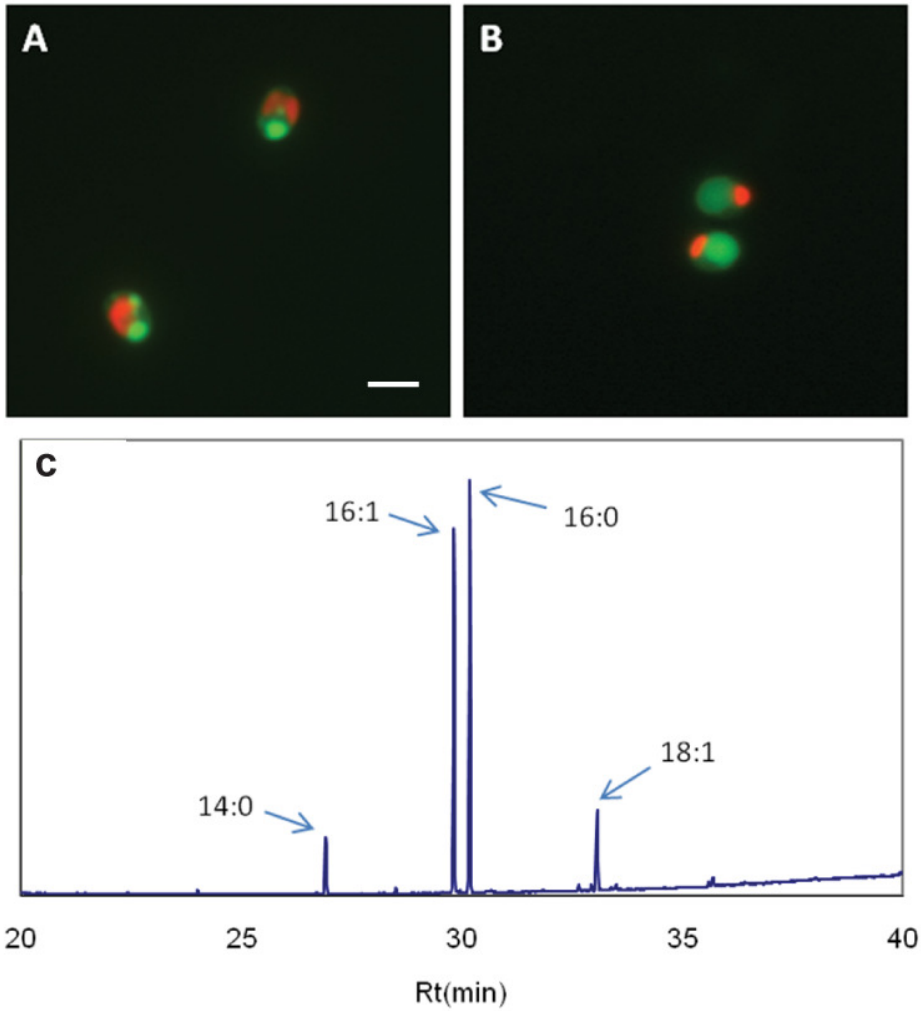
Randor Radakovits[1,a], Robert E. Jinkerson[1,a], Susan I. Fuerstenberg[2], Hongseok Tae[3], Robert E. Settlage[4], Jeffrey L. Boore[2,4], Matthew C. Posewitz[1,b]

[1] Department of Chemistry and Geochemistry, Colorado School of Mines, Golden, CO 80401
[2] Genome Project Solutions, 1024 Promenade Street, Hercules, CA 94547
[3] Data Analysis Core, Virginia Bioinformatics Institute, Virginia Tech, 1 Washington Street, VA 24060
[4] Department of Integrative Biology, University of California, Berkeley, CA 94720

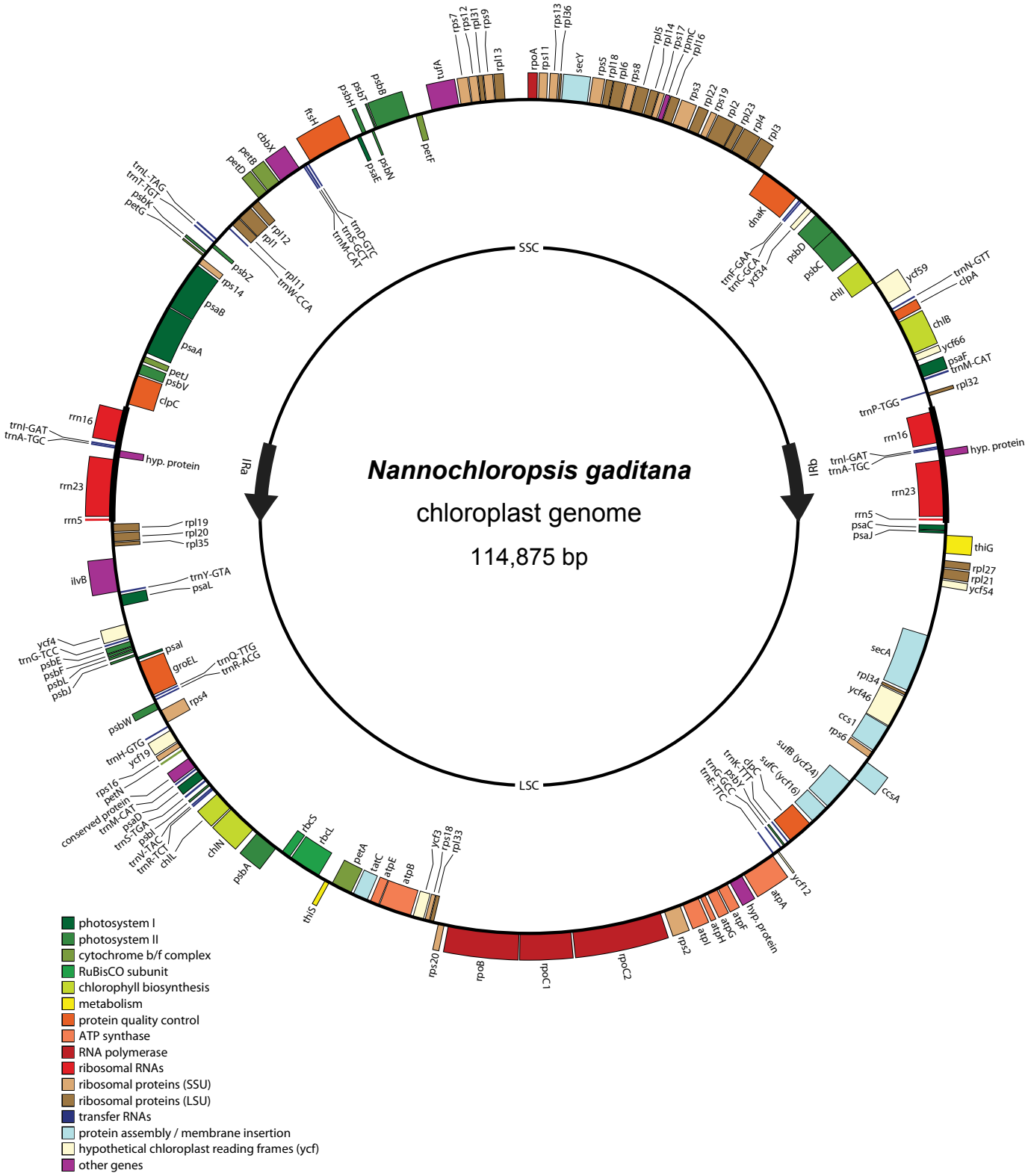
[a]These authors contributed equally to this work.
[b]Correspondence should be directed to M.C.P. (mposewit@mines.edu).

Supplementary Information Table of Contents

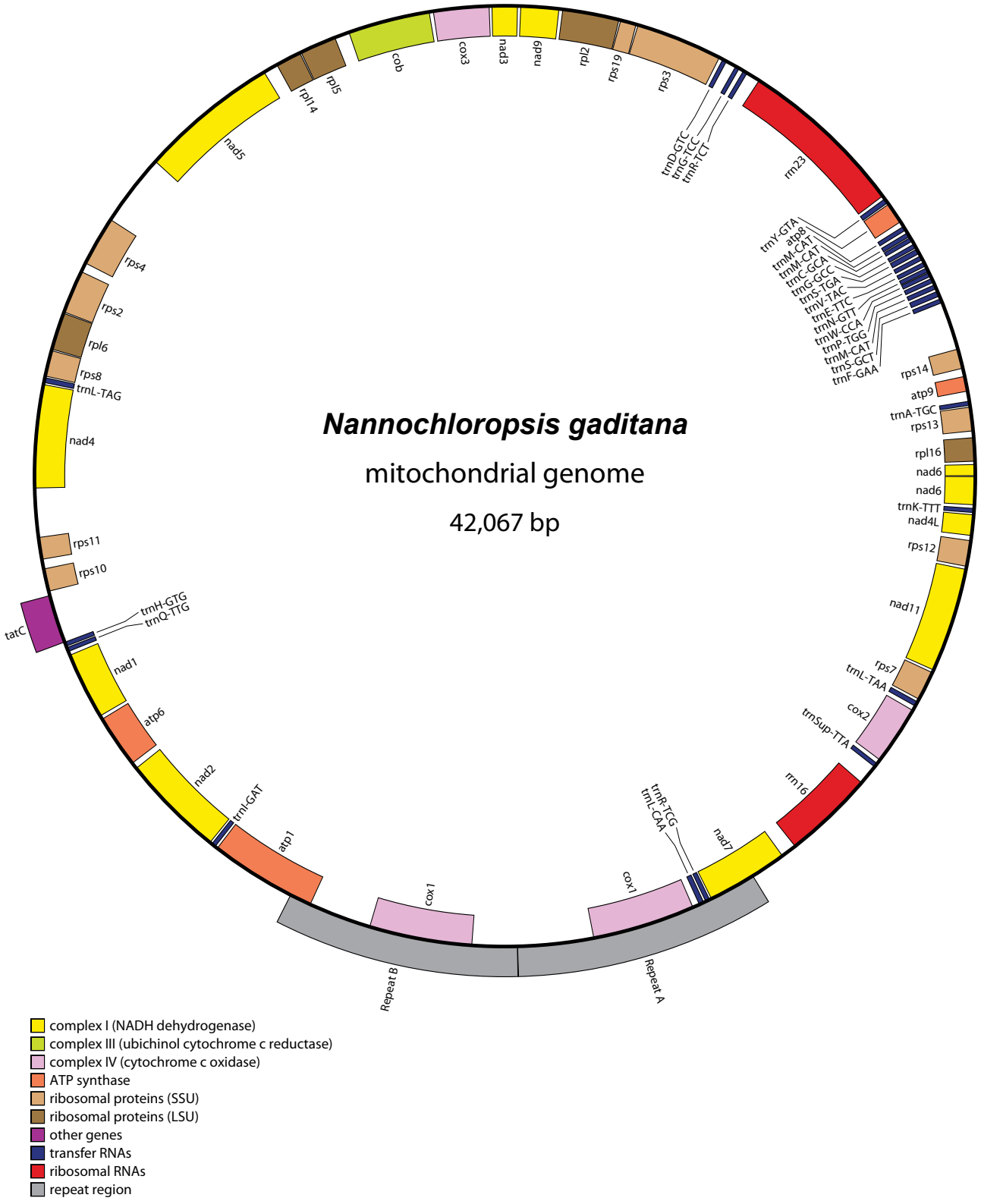
SUPPLEMENTARY FIGURES



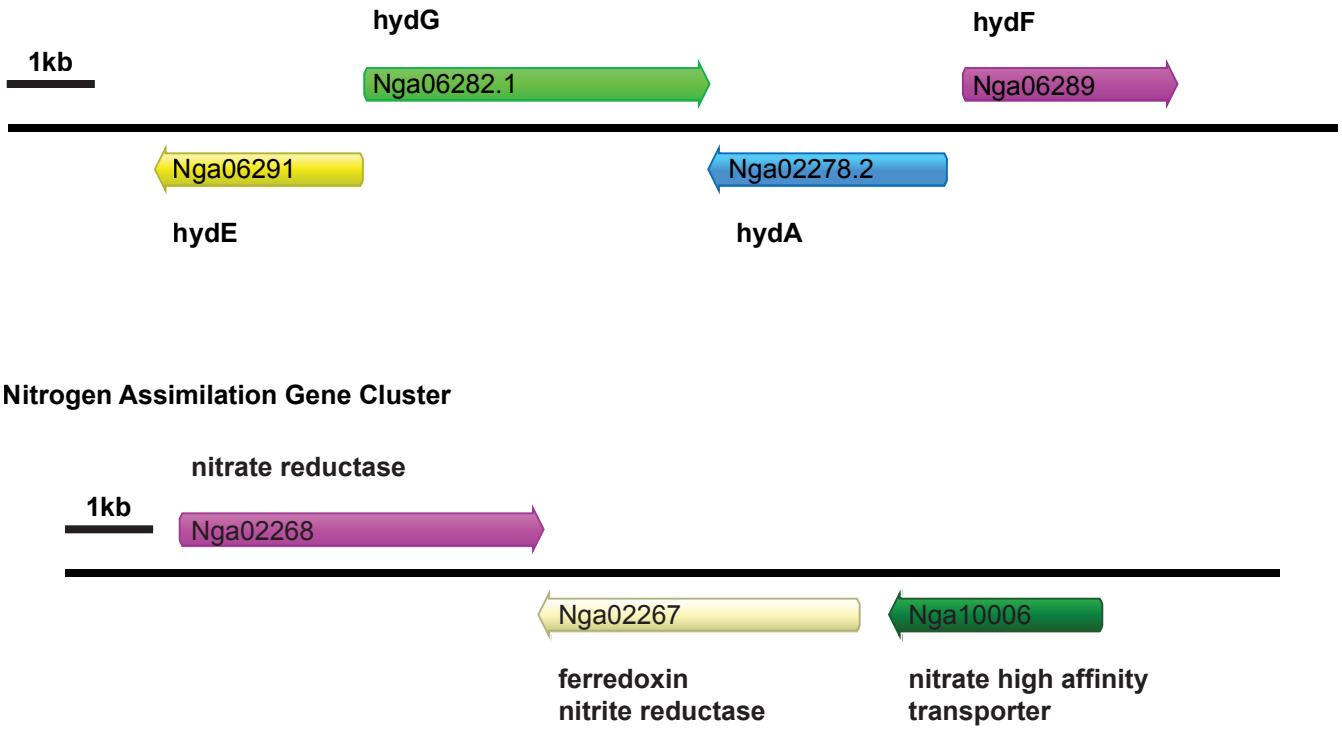
Supplementary Figure S1. Characterization of *N. gaditana* lipid content. (a) Lipid droplets in cells during logarithmic growth. The scale bar represents 3 µm. **(b)** Lipid droplets in cells during stationary phase. Lipid droplets are fluorescently labeled with BODIPY (493/503) (green), chlorophyll autofluorescence (red). **(c)** GC-FID chromatogram showing a typical *N. gaditana* fatty acid profile.



