PNAS
www.pnas.org

Supplementary Information for

Principles of plastid reductive evolution illuminated by non-photosynthetic chrysophytes

Richard G. Dorrell[1+], Tomonori Azuma[2], Mami Nomura[2], Guillemette Audren de Kerdrel[1], Lucas Paoli[1], Shanshan Yang[3], Chris Bowler[1], Ken-ichiro Ishii[2], Hideaki Miyashita[2], Gillian Gile[3], and Ryoma Kamikawa[2+]
[1]IBENS, Département de Biologie, École Normale Supérieure, CNRS, Inserm, PSL Research University, Paris, France
[2]Graduate School of Human and Environmental Studies, Kyoto University, Kyoto, Japan
[3]School of Life Sciences, Arizona State University, Tempe, United States

Paste corresponding author name here
Email: dorrell@biologie.ens.fr , kamikawa.ryoma.7v@kyoto-u.ac.jp

This PDF file includes:

Figs. S1 to S22

www.pnas.org/cgi/doi/10.1073/pnas.1819976116

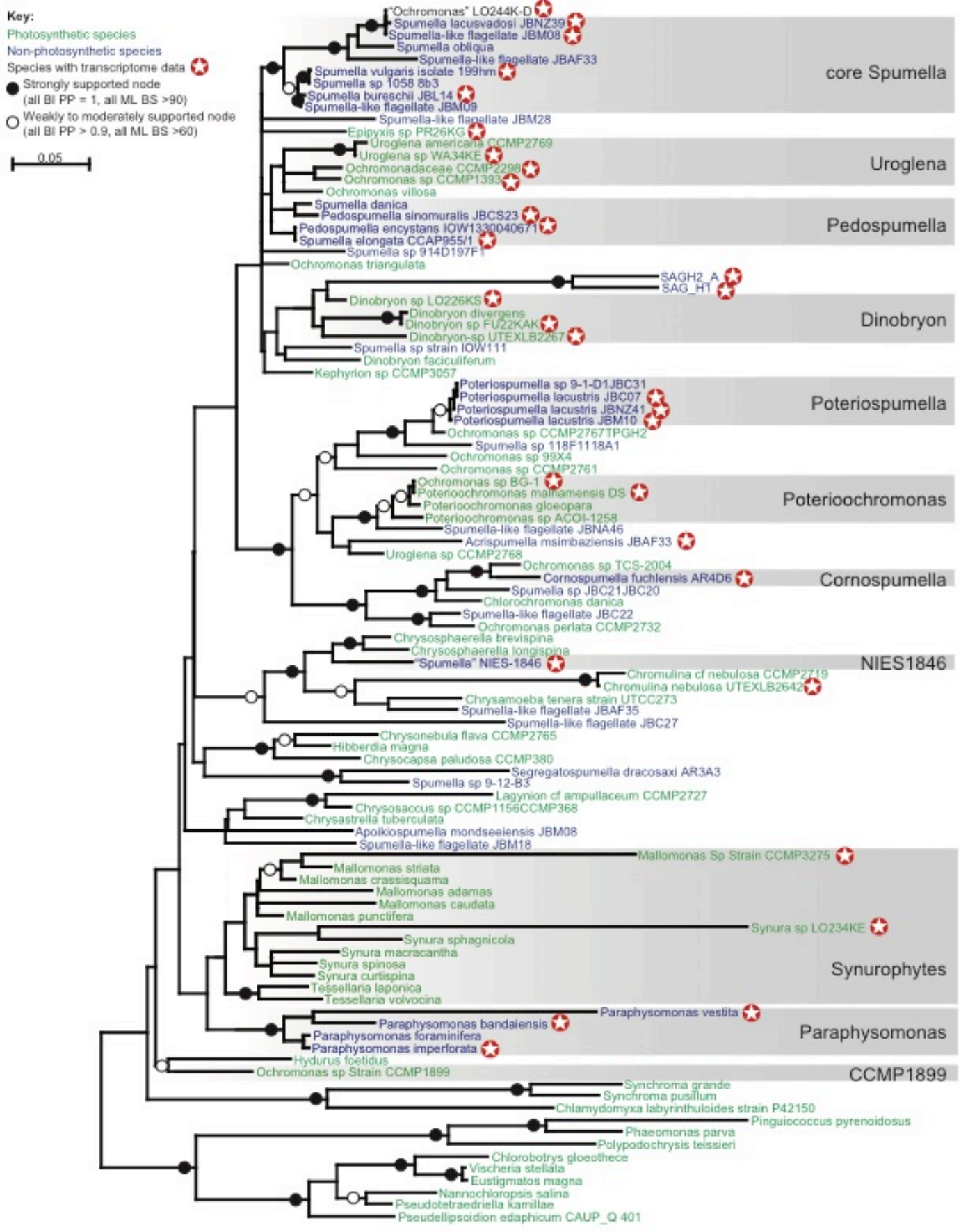


Fig. S1. Multiple independent losses of photosynthesis in chrysophytes.
This tree shows the RAxML consensus topology inferred for a 96 taxa alignment of 18S sequences from cultured « PESC clade » members, trimmed to include only sites with >50% (1734 nt), >80% (1619 nt), and >90% (1527 nt) occupancy. The tree topology is artificially rooted betwen the pinguiophyte/ eustigmatophyte/ synchromophyte outgroup, and all chrysophyte sequences. Taxon names are coloured by life-strategy as per fig. 1; taxa with transcriptome or genome sequence data are asterisked; clades corresponding to those in fig. 1 are shaded and labelled.

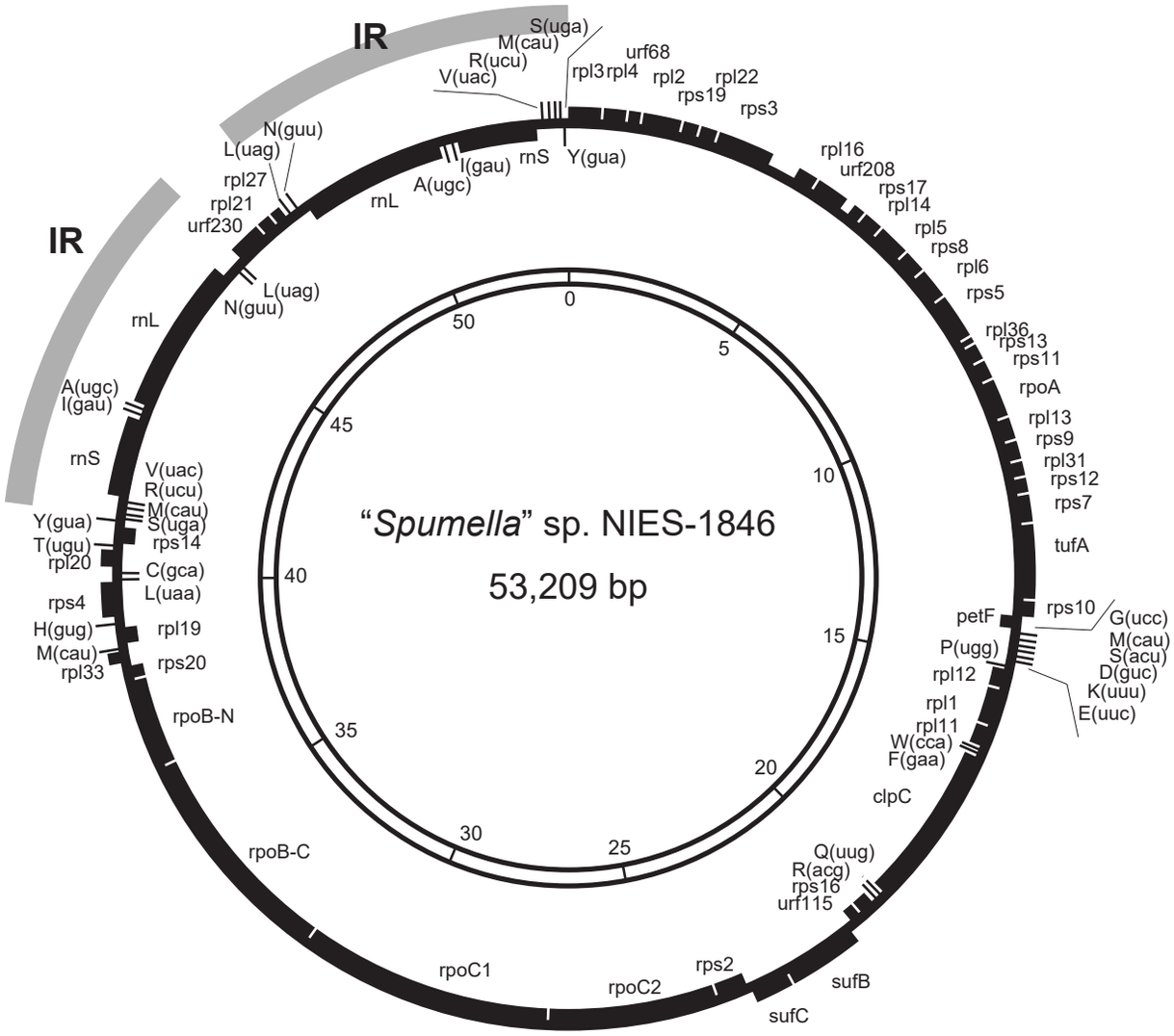


Fig. S2. Circular map of the "*Spumella*" sp. NIES-1846 plastid genome

Fig. S3. Broad comparison of deduced functions encoded in non-photosynthetic plastid genomes. Data of *Euglena longa*, green algae, plants, and red alga are based on Hadariová et al. (2018) and Su et al. (in press). Red alga-derived plastid lineages are also shown in Fig. 2 panel C. Other details are described in the legend of Fig. 2.

Hadariová L, Vesteg M, Hampl V, Krajčovič J. 2018. Reductive evolution of chloroplasts in non-photosynthetic plants, algae and protists. Curr Genet 64:365–387.

Su H, Barkman T, Hao W, Jones SS, Naumann J, Skippington E, Wafula EC, Hu J, Palmer JD, dePamphilis CW. Novel genetic code and record-setting AT-richness in the highly reduced plastid genome of the holoparasitic plant *Balanophora*. Proc Natl Acad Sci USA in press

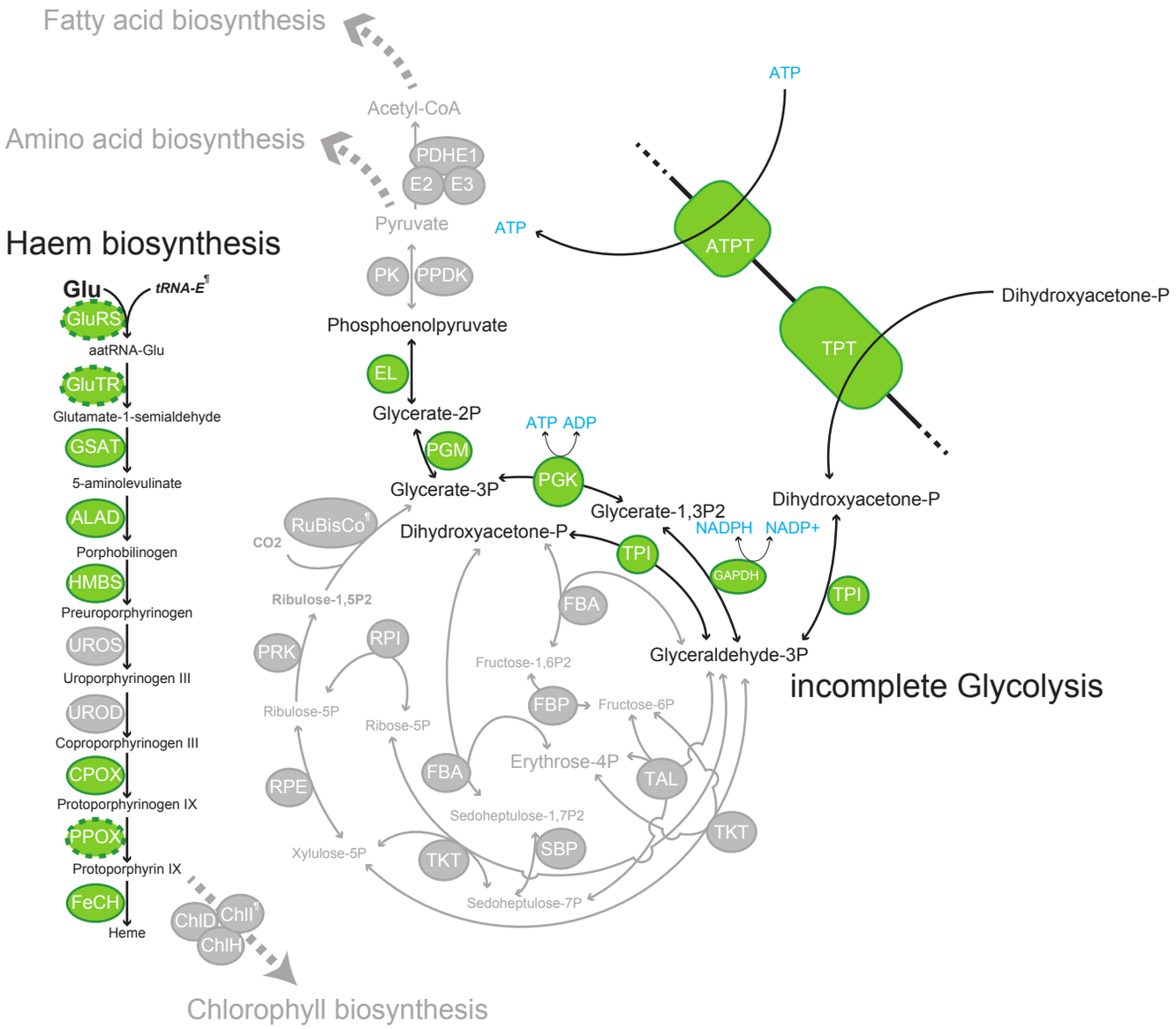


Fig. S4. Main metabolism of non-photosynthetic plastids in "*Spumella*" sp. NIES-1846.

