


SCIENTIFIC REPORTS



OPEN

Peculiar features of the plastids of the colourless alga *Euglena longa* and photosynthetic euglenophytes unveiled by transcriptome analyses

Kristína Záhonová¹ , Zoltán Füßy², Erik Birčák³, Anna M. G. Novák Vanclová⁴, Vladimír Klimesš¹, Matej Vesteg⁵, Juraj Krajčovič⁶, Miroslav Oborník^{2,7} & Marek Eliáš¹

Euglenophytes are a familiar algal group with green alga-derived secondary plastids, but the knowledge of euglenophyte plastid function and evolution is still highly incomplete. With this in mind we sequenced and analysed the transcriptome of the non-photosynthetic species *Euglena longa*. The transcriptomic data confirmed the absence of genes for the photosynthetic machinery, but provided candidate plastid-localised proteins bearing N-terminal bipartite topogenic signals (BTSs) of the characteristic euglenophyte type. Further comparative analyses including transcriptome assemblies available for photosynthetic euglenophytes enabled us to unveil salient aspects of the basic euglenophyte plastid infrastructure, such as plastidial targeting of several proteins as C-terminal translational fusions with other BTS-bearing proteins or replacement of the conventional eubacteria-derived plastidial ribosomal protein L24 by homologs of archaeo-eukaryotic origin. Strikingly, no homologs of any key component of the TOC/TIC system and the plastid division apparatus are discernible in euglenophytes, and the machinery for intraplastidial protein targeting has been simplified by the loss of the cpSRP/cpFtsY system and the SEC2 translocon. Lastly, euglenophytes proved to encode a plastid-targeted homolog of the termination factor Rho horizontally acquired from a Lambdaproteobacteria-related donor. Our study thus further documents a substantial remodelling of the euglenophyte plastid compared to its green algal progenitor.

Euglenophytes, exemplified by the highly studied mixotrophic alga *Euglena gracilis*, are a group of flagellated algae constituting one of the many lineages of the phylum Euglenozoa¹. Euglenophytes and other euglenozoans share many unusual features, such as regulation of gene primarily at the post-transcriptional level^{2–5}. Furthermore, euglenozoans employ *trans*-splicing to process mRNA molecules, whereby the 5'-end of pre-mRNA is replaced by the 5'-end of the specialised spliced leader (SL) RNA, resulting in the presence of an invariant SL sequence (ACTTCTGAGTGTCTATTTTTTTTCG in *E. gracilis*) at the 5'-end of mature mRNAs^{6,7}.

Despite their interesting biology, euglenophytes have not yet been properly studied by genome-wide approaches. A few studies employed transcriptome sequencing of *E. gracilis* to investigate particular aspects of its gene repertoire and selected functional pathways^{8–10}. Two independent transcriptome assemblies are available in the GenBank database, and a nuclear genome draft has been announced, although not yet made public at the time of writing of this paper⁸. In addition, transcriptome assemblies of two different isolates of the marine euglenophyte genus *Eutreptiella* (*E. gymnastica* NIES-381 and *E. gymnastica*-like CCMP1597) were sequenced as part of the MMETSP project¹¹, but no specific analyses of this data resource have been reported. Hence, further

¹Life Science Research Centre, Department of Biology and Ecology and Institute of Environmental Technologies, Faculty of Science, University of Ostrava, 701 00, Ostrava, Czech Republic. ²Institute of Parasitology, Biology Centre CAS, 370 05, České Budějovice, Czech Republic. ³Department of Genetics, Faculty of Natural Sciences, Comenius University, 842 15, Bratislava, Slovakia. ⁴Department of Parasitology, Faculty of Science, Charles University, BIOCEV, Prague, Czech Republic. ⁵Department of Biology and Ecology, Faculty of Natural Sciences, Matej Bel University, 974 01, Banská Bystrica, Slovakia. ⁶Department of Biology, Faculty of Natural Sciences, University of ss. Cyril and Methodius in Trnava, 917 01, Trnava, Slovakia. ⁷University of South Bohemia, Faculty of Science, 370 05, České Budějovice, Czech Republic. Kristína Záhonová and Zoltán Füßy contributed equally. Correspondence and requests for materials should be addressed to M.E. (email: marek.elias@osu.cz)

studies are clearly needed to improve our understanding of the molecular underpinnings of the euglenophyte life and evolution.

The defining feature of euglenophytes is a complex three-membrane-bounded plastid derived from a green alga belonging to Pyramimonadales^{1,12–14}. As in other plastid-bearing eukaryotes, only a minority of plastid proteins are encoded by the plastid genome; the nucleus-encoded majority then need to cross the three membranes of the euglenophyte plastid envelope to reach the site of their function. The mechanism of protein targeting to the plastid has been partially characterized in *E. gracilis*¹⁵. The proteins co-translationally enter the endoplasmic reticulum (ER) and are transported further by vesicular trafficking, passing the Golgi apparatus *en route* to the plastid. The *E. gracilis* plastid-targeted proteins bear a discernible presequence, an N-terminal bipartite topogenic signal (BTS), which comes in two main variants¹⁶. Both include an N-terminal signal peptide mediating the import into the ER, followed by a plastid transit peptide that is exposed upon signal peptide cleavage and mediates the import across the two inner chloroplast membranes. In Class I presequences, the transit peptide is followed by a transmembrane domain (TMD) that anchors the transported protein in the membrane during its subcellular relocation, whereas far less frequent Class II presequences lack the anchoring TMD. The signal peptide itself typically has physicochemical properties of a TMD, resulting in a characteristic double-TMD motif in the Class I presequences of euglenophyte plastid proteins¹⁶. The characteristic structure of the *E. gracilis* plastid-targeting BTSs has facilitated *in silico* identification of candidates for plastid-targeted proteins in euglenophytes^{16–19}. However, the proteome of neither euglenophyte plastid has yet been reconstructed in full.

Although most euglenophytes are photosynthetic, several lineages independently lost photosynthesis and became secondarily heterotrophic²⁰. The fate of their plastid is generally unknown, except for *Euglena longa* (originally described as *Astasia longa*), where a non-photosynthetic plastid has been preserved, as evident from the presence of a plastid genome sequenced a long time ago²¹. We have recently demonstrated that an intact plastid genome is essential for the *E. longa* survival, in contrast to the photosynthetic *E. gracilis*²². The *E. longa* plastid genome size (75 kbp) is approximately half of that of *E. gracilis*, with the difference attributed primarily to the absence of photosynthesis-related genes. The only exception is the *rbcl* gene encoding the large subunit of the enzyme ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO LSU) retained in the *E. longa* plastid genome^{21,23}. However, the plastid organelle itself remains elusive. Double-membrane bodies similar to those present in dark-grown *E. gracilis* were observed in *E. longa* and interpreted as plastids²⁴, but this identification is uncertain given that euglenophyte plastids studied in detail possess three bounding membranes¹. Likewise, the physiological role of the *E. longa* plastid remains unknown.

It is conceivable that a lot could be learned about the *E. longa* plastid by analysing its nuclear genome sequence. However, the genome sequencing project initiated by others for *E. gracilis* revealed that the genome of this species is huge and difficult to assemble⁸. As a close relative^{20,25}, *E. longa* most likely shares general genomic features with *E. gracilis*, making genome sequencing impractical. Hence, to build a resource for exploration of the plastid function and other aspects of the *E. longa* biology, we sequenced and assembled the transcriptome of this species. The sequencing data were obtained from cultures grown at two different light regimes (in the dark and in the light) to improve the coverage of differentially expressed genes and to maximize the chance we detect possible traces of genes encoding the photosynthesis-related machinery. Data extracted from the sequenced *E. longa* transcriptome proved instrumental in characterizing the function of the RuBisCO enzyme in this species²³ and enabled identification of a novel Euglenozoa-specific form of the Rheb GTPase²⁶, but publication of the whole transcriptome assembly has been pending.

Here we describe the general characteristics of the transcriptome and demonstrate its utility for unravelling the biology of the *E. longa* plastid. We provide the first insights into the basic infrastructure of the plastid and compare it to the molecular machinery of plastid biogenesis in photosynthetic euglenophytes. Our results demonstrate not only *E. longa*-specific simplification related to the loss of photosynthesis, but also a surprising reduction of the plastid biogenesis machinery in euglenophytes in general. Finally, we report on a case of an expansion of the euglenophyte plastid functions by acquisition of a plastid-targeted homolog of the bacterial transcription termination factor Rho, which is an unprecedented feature among all plastid-bearing eukaryotes studied to date. We believe that the *E. longa* transcriptome, now made available to the whole scientific community, will become an important resource for further research of various aspects of euglenophyte biology.

Results and Discussion

The transcriptome of *Euglena longa*: not all mRNAs bear the 5' end *trans*-spliced leader sequence.

Our transcriptomic assembly of *E. longa* resulted in 65,563 transcript models, a number somewhat smaller than the numbers reported in recent transcriptomic studies for *E. gracilis* (113,152, ref.¹⁰; 72,506, refs.^{4,8}). Since the use of different sequence assembly algorithms certainly contributes to the difference in contig numbers, we carried out a BUSCO search for conserved unique eukaryote orthologs to assess the quality of our data. 89.1% of BUSCO genes were found to be complete in our dataset, and further 4.3% of orthologs to be present as fragments. These characteristics are similar to those of *E. gracilis* transcriptomic data (Supplementary Fig. S1; refs.^{8,10}) and suggest that our assembly covers the majority of genes in the *E. longa* genome.

Inspection of the assembled transcripts revealed that the SL sequence employed by *E. longa* is the same as the one in *E. gracilis*, although it was often truncated. In total, 31,783 *E. longa* transcript models (48.5%) possessed at least a part of the SL sequence (TTTTTCG) within 35 bp from either end (accounting for transcripts that are by chance assembled in the reverse complement orientation with respect to the template mRNA molecule), indicating 5'-end completeness of the contained coding sequences. The percentage of transcripts with the SL sequence in *E. gracilis* was comparable (54%; ref.¹⁰), even though SL absence was so far directly experimentally demonstrated only for the mRNA of a single *E. gracilis* gene, the one encoding the nucleolar protein fibrillarin²⁷. Since *in silico* prediction of protein subcellular localisation requires full-length protein sequences, it was critical to understand whether the high proportion of SL sequence-lacking transcripts implies a high fraction of truncated sequences.