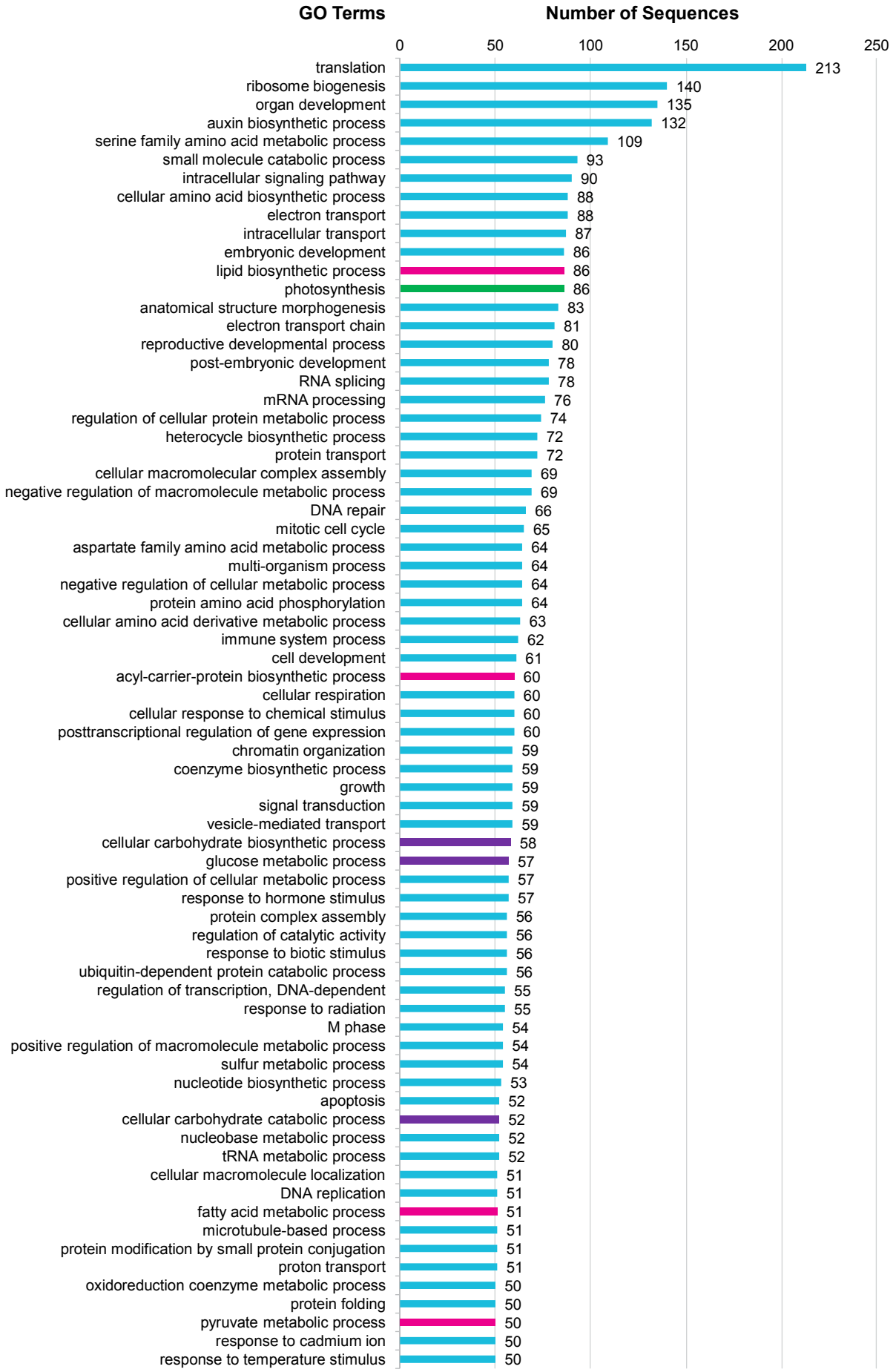
Supplementary Figure S2. Chloroplast genome. Genes on the inside are transcribed clockwise, and genes on the outside are transcribed counter-clockwise.



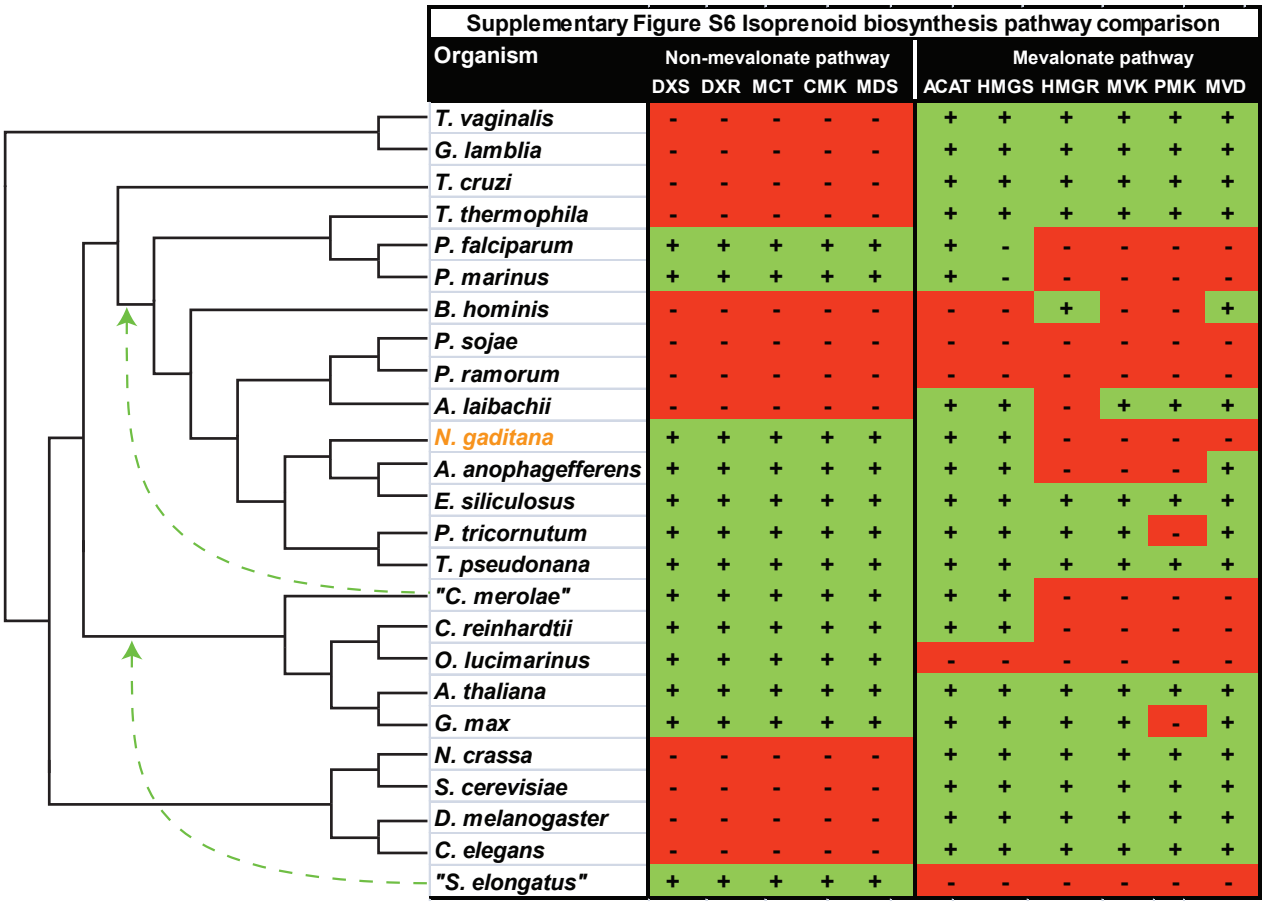
Supplementary Figure S3. Mitochondrial genome. Genes on the inside are transcribed clockwise, and genes on the outside are transcribed counter-clockwise.



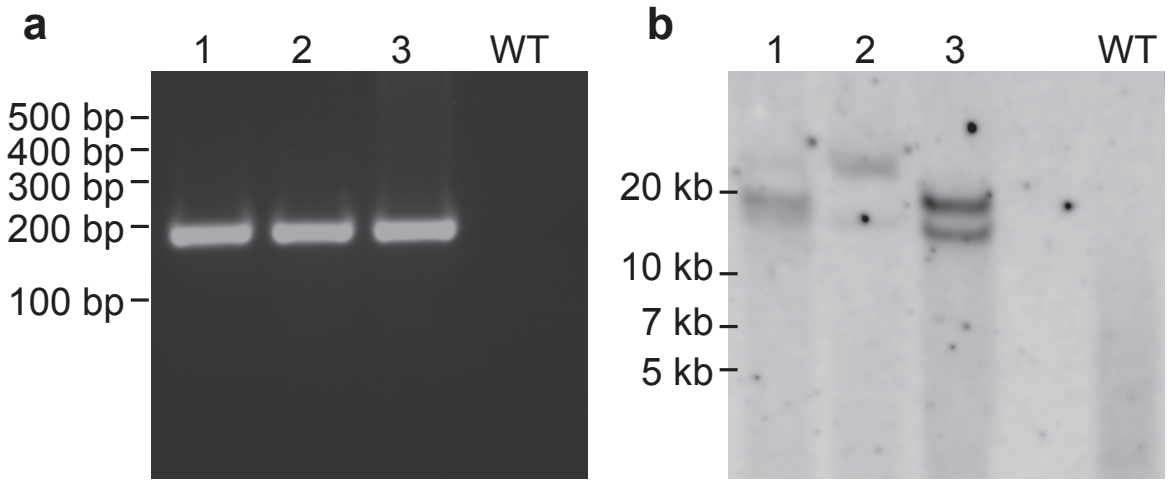
Supplementary Figure S4. Functional gene clusters identified in the *N. gaditana* nuclear genome. Clusters of genes involved in hydrogen metabolism and nitrogen assimilation. Arrows indicate gene direction.


GO Terms
Number of Sequences
0
50
100
150
200
250
translation 213
ribosome biogenesis 140
organ development 135
auxin biosynthetic process 132
serine family amino acid metabolic process 109
small molecule catabolic process 93
intracellular signaling pathway 90
cellular amino acid biosynthetic process 88
electron transport 88
intracellular transport 87
embryonic development 86
lipid biosynthetic process 86
photosynthesis 86
anatomical structure morphogenesis 83
electron transport chain 81
reproductive developmental process 80
post-embryonic development 78
RNA splicing 78
mRNA processing 76
regulation of cellular protein metabolic process 74
heterocycle biosynthetic process 72
protein transport 72
cellular macromolecular complex assembly 69
negative regulation of macromolecule metabolic process 69
DNA repair 66
mitotic cell cycle 65
aspartate family amino acid metabolic process 64
multi-organism process 64
negative regulation of cellular metabolic process 64
protein amino acid phosphorylation 64
cellular amino acid derivative metabolic process 63
immune system process 62
cell development 61
acyl-carrier-protein biosynthetic process 60
cellular respiration 60
cellular response to chemical stimulus 60
posttranscriptional regulation of gene expression 60
chromatin organization 59
coenzyme biosynthetic process 59
growth 59
signal transduction 59
vesicle-mediated transport 59
cellular carbohydrate biosynthetic process 58
glucose metabolic process 57
positive regulation of cellular metabolic process 57
response to hormone stimulus 57
protein complex assembly 56
regulation of catalytic activity 56
response to biotic stimulus 56
ubiquitin-dependent protein catabolic process 56
regulation of transcription, DNA-dependent 55
response to radiation 55
M phase 54
positive regulation of macromolecule metabolic process 54
sulfur metabolic process 54
nucleotide biosynthetic process 53
apoptosis 52
cellular carbohydrate catabolic process 52
nucleobase metabolic process 52
tRNA metabolic process 52
cellular macromolecule localization 51
DNA replication 51
fatty acid metabolic process 51
microtubule-based process 51
protein modification by small protein conjugation 51
proton transport 51
oxidoreduction coenzyme metabolic process 50
protein folding 50
pyruvate metabolic process 50
response to cadmium ion 50
response to temperature stimulus 50

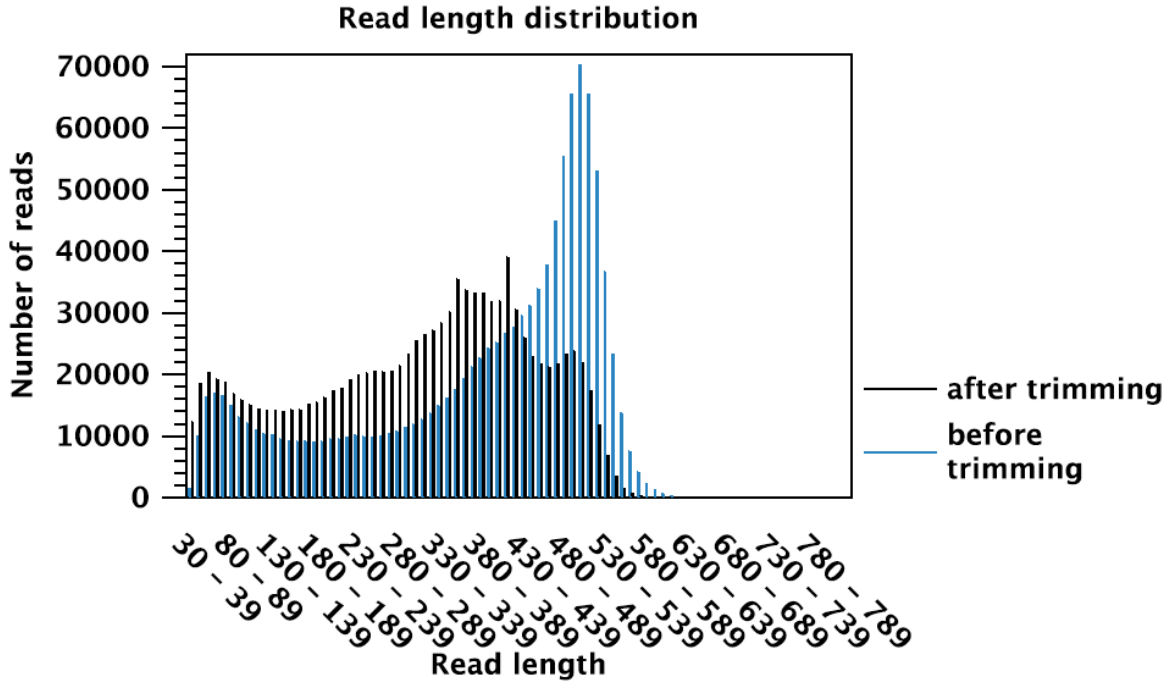
Supplementary Figure S5. Most abundant GO terms assigned to *N. gaditana* gene models. Number of gene models assigned to specified GO term. GO terms associated with lipids (red), carbohydrates (purple), and photosynthesis (green).



