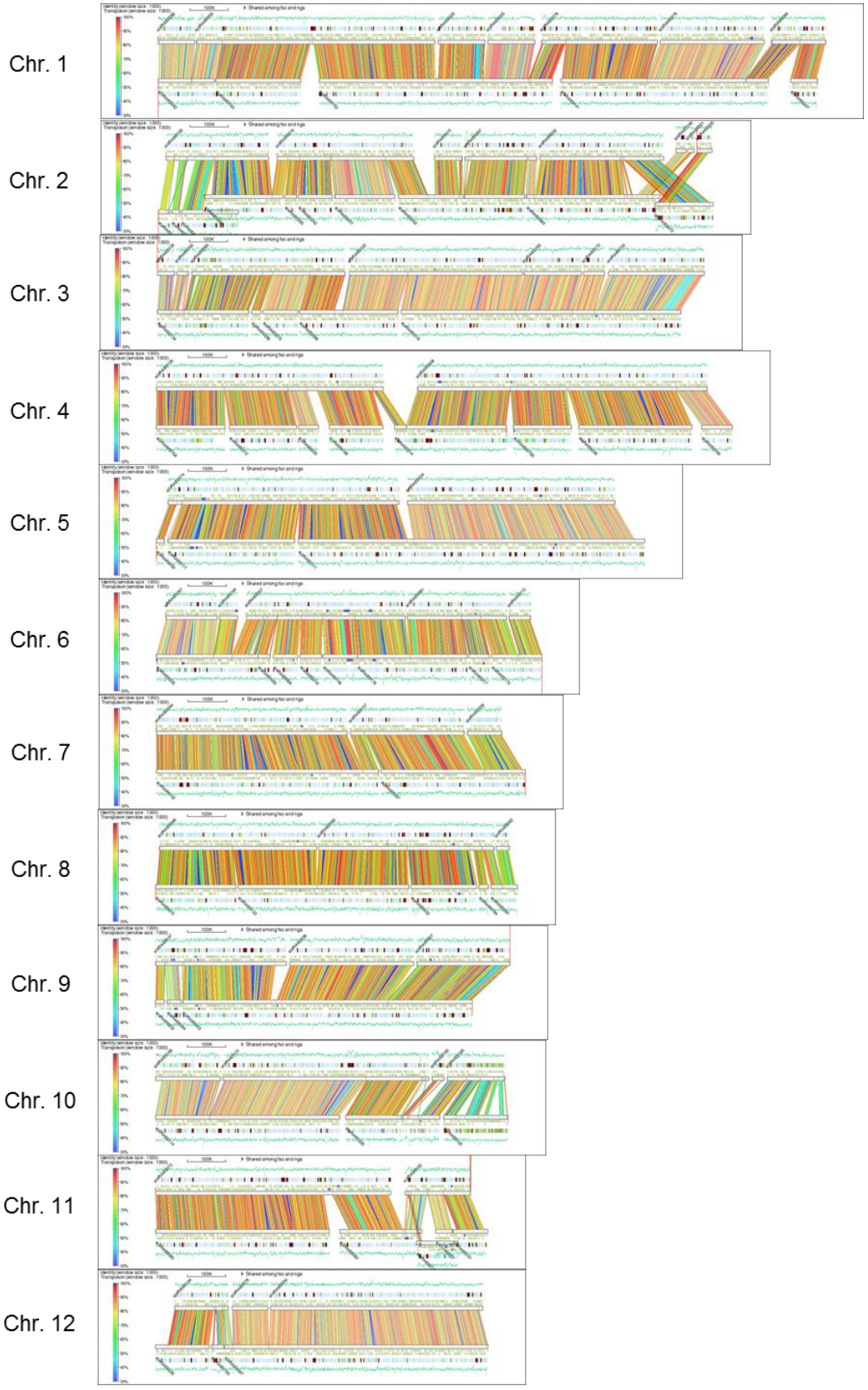
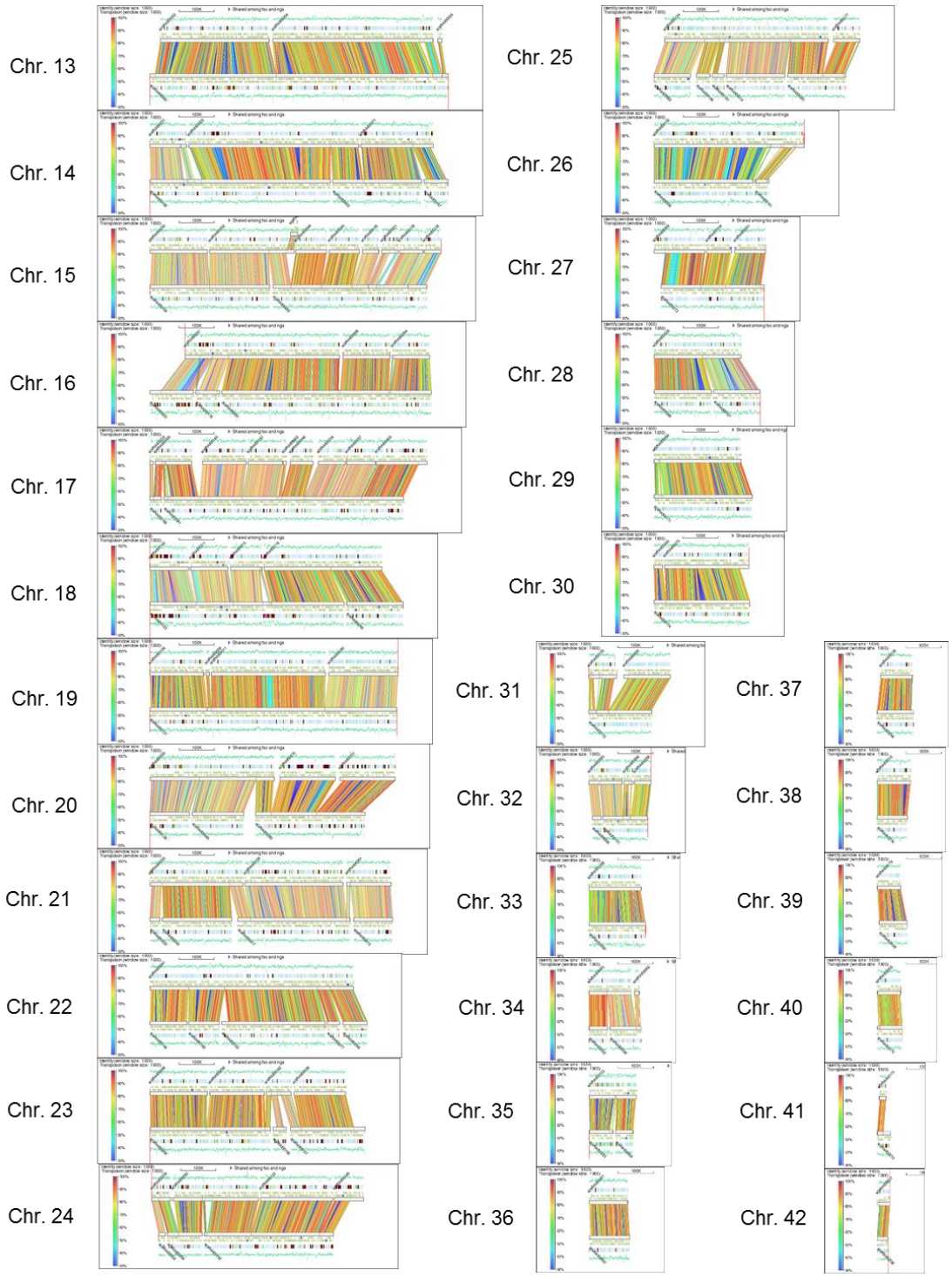
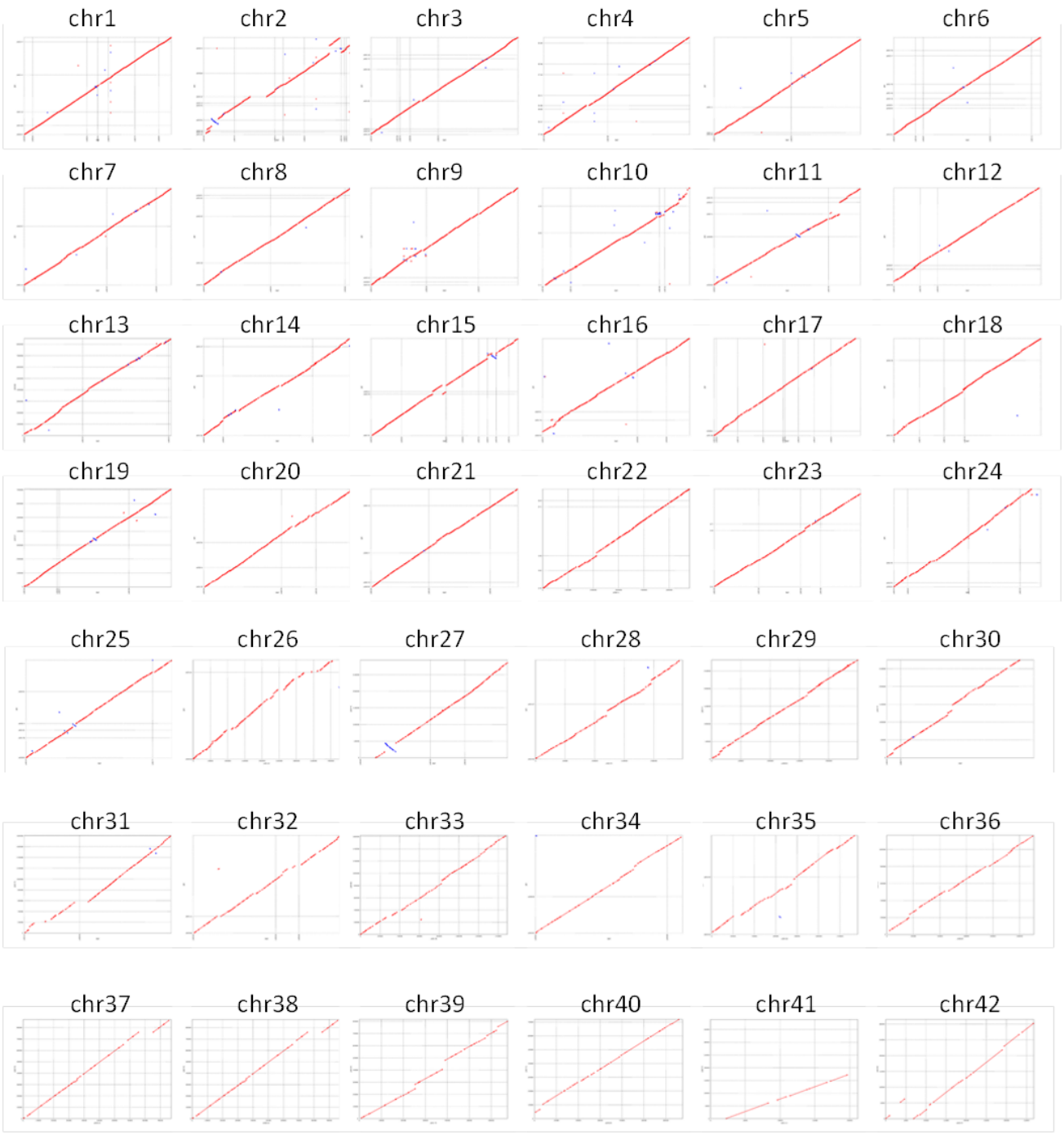