Proteins with plastid-targeting signals are shown as closed circles in green. Proteins not found in the transcriptome data are shown in grey. ¶: plastid genome-encoded proteins. RuBisCO large and small subunits were not present in both plastid genome and transcriptome. ATP/ADP and NADP+/NADPH are highlighted in light blue. EL, enolase; FBA, fructose 1;6-bisphosphate aldolase; FBP, fructose-1,6-bisphosphatase; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; PGK, phosphoglycerate kinase; PGM, phosphoglycerate mutase; PK, pyruvate kinase; PPDK, pyruvate, phosphate dikinase; PRK, phosphoribulokinase; RPE, ribulose 5-phosphate 3-epimerase; RPI, ribose 5-phosphate isomerase; RuBisCO, ribulose 1,5-bisphsphate carboxylase/oxygenase; TAL, transaldolase; TKL, transketolase; TPI, triose phosphate isomerase. PDH, pyruvate dehydrogenase. No metabolic pathway related to chlorophyll, amino acid, and fatty acid biosynthesis are not confidently predicted to locate in the plastids. In contrast, glycolysis and haem biosynthesis are predicted to be located in the plastids. ATP transporter and triose phosphate (TP) transporters are detected in the transcriptome data, indicating the potential of ATP and TP import into the plastids for the plastid metabolic pathways.

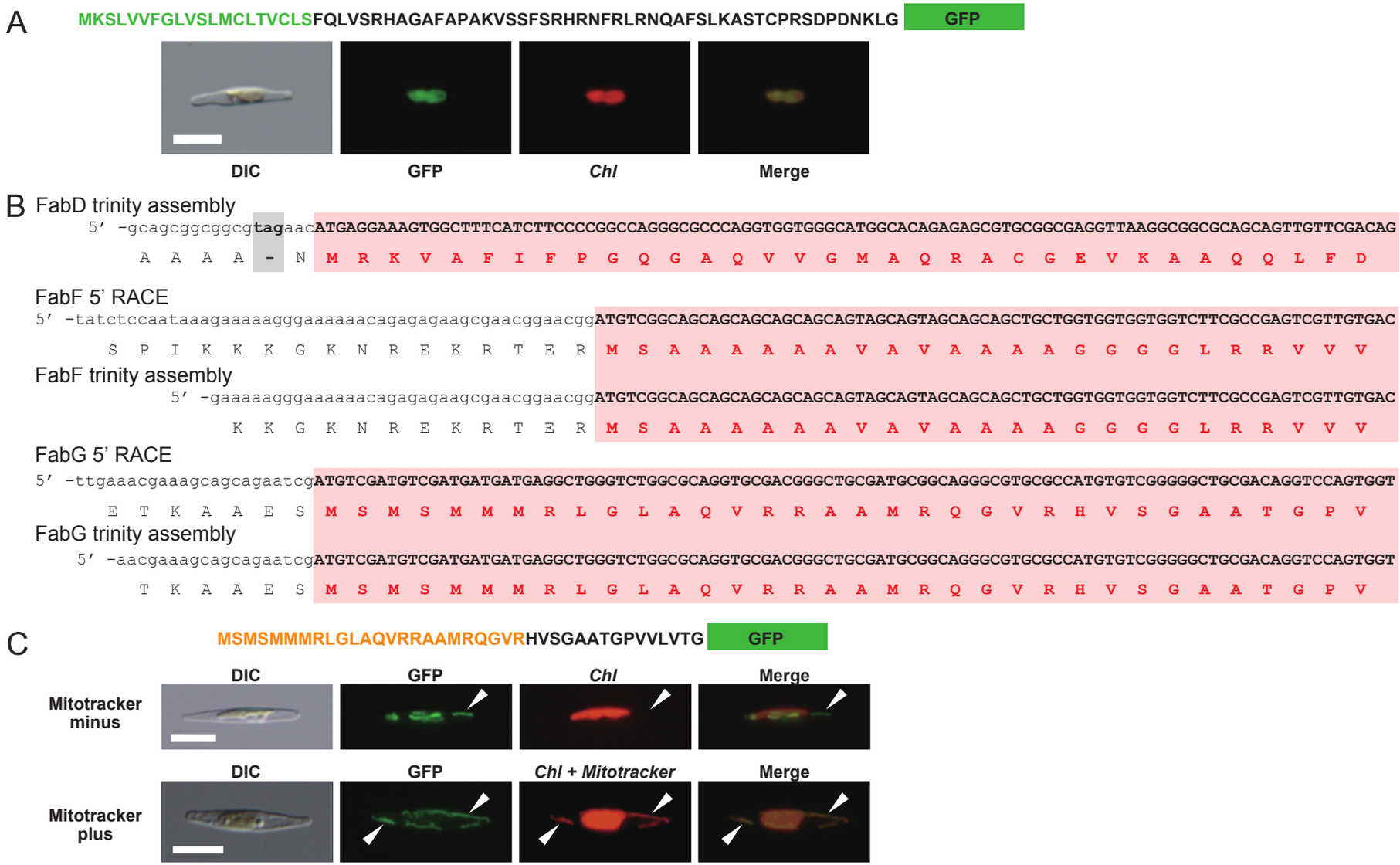


Fig. S5 Localization of FeCH and three Fab proteins of “*Spumella*” sp. NIES-1846

A. N-terminal amino acid sequence of “*Spumella*” FeCH-GFP recombinant protein and localization in the diatom *Phaeodactylum tricornutum*. GFP fluorescence is co-localized with Chlorophyll fluorescence, indicative of the plastid localization of “*Spumella*” FeCH. The deduced signal peptide is highlighted in green. B. Confirmation of 5’ terminal sequences of three fab genes of “*Spumella*” sp. NIES-1846. Deduced coding regions are highighted in red. In-frame termination codon is highighted in grey. Given the in-frame termination codon in the 5’ UTR region, 5’ RACE analysis for fabD was not conducted. The “ATG” codons deduced as the initiation codons here are actually the first methionine codons, confirmed by 5’ RACE analyses. This supports that the “*Spumella*” Fab proteins do not have N-terminal plastid tergeting signals. C. N-terminal amino acid sequence of “*Spumella*” FabG-GFP recombinant protein and localization in *P. tricornutum*. GFP fluorescence is not co-localized with Chlorophyll fluorescence, but co-localized with Mitotracker fluorescence, indicative of mitochondrial localization of “*Spumella*” FabG. Deduced mitochondrial targeting peptide is highlighted in orange. Bar = 10 μm

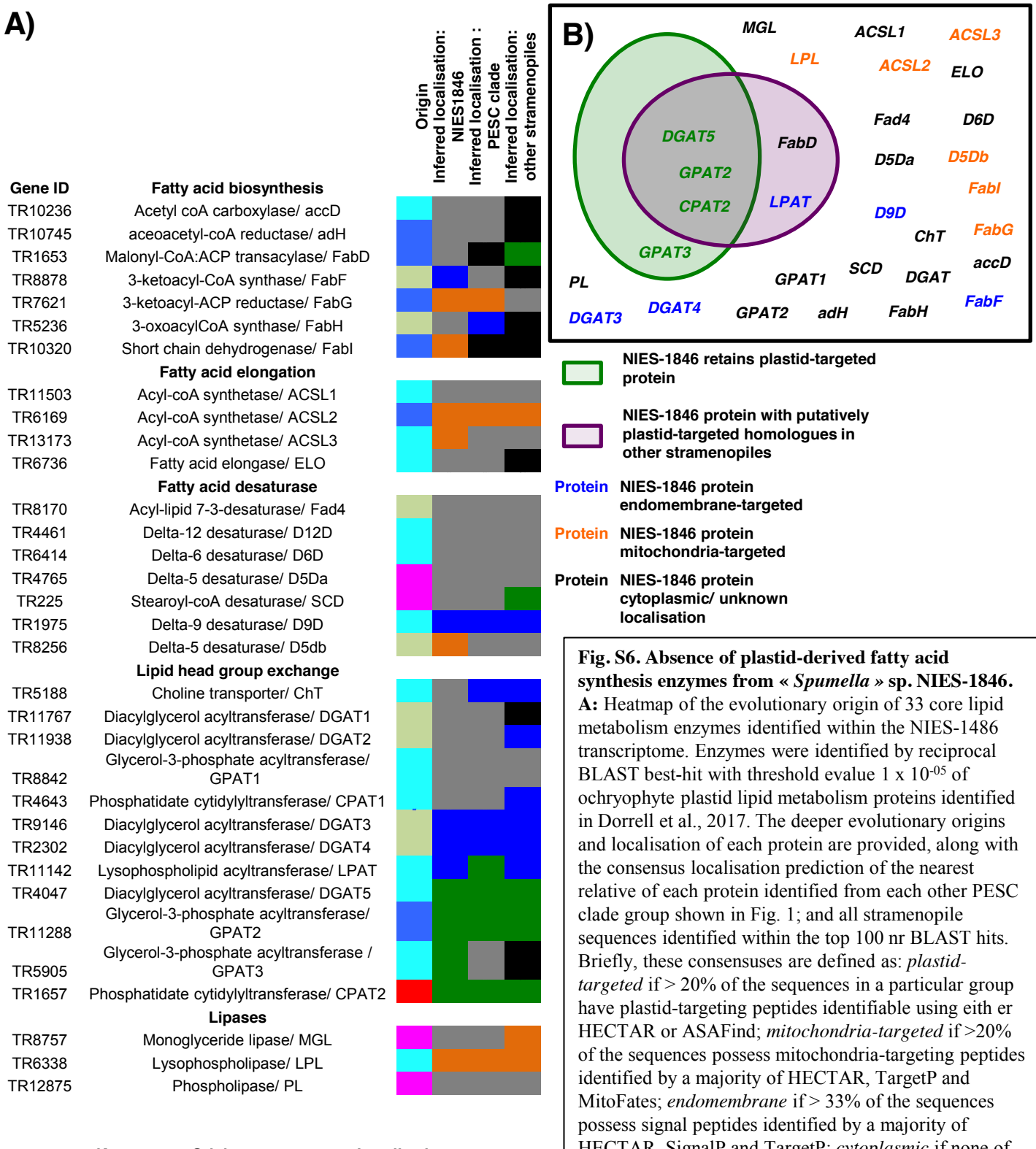


Fig. S6. Absence of plastid-derived fatty acid synthesis enzymes from « *Spumella* » sp. NIES-1846. **A:** Heatmap of the evolutionary origin of 33 core lipid metabolism enzymes identified within the NIES-1486 transcriptome. Enzymes were identified by reciprocal BLAST best-hit with threshold evalue 1 x 10^{-05} of ochryophyte plastid lipid metabolism proteins identified in Dorrell et al., 2017. The deeper evolutionary origins and localisation of each protein are provided, along with the consensus localisation prediction of the nearest relative of each protein identified from each other PESC clade group shown in Fig. 1; and all stramenopile sequences identified within the top 100 nr BLAST hits. Briefly, these consensuses are defined as: *plastid-targeted* if > 20% of the sequences in a particular group have plastid-targeting peptides identifiable using eith er HECTAR or ASAFind; *mitochondria-targeted* if >20% of the sequences possess mitochondria-targeting peptides identified by a majority of HECTAR, TargetP and MitoFates; *endomembrane* if > 33% of the sequences possess signal peptides identified by a majority of HECTAR, SignalP and TargetP; *cytoplasmic* if none of the above criteria are met; and *ambiguous* if more than one of the criteria are met, or if the group is empty. **B:** Venn diagram summarising the results. « *Spumella* » *sp. NIES1846* does not retain enzymes associated with the plastid fatty acid synthesis pathway, with all of the cytoplasmic, mitochondrial and endomembrane-targeted fatty acid synthesis, elongation and desaturation enzymes identified corresponding to enzymes that function in non-plastid compartments in other PESC clade members, and in stramenopiles as a whole.

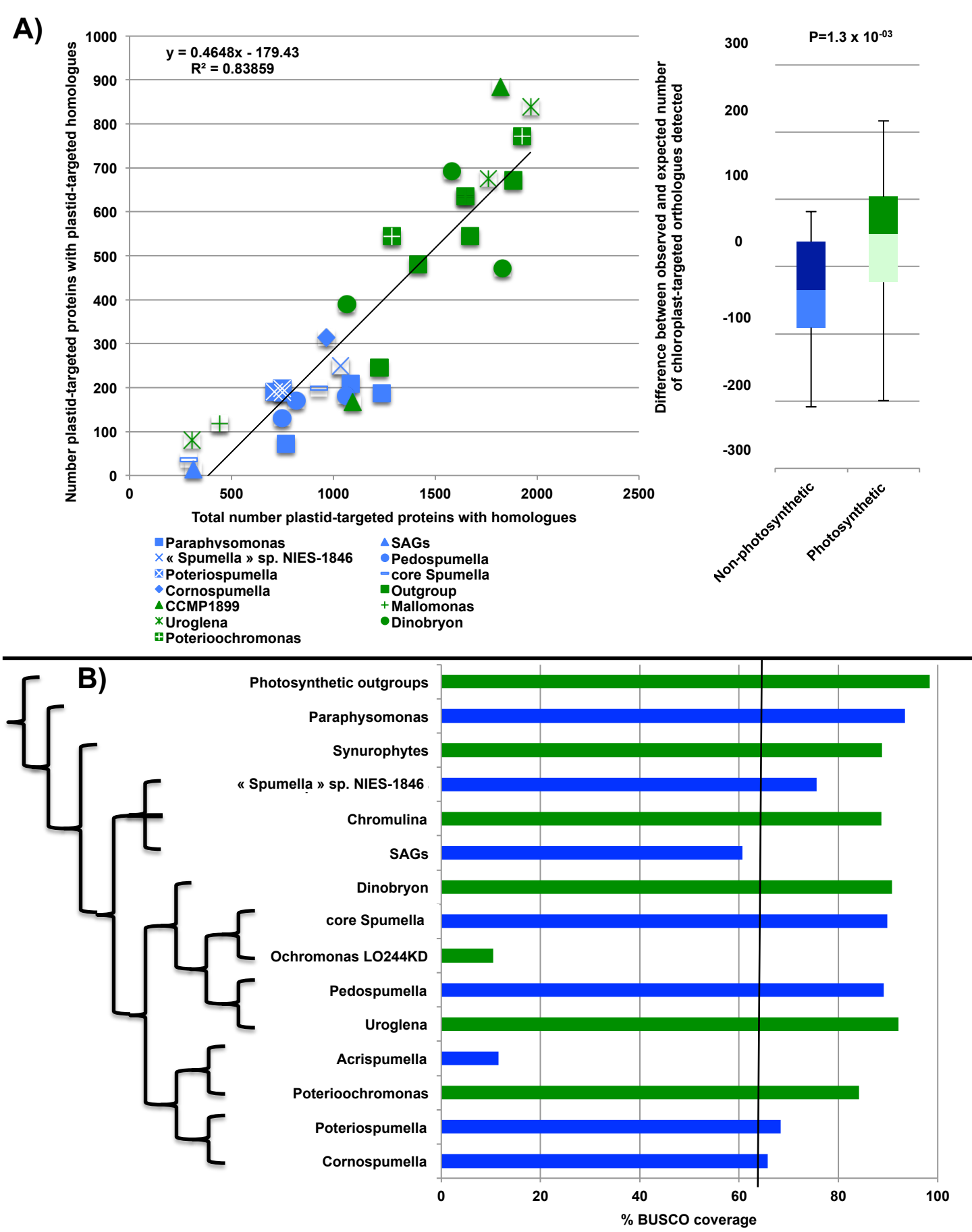


Fig. S7. Features of published PESC clade transcriptomes. A: (left) scatterplot showing the number of query plastid-targeted proteins for which any homologue, or a homologue with a defined plastid-targeted sequence, could be identified using BLAST, for all species for which >250 orthologues in total could be found (Text S1) **(right)** boxplot showing the differences in the number of observed homologues with plastid-targeted proteins, and the number expected from linear regression against the total number of homologues, for photosynthetic (blue) and non-photosynthetic species. **B:** the total BUSCO coverage obtained for different groups of photosynthetic (blue) and non-photosynthetic (red) PESC clade species, defined using the topology in fig. 1. The vertical line shows the 65% coverage threshold for comparative plastid proteome analysis.