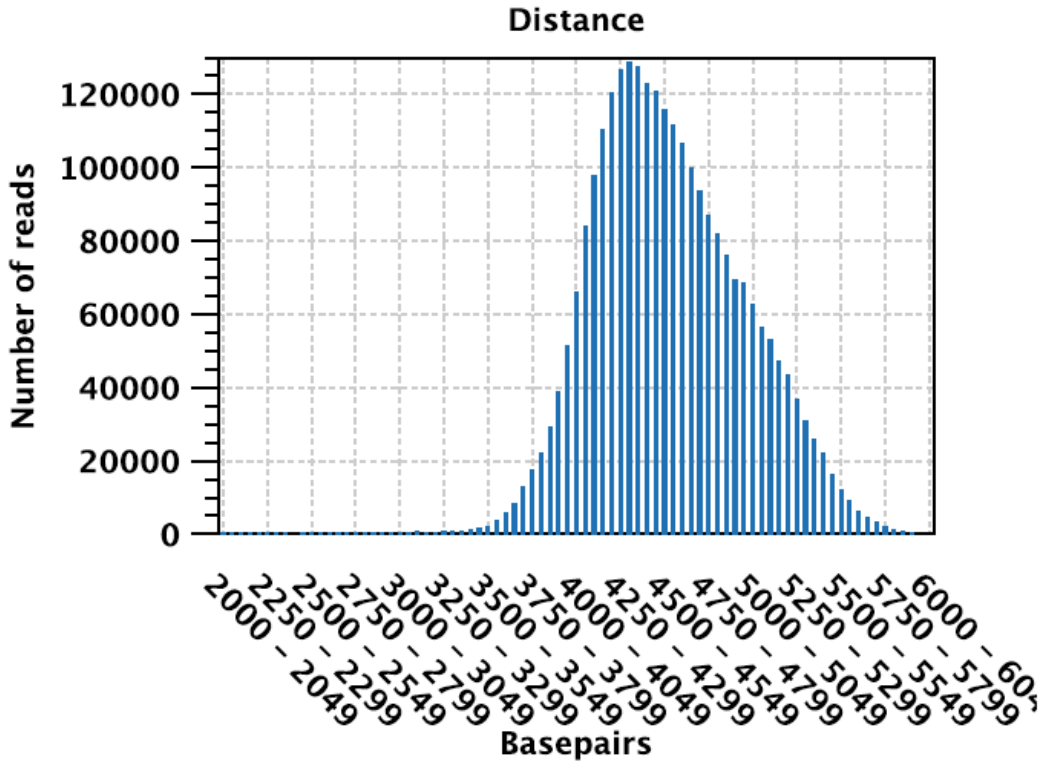
Supplementary Figure S6. Isoprenoid biosynthesis pathway comparison. Simplified phylogenetic cladogram and table showing the relationship between organisms with different sets of isoprenoid biosynthesis pathway genes. Green arrows indicate the acquisition of photosynthetic symbionts thought to have brought the DXP pathway into modern plants and algae. The names of S. elongatus and C. merolae appear in quotes to indicate that the symbiotic events do not refer to the S. elongatus or C. merolae but rather unknown relatives of these species.



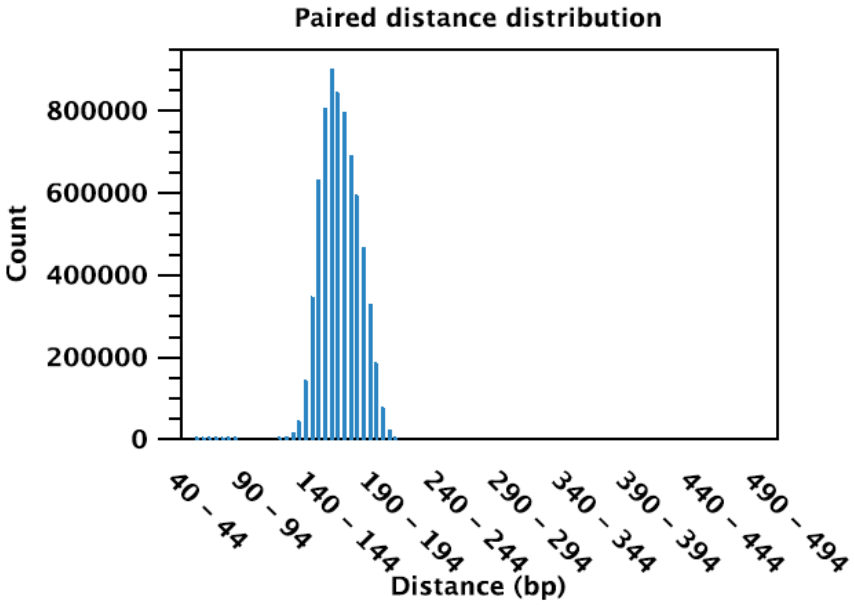
Supplementary Figure S7. Successful transformation of *N. gaditana* by electroporation. **(a)** Genomic PCR confirmation of the presence of the transgene. **(b)** Southern blot indicating the nuclear incorporation and copy number of the Zeocin resistance gene in three different transformants (lanes 1-3), and the lack of the transgene in a wildtype control (WT lane).



Supplementary Figure S8. Roche 454 sequencing read length distribution. Roche 454 sequencing read length distribution before and after trimming.



Supplementary Figure S9. Distribution of paired-end sequencing read separation distances in the genome assembly. Distribution of the separation distances of Illumina LIPES paired-end sequencing reads in the *N. gaditana* genome assembly version 1.1.



Supplementary Figure S10. Distribution of paired-end sequencing read separation distances in the transcriptome assembly. Distribution of paired-end sequencing read separation distances in the assembly of potential transcript contigs from the pooled RNA samples.

SUPPLEMENTARY TABLES

Supplementary Table S1. Chloroplast gene content comparison							
Gene Name	***Guillardia theta***	***Odontella sinensis***	***Phaeodactylum tricornutum***	***Thalassiosira pseudonana***	***Emiliania huxleyi***	***Ectocarpus siliculosus***	***N. gaditana***
acpP	+	+	+	-	-	-	-
atpA	+	+	+	+	+	+	+
atpB	+	+	+	+	+	+	+
atpD	+	+	+	+	+	+	-
atpE	+	+	+	+	+	+	+
atpF	+	+	+	+	+	+	+
atpG	+	+	+	+	+	+	+
atpH	+	+	+	+	+	+	+
atpI	+	+	+	+	+	+	+
cbbX	+	+	+	+	+	+	+
ccs1 (ycf44)	+	+	+	+	+	+	+
ccsA (ycf5)	+	+	+	+	+	+	+
cemA (ycf10)	+	-	-	-	-	-	-
chlB	-	-	-	-	-	+	+
chlI	+	+	+	+	+	+	+
chlJ	-	-	-	-	-	-	-
chlL	-	-	-	-	-	+	+
chlN	-	-	-	-	-	+	+
clpA	-	-	-	-	-	-	+
clpC	+	+	+	+	+	+	+
cpeB	+	-	-	-	-	-	-
dfr	-	-	-	-	+	-	-
dnaB	+	+	+	+	-	+	-
dnaK	+	+	+	+	+	+	+
ftrB	+	-	-	-	-	+	-
ftsH (ycf25)	+	+	+	+	-	+	+
groL	+	+	+	+	+	+	+
hlip (ycf17)	+	-	-	-	-	+	-
hupA	+	-	-	-	-	-	-
ilvB	+	-	-	-	-	+	+

Gene Name	*Guillardia theta*	*Odontella sinensis*	*Phaeodactylum tricornutum*	*Thalassiosira pseudonana*	*Emiliania huxleyi*	*Ectocarpus siliculosus*	*N. gaditana*
ilvH	+	-	-	-	-	+	-
infB	+	-	-	-	-	-	-
minD	+	-	-	-	+	-	-
minE	+	-	-	-	-	-	-
pbsA	+	-	-	-	-	-	-
petA	+	+	+	+	+	+	+
petB	+	+	+	+	+	+	+
petD	+	+	+	+	+	+	+
petF	+	+	+	+	-	+	+
petG	+	+	+	+	+	+	+
petL (ycf7)	+	+	+	+	+	+	-
petM (ycf31)	+	+	+	+	+	+	-
petN (ycf6)	+	+	+	+	+	+	+
psaA	+	+	+	+	+	+	+
psaB	+	+	+	+	+	+	+
psaC	+	+	+	+	+	+	+
psaD	+	+	+	+	+	+	+
psaE	+	+	+	+	-	+	+
psaF	+	+	+	+	+	+	+
psaI	+	+	+	+	+	+	+
psaJ	+	+	+	+	+	+	+
psaK	+	-	-	-	-	-	-
psaL	+	+	+	+	+	+	+
psaM	+	+	+	+	+	+	+
psb28 (ycf79)	+	+	+	+	-	+	+
psbA	+	+	+	+	+	+	+
psbB	+	+	+	+	+	+	+
psbC	+	+	+	+	+	+	+
psbD	+	+	+	+	+	+	+
psbE	+	+	+	+	+	+	+
psbF	+	+	+	+	+	+	+
psbH	+	+	+	+	+	+	+
psbI	+	+	+	+	+	+	+
psbJ	+	+	+	+	+	+	+
psbK	+	+	+	+	+	+	+

Gene Name	*Guillardia theta*	*Odontella sinensis*	*Phaeodactylum tricornutum*	*Thalassiosira pseudonana*	*Emiliania huxleyi*	*Ectocarpus siliculosus*	*N. gaditana*
psbL	+	+	+	+	+	+	+
psbN	+	+	+	+	+	+	+
psbT (ycf8)	+	+	+	+	+	+	+
psbV	+	+	+	+	+	+	+
psbW	-	-	-	-	-	+	+
psbX	+	+	+	+	+	+	-
psbY (ycf32)	+	+	+	+	+	+	+
psbZ (ycf9)	+	+	+	+	+	-	+
rbcL	+	+	+	+	+	+	+
rbcR (ycf30)	+	+	+	+	+	+	-
rbcS	+	+	+	+	+	+	+
rne	+	-	-	-	-	-	-
rpl1	+	+	+	+	-	+	+
rpl11	+	+	+	+	-	+	+
rpl12	+	+	+	+	-	+	+
rpl13	+	+	+	+	-	+	+
rpl14	+	+	+	+	+	+	+
rpl16	+	+	+	+	+	+	+
rpl18	+	+	+	+	-	+	+
rpl19	+	+	+	+	+	+	+
rpl2	+	+	+	+	+	+	+
rpl20	+	+	+	+	+	+	+
rpl21	+	+	+	+	+	+	+
rpl22	+	+	+	+	+	+	+
rpl23	+	+	+	+	+	+	+
rpl24	+	+	+	+	-	+	-
rpl27	+	+	+	+	+	+	+
rpl29	+	+	+	+	-	+	-
rpl3	+	+	+	+	+	+	+
rpl31	+	+	+	+	+	+	+
rpl32	+	+	+	+	-	+	+
rpl33	+	+	+	+	+	+	+
rpl34	+	+	+	+	+	+	+
rpl35	+	+	+	+	-	+	+
rpl36	+	+	+	+	+	+	+

Gene Name	*Guillardia theta*	*Odontella sinensis*	*Phaeodactylum tricornutum*	*Thalassiosira pseudonana*	*Emiliania huxleyi*	*Ectocarpus siliculosus*	*N. gaditana*
rpl4	+	+	+	+	-	+	+
rpl5	+	+	+	+	+	+	+
rpl6	+	+	+	+	+	+	+
rpmC	-	-	-	-	-	+	+
rpoA	+	+	+	+	+	+	+
rpoB	+	+	+	+	+	+	+
rpoC1	+	+	+	+	+	+	+
rpoC2	+	+	+	+	+	+	+
rpoZ (ycf61)	+	-	-	-	-	-	-
rps10	+	+	+	+	+	+	+
rps11	+	+	+	+	+	+	+
rps12	+	+	+	+	+	+	+
rps13	+	+	+	+	+	+	+
rps14	+	+	+	+	+	+	+
rps16	+	+	+	+	+	+	+
rps17	+	+	+	+	+	+	+
rps18	+	+	+	+	+	+	+
rps19	+	+	+	+	+	+	+
rps2	+	+	+	+	+	+	+
rps20	+	+	+	+	-	+	+
rps3	+	+	+	+	+	+	+
rps4	+	+	+	+	+	+	+
rps5	+	+	+	+	+	+	+
rps6	+	+	+	+	+	-	+
rps7	+	+	+	+	+	+	+
rps8	+	+	+	+	+	+	+
rps9	+	+	+	+	+	+	+
secA	+	+	+	+	+	+	+
secG (ycf47)	+	+	+	+	+	+	-
secY	+	+	+	+	+	+	+
sufB (ycf24)	+	+	+	+	+	+	+
sufC (ycf16)	+	+	+	+	-	+	+
syfB	-	-	+	-	-		-
tatC (ycf43)	+	+	+	+	+		+

Gene Name	*Guillardia theta*	*Odontella sinensis*	*Phaeodactylum tricornutum*	*Thalassiosira pseudonana*	*Emiliania huxleyi*	*Ectocarpus siliculosus*	*N. gaditana*
thiG	-	+	+	+	+	+	+
thiS (ycf40)	-	+	+	+	+	+	+
tsf	+	-	+	-	-	+	-
tufA	+	+	+	+	+	+	+
ycf12	+	+	+	+	+	+	+
ycf19	+	-	-	-	+	+	+
ycf20	+	-	-	-	+	-	-
ycf27	+	-	-	-	+	-	-
ycf29	+	-	-	-	-	-	-
ycf3	+	+	+	+	+	+	+
ycf33	+	+	+	+	-	+	-
ycf34	-	-	-	-	-	+	+
ycf35	+	+	+	+	+	+	-
ycf36	+	-	-	-	-	-	-
ycf37	+	-	-	-	-	+	-
ycf39	+	+	+	+	+	+	-
ycf4	+	+	+	+	+	+	+
ycf41	-	+	+	+	-	+	-
ycf42	-	+	+	+	-	+	-
ycf45	-	+	+	+	+	-	-
ycf46	+	+	+	+	+	+	+
ycf54	-	-	-	-	-	+	+
ycf55	-	-	-	-	+	-	-
ycf59	-	-	-	-	-	-	+
ycf60	-	-	-	-	+	+	-
ycf65	+	-	-	-	+	+	-
ycf66	-	+	+	+	-	+	+
ycf80	+	-	-	-	+	-	-
ycf88	-	+	+	+	-	-	-
ycf89	-	+	+	+	-	-	-
ycf90	-	+	+	+	-	-	-

Supplementary Table S2. Gene model naming scheme			
Name Scheme[a]	**Gene model source[b]**	**# of gene models[c]**	**# of nr gene models[d]**
Nga0xxxx	Maker genes	7,315	6,361
Nga1xxxx	Manually added genes	-	-
Nga2xxxx	EST genes	1,310	1,310
Nga3xxxx	EST genes w/o assembly support	1,221	1,221
Nga4xxxx	Chloroplast genes	124	124
Nga5xxxx	Mitochondrial genes	36	36
	Total *N. gaditana* genes (v1)	**10,006**	**9,052**

[a] *N. gaditana* gene model naming scheme where the first digit of the 5-digit number identifies the source of that model. The 'x' represents possible digits (0-9) that uniquely identify the gene model.

