Charles University, Faculty of Science

Univerzita Karlova, Přírodovědecká fakulta

Study program: Parasitology

Studijní program: Parazitologie

Mgr. Anna Novák Vanclová

Evolution of euglenid plastid proteome

Evoluce proteomu plastidu euglenidů

Doctoral thesis

Thesis supervisor: doc. Vladimír Hampl, Ph.D.

Prague, 2019

Declaration of the author:

I declare that I elaborated this thesis independently. I also proclaim that the literary sources were cited properly and neither this work nor a substantial part of it has been used to obtain the same or any other academic degree.

Prohlašuji, že jsem tuto práci zpracovala samostatně a že jsem uvedla všechny použité zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

Mgr. Anna Novák Vanclová

Declaration of the thesis supervisor:

Data presented in this thesis are the result of team collaboration at the Laboratory of evolutionary protistology and cooperation with external associates at the University of Dundee and University of Ostrava. I declare that the involvement and contribution of Anna Novák Vanclová in this work was substantial and sufficient in terms of scope and quality for the award of doctoral degree.

doc. Vladimír Hampl, Ph.D.

Thesis supervisor

Acknowledgements

I would like to give many thanks to my supervisor Vladimír Hampl for sparking my interest in evolutionary protistology in 2010 and for the following years of his support, encouragement, and guidance. Thanks also to my labmates and other colleagues from near and far for their advice and goodwill. Special thanks to Martin Vancl, my brother who was always there to help, and Lukáš Novák, a partner in both science and life who unfailingly supported me through thick and thin.

Table of contents

ABSTRACT

Endosymbiotic gain and transfer of plastids is a widespread evolutionary phenomenon and a major driving force of eukaryotic evolution. The integration of a new organelle is accompanied by changes in its structure, gene content, molecular mechanisms for biogenesis and transport, and re-wiring of the host and organelle metabolic pathways. To understand the course and underlying mechanisms of plastid evolution, it is important to study these processes in variety of secondary algae and notice their differences and similarities.

Euglenophytes gained their plastids from green eukaryotic algae after a long history of heterotrophic lifestyle. In my thesis, I participated in analyses of newly generated sequence datasets: transcriptomes of *Euglena gracilis* and *Euglena longa* and mass spectrometry-determined proteome of *E. gracilis* plastid with especial regard to the potential novelties associated with plastid gain and incorporation. In the resulting publications we particularly focus on plastid protein import machinery and targeting signals and report extremely reduced TIC and completely absent TOC in euglenophyte plastid. Using the proteomic dataset, we predict potential novel plastid protein translocases recruited from ER/Golgi and re-analyze plastid signal domains, characterizing previously overlooked features. Protein inventory of *E. gracilis* plastid suggests complex, in some cases redundant metabolic capacity. Chlorophyll recycling is one of the sources of phytol for reactions not connected to MEP/DOXP pathway. Plastid contribution to amino acid metabolism is very low, if any. We screen the proteome for proteins of other than green algal phylogenetic affiliation and report substantial contribution from "chromists" as well as several cases of LGT from bacteria, including an acquisition of additional SUF pathway.

In summary, the work presented in this thesis provides a solid contribution to plastid proteomics, resource for both basic and applied *Euglena* research and potential foundation for various follow-up studies.

ABSTRAKT

Vznik plastidů endosymbiózou a jejich horizontální šíření je široce rozšířený evoluční jev a jedna z významných hnacích sil evoluce eukaryot. Integrace nové organely je doprovázena změnami v její struktuře, genovém obsahu, biogenezi a importu proteinů a propojením jejích metabolických drah s drahami hostitele. Studium těchto procesů v různých skupinách sekundárních řas a srovnávání mezi nimi je důležité pro porozumění obecným principům evoluce plastidů.

Krásnoočka (Euglenophyta) získala své plastidy od zelených řas po poměrně dlouhém období heterotrofie. V této práci jsem se podílela na analýze nově vygenerovaných sekvenčních datasetů: transkriptomů *Euglena gracilis* a *Euglena longa* a plastidového proteomu *E. gracilis* determinovaného pomocí hmotnostní spektrometrie, a to s ohledem na potenciální inovace související se získáním a integrací plastidu. Ve výsledných publikacích jsme se zaměřili zvláště na složení a evoluci systému pro targeting a import jaderně kódovaných proteinů do plastidu a zjistili, že plastidy krásnooček obsahují extrémně redukovaný TIC a zcela postrádají TOC komplex. Na základě plastidového proteomu jsme identifikovali několik nových potenciálních translokáz odvozených od proteinů endomembránového systému a popsali některé dříve nepovšimnuté vlastnosti N-terminálních targetovacích signálů proteinů importovaných do plastidu. Proteom plastidu *E. gracilis* vypovídá o komplexním, v některých případech redundantním, metabolismu této organely. Recyklace chlorofylu je jedním ze zdrojů fytolu pro reakce nenapojené na plastidovou MEP/DOXP dráhu. Podíl plastidu na metabolismu aminokyselin je velmi nízký, pokud vůbec nějaký. Plastidový proteom jsme rovněž podrobili systematické fylogenetické analýze a zjistili významné množství proteinů původem z „chromist“ a také několik případů laterálního genového přenosu z bakterií, včetně druhé SUF dráhy pro syntézu železosirných center.

Tato práce představuje významný příspěvek k plastidové proteomice, zdroj pro základní i aplikovaný výzkum krásnooček a potenciální základ pro různé typy navazujících studií.

1 INTRODUCTION

1.1 Euglenophyta

1.1.1 Evolutionary position and general characteristics

Euglenids, together with kinetoplastids, diplonemids, and symbiontids, form the phylum Euglenozoa within Discoba, a subgroup of the paraphyletic supergroup Excavata. General characteristics of euglenids shared with other euglenozoans include microtubular corset or pellicle, apical pocket associated with cilliar apparatus comprised of two kinetosomes bearing heteromorphic flagella equipped with paraxonemal rods, and three distinctive microtubular roots, and a single mitochondrion which is often large and reticular and usually has discoidal cristae (Cavalier-Smith 1981; Simpson 1997; Cavalier-Smith 2016; Adl *et al.* 2019). Euglenids have a pellicle composed of proteinaceous strips which enables them to move in a typical slime-like fashion termed metaboly (Triemer & Farmer 1991; Leander *et al.* 2001). They use beta-glucan paramylon as a storage polysaccharide (Barras & Stone 1968) and are quite diverse in regard to nutritional strategies which include eukaryovory, bacteriovory, osmotrophy, and phototrophy (Triemer & Farmer 1991; Leander *et al.* 2001). The latter is the case of euglenophytes, a monophyletic clade of euglenids which harbour from one to few dozen of triple membrane-bound green plastids (Leedale 1967; Gibbs 1981). The early-branching euglenophyte *Rapaza viridis* is an obligatory mixotroph (Yamaguchi *et al.* 2012) which might have substituted these plastids by kleptoplastids from its prey (Karnkowska & Yubuki, personal communication), while several independent lineages such as *Euglena longa* lost the photosynthetic capacity, retain colourless plastids and are secondarily osmotrophic (Gockel & Hachtel 2000; Marin *et al.* 2003).

1.1.2 Morphology and cell structure

Euglenophyte cells are generally large (up to 200 μm), often elongated and capable of some degree of metaboly. The cell plasticity positively correlates with the higher number of pellicle strips and their helical arrangement, while euglenophytes with less flexible or completely rigid cells tend to have fewer strips which are parallel to the cell axis. Some metabolizing euglenophytes have a mineralized extracellular lorica (genera *Trachelomonas*

and *Strombomonas*). Dynamic metaboly is believed to be the ancestral state and its reduction or loss is evolutionarily derived (Leander *et al.* 2007; Karnkowska *et al.* 2015). The more standard and conserved mode of locomotion is swimming using anterior flagellum during which the cell rotates around its longitudinal axis and characteristically sways to the sides. The flagellum contains a paraxonemal rod and is relatively thick. Most euglenophytes have only one emergent flagellum but the ancestral state observed in many primarily heterotrophic euglenids and conserved in the basal paraphylum Eutreptiales are two heteromorphic ones (Leander *et al.* 2001). Euglenophytes possess a typical eyespot formed by carotenoid-rich granula (hence the name from Greek *eu-* ("good") + *glḗnē* ("eye")). It is associated with the paraflagellar pocket and kinetosomes and mediates positive and negative phototaxis, possibly using alternating illumination and shading by the flagellum resulting from cell rotation during swimming to sense the direction of light (James *et al.* 1992; Iseki *et al.* 2002). Some representatives of euglenophytes are sessile with mucilaginous stalks (genus *Colacium*) (Karnkowska *et al.* 2015).

Euglenophytes synthesize paramylon which is stored freely in the cytoplasm in form of grains which can be either monomorphic (all of the same shape and size, ancestral) or dimorphic (two types of grains are present, derived) (Monfils *et al.* 2011). Additionally, paramylon can form caps (one or two) on plastid pyrenoids. Type and distribution of paramylon in the cell is used as morphological diagnostic character (Karnkowska *et al.* 2015).

Euglenophyte plastids can be positioned either axially or parietally in the cell and take multifarious sizes and shapes ranging from simply spherical or oval to variously flattened or concave, extensively lobed, or of ribbon-like appearance, which are also considered as morphological diagnostic characters typical for certain groups, genera, or species. Other characters include pyrenoids, dense regions and sites of carbon fixation, which are generally present and visible, and the presence/absence and topology of the aforementioned paramylon cap (Leedale 1967; Ciugulea & Triemer 2010; Karnkowska *et al.* 2015). The ultrastructure of euglenid plastids, their thylakoids, and enveloping membranes will be discussed in detail and compared to that of other algae in chapter 2.2.2.

1.1.3 Diversity, distribution and ecological significance

Due to their ubiquity in fresh waters, relatively large size, distinctive appearance, and easy collection and cultivation, euglenophytes were among the first protists to be observed and described (Harris 1695; Ehrenberg 1830) and the study of their diversity has a long tradition. To this day, almost one thousand species of euglenophytes in 14 genera have been described (Guiry & Guiry 2017), with disproportionally large number of species assigned to the type genus *Euglena*, which is, at this point, almost undeniably paraphyletic or even polyphyletic (Karnkowska *et al.* 2015; Zakryś *et al.* 2017). A deep, albeit somewhat artificial division splits euglenophytes to Eutreptiales nad Euglenales, with the former being a paraphylum at the base and the latter being a crown monophylum, and with *Rapaza* as a separate lineage sister to both. Eutreptiales (genera *Eutreptia* and *Eutreptiella*) and *Rapaza* exhibit several traits which are not shared with any known member of Euglenales and are considered ancestral: they have two emergent flagella and they are marine. Euglenales can be further subdivided into Phacaceae which lack pyrenoids and developed dimorphic paramylon (genera *Phacus*, *Lepocinclis*, and *Discoplastis*) and Euglenaceae which encompass the rest of the diversity (genera *Euglena*, *Monomorphina*, *Cryptoglena*, *Euglenaria*, *Trachelomonas*, *Strombomonas*, and *Colacium*).

Due to their metabolic plasticity and ability to readily switch to heterotrophy if more convenient organic energy source becomes available, and even survive in anaerobic conditions, euglenophytes inhabit a wide variety of environments, including heavily polluted waters and sediments (Mahapatra *et al.* 2013). It is worth notice that vast majority of the described euglenophyte diversity is freshwater, even though the closest known extant relatives of both the plastid and the plastid acceptor ancestors are marine (Turmel *et al.* 2009; Breglia *et al.* 2013), suggesting (although not proving, as discussed in Jackson *et al.* 2018) that the group originated in the sea. This general absence of euglenophytes in oceans appears to be genuine, albeit peculiar, rather than a result of sampling bias as environmental sequencing efforts repeatedly fail to capture novel or more diversified euglenophyte signals in marine samples (unpublished data; Karnkowska *et al.*, personal communication).

1.1.4 *Euglena gracilis* as a model euglenophyte

Euglenophytes have a very versatile metabolism and are able to synthesize numerous compounds usable in pharmaceutics, such as vitamins A, E and C and other antioxidants, including the aforementioned paramylon which was reported to have remedial effects on certain health conditions in animal models in several studies (Sugiyama *et al.* 2009; Watanabe *et al.* 2013; Nakashima *et al.* 2017; Russo *et al.* 2017). They are also in focus of biotechnological research for their potential utilisation in biofuel industry and wastewater treatment through their production of wax esters under heterotrophic and anaerobic conditions (Mahapatra *et al.* 2013; Krajčovič *et al.* 2015; Nakazawa *et al.* 2015; Ogawa *et al.* 2015). Biochemical capacities of euglenophytes were studied extensively in various strains of *Euglena gracilis* which have been cultivated and used in experiments for over a century. However, unlike for example the green alga *Chlamydomonas* or the diatom *Phaeodactylum*, neither *E. gracilis* nor any other euglenophyte has become a classical routinely transformed and phenotypized model organism so far.

1.1.4.1 Genome and gene expression

One of the main reasons for this is the size and complexity of *E. gracilis* genome and its molecular genetic mechanisms in general. Although two largely complete transcriptomes of *E. gracilis* (O'Neill *et al.* 2015; Yoshida *et al.* 2016) and several lower-quality or draft ones from other representatives of euglenids (Keeling *et al.* 2014; Hrdá *et al.*, unpublished data; Lax *et al.*, unpublished data) have been sequenced, their genomic DNA still resists sequencing. The nuclear genome of *E. gracilis* was estimated to be up to 2 Gbp in size and very rich in non-coding DNA including extensive repetitive regions which notably hamper genome assembly efforts (Ebenezer *et al.* 2017), especially in combination with the presence of multiple types of introns with non-canonical borders (Milanowski *et al.* 2016; Gumińska *et al.* 2018) and pseudogenes which make gene prediction problematic. Moreover, gene expression regulation which involves substantial cis- and trans-splicing of transcripts in *E. gracilis* (Tessier *et al.* 1991; Hastings 2005; Yoshida *et al.* 2016) and seems to take place at a post-transcriptional rather than transcriptional level (Yoshida *et al.* 2016) is not fully understood and potentially complex and unique.

1.1.4.2 Genetic tools

Genetic transformation of *E. gracilis* has been attempted for a long time using various methods. For the introduction and expression of exogenous gene results were achieved by biolistic (Doetsch *et al.* 2001; Ogawa *et al.* 2015) and recently *Agrobacterium*-mediated plasmid delivery (Khatiwada *et al.* 2019). It is also possible to achieve transient gene silencing in *E. gracilis* by introduction of dsRNA, taking advantage of its internal anti-viral RNA interference machinery (Iseki *et al.* 2002; Häder *et al.* 2009; Nakazawa *et al.* 2015; Nasir *et al.* 2018; Novák Vanclová *et al.*; unpublished data). However, none of these methods is robust enough to be used routinely and widely at the present day, so the potential of *E. gracilis* as a euglenid model for studies connecting sequence to function remains mostly unexploited.

1.2 Plastids

1.2.1 The origin of plastids

Plastids are present in many groups of eukaryotes, most of which could be termed algae. Although algae are an artificial, polyphyletic taxon, all their plastids trace back to a single evolutionary event: the integration of a cyanobacterium by a common ancestor of Archaeplastida. The existing plastids were then either inherited vertically or spread horizontally to other lineages.

1.2.1.1 Primary endosymbiosis and related phenomena

The evolutionary event which gave rise to primary plastids present in green algae and plants, rhodophytes, and glaucophytes took place approximately 1.5 billion years ago (Yoon *et al.* 2004; Parfrey *et al.* 2011). The freshwater β-cyanobacterium *Gloeomargarita* was identified as the extant lineage most closely related to the primary plastid ancestor, suggesting that plastids originated in freshwater environment (Ponce-Toledo *et al.* 2017; de Vries & Archibald 2017).

However, the *Gloeomargarita*-like cyanobacterial endosymbiont might have not been the only prokaryotic organism implicated in plastid genesis, as this was presumably

a long-term process of balancing the symbiotic relationship involving multiple trials and errors. According to the "shopping bag" model, laterally transferred genes left behind by previous endosymbionts could have played a significant role in the definitive plastid establishment (Larkum *et al.* 2007). Additionally, a noteworthy number of plastidial genes phylogenetically affiliated to Chlamydiae was noticed and it was hypothesised that an organisms of this group was also an important player in plastid genesis (Huang & Gogarten 2007; Moustafa *et al.* 2008), either as a pre-requisite, suppressing the host cell defense against the newly-arrived cyanobacterial endosymbiont since Chlamydiae are well-adapted to intracellular parasitism (Ball *et al.* 2013), or as a long-term co-inhabitant and partner in the endosymbiotic relationship which would in this case have initially taken place as a "ménage à trois" - between three and not just two organisms (Facchinelli *et al.* 2013; Cenci *et al.* 2016). However, other researchers argue that there was no evolutionary need for such affair and the presence of chlamydial genes can be simply explained as a selectively neutral result of intracellular lifestyle of these bacteria (Domman *et al.* 2015).

The only known case of an independent primary endosymbiosis involves the cercozoan amoeba *Paulinella* which harbours a chromatophore, plastid-like organelle acquired only some 60 million years ago, much later than the plastid of Archaeplastida, and more closely related to α-cyanobacteria *Synechococcus* and *Cyanobium* (Marin *et al.* 2005; Yoon *et al.* 2009; Kim & Park 2016). This peculiar organism allows us to study the early stages of plastid-like organelle evolution and perhaps apply these findings to better understand the origin of classical plastids. For instance, recent studies suggest that there was an extensive lateral influx of genes from various prokaryotes to *Paulinella* that might have been crucial for the establishment of chromatophore as a fully integrated organelle dependent on the host nucleus (Nowack *et al.* 2016).

1.2.1.2 Secondary and higher endosymbioses

Secondary or higher endosymbioses are instances of horizontal plastid transfer between two eukaryotes. They are unquestionably more common in the course of eukaryotic evolution than primary endosymbioses and helped to spread plastids and photosynthetic lifestyle to various groups of organisms throughout most of the eukaryotic supergroups: cryptophytes (Cryptista, currently placed as sister lineage to Archaeplastida), chlorarachniophytes (Rhizaria in TSAR

clade), ochrophytes (Stramenopila in TSAR clade), apicomplexans, chromerids and dinoflagellates (Alveolata in TSAR clade), haptophytes (Haptista, currently placed as sister to TSAR), and euglenophytes (Excavata). Apart from euglenophytes, only chlorarachniophytes and the dinoflagellate genus *Lepidodinium* possess green plastids, in this cases derived from ulvophytes and pedinophytes, respectively (Jackson *et al.* 2018), that can be safely interpreted as secondary based on their phylogenetic signal and, in case of chlorarachniophytes, a presence of nucleomorph, a highly reduced nucleus of the plastid donor, in the periplastidial compartment (Gilson *et al.* 2006; Takahashi *et al.* 2007).

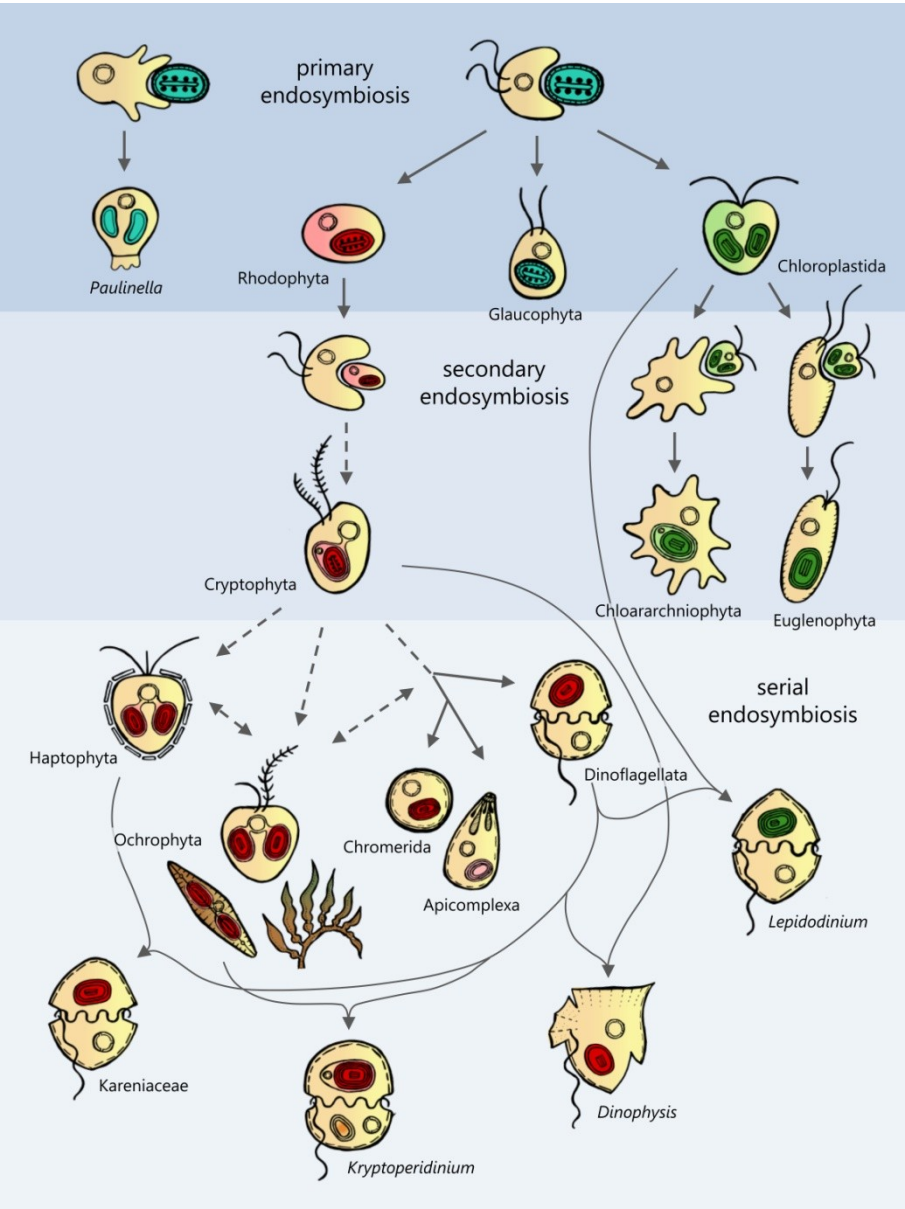

Figure 1: Distribution and possible relationships between known plastids. This model assumes a common origin of all "chromist" plastids and suggests cryptophytes as the original hosts of rhodophyte endosymbiont. The course of most of the following tertiary or higher endosymbiotic events is uncertain (dashed lines).

The relationships between complex red plastids are more complicated and less resolved. It was originally hypothesized that all red secondary algae share a common ancestor that gained its plastid from rhodophytes, and form the group Chromalveolata (Cavalier-Smith 1999). However, the current consensus on pan-eukaryotic tree topology renders such scenario impossible (Baurain *et al.* 2010; Burki *et al.* 2016; Adl *et al.* 2019). The serial hypothesis postulates that plastids of these organisms arose once from rhodophytes and then spread through tertiary, quarternary, or even higher endosymbiotic events (Baurain *et al.* 2010; Dorrell and Smith 2011; Stiller *et al.* 2014; Archibald 2015; Dorrell & Howe 2015; Burki 2017). The order and timing of these transfers are still debated and will require more robust phylogenetic evidence to resolve. However, it is likely that plastids of cryptophytes are truly secondary as they also contain a nucleomorph, in this case identifiable as a remnant of rhodophyte nucleus (Douglas *et al.* 2001; Moore and Archibald 2009). The distribution and relationships between existing plastids are summarized in Figure 1.

1.2.1.3 The origin of euglenid plastids

Euglenophytes harbour green plastids which arose through secondary endosymbiosis (Gibbs 1978). A presence of genes phylogenetically affiliated to plants or cyanobacteria in kinetoplastids led some researchers to hypothesize that plastids were present in the common ancestor of much broader subset of euglenozoans and secondarily lost in most of them (Hannaert *et al.* 2003; Leander 2004; Bodył *et al.* 2010). Current consensus is, however, that the plastid acquisition took place at the basis of euglenophytes (before the split of Rapaza), approximately 500 million years ago (Jackson *et al.* 2018). The prasinophyte alga *Pyramimonas parkeae* was determined as the closest known relative of the plastid donor (Turmel *et al.* 2009).

However, it might have not been the only algal organism that lived in close association or even endosymbiotic relationship with ancestors of euglenophytes, as suggested by the presence of non-negligible amount of genes, often of photosynthetic or otherwise plastid-related function, putatively gained through lateral gene transfer (LGT) from "chromists" (Maruyama *et al.* 2011; Markunas & Triemer 2016; Ponce-Toledo *et al.* 2018). This phenomenon can be observed in chlorarachniophytes as well (Ponce-Toledo *et al.* 2018) and it was proposed that the influx of various plastid-related genes from algal

prey and/or transient symbionts (or kleptoplastids derived from either of these) could have served as a prerequisite for the final establishment of a fully integrated organelle which would be greatly facilitated by an inventory of plastidial genes already present in the host nucleus, rolling out a figurative "red carpet" for the new endosymbiont (Nowack *et al.* 2016; Ponce-Toledo *et al.* 2019).

1.2.2 Plastid membranes and general structure

Plastid structure relates to their photosynthetic function and endosymbiotic origin. The former implies the presence of thylakoids, the sites of light harvesting, electron transport chain, and ATP synthesis dependent on a chemiosmotic potential between plastid stroma and thylakoid lumen. Thylakoids are usually stacked in grana or similar structures, greatly increasing the surface area on which photosynthesis can take place. The endosymbiotic past of plastids resulted in them being surrounded by multiple envelope membranes that are traceable to their respective prokaryotic or eukaryotic donor organisms or organelles since membranes do not arise *de novo* and instead spread epigenetically (Cavalier-Smith 2000).

1.2.2.1 Primary plastids

Primary plastids of Archaeplastida are surrounded by two membranes, both clearly derived from the double cytoplasmic membrane of their cyanobacterial ancestor as suggested by their distinctive lipid and protein composition. Galactolipids monogalactosyldiacylglycerol (MGDG) and digalactosyldiacylglycerol (DGDG), glycolipid sulfoquinovosyldiacylglycerol (SQDG) and phosphatidylglycerol (PG) are the typical constituents of plastid membranes and are implicated in photosynthetic functions and their regulation (Boudière *et al.* 2014). The plastids of glaucophytes exhibit additional structural features linking them to cyanobacteria: they retain a peptidoglycan wall and carboxysomes (Raven 2003). Thylakoids of primary plastids can be either single and free-floating as in rhodophytes and glaucophytes, or stacked in grana (layered discs) connected by lamellae as in most green algae and plants. Pyrenoids, the sites of carbon-fixation, RuBisCO concentration, and starch production, are generally present in rhodophytes and green algae, but absent in multicellular plants. In the plastids of some green algae, storage starch granules can be present, as well as intra-plastidial eyespots comprising of carotenoid granules, rhodopsins, and other proteins.

1.2.2.2 Complex plastids

Secondary and higher plastids are (with one recently described exception) surrounded by more than two membranes which results from their more complex evolutionary journey that lead through one or multiple eukaryotic hosts. The most common number of envelope membranes is four with the innermost two being homologous to the original cyanobacteria-derived membranes of primary plastids, the second outermost one descended from cytoplasmic membrane of the previous eukaryotic host (as demonstrated by the localization of nucleomorph in the groups that possess this residual nucleus (Douglas *et al.* 2001; Gilson *et al.* 2006)), and the outermost being directly derived from and in some cases (e.g. diatoms, haptophytes, and cryptophytes) physically connected to the endomembrane system of the current host organism, forming chloroplast endoplasmic reticulum, CER (Sheiner & Striepen 2013; Flori *et al.* 2016). In some organisms with complex plastids, namely euglenids and peridinin-containing dinoflagellates, one of these membranes was lost (most likely one of the eukaryote-derived ones). Yet, there is one possible exception: the recently described stramenopile lineage *Chrysoparadoxa* which only retains two plastid membranes and appears to have lost one of the cyanobacterial-like ones (Wetherbee *et al.* 2018). On the other hand, plastids of some dinoflagellates underwent even more complex evolutionary history and can have up to five membranes (Dorrell & Howe 2015; Matsuo & Inagaki 2018). Thylakoids of secondary plastids do not usually form high-stacked grana but rather elongated and thinner lamellae, in some cases arranged as a girdle lamella (characteristic for some diatoms). Pyrenoids are conserved in all the major secondary plastid lineages in some form. The structurally simplified non-photosynthetic plastids, such as the apicoplast of Apicomplexa that do not harvest light and assimilate carbon, represent the obvious exception from the abovementioned and contain neither thylakoids nor pyrenoid.