Chr. 1
Chr. 2
Chr. 3
Chr. 4
Chr. 5
Chr. 6
Chr. 7
Chr. 8
Chr. 9
Chr. 10
Chr. 11
Chr. 12

(continued)

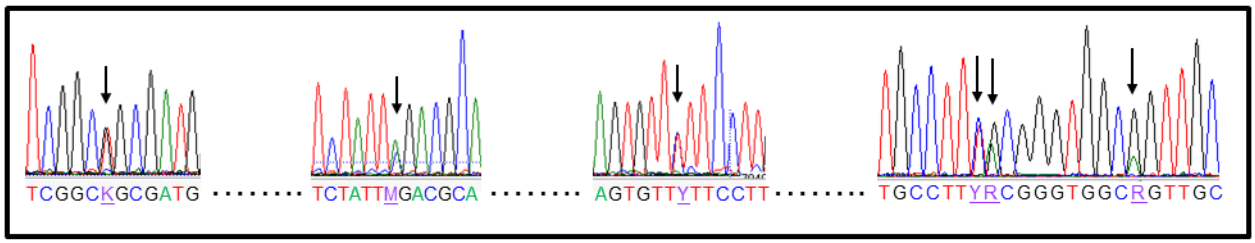


Supplemental Figure 1. Schematic representation of 84 chromosomes in *F. solaris* JPCC DA0580. Chromosome pairs showing high synteny are adjacently aligned. Coloured lines connecting the chromosome pairs depict nucleotide sequence similarity in every thousand bases. Gene distribution, transposable elements, and GC contents are expressed in the innermost, second, and outer columns adjacent to the chromosomes. Genes belonging to the families which are shared with oleaginous microalga *N. gaditana* are shown in purple.

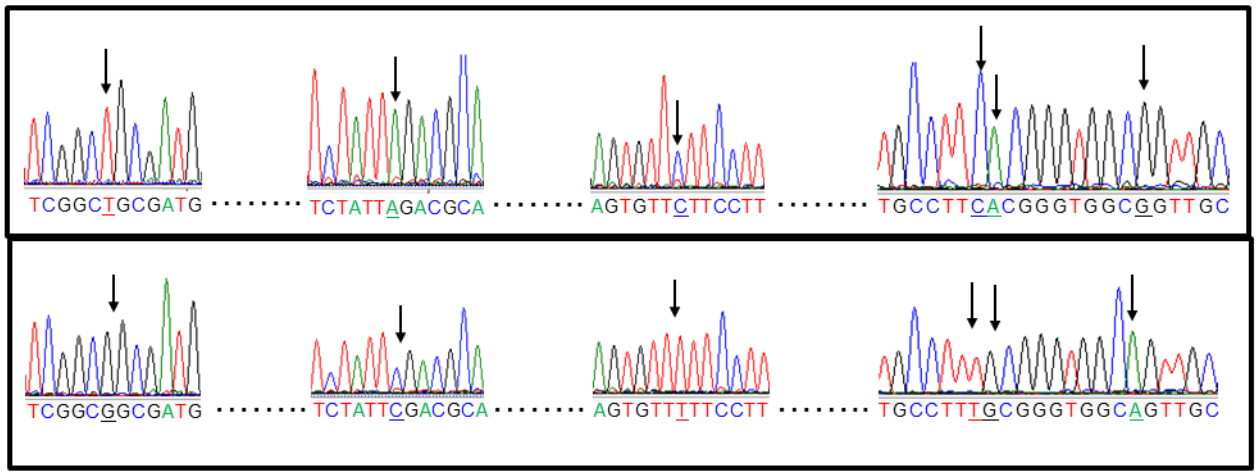


Supplemental Figure 2. Conservation of synteny between homeologous chromosome pairs in *F. solaris* JPCC DA0580