[b] Identifies the methodology used to determine the gene model.

[c] Total number of gene models identified with each method. Gene models that are duplicates are included. For Nga v1 no genes were added manually.

[d] Total number of non-redundant gene models identified with each method. Gene models that are duplicates are not included.

Supplementary Table S3. Gene clusters		
GO-term[a]	**GO-term descritipion[b]**	**Scaffold[c]**
N. gaditana model[d]	Gene model description[e]	Gene #[f]
GO:0003674	**molecular function**	**merged000156**
Nga06291	hydE (radical sam domain protein)	1
Nga06282.1	hydG(biotin and thiamin synthesis associated)	2
Nga02278.2	hydA (fe-only)	3
Nga06289	hydF (small gtp-binding protein)	4
GO:0006807	**nitrogen compound metabolic process**	**merged000381**
Nga02268	nitrate reductase	1
Nga02267	ferredoxin-nitrite reductase	2
Nga10006	nitrate high affinity transporter	3
GO:0010467; GO:0016070	**gene expression; RNA metabolic process**	**merged000024**
Nga00538	small nuclear ribonucleoprotein associated protein b	1
Nga00552	translesion dna polymerase-rev1 deoxycytidyl transferase	14
Nga00540	ash1 (or homeotic)-like	15
Nga20782	50s ribosomal protein l9	17
Nga00535	histone-lysine n-methyltransferase	19
Nga00543	ef-1 guanine nucleotide exchange domain-containing	20
Nga00558	ribosomal protein s16	21
Nga00583	cg9383-pa	26
Nga00593	polyribonucleotide nucleotidyltransferase	27
Nga21110	polyribonucleotide nucleotidyltransferase	28
Nga00596	rna helicase rnase	37
Nga00595	dicer-like protein 2	38
Nga20178	dicer-1	39
Nga00568	cdc2-like protein kinase	43
Nga00562	5 -3 exoribonuclease 2	63
Nga00557	glycyl-trna synthetase	70
Nga00581.01	dna polymerase v family	73
Nga00566.01	protein bud31 homolog	74
GO:0042967	**acyl-carrier-protein biosynthetic process**	**merged000123**
Nga02737	phospholipid:diacylglycerol acyltransferase	1
Nga02743	serine palmitoyltransferase	9
Nga02741	pyruvate dehydrogenase component x	10
GO:0006807	**nitrogen compound metabolic process**	**merged000138**
Nga00713	guanine deaminase	1
Nga00717	cytohesin-1 isoform 2	3
Nga00702	serine threonine-protein kinase tousled-like 1 isoform 2	5
GO:0008610	**lipid biosynthetic process**	**merged000154**
Nga00818	monogalactosyldiacylglycerol synthase	1

Nga00817	fatty acid desaturase	3
GO:0046394	**carboxylic acid biosynthetic process**	**merged000212**
Nga05525	anthranilate synthase	1
Nga05530	cystathionine beta-synthase	5
Nga05522	dihydroxy-acid dehydratase	11
GO:0006778	**porphyrin metabolic process**	**merged000288**
Nga03876	protoporphyrinogen oxidase	1
Nga03879	protoporphyrinogen oxidase	2
Nga03873	protoporphyrinogen oxidase	5
GO:0006397	**mRNA processing**	**merged000340**
Nga20580	rrna processing protein rrp17	1
Nga30171	u4 u6 small nuclear ribonucleoprotein prp3	2
Nga30821	u4 u6 small nuclear ribonucleoprotein prp4	3
Nga05511	prp4 pre-mrna processing factor 4 homolog	4
Nga30351	u4 u6 small nuclear ribonucleoprotein prp4	6
GO:0016070	**RNA metabolic process**	**merged000438**
Nga03294	tfiih subunit	1
Nga30114	protein	2
Nga03302	polyadenlyte binding protein	3
Nga03296	b chain structure of the mlle domain of poly-binding protein in complex with the binding region of paip2	5
GO:0006810	**transport**	**merged000462**
Nga31189	abc transporter	1
Nga03486	abc atp-binding permease protein	3
Nga03479	atp-binding cassette superfamily	6
GO:0034641	**cellular nitrogen compound metabolic process**	**NODE_3127**
Nga03304	s-adenosylmethionine synthetase	1
Nga03314	glutamyl-trna amidotransferase subunit a	2
Nga03312	cytidine and deoxycytidylate deaminase family protein	3
Nga03305	rna pseudouridylate synthase family protein	4
Nga03303	carbamoyl-phosphate small subunit	5
Nga03313	dihydropteroate synthase	6
Nga03316	aspartate aminotransferase	10
Nga03306	thymidylate synthase	11
Nga03310	dihydrofolate reductase	12
Nga03311	threonyl-trna synthetase	13
GO:0006139	**nucleobase, nucleoside, nucleotide and nucleic acid metabolic process**	**NODE_4378**
Nga30490	polymerase (dna directed) epsilon 2 (p59 subunit)	1
Nga03669	uv-damaged dna-binding	2
Nga30761	gtp-binding protein	3
Nga30709	dna damage-binding	4
GO:0006139; GO:0006310	**nucleobase, nucleoside, nucleotide and nucleic acid metabolic process; DNA recombination**	**NODE_5200**

Nga30978	deah (asp-glu-ala-his) box polypeptide 8-like	1
Nga03202	dna mismatch repair protein msh2	3
Nga03203	dna mismatch repair protein msh2	5
Nga30317	dna mismatch repair protein msh2	10
Nga30528	protein	13
GO:0006810	**transport**	**NODE_521**
Nga01558	chromate transporter	1
Nga01560	translocase of inner mitochondrial membrane 13 homolog	4
Nga01563	lipid a export atp-binding permease protein msba	5
Nga20851	at2g36910-like protein	6
Nga20388	atp-binding sub-family b (mdr tap) member 10	7
GO:0042967	**acyl-carrier-protein biosynthetic process**	**NODE_5292**
Nga04041	dihydrolipoamide acetyltransferase	1
Nga04043	branched-chain alpha-keto acid dehydrogenase subunit e2	3
Nga04040	branched-chain alpha-keto acid dehydrogenase subunit e2	6
GO:0006260	**DNA replication**	**NODE_5379**
Nga30513	dna replication licensing factor mcm3	1
Nga01778	dna replication licensing factor mcm3	5
GO:0009081	**branched chain family amino acid metabolic process**	**NODE_5711**
Nga04063	dihydroxy-acid dehydratase	1
Nga30327	dihydroxy-acid dehydratase	3
Nga30644	methylmalonate-semialdehyde dehydrogenase	8
GO:0006536	**glutamate metabolic process**	**NODE_5865**
Nga30142	glutamate--cysteine ligase catalytic subunit	1
Nga30765	glutamate-cysteine catalytic subunit	2
GO:0006732	**coenzyme metabolic process**	**NODE_653**
Nga02623.2	molybdenum cofactor synthesis 1	1
Nga01769.1	5-oxoprolinase	2
Nga01765.01	cystathionine gamma-lyase	6

[a] Gene ontology term that defines gene cluster.

[b] Gene ontology term description.

[c] Name of scaffold that gene cluster is found on.

[d] *N. gaditana* gene model.

[e] *N. gaditana* gene model description.

[f] Gene location in cluster.

Supplementary Table S4. Chlorophyll (tetrapyrrole), carotenoid and sterol biosynthesis genes					
Enzyme	**Description**	**EC number**	***N. gaditana* model[a]**	**Transcript Support[b]**	**Gene Location[c]**
Tetrapyrrole Synthesis					
GTS	glutamyl-tRNA synthetase	6.1.1.17	Nga04989	Y	N
GTS	glutamyl-tRNA synthetase	6.1.1.24	Nga02834	Y	N
GTR	glutamyl-tRNA reductase	1.2.1.70	Nga02604	Y	N
GSA	glutamate-1-semialdehyde aminotransferase / glutamate-1-semialdehyde 21-aminomutase	5.4.3.8	Nga30045	Y	N
ALAD (HemB)	5-aminolevulinic acid dehydratase / porphobilinogen synthase	4.2.1.24	Nga00585	Y	N
PBGD (HemC)	porphobilinogen deaminase / hydroxymethylbilane synthase	2.5.1.61	Nga03248	Y	N
UROS (HemD)	uroporphyrinogen III synthase	4.2.1.75	Nga00807	Y	N
UROD	uroporphyrinogen III decarboxylase	4.1.1.37	Nga04120	Y	N
UROD	uroporphyrinogen III decarboxylase	4.1.1.37	Nga05706	Y	N
CPX1 (HemF)	coproporphyrinogen III oxidase	1.3.3.3	Nga05151	Y	N
CPX1 (HemF)	coproporphyrinogen III oxidase	1.3.3.3	Nga04278 (partial)	Y	N
PPX	protoporphyrinogen IX oxidase	1.3.3.4	Nga03873	Y	N
ChlD	protoporphyrin IX Mg-chelatase subunit D	6.6.1.1	Nga30773	Y	N
ChlI	protoporphyrin IX MG-chelatase subuint I	6.6.1.1	Nga40092	N	C
ChlH1	protoporphyrin IX Mg-chelatase subunit H	6.6.1.1	Nga30995	Y	N
ChlH2	protoporphyrin IX Mg-chelatase subunit H	6.6.1.1	Nga06242	Y	N
PPMT (ChlM)	Mg-protoporphyrin IX methyltransferase	2.1.1.11	Nga04808	Y	N
AcsF (ycf59)	Mg-protoporphyrin IX monomethyl ester (oxidative) cyclase	1.14.13.81	Nga40091	N	C
DVR	divinyl protochlorophyllide a 8-vinyl-reductase	1.3.1.75	Nga05945	Y	N
POR	light-dependent NADPH:protochlorophyllide oxidoreductase	1.3.1.33	Nga04959	Y	N
POR	light-dependent NADPH:protochlorophyllide oxidoreductase	1.3.1.33	Nga00683	Y	N
ChlB	light-independent:protochlorophyllide oxidoreductase subunit B	1.18.-.-	Nga40089	N	C
ChlL	light-independent:protochlorophyllide oxidoreductase subunit L	1.18.-.-	Nga40044	Y	C
ChlN	light-independent:protochlorophyllide oxidoreductase subunit N	1.18.-.-	Nga40045	N	C
CHS (ChlG)	chlorophyll synthase	2.5.1.62	Nga31097	Y	N
GGR	geranylgeranyl reductase	1.3.1.-	Nga04895	Y	N

(ChlP)					
UMT	uroporphyrinogen III C-methyltransferase	2.1.1.107	Nga05160	Y	N
Enzyme	**Description**	**EC number**	***N. gaditana* model[a]**	**Transcript Support[b]**	**Gene Location[c]**
SirB	sirohydrochlorin ferrochelatase	4.99.1.4	Nga00339	Y	N
FC (HemH)	ferrochelatase	4.99.1.1	Nga00748	Y	N
Carotenoid Biosynthesis					
DXS	1-deoxy-D-xylulose-5-phosphate synthase	2.2.1.7	Nga02203	Y	N
DXR (IspC)	1-deoxy-D-xylulose-5-phosphate reductoisomerase	1.1.1.267	Nga30771	Y	N
MCT (IspD)	2-C-methyl-D-erythritol 4-phosphate cytidyltransferase	2.7.7.60	Nga06198	Y	N
CMK (IspE)	4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase	2.7.1.148	Nga04584	Y	N
MDS (IspF)	2-C-methyl-D-erythritol 24-cyclodiphosphate synthase	4.6.1.12	Nga02651	Y	N
HDS (IspG)	4-hydroxy-3-methylbut-2-enyl diphosphate synthase	1.17.4.3	Nga30806	Y	N
HDR (IspH)	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	1.17.1.2	Nga05308	Y	N
IDI	isopentenyl diphosphate:dimethylallyl diphosphate isomerase type I	5.3.3.2	Nga03838	Y	N
IDI	isopentenyl diphosphate:dimethylallyl diphosphate isomerase type I	5.3.3.2	Nga10002	Y	N
GGPPS (CrtE)	geranylgeranyl pyrophosphate synthase	2.5.1.29	Nga02636	Y	N
PSY1 (CrtB)	phytoene synthase	2.5.1.32	Nga02957	Y	N
PDS (CrtP)	phytoene desaturase	1.14.99.-	Nga05064	Y	N
ZDS (CrtQ)	zeta-carotene desaturase	1.14.99.30	Nga07310	Y	N
CRTISO	carotenoid isomerase		No Homolog	-	-
LCYB (CrtL-b)	lycopene β-cyclase	1.14.-.-	Nga00640	Y	N
CYP97E	cytochrome P450 enzyme related to CYP97A carotene β-hydroxylase		Nga30077	Y	N
CYP97F	cytochrome P450 enzyme related to CYP97A carotene β-hydroxylase		Nga00100	Y	N
ZEP	zeaxanthin epoxidase	1.14.13.90	Nga01534	Y	N
VDE	violaxanthin de-epoxidase	1.10.99.3	Nga02700	Y	N
NSY	neoxanthin synthase	5.3.99.9	Nga10001	Y	N
Sterol Synthesis					
ACAT	acetyl-CoA C-acetyltransferase	2.3.1.9	Nga20998	Y	N
ACAT	acetyl-CoA C-acetyltransferase	2.3.1.9	Nga30830	Y	N
HMGS	hydroxymethylglutaryl-CoA synthase	2.3.3.10	Nga06246	Y	N
HMGR	hydroxymethylglutaryl-CoA reductase	1.1.1.34	No	-	-

			Homolog		
MVK	mevalonate kinase	2.7.1.36	No Homolog	-	-
Enzyme	**Description**	**EC number**	***N. gaditana* model[a]**	**Transcript Support[b]**	**Gene Location[c]**
PMK	phosphomevalonate kinase	2.7.4.2	No Homolog	-	-
MVD	diphosphomevalonate decarboxylase	4.1.1.33	No Homolog	-	-
GPPS	geranyl-disphosphate synthase / dimethylallyltranstransferase	2.5.1.1	Nga02865	Y	N
GPPS	geranyl-disphosphate synthase / dimethylallyltranstransferase	2.5.1.1	Nga01978	Y	N
FPPS	farnesyl-diphosphate synthase	2.5.1.68	Nga02874	Y	N
IDI	isopentenyl diphosphate:dimethylallyl diphosphate isomerase type I	5.3.3.2	Nga03838	Y	N
IDI	isopentenyl diphosphate:dimethylallyl diphosphate isomerase type I	5.3.3.2	Nga10002	Y	N
SQE (SQP)	squalene monoxygenase / squalene epoxidase	1.14.99.7	Nga01590	Y	N
CAS	cycloartenol synthase	5.4.99.8	Nga30790	Y	N
CYP51	C14-demethylase (sterol 14-demethylase)	1.14.13.70	Nga03733.1	Y	N
FACKEL	D14-sterol reductase	1.3.1.70	Nga00758	N	N
SMO	Sterol methyl-oxidase (C4-methylsterol monoxygenase)	1.14.13.72	No Homolog	-	-
HSD-D	C4-decarboxylase (sterol-4-alpha-carboxylate 3-dehydrogenase)	1.1.1.170	Nga06114	Y	N
SMT1	sterol-C24-methyl transferase (sterol 24-C-methyltransferase)	2.1.1.41	Nga05943	Y	N
SMT1	sterol-C24-methyl transferase (sterol 24-C-methyltransferase)	2.1.1.41	Nga02534	Y	N
HYD1	D8-D7-sterol-isomerase (cholestenol delta-isomerase)	5.3.3.5	No Homolog	-	-
DWF7 (STE1)	C5-desaturase (lathosterol oxidase)	1.14.21.6	Nga02795	Y	N
DWF5	D7-sterol reductase (7-dehydrocholesterol reductase)	1.3.1.21	Nga03254	Y	N
DWF1	D24-sterol reductase (24-dehydrocholesterol reductase)	1.3.1.72	Nga03764	Y	N
DWF1	D24-sterol reductase (24-dehydrocholesterol reductase)	1.3.1.72	Nga05293	Y	N
CPI (CCI)	cyclopropyl sterol isomerase (cycloeucalenol cycloisomerase)	5.5.1.9	Nga30196	Y	N
DET2	D5-sterol reductase	1.3.1.30	Nga00656	Y	N

[a] Candidate *N. gaditana* gene model encoding corresponding enzyme.