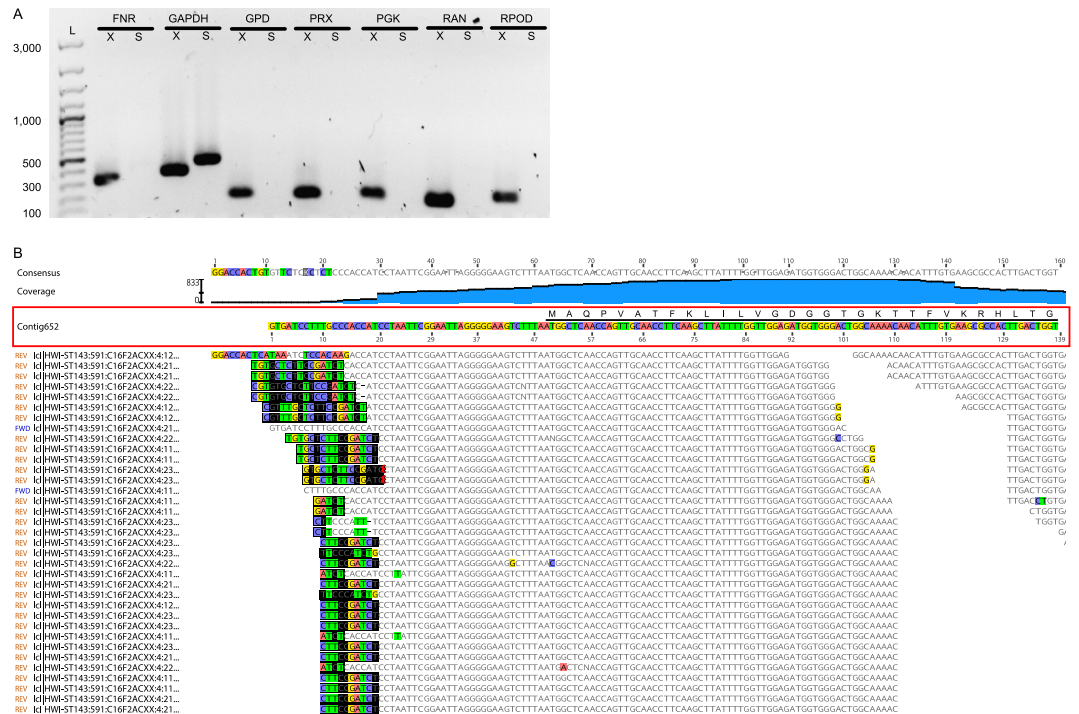


Figure 1. Presence/absence of the spliced leader (SL) in selected transcripts in *E. longa*. **(A)** PCR (with cDNA as the template) with an SL-specific forward primer was done to assay whether the lack of SL in exemplar assembled contigs (listed in Supplementary Table S2) mirrors real SL absence in the respective transcripts. Lanes X show PCR products with gene-specific primers (positive control), lanes S show products with SL forward and gene-specific reverse primers. Lane L is a 100-bp size ladder with sizes shown for selected bands. Assayed transcripts: FNR, ferredoxin-NADP⁺ reductase; GAPDH, glyceraldehyde-3-phosphate dehydrogenase (positive control for the SL-specific primer); GPD, glycerol-phosphate dehydrogenase; PRX, peroxiredoxin; PGK, phosphoglycerate kinase; RAN, RAN GTPase; RPOD, RNA polymerase sigma factor. An unedited full version of the electrophoretic gel is provided as Supplementary Fig. S8. **(B)** Mapping of raw sequence reads to the 5'-end of the RAN GTPase transcript from *E. longa* confirms the absence of the SL sequence. Untrimmed Illumina primer sequences at the 5'-end of several reads are highlighted by black background. The coding sequence is shown by the black horizontal line with the amino acid sequence shown above the Contig652 nucleotide sequence. The apparent discontinuity in the read coverage (around the position 110 of the contig) is only seeming and stems from cropping the read mapping figure (to make the scheme smaller). The coverage plot is shown for comparison.

Therefore, we chose candidate transcripts that lack the SL sequence in our transcriptome assembly (listed in Supplementary Table S1) and tested the presence of the SL sequence at the 5'-end of the respective mRNAs using PCR (with cDNA as the template). Several transcripts failed to be amplified when using the SL-specific forward primer (Fig. 1A, lane S), but were amplified when gene-specific forward primers were used (Fig. 1A, lane X), indicating that they truly lack the SL sequence.

To further test the notion that some *E. longa* mRNAs may not undergo *trans*-splicing, we chose the highly expressed and conserved gene for the GTPase RAN (implicated in nucleocytoplasmic transport; ref.²⁸) and attempted to extend the SL sequence-lacking 5'-end of the respective transcript contig by using all available RNA-seq reads. While the read coverage of the 5'-end is high, no extension to recruit the SL sequence to the end is possible (Fig. 1B). Hence, the maturation of the mRNA 5'-end by SL *trans*-splicing is not universal to all genes in euglenophytes. However, not all SL-sequence lacking contigs in the transcriptome assembly necessarily attest to the lack of *trans*-splicing, as some of them are truly truncated and others may represent un-spliced variants of normally *trans*-spliced mRNAs. Indeed, we found examples of both cases (Supplementary Fig. S2). The genome sequence of *E. gracilis* that should soon become available will enable to carry out a systematic analysis of the occurrence of SL *trans*-splicing across the transcriptome.

No traces of the photosynthetic machinery in the transcriptome of *Euglena longa*. Although polyA-selection was employed during the RNA-seq library preparation, we detected in our assembly transcripts of some genes of the plastid genome (Supplementary Fig. S3; Supplementary Table S2). This may mean that some plastid mRNAs are polyadenylated in *E. longa*, as demonstrated for *E. gracilis*²⁹, but inefficient removal of non-polyadenylated RNA molecules cannot be ruled out as an alternative explanation. Comparison of the transcript sequences with the plastid genome sequence²¹ revealed that the second intron in the *rps2* gene has an incorrectly delimited 3'-border in the current genome annotation, rendering the coding sequence shorter by one amino acid. No signs of plastid mRNA editing were observed in *E. longa*.

The *E. longa* plastid genome lacks many of the genes found in plastid genomes of other euglenophytes (Supplementary Fig. S3). Except for the *rps18* gene encoding the ribosomal protein S18, all missing genes code for components of the photosynthetic machinery, i.e. photosystems I and II, the cytochrome *b₆f* complex, membrane ATP synthase, and the enzyme Mg-protoporphyrin IX chelatase involved in chlorophyll synthesis. We searched the *E. longa* transcriptome assembly to investigate possible transfer of these genes into the nuclear genome, but we did not find any of them, suggesting that no endosymbiotic gene transfer occurred in the *E. longa* lineage after its separation from the *E. gracilis* lineage. We likewise failed to find homologs of other conserved components of the main photosynthetic complexes encoded by the nuclear genome in *E. gracilis* or other photosynthetic euglenophytes (Supplementary Table S2). We assume that if the photosynthesis-related genes were present in the *E. longa* nuclear genome, we would detect transcripts of at least some of them, since one of the sequenced cultures was grown in the light. This corroborates that photosynthesis is truly missing in *E. longa*.

The apparent loss of the gene for the plastid ribosomal protein S18, otherwise broadly conserved in photosynthetic eukaryotes³⁰, raises the question how the plastid ribosome small subunit is affected by the absence of this protein. However, *rps18* is missing from some other colourless plastid genomes, such as those of apicomplexans, the chlorophytes *Helicosporidium* sp., *Prototheca stagnora*, and *Polytoma uvella*, and the diatom *Nitzschia* sp. NIES-3581^{31–33}. Apicomplexans were reported to lack even a nucleus-encoded apicoplast-targeted version of S18 (while keeping a mitochondrion-targeted version³⁴), and we likewise could not find a plastid-targeted S18 protein in the available nuclear genome data from *Helicosporidium*, suggesting that S18 is not absolutely essential for translation in the plastid.

Probing for nucleus-encoded plastid proteins in *E. longa*: aminoacyl-tRNA synthetases and ribosomal proteins.

Several *E. longa* proteins predicted to be targeted to its plastid were previously reported, including the small RuBisCO subunit (RBCS), RuBisCO activase, the chaperonins GroEL/GroES, and the assembly factor RAF (see also Supplementary Table S3), but their targeting sequences were not investigated in detail²³. To get a broader representative set of nucleus-encoded proteins likely imported into the *E. longa* plastid, we searched the transcriptome assembly for proteins from two functional categories expected to be present in the *E. longa* plastid: (1) aminoacyl-tRNA synthetases needed for charging tRNAs specified by the plastid genome; and (2) ribosomal proteins not encoded by the plastid genome. The same search was done for *E. gracilis* to further assess the representativeness of the *E. longa* transcriptome assembly with special regard to nuclear genes encoding plastidial proteins. Plastidial aminoacyl-tRNA synthetases and ribosomal proteins were discriminated from homologs functioning in other compartments (the cytosol and the mitochondrion) by virtue of their closer sequence similarity to plastidial homologs in other eukaryotes and/or by the presence of an N-terminal extension bearing general characteristics of BTSs as previously defined in *E. gracilis*.

We found the same set of these two protein categories in both species, although in several cases the respective sequence was 5'-truncated in one or the other species (Supplementary Tables S4 and S5). Some of these incomplete sequences could be extended by manual iterative recruitment of sequencing reads to the 5'-end of the transcript (often up to the SL sequence), providing the missing part of the coding sequence and consequently the BTS in the encoded protein. Although sequences of some putative plastid-targeted proteins remained truncated even when using this approach, their plastid localisation can be assumed based on the premise that the missing N-terminal region of the protein has similar features as its ortholog from the other *Euglena* species. With this premise, plastidial aminoacyl-tRNA synthetases cognate to all twenty amino acids exist in both *Euglena* species, two of them being included in one fusion protein (see below). The presence of a plastidial version of Gln-tRNA synthetase in *Euglena* spp. is noteworthy, because plastids of different plant and algal groups lack this enzyme and instead rely on an alternative two-step process of Gln-tRNA synthesis inherited from the cyanobacterial progenitor of the plastid and mediated by Glu-tRNA synthetase and Glu-tRNA amidotransferase^{35,36}. However, putative plastid-targeted glutamyl-tRNA synthetases were recently found in diatoms and the cryptophyte *Guillardia theta*³⁷, so plastids may be more diverse in their mechanism of Gln-tRNA synthesis than previously thought. The euglenophyte plastid-targeted Gln-tRNA synthetase is evidently not directly related to the enzymes from other algae and was likely gained by horizontal gene transfer (HGT) from a bacterium, but the exact donor group cannot be resolved by phylogenetic analyses of presently available sequences (data not shown).