1.2.2.3 Euglenid plastids

Euglenids are one of the groups with only three plastid envelope membranes (Gibbs 1978). It is presumed that the innermost two membranes of cyanobacterial origin are conserved while the third one is either symbiont plasma membrane- or host phagosome-derived, but it is ultimately unclear which membrane was lost and no solid evidence on this matter was brought to this day. The outermost plastid membrane is not continuous with the endomembrane system but substantial vesicle traffic was observed between plastid and Golgi, which led some

researchers to hypothesize that the outermost plastid membrane is connected to ER and Golgi functionally albeit not physically and could be a *de facto* part of the secretory pathway (Sulli & Schwartzbach 1996; Enomoto *et al.* 1997). Plastids are often seen in close proximity of the mitochondrion suggesting there might be a metabolite or vesicle exchange or even instances of physical connection between the two organelles. Thylakoids of euglenid plastids are stacked in long lamellae, usually in threes. The pyrenoids can be naked, fitted with one or two paramylon caps or cluster of paramylon grains in the center, or absent. The latter is the case of the family Phacaceae and the genus *Euglenaformis* and a result of secondary loss (Karnkowska *et al.* 2015). The plastids of photosynthetic euglenids are also very variable in shape, size, and amount per cell, some of these characteristics can be used as taxonomical markers for identifying particular genera or species (Leedale 1967; Karnkowska *et al.* 2015). Neither the storage paramylon grains nor the eyespot are part of the plastid in euglenids, unlike primary algae.

1.2.3 Plastid genomes and genetic housekeeping

Genomes of semiautonomous organelles, mitochondria and plastids, are clearly traceable to their respective prokaryotic ancestors, although they underwent a massive reduction in size and code mere units of percent of their original gene content. They usually assemble as single circular molecules reminiscent of bacterial nucleoids, although there are exceptions. These are more common in mitochondrial genomes, with some being very particular and well known – for example the generally fragmented and in some cases linear mtDNA of Euglenozoa with the most extreme deviation represented by the enormous and heavily encrypted kinetoplast of Kinetoplastida (Flegontov *et al.* 2011). However, it was proposed that the circular nature of plastidial genomes is not so universal and stringent either, and that the actual physical arrangement of cpDNA may be more complex and fluid (Bendich 2004). Organellar genomes have low CG content, they are compact, contain group I and II introns (many of which are self-splicing), and have identifiable replication origin (although their replication can be also initiated by recombination). The size of mitochondrial genomes is much more varied than the size of plastidial ones (Barbrook *et al.* 2010).

1.2.3.1 Primary plastids

Genomes of primary plastids of green algae and plants are usually between 100 and 200 kbp in size and code between 90 and 130 genes, while their counterparts in rhodophytes

and glaucophytes are generally larger (136 and 150-200 kbp, respectively) and code more, often over 200 genes (Stirewalt *et al.* 1995; Sato *et al.* 1999; Maul *et al.* 2002; Hagopian *et al.* 2004; Green 2011). The smallest genome of photosynthetic primary plastid of 72 kbp and 86 genes belongs to the picoeukaryote *Ostreococcus tauri* (Robbens *et al.* 2007), only plastids of some secondarily non-photosynthetic green algae and plants were able to shrink their genome even further (Knauf & Hachtel 2002; de Koning & Keeling 2006; Naumann *et al.* 2013; Schelkunov *et al.* 2015) or even completely lose it (Molina *et al.* 2014; Smith & Lee 2014). The almost universal feature of plastid genomes is the inverted repeat region that separates long and short single-copy regions and typically contains genes for three rRNAs and two tRNAs. Plastids code their own RNA polymerase (PEP) which cooperates with sigma factors in transcription regulation. In land plants, plastidial transcription is divided between PEP and its nuclear-encoded viral-like counterpart (NEP) which also transcribes PEP itself (Hajdukiewicz *et al.* 1997), while in algae only PEP is present and the nucleus influences plastidial expression through other genes, such as nuclear-encoded sigma factors. Some plastidial transcripts require splicing, polyadenylation, or other modifications mediated by nuclear-encoded enzymes for correcttranslation (Herrin & Nickelsen 2004; Barbrook *et al.* 2010). Both transcriptional and post-transcriptional regulation responds mainly to light and redox conditions.

Genome of the chromatophore of *Paulinella* is much larger than those of classical primary plastids, 1.02 Mbp in size and coding 867 genes. This, however, represents a mere quarter of the genome of its cyanobacterial predecessor with most of transcription regulation and DNA repair machinery missing which clearly demonstrates its dependence on host nucleus (Nowack *et al.* 2008).

1.2.3.2 Complex plastids

Genomes of complex plastids are generally structured in the same way that those of their respective primary algal ancestors while being smaller and coding less genes, but their reduction is usually not excessive. In red plastids of photosynthetic cryptophytes and stramenopiles, genome size ranges between 115 and 135 kbp while coding for 150-180 genes (Kowallik *et al.* 1995; Douglas & Penny 1999; Khan *et al.* 2007; Oudot-Le Secq *et al.* 2007; Zhang *et al.* 2015). In the haptophyte *Emiliania huxleyi* it is slightly smaller with 105 kbp and 143 genes (Puerta *et al.* 2005) and in the chromerids

Chromera velia and *Vitrella brassicaformis* it is 121 and 84 kpb, respectively. Interestingly, the cpDNA of *Chromera* is linear (Oborník & Lukeš 2015). The genomes of secondarily non-photosynthetic apicoplasts of the apicomplexan parasites are the obvious outliers in both size (up to 35 kbp) and gene content (30-45) (Wilson *et al.* 1996; Cai *et al.* 2003; Imura *et al.* 2014), while plastids of dinoflagellates represent their own chapter. Some dinoflagellate plastids are the result of recent high-tier serial endosymbioses and their genomes have relatively ordinary "chromist" characteristics (Gabrielsen *et al.* 2011) but genomes of the ancestral plastids conserved in peridinin-containing dinoflagellates are very unique: highly reduced (coding less than 20 genes) and structured as multiple minicircles coding from one to few genes each (Zhang *et al.* 1999; Howe *et al.* 2008). Genomes of the green secondary plastids of chlorarachniophytes are only around 70 kbp in size and code around 95 genes (Rogers *et al.* 2007; Tanifuji *et al.* 2014; Suzuki *et al.* 2016). However, their smaller size in comparison to genomes of most red secondary plastids is congruent with the aforementioned general difference in genome sizes and gene content between their primary ancestors.

Most secondary plastid genomes do not contain any introns. The cpDNA of *Rhodomonas salina* (and possibly some other cryptophytes) contain two introns (Khan *et al.* 2007), while a small number of both group I and II introns is generally conserved in cpDNAs of chlorarachniophytes (Suzuki *et al.* 2016). Euglenophyte cpDNAs (discussed below) represent a remarkable exception in this regard.

Editing of plastidial transcripts occurs in dinoflagellates. This process is controlled by nuclear-encoded proteins, it is of host origin and functions even in relatively recently established plastids gained from other secondary algal groups which do not perform any plastidial transcript editing (Zauner *et al.* 2004; Gabrielsen *et al.* 2011).

1.2.3.3 Euglenid plastids

To this day, 17 euglenid plastid genomes have been sequenced. Most of these assembled as circular molecules, with a few exceptions which are most likely caused by incompleteness of the data in the repetitive regions. The gene content is quite conserved in these (around 90, with the exception of non-photosynthetic *E. longa* with 57), but their sizes vary to a significant degree: while the plastid genomes of early-branching Eutreptiales are around 66 kbp, those of some of the "crown" lineages including *E. gracilis* are more than twice

as large, over 140 kbp, and have very low CG content (Hallick *et al.* 1993; Gockel & Hachtel 2000; Hrdá *et al.* 2012; Wiegert *et al.* 2012). This is the result of a massive proliferation of introns: while the plastid genome of *Pyramimonas parkeae* only contains one intron (Turmel *et al.* 2009), two members of Eutreptiales, *Eutreptiella* and *Eutreptia*, have 8 and 27, respectively, and the number steadily grows in other lineages with the maximum of 145 in *E. gracilis*. The introns are also very diverse, including not only the classical organellar types I and II, but also euglenid-specific type III and numerous instances of twintrons – introns nested inside other introns which are spliced subsequently (Copertino & Hallick 1993; Doetsch 2000; Sheveleva & Hallick 2004).

The inverted repeat regions present in the plastid genome of *Pyramimonas parkeae* and typical for primary plastids are absent from Euglenaceae but conserved in *Eutreptiella* and some Phacaceae (Hrdá *et al.* 2012; Karnkowska *et al.* 2018). Euglenophyte plastid genomes contain a VNTR (variable number of tandem repeats) region and 15 gene clusters which are themselves conserved but occur in different order and orientation throughout euglenophytes (Dabbagh & Preisfeld 2016).

While the trans-splicing of nuclear transcripts is widespread in euglenids, no such phenomenon was reported in plastids.

1.2.4 Plastid biogenesis

To control its semi-autonomous organelle, the host organism has to be able to supply it with proteins coded in its nucleus, being it the ones laterally transferred from the endosymbiont or host-originated and newly recruited for organellar function. This requires a) a targeting signal on the nascent protein and b) molecular machinery that can recognize said signal and transfer the protein to its destination. In the case of mitochondria and primary plastids which are direct descendants of prokaryotic cells, the host organism took advantage of the existing bacterial secretion system which was substantially modified by new eukaryote-specific structural, functional, and regulatory subunits. In the case of secondary and higher plastids, the protein importing machinery had to undergo some changes with each transfer of the organelle to the new host and each additional membrane it gained in the process.

1.2.4.1 Primary plastids

Targeting signals

The signal for import to plastid is an N-terminal extension on a newly synthesized protein termed transit peptide (TP) which is eventually removed by processing peptidase during the protein maturation in plastid stroma. The average length of TP is approximately 50 amino acids but it can be up to three times as long or, on the other hand, as short as 13 amino acids. It is not conserved at sequence level, except for the cleavage site, but in turn exhibits certain overall biochemical properties stemming from its amino acid composition: enrichment in hydroxylated residues (mainly serine and threonine) and alanine and very low content of acidic/negatively changed residues (Gavel & von Heijne 1990; Bruce 2000; Li & Teng 2013). It contains semi-conserved physicochemical motifs termed FGLK which presumably mediate its interactions with different parts of the importing machinery (Holbrook *et al.* 2016). However, some plastidial proteins possess neither N- nor C-terminal TP and are transported to their destination based on an alternative, possibly internal signal (Miras *et al.* 2002).

TOC and TIC complexes

Most plastidial proteins are imported by translocase complexes TOC and TIC (translocon of outer and inner chloroplast membrane, respectively). These complexes are partially derived from components of the bacterial secretion system. This is particularly conspicuous in the case of its subunits Toc75 and OEP80 belonging to highly conserved family of bacterial outer membrane β-barrel proteins (Omp85/BamA) which also includes mitochondrial Tob55/SAM50 (Gentle *et al.* 2005; Sommer *et al.* 2011). Other subunits, on the other hand, represent eukaryotic innovations and have no homologs in bacteria and mitochondria (Shi & Theg 2013; Day & Theg 2018). A minor portion of plastidial proteins reach their destination by alternative mechanisms: tail- or signal-anchored membrane proteins are inserted to the outermost membrane, either directly from cytosol or through ER vesicles (Lee *et al.* 2013), other proteins do not possess TP and pass the membranes by yet unknown route (Miras *et al.* 2002; Nada & Soll 2004). Mistargeted proteins that remain in cytosol are recognized by Hsp75-4 based on a motif in their TP and flagged for proteasomal degradation by E3 ubiquitin ligase (Lee *et al.* 2009).

Plastidial proteins are imported post-translationally and in unfolded state. Cytosolic chaperones Hsp90 and Hsp70 (in cooperation with 14-3-3 protein of unknown identity) mediate their transfer to the TOC complex and prevent them from misfolding. Toc64, a subunit with TPR domain, and possibly also OEP61 are believed to interact with these chaperones and pass the transported preproteins further onto the TOC complex. Alternatively, some preproteins pass directly from the cytosol without the help of chaperones. The core of the TOC complex comprises of two subunits mediating TP recognition and preprotein binding, GTPases Toc34 and Toc159 (both of eukaryotic origin) and the channel subunit Toc75 (belonging to the aforementioned Omp85 family) which also contains an intermembrane space-facing triple POTRA domain with TP-binding interface (May & Soll 2000; Shi & Theg 2013; Paila *et al.* 2015; Schwenkert *et al.* 2018). A recently identified kinase KOC1 associated with Toc159 appears to be an essential part of the machinery as well (Zufferey *et al.* 2017). Additionally, isoforms of the main subunits (e.g. Toc132, Toc120, or Toc90) are differentially expressed in some systems and serve a regulatory purpose (Demarsy *et al.* 2014). Toc75 forms a channel with 14-26Å in diameter and can safely accommodate an unfolded protein chain (Hinnah *et al.* 2002). However, certain proteins can pass through the channel in a folded state as well which could be related to a structurally flexible subdomain of Toc75 and/or it forming a homo-oligomer in order to transport larger substrates (Clark & Theg 1997; Paila *et al.* 2015).

The organization and functionality of the TIC complex is more complex and somewhat controversial to this day. Tic20, Tic40, and Tic110 were long considered the core of the complex, with Tic20 (containing four transmembrane helices) forming a channel, Tic40 mediating interaction with stromal chaperones, and Tic110 (containing two transmembrane helices and stroma-facing extension) performing and connecting both of these functions and representing the main subunit (Kouranov *et al.* 1998; Heins *et al.* 2002). However, the isolation of 1 MDa TIC subcomplex showed that the situation is not as straightforward (Kikuchi *et al.* 2009; Kikuchi *et al.* 2013). The 1 MDa complex comprised of Tic20 plus a completely different set of interacting subunits: Tic56, Tic100, and Tic214 (relatively recently discovered protein product of a long-enigmatic plastidial ORF Ycf1; de Vries *et al.* 2015), and also Tic21 which is structurally and functionally similar to Tic20 but appeared to be only partially involved in the complex (Kikuchi *et al.* 2009). However, other co-purification experiments showed that Tic56 is also only loosely associated with the 1 MDa complex and non-essential (Köhler *et al.* 2015), and the significance

of Tic214 was challenged as well, arguing that it is not by any means ubiquitous in plants and therefore cannot carry out a central function (Bölter & Soll 2017). On the other hand, some studies argue that Tic110, despite being able to form cation-selective channel in heterologous experimental system (Heins *et al.* 2002), cannot serve as such in native conditions because of its stromal-facing extension, that there is no solid evidence of it interacting with Tic20, and that its main and possibly only function is indeed to serve as a scaffold for the assembly of stromal chaperones complex (Inaba *et al.* 2003; Tsai *et al.* 2013; Nakai 2018), together with Tic40. Meanwhile, Tic22 interacts with chaperones binding to Toc64 in the intermembrane space and indirectly connects TIC to TOC. The remaining subunits Tic32, Tic55, and Tic62 are all redox-sensing enzymes that are not directly implicated in the protein import but rather regulate the function and assembly of other subunits of the translocon complex.

Stromal chaperones include Hsp93, ClpC, and plastidial homologs of Hsp90 and Hsp70 docking on the stromal extension of TIC. They fold and subsequently pull in the preprotein chain, representing the main ATP-dependent motor of the import process which is relatively costly, consuming 650 ATP molecules on average per protein imported (Shi & Theg 2013; Paila *et al.* 2015). Concurrently, TP is cleaved by stromal processing peptidase (SPP), this step is likely essential (Trösch & Jarvis 2011).

Intraplastial protein sorting

Proteins destined to thylakoid lumen or membrane are further sorted based on additional signals. Proteins with thylakoid lumen-targeting domains (LTD) that are N-terminal (following immediately after the TP) are translocated by either Sec or TAT (twin-arginine translocase) systems which are analogous to their prokaryotic counterparts. Their LTDs are then cleaved by respective peptidases in the lumen. Proteins that are to be inserted into thylakoid membrane, mostly parts of the photosynthetic electron transport chain and light-harvesting complexes, have internal signals, SP-like region, or distinctive overall properties that ensures their insertion which takes part either “spontaneously”, in an energetically neutral way taking advantage of pH gradient and without the aid of any known transporters, or using plastidial signal recognition particle (cpSRP), its interaction partner FtsY, and a thylakoid membrane insertase Alb3 (Jarvis & Robinson 2004).

1.2.4.2 Complex plastids

In complex plastids, a need for significant innovation in protein import is implicit in the presence of one or more additional envelope membranes. These membranes are of eukaryotic origin, derived from cytoplasmatic membranes or endomembrane system to which they are physically connected in some cases (i.e. cryptophytes, haptophytes, and stramenopiles), forming a chloroplast endoplasmic reticulum (CER). In these cases, the plastid lies within the ER lumen and represents its *de facto* subcompartment, so it is not at all surprising that the protein import across the outermost membrane uses the same mechanism as transport of secreted proteins to ER lumen. The protein import starts as a co-translational process employing an N-terminal signal peptide (SP) recognized by cytosolic signal-recognition particle (SRP) as soon as it protrudes from a ribosome on which it is synthesized. The whole ribosome is then docked on ER by an interaction of SRP with its receptor and the nascent protein chain is imported through a coupled protein channel (Keenan *et al.* 2001). Signal peptidase localized in ER lumen then removes the SP at its cleavage site, revealing next part of the N-terminal domain which destines the preprotein to plastid and distinguishes it from regular SP-bearing proteins designated for secretion. The N-terminal targeting signal of complex plastids is therefore at least bipartite, consisting of SP followed by plastidial TP. This general mechanism and the involved signals and translocases are the same in other complex plastids that are not directly connected to the host ER as their outermost membrane is still evolutionarily and functionally continuous with it. It functions as a part of the secretory system and the plastidial proteins imported to the ER lumen reach it via vesicles, either directly from ER (in apicomplexans and chlorarachniophytes; Tonkin *et al.* 2006; Bolte *et al.* 2009; Hirakawa *et al.* 2012) or through Golgi cisterns (in peridinin dinoflagellates and euglenophytes; Inagaki *et al.* 2000; Nassoury *et al.* 2003; Sláviková *et al.* 2005). The conservation of this mechanisms is not surprising among complex plastids of the red lineage as they are all most likely ultimately traceable to a common ancestor (see chapter 2.2.1.2), however, it is notable that this initial stage of plastid protein import across the outermost membrane is the same in the two groups with independently acquired green complex plastids and therefore represents an evolutionarily convergent result of general constraints in the molecular and cellular tools available to a cell in the process of integrating an algal endosymbiont.

Protein import across the second eukaryote-derived envelope membrane, if present, is also notably conserved in the representatives of red plastid lineage. Cryptophytes, haptophytes, stramenopiles, chromerids, and apicomplexans all recruited a set of components of the ER-associated degradation pathway (ERAD) transporting ubiquitin-tagged proteins for degradation in cytosol (Meusser *et al.* 2005) originated by paralog duplication which presumably took place as deep in the evolution as in their last common ancestor (Sommer *et al.* 2007; Hempel *et al.* 2009; Felsner *et al.* 2011; Gould *et al.* 2015). This plastidial version of the ERAD pathway is termed symbiont-specific ERAD-like machinery (SELMA) and includes ubiquitylation enzymes Hrd1 (ubiquitin-ligase, E3), Uba1 (ubiquitin-activating protein, E1) and Ubc4 (ubiquitin-conjugating protein, E2), AAA-ATPase Cdc48 and its cofactor Ufd1 which bind the tagged protein and mediate its expulsion through a channel formed by Der1-1 and Der1-2 (Hempel *et al.* 2010; Felsner *et al.* 2011; Stork *et al.* 2012). Additional subunits, so far identified in diatoms, include a second copy of Cdc48, its interaction partner PUB, and a rhomboid protease Rhom3 (Lau *et al.* 2016). Translocation by SELMA is dependent on the recognition of a sequence motif at the boundary between SP and TP consisting of a phenylalanine residue surrounded by several hydrophobic ones (most common consensus is ASAFAP) which is more or less conserved throughout algae with red complex plastids (Kilian & Kroth 2005; Gruber *et al.* 2007; Hempel *et al.* 2010), and N-terminal lysine residues, probably as sites for ubiquitylation (Lau *et al.* 2015).

The inner two membranes of complex plastids are presumably homologous to the ones of primary plastids. As a result, systems derived from TOC and TIC complexes are expected to function on these membranes, and the available genomic and transcriptomic data partially confirm this. At least the central TIC subunits, Tic110, Tic20 and/or Tic22 were identified in a representative of almost every group with secondary plastids with almost full set in cryptophytes and chlorarachniophytes, coded partially in nuclear and partially in nucleomorph genomes (Douglas *et al.* 2001; McFadden & van Dooren 2004; Gilson *et al.* 2006; Hirakawa *et al.* 2012; Hehenberger *et al.* 2014). On the other hand, nucleomorph-encoded Toc75 of chlorarachniophytes and nucleus-encoded Toc34 of apicomplexans were the only known components of TOC of complex plastids for some time before the discovery of plastidial Omp85-like protein which is widespread in complex algae and, as a member of the same highly conserved protein family, might represent a homolog and/or functional analog of Toc75 or SAM50 (Bullmann *et al.* 2010).

In summary, the core channel-forming subunits of TOC and TIC are generally present in complex plastids, but other components of these translocase complexes are often absent and/or divergent and not readily identified by standard homology detection in transcriptomic or genomic studies. It is also worth notice that the most complete sets of TIC subunits are present in representatives of the two groups of complex algae with nucleomorphs, suggesting that the reduction and/or modification of the translocase complex might be greater the further down the series of endosymbioses it travelled. At the same time, a TP exhibiting only moderate deviations from the canonical plant-like properties formula is retained as a part of a multipartite signal domain of complex plastid proteins, suggesting an analogous TP-recognizing mechanism which, however, might employ alternative receptor as the canonical Toc34 was detected in apicomplexans only (Waller & McFadden 2005) and Toc159 was not identified in any complex plastid so far.

Intraplastidial sorting is generally conserved in complex plastids, with the bacterial-like thylakoid transporters detectable in the genomic and transcriptomic data, and the spontaneous insertion pathway working canonically and even heterologously across different groups of algae (Broughton *et al.* 2006; Gould *et al.* 2007).

1.2.4.3 Euglenid plastids

Protein import mechanism of euglenophyte plastids is one of the less investigated ones but it generally follows the formula common to other complex plastids. The third, outermost plastid membrane is not directly connected to ER or Golgi but frequent vesicular trafficking between these compartments can be observed under electron microscopy (Sulli *et al.* 1999; Inagaki *et al.* 2000) and was demonstrated by *in vitro* Golgi-to-plastid import experiments (Sláviková *et al.* 2005). The mechanism of vesicle recognition and docking is unknown, however, the experimental *in vitro* import from Golgi to isolated plastids was unaffected by N-ethylmaleimid treatment, suggesting N-ethylmaleimid-sensitive fusion proteins (NSF) and, by extension, SNARE proteins are not involved (Sláviková *et al.* 2005).

Nuclear-coded plastid proteins have bipartite or tripartite signal sequence comprising of canonical SP, plant-like TP, and in roughly half of the preproteins described as “class I”, a second hydrophobic domain termed stop-transfer signal (STS) (Durnford & Gray 2006). The STS is believed to form a transmembrane domain which docks the preprotein in transport vesicle membrane while preproteins without STS (“class II”) are transported in soluble form

in vesicle lumen (Sulli *et al.* 1999; Durnford and Gray 2006). Interestingly, analogous plastid preprotein classes, one with additional hydrophobic domain, are observed in peridinin dinoflagellates (Patron *et al.* 2005). Considering these two groups of algae are the only ones sharing the plastid layout with just three, not four (or more) envelope membranes, this might point at some as of yet unknown underlying mechanism resulting in such evolutionary convergence.

Translocation across the two innermost membranes which are presumably homologous to the cyanobacterial-like ones of primary plastids is believed to be carried out by a system homologous to TOC and TIC, possibly reduced or modified to some degree as observed in other complex plastids. This is also suggested by the relatively canonical plant-like properties of TP of euglenophyte plastid preproteins and its ability to mediate a delivery of a *Euglena gracilis* protein into isolated pea plastids (Sláviková *et al.* 2005). However, no molecular or sequence evidence for any Toc, Tic, or other canonical plastid envelope or thylakoid transporter was brought so far.

Figure 2 represents a graphical comparison of protein import machineries and summary of the previous chapters.

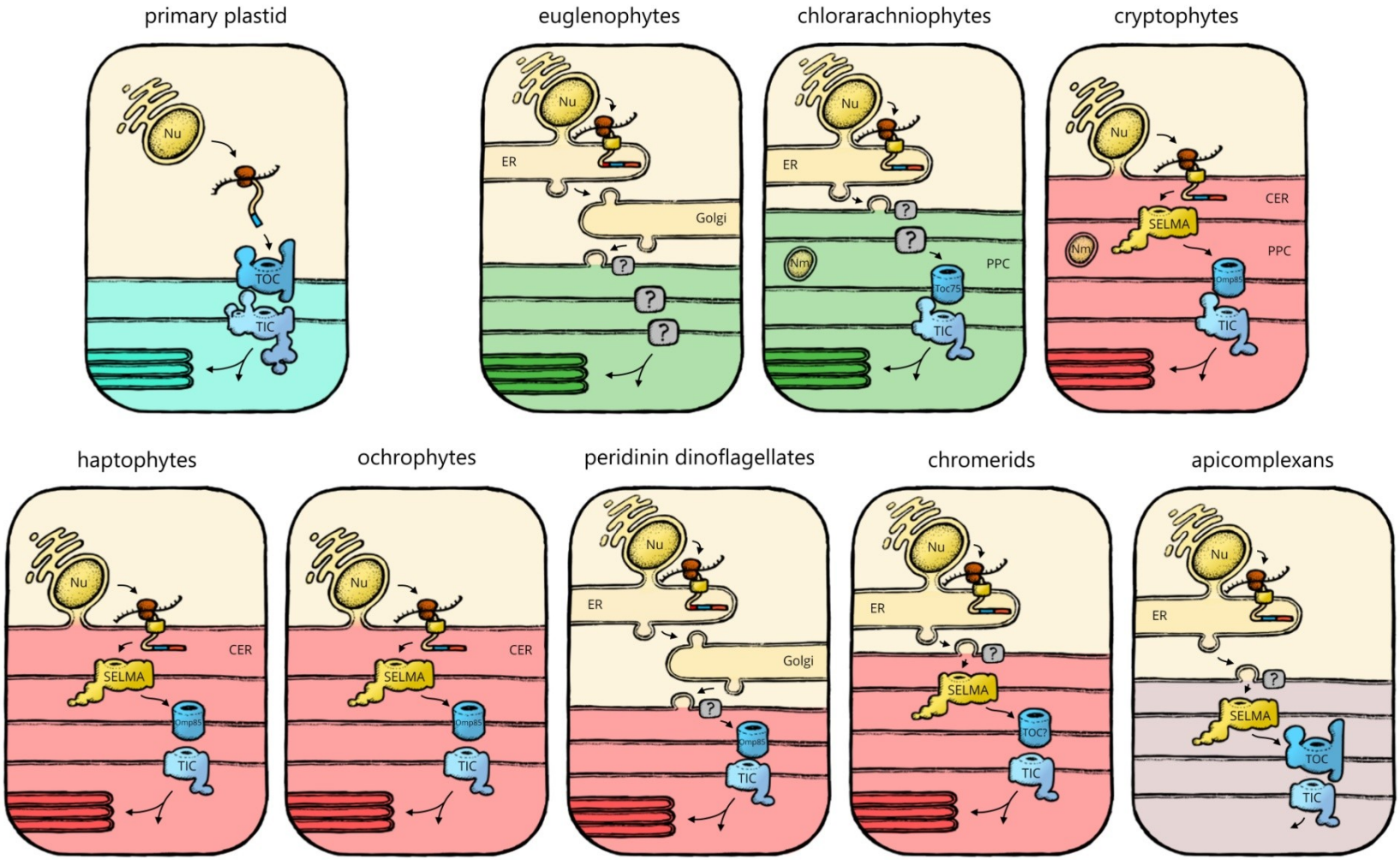

Figure 2: Comparison of protein import machineries of primary and complex plastids. In plants and primary algae, plastidial proteins are synthesized on free ribosomes and imported across two membranes via TOC and TIC complexes based on transit peptide (TP, blue). In complex plastids, additional membranes and signals are involved. Plastidial proteins possess signal peptide (SP, red) and are imported co-translationally into ER lumen. In groups with chloroplast ER (CER) this is equivalent to transport across the outermost plastid membrane, while in other groups, proteins must travel in vesicles (either directly or through Golgi) that fuse with the outermost membrane. In most "chromists", SELMA (translocase complex derived from ERAD pathway) mediates protein import across the second outermost membrane, while TOC and TIC are generally conserved in some, usually reduced, form on the innermost two membranes. However, the nature of vesicle-docking and translocases of some of the membranes in some groups is unknown (grey boxes with "?"); (design of the illustration based on Bolte *et al.* 2009).