[b] Indicates if given gene model has transcript support from RNAseq of a pool of conditions including: +/- nitrate, logarithmic phase, stationary phase, heat shocked culture (2h at 37ºC), cold treated culture (2h at 4ºC), culture after 12h dark, +/- CO_2..

[c] Indicates if the gene model is located in the nuclear genome (N) or chloroplast genome (C).

Supplementary Table S5. Lipid metabolic pathway genes				
Enzyme	**Description**	**EC number**	***N. gaditana* model[a]**	**Transcript support[b]**
Fatty acid biosynthesis				
ACS	acetyl-CoA synthetase	6.2.1.1	Nga30020	Y
ACS	acetyl-CoA synthetase	6.2.1.1	Nga02166	Y
ACC	acetyl-CoA carboxylase	6.4.1.2	Nga30028	Y
ACC	acetyl-CoA carboxylase	6.4.1.2	Nga30783	Y
MAT	malonyl-CoA:ACP-trans-acylase	2.3.1.39	Nga00899	Y
KASIII	3-oxoacyl-ACP synthase III	2.3.1.180	Nga30106	Y
KASII	3-oxoacyl-ACP synthase II	2.3.1.179	Nga02138	Y
KASII	3-oxoacyl-ACP synthase II	2.3.1.179	Nga30755	Y
KASII	3-oxoacyl-ACP synthase II	2.3.1.179	Nga04201.01	Y
KASI/II	fatty acid synthase	2.3.1.-	Nga05827	Y
KAR	3-oxoacyl-ACP reductase	1.1.1.100	Nga20022	Y
KAR	3-oxoacyl-ACP reductase	1.1.1.100	Nga01542	Y
HD	3-hydroxy acyl-CoA dehydratase	4.2.1.-	Nga05985.01	Y
ENR	enoyl-ACP reductase I	1.3.1.9	Nga05091	Y
ENR	enoyl-ACP reductase I	1.3.1.9	Nga10004	Y
KAR	3-oxoacyl-ACP reductase	1.1.1.100	Nga00602	Y
KAR	3-oxoacyl-ACP reductase	1.1.1.100	Nga05627	Y
FAT	oleyl-ACP hydrolase	3.1.2.14	Nga01045.01	Y
ELOVL	elongation of very long chain fatty acids	2.3.1.-	Nga03084	Y
ELOVL	elongation of very long chain fatty acids	2.3.1.-	Nga00451	Y
DESA	acyl-ACP desaturase	1.14.19.2	Nga01458.01	Y
DESC	stearoyl-CoA desaturase	1.14.19.1	Nga00524	Y
DES	omega-6 fatty acid desaturase delta-12	1.14.19.-	Nga02019	Y
DES	omega-6 fatty acid desaturase delta-12	1.14.19.-	Nga00817	Y
DEGS	sphingolipid delta-4 desaturase	1.14.-.-	Nga06739.1	Y
FAD7	glycerolipid omega-3 fatty acid desaturase	1.14.19.-	Nga30138	Y
TAG assembly				
G3PDH	glycerol-3-phosphate dehydrogenase	1.1.1.8	Nga06665.1	Y
G3PDH	glycerol-3-phosphate dehydrogenase	1.1.1.8	Nga01869	Y
G3PDH	glycerol-3-phosphate dehydrogenase	1.1.1.8	Nga04914	Y
G3PDH	glycerol-3-phosphate dehydrogenase	1.1.1.8	Nga05226	Y
G3PDH	glycerol-3-phosphate dehydrogenase	1.1.1.8	Nga30015	Y
GPAT	glycerol-3-phosphate acyltransferase	2.3.1.15	Nga04759	Y
GPAT	glycerol-3-phosphate acyltransferase	2.3.1.15	Nga10005	Y

Enzyme	Description	EC number	*N. gaditana* model[a]	Transcript support[b]
GPAT	glycerol-3-phosphate acyltransferase	2.3.1.15	Nga04850	Y
LPAAT	1-acylglycerol-3-phosphate O-acyltransferase	2.3.1.51	Nga00059	Y
LPAAT	1-acylglycerol-3-phosphate O-acyltransferase	2.3.1.51	Nga30581	Y
LPAAT	1-acylglycerol-3-phosphate O-acyltransferase	2.3.1.51	Nga30809	Y
LPAAT	1-acylglycerol-3-phosphate O-acyltransferase	2.3.1.51	Nga02265	Y
LPAAT	1-acylglycerol-3-phosphate O-acyltransferase	2.3.1.51	Nga21122.1	Y
PAP	phosphatidic acid phosphatase	3.1.3.4	Nga21116	Y
DAGK	diacylglycerol kinase	2.7.1.107	Nga02796	Y
DAGK	diacylglycerol kinase	2.7.1.107	Nga05586.1	Y
PDAT	phospholipid:diacylglycerol acyltransferase	2.3.1.158	Nga02737	Y
DGAT	diacylglycerol acyltransferase	2.3.1.20	Nga10003	Y
DGAT	diacylglycerol acyltransferase	2.3.1.20	Nga30544	Y
DGAT2	diacylglycerol acyltransferase 2	2.3.1.20	Nga00576	Y
DGAT2	diacylglycerol acyltransferase 2	2.3.1.20	Nga00753.01	Y
DGAT2	diacylglycerol acyltransferase 2	2.3.1.20	Nga00933	Y
DGAT2	diacylglycerol acyltransferase 2	2.3.1.20	Nga02378.1	Y
DGAT2	diacylglycerol acyltransferase 2	2.3.1.20	Nga02867	Y
DGAT2	diacylglycerol acyltransferase 2	2.3.1.20	Nga03484	Y
DGAT2	diacylglycerol acyltransferase 2	2.3.1.20	Nga06249	Y
DGAT2	diacylglycerol acyltransferase 2	2.3.1.20	Nga30139	Y
LPAT	lysophosphatidylglycerol acyltransferase	2.3.1.-	Nga05465	Y
PKS5	polyketide synthase 5	KO K12433	Nga00335	Y
MGD	monogalactosyldiacylglycerol synthase	2.4.1.46	Nga00818	Y
SQD2	sulfoquinovosyltransferase	2.4.1.-	Nga00561	Y
SQD1	UDP-sulfoquinovose synthase	3.13.1.1	Nga02686	Y
Lipid activation				
TAGL	TAG-lipase	3.1.1.3	Nga30958	Y
TAGL	TAG-lipase	3.1.1.3	Nga30749	Y
AASDH	acyl-CoA synthetase	6.2.1.-	Nga03422	Y
AASDH	acyl-CoA synthase	6.2.1.-	Nga05597	Y
ACSL	long-chain acyl-CoA synthetase	6.2.1.3	Nga30170	Y
ACSL	long-chain acyl-CoA synthetase	6.2.1.3	Nga03113	Y
ACSL	long-chain acyl-CoA synthetase	6.2.1.3	Nga00675	Y
ACSL	long-chain acyl-CoA synthetase	6.2.1.3	Nga30631	Y
ACSL	long-chain acyl-CoA ligase	6.2.1.3	Nga00919	Y
ACSL	long-chain acyl-CoA ligase	6.2.1.3	Nga02299	Y
FADD9	long-chain acyl-CoA ligase	6.2.1.-	Nga00006	Y

ACSL	long-chain acyl-CoA ligase	6.2.1.3	Nga30568	Y
Enzyme	**Description**	**EC number**	***N. gaditana* model[a]**	**Transcript support[b]**
HADH	3-hydroxyacyl-CoA dehydrogenase	1.1.1.35	Nga00482.01	Y
HADH	3-hydroxyacyl-CoA dehydrogenase	1.1.1.35	Nga02679	Y
HADH	3-hydroxyacyl-CoA dehydrogenase	1.1.1.35	Nga06144.1	Y
LCD	long-chain hydroxyacyl-CoA dehydrogenase	1.1.1.211	Nga03480	Y
ACADS	short-chain acyl-CoA dehydrogenase	1.3.8.1	Nga31130	Y
ACADSB	short-branched-chain acyl-CoA dehydrogenase	1.3.99.12	Nga05705	Y
ACDH	Acyl-CoA dehydrogenase	1.3.99.3	Nga04204.01	Y
ACOX	acyl-CoA oxidase	1.3.3.6	Nga03053	Y
PTER	peroxisomal trans-2-enoyl-CoA reductase	1.3.1.38	Nga06128	Y
AOX	alternative oxidase	1.-.-.-	Nga03545	Y
AOX	alternative oxidase	1.-.-.-	Nga30837	Y
AOX	alternative oxidase	1.-.-.-	Nga03289	Y
ACOX	acyl-CoA oxidase	1.3.3.6	Nga04370.1	Y
ACOX	acyl-CoA oxidase	1.3.3.6	Nga30819	Y
ECH	enyol-CoA hydratase	4.2.1.17	Nga01761.01	Y
ECH	enyol-CoA hydratase	4.2.1.17	Nga00171	Y
ECH	enyol-CoA hydratase	4.2.1.17	Nga20152	Y
ECH	enyol-CoA hydratase	4.2.1.17	Nga06135	Y
KCT1	beta-ketoacyl-CoA thiolase	2.3.1.16	Nga01710	Y
KCT2	beta-ketoacyl-CoA thiolase	2.3.1.16	Nga04504.01	Y
KCT3	beta-ketoacyl-CoA thiolase	2.3.1.16	Nga30830	Y
ACAT	acetyl-CoA acetyltransferase	2.3.1.9	Nga20998	Y
KCT3	beta-ketoacyl-CoA thiolase	2.3.1.16	Nga30830	Y
FATP2	solute carrier family (fatty acid transporter)	6.2.1.-	Nga06551	Y
DECR2	peroxisomal 2,4-dienoyl-CoA reductase	1.3.1.34	Nga05502.01	Y
PNPLA	patatin-like phospholipase	3.1.1.-	Nga03028.01	Y

[a] Candidate *N. gaditana* gene model encoding corresponding enzyme.

[b] Indicates if given gene model has transcript support from RNAseq of a pool of conditions including: +/- nitrate, logarithmic phase, stationary phase, heat shocked culture (2h at 37ºC), cold treated culture (2h at 4ºC), culture after 12h dark, +/- CO_2..

Supplementary Table S6. Lipid metabolic gene comparison[a]					
	N. gaditana	*E. siliculosus*	*P. tricornutum*	*C. merolae*	*C. reinhardtii*
Total # of genes[b]	**9,052**	**16,256**	**10,402**	**5,331**	**15,143**
Fatty acid biosynthesis[c]					
ACC	2	5	1	1	1
MAT	1	1	1	1	2
KASIII	1	1	1	0	1
KASI/II	4	6	4	2	3
KAR	4	5	2	4	4
HD	1	1	1	1	1
ENR	2	2	2	1	1
Total[d]	**15**	**21**	**12**	**10**	**13**
TAG assembly[c]					
G3PDH	5	1	2	3	6
GPAT	3	2	1	0	0
LPAAT	5	5	4	6	1
PAP	1	1	1	0	2
DAGK	2	2	1	1	2
PDAT	1	0	1	0	1
DGAT	10	6	4	1	5
Total[d]	**27**	**17**	**14**	**11**	**17**
Lipid degradation[c]					
TAGL	2	1	2	1	1
AASDH	2	0	0	0	1
ACSL	7	6	5	4	4
FADD9	1	0	0	0	0
HADH	3	2	3	2	2
LCD	1	1	1	0	0
ACADS	1	1	5	1	3
ACADSB	1	0	0	0	0
ACDH	1	1	1	0	0
ACOX	3	0	1	1	3
AOX	3	2	2	3	5
ECH	4	8	8	3	3
KCT1	1	1	1	0	1
KCT2	1	1	1	0	0
KCT3	2	1	1	0	0
ACAT	1	0	1	2	1
Total[d]	**34**	**25**	**32**	**17**	**24**
Total in all categories[e]	**76**	**63**	**58**	**38**	**54**

[a] Comparison of the number of copies of lipid metabolic genes that are homologous between *N. gaditana*, brown algae (*E. siliculosus*), diatoms (*P. tricornutum*), red algae (*C. merolae*) and green algae (*C. reinhardtii*).

[b] Total number of genes in this organism.

[c] Category of lipid metabolic genes, sorted by fatty acid biosynthesis, TAG assembly and lipid degradation.

[d] Total number of genes that are listed in specified category of the lipid metabolism.

[e] Total number of genes in all listed categories of lipid metabolic pathways.