An even more interesting picture emerged from the analysis of plastidial ribosomal proteins. The set of ribosomal proteins with apparent plastid-targeting presequences identified in the transcriptome data complemented the set encoded by the plastid genomes, such that all subunits of the plastid ribosome known from other algal groups (see ref.³⁴) are conserved in both *Euglena* species, with two exceptions. The first is the lack of S18 in *E. longa* discussed in the previous section, and the second is the absence of the expected eubacteria-like L24. Both *Euglena* species each instead encode two other proteins of the same ortholog family (called uL24 according to the latest unified nomenclature of ribosomal proteins³⁸) that exhibit N-terminal extensions fitting the structure of the BTS (Supplementary Table S5). The two *Eutreptiella* species each possess only one such protein, which however represent two different paralogs that originated before the euglenophyte radiation (with possible subsequent differential loss in the *Eutreptiella* lineage; Fig. 2). These two paralogs share a common ancestor apparently belonging to the archaeo-eukaryotic uL24 family branch, often called L26. A phylogenetic analysis did not suggest a more specific scenario, as the position of the paralog pair outside the eukaryotic and archaeal clades in the tree (Fig. 2) is most likely due to their rapid divergence erasing the phylogenetic signal in these short proteins.

It is tempting to speculate that the plastid-targeted L26-related proteins functionally compensate for the absence of the eubacteria-like L24 in the euglenophyte plastidial ribosome, despite considerable sequence divergence between the eubacterial and archaeo-eukaryotic homologs. Such a replacement of an organellar ribosomal protein by a homolog from a different phylogenetic domain may seem unlikely, but is apparently possible. At least two similar cases have been documented, both featuring a novel paralog of the eukaryote-type cytosolic ribosomal protein replacing the homologous eubacteria-like counterpart of the organellar ribosome: the

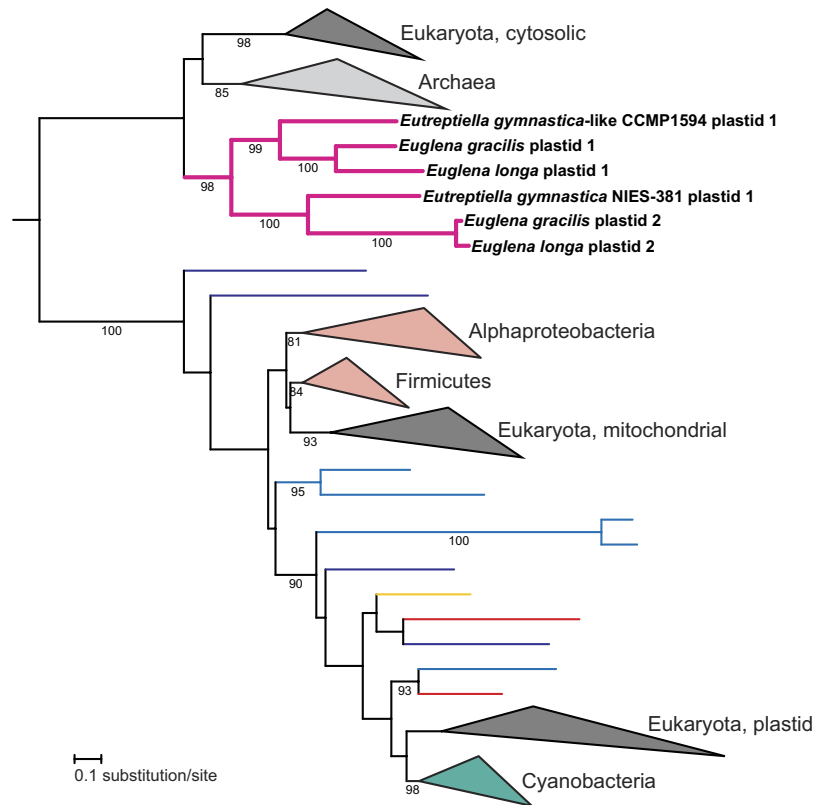


Figure 2. Phylogenetic analysis of the uL24 family of ribosomal proteins. The tree shows the phylogenetic position of the presumably plastid-localised L26-related proteins in euglenophytes. Bootstrap support values are given when ≥ 80 .

eubacteria-type S8 protein in angiosperm mitochondria replaced by the eukaryotic protein S15A³⁹ and the ancestral eubacterial L23 replaced by the eukaryotic homolog in the plastid of spinach (but perhaps also in other plants^{40,41}). The euglenophyte plastidial L26-related proteins thus may be another such case. Why *Euglena* spp. exhibit two different plastid-targeted L26-related proteins is unclear, but it is possible that one paralog is not a part of the ribosome itself and performs another role. Indeed, the cytosolic L26 has a ribosome-independent function as a regulator of translation of the p53 family proteins in mammals^{42,43}.

Protein import into the euglenophyte plastids: targeting sequences and translational fusions.

The collection of the high-confidence candidates for *E. longa* proteins targeted to the plastid established in our previous study²³ and by the analyses described above enabled us to evaluate the general characteristics of the plastid-targeting presequences in this species. The analysis revealed that the N-terminal extensions of these proteins exhibit the same characteristics as the BTS defined in *E. gracilis* (see above). Specifically, we could find presequences of both class I and class II (Supplementary Fig. S4), with the relative abundance of class I being lower (40 class I presequences vs. 15 class II presequences) than in *E. gracilis* (89% according to ref.¹⁶; Supplementary Tables S3–S5). The difference might reflect a preference of photosynthesis-related proteins to utilise Class I presequences, but this needs to be confirmed by a broader analysis of plastid-targeted proteins in euglenophytes. The class I presequences typically exhibited the pattern of two predicted TMDs separated by a hydrophilic amino acid stretch (corresponding to the plastid transit peptide) according to the 60 ± 8 rule¹⁶, and the N-terminal signal peptide was predicted in most proteins by all tools employed. This suggests that both *Euglena* species share a similar route of plastid protein import and, presumably, that the *E. longa* plastid envelope also consists of three membranes.

We previously documented that the RBCS transcript in *E. longa* encodes a polyprotein comprising a single targeting presequence followed by several monomers of the mature RBCS separated by a conserved decapeptide linker²³. A similar organization is found also in the *E. gracilis* RBCS and is characteristic of a handful of other photosynthesis-related proteins in *E. gracilis*^{44–47} and the dinoflagellate *Prorocentrum minimum*⁴⁸. In *E. gracilis*, individual RBCS units are processed by proteolytic cleavage upon the import of the polyprotein into the plastid⁴⁵, while RBCS polymer processing was not detected in *E. longa*²³. Interestingly, we now observed that the translation elongation factor EF-Ts is encoded in a similar fashion in both *E. longa* and *E. gracilis*, with a plastid-targeting presequence followed by two EF-Ts monomers separated by a linker (Fig. 3).

Moreover, we found four *E. longa* plastid proteins that apparently reach the organelle as translational fusions with other proteins endowed with an N-terminal targeting presequence. Specifically, we observed ribosomal proteins L10 and L17 fused to the C-terminus of ribosomal proteins L15 and L28, respectively, methionyl-tRNA

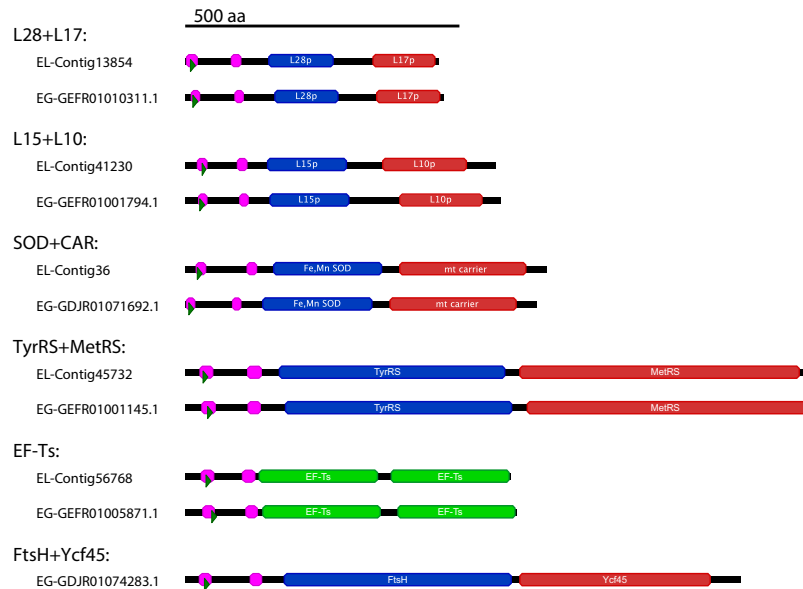


Figure 3. Domain structure of fusion proteins in *E. longa* and *E. gracilis*. Regions corresponding to separate mature proteins presumably released by processing of the fusion protein are shown as boxes in blue and red (if different) or in green (if the fusion comprises a repeat of the same monomer). The two purple domains at the N-terminus are transmembrane domains (TMDs) as predicted by TMHMM, the first TMD is a part of the signal peptide, the second domain is the anchoring TMD that follows the chloroplast transit peptide (Class I targeting presequence). Contig IDs from the presented transcriptome are given for *E. longa* models (EL), accessions are listed for their *E. gracilis* orthologs (EG); GEFR and GDJR accessions are from TSA records GEFR00000000.1 (ref.⁸) and GDJR00000000.1 (ref.¹⁰), respectively.

synthetase fused to the C-terminus of tyrosyl-tRNA synthetase, and a putative dicarboxylate carrier fused to the C-terminus of superoxide dismutase (Fig. 3). All of these fusion proteins have a similar structure, comprised of the N-terminal BTS followed by protein monomers separated by a short peptide linker. Hence, unlike the RBCS and EF-Ts multimers, these transcripts encode two different proteins. All these fusions are encountered also in *E. gracilis* (Fig. 3) and are thus unlikely to be assembly artefacts. The list of proteins delivered to euglenophyte plastids as translational fusions will probably grow with a more in-depth analysis. For example, while investigating the family of FTSH proteases (see the next section), we found out that one of the predicted plastid-targeted paralogs shared by *E. gracilis* and both *Eutreptiella* species (yet absent from *E. longa*) has a C-terminal extension corresponding to the uncharacterized plastid protein Ycf45 (in some algae encoded by the plastid genome) (Fig. 3). Whether the fused proteins are processed in the plastid by cleavage or remain joined together needs to be determined.

Euglenophyte and *E. longa*-specific simplifications of the basic plastid infrastructure. The fact that euglenophyte presequences include a region with characteristics of the plastid transit peptide implies the existence of a plastid import machinery homologous to the translocon of the outer/inner chloroplast membrane (TOC/TIC) of other plastids¹⁵. However, we failed to identify homologs of most of the TOC/TIC components in the *E. longa* transcriptome even when using HMMER and profile HMMs for the respective protein families (i.e. an approach substantially more sensitive than conventional BLAST). The only exceptions were the proteins TIC32 and TIC62, which belong to a large family of short-chain dehydrogenases^{49,50}. TIC32 was described as a calmodulin-binding, NADPH-dependent regulator of the plant TIC, operating in a redox- and calcium-dependent manner⁵⁰. Proteins (with a putative plastid BTS) highly similar to the plant TIC32 are found in *E. longa* as well as other euglenophytes (Fig. 4; Supplementary Table S3), and a phylogenetic analysis places them closer to the plant TIC32 than to other related proteins (data not shown), suggesting their functional equivalence. However, little is known about TIC32 and its TIC-independent function is conceivable. In contrast, even the most similar euglenophyte homologs of the plant TIC62 do not cluster with them in a phylogenetic analysis (data not shown), indicating that they should not be considered as candidates for TIC components.