a



b

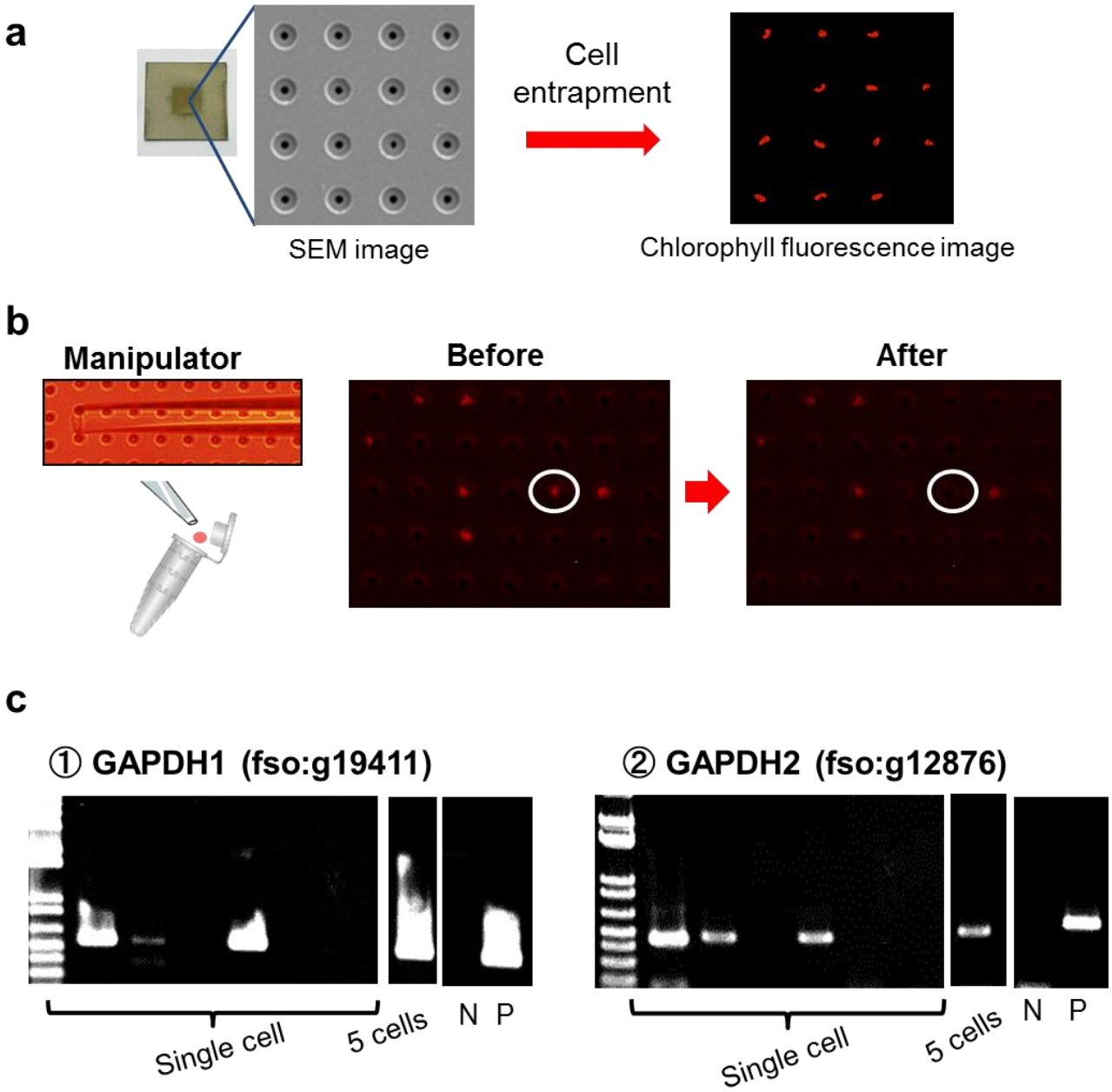


c

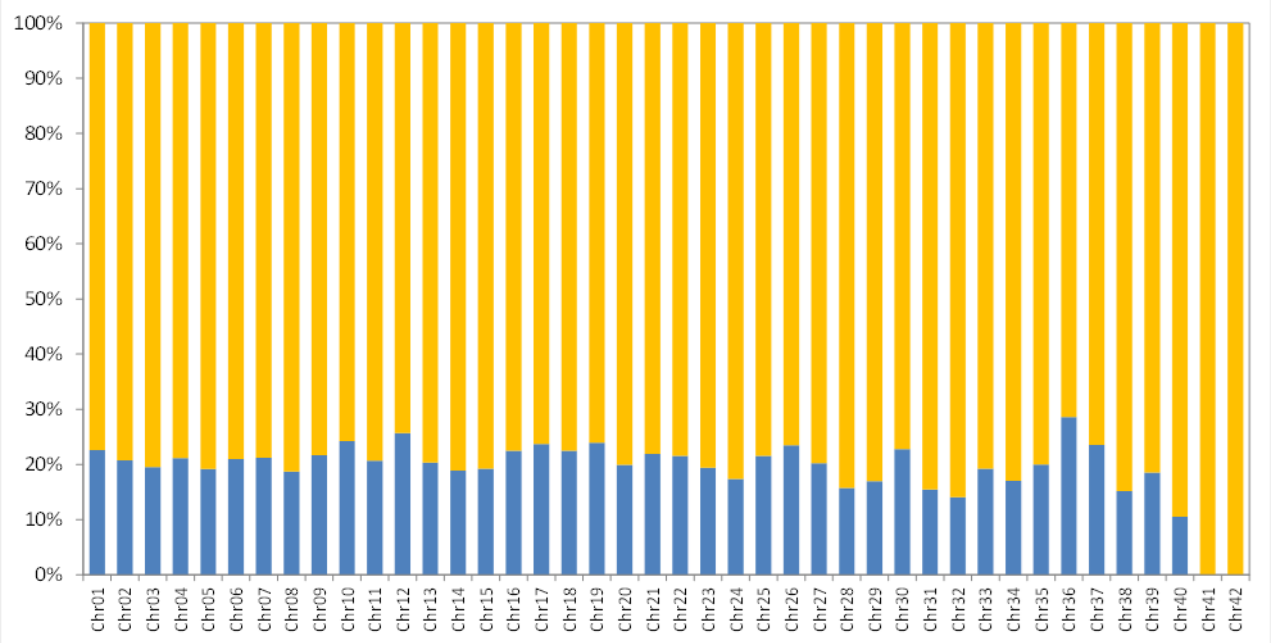
>*Fistulifera* sp. strain JPCC DA 0580 18S ribosomal RNA gene, partial sequence

CCAGTAGTCATATGCTCGTCTCAAAGATTAAGCCATGCATGTCTAAGTATAAATCTTTTACTTTGAAACTGCGAACGGCTCATTATATCAGTTATAGTTTATTTG
ATAGTCCCTTACTACTCGGATAACCGTAGTAATTCTAGAGCTAATACGTGCGTCAATACTCTTCGGAGTAGTATTTATTAGATGGAAACCAACTGCTTCGGCAT
GATGTGGTGATTCATAATAAGCTTGCGGATCGCGTGCTTCGGCKGCGATGGATCATTCAAGTTTCTGCCCTATCAGCTTTGACGGTACTGTATTGGAGTACCG
TGGCGGTAACGGGTAACGGGAAATTAGGGTTTGACACCGGAGAGGGAGCCTGAGAGACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGTAAATTACCC
AATCTTGACACAAGGAGGTAGTGACAATAAATAACAATGCCGGGCCTTTCTAGGTCTGGCAATTGGAATGAGAACAATTTAAACCCCTTATCGAGGATCCAT
TGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTAAAGTTGTTGCAGTTAAAAAGCTCGTAGTTGGATTTGTGGCGTGC
GTTGCGGCGTCCATTGATTTGGTTCTGCCTCGACCGCGCCATCCTTGGGTGGAATCTGTGTGGCATTAGGTTGTCGTGCAGGGGATGCCCATCGTTTACTGTG
AAAAAATTAGAGTGTTCAAAGCAGGCTTATGCCGTTGAATATATTAGCATGGAATAATAAGATAGGTCTAGGGTCCTATTTTGTTGGTTTGCGGTCCTTGGAA
ATGATTAACAGGGACAGTTGGGGGTATTCGTATTCCATTGTCAGAGGTGAAATTCTTGGATTTCTGGAAGACGAACTACTGCGAAAGCATTTACCAAGGATG
TTTTCATTAATCAAGAACGAAAGTTAGGGGATCGAAGATGATTAGATACCATCGTAGTCTTAACCATAAACTATGCCGACAAGGGATTGGCGGGGTTCGTTT
CGTCTCCGTCAGCACCTTATGAGAAATCACAAGTTTTTGGGTTCCGGGGGGGAGTATGGTCGCAAGGCTGAAACTTAAAGAAATTGACGGAAGGGCACCACC
AGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGAAAACTTACCAGGTCCAGACATAGTGAGGATTGACAGATTGAGAGCTCTTTCTTGATTCTATGG
GTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGTGATTTGTCTGGTTAATTCCGTTAACGAACGAGACCCCTGCCTGCTAAATAGTCCCGCGAGTGAATTTCA
CTGGCGCGGTCTTCTTAGAGGGACGTGCGTTCTATTMGACGCAGGAAGATAGGGGCAATAACAGGTCTGTGATGCCCTTAGATGTTCTGGGCCGCACGCGC
GCTACACTGATGCATTCAACGAGTGTTYTTCCTTGGCCGAGAGGCCTGGGCAATCTTTTGAACGTGCATCGTGATAGGGATAGATTATTGCAATTATTAATCTT
GAACGAGGAATTCCTAGTAAACGCAGATCATCAATCTGCATTGATTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCACCTACCGATTGAATGGTCCGGTG
AAGCCTCGGGATTGTGGCTGTTGCCTTYRCGGGTGGCRGTTGCGAGAACTTGTCTAAACCTTATCATTTAGAGGAAGGTGAAGTCGTAACAAGGTTTCCGTA
GGTGAACCTGCGGAAGGAT

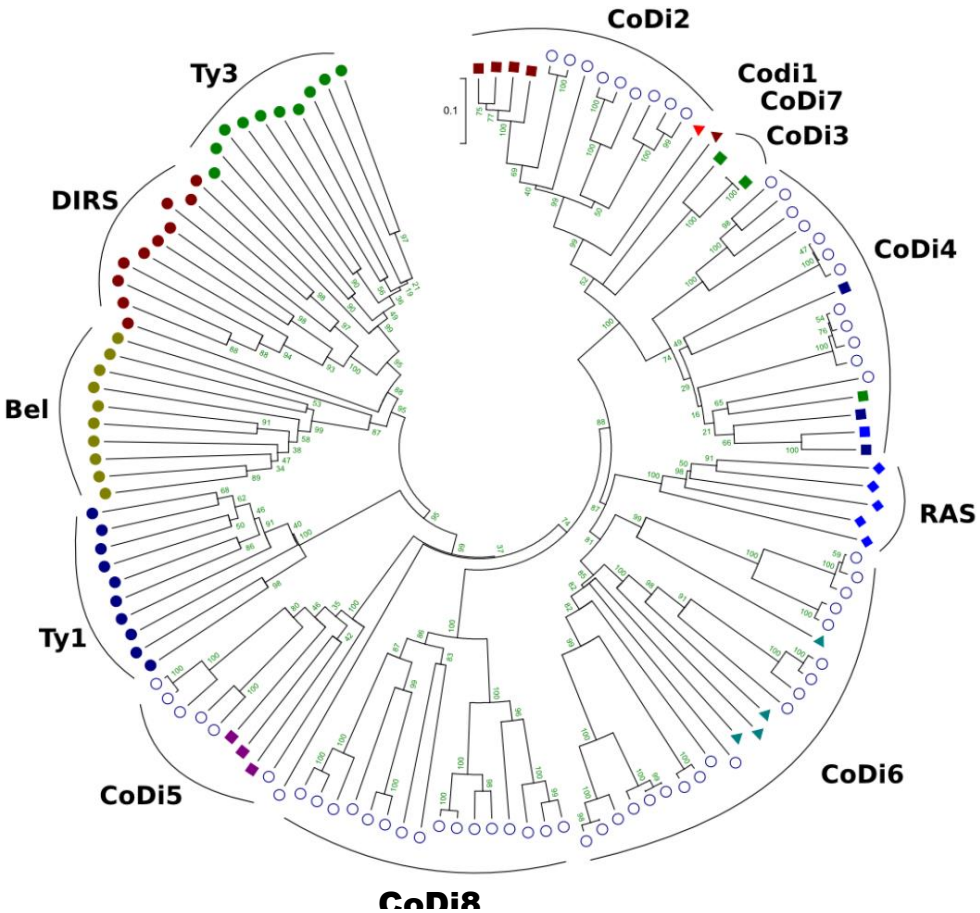
Supplemental Figure 3. The multiple 18S rDNA sequences in the genome of *F. solaris* JPCC DA0580. Sequencing data of the 18S rDNA PCR product (a) showed polymorphisms at six sites; after cloning, pure sequences for each type of rDNA were obtained (b). Polymorphic sites (purple characters) are shown in (c). K represents T or G, M represents A or C, Y represents C or T, R represents G or A.