1.2.5 Metabolic functions of plastids

Plastid represents a significant innovation in metabolic capacities of a host organism. Photosynthesis and carbon fixation bring an obvious switch to autotrophic or facultatively mixotrophic mode of energy acquisition while other metabolic pathways embedded in the organelle allow synthesis of novel compounds, innovative gearing of existing processes to photosynthetic reactions and/or their advantageous compartmentalization. In this chapter, canonical plastidial metabolism which was extensively studied in primary plastids of plants will be briefly summarized and compared to the current findings regarding the derived and younger complex plastids.

1.2.5.1 Primary plastids

Photosynthesis and carbon fixation

The defining feature of energy metabolism of plastids is their ability to utilize light as a source of energy and reducing power. Light-harvesting complexes I and II (LHCI and LHCII) are large apparatuses comprising of multiple protein subunits and photosynthetic pigments embedded in the thylakoid membrane that are able to transfer electrons resulting from light-induced charge separation on chlorophyll to carrier molecules, ferredoxin and plastoquinone, respectively. The former passes the reducing power onto NADPH which serves as a reduction equivalent in various reactions. The latter represents a first step in photosynthetic electron transport chain employing cytochrome b6f complex and plastocyanin and resulting in a formation of proton gradient between thylakoid lumen and plastid stroma which is used to fuel ATP production by F-type ATP synthase complex (Nevo *et al.* 2012).

Carbon for a synthesis of organic molecules is obtained by reductive pentose-phosphate pathway (RPPP), also termed Calvin cycle. The central enzyme of this metabolic pathway is RuBisCO, ribulose-1,5-bisphosphate carboxylase/oxygenase comprising of large (RbcL, plastid-coded) and small (RbcS, nuclear-coded) subunits, eight molecules each. This bifunctional enzyme serves as a carboxylase in Calvin cycle, adding one CO_2 molecule to a ribulose-1,5-bisphosphate (RuBP). This produces two molecules of 3-phosphoglycerate which are then reduced to triose-phosphates glyceraldehyde-3-phosphate (G3P) and dihydroxyacetone phosphate (DHAP), consuming NADPH and ATP. Five out of six molecules of triose-phosphates are used for RuBP regeneration and remain in the cycle while

the sixth molecule represents the net yield usable in subsequent reactions, including synthesis of sugars and storage polysaccharides and isoprenoids. Calvin cycle is activated by carbamoylation of RuBisCO under light conditions that affect the availability of reduced ferredoxin and thioredoxin, Mg^{2+} ions, and pH (Raines 2003; Zakhartsev *et al.* 2016; Tetlow *et al.* 2018).

Under certain conditions, RuBisCO carries out its alternative oxygenase function in a process termed photorespiration, reducing the effectivity of photosynthesis as the 2-phosphoglycolate resulting from RuBP oxygenation cannot continue in the cycle and is instead metabolized by mitochondrial and peroxisomal metabolism evolving CO_2 and ammonia by decarboxylation of glycine, and expending ATP and NADPH (Ogren 1984; Zakhartsev *et al.* 2016; Tetlow *et al.* 2018).

Oxidative pentose phosphate pathway (OPPP) runs in the opposite direction than Calvin cycle, intersects with several reactions of plastidial glycolysis and generates NADPH which is especially important in non-photosynthetic plastids where this is the only source of the reduced form of this cofactor. It also provides ribose for the nucleotide metabolism and erythrose-4-phosphate which enters shikimate pathway (Kruger & von Schaewen 2003; Tetlow *et al.* 2018).

Nitrogen and amino acid and metabolism

Nitrogen in the form of NO_2^- is transported to plastids and converted to ammonia which (as well as ammonia produced by photorespiration) can be assimilated into glutamine and glutamate in glutamine synthase/glutamine oxoglutarate aminotransferase (GS/GOGAT) cycle which represents an important metabolic regulation point and crossroads connecting carbon and nitrogen metabolism. The transfer of an amido group from glutamine to 2-oxoglutarate by GOGAT to form glutamate requires reduction equivalent in the form of ferredoxin or alternatively, in the case of minor and possibly photosynthesis-independent isoform of the enzyme, NADPH. The amido group from glutamate can be then transferred to alpha-ketoacids by transaminase to form aliphatic amino acids (Ohyama & Kumazawa 1980; Tetlow *et al.* 2018).

The biosynthesis of aromatic amino acids requires a specific precursor, chorismate produced by shikimate pathway. The input molecules of shikimate pathway include erythrose-4-phosphate (produced in OPPP) and phosphoenolpyruvate (PEP) that form 3-deoxy-D-arabino heptulosonate 7-phosphate (DAHP) by DAHP synthase in the first

reaction of the pathway. This step, however, must be subjected to regulation by negative feedback by downstream products of the same pathway and thioredoxin, as an uncontrolled entry of erythrose-4-phosphate compromises plastidial pentose phosphate pathway as a whole. Chorismate synthesized by shikimate pathway is then used for a synthesis of phenylalanine and tyrosine (through arogenate) and tryptophan (through anthranilate). In addition to this, it serves as a precursor of folate and vitamin K, indole, and some phytohormones (such as auxin and salicylic acid) and secondary metabolites (such as alkaloids) (Kanehisa & Goto 2000; Rippert *et al.* 2009; Maeda & Dudareva 2012; Tetlow *et al.* 2018).

Biosynthesis of isoprenoids and related compounds

The precursor of isoprenoid compounds, isopentenyl pyrophosphate (IPP) can be synthesized by two alternative pathways. Mevalonate pathway operates in cytosol and uses acetyl-CoA as input molecule, while methylerythritol phosphate/deoxyxylulose phosphate (MEP/DOXP) pathway is present in plastids and uses G3P (produced in Calvin cycle) and PEP to form DOXP in the first step catalyzed by DOXP synthase. Most photosynthetic organisms have lost the cytosolic pathway and rely solely on plastidial MEP/DOXP pathway; this is one of the reasons for plastid retainment and essentiality in secondarily non-photosynthetic algae and plants (Kanehisa & Goto 2000; Janouškovec *et al.* 2015; Tetlow *et al.* 2018).

IPP is a precursor of isoprene and can be converted to geranyl pyrophosphate (GPP) and geranylgeranyl pyrophosphate (GGPP) from which various steroids and mono- and diterpenes, such as the phytohormone gibberellin, are synthesized. GGPP is also a precursor of carotenoids. Phytoene synthesized from GGPP is converted to all-trans-lycopene by desaturation carried out by a single enzyme in most prokaryotes but four enzymes in cyanobacteria and plastids. The resulting isomer of lycopene serves as a central molecule from which other types of carotenoids derive through isomerization and cyclization. Phytohormones abscisate and strigolacton are also derived from carotenoids (Armstrong & Hearst 1996; Kanehisa & Goto 2000). Carotenoids are important as anti-oxidative and photo-protective agents and play a central role in xantophyll cycle which de-epoxydises violaxanthin to zeaxanthin through the intermediate of antheraxanthin which plays major role in non-photochemical quenching (NPQ) to partially dissipate the light energy and reduce damage by ROS and inhibition of photosynthesis by excess light (Latowski *et al.* 2011).

Phytyl pyrophosphate synthesized from GGPP is an important molecule in synthesis of phytol compounds such as tocopherols and tocotrienols (vitamin E) and the photosynthetic cofactor phylloquinone (vitamin K_1). It also constitutes side chains of chlorophyll.

Lipid metabolism

Plastids are the main site of fatty acid synthesis. This pathway uses acetyl-CoA (synthesized *in situ* from glucose-6-phosphate since it cannot be transported across membrane) which is then carboxylated to malonyl-CoA by acetyl-CoA carboxylase (ACC) and passed onto type II fatty acid synthase (FAS) complex which comprises of multiple subunits (as opposed to a single protein of type I FAS) that adds two carbon acyl molecule to the existing chain in every cycle at the expense of two reducing equivalents (NADPH produced by photosynthesis or OPPP). The resulting saturated fatty acids, primarily in the form of oleoyl and palmitoyl residues bound to acyl-carrier protein (ACP) can be exported and modified in ER or stay in plastid and enter prokaryotic-type metabolic machinery to produce plastid-specific lipids crucial for the function of photosynthetic membranes of thylakoids. These are galactolipids monogalactosyl diacylglycerol (MGDG) and digalactosyl diacylglycerol (DGDG), sulfolipid sulfoquinovosyl diacylglycerols (SQDG), and phospholipid phosphatidylglycerol (PG) (Kanehisa & Goto 2000; Wang & Benning 2012; Boudière *et al.* 2014; Tetlow *et al.* 2018).

Tetrapyrrole synthesis

Plastids synthesize chlorophylls, sirohaem, and phytochromobilin, each through a distinct pathway branching from a common backbone of tetrapyrrole synthesis which starts with 5-aminolevulinate. This compound can be synthesized either via Shemin pathway from succinyl-CoA and glycine or via three-step C5 pathway from glutamate, the latter is characteristic for cyanobacteria and plastid-bearing organisms. Chlorophyll biosynthetic branch starts with the chelation of protoporphyrin IX by Mg^{2+} ion by multiunit enzyme magnesium chelatase. The resulting molecule is then methylated and oxidated to protochlorophyllide, which is then converted to chlorophyllide by protochlorophyllide oxidoreductase (POR) which is strictly light-dependent in plants while most algae have a second isoform through which they are able to carry out this reaction in dark. Chlorophyll synthases then conclude the process by addition of a phytol side chain. Sirohaem biosynthesis branches at earlier intermediate, uroporphyrinogen III which is methylated, oxidized to sirohydrochlorin and then chelated by Fe^{2+} ion. Phytochromobilin is synthesized in some

plastids in the haem branch of the tetrapyrrole pathway from biliverdin. The subcellular localization of the synthesis of haem (and subsequently cytochromes) is unclear in plants and algae. It might be divided between plastid and mitochondrion or dual-localized (Von Wettstein *et al.* 1995; Kanehisa & Goto 2000; Tanaka & Tanaka 2007; Heyes & Neil Hunter 2009).

Iron-sulfur cluster assembly

Plastids harbour their own machinery for the assembly of iron-sulfur (FeS) clusters, pyrite-like cofactors essential for the function of redox enzymes, SUF pathway inherited from cyanobacteria. In essence, sulfur is released from cysteine by cysteine desulfurase (SufS) and sulfur transferase (SufE) and transferred onto a protein scaffold (SufBCD) to form a disulfidic bond with an iron molecule (the source of this iron is not clear). Eventually, one of the FeS cluster types (most commonly 2Fe2S, 3Fe4S, or 4Fe4S) is formed, depending on the amino acyl residues involved, and transferred onto a client protein by a specific carrier protein (Pilon *et al.* 2006; Lu 2018).

Starch metabolism

Most plants and algae use starch as a long-term energy reserve. Starch is synthesized in plastids and comprises of two types of polymers of glucose residues, the unbranched amylose connected by alpha 1,4 glycosidic bonds, and amylopectin branched through alpha 1,6 bonds. The glucans are insoluble, form a complex intertwined structure, and can be stored in various cell compartments. ADP-glucose, produced by ADP-glucose pyrophosphorylase from intermediate products of gluconeogenesis, is the precursor of starch synthesis. Multiple starch synthases connect the ADP-glucose units by 1,4 linkages, with different isoforms specializing on different chain lengths, while starch-branching enzymes create the 1,6 branching points in amylopectin chains. The breakdown of starch involves debranching enzymes and endoamylases which break the glycosidic bonds via hydrolysis yielding oligo-, di- and monosaccharides, or starch phosphorylases cleaving off one glucose-1-phosphate at a time via phosphorolysis (Zeeman *et al.* 2010; Tetlow *et al.* 2018).

1.2.5.2 Complex plastids

Notable differences in the metabolic pathways of complex plastids in comparison to that of canonical primary plastids will be summarized in the following chapter and visualized in Figure 3.

Photosynthetic apparatus and pigments

Photosynthetic complexes are generally quite conserved throughout complex plastid-bearing algae, differing in the combination of light-harvesting pigments stemming from their respective red or green origin with "chromists" lacking chlorophyll b but possessing chlorophyll c_1 and/or c_2. Remarkably, biliproteins phycocyanin or phycoerythrin containing open-chain tetrapyrolle chromophores and typical for rhodophytes are conserved in photosystems of cryptophytes (Hill & Rowan 1989; Wedemayer *et al.* 1996). As a result, light-harvesting complexes of these algae have a distinct, robust structure. Ochrophytes possess an inventory of unique PSI and PSII subunits which are conserved throughout this group but not shared with any other algae (Dorrell *et al.* 2017). Photosynthetic apparatus of chlorarachniophytes is also unique as PSI is dramatically reduced and lacking typical antennae-forming subunits which might be in turn substituted by PsbY which is surprisingly abundant in plastids (Hopkins *et al.* 2012; A.D. Neilson *et al.* 2017). PSII, on the other hand, remains canonical.

Xantophyll cycle typical for green plastid lineage is present in some red complex plastids, probably representing a lateral acquisition providing advantage in harsher light conditions in comparison to the ancestral rhodophyte-like state. Chrysophytes, xantophytes, and eustigmatophytes possess a canonical pathway cycling between violaxanthin and zeaxanthin, while diatoms (and possibly bolidophytes, dictyochophytes, and pelagophytes) have a distinctive carotenoid diadinoxanthin which is de-epoxidised to diatoxanthin. This system is partially shared with peridinin dinoflagellates which also use diadinoxanthin but convert it to a different de-epoxydised carotenoid termed dinoxanthin (Cao *et al.* 2013; Dorrell & Bowler 2017). Another interesting adaptation of ochrophytes not shared with typical red algae is the presence of plastocyanin, otherwise restricted to green plastid lineage. This is likely a response to iron-poor environment since plastocyanin using Cu^{2+} ion as cofactor can functionally substitute iron-dependent cytochrome c6 (Peers & Price 2006).

Carbon and nitrogen metabolism

There are some unique modifications to carbon metabolism in some complex algae. In dinoflagellates and chromerids, an unusual type II RuBisCO is present. This is a prokaryotic-like enzyme forming homodimers instead of heterodimers like the canonical

type I, and it is likely a result of lateral gene transfer from bacteria (Morse *et al.* 1995; Janouškovec *et al.* 2010).

In ochrophytes, plastidial glycolysis is missing hexokinase, its phosphofruktokinase is not ATP- but pyrophosphate-dependent, and the whole pathway appears to be specifically adapted to function in dark (Kim *et al.* 2016). Carbon fixation probably uses a modified C4 pathway which is wired to some mitochondrial enzymes and subjected to atypical regulation (Kroth *et al.* 2008; Haimovich-Dayan *et al.* 2013; Kustka *et al.* 2014). Additionally, CO_2 needs to be biophysically concentrated as it would otherwise readily diffuse to environment. This might be mediated by a set of carbonic anhydrases (Kroth *et al.* 2008).

In haptophytes, a unique chemical environment is a direct result of the calcification these organisms carry out. Haptophytes can use both CO_2 and HCO_3^- as carbon source, the latter in fact dominates in some representatives (Rost *et al.* 2003). Active carbon concentration is probably mediated by specialized translocases and might be relatively wasteful (Tsuji & Yoshida 2017). It was theorized that CO_2 evolving during calcification may be directly taken up by the carbon fixation pathway but it is not very likely due to the spatial separation of the processes (Bach *et al.* 2013).

In ochrophytes, a remarkably close association and metabolic flux between plastids and mitochondria is present. The mitochondrial urea cycle likely functions as a sink for excess nitrogen produced in plastid photorespiration (Allen *et al.* 2011). Ornithin cycle putatively present in plastid, might represent the second half of this shuttle (Bailleul *et al.* 2015; Levering *et al.* 2016). Mitochondrion also takes up the excess reducing potential in the form of NADPH in exchange for ATP. This unique metabolic wiring greatly improves the efficiency of photosynthesis and reduces photo-damage (Bailleul *et al.* 2015).

Tetrapyrrole pathway

Plastidial tetrapyrrole pathway is canonical in ochrophytes, cryptophytes, and chlorarachniophytes. In the latter case, there are two parallel versions present, likely resulting from the relatively recent plastid acquisition: “heterotrophic” one localized in mitochondria and using 5-aminolevulinate synthesized by Shemin pathway, and a plastidial one using 5-aminolevulinate from glycine (Hopkins *et al.* 2012; Cihlář *et al.* 2016). In cryptophytes, the plastidial pathway is present and the mitochondrial

one is lost except for ferrochelatase, the final enzyme leading to haem. As a result, all tetrapyrrole synthesis starts in plastid and only bifurcates between plastid and mitochondrion before the last reaction (Cihlář *et al.* 2016). On the other hand, in apicomplexans and chromerids, there is a single pathway of chimeric origin and complicated spatial division of the respective steps. It starts in mitochondrion, using 5-aminolevulinate from Shemin pathway, but continues in cytosol and plastid. As a result, chromerids are the only known organisms using Shemin pathway-derived precursor for chlorophyll synthesis. The haem branch of the pathway, on the other hand, terminates back in the mitochondrion (Koreny *et al.* 2011; van Dooren *et al.* 2012).

Isoprenoid synthesis

MEP/DOXP pathway is generally conserved and canonical in complex plastids. It might be the last essential function of non-photosynthetic apicoplast. Since there is no core carbon metabolism in this organelle, the precursors for IPP synthesis must be imported (Guggisberg *et al.* 2014).

Amino acid metabolism

While ochrophyte plastids are capable of synthesizing a wide variety of amino acids including aromatic ones, other complex plastids are more reduced in this regard. A moderate inventory of amino acid biosynthetic enzymes is present in plastids of chlorarachniophytes and dinoflagellates, with shikimate pathway missing, and only a few are retained in cryptophytes, haptophytes, and chromerids, with apicoplast of apicomplexans synthesizing no amino acids whatsoever (Guedes *et al.* 2011; Hopkins *et al.* 2012; Dorrell *et al.* 2017).

Iron-sulfur cluster assembly

SUF pathway is conserved in complex plastids, including the highly reduced non-photosynthetic apicoplast where Suf subunits are even among the very few genes coded in its genome. Interestingly, a second FeS cluster assembly pathway, CIA, canonically localized to cytosol, is present and active in the periplastidial compartment of cryptophytes and possibly also chlorarachniophytes (Hjorth *et al.* 2005; Grosche *et al.* 2018).

Storage molecules

As a storage compound, most complex algae use starch or a very similar glucan. Interestingly, in cryptophytes, it is synthesized from UDP-glucose units instead of ADP-glucose

(Haferkamp *et al.* 2006; Beardall & Raven 2012). Haptophytes and ochrophytes use a beta-glucan chrysolaminarin. In addition to this, ochrophytes also use neutral lipids for energy storage (Beardall & Raven 2012). Haptophytes, on the other hand, rely on triacylglycerol (TAG) and alkenones (Tsuji *et al.* 2009; Tsuji *et al.* 2015). In addition to this, their small molecule sugar content is heavily skewed towards mannitol in comparison to sucrose (Obata *et al.* 2013).

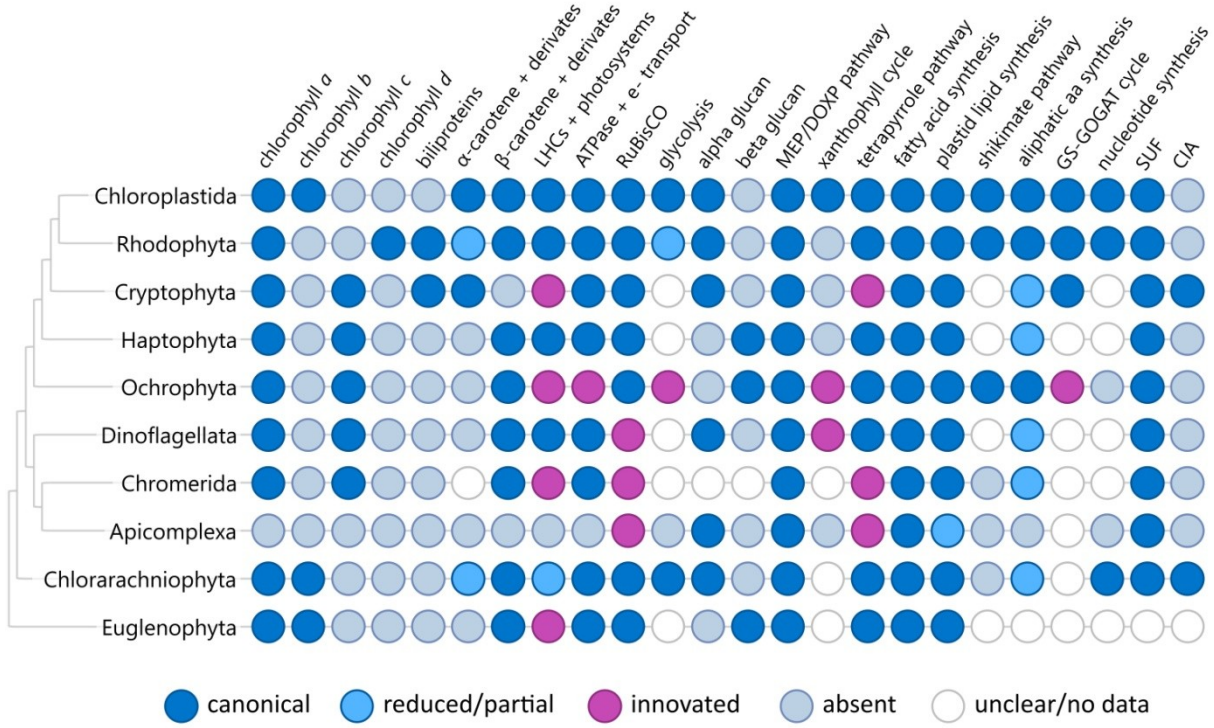

Figure 3: A graphical summary of metabolic capacities of plastids of different groups of primary and secondary algae as presented in chapters 1.2.5.2 and 1.2.5.3. Each dot represents either specific reaction/enzyme or metabolite, or a whole metabolic pathway or functional complex, blue means full or partial conservation, purple marks some degree of novelty or deviation from the canonical version, light blue-gray means presumed or proven absence, and white/colourless denotes uncertainty or missing data.

1.2.5.3 Euglenid plastids

The photosynthetic apparatus of euglenophytes is interesting from the genetic point of view, as its LHC subunits are very diverse and synthesized from large mRNAs in form of polyproteins which are cleaved to their mature size after import to plastid (Koziol & Durnford 2008). Similarly, the nuclear-coded RbcS is also synthesized as polyprotein comprising of eight units and maturated in the organelle (Tessier *et al.* 1995). These arrangements could have originated via uneven crossing-over and might be important for expression regulation. Energy metabolism of euglenophytes is quite adaptable, allowing these organisms to thrive in photoautotrophic, mixotrophic,

and heterotrophic conditions and even under anaerobiosis. In mixotrophic and heterotrophic conditions, the cells rely on OPPP which is likely cytosolic (Bégin-Heick 1973). Glycolytic enzymes, including both hexokinase and glucokinase, are present in *Euglena gracilis*, and some of these putatively localize to plastids (Belsky & Schultz 1962; Hasan *et al.* 2019). Interestingly, both RbcS and RbcL are retained in the secondarily non-photosynthetic *Euglena longa*. The sequences of both proteins are very divergent, however, and their function is unclear (Záhonová *et al.* 2016).

As a main storage molecule, euglenids use paramylon, an unbranched glucan with beta 1,3 glycosidic bonds. Its synthesis seems to be associated with plastids in plastid-bearing euglenids, but its origin in not algal or plastidial as it is present throughout all euglenids, including primarily heterotrophic ones, and not elsewhere (Calvayrac *et al.* 1981; Šantek *et al.* 2009; Tanaka *et al.* 2017; Hasan *et al.* 2019). Additionally, euglenids are able to produce wax esters under anaerobic conditions, but this biosynthesis is completely plastid-independent (Inui *et al.* 1982; Teerawanichpan & Qiu 2010).

Euglenophytes harbour two sets of tetrapyrrole pathway – one following 5-aminolevulinate synthesis by Shemin pathway, presumably in mitochondrion and/or cytosol, second using 5-aminolevulinate from C5 pathway and probably localized in plastid, providing both haem and chlorophyll (Kořený & Oborník 2011).

Additionally, both mevalonate and MEP/DOXP pathways for IPP synthesis are present in euglenophytes, again most likely spatially distributed between mitochondrion and plastid. Interestingly, it was proven biochemically that the IPP synthesized in plastidial MEP/DOXP pathway is used for carotenoid synthesis in plastid but does not contribute to phytol and sterol metabolism in the same compartment. Instead, the IPP produced in mitochondrion is used. Why and how are these two isolated pools of IPP maintained in *Euglena* is unclear (Kim *et al.* 2004). *Euglena* synthesizes a variety of carotenoids, with antheraxanthin being the most abundant one (Krinsky & Goldsmith 1960).

1.3 Plastid proteomics

1.3.1 General methods in proteomics

Protein mass spectrometry (MS) allows a high-throughput identification of proteins in a biological sample (reviewed in Walther & Mann 2010). Identity (mass and subsequently amino acid sequence) of analyzed molecules is inferred from their m/z (mass-to-charge) values measured by a mass spectrometer from their trajectory and/or speed in electro-magnetic field. For this, the molecules need to be ionized and in a gas phase which is often achieved using electrospray ionization method (Fenn *et al.* 1989). Characterization of protein molecules based on their mass becomes problematic in longer chains of similar amino acid frequency and length as their m/z values can overlap. Because of this, bottom-up proteomic methods are widely used, based on enzymatic digestion (usually by trypsin) of analyzed proteins to short peptides and their retrograde *in silico* mapping to full sequences. Alternative, top-down approach is only applied scarcely, in cases that allow it (McLafferty *et al.* 2007). For even higher resolution, a second level of MS detection is employed by further fragmentation of the peptide by a collision with gas molecules (collision-induced dissociation, CID) and MS analysis of the resulting fragments. This approach is termed tandem mass spectrometry (MS/MS or MS^2) and it is potentially able to perform *de novo* protein sequencing.

However, this is not enough to prevent mass spectra overlaps in case of rich protein mixtures (e.g. whole organelles or even cell lysates), so the complexity of these samples must be reduced even before the enzymatic digestion. This is achieved by separation on 1D or 2D gel or liquid chromatography. The most common setup of current machines is the liquid chromatography column being directly connected to the electrospray ionizer and mass spectrometer (LC-MS) with the analyzed molecules passing continuously through the system. It is also possible to include additional steps for enrichment of specific peptides (e.g. transmembrane or containing specific post-translational modifications).

There are several methods to determine m/z of the peptides. In a quadrupole mass filter, only particles of certain m/z are able to reach a detector, depending on the current setting. In a TOF (time-of-flight) analyzer, the speed of the peptides is detected and used for m/z inferrence (Ens & Standing 2005). An ion trap allows capturing and manipulation with a particular type of molecules. In orbital ion trap (orbitrap), the captured molecules are

orbiting around an electrode with a certain frequency which can be measured very accurately and used for the calculation of m/z (Makarov 2000).