Surprisingly, the transcriptomes of *E. gracilis* and *Eutreptiella* spp. proved to encode discernible homologs of only two additional plastid translocon subunits, TIC21 and TIC55 (Fig. 4; Supplementary Table S3). TIC21 (three copies in *E. gracilis* and one in *Eutreptiella* spp.) is only loosely associated with the central translocon subunits and is employed mainly for the import of photosynthesis-related proteins, whereas the import of several non-photosynthetic housekeeping proteins was shown to be unimpaired in TIC21-depleted plant plastids⁵¹. TIC55 was recently shown to serve as phyllobilin hydroxylase in the chlorophyll breakdown pathway and its role in plastid protein import was questioned⁵². Regardless, neither of the euglenophyte TIC55-related proteins is a *bona fide* TIC55 ortholog, as demonstrated by our phylogenetic analysis (Supplementary Fig. S5). Three sequences correspond to chlorophyllide *a* oxygenase (CAO), an enzyme of chlorophyll *b* synthesis, whereas

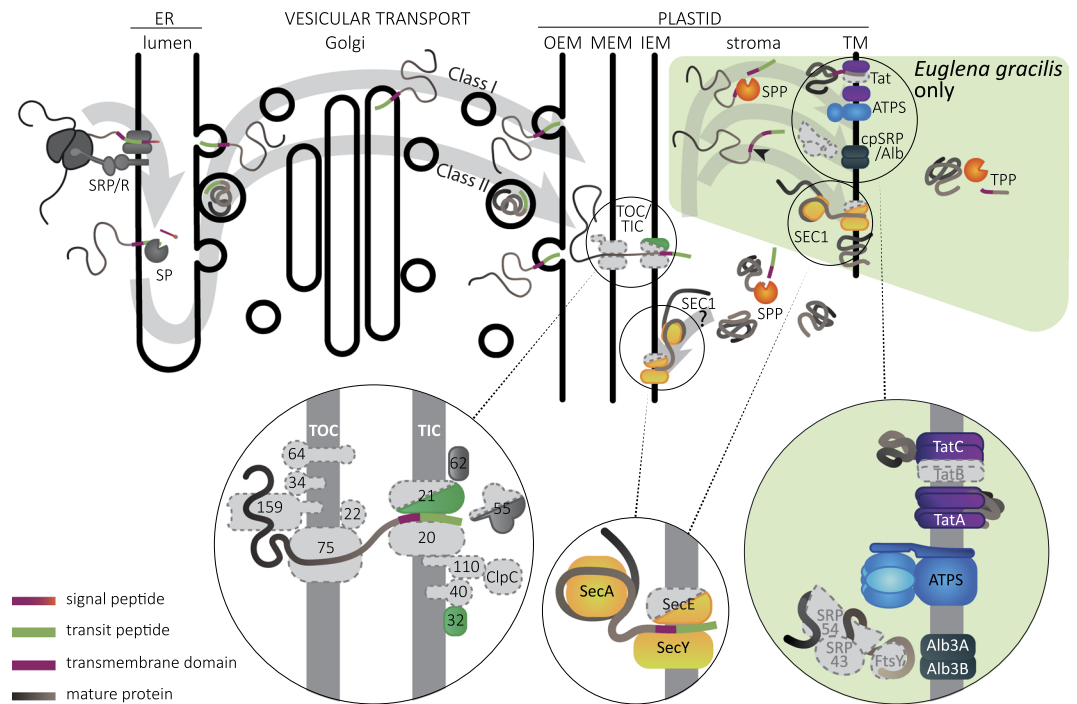


Figure 4. A hypothetical scheme of translocation of plastidial proteins in *Euglena*. The route of plastidial proteins in the *Euglena* cell is schematically depicted, with a zoom-in on individual translocon complexes and protein subunits found in the transcriptomic data analysed. Note that transport across the thylakoid membrane presumably occurs only in the photosynthetic *E. gracilis*, as it is unknown whether the *E. longa* plastid has thylakoids, too. This transport path is taken by proteins with an extended BTS including a third transmembrane domain-like region (arrowhead). The colour code for peptide subdomains is shown in the lower left corner. Note that receptor GTPases Toc34 and Toc159 represent in the figure broader families of paralogous proteins including Toc33 and Toc120/Toc132, respectively. Proteins with similarity to Tic62 and Tic55 were found in euglenophytes, but phylogenetic analyses suggest they are not bona fide orthologs (see main text). Proteins Tic21, Tic55, and SecE are missing in *E. longa*, which is indicated by the half-shape. Proteins without discernible homologs in euglenophytes are shown in grey with a dashed outline. OEM/MEM/IEM/TM: plastid outer, middle, inner envelope membranes and thylakoid membrane; SRP/R: signal recognition particle (receptor) complex; SP: signal peptidase; TOC/TIC: translocon of the outer/inner chloroplast membrane; SPP/TPP: stromal/thylakoid processing peptidase; ATPS: ATP synthase.

the others represent different branches within a broader radiation of plant, algal and cyanobacterial proteins including not only TIC55, but also PAO (pheophorbide *a* oxygenase) and PTC52 (a potential chlorophyllide *a* oxygenase).

While the apparent absence of TIC21 and TIC55-related proteins in *E. longa* is obviously related to the loss of photosynthesis, the lack of discernible homologs of the core TOC/TIC components in euglenophytes in general is striking. It is possible that euglenophytes still possess a form of the TOC/TIC translocon, yet with its components diverged beyond recognition by the bioinformatics tools employed by us. Another possibility is that the original protein import machinery of the green algal progenitor of the euglenophyte plastid was replaced by a novel apparatus that acquired the ability to sort proteins according to similar characteristics (i.e. the presence of an N-terminal plastid transit peptide) as the conventional TOC/TIC translocon. This would not be without a precedent. In algae with rhodophyte-derived four membrane-bound plastids, the N-terminal transit peptide-like region not only enables import into the plastid stroma via the TOC/TIC translocon, but first serves as a sorting signal (recognized by a hitherto uncharacterized receptor) for translocation of the preprotein across the second outermost plastid membrane into the periplastid space mediated by a unique machinery called SELMA⁵³.

After a protein has passed into the plastid, its presequence needs to be cleaved off by the stromal processing peptidase, conserved in *E. longa* as well as other euglenophytes (Supplementary Table S3). Photosynthetic plastids need to translocate specific proteins further into the lumen or the membrane of thylakoids. Several different machineries mediating this step are known, including the Tat (twin-arginine translocase), SRP/Alb3 (Signal Recognition Particle/Albino3) system, and the SEC translocase⁵⁴. Critical subunits of the Tat translocase (TatA and TatC) and two different forms of the Alb3 protein can be readily identified in the transcriptome of *E. gracilis* and *Eutreptiella* spp., whereas no such homologs can be found in the *E. longa* transcriptome assembly (Fig. 4; Supplementary Table S3). This is consistent with the role of these proteins in translocating exclusively the components of the photosynthetic machinery. In addition, the Tat translocase depends on the electric potential across the thylakoid membrane⁵⁵, and hence would be useless in plastids lacking a mechanism to generate it. In non-photosynthetic plastids the transmembrane electrochemical proton gradient is maintained by the ATP

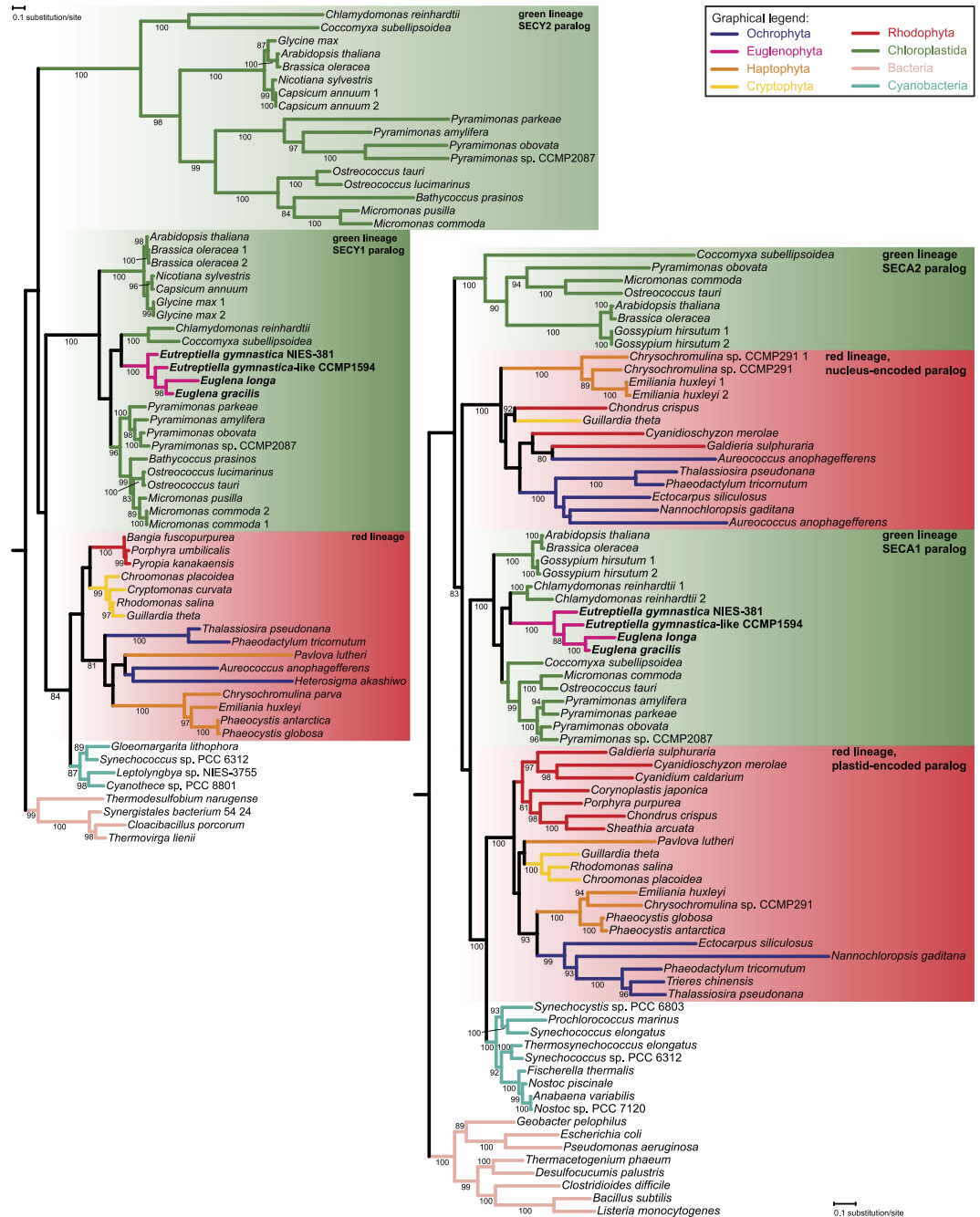


Figure 5. Phylogenetic analysis of the SecA and SecY subunits of the SEC translocon. The maximum likelihood tree of SecA and SecY proteins documents that the euglenophyte proteins are orthologs of the chlorophyte SECA1 and SECY1, respectively. Bootstrap support values are given when ≥ 80 .