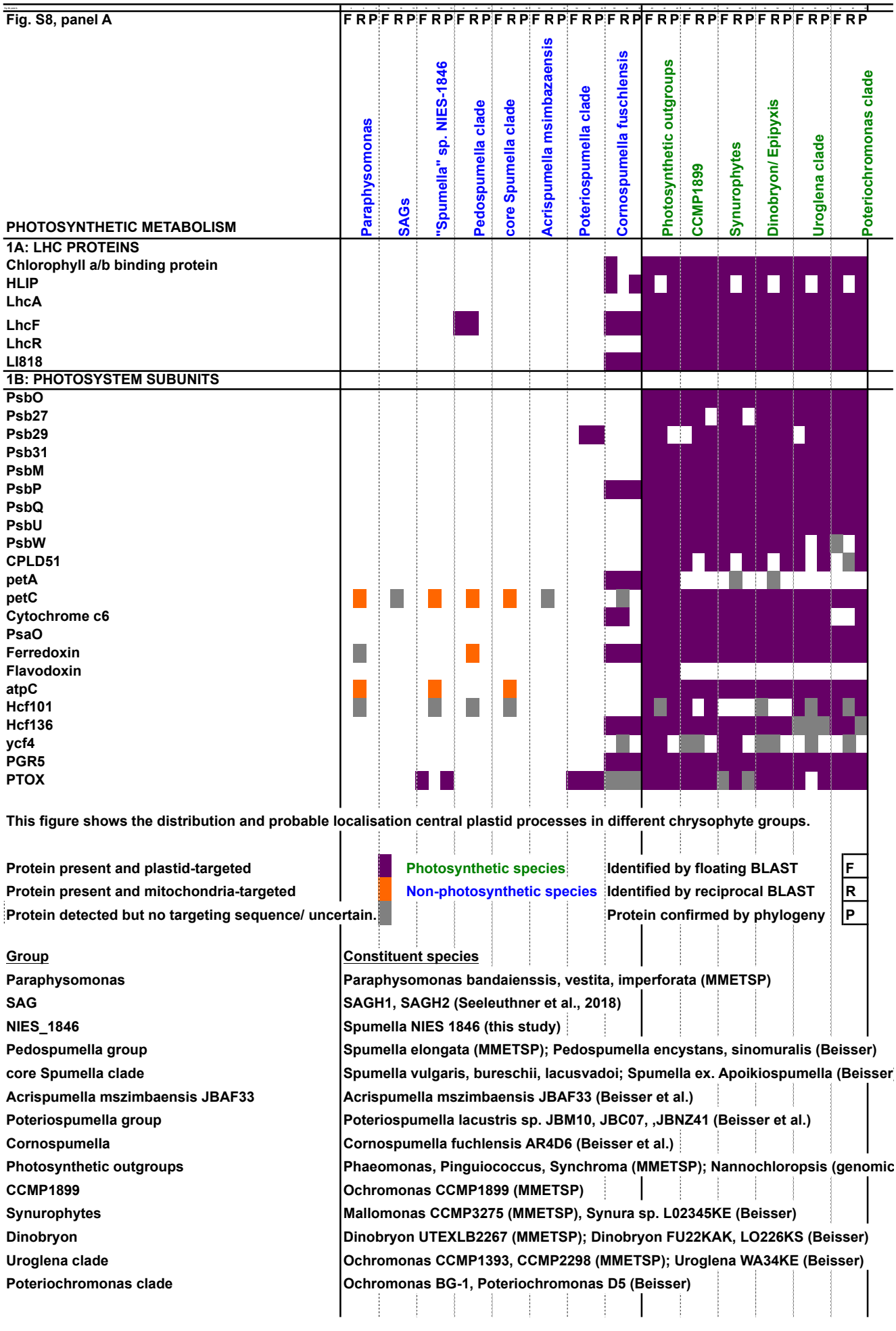


Fig. S8, panel A
F R P F R P F R P F R P F R P F R P F R P F R P F R P F R P F R P F R P F R P
Paraphysomonas
SAGs
"Spumella" sp. NIES-1846
Pedospumella clade
core Spumella clade
Acrispumella msimbazaensis
Poteriospumella clade
Cornospumella fuschlensis
Photosynthetic outgroups
CCMP1899
Synurophytes
Dinobryon/ Epipyxis
Uroglena clade
Poteriochromonas clade
PHOTOSYNTHETIC METABOLISM
1A: LHC PROTEINS
Chlorophyll a/b binding protein
HLIP
LhcA
LhcF
LhcR
LI818
1B: PHOTOSYSTEM SUBUNITS
PsbO
Psb27
Psb29
Psb31
PsbM
PsbP
PsbQ
PsbU
PsbW
CPLD51
petA
petC
Cytochrome c6
PsaO
Ferredoxin
Flavodoxin
atpC
Hcf101
Hcf136
ycf4
PGR5
PTOX
This figure shows the distribution and probable localisation central plastid processes in different chrysophyte groups.
Protein present and plastid-targeted
Protein present and mitochondria-targeted
Protein detected but no targeting sequence/ uncertain.
Photosynthetic species
Non-photosynthetic species
Identified by floating BLAST
Identified by reciprocal BLAST
Protein confirmed by phylogeny
F
R
P
Group
Constituent species
Paraphysomonas
Paraphysomonas bandaienssis, vestita, imperforata (MMETSP)
SAG
SAGH1, SAGH2 (Seeleuthner et al., 2018)
NIES_1846
Spumella NIES 1846 (this study)
Pedospumella group
Spumella elongata (MMETSP); Pedospumella encystans, sinomuralis (Beisser)
core Spumella clade
Spumella vulgaris, bureschii, lacusvadoi; Spumella ex. Apoikiospumella (Beisser
Acrispumella mszimbaensis JBAF33
Acrispumella mszimbaensis JBAF33 (Beisser et al.)
Poteriospumella group
Poteriospumella lacustris sp. JBM10, JBC07, ,JBNZ41 (Beisser et al.)
Cornospumella
Cornospumella fuchlensis AR4D6 (Beisser et al.)
Photosynthetic outgroups
Phaeomonas, Pinguiococcus, Synchroma (MMETSP); Nannochloropsis (genomic
CCMP1899
Ochromonas CCMP1899 (MMETSP)
Synurophytes
Mallomonas CCMP3275 (MMETSP), Synura sp. L02345KE (Beisser)
Dinobryon
Dinobryon UTEXLB2267 (MMETSP); Dinobryon FU22KAK, LO226KS (Beisser)
Uroglena clade
Ochromonas CCMP1393, CCMP2298 (MMETSP); Uroglena WA34KE (Beisser)
Poteriochromonas clade
Ochromonas BG-1, Poteriochromonas D5 (Beisser)

Fig. S8, panel B

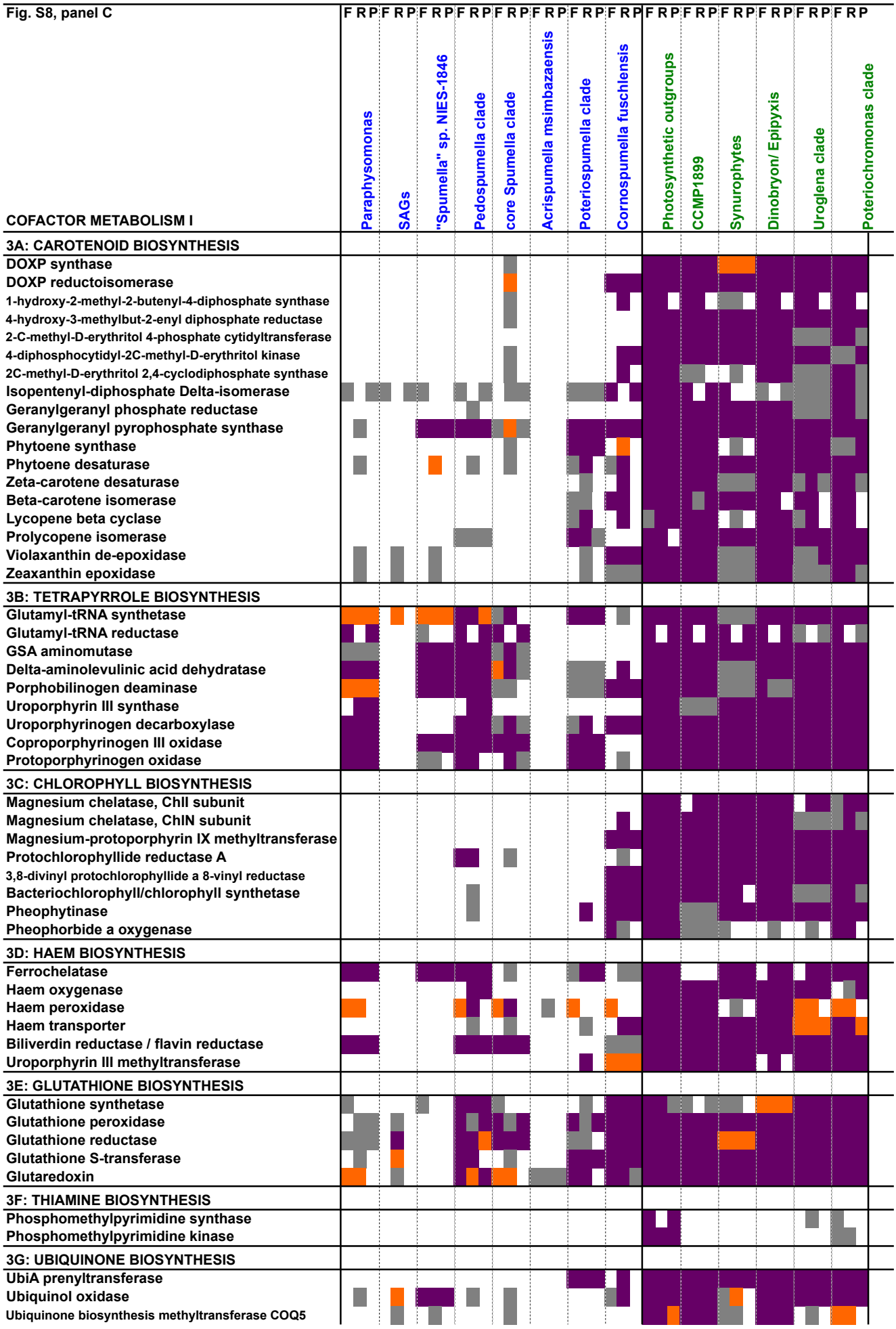


Fig. S8, panel C
COFACTOR METABOLISM I
F R P
Paraphysomonas
SAGs
"Spumella" sp. NIES-1846
Pedospumella clade
core Spumella clade
Acrispumella msimbazaensis
Poteriospumella clade
Cornospumella fuschlensis
Photosynthetic outgroups
CCMP1899
Synurophytes
Dinobryon/ Epipyxis
Uroglena clade
Poteriochromonas clade
3A: CAROTENOID BIOSYNTHESIS
DOXP synthase
DOXP reductoisomerase
1-hydroxy-2-methyl-2-butenyl-4-diphosphate synthase
4-hydroxy-3-methylbut-2-enyl diphosphate reductase
2-C-methyl-D-erythritol 4-phosphate cytidyltransferase
4-diphosphocytidyl-2C-methyl-D-erythritol kinase
2C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
Isopentenyl-diphosphate Delta-isomerase
Geranylgeranyl phosphate reductase
Geranylgeranyl pyrophosphate synthase
Phytoene synthase
Phytoene desaturase
Zeta-carotene desaturase
Beta-carotene isomerase
Lycopene beta cyclase
Prolycopene isomerase
Violaxanthin de-epoxidase
Zeaxanthin epoxidase
3B: TETRAPYRROLE BIOSYNTHESIS
Glutamyl-tRNA synthetase
Glutamyl-tRNA reductase
GSA aminomutase
Delta-aminolevulinic acid dehydratase
Porphobilinogen deaminase
Uroporphyrin III synthase
Uroporphyrinogen decarboxylase
Coproporphyrinogen III oxidase
Protoporphyrinogen oxidase
3C: CHLOROPHYLL BIOSYNTHESIS
Magnesium chelatase, ChlI subunit
Magnesium chelatase, ChlN subunit
Magnesium-protoporphyrin IX methyltransferase
Protochlorophyllide reductase A
3,8-divinyl protochlorophyllide a 8-vinyl reductase
Bacteriochlorophyll/chlorophyll synthetase
Pheophytinase
Pheophorbide a oxygenase
3D: HAEM BIOSYNTHESIS
Ferrochelatase
Haem oxygenase
Haem peroxidase
Haem transporter
Biliverdin reductase / flavin reductase
Uroporphyrin III methyltransferase
3E: GLUTATHIONE BIOSYNTHESIS
Glutathione synthetase
Glutathione peroxidase
Glutathione reductase
Glutathione S-transferase
Glutaredoxin
3F: THIAMINE BIOSYNTHESIS
Phosphomethylpyrimidine synthase
Phosphomethylpyrimidine kinase
3G: UBIQUINONE BIOSYNTHESIS
UbiA prenyltransferase
Ubiquinol oxidase
Ubiquinone biosynthesis methyltransferase COQ5