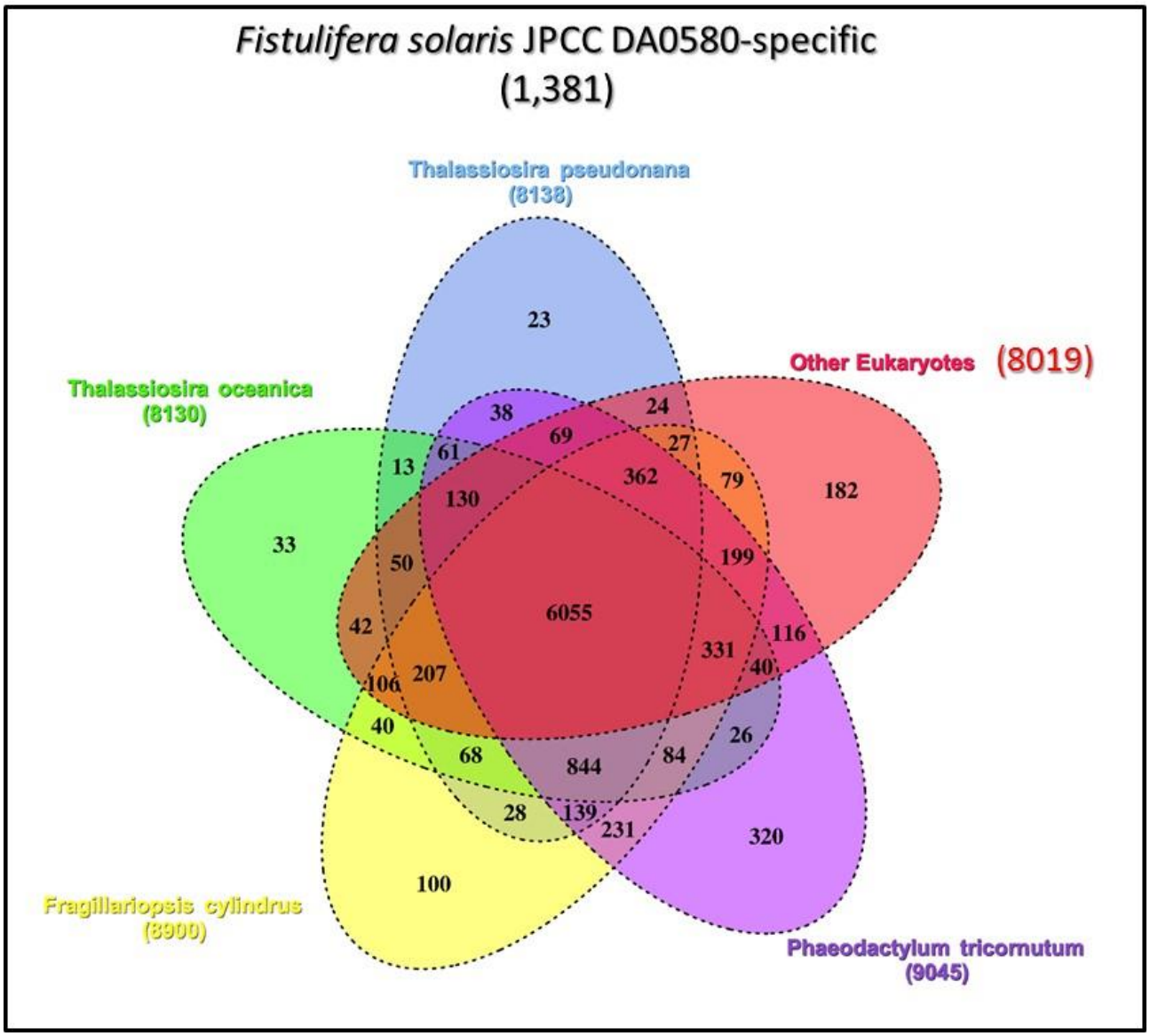
Supplemental Figure 4. Single-cell RT-PCR. (a) Microcavity array for entrapment and alignment of single cells; (b) Recovery of a single diatom from the microcavity array with a micromanipulator; (c) Single-cell RT-PCR of a putative homeologous GAPDH gene pair (fso:g19411 and fso:g12876). P: Positive control, N: Negative control



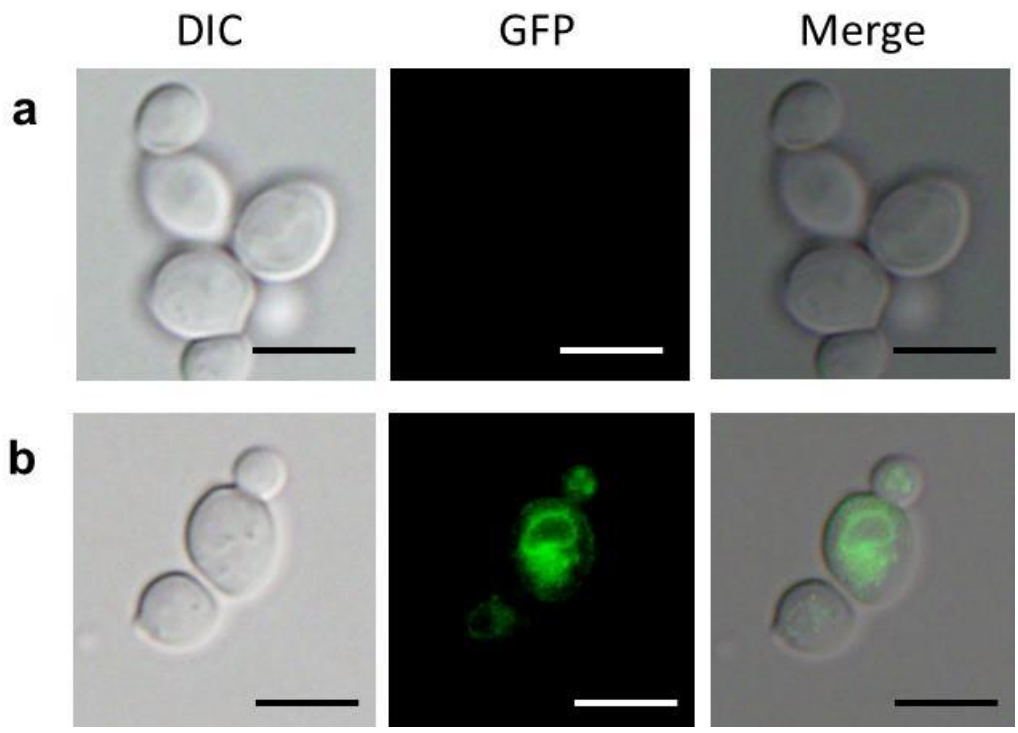
Supplemental Figure 5. Expression patterns of homeologous gene pairs throughout the 42 chromosome pairs. Abundance ratio of gene pairs showing synchronised (blue bars) and differential (orange bars) expression in each chromosome pair.



Supplemental Figure 6. CoDi classification in *F. solaris* JPCC DA0580 based on the RT domains. RAS refers to Red and Aquatic species. Unfilled circles are sequences from *F. solaris* JPCC DA0580.