The determined peptide masses and, by extension, sequences need to be mapped to a reference protein database using an identity detecting software. The most common include Mascot (Perkins *et al.* 1999), SEQUEST (Eng *et al.* 1994), and X! Tandem (Craig & Beavis 2004). Another identity detecting program, Andromeda, exists as a part of MaxQuant package for quantitative proteomics (Cox & Mann 2008).

There are multiple methods of relative quantification in proteomics. For very precise quantification, labelling is required. The most common methods include SILAC (stable isotope labelling with amino acids in cell mixture) where the cells are pre-cultivated in a medium containing isotope-labelled amino acids (Ong *et al.* 2002), iTRAQ (stable isotope-containing tags), and ICAT (isotope-coded affinity tag) which are both added later in the workflow and represent physical tags added onto the peptides as unique identifiers (Ross *et al.* 2004). The LOPIT method which allows high-throughput analysis of multiple cell fractions and precise determination of protein localization uses isotope-tagging combined with protein correlation profiling (PCP) based on known markers for different cell compartments (Dunkley *et al.* 2004; Drissi *et al.* 2013).

Label-free quantification (LFQ) methods represent a less precise but generally respected, cheaper, and more universal alternative. The mass spectra of a certain peptide can be compared across multiple samples through spectral counting or AUC (area-under-curve), both of these methods require recalculation and normalization (potentially taking to account various variables such as sample size, probability of detection, or peptide length) and validation by statistics (Old *et al.* 2005; Neilson *et al.* 2011).

1.3.2 Plastid proteomes

Most of the plastid proteomic studies carried out to this day focused on higher plants, with the proteomic databases of chloroplasts or other types of plastids generated for *Arabidopsis thaliana* (thale cress), maize, rice, potato, tobacco, bell pepper and *Medicago truncatula* (barrel medic) (Schubert *et al.* 2002; Watson *et al.* 2003; Kleffmann *et al.* 2004; van Wijk 2004; Baginsky *et al.* 2004; Siddique *et al.* 2006; Bräutigam *et al.* 2008; Stensballe *et al.* 2008; Daher *et al.* 2010; Wienkoop *et al.* 2010; Huang *et al.* 2013; J. Lee *et al.* 2013). Plant proteome database (PPDB) integrates proteomic

resources for *A. thaliana* and maize and is one of the most comprehensive and widely used databases regarding plant plastid proteomes (Sun *et al.* 2009). In general, up to around 3000 proteins localize to the organelle in plants, however, only a fraction of these is expected to be constitutively expressed in all types of tissues, life stages and growth conditions. In unicellular algae, the situation is less complicated as these variables boil down to strain and cultivation conditions.

However, not many whole plastid proteomes of unicellular algae were generated to this day, even less so outside the Archaeplastida group. The chloroplast proteome of *Chlamydomonas reinhardtii* determined by Terashima *et al.* (2011) comprises of 966 proteins and is generally very similar to that of plants. However, it revealed some differences in central carbon metabolism, with second half of glycolysis localized outside plastid, and pyruvate-ferredoxin oxidoreductase (PFOR) and several other iron-dependent enzymes of bacterial-like fermentative metabolism present, unlike in plants. There were also slight differences in PSII and LHC subunit composition (Terashima *et al.* 2011). Proteome of the cyanelle of glaucophyte *Cyanophora paradoxa* brough by Facchinelli *et al.* (2013) comprised of 510 proteins and also exhibited similar differences in photosystem organization and carbon metabolism, as well as several other unique and presumably ancestral features (Facchinelli *et al.* 2013).

In the case of complex plastids, the as of yet only mass spectrometry-determined proteomic datasets are available for the highly reduced non-photosynthetic apicoplast of *Plasmodium falciparum* (Boucher *et al.* 2018) and for the chlorarachniophyte *Bigelowiella natans* (Hopkins *et al.* 2012). The former consists of 346 proteins while the latter includes 302 proteins only and might be incomplete; especially proteins of periplastidial compartment might be absent due to their low abundance. Nevertheless, it provides a basis for many claims regarding the metabolic capacities of chlorarachniophyte plastid, some of which were presented in chapter 2.2.5.2, and an important contribution to plastid proteomics which is still lacking in the realm of unicellular algae.

2 AIMS

1. To annotate plastid proteome of *E. gracilis* and estimate its metabolic potential.
2. To reconstruct protein import pathway of euglenophyte plastids using transcriptomic data of *E. gracilis* and *E. longa* and proteomic data of *E. gracilis*.
3. To analyze characteristics of euglenophyte plastid-targeting domains based on the proteomic dataset determined by mass spectrometry.

3 LIST OF PUBLICATIONS AND AUTHOR CONTRIBUTION

Vanclová AMG, Hadariová L, Hrdá Š, Hampl V. **Secondary Plastids of Euglenophytes**. In: Y. Hirakawa (ed.), Advances in Botanical Research, Academic Press 2017. doi: 10.1016/bs.abr.2017.06.008

General layout and writing most of the chapter.

Záhonová K, Füssy Z, Birčák E, Novák Vanclová AMG, Klimeš V, Vesteg M, Krajčovič J, Oborník M, Eliáš M. **Peculiar features of the plastids of the colourless alga *Euglena longa* and photosynthetic euglenophytes unveiled by transcriptome analyses**. Sci Rep. 2018 Nov 19;8(1):17012. doi: 10.1038/s41598-018-35389-1.

Targeted search and annotation of proteins involved in plastid protein import.

Ebenezer TE, Zoltner M, Burrell A, Nenarokova A, Novák Vanclová AMG, Prasad B, Soukal P, Santana-Molina C, O'Neill E, Nankissoor NN, Vadakedath N, Daiker V, Obado S, Silva-Pereira S, Jackson AP, Devos DP, Lukeš J, Lebert M, Vaughan S, Hampl V, Carrington M, Ginger ML, Dacks JB, Kelly S, Field MC. **Transcriptome, proteome and draft genome of *Euglena gracilis*.** BMC Biol. 2019 Feb 7;17(1):11. doi: 10.1186/s12915-019-0626-8.

In silico prediction and annotation of plastid proteome from transcriptomic data.

Novák Vanclová AMG, Zoltner M, Kelly S, Soukal P, Záhonová K, Füssy Z, Ebenezer TE, Lacová Dobáková E, Eliáš M, Lukeš J, Field MC, Hampl V. **Metabolic quirks and the colourful history of the *Euglena gracilis* secondary plastid**. Under review.

Annotation of the mass spectrometry-based plastid proteome, reconstruction of metabolic pathways, and analysis of the plastid-targeting signals.

4 SUMMARY

Determination of plastid proteomic datasets, both mass spectrometry-based and predicted *in silico* from transcriptomic or genomic data, is important for the understanding of biogenesis, maintenance, metabolic capacity, and evolutionary history of these organelles. The latter is especially convoluted and intriguing in the case of complex plastids derived from secondary or higher-tier endosymbiotic events that are very common throughout eukaryotes and represent one of the major driving forces behind their diversification and evolutionary success. Euglenophytes and their plastids were characterized well morphologically and biochemically, but the lack of reliable genetic transformation system and genomic data makes further analysis, especially investigation of protein functions, difficult.

In this thesis, we summarized the state of knowledge regarding euglenophyte plastids (Vanclová et al. 2017) and analyzed newly generated transcriptomic and proteomic datasets (Záhonová *et al.* 2018; Ebenezer *et al.* 2019; Novák Vanclová *et al.* 2019), providing new resources usable in various fields of *Euglena* research, from evolutionary biology investigating the origin and spread of plastids, to applied sciences looking to harvest biotechnological potential of this organism.

The *E. gracilis* draft genome reported in Ebenezer *et al.* 2019 is 300-500 Mbp in size, in accord with previous estimates (Ebenezer *et al.* 2017). It is highly fragmented and incomplete, but its partial assembly suggests extreme expansion in non-coding sequence (>99%). The *E. gracilis* transcriptome is over 38 Mbp in size with CEGMA score of 87.9% and thus represents a comprehensive and largely complete dataset.

Some protein families, especially those related to cell signalling, underwent massive paralog duplication in *E. gracilis*. Less extensive duplications were noticed in some of the ER/Golgi-related proteins which might be related to plastid integration. Differential transcriptomics and proteomics revealed remarkably low correlation between light vs. dark up- or down-regulation of particular genes at transcriptional and translational level, with majority of the regulation taking place post-transcriptionally.

The transcriptome was used for *in silico* prediction of plastid proteome based on N-terminal signal domain composition and topology, yielding around 1,900 plastid candidate proteins reported in Ebenezer *et al.* 2019. This was later followed by determination based on liquid chromatography tandem mass-spectrometry (LC-MS/MS) of the isolated plastid

and mitochondrial fractions using the translated transcriptomic database as a reference. The relative label-free quantification (LFQ) of the proteins captured in at least two of the three technical replicates and in either of the fractions was used for calculation of CP/MT ratio reflecting the credibility of plastidial or mitochondrial localization of the particular protein to avoid cross-contaminations.

The resulting plastid proteomic dataset is reported in Novák Vanclová *et al.* 2019 and contains 1,345 protein groups, 43% of which could not be assigned a clear functional annotation or lacked homologs in other organisms whatsoever, suggesting a considerable potential for functional novelty and/or plasticity. A metabolic map of *E. gracilis* plastid reconstructed from the proteome provides additional evidence for some enzymatic processes described or proposed previously based on biochemical evidence, but also contains a number of novelties. Our data supports the existence of two distinct pools of IPP from plastidial MEP/DOXP and non-plastidial mevalonate pathways in *Euglena* and reveals a potential source of some of the phytol used for tocopherol, tocotrienol, and phylloquinone synthesis: recycling from chlorophyll mediated by VTE5, VTE6, and CLD1 (Lin *et al.* 2016). We recovered a full tetrapyrrole and chlorophyll synthesis pathway, including the C5 pathway for 5-aminolevulinate synthesis from glutamate, as well as carotenoid synthesis and part of xantophyll cycle. Enzymes of glutathion cycle and several reactions in polyamine metabolism are present but there is only very little enzymes and no complete pathway for amino acid synthesis. Shikimate pathway is also missing from the plastid proteome and transcriptome evidence suggests its localization in cytosol. This very low contribution of the plastid to amino acid metabolism is remarkable, albeit not completely exotic given the relatively common reductions in this part of metabolism putatively occurring in other complex plastids (see chapter 2.2.5.2). We also identified a second plastid-localized set of SUF proteins which differ from the standard plastidial ones in their generally shorter and less-conserved N-terminal domains and, more importantly, in phylogenetic origin as they are affiliated to Chlamydiae and likely represent a horizontal acquisition. This second SUF is conserved in other euglenophytes, suggesting it is not a transient state but an adaptive and functionally and/or spatially specialized machinery.

We performed a systematic phylogenetic screening of the whole set of plastidial proteins and provide a semi-quantitative evaluation of proportions of plastidial proteins evolutionarily associated with other than green algae, most notably haptophytes and ochrophytes, complementing the previously published studies on this matter (Markunas & Triemer 2016;

Ponce-Toledo *et al.* 2018; Ponce-Toledo *et al.* 2019) and bringing further support for the "shopping bag" (Larkum *et al.* 2007) and "red carpet" hypotheses (Ponce-Toledo *et al.* 2019).

Transcriptome was also generated for *Euglena longa*, a close relative of *E. gracilis* which in secondarily non-photosynthetic but still possesses a reduced plastid with 75 kbp genome which is (in contrast to *E. gracilis*) essential for its survival. It is reported in Záhonová *et al.* 2018, comprises of more than 65,000 transcripts, has BUSCO score of 89.1% (Simão *et al.* 2015), and provides important context to the conclusions based on the analyses of *E. gracilis* and other phototrophic euglenophytes.

The plastid protein import pathway was thoroughly searched in the new transcriptomes of *E. gracilis* and *E. longa* as well as available datasets from *Eutreptiella gymnastica* NIES-381 and *E. gymnastica*-like strain CCMP1597, revealing surprisingly reduced set of transporters. While the machineries mediating protein import to thylakoids were generally conserved in the three phototrophs and absent in *E. longa* which is believed (although this remains to be demonstrated by microscopy) to lack thylakoids, the expected translocases of plastid envelope, namely TOC and TIC components, were largely absent in all organisms. Only a homolog of Tic32 and a protein similar (although not equivalent) to Tic62 were present in all four, and Tic21 and a chlorophyll biogenesis protein related to Tic55 were conserved in phototrophs and absent from *E. longa*. No TOC subunit or other known chloroplast outer membrane protein was recovered in either of the transcriptomes, suggesting this is a genuine feature of euglenophyte plastid and not false negative result stemming from dataset incompleteness. The *E. gracilis* plastid proteome presented in Novák Vanclová *et al.* 2019 provides mass spectrometry-based back-up for these findings by recovering three isoforms of Tic21 (and the Tic55-related protein) as a high-credibility plastid proteins, but no evidence for Tic32 being involved in plastidial function. In addition to this, we propose several plastid-localized but originally ER- or Golgi-derived proteins, namely plastidial isoforms of Rab5 GTPase and SNARE protein GOSR1, as newly recruited components of plastid protein import pathway, potentially operating at the outermost membrane and involved in vesicle docking. The proteome also contains two rhomboid pseudoproteases similar to derlins which we speculate might serve an analogous purpose in euglenophyte plastids as their *bona fide* derlin counterparts in SELMA of "chromist" plastids, possibly substituting the missing TOC and even hinting at the eukaryotic origin of two, not just one envelope membrane of euglenophyte plastids.

We also took advantage of the mass-spectrometry determined set of nuclear-coded plastid-targeted *E. gracilis* proteins with high CP/MT ratio and complete N-termini and revisited their targeting domains in search of features which might have been overlooked previously due to the reliance on *in silico* predictions by algorithms trained on plant transit peptides or based on homology with plant plastid proteins. By using custom script for determination of hydrophobic regions and statistical analysis of amino acid composition of the expected TP region we indeed report TP characteristics not described elsewhere, most remarkably a substantial enrichment in proline which might affect its secondary and tertiary structure significantly, and a surprisingly large proportion of plastid-targeted proteins lacking the typical bipartite domain whatsoever. Both of this is in accord with the significant rearrangements in protein import pathway in euglenophyte plastid.

In summary, we bring protein-level support for conclusions of previous studies as well as numerous novel findings based on a new comprehensive set of *E. gracilis* plastid proteins determined by mass spectrometry and also *in silico* predictions from new high-quality transcriptomes of *E. gracilis* and *E. longa*. The plastid proteome of *E. gracilis* represents only second such dataset from an organism with photosynthetic complex plastid (the first being chlorarachniophyte *B. natans*, Hopkins *et al.* 2012) and one of the few full plastid proteomes of unicellular algae (Terashima *et al.* 2011; Facchinelli *et al.* 2013). Based on these data, we report some metabolic peculiarities of the euglenophyte plastid, including very low contribution to amino acid metabolism and additional SUF system of chlamydial origin. We describe an extensive reduction in plastid import machinery of the inner two plastid membranes, propose candidate translocases of the outer, host-derived membrane, and re-evaluate the characteristics of plastid-targeting signal domains.

5 REFERENCES

A.D. Neilson J, Rangsrikitphoti P, Durnford DG. 2017. Evolution and regulation of Bigelowiella natans light-harvesting antenna system. J Plant Physiol. 217:68–76. doi:10.1016/J.JPLPH.2017.05.019.

Adl SM, Bass D, Lane CE, Lukeš J, Schoch CL, Smirnov A, Agatha S, Berney C, Brown MW, Burki F, et al. 2019. Revisions to the Classification, Nomenclature, and Diversity of Eukaryotes. J Eukaryot Microbiol. 66(1):4–119. doi:10.1111/jeu.12691.

Allen AE, Dupont CL, Oborník M, Horák A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu H, Fernie AR, et al. 2011. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. Nature. 473(7346):203–207. doi:10.1038/nature10074.

Archibald JM. 2015. Genomic perspectives on the birth and spread of plastids. Proc Natl Acad Sci U S A. 112(33):1421374112-. doi:10.1073/pnas.1421374112.

Armstrong GA, Hearst JE. 1996. Carotenoids 2: Genetics and molecular biology of carotenoid pigment biosynthesis. FASEB J. 10(2):228–37. doi:10.1096/fasebj.10.2.8641556.

Bach LT, Mackinder LCM, Schulz KG, Wheeler G, Schroeder DC, Brownlee C, Riebesell U. 2013. Dissecting the impact of CO_2 and pH on the mechanisms of photosynthesis and calcification in the coccolithophore *Emiliania huxleyi*. New Phytol. 199(1):121–134. doi:10.1111/nph.12225.

Baginsky S, Siddique A, Gruissem W. 2004. Proteome Analysis of Tobacco Bright Yellow-2 (BY-2) Cell Culture Plastids as a Model for Undifferentiated Heterotrophic Plastids. doi:10.1021/PR0499186.

Bailleul B, Berne N, Murik O, Petroutsos D, Prihoda J, Tanaka A, Villanova V, Bligny R, Flori S, Falconet D, et al. 2015. Energetic coupling between plastids and mitochondria drives CO2 assimilation in diatoms. Nature. 524(7565):366–369. doi:10.1038/nature14599.

Ball SG, Subtil A, Bhattacharya D, Moustafa A, Weber APM, Gehre L, Colleoni C, Arias M-C, Cenci U, Dauvillée D. 2013. Metabolic Effectors Secreted by Bacterial Pathogens: Essential Facilitators of Plastid Endosymbiosis? Plant Cell. 25(1):7–21. doi:10.1105/TPC.112.101329.

Barbrook AC, Howe CJ, Kurniawan DP, Tarr SJ. 2010. Organization and expression of organellar genomes. Philos Trans R Soc B Biol Sci. 365(1541):785–797. doi:10.1098/rstb.2009.0250.

Barras DR, Stone BA. 1968. Carbohydrate composition and metabolism in Euglena. Biol Euglena. 2:149–191.

Baurain D, Brinkmann H, Petersen J, Rodríguez-Ezpeleta N, Stechmann A, Demoulin V, Roger AJ, Burger G, Lang BF, Philippe H. 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. Mol Biol Evol. 27(7):1698–709. doi:10.1093/molbev/msq059.

Beardall J, Raven JA. 2012. Algal Metabolism. In: eLS. Chichester, UK: John Wiley & Sons, Ltd.

Bégin-Heick N. 1973. The localization of enzymes of intermediary metabolism in Astasia and Euglena. Biochem J. 134(2):607–16. doi:10.1042/bj1340607.

Belsky M, Schultz J. 1962. Partial Characterization of Hexokinase from Euglena gracilis var. bacillaris. J Protozool. 9(2):195–200. doi:10.1111/j.1550-7408.1962.tb02605.x.

Bendich AJ. 2004. Circular chloroplast chromosomes: the grand illusion. Plant Cell. 16(7):1661–6. doi:10.1105/tpc.160771.

Bodył A, Mackiewicz P, Milanowski R. 2010. Did Trypanosomatid Parasites Contain a Eukaryotic Alga-Derived Plastid in Their Evolutionary Past? J Parasitol. 96(2):465–475. doi:Doi 10.1645/Ge-1810.1.

Bolte K, Bullmann L, Hempel F, Bozarth A, Zauner S, Maier U-G. 2009. Protein targeting into secondary plastids. J Eukaryot Microbiol. 56(1):9–15. doi:10.1111/j.1550-7408.2008.00370.x.

Bölter B, Soll J. 2017. Ycf1/Tic214 Is Not Essential for the Accumulation of Plastid Proteins. doi:10.1016/j.molp.2016.10.012.

Boucher MJ, Ghosh S, Zhang L, Lal A, Jang SW, Ju A, Zhang S, Wang X, Ralph SA, Zou J, et al. 2018. Integrative proteomics and bioinformatic prediction enable a high-confidence apicoplast proteome in malaria parasites. Striepen B, editor. PLOS Biol. 16(9):e2005895. doi:10.1371/journal.pbio.2005895.

Boudière L, Michaud M, Petroutsos D, Rébeillé F, Falconet D, Bastien O, Roy S, Finazzi G, Rolland N, Jouhet J, et al. 2014. Glycerolipids in photosynthesis: Composition, synthesis and trafficking. Biochim Biophys Acta - Bioenerg. 1837(4):470–480. doi:10.1016/J.BBABIO.2013.09.007.

Bräutigam A, Hoffmann-Benning S, Hofmann-Benning S, Weber APM. 2008. Comparative proteomics of chloroplast envelopes from C3 and C4 plants reveals specific adaptations of the plastid envelope to C4 photosynthesis and candidate proteins required for maintaining C4 metabolite fluxes. Plant Physiol. 148(1):568–79. doi:10.1104/pp.108.121012.

Breglia SA, Yubuki N, Leander BS. 2013. Ultrastructure and Molecular Phylogenetic Position of *Heteronema scaphurum* : A Eukaryovorous Euglenid with a Cytoproct. J Eukaryot Microbiol. 60(2):107–120. doi:10.1111/jeu.12014.

Broughton MJ, Howe CJ, Hiller RG. 2006. Distinctive organization of genes for light-harvesting proteins in the cryptophyte alga Rhodomonas. Gene. 369:72–9. doi:10.1016/j.gene.2005.10.026.

Bruce BD. 2000. Chloroplast transit peptides: structure, function and evolution. Trends Cell Biol. 10(10):440–447. doi:10.1016/S0962-8924(00)01833-X.

Bullmann L, Haarmann R, Mirus O, Bredemeier R, Hempel F, Maier UG, Schleiff E. 2010. Filling the gap, evolutionarily conserved Omp85 in plastids of chromalveolates. J Biol Chem. 285(9):6848–56. doi:10.1074/jbc.M109.074807.

Burki F. 2017. The Convoluted Evolution of Eukaryotes With Complex Plastids. 1st ed. Elsevier Ltd.

Burki F, Kaplan M, Tikhonenkov D V., Zlatogursky V, Minh BQ, Radaykina L V., Smirnov A, Mylnikov AP, Keeling PJ. 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. Proc R Soc B Biol Sci. 283(1823):20152802. doi:10.1098/rspb.2015.2802.

Cai X, Fuller AL, McDougald LR, Zhu G. 2003. Apicoplast genome of the coccidian Eimeria tenella. Gene. 321:39–46. doi:10.1016/j.gene.2003.08.008.

Calvayrac R, Laval-Martin D, Briand J, Farineau J. 1981. Paramylon synthesis by Euglena gracilis photoheterotrophically grown under low O2 pressure. Planta. 153(1):6–13. doi:10.1007/BF00385311.

Cao S, Zhang X, Xu D, Fan X, Mou S, Wang Y, Ye N, Wang W. 2013. A transthylakoid proton gradient and inhibitors induce a non-photochemical fluorescence quenching in unicellular algae Nannochloropsis sp. FEBS Lett. 587(9):1310–1315. doi:10.1016/J.FEBSLET.2012.12.031.

Cavalier-Smith T. 1981. Eukaryote kingdoms: Seven or nine? Biosystems. 14(3–4):461–481. doi:10.1016/0303-2647(81)90050-2.

Cavalier-Smith T. 1999. Principles of Protein and Lipid Targeting in Secondary Symbiogenesis: Euglenoid, Dinoflagellate, and Sporozoan Plastid Origins and the Eukaryote Family Tree. J Eukaryot Microbiol. 46(4):347–366. doi:10.1111/j.1550-7408.1999.tb04614.x.

Cavalier-Smith T. 2000. Membrane heredity and early chloroplast evolution. Trends Plant Sci. 5(4):174–182.

Cavalier-Smith T. 2016. Higher classification and phylogeny of Euglenozoa. Eur J Protistol. 56:250–276. doi:10.1016/j.ejop.2016.09.003.

Cenci U, Ducatez M, Kadouche D, Colleoni C, Ball SG. 2016. Was the Chlamydial Adaptative Strategy to Tryptophan Starvation an Early Determinant of Plastid Endosymbiosis? Front Cell Infect Microbiol. 6:67. doi:10.3389/fcimb.2016.00067.

Cihlář J, Füssy Z, Horák A, Oborník M, Zdobnov E, Yuan J. 2016. Evolution of the Tetrapyrrole Biosynthetic Pathway in Secondary Algae: Conservation, Redundancy and Replacement. Prigent C, editor. PLoS One. 11(11):e0166338. doi:10.1371/journal.pone.0166338.

Ciugulea I, Triemer RE. 2010. A color atlas of photosynthetic euglenoids. Michigan State University Press.

Clark SA, Theg SM. 1997. A folded protein can be transported across the chloroplast envelope and thylakoid membranes. Mol Biol Cell. 8(5):923–34. doi:10.1091/mbc.8.5.923.

Copertino DW, Hallick RB. 1993. Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. Trends Biochem Sci. 18(12):467–471. doi:10.1016/0968-0004(93)90008-B.

Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 26(12):1367–1372. doi:10.1038/nbt.1511.

Craig R, Beavis RC. 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics. 20(9):1466–1467. doi:10.1093/bioinformatics/bth092.

Dabbagh N, Preisfeld A. 2016. The Chloroplast Genome of Euglena mutabilis-Cluster Arrangement, Intron Analysis, and Intrageneric Trends. J Eukaryot Microbiol.(1993):31–44. doi:10.1111/jeu.12334.

Daher Z, Recorbet G, Valot B, Robert F, Balliau T, Potin S, Schoefs B, Dumas-Gaudot E. 2010. Proteomic analysis of Medicago truncatula root plastids. Proteomics. 10(11):2123–2137. doi:10.1002/pmic.200900345.

Day PM, Theg SM. 2018. Evolution of protein transport to the chloroplast envelope membranes. Photosynth Res. 138(3):315–326. doi:10.1007/s11120-018-0540-x.

Demarsy E, Lakshmanan AM, Kessler F. 2014. Border control: selectivity of chloroplast protein import and regulation at the TOC-complex. Front Plant Sci. 5:483. doi:10.3389/fpls.2014.00483.

Doetsch NA. 2000. Group III intron structure and evolutionary analysis in euglenoid chloroplast genomes.

Doetsch NA, Favreau MR, Kuscuoglu N, Thompson MD, Hallick RB. 2001. Chloroplast transformation in Euglena gracilis: splicing of a group III twintron transcribed from a transgenic psbK operon. Curr Genet. 39(1):49–60.

Domman D, Horn M, Embley TM, Williams TA. 2015. Plastid establishment did not require a chlamydial partner. Nat Commun. 6(1):6421. doi:10.1038/ncomms7421.

van Dooren GG, Kennedy AT, McFadden GI. 2012. The Use and Abuse of haem in Apicomplexan Parasites. Antioxid Redox Signal. 17(4):634–656. doi:10.1089/ars.2012.4539.

Dorrell RG, Bowler C. 2017. Secondary Plastids of Stramenopiles. 1st ed. Elsevier Ltd.

Dorrell RG, Gile G, McCallum G, Méheust R, Bapteste EP, Klinger CM, Brillet-Guéguen L, Freeman KD, Richter DJ, Bowler C. 2017. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. Elife. 6. doi:10.7554/eLife.23717.

Dorrell RG, Howe CJ. 2015. Integration of plastids with their hosts: Lessons learned from dinoflagellates. Proc Natl Acad Sci U S A. 112(33):10247–54. doi:10.1073/pnas.1421380112.

Dorrell RG, Smith AG. 2011. Do red and green make brown?: perspectives on plastid acquisitions within chromalveolates. Eukaryot Cell. 10(7):856–68. doi:10.1128/EC.00326-10.

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG. 2001. The highly reduced genome of an enslaved algal nucleus. Nature. 410(6832):1091–6. doi:10.1038/35074092.

Douglas SE, Penny SL. 1999. The Plastid Genome of the Cryptophyte Alga, Guillardia theta: Complete Sequence and Conserved Synteny Groups Confirm Its Common Ancestry with Red Algae. J Mol Evol. 48(2):236–244. doi:10.1007/PL00006462.

Drissi R, Dubois M, Boisvert F. 2013. Proteomics methods for subcellular proteome analysis. FEBS J. 280(22):5626–5634. doi:10.1111/FEBS.12502

Dunkley TPJ, Watson R, Griffin JL, Dupree P, Lilley KS. 2004. Localization of organelle proteins by isotope tagging (LOPIT). Mol Cell Proteomics. 3(11):1128–34. doi:10.1074/mcp.T400009-MCP200.