Fig. S8, panel D

Fig. S8, panel E

Fig. S8, panel F

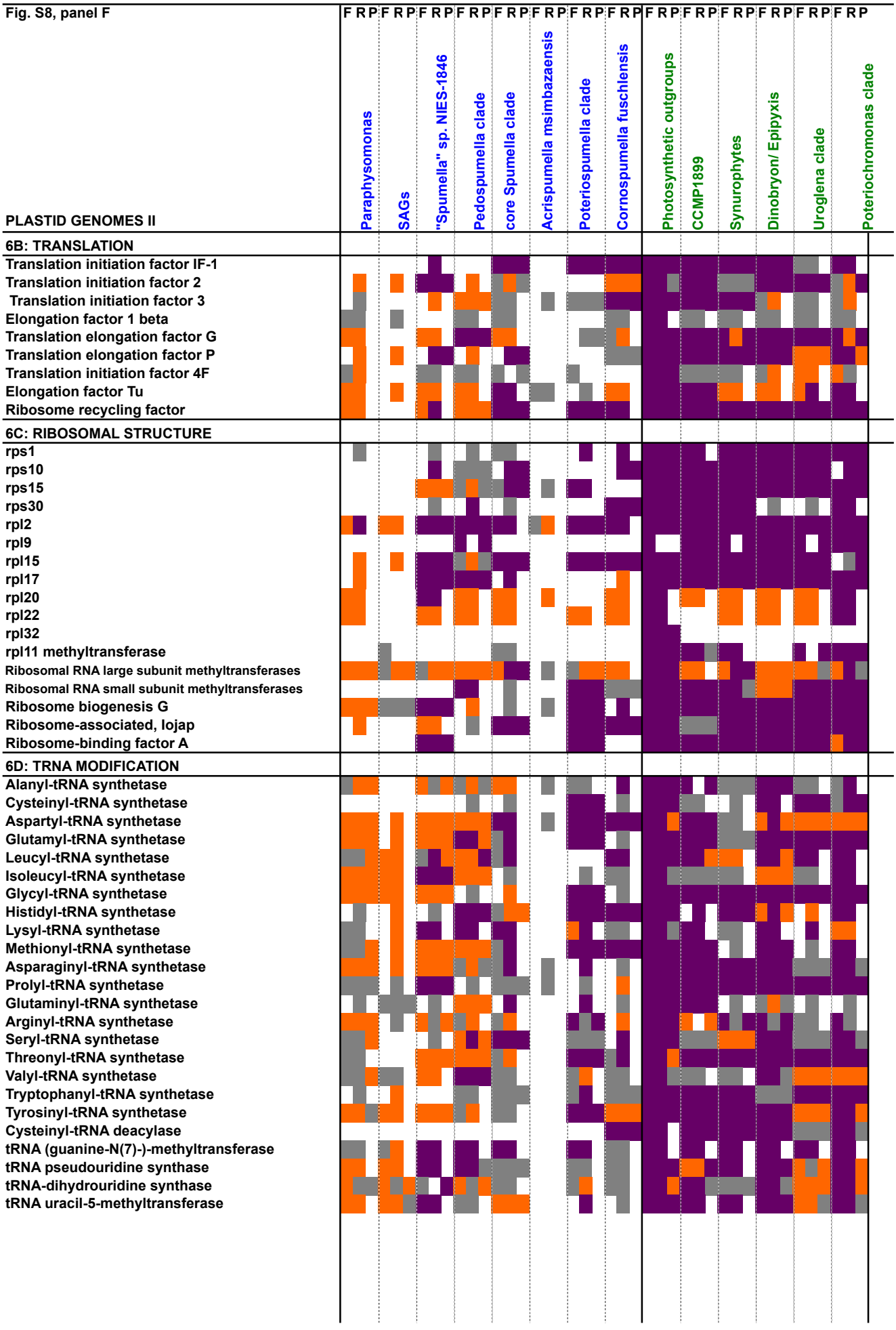


Fig. S8, panel G

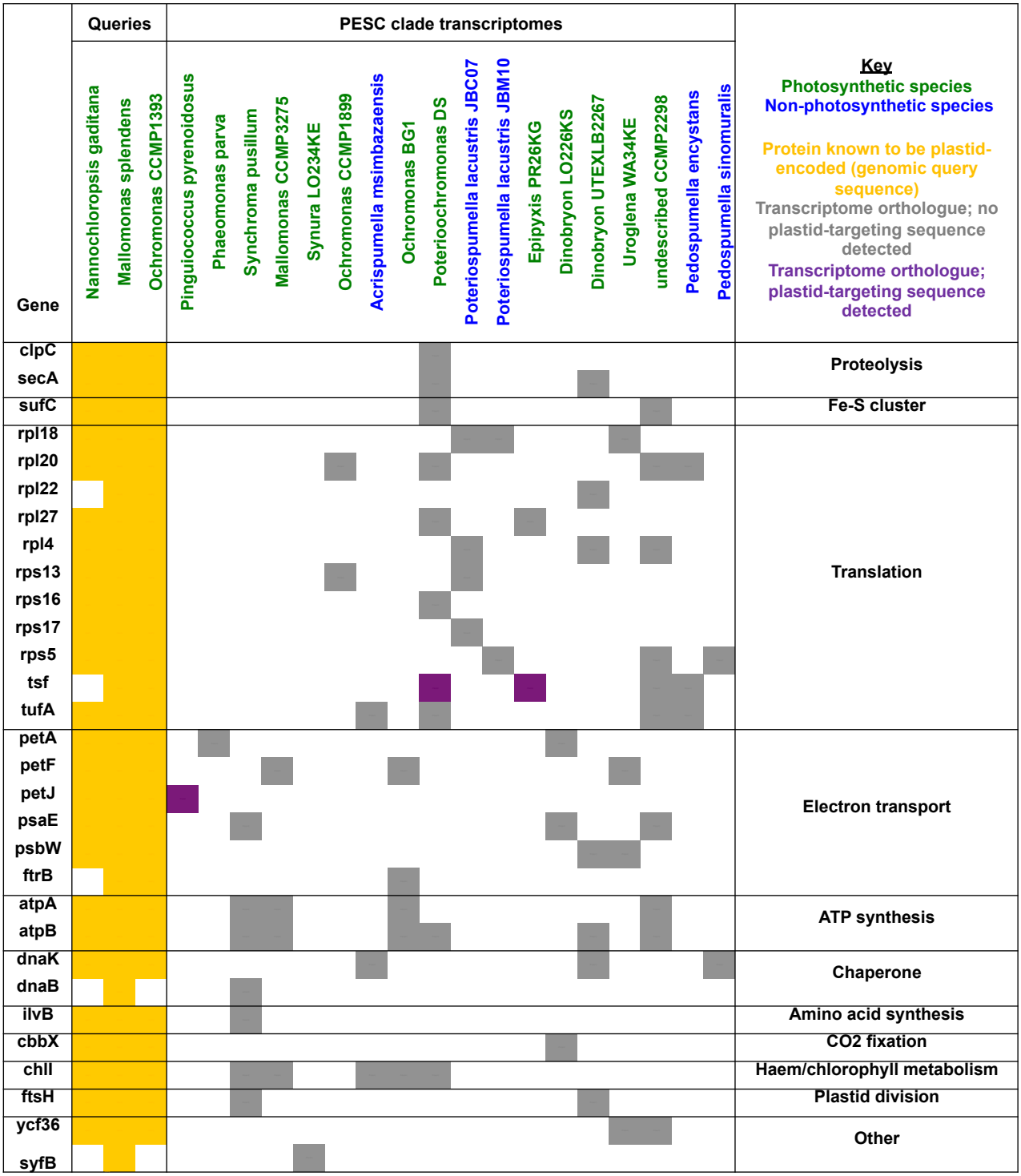
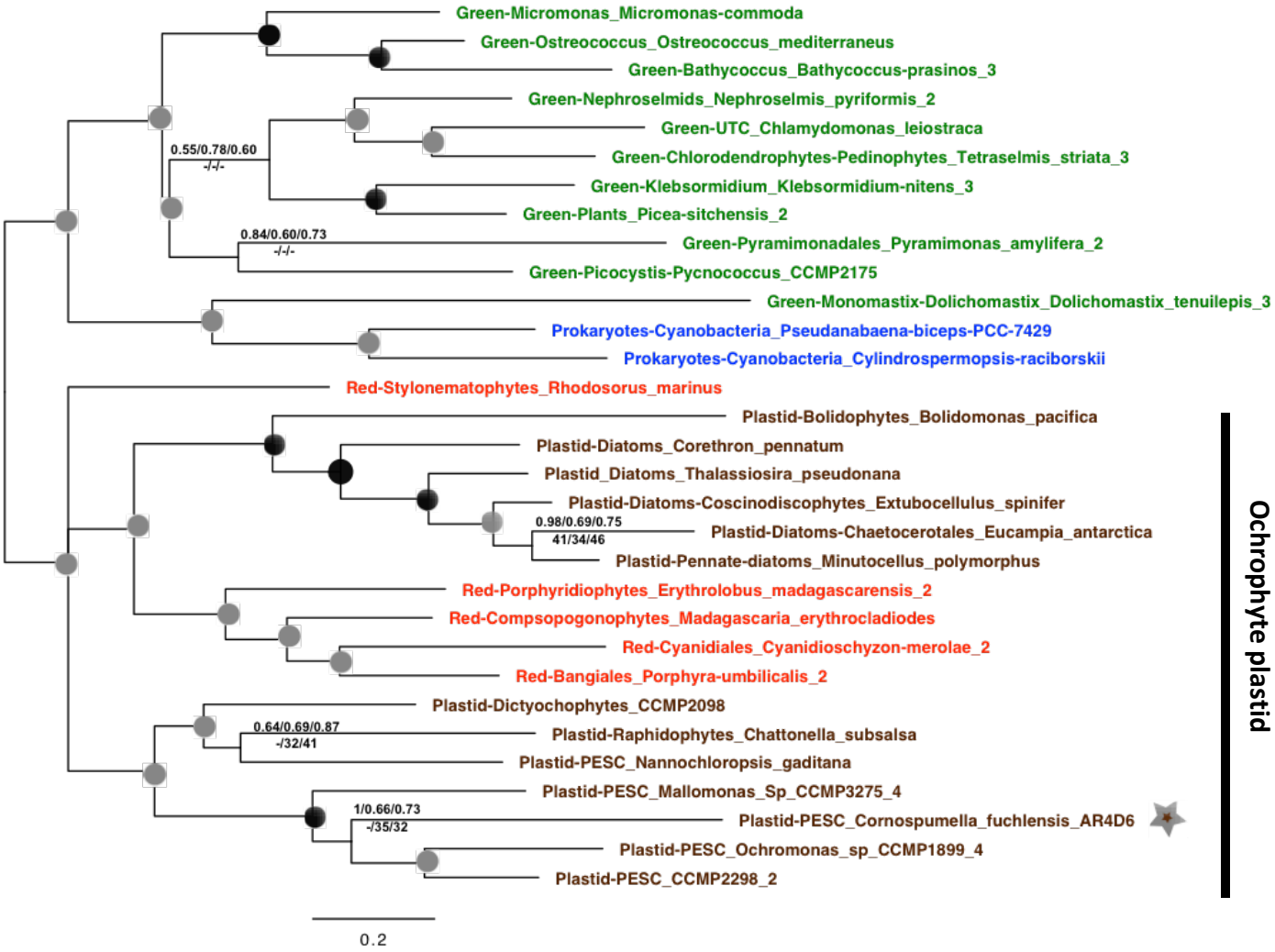


Fig. S9. Plastid-encoded functions detected in PESC clade transcriptomes.
This heatmap shows the distribution of sequences orthologous to genes in published chrysophyte plastid genomes that are detectable through reciprocal BLAST and alignment in published PESC clade transcriptomes. The left hand three columns show the distribution of these proteins encoded in three completed PESCclade plastid genome sequences Some of the sequences identified may correspond to fragments of plastid-encoded transcripts that survived poly(A) selection within the corresponding transcriptome; others may correspond to nucleus-encoded transcripts of plastid origin, identifiable by the presence of N-terminal targeting sequences. Plastid-encoded genes, and taxa for which no such orthologous sequences could be found (e.g. *Paraphysomonas sp.*) are not shown.