synthase at the expense of ATP. Indeed, Tat is missing from non-photosynthetic plastids that lost ATP synthase³². The absence of both Tat and ATP synthase in *E. longa* is thus fully consistent with these insights. More unexpected is the absence of the chloroplast signal recognition particle (cpSRP) components (cpSRP54 and cpSRP43) and its receptor (cpFtsY) in all euglenophytes (Fig. 4), since these proteins are conserved in photosynthetic plastids in general (cpSRP54 and cpFtsY) or in plants and green algae (cpSRP43)^{56,57} and are important for the delivery of substrate proteins to the Alb3 insertase⁵⁴. How the absence of cpSRP/cpFtsY affects the function of the euglenophyte Alb3 remains to be elucidated.

In contrast, the third plastid translocase, SEC, is conserved in both *Euglena* species as well as in *Eutreptiella* spp. (Fig. 4; Supplementary Table S3). Two different plastid SEC systems were described in plants, one located in thylakoid membranes and the other in the chloroplast inner envelope membrane (IEM)⁵⁸. We retrieved only a single SecA and SecY subunit homolog in each *Euglena* species, and in *E. gracilis* we found a single homolog of the third SEC subunit, SecE, whereas *E. longa* apparently lacks it (its absence was confirmed by searching raw

RNA-seq reads). Our phylogenetic analyses revealed that the euglenophyte SecY and SecA proteins are related to the plant and green algal components of the thylakoid-associated SEC1 system (Fig. 5); the phylogeny of the short and poorly conserved SecE protein was not analysed. The apparent absence of the SEC2 complex in euglenophytes is intriguing, but can be explained by at least three different scenarios: (1) the SEC1 complex is in fact not exclusive for thylakoids and operates also in the IEM to facilitate insertion of some IEM proteins, whereas the SEC2-specific substrates have been lost from euglenophytes; (2) SEC1 operates also at the IEM and has taken over some of the SEC2 substrates; (3) there is no SEC machinery at the IEM.

Studies in plant plastids have so far identified only three putative SEC2 substrates, the TIC complex components TIC40 and TIC110 (both apparently missing from euglenophytes) and the FTSH12 protein, one of the paralogs of an expanded family of membrane-bound proteases⁵⁹. We identified FTSH protease homologs in *E. longa* and the three photosynthetic euglenophytes and performed a phylogenetic analysis by including the well-annotated FTSH protease set from *Arabidopsis thaliana* (Supplementary Fig. S6). Like their *A. thaliana* homologs⁶⁰, the euglenophyte FTSH proteases are presumably mitochondrion- or plastid-targeted and their predicted localisation is highly congruent with the phylogenetic relationships of the proteins. *E. longa* and *E. gracilis* proved to encode essentially the same set of mitochondrial FTSH proteases (a difference being the existence of two highly similar variants of one homolog in *E. longa* and both *Eutreptiella* species, possibly due to very recent gene duplications). In comparison, *E. longa* possesses a reduced complement of plastid-localised FTSH proteases compared with the photosynthetic euglenophytes (three versus six to nine), most likely due to the loss of paralogs specialised to act on photosynthesis-related proteins. Interestingly, all euglenophyte plastid-localised FTSH proteases group with *A. thaliana* homologs associated with the thylakoid membrane (FTSH 1, 2, 5, and 8)⁶⁰, hence our *in silico* analysis does not recover any obvious FTSH protease candidates to localise to the euglenophyte IEM. Crucially, the absence of a euglenophyte ortholog of FTSH12 (Supplementary Fig. S6) is consistent with the lack of the SEC2 complex.

Nevertheless, it is still possible that some of the hitherto unidentified SEC2 substrates have been preserved in euglenophytes, but their import was taken over by SEC1. Operation of the same SEC translocon in both the thylakoids and the IEM was the primitive state in plastid evolution as documented by the arrangement in cyanobacteria⁶¹ and glaucophytes⁶². Furthermore, the presence of the SEC1 complex in *E. longa* also supports its localisation to the IEM, since this non-photosynthetic species is unlikely to have thylakoids (although this needs to be proven by electron microscopy). The apparent lack of a plastidial SecE homolog in *E. longa* may reflect functional simplification of the translocase associated with the loss of its predominant substrates (i.e. proteins of the photosynthetic machinery), although we cannot rule out the possibility that it was missed due to potentially incomplete representation of *E. longa* genes in the transcriptome assembly. Finally, it is possible that no SEC machinery is located in the IEM of the euglenophyte plastids and all proteins residing in this membrane (presumably a number of metabolite transporters and components of the elusive protein import machinery) reach their destination via the so-called stop-transfer pathway, i.e. lateral insertion into the IEM during import of the protein⁵⁴. Whereas in most studied plastids this pathway is utilised by only a subset of IEM proteins, the apparently unusual protein import apparatus in euglenophyte plastids (see above) might suggest that this mechanism serves as a general route for the IEM proteins delivery.

We also used our transcriptome assembly to investigate whether *E. longa* has preserved the conventional plastid division machinery, comprising proteins functioning outside (e.g. the dynamin family GTPase Arc5/DRP5B) and inside (e.g. the tubulin homolog FtsZ and Min proteins) the plastid⁶³. None of these proteins could be identified in our data, and *E. gracilis* and *Eutreptiella* spp. also appear to lack them (Supplementary Table S3). The absence of Arc5/DRP5B is not extraordinary, since this protein is also missing in the sequenced representatives of glaucophytes, cryptophytes, chlorarachniophytes, and myxozoans⁶⁴. The absence of FtsZ in *E. longa* would not be particularly surprising either, since myxozoans with non-photosynthetic plastids (apicomplexans and *Perkinsus marinus*) are devoid of it, too. However, the apparent lack of FtsZ in euglenophytes in general is noteworthy, because all organisms with photosynthetic plastids studied to date do keep FtsZ, typically as multiple paralogs⁶⁴. Our analyses thus suggest that the original plastid division mechanism of the green algal donor was substantially simplified or modified during the endosymbiotic integration of the euglenophyte secondary plastid.

A plastid-targeted Rho factor homolog in euglenophytes acquired by HGT from bacteria. The plastid genomes of euglenophytes including *E. longa* encode four subunits of the RNA polymerase responsible for its transcription, namely alpha (RpoA), beta (RpoB), beta' (RpoC1), and beta'' (RpoC2) (Supplementary Fig. S3). The fifth putative subunit, i.e. the sigma factor (RpoD), is found in the transcriptome of *E. longa*, *E. gracilis*, and both *Eutreptiella* species (Supplementary Table S3), completing the conventional cyanobacteria-derived RNA polymerase holoenzyme conserved in plastids in general⁶⁵. However, while surveying the *E. longa* transcriptome for potential plastid-targeted proteins we unexpectedly encountered a homolog of the Rho factor, a highly conserved and widespread component of eubacterial transcription machinery⁶⁶ that, to the best of our knowledge, has never been reported from eukaryotes. This is evidently not due to a bacterial contamination. The respective contig carries the characteristic SL sequence at its 5'-end (Supplementary Table S3), the encoded protein has an N-terminal, BTS-like extension compared to the bacterial homologs (Supplementary Fig. S7), and closely related homologs exist in *E. gracilis* and both *Eutreptiella* species (although the *E. gymnastica*-like CCMP1594 sequence is truncated and the 5'-end could not be completed by iterative read mapping) (Fig. 6; Supplementary Table S3).

The Rho factor is a homohexameric ATP-driven RNA helicase that is critical for proper termination of transcription of a sizeable proportion of genes in bacteria⁶⁶. Despite being so important and prevalent, it has not been retained in the endosymbiotic organelles of eukaryotes; the few eukaryotic hits retrieved by a blastp search against the NCBI nr protein database are all contaminants from bacteria (Supplementary Table S6). In the case of the plastid this seems to be due to an earlier Rho factor loss in Cyanobacteria, as documented by our searches that failed to identify convincing Rho factor homologs in this bacterial phylum (the few hits all seem to be contaminants from other bacteria; Supplementary Table S6). Hence, the emergence of the Rho factor in the euglenophyte