Durnford DG, Gray MW. 2006. Analysis of Euglena gracilis plastid-targeted proteins reveals different classes of transit sequences. Eukaryot Cell. 5(12):2079–2091. doi:10.1128/EC.00222-06.

Ebenezer TE, Carrington M, Lebert M, Kelly S, Field MC. 2017. Euglena gracilis Genome and Transcriptome: Organelles, Nuclear Genome Assembly Strategies and Initial Features. Springer, Cham. p. 125–140.

Ebenezer TE, Zoltner M, Burrell A, Nenarokova A, Novák Vanclová AMG, Prasad B, Soukal P, Santana-Molina C, O'Neill E, Nankissoor NN, et al. 2019. Transcriptome, proteome and draft genome of Euglena gracilis. BMC Biol. 17(1):11. doi:10.1186/s12915-019-0626-8.

Ehrenberg CG. 1830. Organisation, Systematik und geographisches Verhältniss der Infusionsthierchen. Berlin: Druckerei der Königlichen Akademie der Wissenschaften.

Eng JK, McCormack AL, Yates JR. 1994. An approach to correlate tandem mass spectral data

of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom. 5(11):976–989. doi:10.1016/1044-0305(94)80016-2.

Enomoto T, Sulli C, Schwartzbach SD. 1997. 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(6):743–746. doi:10.1093/oxfordjournals.pcp.a029229.

Ens W, Standing KG. 2005. Hybrid Quadrupole/Time-of-Flight Mass Spectrometers for Analysis of Biomolecules. Methods Enzymol. 402:49–78. doi:10.1016/S0076-6879(05)02002-1.

Facchinelli F, Colleoni C, Ball SG, Weber APM. 2013. Chlamydia, cyanobiont, or host: who was on top in the ménage à trois? Trends Plant Sci. 18(12):673–9. doi:10.1016/j.tplants.2013.09.006.

Facchinelli F, Pribil M, Oster U, Ebert NJ, Bhattacharya D, Leister D, Weber APM. 2013. Proteomic analysis of the Cyanophora paradoxa muroplast provides clues on early events in plastid endosymbiosis. Planta. 237(2):637–651. doi:10.1007/s00425-012-1819-3.

Felsner G, Sommer MS, Gruenheit N, Hempel F, Moog D, Zauner S, Martin W, Maier UG. 2011. ERAD components in organisms with complex red plastids suggest recruitment of a preexisting protein transport pathway for the periplastid membrane. Genome Biol Evol. 3(0):140–50. doi:10.1093/gbe/evq074.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. 1989. Electrospray ionization for mass spectrometry of large biomolecules. Science. 246(4926):64–71. doi:10.1126/science.2675315.

Flegontov P, Gray MW, Burger G, Lukeš J. 2011. Gene fragmentation: a key to mitochondrial genome evolution in Euglenozoa? Curr Genet. 57(4):225–32. doi:10.1007/s00294-011-0340-8.

Flori S, Jouneau PH, Finazzi G, Maréchal E, Falconet D. 2016. Ultrastructure of the Periplastidial Compartment of the Diatom Phaeodactylum tricornutum. Protist. 167(3):254–267. doi:10.1016/j.protis.2016.04.001.

Gabrielsen TM, Minge MA, Espelund M, Tooming-Klunderud A, Patil V, Nederbragt AJ, Otis C, Turmel M, Shalchian-Tabrizi K, Lemieux C, et al. 2011. Genome Evolution of a Tertiary Dinoflagellate Plastid. Nikolaidis N, editor. PLoS One. 6(4):e19132. doi:10.1371/journal.pone.0019132.

Gavel Y, von Heijne G. 1990. A conserved cleavage-site motif in chloroplast transit peptides. FEBS Lett. 261(2):455–458. doi:10.1016/0014-5793(90)80614-O.

Gentle IE, Burri L, Lithgow T. 2005. Molecular architecture and function of the Omp85 family of proteins. Mol Microbiol. 58(5):1216–1225. doi:10.1111/j.1365-2958.2005.04906.x.

Gibbs SP. 1978. The chloroplasts of Euglena may have evolved from symbiotic green algae. Can J Bot. 56(22):2883–2889. doi:10.1139/b78-345.

Gibbs SP. 1981. The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. Ann N Y Acad Sci. 361(1 Origins and E):193–208. doi:10.1111/j.1749-6632.1981.tb54365.x.

Gilson PR, Su V, Slamovits CH, Reith ME, Keeling PJ, McFadden GI. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: nature's smallest nucleus. Proc Natl Acad Sci U S A. 103(25):9566–71. doi:10.1073/pnas.0600707103.

Gockel G, Hachtel W. 2000. Complete Gene Map of the Plastid Genome of the Nonphotosynthetic Euglenoid Flagellate Astasia longa. Protist. 151(4):347–351. doi:10.1078/S1434-4610(04)70033-4.

Gould SB, Fan E, Hempel F, Maier U-G, Klösgen RB. 2007. Translocation of a phycoerythrin alpha subunit across five biological membranes. J Biol Chem. 282(41):30295–302. doi:10.1074/jbc.M701869200.

Gould SB, Maier U-G, Martin WF. 2015. Protein Import and the Origin of Red Complex Plastids. Curr Biol. 25(12):R515–R521. doi:10.1016/J.CUB.2015.04.033.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. Plant J. 66(1):34–44. doi:10.1111/j.1365-313X.2011.04541.x.

Grosche C, Diehl A, Rensing SA, Maier UG. 2018. Iron–Sulfur Cluster Biosynthesis in Algae with Complex Plastids. Embley M, editor. Genome Biol Evol. 10(8):2061–2071. doi:10.1093/gbe/evy156.

Gruber A, Vugrinec S, Hempel F, Gould SB, Maier U-G, Kroth PG. 2007. Protein targeting into complex diatom plastids: functional characterisation of a specific targeting motif. Plant Mol Biol. 64(5):519–530. doi:10.1007/s11103-007-9171-x.

Guedes R, Prosdocimi F, Fernandes G, Moura L, Ribeiro H, Ortega J. 2011. Amino acids biosynthesis and nitrogen assimilation pathways: a great genomic deletion during eukaryotes evolution. BMC Genomics. 12(Suppl 4):S2. doi:10.1186/1471-2164-12-S4-S2.

Guggisberg AM, Amthor RE, Odom AR. 2014. Isoprenoid biosynthesis in Plasmodium falciparum. Eukaryot Cell. 13(11):1348–59. doi:10.1128/EC.00160-14.

Guiry MD, Guiry GM. 2017. AlgaeBase. World-wide electronic publication. Natl Univ Ireland, Galw.:http://www.algaebase.org; searched on 06 June 2017.

Gumińska N, Płecha M, Zakryś B, Milanowski R. 2018. Order of removal of conventional and nonconventional introns from nuclear transcripts of Euglena gracilis. Field MC, editor. PLOS Genet. 14(10):e1007761. doi:10.1371/journal.pgen.1007761.

Häder D-P, Richter PR, Schuster M, Daiker V, Lebert M. 2009. Molecular analysis of the graviperception signal transduction in the flagellate Euglena gracilis: Involvement of a transient receptor potential-like channel and a calmodulin. Adv Sp Res. 43(8):1179–1184. doi:10.1016/j.asr.2009.01.029.

Haferkamp I, Deschamps P, Ast M, Jeblick W, Maier U, Ball S, Neuhaus HE. 2006. Molecular and biochemical analysis of periplastidial starch metabolism in the cryptophyte Guillardia theta. Eukaryot Cell. 5(6):964–71. doi:10.1128/EC.00381-05.

Hagopian JC, Reis M, Kitajima JP, Bhattacharya D, de Oliveira MC. 2004. Comparative Analysis of the Complete Plastid Genome Sequence of the Red Alga Gracilaria tenuistipitata var. liui Provides Insights into the Evolution of Rhodoplasts and Their Relationship to Other Plastids. J Mol Evol. 59(4):464–477. doi:10.1007/s00239-004-2638-3.

Haimovich-Dayan M, Garfinkel N, Ewe D, Marcus Y, Gruber A, Wagner H, Kroth PG, Kaplan A. 2013. The role of C $_4$ metabolism in the marine diatom *Phaeodactylum tricornutum*. New Phytol. 197(1):177–185. doi:10.1111/j.1469-8137.2012.04375.x.

Hajdukiewicz PTJ, Allison LA, Maliga P. 1997. The two RNA polymerases encoded by the nuclear and the plastid compartments transcribe distinct groups of genes in tobacco plastids. EMBO J. 16(13):4041–4048. doi:10.1093/emboj/16.13.4041.

Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E. 1993. Complete sequence of Euglena gracilis chloroplast DNA. Nucleic Acids Res. 21(15):3537–3544. doi:10.1093/nar/21.15.3537.

Hannaert V, Saavedra E, Duffieux F, Szikora J-P, Rigden DJ, Michels PAM, Opperdoes FR. 2003. Plant-like traits associated with metabolism of Trypanosoma parasites. Proc Natl Acad Sci U S A. 100(3):1067–71. doi:10.1073/pnas.0335769100.

Harris J. 1695. Some Microscopical Observations of Vast Numbers of Animalcula Seen in Water by John Harris, M. A. Rector of Winchelsea in Sussex, and F. R. S. Philos Trans R Soc London. 19(215–235):254–259. doi:10.1098/rstl.1695.0036.

Hasan MT, Sun A, Khatiwada B, McQuade L, Mirzaei M, Te'o J, Hobba G, Sunna A, Nevalainen H. 2019. Comparative proteomics investigation of central carbon metabolism in Euglena gracilis grown under predominantly phototrophic, mixotrophic and heterotrophic cultivations. Algal Res. 43:101638. doi:10.1016/J.ALGAL.2019.101638.

Hastings KEM. 2005. SL trans-splicing: easy come or easy go? Trends Genet. 21(4):240–7. doi:10.1016/j.tig.2005.02.005.

Hehenberger E, Imanian B, Burki F, Keeling PJ. 2014. Evidence for the retention of two evolutionary distinct plastids in dinoflagellates with diatom endosymbionts. Genome Biol Evol. 6(9):2321–34. doi:10.1093/gbe/evu182.

Heins L, Mehrle A, Hemmler R, Wagner R, Küchler M, Hörmann F, Sveshnikov D, Soll J. 2002. The preprotein conducting channel at the inner envelope membrane of plastids. EMBO J. 21(11):2616–25. doi:10.1093/emboj/21.11.2616.

Hempel F, Bullmann L, Lau J, Zauner S, Maier UG. 2009. ERAD-derived preprotein transport across the second outermost plastid membrane of diatoms. Mol Biol Evol. 26(8):1781–90. doi:10.1093/molbev/msp079.

Hempel F, Felsner G, Maier UG. 2010. New mechanistic insights into pre-protein transport across the second outermost plastid membrane of diatoms. Mol Microbiol. 76(3):793–801. doi:10.1111/j.1365-2958.2010.07142.x.

Herrin DL, Nickelsen J. 2004. Chloroplast RNA processing and stability. Photosynth Res. 82(3):301–314. doi:10.1007/s11120-004-2741-8.

Heyes DJ, Neil Hunter C. 2009. Biosynthesis of Chlorophyll and Bacteriochlorophyll. In: Tetrapyrroles. New York, NY: Springer New York. p. 235–249.

Hill DRA, Rowan KS. 1989. The biliproteins of the Cryptophyceae. Phycologia. 28(4):455–463. doi:10.2216/i0031-8884-28-4-455.1.

Hinnah SC, Wagner R, Sveshnikova N, Harrer R, Soll J. 2002. The Chloroplast Protein Import Channel Toc75: Pore Properties and Interaction with Transit Peptides. Biophys J. 83(2):899–911. doi:10.1016/S0006-3495(02)75216-8.

Hirakawa Y, Burki F, Keeling PJ. 2012. Genome-based reconstruction of the protein import machinery in the secondary plastid of a chlorarachniophyte alga. Eukaryot Cell. 11(3):324–33. doi:10.1128/EC.05264-11.

Hjorth E, Hadfi K, Zauner S, Maier U-G. 2005. Unique genetic compartmentalization of the SUF system in cryptophytes and characterization of a SufD mutant in *Arabidopsis thaliana*. FEBS Lett. 579(5):1129–1135. doi:10.1016/j.febslet.2004.12.084.

Holbrook K, Subramanian C, Chotewutmontri P, Reddick LE, Wright S, Zhang H, Moncrief

L, Bruce BD. 2016. Functional Analysis of Semi-conserved Transit Peptide Motifs and Mechanistic Implications in Precursor Targeting and Recognition. Mol Plant. 9(9):1286–1301. doi:10.1016/J.MOLP.2016.06.004.

Hopkins JF, Spencer DF, Laboissiere S, Neilson JAD, Eveleigh RJM, Durnford DG, Gray MW, Archibald JM. 2012. Proteomics Reveals Plastid- and Periplastid-Targeted Proteins in the Chlorarachniophyte Alga Bigelowiella natans. Genome Biol Evol. 4(12):1391–1406. doi:10.1093/gbe/evs115.

Howe C J, Barbrook a C, Nisbet RER, Lockhart PJ, Larkum a WD. 2008. The origin of plastids. Philos Trans R Soc Lond B Biol Sci. 363(1504):2675–85. doi:10.1098/rstb.2008.0050.

Howe C. J., Nisbet RER, Barbrook AC. 2008. The remarkable chloroplast genome of dinoflagellates. J Exp Bot. 59(5):1035–1045. doi:10.1093/jxb/erm292.

Hrdá Š, Fousek J, Szabová J, Hampl V, Vlček Č. 2012. The plastid genome of eutreptiella provides a window into the process of secondary endosymbiosis of plastid in euglenids. PLoS One. 7(3):e33746. doi:10.1371/journal.pone.0033746.

Huang J, Gogarten J. 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biol. 8(6):R99. doi:10.1186/gb-2007-8-6-r99.

Huang M, Friso G, Nishimura K, Qu X, Olinares PDB, Majeran W, Sun Q, van Wijk KJ. 2013. Construction of Plastid Reference Proteomes for Maize and *Arabidopsis* and Evaluation of Their Orthologous Relationships; The Concept of Orthoproteomics. J Proteome Res. 12(1):491–504. doi:10.1021/pr300952g.

Imura T, Sato S, Sato Y, Sakamoto D, Isobe T, Murata K, Holder AA, Yukawa M. 2014. The apicoplast genome of Leucocytozoon caulleryi, a pathogenic apicomplexan parasite of the chicken. Parasitol Res. 113(3):823–8. doi:10.1007/s00436-013-3712-9.

Inaba T, Li M, Alvarez-Huerta M, Kessler F, Schnell DJ. 2003. atTic110 Functions as a Scaffold for Coordinating the Stromal Events of Protein Import into Chloroplasts. J Biol Chem. 278(40):38617–38627. doi:10.1074/jbc.M306367200.

Inagaki J, Fujita Y, Hase T, Yamamoto Y. 2000. Protein translocation within chloroplast is similar in Euglena and higher plants. Biochem Biophys Res Commun. 277(2):436–442. doi:10.1006/bbrc.2000.3702.

Inui H, Miyatake K, Nakano Y, Kitaoka S. 1982. Wax ester fermentation in *Euglena gracilis*. FEBS Lett. 150(1):89–93. doi:10.1016/0014-5793(82)81310-0.

Iseki M, Matsunaga S, Murakami A, Ohno K, Shiga K, Yoshida K, Sugai M, Takahashi T, Hori T, Watanabe M. 2002. A blue-light-activated adenylyl cyclase mediates photoavoidance in Euglena gracilis. Nature. 415(6875):1047–1051. doi:10.1038/4151047a.

Jackson C, Knoll AH, Chan CX, Verbruggen H. 2018. Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green plastids. Sci Rep. 8(1):1523. doi:10.1038/s41598-017-18805-w.

James TW, Crescitelli F, Loew ER, McFarland WN. 1992. The eyespot of euglena gracilis: a microspectrophotometric study. Vision Res. 32(9):1583–1591. doi:10.1016/0042-6989(92)90151-8.

Janouškovec J, Horák A, Oborník M, Lukeš J, Keeling PJ. 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proc Natl Acad Sci. 107(24):10949–10954. doi:10.1073/PNAS.1003335107.

Janouškovec J, Tikhonenkov D V., Burki F, Howe AT, Kolísko M, Mylnikov AP, Keeling PJ. 2015. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. Proc Natl Acad Sci. 112(33):10200–10207. doi:10.1073/pnas.1423790112.

Jarvis P, Robinson C. 2004. Mechanisms of protein import and routing in chloroplasts. Curr Biol. 14(24):R1064-77. doi:10.1016/j.cub.2004.11.049.

Kanehisa M, Goto S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. 28(1):27–30. doi:10.1093/nar/28.1.27.

Karnkowska A, Bennett MS, Triemer RE. 2018. Dynamic evolution of inverted repeats in Euglenophyta plastid genomes. Sci Rep. 8(1):16071. doi:10.1038/s41598-018-34457-w.

Karnkowska A, Bennett MS, Watza D, Kim JI, Zakryś B, Triemer RE. 2015. Phylogenetic Relationships and Morphological Character Evolution of Photosynthetic Euglenids (Excavata) Inferred from Taxon-rich Analyses of Five Genes. J Eukaryot Microbiol. 62(3):362–373. doi:10.1111/jeu.12192.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, et al. 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. Roberts RG, editor. PLoS Biol. 12(6):e1001889. doi:10.1371/journal.pbio.1001889.

Keenan RJ, Freymann DM, Stroud RM, Walter P. 2001. The Signal Recognition Particle. Annu Rev Biochem. 70(1):755–775. doi:10.1146/annurev.biochem.70.1.755.

Khan H, Parks N, Kozera C, Curtis BA, Parsons BJ, Bowman S, Archibald JM. 2007. Plastid Genome Sequence of the Cryptophyte Alga Rhodomonas salina CCMP1319: Lateral Transfer of Putative DNA Replication Machinery and a Test of Chromist Plastid Phylogeny. Mol Biol Evol. 24(8):1832–1842. doi:10.1093/molbev/msm101.

Khatiwada B, Kautto L, Sunna A, Sun A, Nevalainen H. 2019. Nuclear transformation of the versatile microalga Euglena gracilis. Algal Res. 37:178–185. doi:10.1016/J.ALGAL.2018.11.022.

Kikuchi S, Bédard J, Hirano M, Hirabayashi Y, Oishi M, Imai M, Takase M, Ide T, Nakai M. 2013. Uncovering the Protein Translocon at the Chloroplast Inner Envelope Membrane. Science (80-). 339(6119):571–574. doi:10.1126/science.1229262.

Kikuchi S, Oishi M, Hirabayashi Y, Lee DW, Hwang I, Nakai M. 2009. A 1-megadalton translocation complex containing Tic20 and Tic21 mediates chloroplast protein import at the inner envelope membrane. Plant Cell. 21(6):1781–97. doi:10.1105/tpc.108.063552.

Kilian O, Kroth PG. 2005. Identification and characterization of a new conserved motif within the presequence of proteins targeted into complex diatom plastids. Plant J. 41(2):175–83. doi:10.1111/j.1365-313X.2004.02294.x.

Kim D, Filtz MR, Proteau PJ. 2004. The methylerythritol phosphate pathway contributes to carotenoid but not phytol biosynthesis in Euglena gracilis. J Nat Prod. 67(6):1067–1069. doi:10.1021/np049892x.

Kim J, Fabris M, Baart G, Kim MK, Goossens A, Vyverman W, Falkowski PG, Lun DS. 2016. Flux balance analysis of primary metabolism in the diatom *Phaeodactylum tricornutum*. Plant J. 85(1):161–176. doi:10.1111/tpj.13081.

Kim S, Park MG. 2016. Paulinella longichromatophora sp. nov., a New Marine

Photosynthetic Testate Amoeba Containing a Chromatophore. Protist. 167(1):1–12. doi:10.1016/J.PROTIS.2015.11.003.

Kleffmann T, Russenberger D, von Zychlinski A, Christopher W, Sjölander K, Gruissem W, Baginsky S. 2004. The Arabidopsis thaliana Chloroplast Proteome Reveals Pathway Abundance and Novel Protein Functions. Curr Biol. 14(5):354–362. doi:10.1016/J.CUB.2004.02.039.

Knauf U, Hachtel W. 2002. The genes encoding subunits of ATP synthase are conserved in the reduced plastid genome of the heterotrophic alga Prototheca wickerhamii. Mol Genet Genomics. 267(4):492–497. doi:10.1007/s00438-002-0681-6.

Köhler D, Montandon C, Hause G, Majovsky P, Kessler F, Baginsky S, Agne B. 2015. Characterization of chloroplast protein import without Tic56, a component of the 1-megadalton translocon at the inner envelope membrane of chloroplasts. Plant Physiol. 167(3):972–90. doi:10.1104/pp.114.255562.

de Koning AP, Keeling PJ. 2006. The complete plastid genome sequence of the parasitic green alga Helicosporidium sp. is highly reduced and structured. BMC Biol. 4:12. doi:10.1186/1741-7007-4-12.

Kořený L, Oborník M. 2011. Sequence evidence for the presence of two tetrapyrrole pathways in Euglena gracilis. Genome Biol Evol. 3(1):359–364. doi:10.1093/gbe/evr029.

Koreny L, Sobotka R, Janouskovec J, Keeling PJ, Oborník M. 2011. Tetrapyrrole synthesis of photosynthetic chromerids is likely homologous to the unusual pathway of apicomplexan parasites. Plant Cell. 23(9):3454–62. doi:10.1105/tpc.111.089102.

Kouranov A, Chen X, Fuks B, Schnell DJ. 1998. Tic20 and Tic22 Are New Components of the Protein Import Apparatus at the Chloroplast Inner Envelope Membrane. J Cell Biol. 143(4):991–1002. doi:10.1083/jcb.143.4.991.

Kowallik K V., Stoebe B, Schaffran I, Kroth-Pancic P, Freier U. 1995. The chloroplast genome of a chlorophylla+c-containing alga,Odontella sinensis. Plant Mol Biol Report. 13(4):336–342. doi:10.1007/BF02669188.

Koziol AG, Durnford DG. 2008. Euglena Light-Harvesting Complexes Are Encoded by Multifarious Polyprotein mRNAs that Evolve in Concert. Mol Biol Evol. 25(1):92–100. doi:10.1093/molbev/msm232.

Krajčovič J, Vesteg M, Schwartzbach SD. 2015. Euglenoid flagellates: A multifaceted biotechnology platform. J Biotechnol. 202:135–145. doi:10.1016/j.jbiotec.2014.11.035.

Krinsky NI, Goldsmith TH. 1960. The carotenoids of the flagellated alga, Euglena gracilis. Arch Biochem Biophys. 91(12):271–279. doi:10.1016/0003-9861(60)90501-4.

Kroth PG, Chiovitti A, Gruber A, Martin-Jezequel V, Mock T, Parker MS, Stanley MS, Kaplan A, Caron L, Weber T, et al. 2008. A Model for Carbohydrate Metabolism in the Diatom Phaeodactylum tricornutum Deduced from Comparative Whole Genome Analysis. Kroymann J, editor. PLoS One. 3(1):e1426. doi:10.1371/journal.pone.0001426.

Kruger NJ, von Schaewen A. 2003. The oxidative pentose phosphate pathway: structure and organisation. Curr Opin Plant Biol. 6(3):236–246. doi:10.1016/S1369-5266(03)00039-6.

Kustka AB, Milligan AJ, Zheng H, New AM, Gates C, Bidle KD, Reinfelder JR. 2014. Low CO_2 results in a rearrangement of carbon metabolism to support C_4 photosynthetic carbon assimilation in *Thalassiosira pseudonana*. New Phytol. 204(3):507–520. doi:10.1111/nph.12926.

Larkum AWD, Lockhart PJ, Howe CJ. 2007. Shopping for plastids. Trends Plant Sci. 12(5):189–195. doi:10.1016/j.tplants.2007.03.011.

Latowski D, Kuczyńska P, Strzałka K. 2011. Xanthophyll cycle – a mechanism protecting plants against oxidative stress. Redox Rep. 16(2):78–90. doi:10.1179/174329211X13020951739938.

Lau JB, Stork S, Moog D, Schulz J, Maier UG. 2016. Protein-protein interactions indicate composition of a 480 kDa SELMA complex in the second outermost membrane of diatom complex plastids. Mol Microbiol. 100(1):76–89. doi:10.1111/mmi.13302.

Lau JB, Stork S, Moog D, Sommer MS, Maier UG. 2015. N-terminal lysines are essential for protein translocation via a modified ERAD system in complex plastids. Mol Microbiol. 96(3):609–620. doi:10.1111/mmi.12959.

Leander BS. 2004. Did trypanosomatid parasites have photosynthetic ancestors? Trends Microbiol. 12(6):251–258. doi:10.1016/j.tim.2004.04.001.

Leander BS, Esson HJ, Breglia SA. 2007. Macroevolution of complex cytoskeletal systems in euglenids. Bioessays. 29(10):987–1000. doi:10.1002/bies.20645.

Leander BS, Triemer RE, Farmer M a. 2001. Character evolution in heterotrophic euglenids. Eur J Protistol. 37:337–356.

Lee DW, Jung C, Hwang I. 2013. Cytosolic events involved in chloroplast protein targeting. Biochim Biophys Acta - Mol Cell Res. 1833(2):245–252. doi:10.1016/j.bbamcr.2012.03.006.

Lee J, Lei Z, Watson BS, Sumner LW. 2013. Sub-cellular proteomics of Medicago truncatula. Front Plant Sci. 4:112. doi:10.3389/fpls.2013.00112.

Lee Sookjin, Lee DW, Lee Y, Mayer U, Stierhof Y-D, Lee Sumin, Jürgens G, Hwang I. 2009. Heat Shock Protein Cognate 70-4 and an E3 Ubiquitin Ligase, CHIP, Mediate Plastid-Destined Precursor Degradation through the Ubiquitin-26S Proteasome System in *Arabidopsis*. Plant Cell. 21(12):3984–4001. doi:10.1105/tpc.109.071548.

Leedale GF. 1967. Euglenoid Flagellates. Prentice-Hall.

Levering J, Broddrick J, Dupont CL, Peers G, Beeri K, Mayers J, Gallina AA, Allen AE, Palsson BO, Zengler K. 2016. Genome-Scale Model Reveals Metabolic Basis of Biomass Partitioning in a Model Diatom. Ianora A, editor. PLoS One. 11(5):e0155038. doi:10.1371/journal.pone.0155038.

Li H min, Teng YS. 2013. Transit peptide design and plastid import regulation. Trends Plant Sci. 18(7):360–366. doi:10.1016/j.tplants.2013.04.003.

Lin Y-P, Wu M-C, Charng Y-Y. 2016. Identification of a Chlorophyll Dephytylase Involved in Chlorophyll Turnover in Arabidopsis. Plant Cell. 28(12):2974–2990. doi:10.1105/tpc.16.00478.

Lu Y. 2018. Assembly and Transfer of Iron–Sulfur Clusters in the Plastid. Front Plant Sci. 9:336. doi:10.3389/fpls.2018.00336.

Maeda H, Dudareva N. 2012. The Shikimate Pathway and Aromatic Amino Acid Biosynthesis in Plants. Annu Rev Plant Biol. 63(1):73–105. doi:10.1146/annurev-arplant-042811-105439.

Mahapatra DM, Chanakya HN, Ramachandra T V. 2013. Euglena sp. as a suitable source of lipids for potential use as biofuel and sustainable wastewater treatment. J Appl Phycol. 25(3):855–865. doi:10.1007/s10811-013-9979-5.

Makarov A. 2000. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. doi:10.1021/AC991131P.

Marin B, Nowack ECM, Melkonian M. 2005. A plastid in the making: evidence for a second primary endosymbiosis. Protist. 156(4):425–32. doi:10.1016/j.protis.2005.09.001.

Marin B, Palm A, Klingberg M, Melkonian M. 2003. 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(1):99–145. doi:10.1078/143446103764928521.

Markunas CM, Triemer RE. 2016. Evolutionary History of the Enzymes Involved in the Calvin-Benson Cycle in Euglenids. J Eukaryot Microbiol. 63(3):326–339. doi:10.1111/jeu.12282.

Maruyama S, Suzaki T, Weber APM, Archibald JM, Nozaki H. 2011. Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. BMC Evol Biol. 11(1):105. doi:10.1186/1471-2148-11-105.