Supplementary Table S7 Selected GO-term expansions/reductions[a]			
GO term[b]	Over- or Under-represented[c]	p-value[d] *P. tricornutum*	p-value[d] *C. reinhardtii*
auxin biosynthetic process	over	2.20E-53	1.26E-55
acyl-carrier-protein biosynthetic process	over	2.22E-25	1.35E-30
pyruvate metabolic process	over	1.96E-06	1.45E-09
carbon utilization	over	2.44E-07	2.54E-04
response to stress	over	8.46E-07	1.87E-12
response to chemical stimulus	over	6.11E-08	1.97E-16
posttranscriptional regulation of gene expression	over	2.76E-04	2.38E-04
halogenated hydrocarbon metabolic process	over	3.15E-05	3.51E-06
chlorinated hydrocarbon metabolic process	over	3.15E-05	3.51E-06
cellular ketone metabolic process	over	3.88E-14	2.63E-21
carboxylic acid metabolic process	over	5.51E-13	7.07E-20
organic acid metabolic process	over	1.55E-12	8.11E-20
response to temperature stimulus	over	-	5.55E-06
response to salt stress	over	-	1.95E-04
acetyl-CoA catabolic process	over	-	2.46E-03
C-acyltransferase activity	over	-	3.58E-03
regulation of growth rate	under	6.37E-13	1.15E-09
regulation of growth	under	2.89E-09	1.92E-08
transcription factor activity	under	1.33E-10	6.07E-04
reproduction	under	1.64E-18	2.00E-19
meiotic cell cycle	under	1.27E-03	3.04E-03
multicellular organismal development	under	1.17E-27	1.18E-37
cell wall	under	6.24E-07	3.30E-12
vitamin binding	under	2.52E-13	2.24E-07
transmembrane transport	under	1.96E-32	3.09E-25
phagocytosis	under	6.78E-06	2.14E-05
cytokinesis	under	1.82E-03	5.65E-04
response to water deprivation	under	-	8.40E-04

[a] Expansions and reduction of GO-terms. Green indicates over-representation in comparison with both *P. tricornutum* and *C. reinhardtii*, while red indicates under-representation. Lighter green and red indicates over-representation and under-representation in comparison with *C. reinhardtii* alone.

[b] Gene ontology term description.

[c] Signifies whether the GO-term is over- or under-represented in comparison with *P. tricornutum* and *C. reinhardtii.*

[d] Probability for over/under-representation in comparison with *C. reinhardtii* and *P. tricornutum* calculated by Fisher's exact test.

Supplementary Table S8. Carbon assimilation genes	
Gene model[a]	**Targeting**[b]
Carbonic anhydrase	
Nga01240	Cytosolic/Chloroplast?[c]
Nga01717	Mitochondria
Nga03728	Cytosolic
Nga30848	Cytosolic
Nga10007	Chloroplast
Nga21222	Extracellular
Bicarbonate transporter	
Nga00165.01	Plasma membrane
Nga06584	Chloroplast

[a] Gene models involved in carbon assimilation.

[b] Signal peptide targeting predicted by TargetP and HECTAR.

[c] No clear signal peptide was determined for Nga01240 but it has high homology with a chloroplast targeted β-carbonic anhydrase in both *P. tricornutum* and *E. siliculosus*.

Supplementary Table S9. Genes important for carbon fixation			
Gene	**Number**	**EC number**	**Targeting[a]**
Calvin cycle			
Rubisco[b]		4.1.1.39	Nt
phosphoglycerate kinase	3	2.7.2.3	Nt/M
glyceraldehyde-3-phosphate dehydrogenase	4	1.2.1.13	Nt/C
fructose-bisphosphate aldolase	4	4.1.2.13	Nt/C
fructose-1,6-bisphosphatase	5	3.1.3.11	Nt/C
transketolase	2	2.2.1.1	C
seduheptulose-bisphosphatase	3	3.1.3.37	Nt
ribose 5-phosphate isomerase	1	5.3.1.6	Nt
phosphoribulokinase	1	2.7.1.19	C
triose-phosphate isomerase	3	5.3.1.1	Nt/C/M
ribulose-phosphate 3-epimerase	2	5.1.3.1	Nt
C4 cycle			
malate dehydrogenase (NAD/NADP)	2	1.1.1.37/82	Nt/M
malic enzyme (NAD/NADP)	2	1.1.1.39/40	Nt/C
aspartate aminotransferase	4	2.6.1.1	Nt/M
phosphoenolpyruvate carboxykinase	1	4.1.1.49	M
phosphoenolpyruvate carboxylase	1	4.1.1.31	Nt
malate dehydrogenase (oxaloacetate-decarboxylating) (NADP+)	1	1.1.1.40	Nt
pyruvate, phosphate dikinase	3	2.7.9.1	Nt
alanine transaminase	2	2.6.1.2	Nt/M
fructokinase	1	2.7.1.4	Nt

[a] TargetP prediction of targeting. Nt, no target peptide identified; M, mitochondria; C, chloroplast.

[b] Rubisco is chloroplast encoded, all the other genes are nuclear encoded.

Supplementary Table S10. Details of the sequencing data for the *N. gaditana* genome using Roche 454 and Illumina sequencing

Format	Number of raw reads	Number of reads after trimming	Total nts after trimming	Mean trimmed read length	Genome coverage after trimming
Roche unpaired	1,123,414	1,103,775	338,614,903	307	9
Illumina unpaired	24,709,613	24,484,194	2,313,756,333	95	63
Illumina LIPES	57,978,246	56,644,602	5,845,722,926	103	158
TOTAL	83,811,273	82,232,571	8,498,094,162	N/A	230

Supplementary Table S11. Over-representation of amino acid metabolic GO-terms[a]			
GO term[b]	Over- or Under-represented[c]	p-value[d] *P. tricornutum*	p-value[d] *C. reinhardtii*
serine family amino acid metabolic process	over	3.39E-23	3.37E-28
isoleucine metabolic process	over	2.81E-12	3.10E-16
cellular amino acid metabolic process	over	2.10E-11	8.37E-22
valine metabolic process	over	6.23E-10	3.86E-15
cellular amino acid and derivative metabolic process	over	3.47E-09	2.03E-17
tyrosine metabolic process	over	7.89E-09	4.25E-10
leucine metabolic process	over	5.65E-08	5.39E-12
isoleucine catabolic process	over	8.54E-08	1.19E-10
aspartate family amino acid metabolic process	over	4.62E-07	1.57E-10
valine catabolic process	over	5.92E-07	1.74E-09
valine biosynthetic process	over	7.25E-07	3.62E-08
leucine catabolic process	over	2.86E-06	1.33E-08
tryptophan metabolic process	over	2.86E-06	1.63E-08
tyrosine biosynthetic process	over	7.36E-06	2.42E-06
isoleucine biosynthetic process	over	7.36E-06	4.06E-07
lysine catabolic process	over	1.23E-05	1.12E-06
aromatic amino acid family metabolic process	over	2.90E-05	3.46E-07
lysine metabolic process	over	7.78E-05	1.99E-05
L-phenylalanine metabolic process	over	1.36E-04	1.43E-06
cellular amino acid derivative metabolic process	over	3.08E-04	1.43E-03
aspartate metabolic process	over	4.60E-04	5.07E-05
branched chain family amino acid metabolic process	over	5.22E-04	2.01E-08
tryptophan biosynthetic process	over	5.43E-04	7.31E-04
leucine biosynthetic process	over	5.43E-04	3.41E-05
branched chain family amino acid catabolic process	over	6.73E-04	2.98E-07
L-serine metabolic process	over	6.95E-04	8.19E-05
threonine metabolic process	over	8.29E-04	7.83E-07
glycine metabolic process	over	9.27E-04	5.07E-05
aspartate family amino acid catabolic process	over	2.13E-03	3.51E-06
alanine metabolic process	over	2.13E-03	1.86E-05
aromatic amino acid family biosynthetic process, prephenate pathway	over	1.73E-05	8.66E-07
L-phenylalanine biosynthetic process	over	7.36E-06	4.06E-07
pyruvate family amino acid metabolic process	over	2.13E-03	1.86E-05
amino acid binding	under	2.65E-03	1.14E-03
cysteine-type endopeptidase activity	under	1.27E-03	3.59E-05
serine-type peptidase activity	under	1.79E-04	1.32E-03
serine hydrolase activity	under	1.79E-04	1.32E-03
serine-type endopeptidase activity	under	8.53E-04	5.59E-04

[a] Expansions and reduction of GO-terms relating to amino acid metabolism. Green indicates overrepresentation in comparison with both *P. tricornutum* and *C. reinhardtii*, while red indicates under-representation.

[b] Gene ontology term description.

[c] Signifies whether the GO-term is over- or under-represented in comparison with *P. tricornutum* and *C. reinhardtii.*

[d] Probability for over/under-representation in comparison with *C. reinhardtii* and *P. tricornutum* calculated by Fisher exact test.

Supplementary Table S12. Genes involved in nitrogen assimilation up to ammonium[a]	
Function[b]	**Gene model[c]**
Assimilation	
NAD(P)H nitrate reductase	Nga02268
Ferredoxin nitrite reductase	Nga02267
NAD(P)H nitrite reductase	No homolog[d]
Urease	Nga03713
Urease accessory protein ureF	Nga20646.1
Urease accessory protein ureG	Nga00207.01
Urease accessory protein ureD	Nga01342.01
Transport	
Nitrate high affinity transporter	Nga30904
Nitrate high affinity transporter	Nga10006
Putative nitrate / peptide transporter	Nga04130
Nitrite transporter	Nga00897
Nitrite transporter	Nga06677
Ammonium transporter	Nga20972
Ammonium transporter	Nga30207
Urea/Na+ high affinity symporter	Nga06186
Urea Cycle	
carbamoyl-phosphate synthetase I large subunit	Nga00381
carbamoyl-phosphate synthetase I small subunit	Nga03303
ornithine carbamoyltransferase	Nga20220
argininosuccinate synthetase	Nga04487
argininosuccinate lyase	Nga03647
arginase	Nga00526

[a] Genes involved in nitrogen assimilation in *N. gaditana*. Proteins from *P. tricornutum* and *E. siliculosus* were used as BLAST queries.

[b] Functional role in nitrogen assimilation.

[c] Candidate *N. gaditana* gene model encoding corresponding enzyme.

[d] No NAD(P)H nitrite reductase were found in *N. gaditana* gene models or in Assembly v1.1 (tBLASTn, E-value less than 1E-03).

Supplementary Table S13. Meiosis Genes		
Gene	***N. gaditana* gene model[a]**	**Transcript Support[b]**
RAD51	Nga03590	yes
RAD51A	Nga02535.01	no
DMC1	Nga00201.01	no
recA	Nga04950	no
Msh4	Nga10008	no
Msh5	Nga01439.01	no
Mer3	Nga03449	no
Hop1	no homolog[c]	-
Hop2	no homolog	-
mnd1	Nga01637.01	yes
Spo11	Nga02765	no
Spo11/Top6A	Nga06442	yes
Top6B	Nga05494.01	no
ERCC4	Nga10010.1	no
Mus81	no homolog	-
Mms4	no homolog	-

[a] Candidate *N. gaditana* gene model encoding corresponding enzyme.

[b] Indicates if the gene model has transcript support from the assembled transcripts from the following pooled conditions, +/- nitrate, logarithmic phase, stationary phase, heat shocked culture (2h at 37ºC), cold treated culture (2h at 4ºC), 12h dark acclimation, and +/- supplemental CO2.

[c] Weak partial hit.

Supplementary Table S14. Genome data sources				
Organism	**Number of proteins**	**Version / Date accessed**	**Source institution**	**Accessed from**
Nannochloropsis gaditana	9,052	v1 (this study)	Colorado School of Mines	link
Photosynthetic stramenopiles				
Ectocarpus siliculosus	12,364	8/16/2011	Genoscope	link
Phaeodactylum tricornutum	10,108	7/12/2011	JGI	link
Thalassiosira pseudonana	11,625	7/12/2011	JGI	link
Fragilariopsis cylindrus	27,137	v1 (filtered models 1)	JGI	link
Aureococcus anophagefferens	11,510	8/23/2011	JGI	link
Non-photosynthetic stramenopiles				
Phytophthora infestans T304	17,602	9/3/2011	BROAD Inst.	link
Phytophthora ramorum	15,743	v1.1	JGI	link
Phytophthora sojae	19,027	v1.1	JGI	link
Albugo laibachii Nc14	12,983	9/3/2011	The Sainsbury Laboratory	link
Blastocystis hominis	5,831	7/12/2011	Genoscope	link
Hyaloperonospora arabidopsidis	14,543	V8.3.2	VBI	link
Pythium ultimum BR144	15,290	9/29/2011	MSU	link
Chromalveolates				
Plasmodium falciparum	5,349	9/3/2011	NCBI	link
Tetrahymena thermophila	24,725	tta1_oct2008_finalrelease	TIGR	link
Green/Red algae				
Chlamydomonas reinhardtii	15,075	7/12/2011	JGI	link
Ostreococcus tauri	7,969	8/5/2011	Genopole LR + Ghent Univ.	link
Chlorella sp. NC64A	9,791	v1 (filtered models)	JGI	link
Cyanidioschyzon merolae	4,950	v1	*C. merolae* consortium	link

SUPPLEMENTARY NOTES

Supplementary Note 1. Plastid and mitochondrial genomes.

The circular chloroplast genome is 114,785 bp, which is similar in size to the chloroplast genomes of *P. tricornutum*, *T. pseudonana*, and *E. siliculosus*[32,61], and contains 124 protein-encoding genes as well as those for 5S, 16S and 23S rRNAs, and 27 tRNAs, which satisfy all translational requirements. Due to the close phylogenetic relationship between *N. gaditana* and diatoms and brown algae we compared the plastid and mitochondrial genomes with *P. tricornutum*, *T.* pseudonana, and *E. siliculosus* (**Supplementary Table S1**). The gene order is not strictly conserved but there are a number of conserved gene clusters, such as that for ribosomal genes. Interestingly, the *N. gaditana* chloroplast genome contains all the subunits of the light-independent protochlorophyllide reductase (chlB, chlL, and chlN) which is needed for chlorophyll synthesis. Neither chloroplast genomes of *P. tricornutum* or *T. pseudonana* encode these subunits, while the chloroplast genome of the evolutionarily more distant green alga *C. reinhardtii* does[62].

The circular mitochondrial genome is 42,067 bp, which is similar in size to the mitochondrial genomes of *T. pseudonana* and *P. tricornutum*, and contains 36 protein-encoding genes as well as those for 16S and 23S rRNAs and 27 tRNAs, which satisfy all translational requirements except for tRNA-Thr, which is also missing in other heterokonts[63]. Interestingly, the *N. gaditana* mitochondrial genome has two complete copies of the *cox1* gene lacking introns in two duplicated regions. This is a different configuration than what is found in *P. tricornutum* or *T. pseudonana* where the *cox1* gene is instead split into exons and introns. This change in configuration may suggest that the duplicated *cox1* gene in *N. gaditana* is an older configuration which has been modified in *P. tricornutum* and *T. pseudonana*. Heterokontophyte brown algae have only a single intron-less copy of *cox1*[64].

Supplementary Note 2. Clusters of genes with similar functions.

Some stramenopiles are known to have clusters of genes with similar functions[36]. To detect if *N. gaditana* has genes arranged in this fashion we linked GO-terms to gene model locations on assembly version 1.1. GO-terms were then sorted by location and those terms that were represented by three or more gene models were investigated. Clusters of genes with similar functions were determined and a selection of these can be found in **Supplementary Table S3**. These clusters may include genes of a similar biological process of independent evolutionary origin or genes that have been duplicated at a given locus. Some of the functions found within these gene clusters include acyl-carrier protein biosynthesis, cellular nitrogen compound metabolic processes, gene expression, branched chain amino acid metabolism, mRNA processing, and transport. **Supplementary Fig. S4** details gene clusters involved in hydrogen metabolism and nitrogen assimilation. The hydrogen gene cluster is made up of the [FeFe]-hydrogenase (*hydA*) and three genes needed for [FeFe]-hydrogenase maturation (*hydE*, *hydF*, and *hydG*). This cluster appears to have two bi-directional promoters found between *hydE* and *hydG* and also between *hydA* and *hydF*. This configuration of hydrogenase genes has not been observed in other algae sequenced to date[65]. The nitrogen assimilation cluster includes genes for the transport of nitrate, nitrate reduction to nitrite, and finally nitrite reduction.