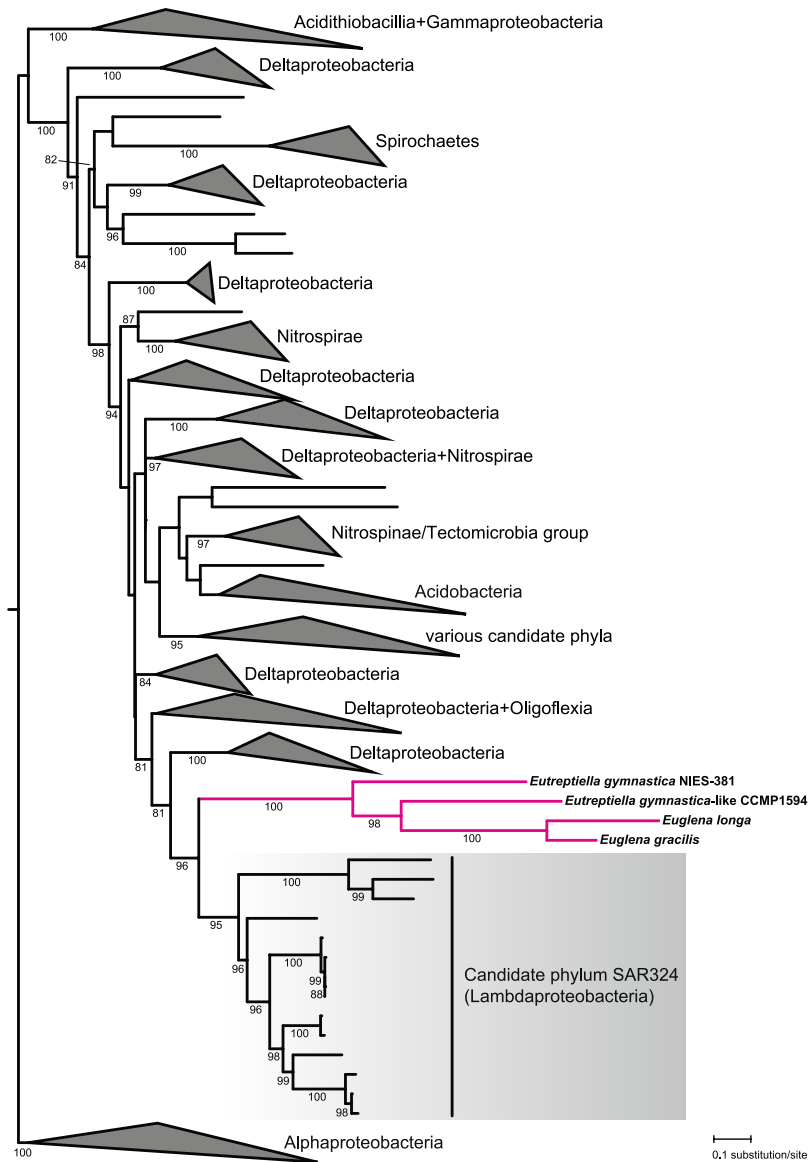


Figure 6. The phylogenetic position of the euglenophyte transcription-termination factor Rho among bacterial homologs. The tree was inferred by using the maximum likelihood method. Bootstrap support values are given when ≥ 80 . For simplicity, various broader clades were collapsed and the taxonomic provenance of the sequences included in them is indicated at the triangles. A full version of the tree is provided in the Supplementary Data S1.

plastid is indeed striking. Our phylogenetic analysis indicates that the donor of the euglenophyte Rho factor was related to the recently recognized bacterial phylum SAR324⁶⁷ (see also <http://gtdb.ecogenomic.org/>) (Fig. 6; Supplementary Data S1). This bacterial lineage (also called Candidatus Lambdaproteobacteria) so far lacks cultured representatives and all available genomic data are derived from metagenomes or single-cell genome sequencing (e.g. refs^{67–71}). The euglenophyte Rho factor thus represents an interesting example of a HGT-derived eukaryotic gene whose actual bacterial source could be properly identified only owing to the recent substantial improvement of the genome sampling of the bacterial phylogenetic diversity.

What might be the function of the Rho factor in the euglenophyte plastid? Sequence comparison of the euglenophyte and bacterial Rho proteins reveals that the motifs critical for their function have not been affected by the gene transfer into the eukaryotic lineage (Supplementary Fig. S7). Therefore, there is no reason to assume that the function of the euglenophyte Rho factor is different from the bacterial prototype at the biochemical level and the obvious null hypothesis is that the euglenophyte Rho factor is involved in transcription termination in the plastid. The Rho factor in bacteria terminates transcription in only a specific subset of genes, but a common Rho-specific termination signal was not found⁶⁶. Hence, it cannot be decided at the moment which euglenophyte plastid genes are the best candidates for being regulated by the Rho factor. Future biochemical experiments should enable us to test whether the Rho factor is involved in transcription termination in euglenophyte plastids and if so, which genes are its specific targets.

Conclusions

We have sequenced and assembled the transcriptome of an interesting organism and a useful model for studying plastid reduction accompanying the loss of photosynthesis – a widespread phenomenon among non-photosynthetic plants, algae and protists⁷². Our analyses suggest that our *E. longa* transcriptome assembly provides a good representation of nucleus-encoded plastid-targeted proteins in general. We also confirmed that the N-terminal plastid-targeting presequences in *E. longa* exhibit the same characteristic structure as in *E. gracilis*, which opens up a possibility of a systematic bioinformatic survey of the *E. longa* plastid proteome. Complete reconstruction of the metabolic pathways localised in the non-photosynthetic plastid of *E. longa* will help to understand its physiological role(s). Work on this task is in progress in our laboratories.

In this paper, we exploited the transcriptome assemblies of *E. longa* and its photosynthetic relatives to illuminate the molecular machinery responsible for plastid biogenesis and division. As expected, we observed selective loss of components linked to the loss of photosynthesis in *E. longa*, but strikingly, our results revealed that euglenophytes as a whole have lost many components present in the green algal donors of their plastid and conserved in plants and algae in general. Most notable is the elusive nature of the euglenophyte mechanisms of plastid protein import and plastid division. We cannot formally rule out that some of the apparent absences of homologs of common plastidial components are due to the character of the sequence data available for euglenophytes, i.e. transcriptome assemblies rather than full genome sequences. However, the consistent pattern of these absences (from all four species analysed or, in cases of components directly or indirectly related to photosynthesis, from the non-photosynthetic species *E. longa*), makes this explanation unlikely and points to an unexpected degree of simplification (or replacement) of the original plastid-associated molecular machineries during the integration of the green alga-derived secondary plastid into the euglenophyte lineage.

On the other hand, the identification of the plastid-targeted Rho factor is a manifestation of the well documented significance of HGT from bacteria in the evolution of photosynthetic eukaryotes and their plastid^{73,74}, and points to a functional enrichment of the euglenophyte plastid that is unprecedented among eukaryotes. Previous phylogenetic analyses unveiled a mosaic nature of the euglenophyte plastid proteome, indicating that many proteins, such as some enzymes of the Calvin cycle or the MEP pathway of isoprenoid biosynthesis, were gained by HGT from various algal sources different from the donor of the plastid itself^{18,19,75}. Our results suggest that the euglenophyte plastid proteome has an even more complex evolutionary origin, including a contribution from bacteria. Our results thus emphasize the need to revive the interest in how the euglenophyte plastids have evolved and function as cellular organelles.

Methods

Culture conditions, RNA isolation and mRNA extraction, cDNA synthesis, and PCR. *Euglena longa* strain CCAP 1204-17a was cultivated statically in the dark or under constant illumination at 26 °C in Cramer-Myers medium⁷⁶ supplemented with ethanol (0.8% v/v). The cultures were not completely axenic, but the contaminating bacteria were kept at as low level as possible. RNA was isolated using TRIzol[®] Reagent (Invitrogen, Carlsbad, USA) and mRNA was then extracted using PolyATtract mRNA Isolation Systems III (Promega, Madison, USA). cDNA synthesis was carried out with an oligo(dT) primer using Transcriptor First Strand cDNA Synthesis Kit (Roche, Basel, Switzerland). Sequences of all primers used in PCR experiments are listed in Supplementary Table S1. PCR products were amplified from 30 ng of *E. longa* cDNA using MyTaq[™] Red DNA Polymerase (Bioline, London, UK). The presence of SL-sequence at the 5'-end of transcripts was tested using the same PCR experimental design as described previously^{29,77}. PCR conditions were as follows: 95 °C for 1 min; 35 cycles of 95 °C for 15 sec, 50 °C for 15 sec, 72 °C for 1 min, and the final extension at 72 °C for 5 min. The PCR products were purified (Gel/PCR DNA Fragments Extraction Kit, Geneaid Biotech, New Taipei City, Taiwan) and their identity was verified by sequencing (Macrogen Europe, Amsterdam, Netherlands).

***E. longa* transcriptome sequencing and assembly.** Library preparation and sequencing was performed by GATC Biotech (Germany). Briefly, libraries were prepared from mRNA isolated from dark-grown and light-grown *E. longa* using random-primed strand-specific cDNA synthesis and sequenced on an Illumina HiSeq2000 platform. A total of 47,442,811 paired-end 80-bp reads were obtained. Contamination from *Homo sapiens* and *Capsicum annuum* identified in a preliminary transcriptome assembly was removed by mapping the reads to the genome sequences of the respective species using Deconseq 0.43⁷⁸. The remaining reads were adapter- and quality-trimmed by Trimmomatic 0.33⁷⁹. The final read assembly was performed using the ABySS software 1.52⁸⁰ (k-mers 31–51), then fused with Trans-ABYSS 1.48⁸¹, Trinity r20140717⁸² (k-mers 31 and 25), and SOAPdenovo-Trans 1.04⁸³ (k-mers 31 and 33), followed by merging the contigs (>99% sequence identity over 150 nt) using CAP3 12/21/07⁸⁴. The completeness of the assembly was assessed by a BUSCO search of conserved eukaryotic orthologs using the transcript mode and eukaryotaV1 and eukaryotaV2 sets of orthologs⁸⁵.

Sequence searches and phylogenetic analyses. Homologs of proteins of interest were searched in the final transcriptome assembly using local tBLASTn⁸⁶. The contigs representing candidate hits were translated in all six frames and the corresponding protein model was selected. To possibly detect sequences not identified by tBLASTn, we employed HMMER 3.1b2, a more sensitive method of homology detection based on profile hidden Markov models⁸⁷. Profile HMMs were built from seed alignments of the proteins families of interest obtained from the Pfam database and used to search the transcriptome of *E. longa* and both available *E. gracilis* transcriptomes (accession numbers GDJR00000000.1, ref.¹⁰, and GEFR00000000.1, ref.⁸) translated in all six frames by an in-house python script, and of both *Eutreptiella* species (reassemblies available at zenodo.org, <https://doi.org/10.5281/zenodo.257410>). Searches for candidates for TOC/TIC machinery components in euglenophytes were done in parallel by iterative HMMER searches to enable identification of even more distant homologs. Specifically, alignments of full proteins from the RefSeq database and alignments of separate domains from the

Conserved Domains Database⁸⁸ (CDD) were used to construct the initial profile HMMs for searching transcriptomes of several chlorophytes including *Pyramimonas parkeae* and *Pyramimonas obovata* (sequenced in frame of the MMETSP project¹¹), which represent close relatives of the putative euglenophyte plastid donor^{1,12–14}. The identified chlorophyte homologs were re-aligned with sequences of the initial reference set using ClustalW⁸⁹, new profile HMMs were built and euglenophyte sequence data were searched with them. To further confirm the absence of some genes in the transcriptome of *E. longa*, tBLASTn searches were carried out against the unassembled raw reads using the respective protein sequences from *E. gracilis* as queries. Iterative searches of raw reads were also used in attempts to extend termini of contigs that proved to have truncated coding sequences (taking into account also linking information provided by pair-end reads).

Based on the analysis of high-confidence candidates for *E. longa* plastid-targeted proteins and the known structure of plastidial BTSs in *E. gracilis*, the criteria for identifying a protein as plastid-targeted were set as follows: (1) the signal peptide was predicted by the PrediSi⁹⁰ or PredSL⁹¹ programs; (2) one or two transmembrane domains at the N-terminus of the protein were predicted by the TMHMM program⁹² available online or implemented in the Geneious 10.1.3 software⁹³. The resulting set of sequences was further filtered by checking for the presence of a plastid transit peptide, which was predicted by MultiLoc2⁹⁴ after *in silico* removal of the signal peptide or the first transmembrane domain.