Key

Ochrophyte plastid-targeted protein
Red algae
Green algae
Prokaryote

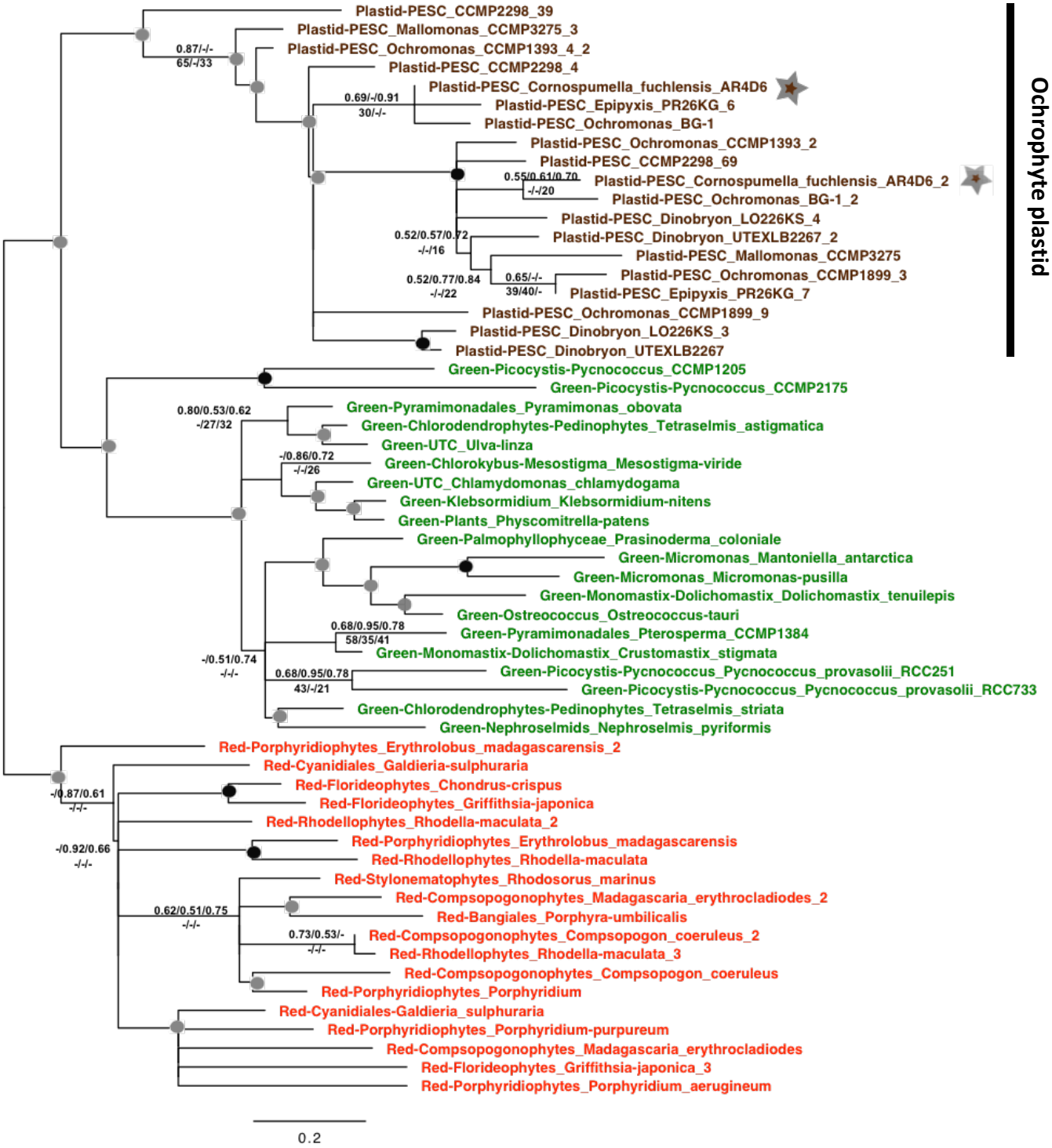
Plastid-derived protein retained in non-photosynthetic chrysophyte

Node with strong support (Bayesian PP = 1.0, and RAxML best tree bootstrap support > 80% in all alignments tested)

Node with moderate support (Bayesian PP > 0.8, or RAxML best tree bootstrap support > 50% in 2/3 alignments tested)

A/B/C Consensus support: Bayesian PP (GTR/ Jones/ WAG)
x/y/z RAxML bootstrap (GTR/ JTT/ WAG)

Fig. S10. Consensus tree of PESC clade PsbP sequences.
This tree shows the Bayesian consensus topology inferred for a 31 taxa x 96 aa alignment corresponding to a plastid-targeted PsbP-type protein identified across PESC clade members. Only proteins from ochrophytes with inferred plastid-targeting sequences; red algae, green algae, and prokaryotes are shown. Sequences are labelled by taxonomic origin, and a paraphyletic clade of ochrophyte-plastid targeted sequences are labelled with a horizontal bar. A plastid-targeted protein of clear chrysophyte origin, retained in the non-photosynthetic species *Cornospumella fuschlensis,* is asterisked.



Key

Ochrophyte plastid-targeted protein
Red algae
Green algae

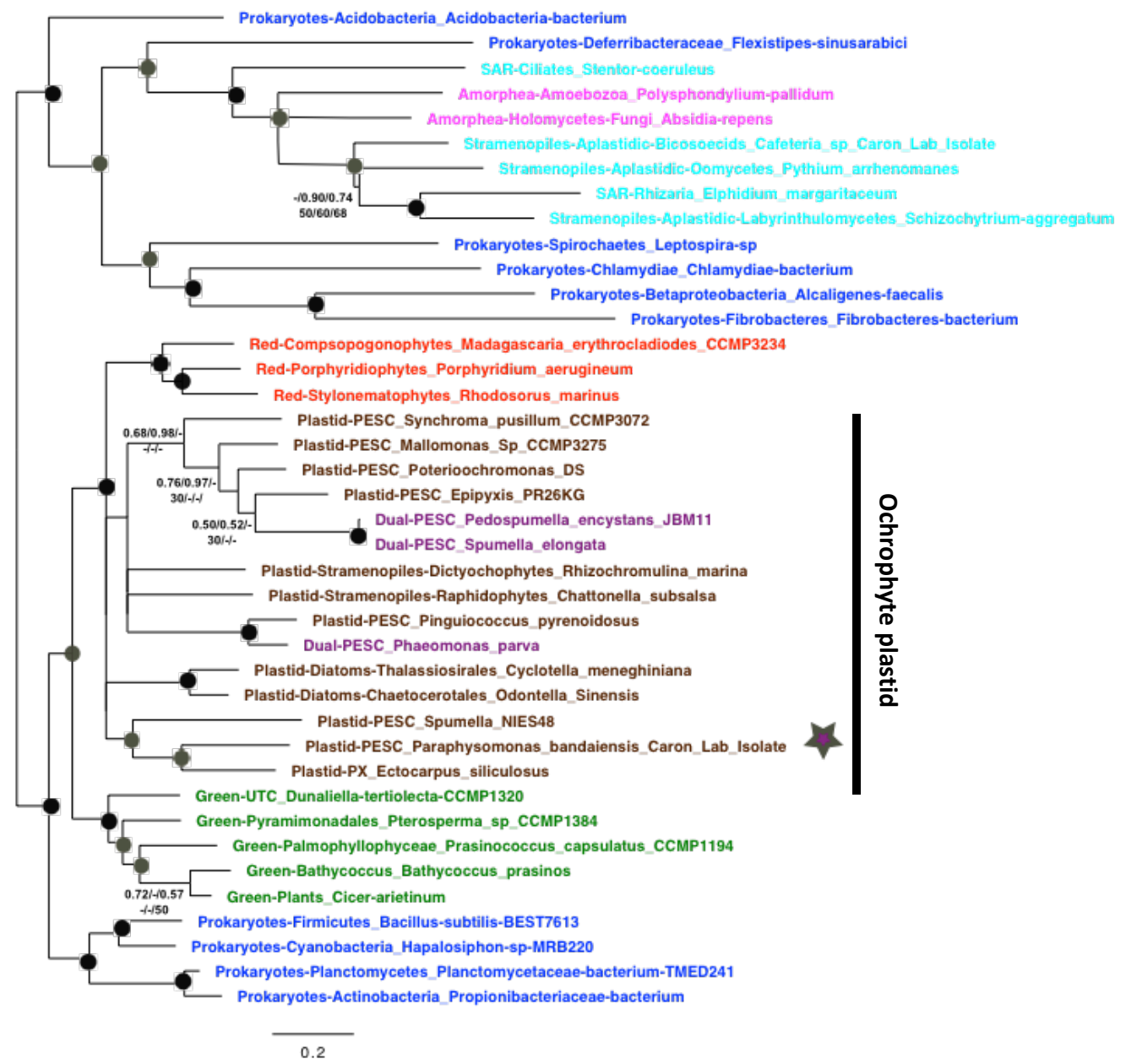
Plastid-derived protein retained in non-photosynthetic chrysophyte

Node with strong support (Bayesian PP = 1.0, and RAxML best tree bootstrap support > 80% in all alignments tested)

Node with moderate support (Bayesian PP > 0.8, or RAxML best tree bootstrap support > 50% in 2/3 alignments tested)

A/B/C Consensus support: Bayesian PP (GTR/ Jones/ WAG)
x/y/z RAxML bootstrap (GTR/ JTT/ WAG)

Fig. S11. Consensus tree of PESC clade LI818/ Lhcx sequences.
This tree shows the Bayesian consensus topology inferred for a 58 taxa x 72 aa alignment corresponding to a plastid-targeted Lhcx/Li818-type protein identified across PESC clade members, shown as per fig. S10.



Key

Ochrophyte plastid-targeted protein
Ochrophyte dual plastid/ mitochondria-targeted protein
Red algae
Green algae
Prokaryote
Plastid-lacking SAR clade member
Other aplastidic eukaryote

★ **Plastid-derived protein retained in *Paraphysomonas* sp.**

● **Node with strong support (Bayesian PP = 1.0, and RAxML best tree bootstrap support > 80% in all alignments tested)**

● **Node with moderate support (Bayesian PP > 0.8, or RAxML best tree bootstrap support > 50% in 2/3 alignments tested)**

A/B/C Consensus support: Bayesian PP (GTR/ Jones/ WAG)
x/y/z RAxML bootstrap (GTR/ JTT/ WAG)

Fig. S12. Consensus tree of PESC ferrochelatase sequences
This tree shows the Bayesian consensus topology inferred for a 40 taxa x 318 aa alignment corresponding to ferrochelatase sequences from across the tree of life. Ochrophyte sequences are shaded by predicted subcellular localisation, and all remaining sequences are shade by taxonomy. The ochrophyte sequences, which include a plastid-targeted protein from *Paraphysomonas*, resolve as a monophyletic clade with other plastid-bearing eukaryotes, separate to the mitochondrial/ cytoplasmic-type enzymes found in plastid-lacking members of the SAR clade (oomycetes, ciliates, rhizarians) and opisthokonts.

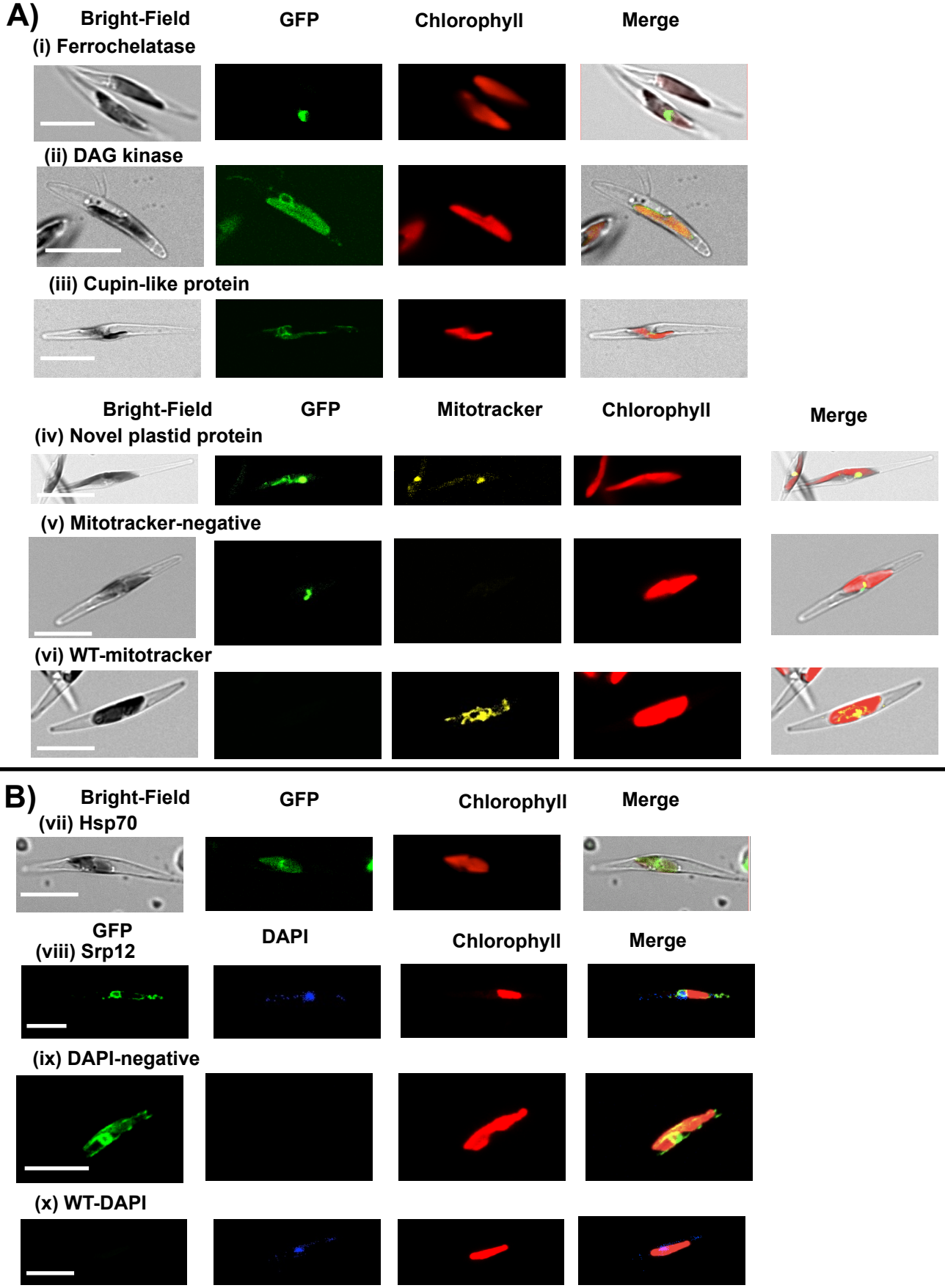


Fig. S13. Exemplar plastid-targeted proteins in *Paraphysomonas*.
A *:* images of *Phaeodactylum* lines transformed with GFP-linked constructs of the N-terminal regions of proteins associated with *Paraphysomonas bandaiensis* plastid metabolism*:* (i) Ferrochelatase (haem synthesis), (ii) DAG kinase (lipid metabolism), (iii, iv) two novel protein widely conserved across ochrophyte plastid proteomes (Dorrell et al., 2017). Each construct localises to the periplastid compartment (i) other regions within the *Phaeodactylum* plastid (ii, iii); or dual localises to the plastid and mitochondria (iv), as verified by Mitotracker Orange (v-vi). **B:** heterologous expression GFP-linked constructs of the N-terminal regions of *Paraphysomonas bandaiensis* plastid protein import subunits: Hsp70, which localises to the periplastid compartment (vii), and Srp12 (viii)**,** which localises to the plastid endoplasmic reticulum, as verified with DAPI staining (ix, x). Scale bars are to 10 µm**.**