Matsuo E, Inagaki Y. 2018. Patterns in evolutionary origins of heme, chlorophyll a and isopentenyl diphosphate biosynthetic pathways suggest non-photosynthetic periods prior to plastid replacements in dinoflagellates. PeerJ. 6:e5345. doi:10.7717/peerj.5345.

Maul JE, Lilly JW, Cui L, dePamphilis CW, Miller W, Harris EH, Stern DB. 2002. The Chlamydomonas reinhardtii Plastid Chromosome. Plant Cell. 14(11):2659–2679. doi:10.1105/TPC.006155.

May T, Soll J. 2000. 14-3-3 Proteins Form a Guidance Complex with Chloroplast Precursor Proteins in Plants. Plant Cell. 12(1):53. doi:10.2307/3871029.

McFadden GI, van Dooren GG. 2004. Evolution: red algal genome affirms a common origin of all plastids. Curr Biol. 14(13):R514-6. doi:10.1016/j.cub.2004.06.041.

McLafferty FW, Breuker K, Jin M, Han X, Infusini G, Jiang H, Kong X, Begley TP. 2007. Top-down MS, a powerful complement to the high capabilities of proteolysis proteomics. FEBS J. 274(24):6256–6268. doi:10.1111/j.1742-4658.2007.06147.x.

Meusser B, Hirsch C, Jarosch E, Sommer T. 2005. ERAD: the long road to destruction. Nat Cell Biol. 7(8):766–72. doi:10.1038/ncb0805-766.

Milanowski R, Gumińska N, Karnkowska A, Ishikawa T, Zakryś B. 2016. Intermediate introns in nuclear genes of euglenids – are they a distinct type? BMC Evol Biol. 16(1):49. doi:10.1186/s12862-016-0620-5.

Miras S, Salvi D, Ferro M, Grunwald D, Garin J, Joyard J, Rolland N. 2002. Non-canonical transit peptide for import into the chloroplast. J Biol Chem. 277(49):47770–8. doi:10.1074/jbc.M207477200.

Molina J, Hazzouri KM, Nickrent D, Geisler M, Meyer RS, Pentony MM, Flowers JM, Pelser P, Barcelona J, Inovejas SA, et al. 2014. Possible loss of the chloroplast genome in the parasitic flowering plant Rafflesia lagascae (Rafflesiaceae). Mol Biol Evol. 31(4):793–803. doi:10.1093/molbev/msu051.

Monfils AK, Triemer RE, Bellairs EF. 2011. Characterization of paramylon morphological diversity in photosynthetic euglenoids (Euglenales, Euglenophyta). Phycologia. 50(2):156–169. doi:10.2216/09-112.1.

Moore CE, Archibald JM. 2009. Nucleomorph genomes. Annu Rev Genet. 43:251–64.

doi:10.1146/annurev-genet-102108-134809.

Morse D, Salois P, Markovic P, Hastings J. 1995. A nuclear-encoded form II RuBisCO in dinoflagellates. Science (80-). 268(5217):1622–1624. doi:10.1126/science.7777861.

Moustafa A, Reyes-Prieto A, Bhattacharya D. 2008. Chlamydiae Has Contributed at Least 55 Genes to Plantae with Predominantly Plastid Functions. DeSalle R, editor. PLoS One. 3(5):e2205. doi:10.1371/journal.pone.0002205.

Nada A, Soll J. 2004. Inner envelope protein 32 is imported into chloroplasts by a novel pathway. J Cell Sci. 117(Pt 17):3975–82. doi:10.1242/jcs.01265.

Nakai M. 2018. New Perspectives on Chloroplast Protein Import. Plant Cell Physiol. 59(6):1111–1119. doi:10.1093/pcp/pcy083.

Nakashima A, Suzuki K, Asayama Y, Konno M, Saito K, Yamazaki N, Takimoto H. 2017. Oral administration of Euglena gracilis Z and its carbohydrate storage substance provides survival protection against influenza virus infection in mice. Biochem Biophys Res Commun. 494(1–2):379–383. doi:10.1016/J.BBRC.2017.09.167.

Nakazawa M, Andoh H, Koyama K, Watanabe Y, Nakai T, Ueda M, Sakamoto T, Inui H, Nakano Y, Miyatake K. 2015. Alteration of wax ester content and composition in euglena gracilis with gene silencing of 3-ketoacyl-coa thiolase isozymes. Lipids. 50(5):483–492. doi:10.1007/s11745-015-4010-3.

Nasir A, Le Bail A, Daiker V, Klima J, Richter P, Lebert M. 2018. Identification of a flagellar protein implicated in the gravitaxis in the flagellate Euglena gracilis. Sci Rep. 8(1):7605. doi:10.1038/s41598-018-26046-8.

Nassoury N, Cappadocia M, Morse D. 2003. Plastid ultrastructure defines the protein import pathway in dinoflagellates. J Cell Sci. 116(Pt 14):2867–74. doi:10.1242/jcs.00517.

Naumann J, Salomo K, Der JP, Wafula EK, Bolin JF, Maass E, Frenzke L, Samain M-S, Neinhuis C, dePamphilis CW, et al. 2013. Single-Copy Nuclear Genes Place Haustorial Hydnoraceae within Piperales and Reveal a Cretaceous Origin of Multiple Parasitic Angiosperm Lineages. Corradi N, editor. PLoS One. 8(11):e79204. doi:10.1371/journal.pone.0079204.

Neilson KA, Ali NA, Muralidharan S, Mirzaei M, Mariani M, Assadourian G, Lee A, van Sluyter SC, Haynes PA. 2011. Less label, more free: Approaches in label-free quantitative mass spectrometry. Proteomics. 11(4):535–553. doi:10.1002/pmic.201000553.

Nevo R, Charuvi D, Tsabari O, Reich Z. 2012. Composition, architecture and dynamics of the photosynthetic apparatus in higher plants. Plant J. 70(1):157–176. doi:10.1111/j.1365-313X.2011.04876.x.

Novák Vanclová AMG, Zoltner M, Kelly S, Soukal P, Záhonová K, Füssy Z, Ebenezer TE, Lacová Dobáková E, Eliáš M, Lukeš J, et al. 2019. Metabolic quirks and the colourful history of the Euglena gracilis secondary plastid. Under review.

Nowack ECM, Melkonian M, Glöckner G. 2008. Chromatophore Genome Sequence of Paulinella Sheds Light on Acquisition of Photosynthesis by Eukaryotes. Curr Biol. 18(6):410–418. doi:10.1016/j.cub.2008.02.051.

Nowack ECM, Price DC, Bhattacharya D, Singer A, Melkonian M, Grossman AR. 2016. Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of Paulinella chromatophora. Proc Natl Acad Sci U S A. 113(43):12214–12219. doi:10.1073/pnas.1608016113.

O'Neill EC, Trick M, Hill L, Rejzek M, Dusi RG, Hamilton CJ, Zimba P V., Henrissat B, Field RA. 2015. The transcriptome of Euglena gracilis reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. Mol Biosyst. 11(10):2808–2820. doi:10.1039/C5MB00319A.

Obata T, Schoenefeld S, Krahnert I, Bergmann S, Scheffel A, Fernie A. 2013. Gas-Chromatography Mass-Spectrometry (GC-MS) Based Metabolite Profiling Reveals Mannitol as a Major Storage Carbohydrate in the Coccolithophorid Alga Emiliania huxleyi. Metabolites. 3(1):168–184. doi:10.3390/metabo3010168.

Oborník M, Lukeš J. 2015. The Organellar Genomes of Chromera and Vitrella, the Phototrophic Relatives of Apicomplexan Parasites. Annu Rev Microbiol. 69(1):129–144. doi:10.1146/annurev-micro-091014-104449.

Ogawa T, Tamoi M, Kimura A, Mine A, Sakuyama H, Yoshida E, Maruta T, Suzuki K, Ishikawa T, Shigeoka S. 2015. Enhancement of photosynthetic capacity in Euglena gracilis by expression of cyanobacterial fructose-1,6-/sedoheptulose-1,7-bisphosphatase leads to increases in biomass and wax ester production. Biotechnol Biofuels. 8(1):80. doi:10.1186/s13068-015-0264-5.

Ogren WL. 1984. Photorespiration: Pathways, Regulation, and Modification. Annu Rev Plant Physiol. 35(1):415–442. doi:10.1146/annurev.pp.35.060184.002215.

Ohyama T, Kumazawa K. 1980. Nitrogen assimilation in soybean nodules. Soil Sci Plant Nutr. 26(1):109–115. doi:10.1080/00380768.1980.10433217.

Old WM, Meyer-Arendt K, Aveline-Wolf L, Pierce KG, Mendoza A, Sevinsky JR, Resing KA, Ahn NG. 2005. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. Mol Cell Proteomics. 4(10):1487–502. doi:10.1074/mcp.M500084-MCP200.

Ong S-E, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M. 2002. Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. Mol Cell Proteomics. 1(5):376–86. doi:10.1074/mcp.m200025-mcp200.

Oudot-Le Secq M-P, Grimwood J, Shapiro H, Armbrust EV, Bowler C, Green BR. 2007. Chloroplast genomes of the diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana: comparison with other plastid genomes of the red lineage. Mol Genet Genomics. 277(4):427–439. doi:10.1007/s00438-006-0199-4.

Paila YD, Richardson LGL, Schnell DJ. 2015. New Insights into the Mechanism of Chloroplast Protein Import and Its Integration with Protein Quality Control, Organelle Biogenesis and Development. J Mol Biol. 427(5):1038–1060. doi:10.1016/J.JMB.2014.08.016.

Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci. 108(33):13624–13629. doi:10.1073/pnas.1110633108.

Patron NJ, Waller RF, Archibald JM, Keeling PJ. 2005. Complex protein targeting to dinoflagellate plastids. J Mol Biol. 348(4):1015–24. doi:10.1016/j.jmb.2005.03.030.

Peers G, Price NM. 2006. Copper-containing plastocyanin used for electron transport by an oceanic diatom. Nature. 441(7091):341–344. doi:10.1038/nature04630.

Perkins DN, Pappin DJC, Creasy DM, Cottrell JS. 1999. Probability-based protein

identification by searching sequence databases using mass spectrometry data. Electrophoresis. 20(18):3551–3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2.

Pilon M, Abdel-Ghany SE, Hoewyk D, Ye H, Pilon-Smits EAH. 2006. Biogenesis of Iron-Sulfur Cluster Proteins in Plastids. In: Genetic Engineering. Boston: Kluwer Academic Publishers. p. 101–117.

Ponce-Toledo RI, Deschamps P, López-García P, Zivanovic Y, Benzerara K, Moreira D. 2017. An Early-Branching Freshwater Cyanobacterium at the Origin of Plastids. Curr Biol. 27(3):386–391. doi:10.1016/J.CUB.2016.11.056.

Ponce-Toledo RI, Moreira D, López-García P, Deschamps P, Ruiz-Trillo I. 2018 Jun 19. Secondary Plastids of Euglenids and Chlorarachniophytes Function with a Mix of Genes of Red and Green Algal Ancestry. Ruiz-Trillo I, editor. Mol Biol Evol. doi:10.1093/molbev/msy121.

Ponce-Toledo RI, López-García P, Moreira D. 2019 May 28. Horizontal and endosymbiotic gene transfer in early plastid evolution. New Phytol.:nph.15965. doi:10.1111/nph.15965.

Puerta MVS, Bachvaroff TR, Delwiche CF. 2005. The Complete Plastid Genome Sequence of the Haptophyte Emiliania huxleyi: a Comparison to Other Plastid Genomes. DNA Res. 12(2):151–156. doi:10.1093/dnares/12.2.151.

Raines CA. 2003. The Calvin cycle revisited. Photosynth Res. 75(1):1–10. doi:10.1023/A:1022421515027.

Raven J a. 2003. Carboxysomes and peptidoglycan walls of cyanelles: possible physiological functions. Eur J Phycol. 38(1):47–53. doi:10.1080/0967026031000096245.

Rippert P, Puyaubert J, Grisollet D, Derrier L, Matringe M. 2009. Tyrosine and Phenylalanine Are Synthesized within the Plastids in Arabidopsis. PLANT Physiol. 149(3):1251–1260. doi:10.1104/pp.108.130070.

Robbens S, Derelle E, Ferraz C, Wuyts J, Moreau H, Van de Peer Y. 2007. The Complete Chloroplast and Mitochondrial DNA Sequence of Ostreococcus tauri: Organelle Genomes of the Smallest Eukaryote Are Examples of Compaction. Mol Biol Evol. 24(4):956–968. doi:10.1093/molbev/msm012.

Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. 2007. The complete chloroplast genome of the chlorarachniophyte Bigelowiella natans: Evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. Mol Biol Evol. 24(1):54–62. doi:10.1093/molbev/msl129.

Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, et al. 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics. 3(12):1154–69. doi:10.1074/mcp.M400129-MCP200.

Rost B, Riebesell U, Burkhardt S, Sültemeyer D. 2003. Carbon acquisition of bloom-forming marine phytoplankton. Limnol Oceanogr. 48(1):55–67. doi:10.4319/lo.2003.48.1.0055.

Russo R, Barsanti L, Evangelista V, Frassanito AM, Longo V, Pucci L, Penno G, Gualtieri P. 2017. *Euglena gracilis* paramylon activates human lymphocytes by upregulating pro-inflammatory factors. Food Sci Nutr. 5(2):205–214. doi:10.1002/fsn3.383.

Šantek B, Felski M, Friehs K, Lotz M, Flaschel E. 2009. Production of paramylon, a β-1,3-glucan, by heterotrophic cultivation of *Euglena gracilis* on a synthetic medium. Eng Life Sci.

9(1):23–28. doi:10.1002/elsc.200700032.

Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S. 1999. Complete Structure of the Chloroplast Genome of Arabidopsis thaliana. DNA Res. 6(5):283–290. doi:10.1093/dnares/6.5.283.

Schelkunov MI, Shtratnikova VY, Nuraliev MS, Selosse M-A, Penin AA, Logacheva MD. 2015. Exploring the limits for reduction of plastid genomes: a case study of the mycoheterotrophic orchids Epipogium aphyllum and Epipogium roseum. Genome Biol Evol. 7(4):1179–91. doi:10.1093/gbe/evv019.

Schubert M, Petersson UA, Haas BJ, Funk C, Schröder WP, Kieselbach T. 2002. Proteome map of the chloroplast lumen of Arabidopsis thaliana. J Biol Chem. 277(10):8354–8365. doi:10.1074/jbc.M108575200.

Schwenkert S, Dittmer S, Soll J. 2018. Structural components involved in plastid protein import. Essays Biochem. 62(1):65–75. doi:10.1042/EBC20170093.

Sheiner L, Striepen B. 2013. Protein sorting in complex plastids. Biochim Biophys Acta - Mol Cell Res. 1833(2):352–359. doi:10.1016/j.bbamcr.2012.05.030.

Sheveleva E V., Hallick RB. 2004. Recent horizontal intron transfer to a chloroplast genome. Nucleic Acids Res. 32(2):803–810. doi:10.1093/nar/gkh225.

Shi L-X, Theg SM. 2013. The chloroplast protein import system: from algae to trees. Biochim Biophys Acta. 1833(2):314–31. doi:10.1016/j.bbamcr.2012.10.002.

Siddique MA, Grossmann J, Gruissem W, Baginsky S. 2006. Proteome Analysis of Bell Pepper (Capsicum annuum L.) Chromoplasts. Plant Cell Physiol. 47(12):1663–1673. doi:10.1093/pcp/pcl033.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31(19):3210–3212. doi:10.1093/bioinformatics/btv351.

Simpson AGB. 1997. The identity and composition of the Euglenozoa. Arch für Protistenkd. 148(3):318–328. doi:10.1016/S0003-9365(97)80012-7.

Sláviková S, Vacula R, Fang Z, Ehara T, Osafune T, Schwartzbach SD. 2005. Homologous and heterologous reconstitution of Golgi to chloroplast transport and protein import into the complex chloroplasts of Euglena. J Cell Sci. 118(Pt 8):1651–1661. doi:10.1242/jcs.02277.

Smith DR, Lee RW. 2014. A plastid without a genome: evidence from the nonphotosynthetic green algal genus Polytomella. Plant Physiol. 164(4):1812–9. doi:10.1104/pp.113.233718.

Sommer MS, Daum B, Gross LE, Weis BLM, Mirus O, Abram L, Maier U-G, Kühlbrandt W, Schleiff E. 2011. Chloroplast Omp85 proteins change orientation during evolution. Proc Natl Acad Sci U S A. 108(33):13841–6. doi:10.1073/pnas.1108626108.

Sommer MS, Gould SB, Lehmann P, Gruber A, Przyborski JM, Maier U-G. 2007. Der1-mediated Preprotein Import into the Periplastid Compartment of Chromalveolates? Mol Biol Evol. 24(4):918–928. doi:10.1093/molbev/msm008.

Stensballe A, Hald S, Bauw G, Blennow A, Welinder KG. 2008. The amyloplast proteome of potato tuber. FEBS J. 275(8):1723–1741. doi:10.1111/j.1742-4658.2008.06332.x.

Stiller JW, Schreiber J, Yue J, Guo H, Ding Q, Huang J. 2014. The evolution of photosynthesis in chromist algae through serial endosymbioses. Nat Commun. 5(1):5764. doi:10.1038/ncomms6764.

Stirewalt VL, Michalowski CB, Löffelhardt W, Bohnert HJ, Bryant DA. 1995. Nucleotide sequence of the cyanelle genome fromCyanophora paradoxa. Plant Mol Biol Report. 13(4):327–332. doi:10.1007/BF02669186.

Stork S, Moog D, Przyborski JM, Wilhelmi I, Zauner S, Maier UG. 2012. Distribution of the SELMA Translocon in Secondary Plastids of Red Algal Origin and Predicted Uncoupling of Ubiquitin-Dependent Translocation from Degradation. Eukaryot Cell. 11(12):1472–1481. doi:10.1128/EC.00183-12.

Sugiyama A, Suzuki K, Mitra S, Arashida R, Yoshida E, Nakano R, Yabuta Y, Takeuchi T. 2009. Hepatoprotective effects of paramylon, a beta-1, 3-D-glucan isolated from Euglena gracilis Z, on acute liver injury induced by carbon tetrachloride in rats. J Vet Med Sci Japanese Soc Vet Sci. 71(7):885–890. doi:10.1292/jvms.71.885.

Sulli C, Fang ZW, Muchhal U, Schwartzbach SD. 1999. Topology of Euglena chloroplast protein precursors within endoplasmic reticulum to Golgi to chloroplast transport vesicles. J Biol Chem. 274(1):457–463. doi:10.1074/jbc.274.1.457.

Sulli C, Schwartzbach SD. 1996. A Soluble Protein Is lmported into Euglena Chloroplasts as a Membrane-Bound Precursor. Plant Cell. 8(1):43–53.

Sun Q, Zybailov B, Majeran W, Friso G, Olinares PDB, van Wijk KJ. 2009. PPDB, the Plant Proteomics Database at Cornell. Nucleic Acids Res. 37(suppl_1):D969–D974. doi:10.1093/nar/gkn654.

Suzuki S, Hirakawa Y, Kofuji R, Sugita M, Ishida K. 2016. Plastid genome sequences of Gymnochlora stellata, Lotharella vacuolata, and Partenskyella glossopodia reveal remarkable structural conservation among chlorarachniophyte species. J Plant Res. 129(4):581–590. doi:10.1007/s10265-016-0804-5.

Takahashi F, Okabe Y, Nakada T, Sekimoto H, Ito M, Kataoka H, Nozaki H. 2007. Origins of the secondary plastids of Euglenophyta and Chlorarachniophyta as revealed by an analysis of the plastid-targeting, nuclear-encoded gene psbO 1. J Phycol. 43(6):1302–1309. doi:10.1111/j.1529-8817.2007.00411.x.

Tanaka R, Tanaka A. 2007. Tetrapyrrole Biosynthesis in Higher Plants. Annu Rev Plant Biol. 58(1):321–346. doi:10.1146/annurev.arplant.57.032905.105448.

Tanaka Y, Ogawa T, Maruta T, Yoshida Y, Arakawa K, Ishikawa T. 2017. Glucan synthase-like 2 is indispensable for paramylon synthesis in Euglena gracilis. FEBS Lett. 591(10):1360–1370. doi:10.1002/1873-3468.12659.

Tanifuji G, Onodera NT, Brown MW, Curtis BA, Roger AJ, Ka-Shu Wong G, Melkonian M, Archibald JM. 2014. Nucleomorph and plastid genome sequences of the chlorarachniophyte Lotharella oceanica: convergent reductive evolution and frequent recombination in nucleomorph-bearing algae. BMC Genomics. 15(1):374. doi:10.1186/1471-2164-15-374.

Teerawanichpan P, Qiu X. 2010. Fatty Acyl-CoA Reductase and Wax Synthase from Euglena gracilis in the Biosynthesis of Medium-Chain Wax Esters. Lipids. 45(3):263–273. doi:10.1007/s11745-010-3395-2.

Terashima M, Specht M, Hippler M. 2011. The chloroplast proteome: a survey from the Chlamydomonas reinhardtii perspective with a focus on distinctive features. Curr Genet. 57(3):151–68. doi:10.1007/s00294-011-0339-1.

Tessier LH, Keller M, Chan RL, Fournier R, Weil JH, Imbault P. 1991. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in

Euglena. EMBO J. 10(9):2621–5.

Tessier LH, Paulus F, Keller M, Vial C, Imbault P. 1995. Structure and expression of Euglena gracilis nuclear rbcS genes encoding the small subunits of the ribulose 1,5-bisphoshate carboxylase/oxygenase: A novel splicing process for unusual intervening sequences? J Mol Biol. 245(1):22–33. doi:10.1016/S0022-2836(95)80035-2.

Tetlow IJ, Rawsthorne S, Raines C, Emes MJ. 2018. Plastid Metabolic Pathways. In: Annual Plant Reviews online. Chichester, UK: John Wiley & Sons, Ltd. p. 60–125.

Tonkin CJ, Struck NS, Mullin K a, Stimmler LM, McFadden GI. 2006. Evidence for Golgi-independent transport from the early secretory pathway to the plastid in malaria parasites. Mol Microbiol. 61(3):614–30. doi:10.1111/j.1365-2958.2006.05244.x.

Triemer RE, Farmer MA. 1991. An ultrastructural comparison of the mitotic apparatus, feeding apparatus, flagellar apparatus and cytoskeleton in euglenoids and kinetoplastids. Protoplasma. 164(1–3):91–104. doi:10.1007/BF01320817.

Trösch R, Jarvis P. 2011. The Stromal Processing Peptidase of Chloroplasts is Essential in Arabidopsis, with Knockout Mutations Causing Embryo Arrest after the 16-Cell Stage. Tsiantis M, editor. PLoS One. 6(8):e23039. doi:10.1371/journal.pone.0023039.

Tsai J-Y, Chu C-C, Yeh Y-H, Chen L-J, Li H-M, Hsiao C-D. 2013. Structural characterizations of the chloroplast translocon protein Tic110. Plant J. 75(5):847–57. doi:10.1111/tpj.12249.

Tsuji Y, Suzuki I, Shiraiwa Y. 2009. Photosynthetic Carbon Assimilation in the Coccolithophorid Emiliania huxleyi (Haptophyta): Evidence for the Predominant Operation of the C3 Cycle and the Contribution of β-Carboxylases to the Active Anaplerotic Reaction. Plant Cell Physiol. 50(2):318–329. doi:10.1093/pcp/pcn200.

Tsuji Y, Yamazaki M, Suzuki I, Shiraiwa Y. 2015. Quantitative Analysis of Carbon Flow into Photosynthetic Products Functioning as Carbon Storage in the Marine Coccolithophore, Emiliania huxleyi. Mar Biotechnol. 17(4):428–440. doi:10.1007/s10126-015-9632-1.

Tsuji Y, Yoshida M. 2017. Biology of Haptophytes: Complicated Cellular Processes Driving the Global Carbon Cycle. 1st ed. Elsevier Ltd.

Turmel M, Gagnon M-C, O'Kelly CJ, Otis C, Lemieux C. 2009. 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(3):631–48. doi:10.1093/molbev/msn285.

Vanclová AMG, Hadariová L, Hrdá Š, Hampl V. 2017. Secondary Plastids of Euglenophytes. In: Hirakawa Y, editor. Advances in Botanical Research. 1st ed. Academic Press. p. 321–358.

de Vries J, Archibald JM. 2017. Endosymbiosis: Did Plastids Evolve from a Freshwater Cyanobacterium? Curr Biol. 27(3):R103–R105. doi:10.1016/J.CUB.2016.12.006.

de Vries J, Sousa FL, Bölter B, Soll J, Gould SB. 2015. YCF1: A Green TIC? Plant Cell. 27(7):1827–33. doi:10.1105/tpc.114.135541.

Waller RF, McFadden GI. 2005. The apicoplast: a review of the derived plastid of apicomplexan parasites. Curr Issues Mol Biol. 7(1):57–79.

Walther TC, Mann M. 2010. Mass spectrometry-based proteomics in cell biology. J Cell Biol. 190(4):491–500. doi:10.1083/jcb.201004052.

Wang Z, Benning C. 2012. Chloroplast lipid synthesis and lipid trafficking through ER–

plastid membrane contact sites. Biochem Soc Trans. 40(2):457–463. doi:10.1042/BST20110752.

Watanabe T, Shimada R, Matsuyama A, Yuasa M, Sawamura H, Yoshida E, Suzuki K. 2013. Antitumor activity of the β-glucan paramylon from Euglena against preneoplastic colonic aberrant crypt foci in mice. Food Funct. 4(11):1685. doi:10.1039/c3fo60256g.

Watson BS, Asirvatham VS, Wang L, Sumner LW, Fu A, Rodermel SR. 2003. Mapping the Proteome of Barrel Medic (*Medicago truncatula*). Plant Physiol. 131(3):1104–1123. doi:10.1104/pp.102.019034.

Wedemayer GJ, Kidd DG, Glazer AN. 1996. Cryptomonad biliproteins: Bilin types and locations. Photosynth Res. 48(1–2):163–170. doi:10.1007/BF00041006.

Wetherbee R, Jackson CJ, Repetti SI, Clementson LA, Costa JF, van de Meene A, Crawford S, Verbruggen H. 2018 Dec 9. The golden paradox – a new heterokont lineage with chloroplasts surrounded by two membranes. J Phycol.:jpy.12822. doi:10.1111/jpy.12822.

Von Wettstein D, Gough S, Kannangara CG. 1995. Chlorophyll Biosynthesis. American Society of Plant Physiologists.

Wiegert KE, Bennett MS, Triemer RE. 2012. Evolution of the chloroplast genome in photosynthetic euglenoids: a comparison of Eutreptia viridis and Euglena gracilis (Euglenophyta). Protist. 163(6):832–43. doi:10.1016/j.protis.2012.01.002.

Wienkoop S, Baginsky S, Weckwerth W. 2010. Arabidopsis thaliana as a model organism for plant proteome research. J Proteomics. 73(11):2239–2248. doi:10.1016/j.jprot.2010.07.012.

van Wijk KJ. 2004. Plastid proteomics. Plant Physiol Biochem. 42(12):963–977. doi:10.1016/J.PLAPHY.2004.10.015.

Wilson RJ, Denny PW, Preiser PR, Rangachari K, Roberts K, Roy A, Whyte A, Strath M, Moore DJ, Moore PW, et al. 1996. Complete Gene Map of the Plastid-like DNA of the Malaria ParasitePlasmodium falciparum. J Mol Biol. 261(2):155–172. doi:10.1006/jmbi.1996.0449.

Yamaguchi A, Yubuki N, Leander BS. 2012. Morphostasis in a novel eukaryote illuminates the evolutionary transition from phagotrophy to phototrophy: description of Rapaza viridis n. gen. et sp. (Euglenozoa, Euglenida). BMC Evol Biol. 12(1):29. doi:10.1186/1471-2148-12-29.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol. 21(5):809–18. doi:10.1093/molbev/msh075.

Yoon HS, Nakayama T, Reyes-Prieto A, Andersen RA, Boo SM, Ishida K-I, Bhattacharya D. 2009. A single origin of the photosynthetic organelle in different Paulinella lineages. BMC Evol Biol. 9(1):98. doi:10.1186/1471-2148-9-98.