Phylogenetic analyses were carried out for selected proteins. Homologs were identified by BLAST searches in the non-redundant protein sequence database at NCBI and protein models of selected organisms from JGI (Joint Genome Institute, jgi.doe.gov), Ensembl (www.ensembl.org), and MMETSP (Marine Microbial Eukaryote Transcriptome Sequencing Project¹¹; original assemblies from marinemicroeukaryotes.org and reassemblies currently available at zenodo.org, <https://doi.org/10.5281/ZENODO.257410>). Sequences were aligned using the MAFFT 7 tool⁹⁵ and poorly aligned positions were eliminated with the trimAL tool⁹⁶. The alignments were manually refined using AliView⁹⁷ and ambiguously aligned positions were removed. For presentation purposes, alignments were processed using the program CHROMA⁹⁸. Maximum likelihood (ML) trees were inferred from the alignments using the best-fitting substitution model as determined by the IQ-TREE software⁹⁹ and employing the strategy of rapid bootstrapping followed by a “thorough” ML search with 1,000 bootstrap replicates. The list of species, and the number of sequences and amino acid positions are present in Supplementary Tables S7–12 or each phylogenetic tree. The multiple sequence alignments used for phylogenetic analyses are available upon request from the corresponding author.

Data Availability

The raw sequencing data and the final assembly of the *E. longa* transcriptome are available at NCBI (www.ncbi.nlm.nih.gov) as BioProject PRJNA471257.

References

- Leander, B. S., Lax, G., Karnkowska, A. & Simpson, A. G. B. Euglenida in Handbook of the Protists (eds John M. Archibald *et al.*), 1–42 (Springer International Publishing, 2017).
- Campbell, D. A., Thomas, S. & Sturm, N. R. Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect.* **5**, 1231–1240 (2003).
- Clayton, C. E. Gene expression in kinetoplastids. *Curr Opin Microbiol.* **32**, 46–51 (2016).
- Ebenezer, T. E. *et al.* Unlocking the biological potential of *Euglena gracilis*: evolution, cell biology and significance to parasitism. bioRxiv, <https://doi.org/10.1101/228015>, (2017).
- Hoffmeister, M. *et al.* *Euglena gracilis* rhoDoquinone:ubiquinone ratio and mitochondrial proteome differ under aerobic and anaerobic conditions. *J Biol Chem.* **279**, 22422–22429 (2004).
- Frantz, C., Ebel, C., Paulus, F. & Imbault, P. Characterization of *trans*-splicing in Euglenoids. *Curr Genet.* **37**, 349–355 (2000).
- Liang, X. H., Haritan, A., Uliel, S. & Michaeli, S. *Trans* and *cis* splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot Cell.* **2**, 830–840 (2003).
- Ebenezer, T. E., Carrington, M., Lebert, M., Kelly, S. & Field, M. C. *Euglena gracilis* genome and transcriptome: Organelles, nuclear genome assembly strategies and initial features. *Adv Exp Med Biol.* **979**, 125–140 (2017).
- O’Neill, E. C. *et al.* The transcriptome of *Euglena gracilis* reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. *Mol Biosyst.* **11**, 2808–2820 (2015).
- Yoshida, Y. *et al.* *De novo* assembly and comparative transcriptome analysis of *Euglena gracilis* in response to anaerobic conditions. *BMC Genomics.* **17**, 182 (2016).
- Keeling, P. J. *et al.* The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol.* **12**, e1001889 (2014).
- Jackson, C., Knoll, A. H., Chan, C. X. & Verbruggen, H. Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green plastids. *Sci Rep.* **8**, 1523 (2018).
- Turmel, M., Gagnon, M. C., O’Kelly, C. J., Otis, C. & Lemieux, C. The chloroplast genomes of the green algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. *Mol Biol Evol.* **26**, 631–648 (2009).
- Vanclová, A. M. G., Hadariová, L., Hrdá, Š. & Hampl, V. Chapter Nine - Secondary Plastids of Euglenophytes in Advances in Botanical Research Vol. 84 (ed. Yoshihisa Hirakawa), 321–358 (Academic Press, 2017).
- Durnford, D. G. & Schwartzbach, S. D. Protein targeting to the plastid of *Euglena*: Biochemistry, cell and molecular biology Vol. 979 (eds Steven D. Schwartzbach & Shigeru Shigeoka), 183–205 (Springer International Publishing, 2017).
- Durnford, D. G. & Gray, M. W. Analysis of *Euglena gracilis* plastid-targeted proteins reveals different classes of transit sequences. *Eukaryot Cell.* **5**, 2079–2091 (2006).
- Kořený, L. & Oborník, M. Sequence evidence for the presence of two tetrapyrrole pathways in *Euglena gracilis*. *Genome Biol Evol.* **3**, 359–364 (2011).
- Lakey, B. & Triemer, R. The tetrapyrrole synthesis pathway as a model of horizontal gene transfer in euglenoids. *J Phycol.* **53**, 198–217 (2017).
- Markunas, C. M. & Triemer, R. E. Evolutionary history of the enzymes involved in the Calvin-Benson cycle in euglenids. *J Eukaryot Microbiol.* **63**, 326–339 (2016).

20. Marin, B., Palm, A., Klingberg, M. & Melkonian, M. Phylogeny and taxonomic revision of plastid-containing euglenophytes based on SSU rDNA sequence comparisons and synapomorphic signatures in the SSU rRNA secondary structure. *Protist*. **154**, 99–145 (2003).
21. Gockel, G. & Hachtel, W. Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa*. *Protist*. **151**, 347–351 (2000).
22. Hadariová, L., Vesteg, M., Birčák, E., Schwartzbach, S. D. & Krajčovič, J. An intact plastid genome is essential for the survival of colorless *Euglena longa* but not *Euglena gracilis*. *Curr Genet*. **63**, 331–341 (2017).
23. Záhonová, K., Füssy, Z., Oborník, M., Eliáš, M. & Yurchenko, V. RuBisCO in non-photosynthetic alga *Euglena longa*: divergent features, transcriptomic analysis and regulation of complex formation. *PLoS ONE*. **11**, e0158790 (2016).
24. Webster, D. A., Hackett, D. P. & Park, R. B. The respiratory chain of colorless algae: III. Electron microscopy. *J Ultrastruct Res*. **21**, 514–523 (1967).
25. Nudelman, M. A., Rossi, M. S., Conforti, V. & Triemer, R. E. Phylogeny of Euglenophyceae based on small subunit rDNA sequences: Taxonomic implications. *J Phycol*. **39**, 226–235 (2003).
26. Záhonová, K. *et al.* Extensive molecular tinkering in the evolution of the membrane attachment mode of the Rheb GTPase. *Sci Rep*. **8**, 5239 (2018).
27. Russell, A. G., Watanabe, Y., Charette, J. M. & Gray, M. W. Unusual features of fibrillarin cDNA and gene structure in *Euglena gracilis*: evolutionary conservation of core proteins and structural predictions for methylation-guide box C/D snoRNPs throughout the domain Eucarya. *Nucleic Acids Res*. **33**, 2781–2791 (2005).
28. Nagai, M. & Yoneda, Y. Small GTPase Ran and Ran-binding proteins. *Biomol Concepts*. **3**, 307–318 (2012).
29. Záhonová, K. *et al.* A small portion of plastid transcripts is polyadenylated in the flagellate *Euglena gracilis*. *FEBS Lett*. **588**, 783–788 (2014).
30. Maier, U. G. *et al.* Massively convergent evolution for ribosomal protein gene content in plastid and mitochondrial genomes. *Genome Biol Evol*. **5**, 2318–2329 (2013).
31. Figueroa-Martínez, F., Nedelcu, A. M., Smith, D. R. & Reyes-Prieto, A. The plastid genome of *Polytoma uvella* is the largest known among colorless algae and plants and reflects contrasting evolutionary paths to nonphotosynthetic lifestyles. *Plant Physiol*. **173**, 932–943 (2017).
32. Kamikawa, R. *et al.* Proposal of a twin arginine translocator system-mediated constraint against loss of ATP synthase genes from nonphotosynthetic plastid genomes. *Mol Biol Evol*. **32**, 2598–2604 (2015).
33. Suzuki, S., Endoh, R., Manabe, R. I., Ohkuma, M. & Hirakawa, Y. Multiple losses of photosynthesis and convergent reductive genome evolution in the colourless green algae *Prototheca*. *Sci Rep*. **8**, 940 (2018).
34. Habib, S., Vaishya, S. & Gupta, K. Translation in organelles of apicomplexan parasites. *Trends Parasitol*. **32**, 939–952 (2016).
35. Mailu, B. M. *et al.* *Plasmodium* apicoplast Gln-tRNA^{Gln} biosynthesis utilizes a unique GatAB amidotransferase essential for erythrocytic stage parasites. *J Biol Chem*. **290**, 29629–29641 (2015).
36. Sheppard, K. *et al.* From one amino acid to another: tRNA-dependent amino acid biosynthesis. *Nucleic Acids Res*. **36**, 1813–1825 (2008).
37. Gile, G. H., Moog, D., Slamovits, C. H., Maier, U. G. & Archibald, J. M. Dual organellar targeting of aminoacyl-tRNA synthetases in diatoms and cryptophytes. *Genome Biol Evol*. **7**, 1728–1742 (2015).
38. Ban, N. *et al.* A new system for naming ribosomal proteins. *Curr Opin Struct Biol*. **24**, 165–169 (2014).
39. Adams, K. L., Daley, D. O., Whelan, J. & Palmer, J. D. Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts. *Plant Cell*. **14**, 931–943 (2002).
40. Bieri, P., Leibundgut, M., Saurer, M., Boehringer, D. & Ban, N. The complete structure of the chloroplast 70S ribosome in complex with translation factor pY. *EMBO J*. **36**, 475–486 (2017).
41. Bubunenko, M. G., Schmidt, J. & Subramanian, A. R. Protein substitution in chloroplast ribosome evolution. A eukaryotic cytosolic protein has replaced its organelle homologue (L23) in spinach. *J Mol Biol*. **240**, 28–41 (1994).
42. Takagi, M., Absalon, M. J., McLure, K. G. & Kastan, M. B. Regulation of p53 translation and induction after DNA damage by ribosomal protein L26 and nucleolin. *Cell*. **123**, 49–63 (2005).
43. Zhang, M., Zhang, J., Yan, W. & Chen, X. p73 expression is regulated by ribosomal protein RPL26 through mRNA translation and protein stability. *Oncotarget*. **7**, 78255–78268 (2016).
44. Chan, R. L., Keller, M., Canaday, J., Weil, J. H. & Imbault, P. Eight small subunits of *Euglena* ribulose 1-5 bisphosphate carboxylase/oxygenase are translated from a large mRNA as a polyprotein. *EMBO J*. **9**, 333–338 (1990).
45. Enomoto, T., Sulli, C. & Schwartzbach, S. D. A soluble chloroplast protease processes the *Euglena* polyprotein precursor to the light harvesting chlorophyll a/b binding protein of photosystem II. *Plant Cell Physiol*. **38**, 743–746 (1997).
46. Koziol, A. G. & Durnford, D. G. *Euglena* light-harvesting complexes are encoded by multifarious polyprotein mRNAs that evolve in concert. *Mol Biol Evol*. **25**, 92–100 (2008).
47. Nowitzki, U., Gelius-Dietrich, G., Schwieger, M., Henze, K. & Martin, W. Chloroplast phosphoglycerate kinase from *Euglena gracilis*: endosymbiotic gene replacement going against the tide. *Eur J Biochem*. **271**, 4123–4131 (2004).
48. Zhang, H. & Lin, S. Complex gene structure of the the form II RuBisCO in the dinoflagellate *Prorocentrum minimum* (Dinophyceae). *J Phycol*. **39**, 1160–1171 (2003).
49. Benz, J. P. *et al.* Arabidopsis Tic62 and ferredoxin-NADP(H) oxidoreductase form light-regulated complexes that are integrated into the chloroplast redox poise. *Plant Cell*. **21**, 3965–3983 (2009).
50. Chigri, F. *et al.* Calcium regulation of chloroplast protein translocation is mediated by calmodulin binding to Tic32. *Proc Natl Acad Sci USA*. **103**, 16051–16056 (2006).
51. Kikuchi, S. *et al.* A 1-megadalton translocation complex containing Tic20 and Tic21 mediates chloroplast protein import at the inner envelope membrane. *Plant Cell*. **21**, 1781–1797 (2009).
52. Hauenstein, M., Christ, B., Das, A., Aubry, S. & Hortensteiner, S. A role for TIC55 as a hydroxylase of phyllobilins, the products of chlorophyll breakdown during plant senescence. *Plant Cell*. **28**, 2510–2527 (2016).
53. Maier, U. G., Zauner, S. & Hempel, F. Protein import into complex plastids: Cellular organization of higher complexity. *Eur J Cell Biol*. **94**, 340–348 (2015).
54. Lee, D. W., Lee, J. & Hwang, I. Sorting of nuclear-encoded chloroplast membrane proteins. *Curr Opin Plant Biol*. **40**, 1–7 (2017).
55. Braun, N. A., Davis, A. W. & Theg, S. M. The chloroplast Tat pathway utilizes the transmembrane electric potential as an energy source. *Biophys J*. **93**, 1993–1998 (2007).
56. Träger, C. *et al.* Evolution from the prokaryotic to the higher plant chloroplast signal recognition particle: the signal recognition particle RNA is conserved in plastids of a wide range of photosynthetic organisms. *Plant Cell*. **24**, 4819–4836 (2012).
57. Ziehe, D., Dünschede, B. & Schünemann, D. From bacteria to chloroplasts: evolution of the chloroplast SRP system. *Biol Chem*. **398**, 653–661 (2017).
58. Skaltzky, C. A. *et al.* Plastids contain a second sec translocase system with essential functions. *Plant Physiol*. **155**, 354–369 (2011).
59. Li, Y., Martin, J. R., Aldama, G. A., Fernandez, D. E. & Cline, K. Identification of putative substrates of SEC. 2, a chloroplast inner envelope translocase. *Plant Physiol*. **173**, 2121–2137 (2017).
60. Nishimura, K., Kato, Y. & Sakamoto, W. Chloroplast proteases: Updates on proteolysis within and across suborganellar compartments. *Plant Physiol*. **171**, 2280–2293 (2016).