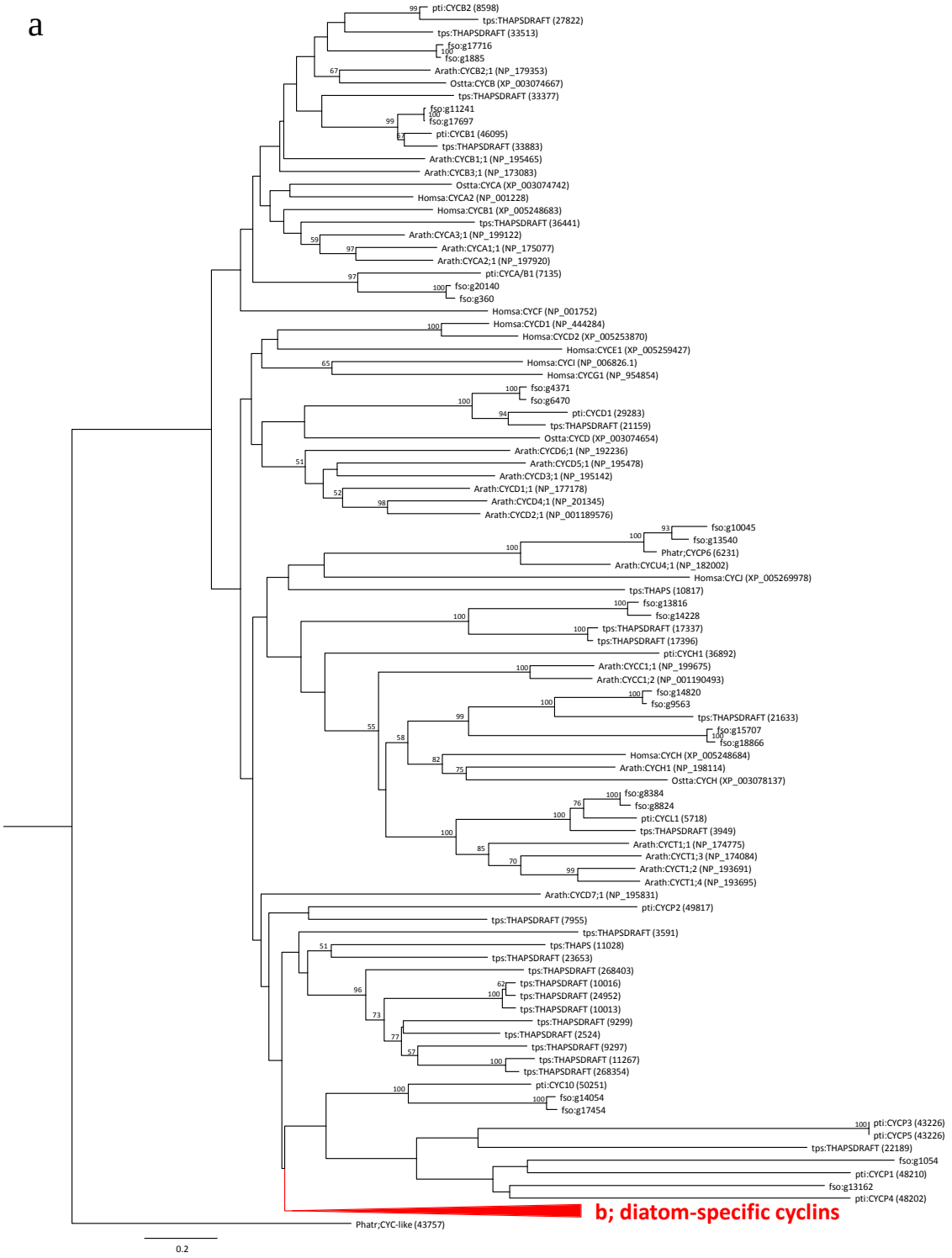


Supplemental Figure 7. Venn diagram of shared/unique genes of *F. solaris* JPCC DA0580. The outer square represents the whole set of *F. solaris* JPCC DA0580 (pennate) genes. Ellipses represent genes shared with *P. tricornutum* (pennate), *T. pseudonana* (centric), *T. oceanica* (centric), *F. cylindrus* (pennate), and other eukaryotes. The numbers of *F. solaris* JPCC DA0580 genes shared with other organisms are specified in brackets. The numbers of diatom-, pennate-, and *F. solaris*-specific genes are 844, 231, and 1,381, respectively.



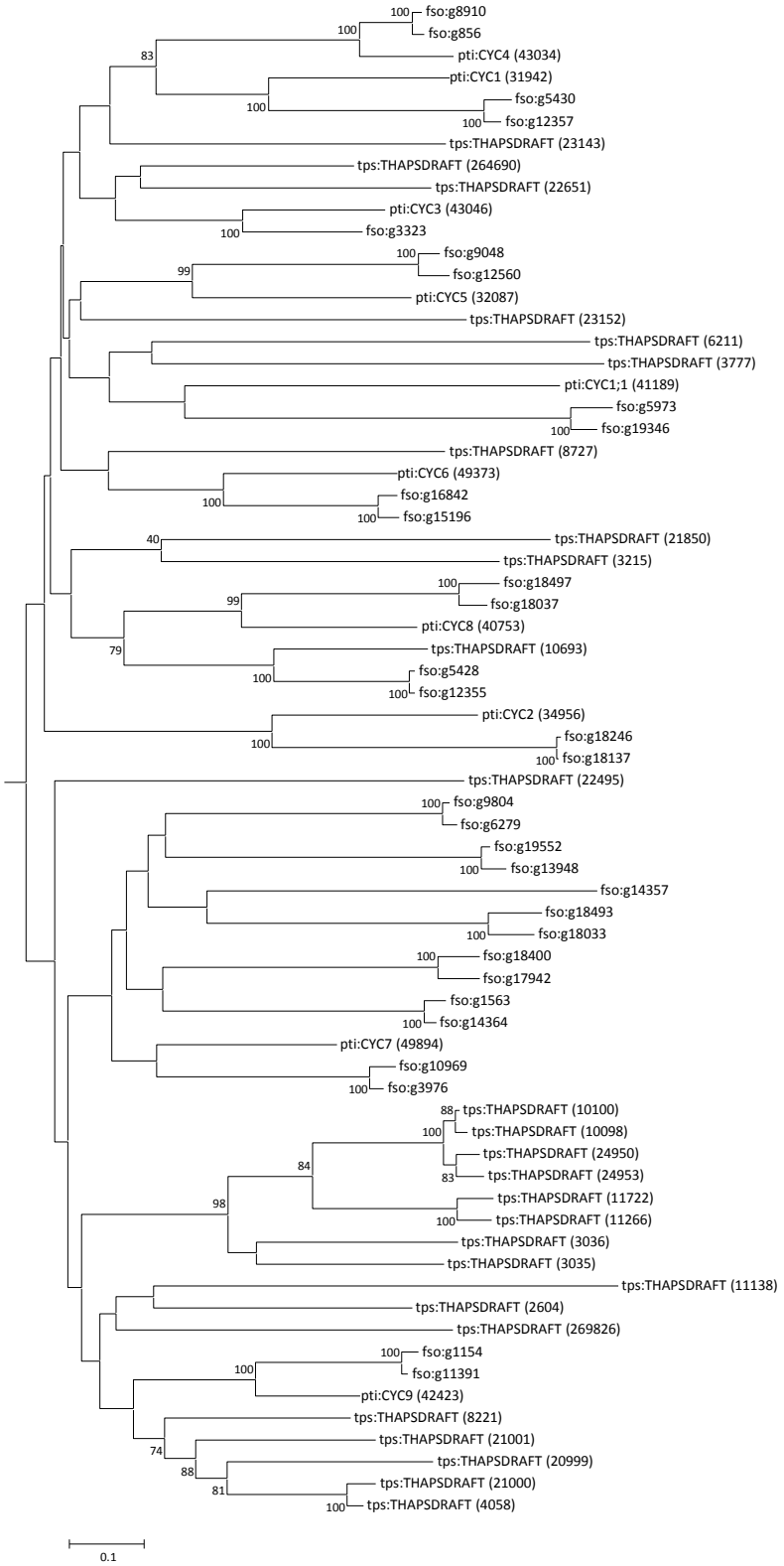
Supplemental Figure 8. Fluorescence microscopy of *Saccharomyces cerevisiae* cells expressing FIT-like protein. (a) The cells harboring an empty vector pYES2.1/V5-His/*lacZ*, and (b) the cells harboring FIT-like protein-GFP expression vector pYES2.1/V5-g11329-gfp. (scale bar = 5 μm)

a


pti:CYCB2 (8598)
tps:THAPSDRAFT (27822)
tps:THAPSDRAFT (33513)
fso:g17716
fso:g1885
Arath:CYCB2;1 (NP_179353)
Ostta:CYCB (XP_003074667)
tps:THAPSDRAFT (33377)
fso:g11241
fso:g17697
pti:CYCB1 (46095)
tps:THAPSDRAFT (33883)
Arath:CYCB1;1 (NP_195465)
Arath:CYCB3;1 (NP_173083)
Ostta:CYCA (XP_003074742)
Homsa:CYCA2 (NP_001228)
Homsa:CYCB1 (XP_005248683)
tps:THAPSDRAFT (36441)
Arath:CYCA3;1 (NP_199122)
Arath:CYCA1;1 (NP_175077)
Arath:CYCA2;1 (NP_197920)
pti:CYCA/B1 (7135)
fso:g20140
fso:g360
Homsa:CYCF (NP_001752)
Homsa:CYCD1 (NP_444284)
Homsa:CYCD2 (XP_005253870)
Homsa:CYCE1 (XP_005259427)
Homsa:CYCI (NP_006826.1)
Homsa:CYCG1 (NP_954854)
fso:g4371
fso:g6470
pti:CYCD1 (29283)
tps:THAPSDRAFT (21159)
Ostta:CYCD (XP_003074654)
Arath:CYCD6;1 (NP_192236)
Arath:CYCD5;1 (NP_195478)
Arath:CYCD3;1 (NP_195142)
Arath:CYCD1;1 (NP_177178)
Arath:CYCD4;1 (NP_201345)
Arath:CYCD2;1 (NP_001189576)
fso:g10045
fso:g13540
Phatr;CYCP6 (6231)
Arath:CYCU4;1 (NP_182002)
Homsa:CYCJ (XP_005269978)
tps:THAPS (10817)
fso:g13816
fso:g14228
tps:THAPSDRAFT (17337)
tps:THAPSDRAFT (17396)
pti:CYCH1 (36892)
Arath:CYCC1;1 (NP_199675)
Arath:CYCC1;2 (NP_001190493)
fso:g14820
fso:g9563
tps:THAPSDRAFT (21633)
fso:g15707
fso:g18866
Homsa:CYCH (XP_005248684)
Arath:CYCH1 (NP_198114)
Ostta:CYCH (XP_003078137)
fso:g8384
fso:g8824
pti:CYCL1 (5718)
tps:THAPSDRAFT (3949)
Arath:CYCT1;1 (NP_174775)
Arath:CYCT1;3 (NP_174084)
Arath:CYCT1;2 (NP_193691)
Arath:CYCT1;4 (NP_193695)
Arath:CYCD7;1 (NP_195831)
pti:CYCP2 (49817)
tps:THAPSDRAFT (7955)
tps:THAPSDRAFT (3591)
tps:THAPS (11028)
tps:THAPSDRAFT (23653)
tps:THAPSDRAFT (268403)
tps:THAPSDRAFT (10016)
tps:THAPSDRAFT (24952)
tps:THAPSDRAFT (10013)
tps:THAPSDRAFT (9299)
tps:THAPSDRAFT (2524)
tps:THAPSDRAFT (9297)
tps:THAPSDRAFT (11267)
tps:THAPSDRAFT (268354)
pti:CYC10 (50251)
fso:g14054
fso:g17454
pti:CYCP3 (43226)
pti:CYCP5 (43226)
tps:THAPSDRAFT (22189)
fso:g1054
pti:CYCP1 (48210)
fso:g13162
pti:CYCP4 (48202)
b; diatom-specific cyclins
Phatr;CYC-like (43757)
0.2