Yoshida Y, Tomiyama T, Maruta T, Tomita M, Ishikawa T, Arakawa K. 2016. De novo assembly and comparative transcriptome analysis of Euglena gracilis in response to anaerobic conditions. BMC Genomics. 17(1):182. doi:10.1186/s12864-016-2540-6.

Záhonová K, Füssy Z, Birčák E, Novák Vanclová AMG, Klimeš V, Vesteg M, Krajčovič J, Oborník M, Eliáš M. 2018. Peculiar features of the plastids of the colourless alga Euglena longa and photosynthetic euglenophytes unveiled by transcriptome analyses. Sci Rep. 8(1):17012. doi:10.1038/s41598-018-35389-1.

Záhonová K, Füssy Z, Oborník M, Eliáš M, Yurchenko V. 2016. RuBisCO in non-photosynthetic alga Euglena longa: Divergent features, transcriptomic analysis and regulation of complex formation. PLoS One. 11(7):1–15. doi:10.1371/journal.pone.0158790.

Zakhartsev M, Medvedeva I, Orlov Y, Akberdin I, Krebs O, Schulze WX. 2016. Metabolic model of central carbon and energy metabolisms of growing Arabidopsis thaliana in relation to sucrose translocation. BMC Plant Biol. 16(1):262. doi:10.1186/s12870-016-0868-3.

Zakryś B, Milanowski R, Karnkowska A. 2017. Evolutionary Origin of Euglena. Springer International Publishing. p. 3–17.

Zauner S, Greilinger D, Laatsch T, Kowallik K V., Maier U-G. 2004. Substitutional editing of transcripts from genes of cyanobacterial origin in the dinoflagellate *Ceratium horridum*. FEBS Lett. 577(3):535–538. doi:10.1016/j.febslet.2004.10.060.

Zeeman SC, Kossmann J, Smith AM. 2010. Starch: Its Metabolism, Evolution, and Biotechnological Modification in Plants. Annu Rev Plant Biol. 61(1):209–234. doi:10.1146/annurev-arplant-042809-112301.

Zhang L, Wang X, Liu T, Wang H, Wang G, Chi S, Liu C. 2015. Complete Plastid Genome of the Brown Alga Costaria costata (Laminariales, Phaeophyceae). Escriva H, editor. PLoS One. 10(10):e0140144. doi:10.1371/journal.pone.0140144.

Zhang Z, Green BR, Cavalier-Smith T. 1999. Single gene circles in dinoflagellate chloroplast genomes. Nature. 400(6740):155–159. doi:10.1038/22099.

Zufferey M, Montandon C, Douet V, Demarsy E, Agne B, Baginsky S, Kessler F. 2017. The novel chloroplast outer membrane kinase KOC1 is a required component of the plastid protein import machinery. J Biol Chem. 292(17):6952–6964. doi:10.1074/jbc.M117.776468.

CHAPTER NINE

Secondary Plastids of Euglenophytes

Anna M.G. Vanclová[1], Lucia Hadariová, Štěpánka Hrdá, Vladimír Hampl[1]

Faculty of Science, Charles University in Prague, Prague, Czech Republic
[1]Corresponding authors: e-mail address: vanclova@gmail.com; vlada@natur.cuni.cz

Contents

Abstract

Euglenophytes obtained their plastids from a primary green alga related to extant genus *Pyramimonas*. The relatively recent establishment of this new organelle is an intriguing evolutionary phenomenon worth studying and comparing with other secondary plastids with a regard to their similarities and differences. A remarkably fast evolution driven by rapid intron gain and diversification is observed in euglenid plastid genomes which often tend to swell in size and rearrange while keeping the gene content stable. As a result of the secondary endosymbiosis, the plastid is wrapped in an additional membrane which makes any protein, metabolite, or ion transporting routes more complicated. In the case of protein import, secretory pathway-derived, signal peptide-dependent mechanism involving the endoplasmic reticulum, Golgi, and vesicular transport were recruited. The plastid endosymbiosis also served as a source of various oddities concerning metabolic pathways as the new organelle contained some of the enzymes and pathways already present in the host. Thus, several cases of division of labour and specialization can be observed, as well as simple redundancies which might be in fact just transitory and will eventually disappear in the future course of evolution. Endosymbiotic and lateral gene transfers were quite common in the

Advances in Botanical Research, Volume 84
ISSN 0065-2296
http://dx.doi.org/10.1016/bs.abr.2017.06.008

ancestors of euglenophytes, especially in the case of plastid proteins many of which were demonstrated to have originated not only from the green-algal endosymbiont but also from a spectrum of nongreen lineages. The circumstances of the nongreen-algal gene gains are unclear. Another evolutionary phenomenon occurring in euglenophytes is the secondary loss of plastid or its photosynthetic capacity. This process gave rise to a number of distinct species which no longer possess the ability to photosynthesize. Interestingly, this "bleaching" process can be induced in the laboratory, enabling to study the process of plastid loss in vitro.

1. INTRODUCTION: WHAT ARE THE EUGLENOPHYTES AND WHY TO CARE ABOUT THEM

Euglenids are a group of flagellate protists belonging to the phylum Euglenozoa (Excavata), along with kinetoplastids—a group made famous by its important pathogenic members such as trypanosomes—diplonemids, and symbiontids (also termed Postgaardea) (Adl et al., 2012; Cavalier-Smith, 2016). Euglenids are a highly diversified taxon in regard to their nutritional strategies. The ancestral mode of nutrition was probably bacteriovory which was replaced by eukaryovory in a large portion of the diversity; this shift is believed to have taken place once in the evolution. One lineage of eukaryovores then lost the phagotrophic ability and became osmotrophic. This lineage is currently termed primary osmotrophs (Leander, 2004; Leander, Esson, & Breglia, 2007; Leander, Triemer, & Farmer, 2001). However, some researchers propose an alternative hypothesis according to which eukaryovory was the ancestral nutritional strategy common not only to euglenids but also to kinetoplastids, diplonemids, and symbiontids, and all other feeding modes are derived from it (Cavalier-Smith, 2016). One lineage within eukaryovorous euglenids gained a green secondary plastid in presumably one endosymbiotic event. All these photoautotrophs form one robust phylogenetic clade termed the Euglenophyta. The term "euglenophytes" thus refers to phototrophic (or secondarily osmotrophic) euglenids that harbour (or once harboured) a plastid, while the term "euglenids" addresses the whole group of organisms of various nutritional strategies. The euglenophytes were one of the first protists to be discovered and described. The first documented observation of an euglenophyte, presumably the genus *Euglena*, and description of its typical slime-like mode of movement, metaboly, was carried out by John Harris at the end of the 17th century (Harris, 1695) and the first species of euglenophytes were described

in the beginning of the 19th century by Ehrenberg (1830). Several isolated lineages within euglenophytes lost the photosynthetic pigments, switching to osmotrophic mode of nutrition independently of the earlier mentioned primary osmotrophs; these organisms are termed secondary osmotrophs and retain a colourless plastid despite not using it for photosynthesis. Additionally, one obligatory–mixotrophic species, *Rapaza viridis*, has a photosynthetically active green plastid but at the same time requires green-algal (*Tetraselmis*) prey to survive and it is the only euglenid which combines photoautotrophy with phagotrophy. It is suspected to represent a link between the eukaryovorous euglenids and the phototrophic euglenophytes. This is also suggested by its phylogenetic position as sister lineage to all other euglenophytes (Fig. 1; Yamaguchi, Yubuki, & Leander, 2012). However, it is still unclear whether the *R. viridis* plastid is stable or transient and if it is related to the plastids of other euglenophytes. The closest known non-photosynthetic sister of euglenophytes is the eukaryovorous genus *Teloprocta* (former *Heteronema*) *scaphurum* (Cavalier-Smith, 2016; Lax & Simpson, 2013).

Euglenophytes have been intensively studied for a long time thanks to their relatively large size, distinctive and aesthetically pleasing appearance, and easy collection and cultivation, and so approximately 950 species of euglenophytes have been described to this day (Guiry & Guiry, 2017). The inner taxonomy of euglenophytes is currently relatively clearly resolved (Fig. 1; Karnkowska et al., 2015; Kim, Linton, & Shin, 2015; Kim & Shin, 2008; Linton et al., 2010; Triemer et al., 2006; Zakryś, Milanowski, & Karnkowska, 2017). The group splits into two orders: Eutreptiales and Euglenales. The Eutreptiales exhibit a number of features which are considered ancestral, such as the adaptation to marine environment and presence of two emergent flagella, and contain two genera: *Eutreptia* and *Eutreptiella*. The Euglenales are more diverse, they have only one emergent flagellum, and they are thought to be almost exclusively freshwater dwelling. The Eutreptiales are probably paraphyletic, with the Euglenales being their inner group (Cavalier-Smith, 2016). The Euglenales are further split into two distinctive families, Euglenaceae and Phacaceae. The Euglenaceae include eight genera: *Euglena*, *Euglenaria*, *Euglenaformis*, *Cryptoglena*, *Monomorphina*, *Colacium*, *Trachelomonas*, and *Strombomonas*. This group is highly diverse in regard to the shape, number, and position of the plastids, and also overall morphology of the cell which may exhibit various shapes—some members, i.e., genera *Trachelomonas* and *Strombomonas*, are even surrounded

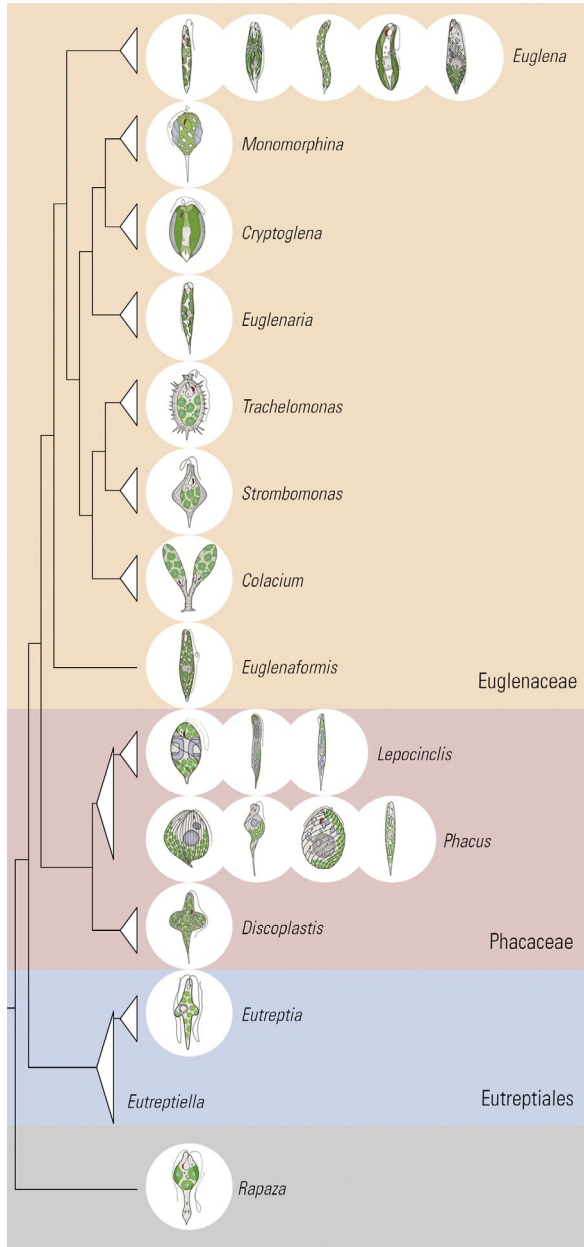

Fig. 1 Simplified phylogenetic tree of the 14 euglenophyte genera (including *Rapaza*) with examples of cell and plastid morphologies. One to five different illustrative species are shown for each genus, depending on its relative intrageneric plastid morphological diversity.

by mineralized extracellular lorica—and types and degrees of motility—some members have pellicles of low flexibility and prefer swimming using flagellum to metaboly, while some species in the genus *Colacium* are completely sessile. The Phacaceae include three genera: *Phacus*, *Lepocinclis*, and *Discoplastis*, the first two of which have a rigid pellicle, which is flattened or even helical in shape, and are not able to move using metaboly (Cavalier-Smith, 2016; de M. Bicudo & Menezes, 2016).

Euglenophytes are currently studied for their possible utilization in biotechnology and other applied sciences, especially as potential producers of biofuels and also nutritional supplements due to their capacity to synthesize various lipids and compounds with antioxidative properties such as tocopherols and carotenoids (Krajčovič, Vesteg, & Schwartzbach, 2015; Ogawa et al., 2015; Teerawanichpan & Qiu, 2010).

2. ORIGIN OF EUGLENOPHYTE PLASTIDS: EARLY OR LATE, GREEN OR RED?

Before the age of molecular phylogenetics, euglenophytes were traditionally classified as a subgroup or sister of green algae due to their plastid morphology and pigment content, especially the combination of chlorophylls *a* and *b*, which is otherwise unique to green plants and algae and chlorarachniophytes, and the chlorophylls *a*- and *b*-containing plastids were correctly assumed to have arisen only once in the evolution. However, it was suspected for a long time that something is amiss here because most of the nonplastid morphology, physiology, and biochemistry of these organisms were strikingly dissimilar to green algae. With the (re)invention of the endosymbiotic theory by Lynn Margulis in 1967 (Sagan, 1993) and its general acceptance, a completely new paradigm opened for evolutionary biology and taxonomy trying to resolve unclear phenomena and phylogenetic relationships such as this one. A decade later, the secondary endosymbiotic origin of the euglenophyte plastid was proposed by Gibbs (1978, 1981) and euglenids were placed along with kinetoplastids into a new phylum Euglenozoa by Cavalier-Smith (1981). The first sequenced genes from the model euglenid *Euglena gracilis* (Douglas & Turner, 1991; Morden & Golden, 1991), followed shortly by the complete sequencing of its plastid genome (Hallick et al., 1993), brought a molecular evidence for this claim. In 2009, a wider analysis of phylogenetic relationship within green algae with a focus on prasinophytes and also including sequences from secondary plastids was performed by Turmel, Gagnon, O'Kelly, Otis, and Lemieux

(2009) which resulted in pinpointing the origin of the euglenid plastid to the Pyramimonadales clade in the prasinophytes with the genus *Pyramimonas* as the most suspect source. Following studies on other euglenid plastid genomes (Bennett & Triemer, 2015; Bennett, Wiegert, & Triemer, 2014; Dabbagh & Preisfeld, 2016; Hrdá, Fousek, Szabová, Hampl, & Vlček, 2012; Pombert, James, Janouškovec, & Keeling, 2012; Wiegert, Bennett, & Triemer, 2012, 2013) confirmed *Pyramimonas parkeae* as the closest living relative of euglenophyte plastids.

The relative position of the plastid acquisition in the euglenid lineage was disputed in the past (Bodył, Mackiewicz, & Milanowski, 2010; Hannaert et al., 2003) and even though it is currently considered resolved and placed at the root of the extant euglenophytes the evolutionary history of euglenids seems more complex and traces of an interesting story about lateral gains and secondary losses of genes—and possibly endosymbionts or even organelles—can be read from the available sequence data of euglenids and their not-so-distant relatives kinetoplastids.

Kinetoplastids, a group known mainly for their parasitic members of considerable epidemiological significance of the genera *Trypanosoma* and *Leishmania*, have been shown to contain genes related to cyanobacteria or green plastids. Moreover, some of these genes are homologues of enzymes of the Calvin cycle which have been supposedly recruited for glycolysis-related functions in the glycosome (Hannaert et al., 2003) (i.e. peroxisome-derived organelle invented specifically by kinetoplastids and diplonemids; Morales et al., 2016; Rybicka, 1996). In reaction to this fact plus other occurrences of cyanobacterial-like genes in heterotrophic protists, a plastid-early hypothesis was formulated. This hypothesis placed the primary plastid acquisition much earlier in the evolution and suggested that all bikonts except the Archaeplastida underwent its secondary loss (Andersson & Roger, 2002; Maruyama, Matsuzaki, Misawa, & Nozaki, 2009). The less controversial theories focused on Euglenozoa suggested that the current secondary plastid of euglenophytes could have originated before the split of kinetoplastids and other euglenozoans and become lost in all lineages but euglenophytes (Bodył et al., 2010; Hannaert et al., 2003) resulting in the presence of "green" genes dispersed throughout the whole group. The possibility of a cryptic, euglenophyte-unrelated plastid gain and loss within the kinetoplastid lineage, was also considered (Bodył et al., 2010; Leander, 2004; Martin & Borst, 2003). The alternative hypothesis to all the previously mentioned suggests a mere lateral gene transfer from cyanobacteria or primary or secondary

plastid-containing eukaryote based solely on the "you are what you eat" notion (Doolittle, 1998) without a requirement for a previously existing stable relationship between the host and the endosymbiont/organelle (Bodył et al., 2010; Leander, 2004; Maruyama et al., 2009).

The mosaicism in the evolutionary origins of euglenophyte genes provokes similar questions and similar approaches to their resolving. Euglenophytes contain undisputable amount of laterally transferred genes originating from rhodophytes or secondary algae containing red-algal plastids (e.g. haptophytes or stramenopiles), i.e. genes gained from other sources than the ancestor of the current plastid related to *Pyramimonas*. Most of these genes play a role in the typical plastid metabolic pathways and processes. The examples of these "nongreen" genes include several enzymes of the Calvin cycle (fructose bisphosphatase, glyceraldehyde 3-phosphate dehydrogenase, phosphoribulokinase, ribose 5-phosphate isomerase, and triosephosphate isomerase), several enzymes of the tetrapyrrole biosynthesis (glutamate 1-semialdehyde 2,1-aminotransferase, uroporphyrinogen decarboxylase, coproporphyrinogen oxidase, and protoporphyrinogen oxidase), glycolytic enzyme glucokinase, and tocopherol biosynthesis enzyme homogentisate phytyltransferase (Lakey & Triemer, 2016; Markunas & Triemer, 2016; Maruyama, Suzaki, Weber, Archibald, & Nozaki, 2011). Whether similar plastid-related genes, green or nongreen, are present in heterotrophic euglenids, and to what extent, is currently unknown.

Where do these genes come from? The least controversial explanation is simple lateral gene transfer from eukaryotic prey, congruent with the "you are what you eat" hypothesis (Doolittle, 1998). Many eukaryovorous euglenids prey on algae, often in a rather generalist way. This nutritional strategy represents a clear prerequisite for the plastid acquisition as well as a logical source of the lateral gene transfer from various phototrophs. But how come that the transferred genes were retained by the euglenid predator despite having photosynthetic or other plastidal function and being virtually useless to a heterotrophic organism? The answer could be that the said euglenid predator was not a pure heterotroph but rather a mixotroph which already possessed the plastid or *Pyramimonas*-related endosymbiont not yet fully transformed into a true organelle, much like the contemporary species *R. viridis* (Yamaguchi et al., 2012). Genes obtained from the algal prey could have been immediately recruited for a function in a preexisting plastid-localized process, allowing the reductive evolution of the plastid/endosymbiont genome and its further integration into the host. This evolutionary

process was recently suggested as a major driving force of the organellogenesis of the chromatophore of *Paulinella chromatophora* (Nowack et al., 2016) (a rhizarian amoeba with a plastid-like organelle, chromatophore, acquired via second independent primary endosymbiosis; Marin, Nowack, & Melkonian, 2005).

Alternative hypotheses count with the possibility of euglenophyte predecessors harbouring other endosymbiont(s) or transient or even stable plastid(s) in their evolutionary past which were lost and replaced with the plastid from *Pyramimonas*-related alga. This model is based on the "shopping bag" hypothesis and reflects the current state of knowledge regarding the transient relationships, repeated gains and losses, and overall plastid fluidity observed in dinoflagellates (Bodył et al., 2010; Howe, Barbrook, Nisbet, Lockhart, & Larkum, 2008; Larkum, Lockhart, & Howe, 2007; O'Neill, Trick, Henrissat, & Field, 2015). If this hypothesis was correct, at least one "red" endosymbiont or preplastid was present in the ancestor of euglenophytes at some point. It is even possible that some of these plastid-like symbioses or similar ecological relationships could have taken place long time before the acquisition of the extant green plastid, explaining the presence of genes of algal origin in heterotrophic euglenids (Bodył et al., 2010; Leander, 2004; Maruyama et al., 2011) and making some of the implications of the plastid-early hypothesis true. The plastid-related genes it had left in the nuclear genome of the euglenid could have facilitated the acquisition of a brand new plastid and its "enslavement" by reductive genome evolution, quite opposite to the *Paulinella*-like model proposed earlier. Another imaginable scenario is the coexistence of red and green endosymbionts/preplastids in a single cell for a certain period of time during which the host had time to gain genes from both before ultimately "deciding" to let the red one go and keep the green one.

These hypotheses are not mutually exclusive and the truth can well lie somewhere between these proposed models—or somewhere else entirely.

3. PLASTID MORPHOLOGY: DISPLAY OF DIVERSITY

Not much can be said about the morphology of euglenophyte plastids in general because their shapes, sizes, numbers, positions, and other characteristics vary greatly among different genera and even species or strains (Ciugulea & Triemer, 2010; Leedale, 1967).

The most basic universal characteristic is perhaps that the euglenophyte plastids are enveloped by three membranes which reflect its secondary endosymbiotic origin and evolutionary history: the inner two membranes are generally believed to be homologous with the two membranes of primary plastid (i.e. originating in the two envelope membranes of cyanobacteria), the additional one is of an eukaryotic ancestry, derived either from the euglenid endomembrane system or from the cytoplasmic membrane of the green-algal ancestor of the plastid (Gibbs, 1978; Lefort-Tran, 1981). The same number of plastid membranes is described in dinoflagellates with peridinin-containing plastid, while four membranes are generally more common occurrence in secondary plastids. The outermost membrane of euglenophyte plastid is not spatially continuous with the ER as in the case of some other organisms with secondary plastids such as cryptophytes, haptophytes, and heterokonts (Bolte et al., 2009; Maier, Zauner, & Hempel, 2015). However, substantial communication via vesicles was observed to take place between the endomembrane system and plastid, and it was proposed that its outermost membrane might in fact act as a part of the secretory system (Sulli & Schwartzbach, 1995, 1996). This would be crucial for the targeting of ER-synthesized proteins and other molecules into the plastid.

Euglenophyte plastids are very variable in terms of shape: they can be spherical, oval, partially flattened, disc-shaped, watch-glass-shaped, biconcave or biconvex, with various numbers of lobes and prominences, or even ribbon- or star-shaped (Leedale, 1967). A certain plastid shape is usually specific to a certain monophyletic or paraphyletic group of euglenophytes. The amount of plastids per cell ranges from several to several dozens and it is also usually species-specific but in this case a certain range is conserved rather than a concrete number: in the model species *E. gracilis*, for example, the number of plastids varies from 6 to 12 (Leedale, 1967). This variability might be linked to the fact that plastid replication can take place either synchronously with the cell division or independently of it.

Euglenophyte plastids have well-developed thylakoids (also termed *discs* in older literature) which are organized in elongated stacks (also termed *lamellae* or *bands*) of varying number of layers (usually three) instead of grana (i.e. relatively high cylindrical stacks of thylakoids present in plant plastids) (Ben-Shaul, Schiff, & Epstein, 1966; Gibbs, 1970; Gibor & Granick, 1962).

Many euglenophyte plastids contain visible pyrenoids (i.e. dense regions where most of the carbon fixation enzymes are localized): these may be

either naked or covered by a watch glass-shaped paramylon cap, either from one side (haplopyrenoids) or from both sides (diplopyrenoids). Under transmission electron microscopy, the pyrenoids appear as conspicuously delimited regions of dense granulation (Gibbs, 1970). Pyrenoids are absent in plastids of the family Phacaceae and also in some members of the Euglenaceae (Gibbs, 1970; Leedale, 1967). The presence/absence and appearance of pyrenoids can also change during the life of a single cell as a result of changes in environmental factors—light intensity, length of the light/dark cycle, starvation, nutritional value of the medium, and growth phase of the culture.

Euglenophytes possess a single eyespot (*stigma*), light-sensing, granular, red/orange-coloured, and carotenoid-containing organelle, which enables positive or negative phototaxis. In contrast to green algae, where eyespots exist within the plastids, the euglenophyte eyespot is located in the cytoplasm near the base of the flagellum (Benedetti & Checcucci, 1975; Iseki et al., 2002; Osafune & Schiff, 1980b; Walne & Arnott, 1967). However, it was probably derived from plastid as well (Walne & Arnott, 1967).

4. PLASTID GENOMES: SPED-UP EVOLUTION AND INTRONS GONE HAYWIRE

To this date, 17 euglenophyte plastid genomes (cpDNAs) have been published. The sampling covers most of the diversity of the Euglenaceae family with 14 cpDNAs (6 from the genus *Euglena* and 8 from other genera), one cpDNA of the Phacaceae family and two cpDNAs from the two genera of the Eutreptiales (Bennet, Wiegert, & Triemer, 2012; Bennett & Triemer, 2015; Bennett et al., 2014; Dabbagh & Preisfeld, 2016; Gockel & Hachtel, 2000; Hallick et al., 1993; Hrdá et al., 2012; Kasiborski, Bennett, Linton, & Lane, 2016; Pombert et al., 2012; Wiegert et al., 2012, 2013). Euglenophyte cpDNAs generally take the form of a circular chromosome. Four genomes are not complete and have not been circularized due to unknown number of repetitive sequences (Kasiborski et al., 2016) or unknown number of the ribosomal operons (Wiegert et al., 2012, 2013). Their basic characteristics including the comparison with the cpDNA of *P. parkeae* (Turmel et al., 2009) are summarized in Table 1.