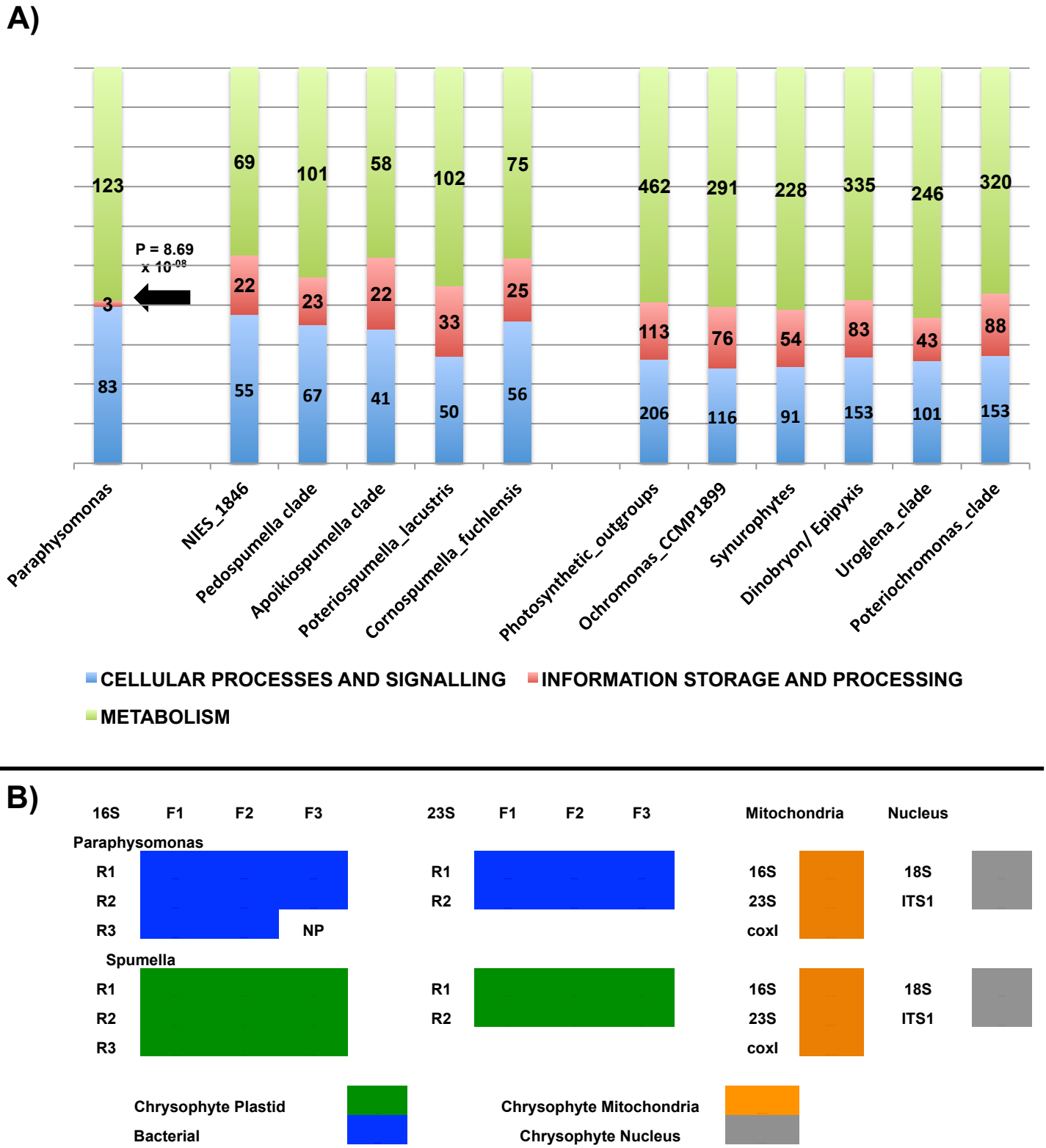


Fig. S14. Evidence for loss of the *Paraphysomonas* plastid genome.
A: KOG family distribution of plastid-targeted orthologues of different ochrophyte plastid-targeted proteins identified in PESC clade members. *Paraphysomonas* possesses substantially fewer proteins with potential KOG functions associated with information storage and processing related functions than any other PESC clade group considered. **B:** results of PCRs using consensus primers designed against chrysophyte plastid, mitochondrial, and nuclear genomes, for *Paraphysomonas bandaiensis* RCC383 and *Spumella elongata* CCAP955/1. Products are shaded by evolutionary origin as inferred by BLAST; « NP » indicates no product was obtained for a given reaction even with reduced annealing temperatures (>15°C below the primer melt temperatures) and two successive rounds of PCR, using the primary reaction product as a template for the second round of amplification. Although mitochondrial and nuclear DNA could be amplified for both species, a plastid genome was only identifiable for *S. elongata* PCRs of plastid contigs for *P. bandaiensis* only yielded bacterial (*Marinobacter, Labrenzia*) contaminants, consistent with an absence of plastid DNA.

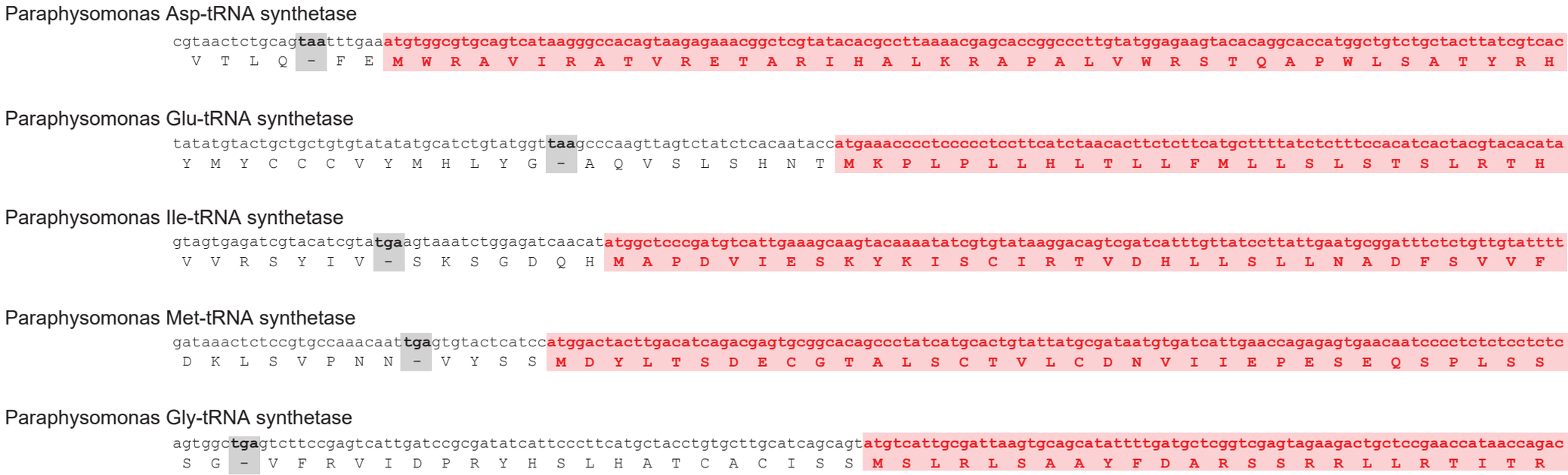


Fig. S15. 5' UTR sequences of five amino acyl-tRNA synthetases of *Paraphysomonas* determined by TAiL-PCRs

Deduced coding regions are highighted in red and are predicted to have N-terminal mitochondrial targeting sequences (see also Fig. S4). In-frame terminal codons are highighted in grey. The "ATG" codons deduced as the initiation codons here are actually the first methionine codons appeared downstream from the in-frame termination codons. This indicates that the *Paraphysomonas* amino acyl-tRNA synthetases with the N-terminal mitochondrial targeting sequences do not have any additional signal peptide at the upstream regions.

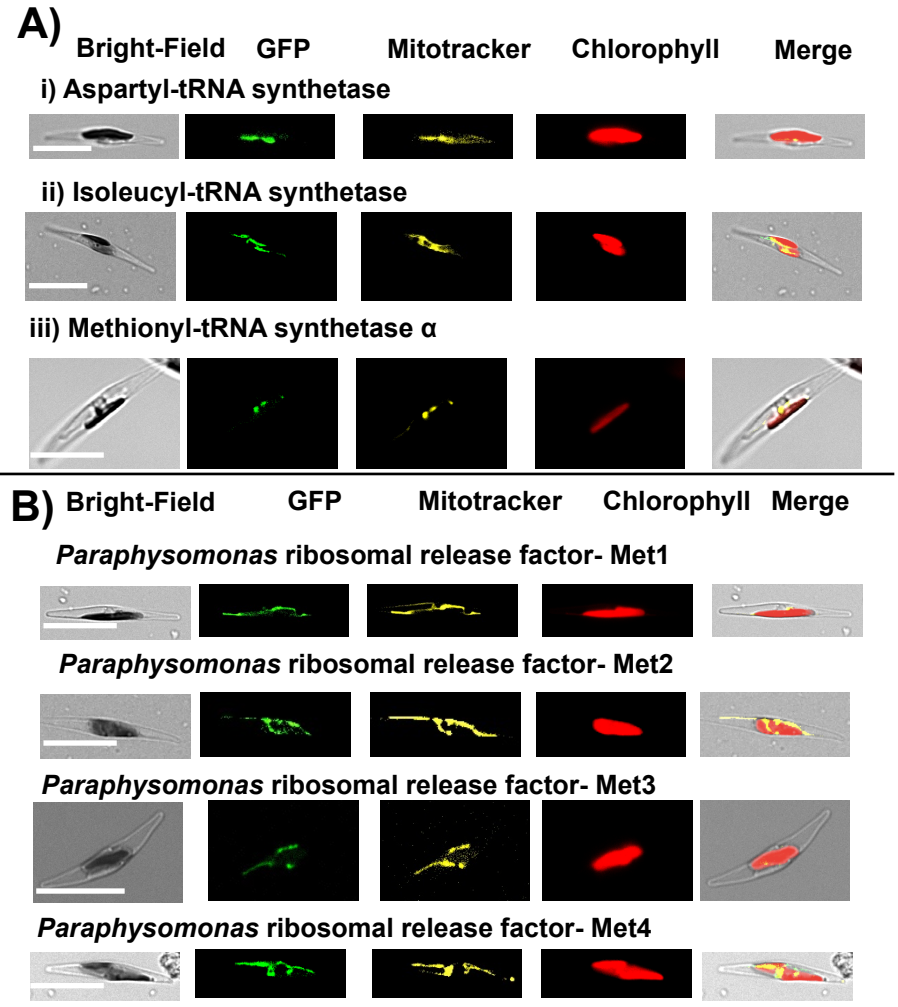
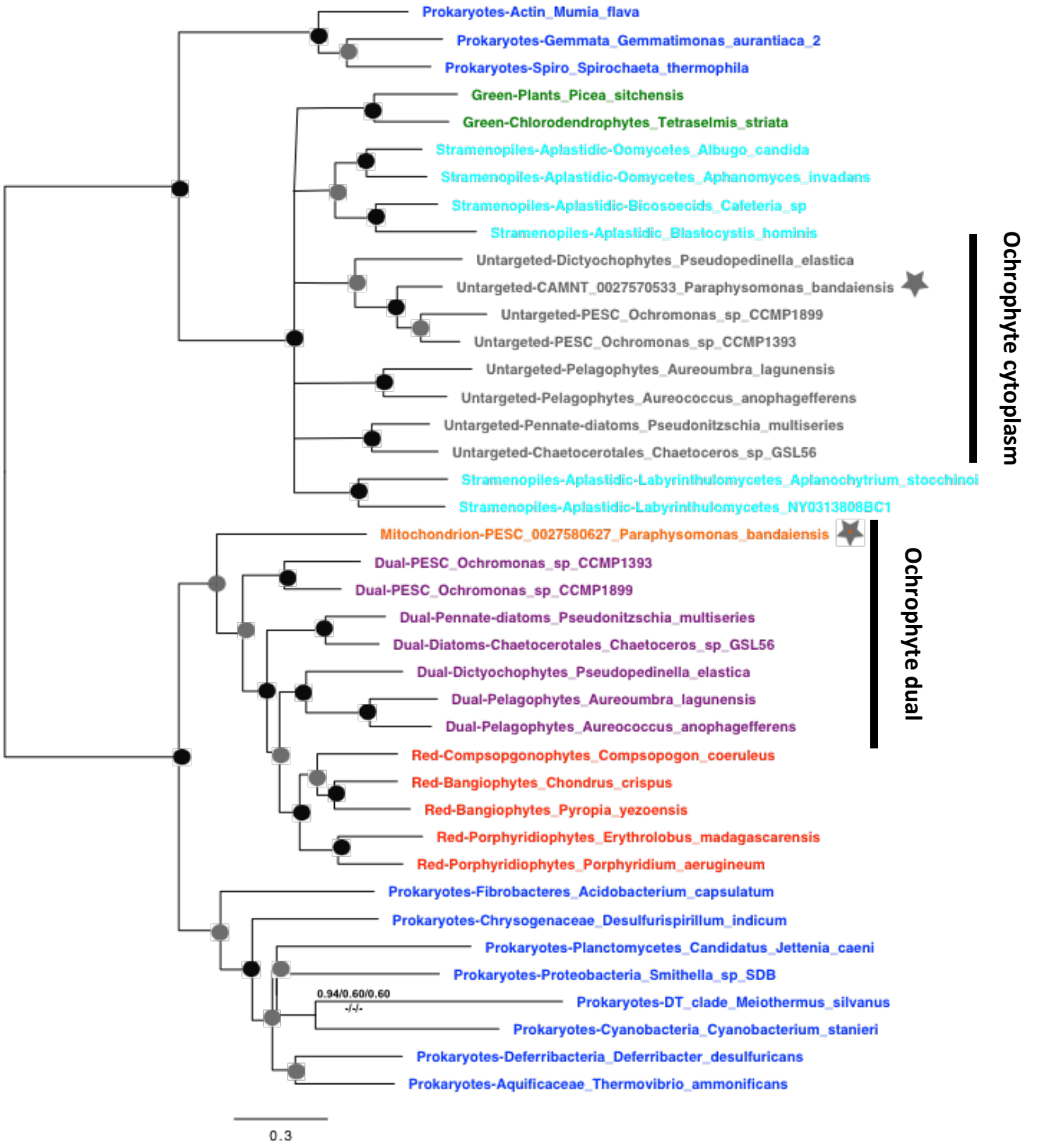


Fig. S16. Mitochondrial retargeting of proteins previously associated with the *Paraphysomonas* plastid genome. A: images of *Phaeodactylum* lines transformed with GFP-linked constructs of the N-terminal regions of *Paraphysomonas bandaiensis* aminoacyl-tRNA synthetases identified by phylogeny to previously have functioned in the expression of the plastid genome, stained with Mitotracker Orange**.** Each construct localises unilaterally to the mitochondria. **B:** analogous images for a mitochondrion-targeted *Paraphysomonas* ribosomal release factor of mitochondrial evolutionary origin. Separate images are provided for four candidate translation initiation codons upstream of the CDD, as per fig. 4B. Stain- and GFP-negative controls for each image are shown in Fig. S13B. Scale bars are to 10 µm.