(continued)

b


fso:g8910
fso:g856
pti:CYC4 (43034)
pti:CYC1 (31942)
fso:g5430
fso:g12357
tps:THAPSDRAFT (23143)
tps:THAPSDRAFT (264690)
tps:THAPSDRAFT (22651)
pti:CYC3 (43046)
fso:g3323
fso:g9048
fso:g12560
pti:CYC5 (32087)
tps:THAPSDRAFT (23152)
tps:THAPSDRAFT (6211)
tps:THAPSDRAFT (3777)
pti:CYC1;1 (41189)
fso:g5973
fso:g19346
tps:THAPSDRAFT (8727)
pti:CYC6 (49373)
fso:g16842
fso:g15196
tps:THAPSDRAFT (21850)
tps:THAPSDRAFT (3215)
fso:g18497
fso:g18037
pti:CYC8 (40753)
tps:THAPSDRAFT (10693)
fso:g5428
fso:g12355
pti:CYC2 (34956)
fso:g18246
fso:g18137
tps:THAPSDRAFT (22495)
fso:g9804
fso:g6279
fso:g19552
fso:g13948
fso:g14357
fso:g18493
fso:g18033
fso:g18400
fso:g17942
fso:g1563
fso:g14364
pti:CYC7 (49894)
fso:g10969
fso:g3976
tps:THAPSDRAFT (10100)
tps:THAPSDRAFT (10098)
tps:THAPSDRAFT (24950)
tps:THAPSDRAFT (24953)
tps:THAPSDRAFT (11722)
tps:THAPSDRAFT (11266)
tps:THAPSDRAFT (3036)
tps:THAPSDRAFT (3035)
tps:THAPSDRAFT (11138)
tps:THAPSDRAFT (2604)
tps:THAPSDRAFT (269826)
fso:g1154
fso:g11391
pti:CYC9 (42423)
tps:THAPSDRAFT (8221)
tps:THAPSDRAFT (21001)
tps:THAPSDRAFT (20999)
tps:THAPSDRAFT (21000)
tps:THAPSDRAFT (4058)
0.1

(continued)

Supplemental Figure 9. Phylogenetic analysis of the cyclins of *F. solaris* JPCC DA0580. The clade including diatom-specific cyclins is depicted as a red triangle in (a), and its detail is shown in (b). Abbreviations: fso, *F. solaris* JPCC DA0580; pti, *P*. *tricornutum*; tps, *T*. *pseudonana*; Arath, *Arabidopsis thaliana*; Homsa, *Homo sapience*; Ostta, *Ostreococcus tauri*

Supplemental Table 1. Sequencing analysis of *F. solaris* JPCC DA0580 by GS FLX Titanium DNA pyrosequencing

Parameters	**Results**
Total # of reads	2,966,486
Total length of reads (b)	1,239,444,594
Redundant sequence coverage	24.8
Total scaffold #	297
Total scaffold length (b)	49,899,580
Avg. scaffold length (b)	168,012
Maximum scaffold length (b)	904,706
Contig #	3,913
Total contig length (b)	49,713,831
GC content (%)	46.13
Contig peak depth	20
Contig (≥500b) #	1,592
Total contig (≥500b) length (b)	49,433,762
Contig (≥4kb) #	1,083
Total contig (≥4kb) length(b)	48,728,239

Supplemental Table 2. The list of genes shared between *F. solaris* JPCC DA0580-specific gene families and the entire *N. gaditana* gene family

PFAM domain ID	PFAM domain	Annotation	Genes in *F. solaris*	Genes in *N. gaditana*
PF00221	PAL	Phenylalanine and histidine ammonia-lyase	g981	Nga30438
PF00346	Complex1_49kDa	Respiratory-chain NADH dehydrogenase, 49 Kd subunit	g16662	Nga50031
PF00932	IF_tail	Intermediate filament tail domain	g15656, g16968, g17044, g5419	Nga05631
PF01056	Myc_N	Myc amino-terminal region	g7178, g8843	Nga05585
PF01094	ANF_receptor	Receptor family ligand binding region	g4832, g4833, g8203, g8204, g8205	Nga04451.1
PF01136	Peptidase_U32	Peptidase family U32	g13254, g16553	Nga06945
PF01632	Ribosomal_L35p	Ribosomal protein L35	g17004	Nga40026
PF02597	ThiS	ThiS family	g12484, g5676	Nga40049
PF03024	Folate_rec	Folate receptor family	g11127, g11130, g11151, g17613, g4184, g4187	Nga30480
PF03066	Nucleoplasmin	Nucleoplasmin	g13670, g13698, g6232, g7851, g8636, g8843	Nga01971.01
PF03552	Cellulose_synt	Cellulose synthase	g16990, g8426	Nga05300
PF03845	Spore_permease	Spore germination protein	g17080, g20313	Nga02631
PF04065	Not3	Not1 N-terminal domain, CCR4-Not complex component	g5473, g605	Nga02973
PF04086	SRP-alpha_N	Signal recognition particle, alpha subunit, N-terminal	g9011	Nga05826.1
PF04148	Erv26	Transmembrane adaptor Erv26	g10909	Nga05247
PF04494	TFIID_90kDa	WD40 associated region in TFIID subunit	g17360, g18307	Nga20113.1
PF04934	Med6	MED6 mediator sub complex component	g11073, g207	Nga01085
PF05432	BSP_II	Bone sialoprotein II (BSP-II)	g10281, g10904, g13743, g5559, g616	Nga00963
PF05915	DUF872	Eukaryotic protein of unknown function (DUF872)	g13246	Nga02193.01
PF05964	FYRN	F/Y-rich N-terminus	g15003, g15532	Nga03750.1
PF05995	CDO_I	Cysteine dioxygenase type I	g4423	Nga06469
PF06101	DUF946	Plant protein of unknown function (DUF946)	g11099, g182	Nga00395
PF06108	DUF952	Protein of unknown function (DUF952)	g14145	Nga07074
PF06179	Med22	Surfeit locus protein 5 subunit 22 of Mediator complex	g15954	Nga00195.01
PF06624	RAMP4	Ribosome associated membrane protein RAMP4	g11153, g17615, g8909	Nga00191
PF06810	Phage_GP20	Phage minor structural protein GP20	g11659, g14906, g2576, g5471	Nga02702
PF07574	SMC_Nse1	Nse1 non-SMC component of SMC5-6 complex	g14836, g9579	Nga06324.01
PF07958	DUF1688	Protein of unknown function (DUF1688)	g10619, g13607	Nga02320.01
PF08432	DUF1742	Fungal protein of unknown function (DUF1742)	g11462	Nga04186
PF08636	Pkr1	ER protein Pkr1	g18804	Nga02572
PF08654	DASH_Dad2	DASH complex subunit Dad2	g19400	Nga07178
PF08696	Dna2	DNA replication factor Dna2	g12853, g153	Nga00884.1
PF09320	DUF1977	Domain of unknown function (DUF1977)	g16580, g6160	Nga02596
PF10153	DUF2361	Uncharacterised conserved protein (DUF2361)	g10048, g13543	Nga04245
PF10174	Cast	RIM-binding protein of the cytomatrix active zone	g15069, g19232, g2858, g9751	Nga07217
PF10229	DUF2246	Uncharacterized conserved protein (DUF2246)	g15366, g8690	Nga04304.01
PF10261	Scs3p	Inositol phospholipid synthesis protein Scs3p	g11329, g13770	Nga07301
PF10303	DUF2408	Protein of unknown function (DUF2408)	g6333	Nga03427
PF10961	DUF2763	Protein of unknown function (DUF2763)	g17706	Nga03702
PF11543	UN_NPL4	Nuclear pore localisation protein NPL4	g16948	Nga01268.01
PF11831	DUF3351	Domain of unknown function (DUF3351)	g2360, g3726	Nga05876, Nga05879
PF11861	DUF3381	Domain of unknown function (DUF3381)	g16169, g7852	Nga04812
PF11891	DUF3411	Domain of unknown function (DUF3411)	g7007, g916	Nga02558
PF12063	DUF3543	Domain of unknown function (DUF3543)	g13365, g1444	Nga00270
PF12118	SprA-related	SprA-related family	g18702	Nga03714
PF12353	eIF3g	Eukaryotic translation initiation factor 3 subunit G	g16545, g19014, g7134	Nga01351.01
PF12463	DUF3689	Protein of unknown function (DUF3689)	g12172, g16697	Nga06368
PF12619	MCM2_N	Mini-chromosome maintenance protein 2	g15355, g8680	Nga04400.1

Supplemental Table 3. The number of diatom genes in the major metabolism pathways

	Fistulifera solaris JPCC DA0580	*P. tricornutum*	*T. pseudonana*
Glycolysis	40	42	38
Glyconeogenesis	6	6	4
TCA cycle	20	20	20
Nitrogen metabolism	17	17	17
Carbobohydrate synthesis	12	13	7
Fatty acid synthesis	11	10	9
Fatty acid degradation	17	20	19
Glucerolipid synthesis	10	10	10
PUFA synthesis	10	10	10
Calvin-Benson Cycle	7	9	8

Supplemental Methods

Repeat identification and annotation

Repeat identification and annotation is an important part of any genome sequencing project. Here, we used three approaches to annotate repeats: (1) structure-based detection of long-terminal repeats (LTRs); (2) sequence-based transposable element (TEs) identification; and (3) a combination of these methods for classification and annotation.

(i) *Structure-based LTR detection*: LTR pairs in the 297 scaffolds of *F. solaris* were detected with LTR_finder (Xu and Wang, 2007). The presence of a primer binding site, integrase, reverse transcriptase domain, or RnaseH was used to facilitate LTR identification. LTR_finder precisely predicts the reverse transcriptase domain which is later used for classification.