Supplementary Note 3. Expansion of gene families.

For further analysis of the expansion of gene families/enrichment of gene ontology terms (GO-terms) in *N. gaditana* we compared the prevalence of GO-terms with *P. tricornutum* and *C. reinhardtii*. Gene Ontology terms were assigned with Blast2GO[66] and the complete gene ontologies for *P. tricornutum* and *C. reinhardtii* were obtained from the B2G-FAR database[67]. Fisher's exact test was used to analyze the significance of the expansions/reductions[68] through the use of the Gossip algorithm[69].

A selected list of over- and under-represented GO-terms with a maximum P-Value of $4x10^{-03}$ (determined by Fisher's exact test) and maximum false discovery rate of $5x10^{-02}$ are shown in **Supplementary Table S7**. Several expanded gene families that may be of importance for the exemplary biomass and lipid production characteristics of *N. gaditana* include those for acyl-carrier protein biosynthetic processes, auxin biosynthetic processes related to the production of plant growth hormones, carbon utilization, response to stress (including chemical, temperature and salt), and pyruvate metabolic processes. A large number of GO-terms associated with amino acid metabolism are also overrepresented (**Supplementary Table S11**). In addition, GO-terms for chlorinated/halogenated hydrocarbon metabolic processes are also expanded, which is also observed in *E. siliculosus*[36] where it is thought that these genes may protect against halogenated compounds produced by kelps as defense molecules, allowing epiphytic growth on these organisms[36]. *Nannochloropsis gaditana* growth does not rely on exogenous sources of vitamins, which is reflected in the under-representation of genes involved in vitamin acquisition.

Supplementary Note 4. Nitrogen assimilation genes.

Nitrogen can be assimilated by *N. gaditana* in several ways. We have determined that growth can be sustained on nitrate, nitrite[70], ammonium, or urea as sole nitrogen sources. The genome encodes nitrate, nitrite, ammonium, and urea transporters (**Supplementary Table S12**). In order to utilize nitrate and nitrite they need to be reduced to ammonium. The *N. gaditana* genome encodes one NAD(P)H-nitrate reductase to reduce nitrate to nitrite. *Ectocarpus siliculosus*, *P. tricornutum*, and *T. pseudonana* contain two types of enzymes that use different reducing substrates to catalyze the reduction of nitrite, ferredoxin dependant as found in plants and NAD(P)H dependant as found in bacteria. Unlike these other stramenopiles, NAD(P)H-nitrite reductase cannot be found in our *N. gaditana* gene models or in Assembly 1.1 (tBLASTn, E-value less than 1E-03). This suggests that *N. gaditana* can only reduce nitrite when reduced ferredoxin is available via the ferredoxin-nitrite reductase (Nga02267), and this activity is diminished during the night or under low-light when reduced ferredoxin levels are attenuated. This is supported by experiments in *Nannochloropsis oculata* that show in the dark or in the presence of DCMU nitrite uptake is negligible, whereas nitrate uptake is substantial. The majority (82%) of this nitrate reappeared in the medium as nitrite[70], indicating the possible use of nitrate as a terminal electron acceptor under these conditions. Genome analysis indicates that nitrate can be reduced in the absence of photosynthesis via NAD(P)H dependant nitrate reductase but nitrite is not readily reduced in the absence of the high levels of reduced ferredoxin generated during photosynthesis.

Urea can also be utilized by *N. gaditana* cells via a urea/Na^+ symporter and can be converted into ammonium with a urease and urease accessory proteins (ureD, ureF, and ureG). Additionally, the *N. gaditana* genome encodes the complete urea cycle including carbamoyl-phosphate synthetase I (large and small subunits), ornithine carbamoyltransferase, argininosuccinate synthase, argininosuccinate lyase, and arginase.

Genes responsible for nitrate assimilation and reduction of nitrate to nitrite to ammonium are found in a gene cluster, which includes NAD(P)H-nitrite reductase, the ferredoxin-nitrite reductase, and a

nitrate high affinity transporter (**Supplementary Fig. S4**). Nitrogen assimilation genes have been found in clusters in *C. reinhardtii*[34,35], but not in *E. siliculosus*, where no such nitrogen assimilation clusters exist, not even as gene pairs[36].

The lack of functional redundancy in several key nitrogen assimilation genes can be exploited for biotechnological purposes. In *C. reinhardtii* nitrate reductase has been used as a selectable marker in nitrate reductase (*nit-1*) deficient strains[71], and a similar approach could be used in *N. gaditana*. Knockouts in NAD(P)H-nitrate reductase (Nga02268) or ferredoxin nitrite reductase (Nga02267) if isolated, could serve as selectable markers.

Supplementary Note 5. Meiosis genes.

Ectocarpus siliculosus and some species of diatoms are known to have sexual cycles[36], so we queried genes known to be associated with meiosis to determine if they were contained within the *N. gaditana* genome. Seven members of the RAD51 family are known to be found in eukaryotes and are important for DNA repair, homologous recombination, and genome stability[72]. Of these, several are implicated in playing a role in meiosis including, DMC1 which is specifically required for meiotic recombination in plants, yeast, and animals and RAD51, RAD51C, and XRCC3 which are essential for meiosis in *Arabidopsis*[73]. The *N. gaditana* genome contains four RAD51 homologs, including one for DMC1 and RAD51 (**Supplementary Table S13**).

Two genes, Hop2 and Mnd1, form a heterodimeric complex that stimulate DMC1 activity and are required for meiotic recombination in *Saccharomyces cerevisiae*[74]. A Mnd1 homolog is found in *N. gaditana*, but no clear Hop2 homologs can be found. Two other genes involved in meiotic recombination process (crossovers) that are found in *N. gaditana* are Msh4 and Msh5. These are meiosis specific proteins and are required for proper segregation of homologues chromosomes during meiosis. Msh4 and Msh5 form a heterodimer and are thought to stabilize chromosomes during meiosis double strand break repair[75]. A DNA helicase required for interference-sensitive meiotic crossover events in *Arabidopsis*, Mer3[76], also has a homolog in *N. gaditana.*

Interestingly many of the meiotic genes found in *N. gaditana* have no evidence of being transcribed under any of the physiological conditions we have assayed, which include +/- nitrate, logarithmic phase, stationary phase, heat shocked culture (2h at 37ºC), cold treated culture (2h at 4ºC), 12h dark acclimation, and +/- supplemental CO_2. More than 98% of the 9,052 gene models have transcript support from these conditions. This lack of transcript support of meiosis-specific genes may suggest that the culture condition that induces the *N. gaditana* sexual cycle was not examined or that these genes are no longer transcribed because *N. gaditana* no longer undergoes meiotic recombination. Further investigation will be needed to clarify whether meiotic recombination is possible in *N. gaditana*.

SUPPLEMENTARY METHODS

Assembly of organellar genomes.

The merged assembly was searched for sequence similarity against chloroplast (Genbank: NC_008589 and NC_008588) and mitochondrial (Genbank: NC_007405 and HQ840789) proteins from *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* using tBLASTn. The organellar genomes were identified on 5 scaffolds. The circular mitochondrial genome was identified on two scaffolds: one that contained the complete mitochondrial genome and the second that contained a partial sequence of the mitochondrial genome that bridged the 5' and 3' ends of the complete scaffold (**Supplementary Fig. S3**). The circular chloroplast genome was identified on three scaffolds which contained the large single copy (LSC), short single copy (SSC), and inverted repeat (IR) regions, respectively. These three scaffolds were not automatically assembled due to the inverted repeat (IR) regions commonly present in plastid genomes[32] and were manually assembled into one complete chloroplast genome (**Supplementary Fig. S2**). The scaffold that contained the IR was found to have approximately twice the sequence coverage as the LSC and SSC containing scaffolds, verifying its duplication. Genes were identified by sequence similarity of open reading frames to known plastid and mitochondrial proteins while transfer RNAs were determined by tRNAscan-SE 1.21[77]. Both mitochondrial and chloroplast genomes were visualized with OGDRAW[78].

Assembly of the transcriptome.

The trimmed Illumina SIPES sequencing reads from the pooled RNA samples were assembled using the commercial package from CLC Bio. This generated a total of 37,055 contigs, which is surely an over-count of the actual number of genes. These were formed by the matching of 16,094,133 reads (91% of the total), 13,921,494 of which matched as pairs (defined as facing one another and separated by 50 to 500 nts). The total length of these contigs is 17,024,428 nts with a G+C content of 56%. These contigs have a fairly tight distribution of paired-end separation distance, as shown in **Supplementary Fig. S10**.

Creating gene models.

Genes were modeled in numerous ways after masking repeated genomic elements: (1) Exonerate[79] gene models using the *N. gaditana* transcriptome assembly compared to the *N. gaditana* genome scaffolds. (2) EST evidence using this same aligned data by BLASTn. (3) EST evidence using 66,106 *Ectocarpus siliculosus* EST sequences[36] aligned to the *N. gaditana* genome scaffolds with reduced stringency of matching using tBLASTx. (4) Gene models created by aligning the RNAseq data to the *N. gaditana* genome scaffolds using Bowtie[80] followed by gene modeling with Tophat and Cufflinks[81]. (5) Protein homology evidence by aligning all proteins of 10 genomes to the *N. gaditana* genome scaffolds using BLASTx. These were the gene sets of *Chlamydomonas reinhardtii*[13] because of its high quality gene annotations and nine stramenopiles (like *N. gaditana*), the diatoms *Phaeodactylum tricornutum*[14] and *Thalassiosira pseudonana*[40], the brown algae *Ectocarpus siliculosus*[36] and *Aureococcus anophagefferens*[37], and the oomycetes *Hyaloperonospora arabidopsidis*[82], *Phytophthora infestans*[83], *Phytophthora ramorum*[60], *Phytophthora sojae*[60], and *Pythium ultimum*[84]. (6) Exonerate gene models using all proteins from these same 10 organisms compared to the *N. gaditana* genome scaffolds. (7) *Ab initio* gene models created by the program Augustus[85], which had been trained on the gene structures of *Neurospora crassa*[86], which had been demonstrated to be the best training data for this genome in earlier tests. (8) *Ab initio* gene models created by the program SNAP[87], which had been trained on Hidden Markov Models (HMMs) of the genes of *Pythium ultimum*[84]. (9) *Ab initio* gene models created by the

program GeneMark[88], which had been trained on HMMs of the genes of *Pythium ultimum*. All of these lines of evidence were reconciled into a single gene set using Maker[33].

Recognizing that some genes may not be modeled in this way because of limitations of the assembly or of our gene modeling process, we also examined the transcriptome assembly for any additional genes that may be present. We searched all transcript contigs (see above) for homology to the gene sets of *P. tricornutum*, *T. pseudonana*, *C. reinhardtii*, *E. siliculosus*, and *A. anophagefferens* using BLASTx with a cutoff of 1E-10 (**Supplementary Table S14**). The subset so identified was then searched against the complete gene set created as described above using BLASTn with a cuttoff of 1E-20, narrowing this to 2,531 transcript assembly contigs that have homology support for being real genes but were not present in our Maker gene set.

We masked all of the regions with Maker-predicted gene models and then ran the program EST2genome[89] to align these 2,531 transcript assemblies to the genome assembly v1.1. Of these, 1,310 could be assigned to the genome and 1,221 could not. Presumably EST2genome could not match these 1,221 genes because they are located in regions that are misassembled, overlaps with a Maker-predicted gene model (and so were masked to avoid artifactual duplication), span two or more scaffolds, or are in regions missing from the assembly altogether. We checked this by running BLASTn of these 1,221 transcript assemblies against the genome assembly and found matches for 784 of these 1,221 transcript contigs. This supports the view that these genes largely do appear in the genome assembly, and only 437 do not. Gene models that are not found in the genome assembly account for only 4.8% of the total gene models.

All gene models were subsequently assigned a unique, short gene designation ("Nga" followed by a unique number) which is followed by the more complete gene name as assigned by Maker or by EST2genome (**Supplementary Table S2**). Of these, an all-by-all BLASTn search identified 619 genes that appear as duplications with identical sequence and 670 with large regions of identical sequence, but in each case being at different loci. These may be actual biological duplications or might be errors in the assembly. Those that are completely identical are distinguished by suffices of '.01' and '.02' and those with only large regions of identity by suffices of '.1' and '.2' (with the longer sequence called .1). This 9,052-member gene set is designated version 1.1 genes models.

Gene expression analysis.

RNA was isolated from cultures in logarithmic growth during nitrogen replete conditions (+N) and from parallel cultures that had been washed with nitrogen deprived medium twice and transferred into nitrogen deprived medium for 16 hours before RNA isolation (–N). RNA was extracted as described previously[19], selected for polyA+ transcripts and converted into cDNA, and then sequenced using the Illumina SIPES protocol.

Sequencing reads were trimmed as described previously, leaving 19,564,728 paired reads of the +N and 18,061,406 paired reads of the –N conditions, with an average read length of 50.87 nts. A count was made of each of these RNAseq reads aligned to each of the gene sequences of gene set version 1.1. This was then normalized by the length of each transcript to get the mean depth of coverage and then normalized for the differing numbers of input reads. Ratios were calculated for each gene to assay the differences in gene expression under these two conditions. There were matches of this RNAseq data to 8,948 of the total of 9,052 gene models; presumably, the 104 genes that were not matched were not expressed in these conditions. There were 48 additional genes that were matched by reads from the cDNA of the +N condition but not of the –N condition and 27 other genes that were matched by reads from the

cDNA of the –N condition but not of the +N condition. Other than these absolute differences in expression, the ratios were found to be as high as 23-fold.

Functional annotation of gene models.

Conserved protein domains were identified within *N. gaditana* gene models by CD-Search, an algorithm which uses RPS-BLAST to compare gene models to NCBI-curated domains, Pfam, SMART, COG, PRK, TIGRFAM databases[90]. Putative subcellular localizations were predicted using TargetP[52] and HECTAR, a statistical prediction method that is specific for heterokonts[53]. Gene Ontology terms (GO-terms) were assigned with Blast2GO[66]. GO-term enrichment analysis was conducted with the Gossip algorithm[69] to perform Fisher exact test against *P. tricornutum* and *C. reinhardtii* GO-terms retrieved from the B2G-FAR database[67]. KEGG orthology (KO) identifiers were assigned using the KEGG Automatic Annotation Server (KAAS) with default settings[91] and KOG numbers were assigned by the mutual best BLASTp hit with an E-value of less than 1E-10 of each gene model to the KOG Database[92]. KO identifiers and KOG numbers were used to populate metabolic maps, which were visualized with iPath2.0: interactive pathway explorer[93].