Key

Ochrophyte mitochondria-targeted protein
Ochrophyte dual plastid/ mitochondria-targeted protein
Ochrophyte cytoplasmic/ untargeted protein
Red algae
Green algae
Prokaryote
Plastid-lacking SAR clade member

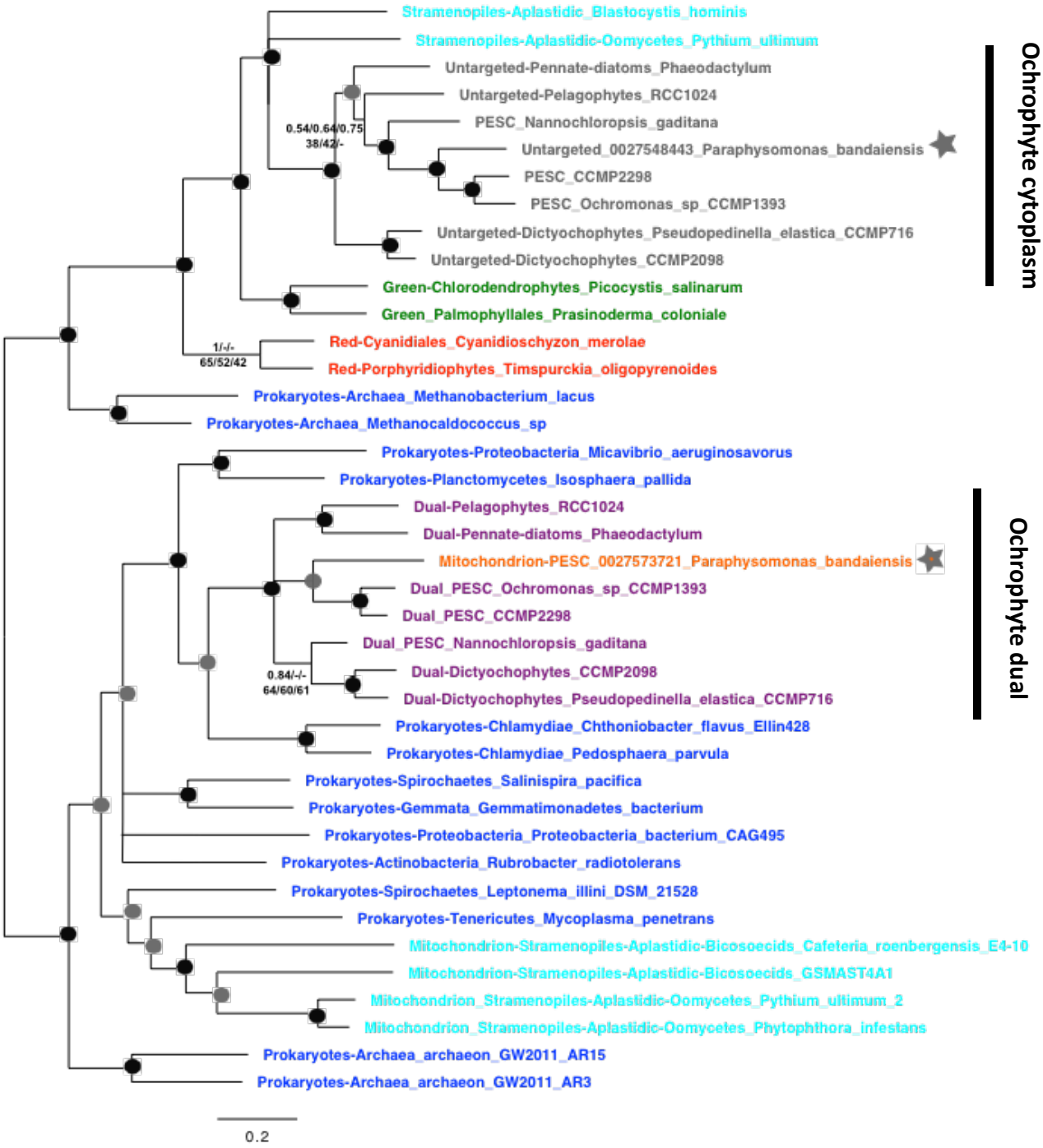
Plastid-derived protein retained in *Paraphysomonas* sp.

Node with strong support (Bayesian PP = 1.0, and RAxML best tree bootstrap support > 80% in all alignments tested)

Node with moderate support (Bayesian PP > 0.8, or RAxML best tree bootstrap support > 50% in 2/3 alignments tested)

A/B/C Consensus support: Bayesian PP (GTR/ Jones/ WAG)
x/y/z RAxML bootstrap (GTR/ JTT/ WAG)

Fig. S17. Consensus tree of PESC glutamyl-tRNA synthetase sequences.
This tree shows the Bayesian consensus topology inferred for a 40 taxa x 433 aa alignment corresponding to glutamyl-tRNA sequences from across the tree of life. Ochrophyte sequences are shaded by predicted subcellular localisation, and all remaining sequences are shade by taxonomy. Two *Paraphysomonas* sequences are identified: an experimentally verified mitochondria-targeted protein, which groups with dual plastid/ mitochondria-targeted isoforms from other PESC clade members; and a putative cytoplasmic version, which groups with other ochrophyte cytoplasmic enzymes.



Key

Ochrophyte mitochondria-targeted protein
Ochrophyte dual plastid/ mitochondria-targeted protein
Ochrophyte cytoplasmic/ untargeted protein
Red algae
Green algae
Prokaryote
Plastid-lacking SAR clade member

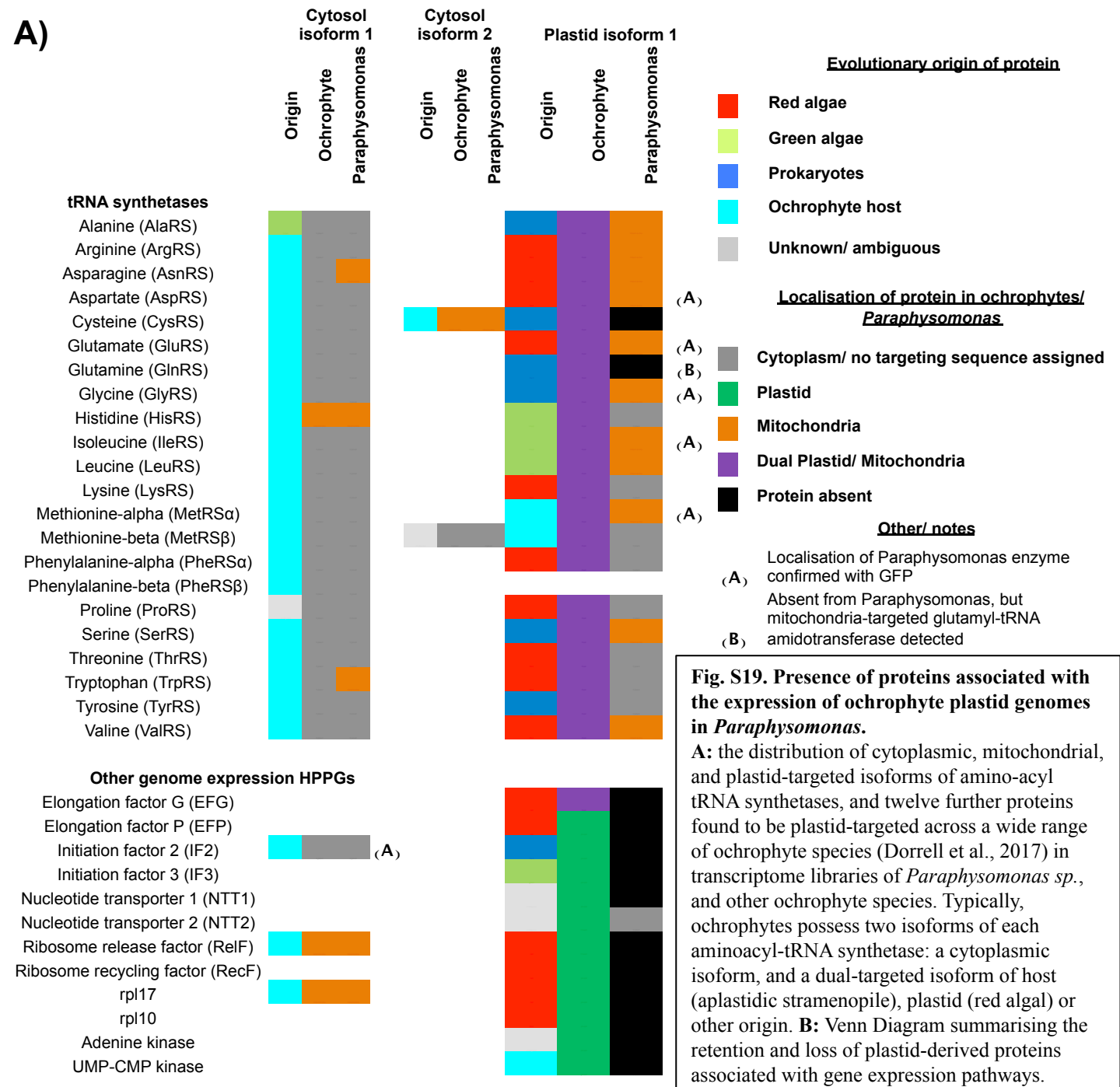
Plastid-derived protein retained in *Paraphysomonas* sp.

Node with strong support (Bayesian PP = 1.0, and RAxML best tree bootstrap support > 80% in all alignments tested)

Node with moderate support (Bayesian PP > 0.8, or RAxML best tree bootstrap support > 50% in 2/3 alignments tested)

A/B/C Consensus support: Bayesian PP (GTR/ Jones/ WAG)
x/y/z RAxML bootstrap (GTR/ JTT/ WAG)

Fig. S18. Consensus tree of PESC glycyl-tRNA synthetase sequences.
This tree shows the Bayesian consensus topology inferred for a 40 taxa x 410 aa alignment corresponding to glycyl-tRNA synthetase sequences from across the tree of life, shown as per fig. S17.



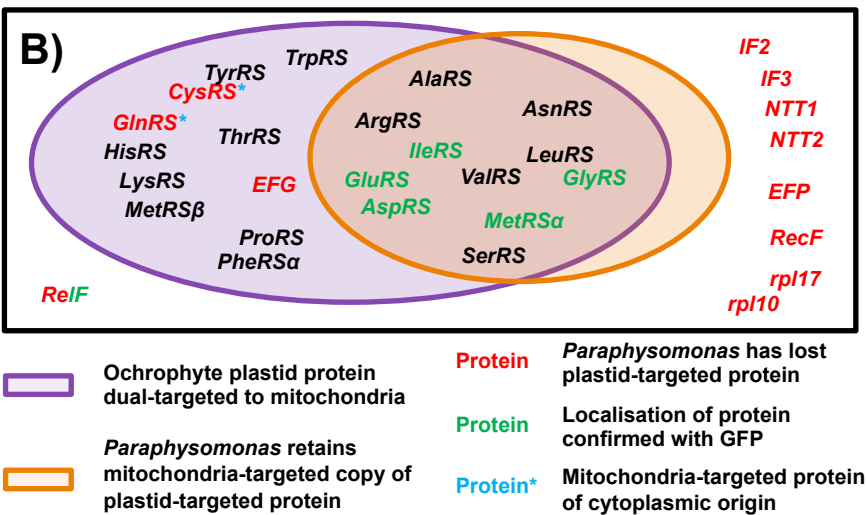
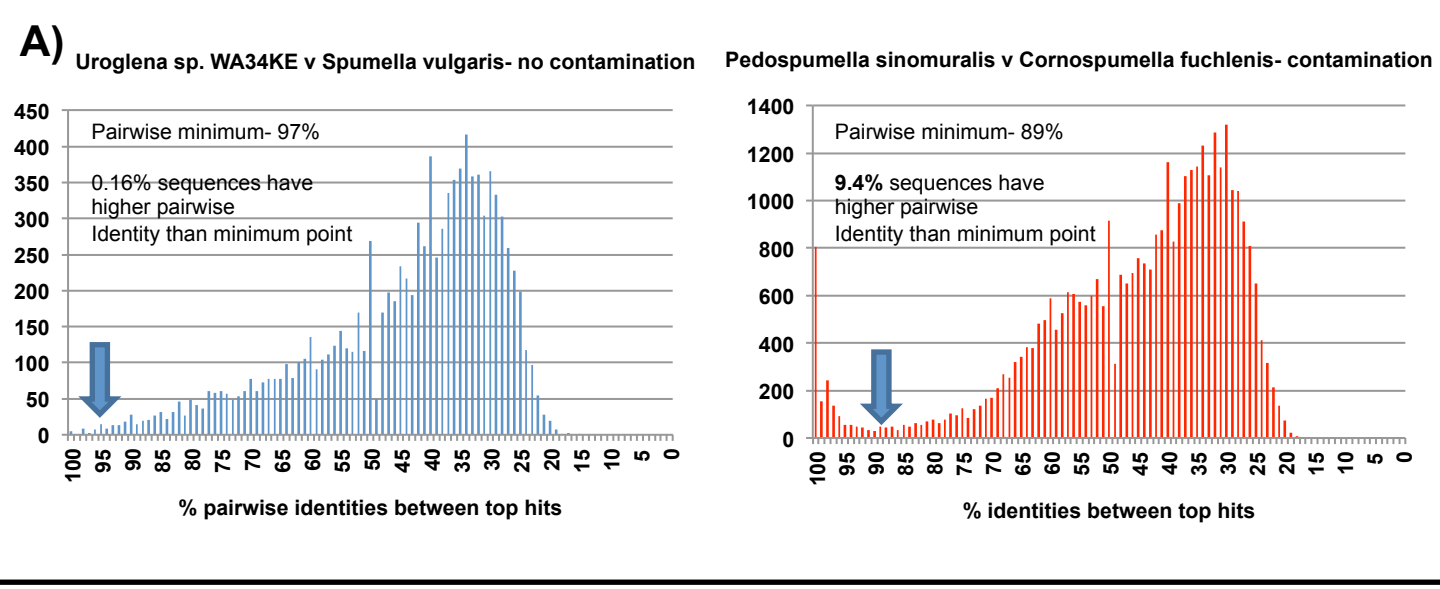


Fig. S19. Presence of proteins associated with the expression of ochrophyte plastid genomes in *Paraphysomonas*.
A: the distribution of cytoplasmic, mitochondrial, and plastid-targeted isoforms of amino-acyl tRNA synthetases, and twelve further proteins found to be plastid-targeted across a wide range of ochrophyte species (Dorrell et al., 2017) in transcriptome libraries of *Paraphysomonas sp.*, and other ochrophyte species. Typically, ochrophytes possess two isoforms of each aminoacyl-tRNA synthetase: a cytoplasmic isoform, and a dual-targeted isoform of host (aplastidic stramenopile), plastid (red algal) or other origin. **B:** Venn Diagram summarising the retention and loss of plastid-derived proteins associated with gene expression pathways. *Paraphysomonas* retains the overwhelming majority of the ancestrally dual-targeted proteins, and in many cases the *Paraphysomonas* isoforms are inferred to possess uniquely mitochondrial targeting sequences. The only exceptions to this rule are cysteinyl-tRNA synthetase, for which *Paraphysomonas* retains an enzyme of solely mitochondrial origin, and glutaminyl-tRNA synthetase, for which *Paraphysomonas* instead apparently uses a mitochondria-targeted glutamyl-tRNA transaminase (Gile et al., 2015). For almost all of the other studied proteins, the ochrophytes apparently possess separate mitochondria- and plastid-targeted isoforms, the latter of which are unilaterally not detected in *Paraphysomonas* transcriptomes.