(ii) *Repeat identification by similarity search and annotation*: Repeat annotation was carried out using the two-step REPET pipeline (Flutre et al., 2011): (a) *de novo* repeat identification in the 297 scaffolds and (b) homology search against known repeats from RepBase-Dec2012 (Jurka et al., 2005). In the *de novo* TE identification, the *F. solaris* genome was split into small units, followed by high scoring pair (HSP) detection within the genome using BLASTER and WU-BLAST with e-value e-300. The second step was based on structural detection of LTRs using LTRharvest (Ellinghaus et al., 2008). Structural parameters included 100–1000 bp LTR, 87% similarity between the two LTRs, and 20-bp target site duplication (TSD) at the two ends. Clustering was performed using Grouper, Recon, and Piler. Multiple sequence alignments were built with MULTALIGN. Classification of the *de novo* library into LTRs, terminal inverted repeats, simple sequence repeat-like, PolyA, etc. was performed with the PASTEC tool according to the Wicker's classification model. The *de novo* and TE libraries from *P. tricornutum* and *T. pseudonana* were combined and clustered to create a consensus full-length diatom TE library. This library was used in the second step of genome annotation, for which BLASTER, RepeatMasker, and CENSOR were used. MATCHER removed overlapping HSPs and made connections with the 'long joint' procedure. LTR_finder results and REPET-derived repeats were combined, curated, and exported to GFF3 format for loading into any genome visualisation tool.

Classification of the CoDi group was based on reverse transcriptase domains, shown in a phylogenetic tree with *Dictyostelium* intermediate repeat sequence (DIRS), Bel (Lineage of insect LTR), Ty1(Copia-type), and red/aquatic species (RAS) as outgroups. Diatom reverse transcriptase sequences were classified into 26 families (Seven CoDi groups) including the DIRS sequences from *Dictyostelium discoideum* and *Xenopus* (*Silurana*) *tropicalis*, Bel sequences from *Trichosurus vulpecula* and *Ovis aries*, Ty1 sequences from *Solanum demissum* and *Oryza sativa*, Ty3 from *Oryza sativa*, and RAS from *Eumunida annulosa* and *Porphyra yezoensis*, as well as the *F. solaris* reverse transcriptase domain sequences. A multiple sequence alignment of the reverse transcriptase domain clusters was generated with MULTALIGN and the phylogenetic tree was created using MEGA5 software. Compared to *T. pseudonana* and *P. tricornutum*, we found an expansion of CoDi elements in *F. solaris* (both in copy number and new class). *F. solaris*-specific CoDi is marked as 'CoDi8'.

Transcription factor detection

HMMER3.0 was used to search for transcription factors with e-value <0.00001. Transcription factor family sequences were obtained from domain database Pfam-A. We performed a comparative analysis of transcription factors for other diatom proteomes with RNAseq support.

Comparative studies of genes

Comparative gene-set analysis in *F. solaris*, four other sequenced diatoms (*P. tricornutum, T. pseudonana, T. oceanica*, and *Fragillariopsis cylindrus*), and other eukaryotic organisms recorded in the National Center for Biotechnology Information (NCBI) database was performed using BLASTP with e-value of 10^{-5} and a 50-amino acid overlap. Sequence searches against publicly available databases and comparative analysis showed that diatom genes originated from at least four phylogenetic groups (Eukaryotic core, Diatom core, Species-specific gene set, and Bacterial origin group). *F. solaris*-specific genes were derived by pairwise comparison against these data sets. Genes (Filtered set) from the four diatoms were obtained from JGI and GEOMAR. The overlap was plotted using VennDiagram library in R. Diatom-specific genes were present in all analysed diatoms but not elsewhere. Pennate-specific genes were present in *P. tricornutum*, *Fragillariopsis cylindrus*, and *F. solaris* but not elsewhere. *F. solaris*-specific genes were found only in the genome of *F. solaris* JPCC DA0580.

Prediction of pathway localisation

Candidate genes in pathways such as carbon and lipid metabolism were screened by BLASTP. Protein sequences with KO numbers in the KEGG database or protein sequences described in published papers were used as query sequences. Candidate genes were submitted to InterProScan to identify protein signatures and provide annotation. In cases where gene annotations based on protein signatures were ambiguous, phylogenetic tree analysis was also used to assess phylogenetic relationships with diatom genes of known function. Finally, the localisation of proteins encoded by these annotated genes was predicted using several bioinformatics programs. TargetP (Emanuelsson et al., 2000) and HECTAR (Gschloessl et al., 2008) were used to predict general protein localisation. Signal peptides for the ER or chloroplast-targeting peptides were screened using SignalP (Bendtsen et al., 2004). When a cleavage site was predicted by SignalP, the presence of an ER retention signal (K(/D)-D(/E)-E-L) in the C-terminus was checked manually (Kroth et al., 2008). If proteins possessed no ER retention signals and contained F, W, Y, or L at the +1 position of cleavage sites, they were considered to be chloroplast-targeted proteins (Gruber et al., 2007). When proteins were predicted by both TargetP and HECTAR to be localised in the mitochondria, the proteins were considered to be transported into mitochondria. In addition, Mitoprot (Claros and Vincens, 1996) was used to detect mitochondrial targeting peptides. If (1) the Mitoprot score was >0.9 or (2) >0.8 and mitochondrial localisation was predicted by either TargetP or HECTAR, the proteins were also considered to be transported into mitochondria. For proteins without chloroplast or mitochondrial targeting peptides, peroxisome-targeting signals at the C-terminus (S(/A/C)-K(/R/H)-L(/M) or S-S-L) were checked manually (Gonzalez et al., 2011). TMHMM (Krogh et al., 2001) was used to

detect transmembrane regions. If transmembrane regions were predicted in proteins without targeting peptides, the possibility of ER localisation could not be excluded. Proteins that conform to none of the above criteria were predicted to be localised in the cytoplasm. This pipeline was also described in previous report (Sunaga et al., 2014).

Single-cell analysis

Single-cell entrapment of *F. solaris* and manipulation were conducted on a microcavity array. The device was microfabricated as described (Hosokawa et al., 2009; Hosokawa et al., 2012). In brief, a poly(ethylene terephthalate) (PET) plate (20 × 20 mm, thickness = 38 μm) was used as the substrate to fabricate the microcavity array. The distance between each cavity was 30 μm, with 44,100 cavities arranged in a 210 × 210 array on a typical microscope glass slide. A poly(dimethylsiloxane) (PDMS) structure equipped with a vacuum microchannel (i.d. = 500 μm) was fitted directly beneath the microcavity array to apply negative pressure for cell entrapment. The vacuum microchannel was connected to a peristaltic pump, and the cell entrapment setup was placed on the computer-operated motorised stage of an upright microscope.

Single F. solaris cell entrapment: *F. solaris*: cells at logarithmic growth phase (10^6 cells) were loaded onto the microcavity array. Negative pressure was applied to the cell suspension in phosphate-buffered saline using a peristaltic pump connected to the vacuum microchannel at a flow rate of 200 μL/min, allowing the cells to be driven towards and entrapped on the array (Supplemental Figure 8-a). Images of the entrapped cells were captured using a fluorescence microscope (BX61; Olympus Corporation, Japan) equipped with a computer-operated motorised stage, FGW filter sets (Olympus Corporation, Japan), and a cooled digital camera (DP70; Olympus Corporation, Japan). Lumina Vision acquisition software (Mitani Corporation, Japan) was used for image acquisition.

Single-cell RT-PCR analysis for the detection of mRNAs from each glyceraldehyde-3-phosphate dehydrogenase (GAPDH) homeologous genes: Single cells were collected from the cavities by using micropipettes (20 □m diameter) made from glass capillaries using a micromanipulator system (Eppendorf Co., Ltd., Tokyo, Japan). The recovered cells were transferred to a 10-□L microtube. The recovered single cells were used directly for reverse transcription and PCR amplification with the Primescript RT-PCR Kit (Takara Bio Inc., Japan). mRNA extraction (65°C for 5 min) and annealing with an Oligo-dT primer (4°C for 10 min) were followed by reverse transcription at 42°C for 60 min and 95°C for 5 min, and by PCR amplification for 50 cycles at 94°C for 30 s and 60°C for 30 s. PCR products were confirmed by 1% agarose gel electrophoresis with ethidium bromide staining and by sequencing. Amplification primers for GAPDH (g12876 and g19411) were as follows: 5′-ACGGAAAAGGCTCAGGCC-3′ (forward primer for g12876), 5′-TCGACGATACTGGAGCGC-3′ (reverse primer for g12876), 5′-CACGGAAAAGGCTCAGGCA-3′ (forward primer for g19411), and 5′-GGACGAAATCTTGCGACACA-3′ (reverse primer for g12876).

Expression of gene encoding FIT-like protein in yeast cells

Comparative gene family analysis revealed a unique family of *FIT* genes (also known as *Scs3p*) shared with *F. solaris* and *N. gaditana* (Figure 2 in the main text). Two genes encoding homeologous FIT-like protein (g11329, g13770) were assigned to this family from the genome (Supplemental Table 2). For heterologous expression of g11329 in yeast cells, an expression vector including a g11329-*gfp* fusion gene was constructed. A g11329 gene fragment without stop codon was amplified from *F. solaris* genomic DNA, then inserted into the *Eco*RI site within the vector, pSP-GFP/GAPDH (Nojima et al., 2013) to fuse the g11329 and *gfp* genes (performed by Takara Bio Inc., Japan). The constructed g11329-*gfp* fusion gene was amplified using a specific primer set (Forward primer; 5' – ATG CCT CCC CGG GAA GCC – 3', Reverse primer; 5' – CTT GTA CAG CTC GTC CAT GC – 3'), and cloned into the yeast expression vector pYES2.1/V5-His-TOPO (Invitrogen). The constructed plasmid is denoted pYES2.1/V5-g11329-GFP. *Saccharomyces cerevisiae* INVSc-1 (*MATa/MATα, his3Δ1/his3Δ1, leu2/leu2, trip1-289/trip1-289,* and *ura3-52/ura3-52*) (Invitrogen) was transformed with the plasmid DNA using *S. c.* EasyComp™ Transformation Kit (Invitrogen) according to the manufacturer's instructions. Yeast cells harboring the pYES2.1/V5-His/*lacZ*, which does not include g11329 gene, were used as a negative control. The transformants were selected by uracil prototrophy on a selective dropout media (SD) plate lacking uracil. For protein expression, SD medium containing 2% (w/v) galactose was used, and the transformants grown at 30°C. The resulting cells were observed using a fluorescence microscope (BX51; Olympus Corporation, Tokyo, Japan) equipped with a cooled digital camera (DP-70; Olympus) with a NIBA filter set for GFP.