Gene clusters were identified by linking GO-terms to gene location on assembly version 1.1. GO-terms were sorted and those terms that were represented by three or more gene models were manually investigated. Select clusters are displayed in **Supplementary Fig. S4** and **Supplementary Table S3**.

Comparative algal genomics.

The 9,052 gene models encoded by *N. gaditana* were compared to the following algal groups 1) 'brown' algae: *E. siliculosus* and *A. anophagefferens*; 2) diatoms: *P. tricornutum* and *T. pseudonana*; 3) red algae: *Cyanidioschyzon merolae*; and 4) green algae: *C. reinhardtii*, *Chlorella variabilis* NC64A, and *Ostreococcus tauri*. For this comparison *A. anophagefferens* was included as a 'brown' algae for simplicity of analysis due to their close phylogenetic relatedness. *Nannochloropsis gaditana* gene models were used as a BLASTp query against the aforementioned algae and hits with an E-value of less than 1E-10 were designated as homologous proteins. Gene models with no BLASTp hits at this cutoff (2,733) were designated as 'unique'. If a *N. gaditana* protein was found to be homologous by at least one member of an algal group, it was designated as being found within that group. This data was compiled for all *N. gaditana* gene models and visualized in a Venn diagram (**Fig. 2c**). *Nannochloropsis gaditana* gene models were also compared to all previously sequenced genomes in the non-redundant (nr) protein database using BLASTp. The organism from the top BLASTp hit (E-value less than 1E-3) of each gene model were aggregated and displayed in **Fig. 2d**. Gene models with no BLASTp hits (E-value less than 1E-3) numbered 2,061.

Generation of the StramenopilePhotoCut.

To determine conserved protein families important to photosynthetic stramenopiles we followed a phylogenomic analysis similar to Merchant *et al.* and Karpowicz *et al.* who established the GreenCut and GreenCut2, respectively, to inventory orthologous sets of proteins common to the green algal and plant lineage[13,41]. All *N. gaditana* protein sequences were used as a BLASTp query against the complete proteomes of photosynthetic stramenopiles, *A. anophagefferens*, *E. siliculosus*, *P. tricornutum* and *T. pseudonana,* and non-photosynthetic stramenopiles, *Albugo laibachii* Nc14, *Blastocystis hominis*, *P. sojae*, *P. ramorum*, and *P. infestans* T304 (**Supplementary Table S14**). To establish orthology to a *N. gaditana* protein, an organism must have a BLASTp hit with an E-value less than 1E-10 to that protein.

Nannochloropsis gaditana proteins that have orthologs in all photosynthetic stramenopiles and no orthologs in any non-photosynthetic stramenopiles were designated the "StramenopilePhotoCut". In total, 363 proteins were included in the "StramenopilePhotoCut". These proteins were also compared against proteins found in the GreenCut2[41] and orthologs were sought (BLASTp hit with E-value less than 1E-10) in the green algae: *C. reinhardtii*, *C. variabilis* NC64A, and *O. tauri*; and red algae: *C. merolae*. Only 39 "StramenopilePhotoCut" proteins were found exclusively in stramenopiles (brown algae and diatoms) and not in the green or red lineage, while 115 are encompassed by the GreenCut2 (**Fig. 3a** and **Supplementary Data 1**).

False Positive Rate. The false positive rate was estimated by comparing the "StramenopilePhotoCut" proteins to the proteome of the diatom *Fragilariopsis cylindrus*, a recently sequenced photosynthetic stramenopile, which ideally should contain all 363 "StramenopilePhotoCut" proteins. A BLASTp hit (less than 1E-10) of a *F. cylindrus* protein to a *N. gaditana* "StramenopilePhotoCut" protein was used to establish orthology. When a "StramenopilePhotoCut" protein was not found in the *F. cylindrus* v1.0 protein set, a tBLASTn search (E-value less than 1E-5)[41] was conducted against the *F. cylindrus* genomic assembly (JGI prepublication). In total, 344 out of 363 "StramenopilePhotoCut" proteins were found in *F. cylindrus*, giving an estimated false positive rate of 5.2%. Additionally, two non-photosynthetic stramenopiles, *Hyaloperonospora arabidopsidis*[82], and *Pythium ultimum*[84], were also used to find false positives, because they should contain no "StramenopilePhotoCut" proteins. *Hyaloperonospora arabidopsidis* has orthology to 3 proteins from the "StramenopilePhotoCut" list, while *P. ultimum* has orthology to 5 proteins, implying a combined (7 unique proteins) false discovery rate of 1.9%. Because these sets of false positives from photosynthetic and non-photosynthetic stramenopiles are mutually exclusive, in combination they suggest a false positive rate of up to ~7.2%.

False Negative Rate. The false negative rate was estimated by determining if a set of proteins known to be involved in photosynthesis were included in the "StramenopilePhotoCut". These proteins are PsaA, PsaB, PsaC, PsaD, PsaE, PsaF, PsaG, PsaL, PsbA, PsbB, PsbC, PsbD, PsbE, PsbH, PsbQ, PsbR, PsbU, PsbV, PsbW, Rubisco small subunit (rbcS), Rubisco large subunit (rbcL), ferredoxin, and ferredoxin-NADP reductase. From this list 20 out of 23 proteins were found in the "StramenopilePhotoCut", with PsaG, PsbR, and PsbW absent, suggesting a false negative rate of ~13%. Identification of orthologous proteins is problematic for short, moderately divergent sequences[41], so these proteins may have been missed because they have BLASTp expect values to *N. gaditana* greater than our cutoff of 1E-10.

Genetic transformation of *N. gaditana*.

Genomic DNA from an axenic culture of *N. gaditana* (CCMP526, Provasoli-Guillard National Center for Culture of Marine Phytoplankton) was purified as previously described using a phenol chloroform extraction protocol (Radakovits) and a full 454 sequencing run was used to generate a preliminary Newbler assembly of the *N. gaditana* genome. BLASTx was used to annotate genes by homology. The *N. gaditana* genome is similar to the genomes of *P. tricornutum* and *T. pseudonana* in that the genes only have 0-2 introns, this characteristic allowed us to identify many full length genes including their upstream promoter regions. From the identified full length genes that included an upstream promoter region we selected three for testing in transformation experiments. We cloned the 608 bp region upstream of the start codon of the heat shock protein 70 gene (HSP), the 520 bp promoter region of beta-tubulin (TUB)

and the 710 bp promoter region of the ubuquitin extension protein (UEP) from genomic DNA using the following primers:

HSP forward primer: ATGCTCCGGAGCC GAAGCCCTGTCG ACCAC
HSP reverse primer: AGCTTGGCCATGTTAGTCTGTCAAAAAATGACGTTGCG
TUB forward primer: ATGCTCCGGAACTGCGCATGGATTGACCGA
TUB reverse primer: AGCTTGGCCATGCTTCACAAAAAAGACAGCTTCTTGAT
UEP forward primer: ATGCTCCGGAGCTGCTGCCCCGACCGTATC
UEP reverse primer: AGCTTGGCCATCCT GCTGTATGATTTTGGCACAACG.

The cloned fragments were ligated in front of the bleomycin (ble) resistance gene in the pPha-T1 plasmid[16] replacing the *P. tricornutum* fcpB promoter to create the pPha-T1-HSP, pPha-T1-TUB and the pPhaT1-UEP plasmids.

Nannochloropsis gaditana was grown in f/2 50% seawater medium under cool white fluorescent lights at 100 µE (24 h illumination). After two weeks of growth $5x10^8$ cells were harvested for each transformation experiment. Cells were washed twice with 375 mM sorbitol before resuspension in 100 µl 375 mM sorbitol containing 5 µg plasmid DNA linearized with ScaI. Electroporation was done using a ECM630 BTX electroporator (Harvard Apparatus, Inc., Holliston, MA) set at 500 Ω, 50 µF and either 900, 1050 or 1200 V using a 1 mm cuvette, resulting in a single 17-20 ms pulse. After electroporation cells were resuspended in 10 ml f/2 medium and kept overnight on a shaker at room temperature in low light (50 µmol m^{-2} s^{-1}) before plating on f/2 + zeocin selection plates. $5x10^7$ cells were plated per 10 cm plate containing 3 µg/ml zeocin. Resistant colonies were detected after 5-6 weeks and picked after 7-8 weeks. No colonies grew on control plates with cells electroporated without plasmid and survival of cells plated without zeocin appeared unaffected even at the highest voltage. The highest number of colonies was generated using 1200 V (12000 V/cm field strength) and the promoter with the highest number of transformants was TUB followed by UEP (**Table 2**).

Confirmation of successful transformation by genomic PCR and Southern blot.

Picked colonies were grown in f/2 liquid media and 10^9 cells were harvested for verification of transgene incorporation into the genome of the zeocin resistant colonies. Genomic DNA was purified as described previously (Radakovits) and either used for genomic PCR (**Supplementary Fig. S7a**) or digested with the StuI and ClaI restriction enzymes over night for Southern blot analysis (**Supplementary Fig. S7b**). The resulting DNA fragments were separated on a 0.7% agarose gel before transfer onto a nitro cellulose membrane which was used for hybridization with a 371 bp DNA probe specific for the ble resistance gene. The ble probe was generated by PCR using the following primers: ble forward primer: CCGGGACTTCGTGGAGGACGAC; ble reverse primer: GCTGCTCGCCGATCTCGGTCAT. Probe synthesis and hybridization were performed using the AlkPhos Direct Labeling and Detection Systems as described previously[94], according to the manufacturer's instructions (Amersham Biosciences). The chemilumiscent signal was detected by a LAS-4010 imaging system (GE Healthcare Life Sciences), 20 h exposures gave the best results. The differences in the size of the bands indicate random insertion of the transgene while the presence of multiple bands in some mutants signifies multiple insertions.

Microscopy.

Lipid accumulation during logarithmic growth and stationary phase (**Supplementary Fig. S1a and b**) was visualized by fluorescence microscopy using the nonpolar lipid fluorophore BODIPY 493/503

(Molecular Probes, Invitrogen Corporation) as described previously[20]. Briefly, a 100x concentrated suspension of cells was stained with 10 µg/ml of BODIPY and the labeled cells were visualized on a Nikon Eclipse 80i microscope with UV fluorescence imaging capacity for BODIPY 493/503 and chlorophyll fluorescence.

SUPPLEMENTARY REFERENCES

61 Le Corguille, G. *et al.* Plastid genomes of two brown algae, *Ectocarpus siliculosus* and *Fucus vesiculosus*: further insights on the evolution of red-algal derived plastids. *BMC Evolutionary Biology* **9**, 253 (2009).

62 Li, J., Goldschmidt-Clermont, M. & Timko, M. P. Chloroplast-encoded chlB is required for light-independent protochlorophyllide reductase activity in *Chlamydomonas reinhardtii*. *The Plant Cell* **5**, 1817-1829 (1993).

63 Oudot-Le Secq, M.-P. & Green, B. R. Complex repeat structures and novel features in the mitochondrial genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. *Gene* **476**, 20-26 (2011).

64 Oudot-Le Secq, M.-P., Loiseaux-de Goër, S., Stam, W. & Olsen, J. Complete mitochondrial genomes of the three brown algae (Heterokonta: Phaeophyceae) *Dictyota dichotoma*, *Fucus vesiculosus*, *Desmarestia viridis*. *Current Genetics* **49**, 47-58 (2006).

65 Meuser, J. *et al.* Evolutionary significance of an algal gene encoding an [FeFe]-hydrogenase with F-domain homology and hydrogenase activity in *Chlorella variabilis* NC64A. *Planta* **234**, 829-843 (2011).

66 Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* **36**, 3420-3435 (2008).

67 Götz, S. *et al.* B2G-FAR, a species centered GO annotation repository. *Bioinformatics* **27**, 919-924. (2011).

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

69 Blüthgen, N. *et al.* Profiling of gene groups utilizing gene ontology – A statistical framework. *arXiv:q-bio* (2004).

70 Hipkin, C. R., Thomas, R. J. & Syrett, P. J. Effects of nitrogen deficiency on nitrate reductase, nitrate assimilation and photosynthesis in unicellular marine algae. *Marine Biology* **77**, 101-105 (1983).

71 Fernández, E. *et al.* Isolation and characterization of the nitrate reductase structural gene of *Chlamydomonas reinhardtii*. *Proceedings of the National Academy of Sciences* **86**, 6449-6453 (1989).

72 Lin, Z., Kong, H., Nei, M. & Ma, H. Origins and evolution of the recA/RAD51 gene family: Evidence for ancient gene duplication and endosymbiotic gene transfer. *Proceedings of the National Academy of Sciences* **103**, 10328-10333 (2006).

73 Li, W. & Ma, H. Double-stranded DNA breaks and gene functions in recombination and meiosis. *Cell Res* **16**, 402-412 (2006).

74 Chen, Y.-K. *et al.* Heterodimeric complexes of Hop2 and Mnd1 function with Dmc1 to promote meiotic homolog juxtaposition and strand assimilation. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 10572-10577 (2004).

75 Snowden, T., Acharya, S., Butz, C., Berardini, M. & Fishel, R. hMSH4-hMSH5 recognizes holliday junctions and forms a meiosis-specific sliding clamp that embraces homologous chromosomes. *Molecular Cell* **15**, 437-451 (2004).

76 Mercier, R. *et al.* Two meiotic crossover classes cohabit in *Arabidopsis*: One is dependent on MER3, whereas the other one is not. *Current Biology* **15**, 692-701 (2005).

77 Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research* **33**, 686-689 (2005).

78 Lohse, M., Drechsel, O. & Bock, R. OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. *Current Genetics* **52**, 267-274 (2007).
79 Slater, G. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
80 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**, R25 (2009).
81 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).
82 Baxter, L. *et al.* Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. *Science* **330**, 1549-1551 (2010).
83 Haas, B. J. *et al.* Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* **461**, 393-398 (2009).
84 Levesque, C. A. *et al.* Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. *Genome Biology* **11**, R73 (2010).
85 Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19**, 215-225 (2003).
86 Galagan, J. E. *et al.* The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**, 859-868 (2003).
87 Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
88 Lukashin, A. V. & Borodovsky, M. GeneMark.hmm: New solutions for gene finding. *Nucleic Acids Research* **26**, 1107-1115 (1998).
89 Mott, R. EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Computer applications in the biosciences* **13**, 477-478 (1997).
90 Marchler-Bauer, A. *et al.* CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research* **39**, 225-229 (2011).
91 Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. & Kanehisa, M. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research* **35**, 182-185 (2007).
92 Tatusov, R. *et al.* The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41 (2003).
93 Yamada, T., Letunic, I., Okuda, S., Kanehisa, M. & Bork, P. iPath2.0: interactive pathway explorer. *Nucleic Acids Research* **39**, 412-415 (2011).
94 Pootakham, W., Gonzalez-Ballester, D. & Grossman, A. R. Identification and regulation of plasma membrane sulfate transporters in *Chlamydomonas*. *Plant Physiology* **153**, 1653-1668 (2010).