B)

% contamination

37.64
30
20
15
10
5

	Pedospumella encystans	Pedospumella sinomuralis JBCS23	Spumella vulgaris	Spumella lacusvadoi JBNZ39	Apoikiospumella mondseensis JBM08*	Spumella bureschii JBL14	Acrispumella msimbazaensis JBAF33	Poteriospumella lacustris JBC07	Poteriospumella lacustris JBM10	Poteriospumella lacustris JBN41	Cornospumella fuchlensis AR4D6	Synura sp. LO234KE	Dinobryon sp. FU22KAK	Dinobryon sp. LO226KS	Epipyxis sp PR26KG	Ochromonas sp. LO244K-D	Uroglena sp. WA34KE	Poterioochromonas sp. DS
Pedospumella encystans			80	159	147	141	107	12	17	10	172	9	21	10	9	78	9	15
Pedospumella sinomuralis JBCS23			207	870	203	27	307	44	43	38	2143	159	14	38	11	144	16	31
Spumella vulgaris	101	181					65	7	10	8	49	37	28	28	62	260	33	29
Spumella lacusvadoi JBNZ39	508	2043					3186	7	6	8	47	195	8	18	50	3560	191	8
Apoikiospumella mondseensis JBM08*	117	347					728	3	3	54	196	10	22	6	22	505	65	1
Spumella bureschii JBL14	137	45					3265	10	15	33	54	33	259	59	35	1651	18	28
Acrispumella msimbazaensis JBAF33	347	96	222	3479	530	3301		151	148	147	524	14	14	10	55	974	17	17
Poteriospumella lacustris JBC07	7	28	17	1	12	5	8				28	14	29	161	25	5	20	890
Poteriospumella lacustris JBM10	8	13	15	14	13	8	9				2	20	33	125	37	5	25	946
Poteriospumella lacustris JBN41	9	92	6	7	29	17	5				4	18	33	172	54	7	24	861
Cornospumella fuchlensis AR4D6	27	261	11	20	15	11	27	20	3	8		15	27	15	3	207	11	3
Synura sp. LO234KE	16	52	30	68	14	11	19	12	20	19	20		92	153	85	1697	51	27
Dinobryon sp. FU22KAK	24	31	23	59	45	69	10	36	26	29	49	112				25	27	25
Dinobryon sp. LO226KS	24	37	54	31	27	58	50	278	228	267	86	120				54	31	305
Epipyxis sp PR26KG	30	39	52	55	6	22	53	25	56	58	7	45				47	222	25
Ochromonas sp. LO244K-D	105	130	413	3746	765	1381	1215	7	7	6	488	2817	24	39	174		23	13
Uroglena sp. WA34KE	15	8	12	13	38	12	15	33	32	33	16	29	35	34	50	15		76
Poterioochromonas sp. DS	44	270	9	10	18	32	18	1530	1617	1589	35	17	36	299	26	7	56	

Fig. S20. Contamination in published chrysophyte transcriptomes.
A: exemplar BLAST top hit frequency distribution plots obtained for two published chrysophyte transcriptomes with no apparent reciprocal contamination (left), and two with contamination (right; Text S1A). **B:** heatmap showing the number (values) and percentage (shaded intensity) of transcripts in libraries reported in Beisser et al., 2017; inferred to be potential contaminants, using the BLAST approach. Each cell shows the number of transcripts in the row species identified to be potential contaminants when searched against the column species library. *Apoikiospumella mondseensis JBM08* is asterisked as the transcriptome identified with this name was found (by 18S analysis) to in fact correspond to a second isolate of *Spumella lacusvadoi JBNZ39.*

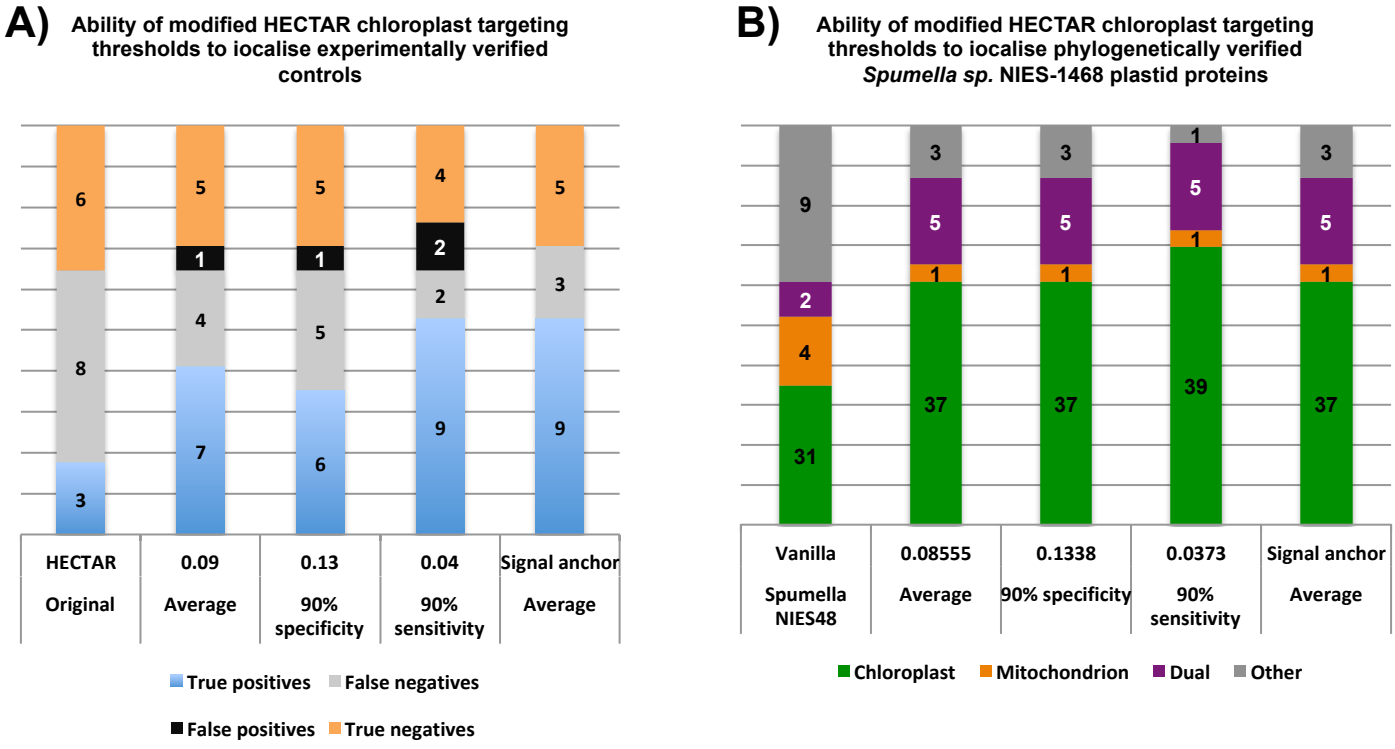


Fig. S21. Custom targeting thresholds for inspecting PESC clade transcriptomes.
A: the number of experimentally verified chrysophyte and eustigmatophyte proteins identified to possess plastid-targeting peptides under default chloroplast score conditions; the average between 90% specificity and sensitivity chloroplast score values; 90% specificity chloroplast score; 90% sensitivity chloroplast score ; and the selected condition of the average between 90% specificity and sensitivity chloroplast score values, with an additional stipulation that the chloroplast score be greater than the signal anchor score calculated. **B:** the targeting predictions made by each threshold for a validation dataset of 48 proteins identified from the « *Spumella » sp. NIES1846* transcriptome to resolve with other PESC clade plastid proteins.

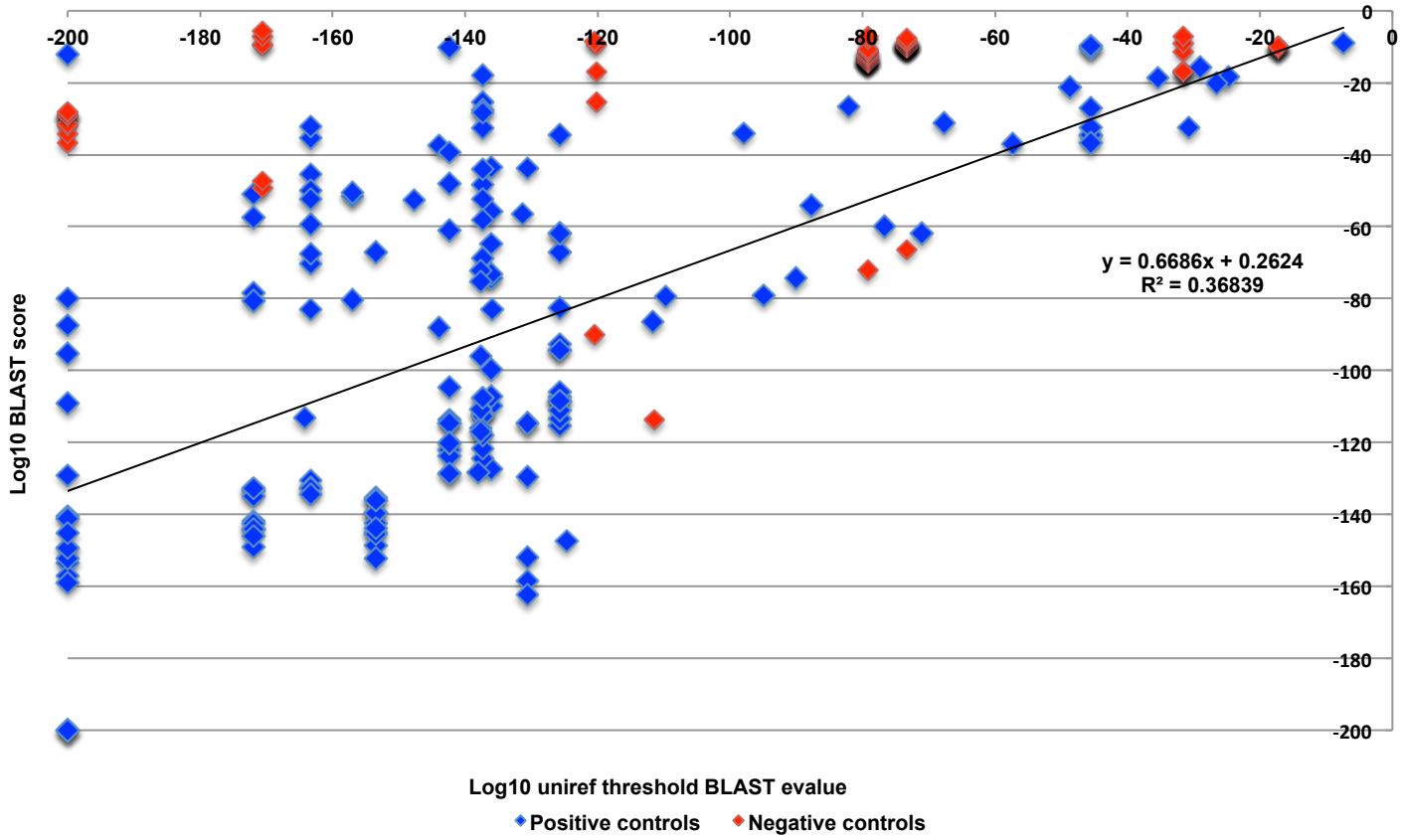


Fig. S22. Floating evalue thresholds for BLAST identification of plastid protein homologues in PESC clade libraires. This figure shows a scatterplot of (horizontal) the BLAST evalues obtained when a subset of the 9531 query proteins conserved across ochrophyte lineages are searched against a modified uniref library, excluding all taxa with a suspected history of serial endosymbiosis; and (vertical) a selected dataset of phylogenetically verified positive control proteins (consisting of proteins inferred to resolve with other ochrophyte plastid orthologues) and negative controls (proteins corresponding to mitochondrial and cytoplasmic homologues of plastid-targeted and/ or plastid-encoded proteins). To facilitate regression calculations, all zero evalues are shown as 1 x 10^{-200}. The best-fit line shows the best possible separation of the positive and negative control datasets: values below the line (i.e. the PESC clade evalue is lower than the expected value from the regression against the uniref evalue) are likely to be orthologues; values above the line show too weak homology for orthology to be confidently assigned.