Phylogenetic genetic analysis of cyclins

The cyclin amino acid sequences were retrieved from the genome sequence data of *P. tricornutum*, *T. pseudonana*, *Arabidopsis thaliana*, *Homo sapience* and *Ostreococcus tauri* in the NCBI database, where cyclins were searched with the following KEGG orthology (KO) number (ko:K15161, ko:K15188, ko:K14505, ko:K12760, ko:K10289, ko:K10145, ko:K10146, ko:K10151, ko:K10152, ko:K06659, ko:K06656, ko:K06657, ko:K06654, ko:K06626, ko:K06627, ko:K06634, ko:K06646, ko:K06650, ko:K06651, ko:K05868, ko:K04503). The amino acid sequences were aligned with that of *F. solaris* using the ClustalW (Larkin et al., 2007). The phylogenetic tree was constructed via the neighbor-joining method and evaluated with 1,000 rounds of bootstrapping using MEGA5.0 (Tamura et al., 2011). The bootstrap values for neighbor-joining analysis was calculated to measure the tree strength.

Abbreviations

ACAT	acetyl-CoA acyltransferase
ACC	acetyl-CoA carboxylase
ACP	acyl carrier protein
ADE	acyl-CoA dehydrogenase
AHY	aconitate hydratase
AMPK	AMP-activated kinase
AOX	acyl-CoA oxidase
Arg	arginase
AsL	argininosuccinate lyase
AsuS	argininosuccinate synthase
BDE	butyryl-CoA dehydrogenase
BGL	beta-glucosidase
BGS	1,3-β-glucan synthase
BLAST	basic local alignment search tool
Chr	chromosome
CAT1	carnitine acyltransferase I (also known as carnitine palmitoyltransferase I)
CPS	carbamoyl phosphate synthase (ammonia-dependent)
CPT	cholinephosphotranseferase
CSY	citrate synthase
DAG	diacylglycerol
DGAT	diacylglycerol acyltransferase
DGD	digalactosyldiacylglycerol synthase
DGDG	digalactosyldiacylglycerol
DGK	diacylglycerol kinase
DIRS	*Dictyostelium* intermediate repeat sequence
EAR	enoyl-ACP reductase
EHY	enoyl-CoA hydratase
EHY/HADE	enoyl-CoA hydratase/beta-hydroxyacyl-CoA dehydrogenase
endo-Glu	endo-1,3-β-glucanase
ENO	enolase
ER	endoplasmic reticulum
exo-Glu	exo-1,3-β-glucanase
FAS	fatty acid synthesis
FBA	fructose-bisphosphate aldolase
FBP	fructose-1,6-bisphosphatase
FFA	free fatty acid

FHY	fumarate hydratase, class I
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
GLK	glucokinase
GOGAT	glutamate synthase (ferredoxin-dependent)
GPAT	glycerol-3-phosphate acyltransferase
GPI	glucose-6-phosphate isomerase
GS	glutamine synthetase
G3P	glycerol 3-phosphate
HAD	beta-hydroxyacyl-ACP dehydrase
HADE	beta-hydroxyacyl-CoA dehydrogenase
HMM	Hidden Markov model
HSP	high scoring pair
ICDE	isocitrate dehydrogenase
KAR	beta-ketoacyl-ACP reductase
KAS	beta-ketoacyl-ACP synthase
KEGG	Kyoto encyclopedia of genes and genomes
KO	KEGG orthology
LACS	long-chain acyl-CoA synthetase
LPAAT	lysophosphatidic acid acyltransferase
LTR	long-terminal repeat
Lyso-PA	lyso-phosphatidic acid
MAT	malonyl-CoA ACP transacylase
MDE	malate dehydrogenase
MGD	monogalactosyldiacylglycerol synthase
MGDG	monogalactosyldiacylglycerol
NCBI	National Center for Biotechnology Information
NiR	nitrite reductase (ferredoxin-dependent)
NR	nitrate reductase (NADH-dependent)
OCD	ornithine cyclodeaminase
OdC	ornithine decarboxylase
ODE	2-oxoglutarate dehydrogenase E1 component
OTC	ornithine carbamoyltransferase
PBS	phosphate-buffered saline
PA	phosphatidic acid
PC	pyruvate carboxylase
PDAT	phospholipid:diacylglycerol acyltransferase
PDE	pyruvate dehydrogenase E1 component subunit alpha
PDMS	poly(dimethylsiloxane)

PEPCK	phosphoenolpyruvate carboxykinase
PET	poly(ethylene terephthalate)
PFK	phosphofructokinase
PG	phosphatidylglycerol
PGAM	phosphoglycerate mutase
PGK	phosphoglycerate kinase
PGPS	phosphatidylglycerol phosphate synthase
PIPLC	phosphoinositide phospholipase C
PK	pyruvate kinase
PP	phosphatidate phosphatase
PUFA	polyunsaturated fatty acid
RAS	red/aquatic species
rDNA	ribosomal RNA gene
RPKM	reads per kilobase of exon model per million mapped reads
RT	reverse transcription
SCS	succinyl-CoA synthetase alpha subunit
SDE	succinate dehydrogenase (ubiquinone) flavoprotein subunit
shRNA	small hairpin RNA
SQD2	sulfoquinovosyltransferase
SQDG	sulfoquinovosyldiacylglycerol
TAG	triacylglycerol
TCA	tricarboxylic acid
TE	transposable element
TPI	triosephosphate isomerase
UGP	UDP-glucose pyrophosphorylase
Ure	urease
Δ5	Δ5 desaturase
Δ6	Δ-6 desaturase
Δ6-ELO	poly unsaturated elongase (Δ6)
Δ9	Δ9 desaturase
Δ12	Δ12 desaturase
Δ15	Δ15 desaturase

Supplemental References

Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. J Mol Biol **340,** 783-795.

Claros, M.G., and Vincens, P. (1996). Computational method to predict mitochondrially imported proteins and their targeting sequences. Eur J Biochem **241,** 779-786.

Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics **9,** 18.

Emanuelsson, O., Nielsen, H., Brunak, S., and von Heijne, G. (2000). Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol **300,** 1005-1016.

Flutre, T., Duprat, E., Feuillet, C., and Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. PloS one **6,** e16526.

Gonzalez, N.H., Felsner, G., Schramm, F.D., Klingl, A., Maier, U.G., and Bolte, K. (2011). A Single Peroxisomal Targeting Signal Mediates Matrix Protein Import in Diatoms. PloS one **6**.

Gruber, A., Vugrinec, S., Hempel, F., Gould, S.B., Maier, U.-G., and Kroth, P.G. (2007). Protein targeting into complex diatom plastids: functional characterisation of a specific targeting motif. Plant Mol Biol **64,** 519-530.

Gschloessl, B., Guermeur, Y., and Cock, J.M. (2008). HECTAR: a method to predict subcellular targeting in heterokonts. BMC Bioinformatics **9,** 393.

Hosokawa, M., Arakaki, A., Takahashi, M., Mori, T., Takeyama, H., and Matsunaga, T. (2009). High-density microcavity array for cell detection: single-cell analysis of hematopoietic stem cells in peripheral blood mononuclear cells. Anal Chem **81,** 5308-5313.

Hosokawa, M., Asami, M., Nakamura, S., Yoshino, T., Tsujimura, N., Takahashi, M., Nakasono, S., Tanaka, T., and Matsunaga, T. (2012). Leukocyte counting from a small amount of whole blood using a size-controlled microcavity array. Biotechnol Bioeng **109,** 2017-2024.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res **110,** 462-467.

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

Kroth, P.G., Chiovitti, A., Gruber, A., Martin-Jezequel, V., Mock, T., Parker, M.S., Stanley, M.S., Kaplan, A., Caron, L., Weber, T., Maheswari, U., Armbrust, E.V., and Bowler, C. (2008). A model for carbohydrate metabolism in the diatom *Phaeodactylum tricornutum* deduced from comparative whole genome analysis. PloS One **3,** e1426.

Larkin, M.A., Blackshields, G., Brown, N.P. Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G. (2007). Clustal

W and clustal X version 2.0. Bioinformatics **23**, 2947-8.

Nojima, D., Yoshino, T., Maeda, Y., Tanaka, M., Nemoto, M., and Tanaka, T. (2013). Proteomics analysis of oil body-associated proteins in the oleaginous diatom. J Proteome Res **12,** 5293-5301.

Sunaga, Y., Maeda, Y., Yabuuchi, T., Muto, M., Yoshino, T., and Tanaka, T. (2014). Chloroplast-targeting protein expression in the oleaginous diatom *Fistulifera solaris* JPCC DA0580 toward metabolic engineering. J Biosci Bioeng. **119**, 28-34

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol **28**, 2731-2739.

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