61. Nakai, M., Sugita, D., Omata, T. & Endo, T. Sec-Y protein is localized in both the cytoplasmic and thylakoid membranes in the cyanobacterium *Synechococcus* PCC7942. *Biochem Biophys Res Commun.* **193**, 228–234 (1993).
62. Yusa, F., Steiner, J. M. & Löffelhardt, W. Evolutionary conservation of dual Sec translocases in the cyanelles of *Cyanophora paradoxa*. *BMC Evol Biol.* **8**, 304 (2008).
63. Chen, C., MacCready, J. S., Ducat, D. C. & Osteryoung, K. W. The molecular machinery of chloroplast division. *Plant Physiol.* **176**, 138–151 (2018).
64. Miyagishima, S. Y., Nakamura, M., Uzuka, A. & Era, A. FtsZ-less prokaryotic cell division as well as FtsZ- and dynamin-less chloroplast and non-photosynthetic plastid division. *Front Plant Sci.* **5**, 459 (2014).
65. Chi, W., He, B., Mao, J., Jiang, J. & Zhang, L. Plastid sigma factors: Their individual functions and regulation in transcription. *Biochim Biophys Acta.* **1847**, 770–778 (2015).
66. Kriner, M. A., Sevostyanova, A. & Groisman, E. A. Learning from the leaders: Gene regulation by the transcription termination factor Rho. *Trends Biochem Sci.* **41**, 690–699 (2016).
67. Parks, D. H. *et al.* A proposal for a standardized bacterial taxonomy based on genome phylogeny. *Nat Biotechnol.*, <https://doi.org/10.1038/nbt.4229>, (2018).
68. Anantharaman, K. *et al.* Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun.* **7**, 13219 (2016).
69. Cao, H. *et al.* Delta-proteobacterial SAR324 group in hydrothermal plumes on the South Mid-Atlantic Ridge. *Sci Rep.* **6**, 22842 (2016).
70. Chitsaz, H. *et al.* Efficient de novo assembly of single-cell bacterial genomes from short-read data sets. *Nat Biotechnol.* **29**, 915–921 (2011).
71. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data.* **5**, 170203 (2018).
72. Hadariová, L., Vesteg, M., Hampl, V. & Krajčovič, J. Reductive evolution of chloroplasts in non-photosynthetic plants, algae and protists. *Curr Genet.* **64**, 365–387 (2018).
73. Huang, J. & Yue, J. Horizontal gene transfer in the evolution of photosynthetic eukaryotes. *J Syst Evol.* **51**, 13–29 (2013).
74. Mackiewicz, P., Bodyl, A. & Moszczyński, K. The case of horizontal gene transfer from bacteria to the peculiar dinoflagellate plastid genome. *Mob Genet Elements.* **3**, e25845 (2013).
75. Maruyama, S., Suzaki, T., Weber, A. P., Archibald, J. M. & Nozaki, H. Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. *BMC Evol Biol.* **11**, 105 (2011).
76. Cramer, M. & Myers, J. Growth and photosynthetic characteristics of *Euglena gracilis*. *Archiv Mikrobiol.* **17**, 384–402 (1952).
77. Mateášiková-Kováčová, B. *et al.* Nucleus-encoded mRNAs for chloroplast proteins GapA, PetA, and PsbO are trans-spliced in the flagellate *Euglena gracilis* irrespective of light and plastid function. *J Eukaryot Microbiol.* **59**, 651–653 (2012).
78. Schmieder, R. & Edwards, R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PLOS ONE.* **6**, e17288 (2011).
79. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* **30**, 2114–2120 (2014).
80. Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res.* **19**, 1117–1123 (2009).
81. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nat Methods.* **7**, 909–912 (2010).
82. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* **29**, 644–652 (2011).
83. Xie, Y. *et al.* SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics.* **30**, 1660–1666 (2014).
84. Huang, X. & Madan, A. CAP3: A DNA sequence assembly program. *Genome Res.* **9**, 868–877 (1999).
85. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* **31**, 3210–3212 (2015).
86. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
87. Eddy, S. R. A new generation of homology search tools based on probabilistic inference. *Genome Inform.* **23**, 205–211 (2009).
88. Marchler-Bauer, A. *et al.* CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* **45**, D200–D203 (2017).
89. Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics.* **23**, 2947–2948 (2007).
90. Hiller, K., Grote, A., Scheer, M., Munch, R. & Jahn, D. PrediSi: prediction of signal peptides and their cleavage positions. *Nucleic Acids Res.* **32**, W375–379 (2004).
91. Petsalaki, E. I., Bagos, P. G., Litou, Z. I. & Hamodrakas, S. J. PredSL: a tool for the N-terminal sequence-based prediction of protein subcellular localization. *Genomics Proteomics Bioinformatics.* **4**, 48–55 (2006).
92. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol.* **305**, 567–580 (2001).
93. Kears, M. *et al.* Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics.* **28**, 1647–1649 (2012).
94. Blum, T., Briesemeister, S. & Kohlbacher, O. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. *BMC Bioinformatics.* **10**, 274 (2009).
95. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* **30**, 772–780 (2013).
96. Capella-Gutierrez, S., Silla-Martinez, J. M. & Gabaldon, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* **25**, 1972–1973 (2009).
97. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics.* **30**, 3276–3278 (2014).
98. Goodstadt, L. & Ponting, C. P. CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics.* **17**, 845–846 (2001).
99. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* **32**, 268–274 (2015).

Acknowledgements

The *Euglena longa* transcriptome sequence data were produced with support of the European Science Foundation, Generation and Analysis of Next Generation Sequence (NGS) data workshop (http://bioinformatics.psb.ugent.be/ngs_workshop/). We thank Lieven Sterck (Department of Plant Biotechnology and Bioinformatics, Ghent University) for his help with obtaining RNA-seq data from *E. longa*. Data analyses were supported by the Czech Science Foundation (17-21409S to ME and 16-24027S to MO), the National Feasibility Programme I of the Czech Republic (TEWEP LO1208), the infrastructure grant “Přístroje IET” (CZ.1.05/2.1.00/19.0388), and the OPVVV project CZ.02.1.01/0.0/0.0/16_019/0000759 (Centre for research of pathogenicity and virulence of parasites). This work was also supported by the Scientific Grant Agency of the Slovak Ministry of Education and the Academy

of Sciences (grant VEGA 1/0535/17 to JK and MV), and by the project ITMS 26210120024 supported by the Research & Development Operational Programme funded by the ERDF. We acknowledge computation resources provided by CERIT-SC and MetaCentrum, Brno, Czech Republic.

Author Contributions

K.Z. and Z.F. performed most bioinformatic analyses and prepared figures and tables. A.M.G.N.V. contributed by analyses of the TOC/TIC system. K.Z., M.V., and J.K. maintained the *E. longa* culture and prepared RNA for transcriptome sequencing. E.B. and V.K. processed and assembled the RNAseq data. M.O. contributed to the design of the study and the paper. M.E. conceived the study, performed some bioinformatics analyses and prepared the first draft of the manuscript. All authors edited the manuscript and approved its final form.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-35389-1>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018