The cpDNA of *P. parkeae,* the closest known relative of the euglenophyte plastid, is 101,605 bp long and contains 110 genes (Turmel et al., 2009). These values were seemingly reduced during the secondary

Table 1 Characteristics of the Plastid Genomes of 17 Euglenophytes and *Pyramimonas parkeae* According to Sequences Deposited in GenBank

Taxonomy	Species/Strain	cp Genome Size (bp)	Number of Genes	Number of Introns	GC Content (%)
	Euglena gracilis Z	143,171	90	145	26.1
	E. gracilis var. *bacillaris*	132,034	90	134	25.8
	Euglena longa	73,345	57	60	22.4
	Euglena viridis epitype	91,616	92	77	26.4
	E. viridis SAG 1224-17d	76,156	92	77	26.2
	Euglena mutabilis	86,975	92	76	26.7
	Monomorphina aenigmatica	74,746	92	53	29.4
	Monomorphina parapyrum	80,147	93	80	28.0
	Cryptoglena skujai	106,843	92	84	26.3
	Euglenaria anabaena	88,487	93	82	28.0
	Trachelomonas volvocina	85,392*	93	94	27.3
	*Strombomonas acuminata**	144,166	93	110	26.6
	*Colacium vesiculosum**	128,892	92	130	26.1
	Euglenaformis proxima	94,185	91	113	26.9
	*Phacus orbicularis**	65,992	90	66	27.2
	*Eutreptia viridis**	65,523	86	27	28.6
	Eutreptiella gymnastica	67,623	87	8	34.3
	Pyramimonas parkeae	101,605	110	1	34.7

The "number of genes" indicates both protein and RNA coding but duplicates and ORFs were not included. The genome sizes marked by *asterisk* denote incomplete genomes. The first three columns marked taxonomy symbolize phylogenetic relationships between the organisms—merged rows represent defined clades; the Euglenales are coloured in *orange* and additionally divided into the Euglenaceae (*light orange*) and the Phacaceae (*light red*), the Eutreptiales are coloured in *yellow* and the prasinophytes are coloured in *purple*.

plastid establishment—22 genes (e.g. all genes of NADH-plastoquinone oxidoreductase of plastidal respiratory chain) were lost or transferred into the nucleus of the common ancestor of euglenophytes (Fig. 2). The highly conserved core content of euglenophyte cpDNAs consists of 89 genes (including *rrn5* not present in *P. parkeae*). These include 32 genes for photosynthetic proteins, 5 genes for transcription/translation proteins, 22 genes for ribosomal proteins, 3 rRNAs, and 27 tRNAs. One to four of these genes were not found in some lineages but their absence does not show any

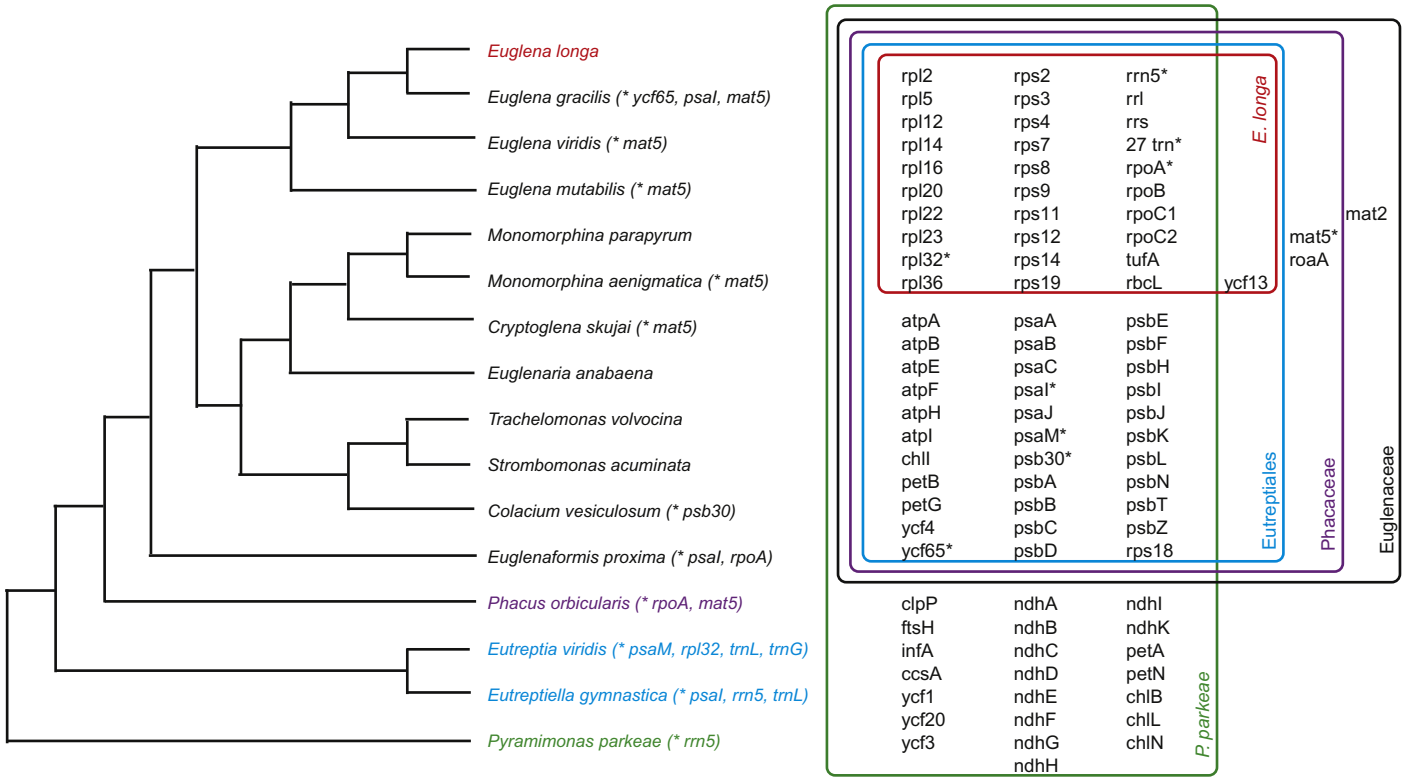

Fig. 2 Gene contents of the plastid genomes in 14 euglenophytes and *Pyraminonas parkeae*. *Asterisks* denote the genes missing in certain plastid genomes. Schematic phylogenetic tree is drawn according to Bennett and Triemer (2015).

phylogenetic pattern. Four genes were gained after the secondary endosymbiosis: the Eureptiales have obtained *ycf13* (synonymous to the intron-encoded maturase *mat1*), the Phacaceae possess three additional genes, *ycf13*, *mat5* (found within the *psbA* gene of *Lepocinclis spirogyroides*—GenBank record), and *roaA* (ribosomal operon-associated gene), and the Euglenaceae possess *mat2* in addition to the three aforementioned genes (Fig. 2; Bennett & Triemer, 2015). The *mat5* gene was lost in several independent instances within the Euglenales (Bennett & Triemer, 2015; Kasiborski et al., 2016). Unsurprisingly, genes for photosynthetic proteins have been lost in *Euglena longa* whose plastid has no photosynthetic activity. The gene contents are summarized in Fig. 2.

The plastid genome organization differs between *P. parkeae*, which contains two inverted repeats with ribosomal operon and large and small single copy region (Turmel et al., 2009), and most euglenophytes, which do not have this quadripartite arrangement. The cpDNA of *Eutreptiella gymnastica* represents a single exception as it contains two inverted repeats (one of them discontinuous), each with rRNA operon (Hrdá et al., 2012). The rRNA operon in *E. gracilis* and *E. longa* is organized in three tandemly repeated copies (Gockel & Hachtel, 2000; Hallick et al., 1993). The replication origins of euglenophyte cpDNAs are presumably located in the VNTR (variable number of tandem repeats) region (Koller & Delius, 1982; Ravel-Chapuis, Heizmann, & Nigon, 1982). Overall, the genes are arranged into 15 conserved gene clusters whose order and orientation have been excessively rearranged and it is hard to trace and reconstruct the course of these rearrangements (Dabbagh & Preisfeld, 2016).

The most noticeable trend in the euglenophyte plastid genome evolution is undoubtedly the rapid intron gain. The numbers of recognized introns in the genomes change over time as the annotations improve and so the numbers provided in Table 1 reflect the status of genome annotations published in the GenBank database as of July 2017. Remarkably, the cpDNA of *Pyramimonas* contains only a single intron like those of other prasinophyte algae such as *Ostreococcus tauri* and *Pycnococcus provasoli* (Turmel et al., 2009). The number of introns started to grow after the plastid acquisition by the ancestral euglenid before the split of Eutreptiales and Euglenales: cpDNAs of *Eutreptiella* and *Eutreptia* which represent the deepest splits from other euglenophytes already contain 8 and 27 introns, respectively (Hrdá et al., 2012; Wiegert et al., 2012). An intron of *psbC* gene, which carries an intron-encoded maturase *ycf13*, is considered the ancestral intron because it is the only homologous intron in euglenophyte cpDNAs (Bennett &

Triemer, 2015; Pombert et al., 2012). The major intron gain and amplification took place within the Euglenales lineage resulting in several species containing over 100 introns in their cpDNAs, the most extreme case being the model *E. gracilis* strain *Z* with at least 145 introns (Bennett & Triemer, 2015; Bennett et al., 2014; Hallick et al., 1993; Wiegert et al., 2013). The intron accumulation seems to be an ongoing process which is afoot at this very moment and it could be potentially very interesting to review some of the "old" sequences and see if they changed during the long-term cultivation by humans during a tiny part of the period of their sped-up evolution.

Not only are the euglenophyte plastid introns numerous, they are also unique in their structure. Majority of them are classified as group II introns, which are known from prokaryotes and mitochondria and plastids of eukaryotes. They function as self-splicing ribozymes and contain six stem-loop-forming domains and a conserved 5′-border motif (GUGYG). Group II introns are mobile elements and their mobility is mediated by maturases (Bonen & Vogel, 2001; Sheveleva & Hallick, 2004). The euglenophyte plastid group II introns are often significantly shorter (the average length of euglenophyte group II intron is 463 nt—approximately 100 nt shorter than the average group II intron of liverwort; Dabbagh & Preisfeld, 2016) with some of the conserved domains being missing and/or divergent beyond recognition. In addition to group II introns, euglenophyte plastid genomes contain group III introns which are exclusive to these organisms. They seem to be extremely derived form of group II introns which is much shorter (of average length around 100 nt) and lacks almost all core structures retaining only one modified conserved domain on 3′-end and a degenerate 5′-border motif (the consensus is NUNNG) (Bonen & Vogel, 2001; Doetsch, 2000; Doetsch, Favreau, Kuscuoglu, Thompson, & Hallick, 2001; Jenkins, Hong, & Hallick, 1995; Thompson, Copertino, Thompson, Favreau, & Hallick, 1995). To make matters even more complicated, euglenophyte plastid genomes also contain a number of twintrons—introns nested inside other introns (termed internal and external introns, respectively) which are spliced subsequently. Twintrons come in different types: group II, group III, and mixed or even complex ones where multiple internal introns (Copertino, Christopher, & Hallick, 1991; Copertino & Hallick, 1991, 1993) or even additional introns inside an internal intron (Drager & Hallick, 1993) can be observed in one external intron. In several cases, the recent conversion of a simple intron into a twintron is traceable on certain insertion sites. For example, six loci containing a twintron in the *E. gracilis* cpDNA contain only a single intron in *Monomorphina aenigmatica*.

In most of these cases, the introns of *M. aenigmatica* are orthologous to the external introns of their twintron counterparts in *E. gracilis* (Pombert et al., 2012). This further supports the notion that intron and twintron propagation in euglenophyte cpDNAs is a recent and probably still active process.

The relatively low GC content of the cpDNAs of euglenophytes is also linked to the intron accumulation since euglenid plastid introns are generally AT-rich and tend to bias the overall GC content of the whole genome. This is especially apparent in the case of *E. longa* whose cpDNA has the lowest number of genes while still being rich in introns, as a result its GC content is extremely low (22.4%) (Gockel & Hachtel, 2000).

5. PLASTID BIOGENESIS AND HOUSEKEEPING: HOW TO MAKE IT AND HOW TO CONTROL IT

The crucial part of organellogenesis is the ability of the host cell to take control over the replication and biogenesis of the formerly independent endosymbiotic hostage. The development of plastids of euglenophytes was studied in detail in the model species *E. gracilis* and the following subchapter will be based mainly on the findings regarding this particular organism.

Euglena plastids develop from proplastids which are nascent or dark-induced regressed stages with low or no photosynthetic activity, small number of thylakoids, and low amount of photosynthetic pigments stored mostly in their precursor forms (protochlorophyll and protochlorophyllide).

The plastid development is triggered by light and influenced by other environmental stimuli, especially availability of organic carbon sources which act as catabolic repressors that modulate the metabolic mode of the cell towards the more efficient one in the given situation. Glucose, ethanol, and acetate were shown to inhibit chlorophyll synthesis, while malate and succinate do not seem to function as catabolic repressors. Ethanol was also shown to inhibit synthesis or activity of various other proteins of carbon fixation pathway, light-harvesting antennae, and plastid housekeeping system (Horrum & Schwartzbach, 1980; Monroy & Schwartzbach, 1984; Schwartzbach, 2017). As a result, the cell performs the energetically expensive switch to the autotrophic nutritional mode and develops mature plastids only when carbon sources consumable by the relatively energetically cheap glyoxylate pathway become unavailable.

The DNA replication in plastids can take place independently of the nuclear DNA replication. However, the process of plastid greening and

development is directed by the nucleus on the transcriptional and translational level. The proteosynthetic apparatus of the developing plastids receives an external boost as its nucleus-encoded components (ribosomal proteins and initiation and elongation factors) are upregulated by light, synthesized in the cytoplasm, and transported into proplastids as well as other nucleus-encoded plastid proteins (Bingham & Schiff, 1979; Bingham & Shiff, 1979; Egan & Carell, 1972; Fox et al., 1980; Kraus & Spremulli, 1988). The amounts of rRNA, tRNA, and other nucleic acids which are fairly low in inactive proplastids increase up to threefold after illumination (Egan & Carell, 1972). The plastid genes are expressed constitutively regardless of light or dark conditions and the light increases the transcriptional activity but does not change the composition and ratio of the expressed genes (Geimer et al., 2009).

Protochlorophyll and protochlorophyllide are phototransformable molecules which are believed to act as photoreceptors on the thylakoid membranes of the developing plastids (perhaps in cooperation with a yet unspecified photoreceptor specialized to blue light sensing). These precursors are converted to chlorophyll and play a role in the initiation of de novo chlorophyll synthesis (Egan, Dorsky, & Schiff, 1975; Kirk, 1970; Stern, Epstein, & Schiff, 1964; Stern, Schiff, & Epstein, 1964).

The proplastids start to grow in size and new thylakoids are formed by the invagination of the innermost membrane and eventually start fusing and stacking into lamellae (Ben-Shaul et al., 1966). Thylakoid formation and chlorophyll synthesis are mutually dependent so the time dynamics of these two processes are concurrent. This development starts soon after the induction by light but it progresses very slowly during the initial lag phase and sets off rapidly after approximately 6 h. The length of the lag can be, however, significantly modified by the adjustment of the light conditions, namely preillumination. While the concentration of chlorophyll grows steadily, the concentration of carotenoids remains more or less the same—the final ratio between these two types of pigments is ca. 2.5:1 in the mature plastids (Stern, Schiff, et al., 1964). The interconnected carbon fixation and oxygen production start several hours after the plastid differentiation induction and grow rapidly after approximately 10 h. Around this time, the plastids swell quickly gaining roughly three times the original mass and thylakoid amount. It was noticed that the number of initial proplastids in *E. gracilis* is usually around 30 per cell while the final number of mature plastids is around 10. The logical deduction is that the proplastids actually form aggregates of three and fuse around this time off the differentiation causing a significant leap in

the plastids size and metabolic capacity. The process of the plastid maturation is completed in about 72 h (Ben-Shaul et al., 1966).

A vast majority of proteins which are synthesized and inserted into thylakoids during this process are coded in the nucleus and translated on cytoplasmic ribosomes—the gene expression is almost completely under the control of the host cell (Bingham & Shiff, 1979), only a small number of proteins are synthesized in the plastid (Schwartzbach & Schiff, 1974). These gene products are imported into plastids based on their rather complex-targeting signals.

The plastid protein import mechanism is much more complicated in euglenophytes in comparison to plants and is far from clearly resolved to this date. A schematic of the protein transport into euglenophyte plastid is shown in Fig. 3. Plastid-targeted precursor proteins (preproteins) coded in nucleus generally have an N-terminal signal peptide which is very similar to the classical signal peptide of proteins destined to the secretory pathway. The signal peptide is supposedly recognized by the signal recognition particle right after its translation and protrusion from the cytoplasmic ribosome and the rest of the translation takes place on the rough ER and the nascent preprotein is cotranslationally imported into the ER lumen. Subsequently, the signal peptide is cleaved by a signal peptidase in the ER revealing the following part of the signal whose amino acid composition is similar to that of a canonical plant chloroplast-targeting signal, the so-called transit peptide: it is rich in serine, threonine, and alanine; depleted in aspartic and glutamic acid; and has a slight positive charge as a result. The length of the transit peptides is quite variable and ranges from 36 to 135 amino acid residues in *E. gracilis* (Durnford & Gray, 2006). The origin of these transit peptides could be related to *cis*- and *trans*-splicing of short introns which were identified in some of these sequences (Vesteg et al., 2010). In euglenophytes, nucleus-encoded plastid preproteins can be divided into two classes based on the presence or absence of the third, highly hydrophobic domain which follows immediately after the transit peptide (Durnford & Gray, 2006) and acts as a stop-transfer transmembrane anchor. Preproteins are transported from the ER to the Golgi apparatus and they are either packed into a transport vesicle (class II) or anchored in its membrane (class I) with the transit peptide inside and the rest of the protein protruding into the cytoplasm (Durnford & Gray, 2006; Sulli, Fang, Muchhal, & Schwartzbach, 1999; Sulli & Schwartzbach, 1996). Vesicles from Golgi fuse with the outermost plastid membrane by a yet unknown mechanism. This process is resistant to *N*-ethylmaleimide that interacts with the *N*-ethylmaleimide-sensitive factor (NSF) and inhibits

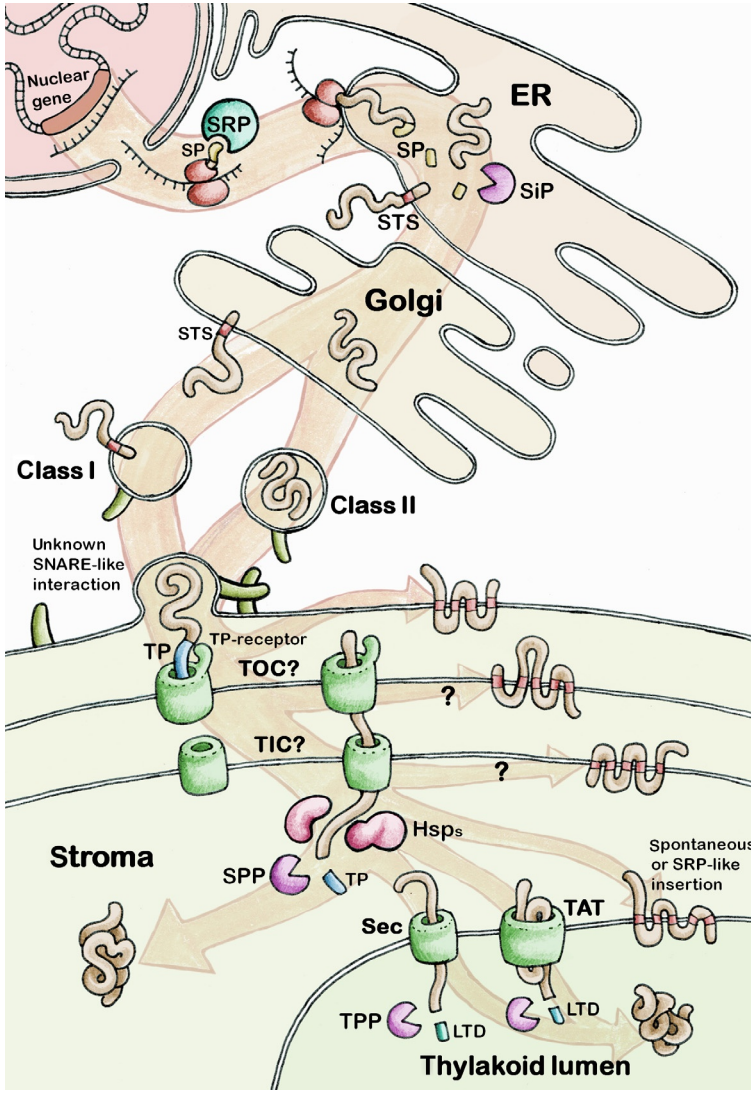

Fig. 3 Schematic of the plastid protein import pathway in euglenophytes. Nuclear-encoded plastid-targeted proteins are cotranslationally transported into the ER by the same mechanism as proteins destined into the secretory pathway based on their signal peptide (SP). They are either completely inserted into the ER lumen or remain anchored in its membrane by a hydrophobic stop-transfer signal (STS). Their SPs are cleaved by the signal peptidase (SiP) in the ER lumen. They pass through Golgi and are loaded onto vesicles which then fuse with the outermost plastid membrane via an unknown, SNARE-independent mechanism. Proteins destined to plastid stroma or thylakoids pass the remaining two membranes via transit peptide (TP)-dependent process which is probably facilitated by plant TIC- and TOC-like complexes. In plastid

soluble NSF attachment protein receptor (SNARE)-dependent membrane fusion. This suggests that SNAREs are not utilized in the vesicular transport of plastid preproteins (Sláviková et al., 2005). After reaching the intermembrane space between the outermost and the middle plastid membrane, the transit peptide is apparently recognized by a yet unknown receptor and the preprotein is pulled through the inner two membranes, most probably by a system homologous to the TOC/TIC (translocons at the outer/inner envelope membrane of chloroplasts) machinery which is present in primary plastids of plants (Shi & Theg, 2013) and generally conserved (albeit reduced in the number of subunits) among other secondary plastids (Maier et al., 2015; Sheiner & Striepen, 2013). The transit peptides of euglenid plastid proteins were repeatedly demonstrated to function in heterologous system and successfully direct the protein into plant plastid both in vitro and in vivo (Inagaki, Fujita, Hase, & Yamamoto, 2000; Shashidhara, Limsli, Shackletonii, Robinsonii, & Smith, 1992; Sláviková et al., 2005). These results suggest the presence of a TOC/TIC-like machinery in euglenophyte plastids. However, no direct proof of existence of these protein complexes in euglenophytes was brought to this date.

The transit peptide is presumably cleaved by stromal processing peptidase and the protein assumes its native conformation with the help of putative translocase-associated and stromal chaperones (Maier et al., 2015; Shi & Theg, 2013). Proteins destined to the thylakoid membrane or lumen are supposedly further sorted based on their additional signals (N-terminal, C-terminal, or internal) and inserted into thylakoids by specialized import pathways known from primary plastids such as the Sec-dependent pathway or the twin-arginine translocation pathway (Jarvis & Robinson, 2004).

The transcription of genes encoded in the plastid genome was observed to take place constitutively with practically all genes including several unknown ORFs and pseudogenes being expressed and present as mRNAs in the plastid, albeit in very low concentration. The transcripts sometimes form very large and relatively stable polycistronic units which are subsequently subjected to extensive posttranscriptional modifications. The plastid

stroma, the TPs are cleaved by the stromal processing peptidase (SPP) and the proteins are sorted into their final localization (stroma, thylakoid membrane, or thylakoid lumen) and folded to their mature conformation. Proteins destined to thylakoid lumen are transported based on their lumen-targeting domains (LTD) which are then cleaved by the thylakoid-processing peptidase (TPP). Proteins of the plastid envelope are inserted into their target membrane via an unknown mechanism or spontaneously.

gene expression does not react to environmental stimuli in terms of quality but rather on quantitative level. Moreover, some genes seem to have very short turnover and to be expressed only to be degraded very quickly. In summary, regulation of the gene expression in euglenophyte plastid might be taking place predominantly on the level of posttranscriptional RNA processing and/or translation rather than transcription (Geimer et al., 2009; Thompson et al., 1995). A small portion of plastid transcripts was shown to have polyadenylated 3′-ends in *E. gracilis*. The low ratios of polyadenylated vs nonpolyadenylated RNAs (only 1/350 to 1/100,000) suggest that the polyadenylation serves as a signal for exonucleolytic degradation as it does in the case of primary plastids (Záhonová et al., 2014).

6. PLASTID METABOLISM: A FACTORY WITH REDUNDANT PRODUCTION LINES

In this chapter, findings on biosynthetic pathways in euglenophyte plastid will be briefly summarized with focus on their potentially unique characteristics and differences from other groups which are interesting from evolutionary point of view. Almost all studies cited in this section were performed on the model euglenophyte *E. gracilis*, and generalizations based on them should be taken with a grain of salt.

Photosynthetic apparatus of the *E. gracilis* plastid is comparable to that of green algae and plants in regard to the function and architecture. Photosynthesis is a subject of tight regulation by light (Yoshida et al., 2016) and it reacts negatively to both extremes in light conditions as the photosynthetic activity is inhibited or even completely compromised by both darkness and the excessive illumination by visible or UV light (Richter, Helbling, Streb, & Häder, 2007). Additionally, the photosynthetic capacity of *E. gracilis* is affected by salinity (Gonzalez-Moreno, Gomez-Barrera, Perales, & Moreno-Sanchez, 1997) and other ion concentrations (Ferroni, Baldisserotto, Fasulo, Pagnoni, & Pancaldi, 2004; Krajčovič et al., 2015).

E. gracilis synthesizes chlorophylls *a* and *b* via the standard plastid-localized pathway from aminolevulinic acid. There are several spectroscopically distinguishable precursor forms of chlorophylls (i.e. protochlorophyll, protochlorophyllide, and phytyl-protochlorophyllide), and developing plastids especially contain these precursors to a great extent, and their ratio, dynamics, and interconversions are believed to play a major role in regulation of biosynthesis of chlorophyll and possibly other compounds (Kirk, 1970).

For example, protochlorophyllide has been shown to inhibit further steps in chlorophyll biosynthesis until its light-induced conversion (Schwartzbach, Schiff, & Goldstein, 1975).

Aminolevulinic acid can be synthesized in two ways: through the C5 pathway from glutamate or through the C4 (Shemin) pathway from glycine and succinyl-CoA; most organisms utilize only one of these pathways. In *E. gracilis*, aminolevulinic acid was proven to be synthesized via the C5 pathway and for some time the organism was thought to lack the C4 pathway (Gomez-Silva, Timko, & Schiff, 1985). However, the situation was recently shown to be more complex as two pathways for the synthesis of aminolevulinic acid followed by the synthesis of haem have been predicted in *Euglena*. One pathway is predicted to be partially localized in either the cytosol or mitochondrion to produce haem for mitochondria. This pathway utilizes aminolevulinic acid synthesized in the C4 pathway. Another pathway is localized in the plastid to produce chlorophyll and it utilizes aminolevulinic acid synthesized in the C5 pathway (Kořený & Oborník, 2011).

Carotenoid synthesis seems to be linked to chlorophyll production and it was even hypothesized that one could be synthesized from another due to the observed reciprocal proportion of these two types of pigments (Wolken & Mellon, 1956). The regulation of biosynthesis of carotenoids and chlorophyll indeed seems to be tuned and subjected to reciprocal feedback but the exact mechanism of it remains unclear. Major *E. gracilis* carotenoids are antheraxanthin (more than 80%), β-carotene (11%), and neoxanthin (7%), which represent 99% of total carotenoids. Minor carotenoids include cryptoxanthin, γ-carotene, ζ-carotene, echinenone, hydroxyechinenone, and its derivate canthaxanthin (also termed euglenanone in older literature as it was believed to be unique to euglenids) (Krinsky & Goldsmith, 1960). The first metabolite in carotenoid synthesis pathway is geranylgeranyl diphosphate (GGPP) which is synthesized in putatively plastidal methylerythritol phosphate/deoxyxylulose phosphate (MEP/DOXP) pathway—as opposed to the GGPP synthesized in mitochondrial mevalonate pathway which is used for the synthesis of phytols and sterols in *E. gracilis*. Therefore, there are two distinct GGPP pools in *E. gracilis* and carotenoids and phytols are synthesized by independent pathways (Kim, Filtz, & Proteau, 2004).

A variety of terpenoids are synthesized in the *E. gracilis* plastid: plastoquinone, α-tocopherol, α-tocopherolquinone, phytylquinone, nonaprenyl, octaprenyl, nonaprenyl toluquinone, octaprenyl toluquinone, and phytyl

pyrophosphate (Griffiths, Threlfall, & Goodwin, 1967; Thomas & Threlfall, 1974). Their production is not directly linked to photosynthesis but it is light dependent; terpenoid formation is mostly inhibited in dark-grown cultures as opposed to ubiquinone which is synthesized in mitochondrion and whose production level does not correlate with different light conditions (Griffiths et al., 1967). Polyprenylation/phytylation or nonoxidative decarboxylation of homogentisate were proposed as a mechanism of this reaction (Thomas & Threlfall, 1974). Tocopherols (compounds collectively termed as vitamin E) are present in all *E. gracilis* cell fractions including plastids where 97% of their bulk is represented by α-tocopherol which is thought to be synthesized in situ from other tocopherols which are transported from outside the plastid (Shigeoka, Onishi, Nakano, & Kitaoka, 1986). This compound functions as a protective agent against reactive oxygen species and its synthesis modulation by various substrates and conditions was studied in the view of potential biotechnological or pharmaceutical application (Fujita, Aoyagi, Ogbonna, & Tanaka, 2008; Fujita, Ogbonna, Tanaka, & Aoyagi, 2009). Ascorbic acid (vitamin C) is also an important antioxidative protectant. Remarkably, in *E. gracilis* it is synthesized by the alternative terminal enzyme, L-galactonolactone dehydrogenase, instead of L-gulonolactone oxidase which was lost multiple times and then replaced by an alternative enzyme in multiple phototrophic lineages (Wheeler, Ishikawa, Pornsaksit, & Smirnoff, 2015).

E. gracilis also synthesizes galactolipids (monogalactosyl and digalactosyl diglycerides) and sulpholipids (mainly sulphoquinovosyl diglyceride) in its plastid (Davies, Mercer, & Goodwin, 1966; Rosenberg & Gouaux, 1967; Rosenberg, Gouaux, & Milch, 1966). Production of these lipids is linked to the plastid development and thylakoid membrane amplification and is modulated by light conditions—much higher amounts of these compounds can be found in photosynthetically active cells in comparison to dark-grown or bleached ones (Matson, Fei, & Chang, 1970). Interestingly and in contrast to plants, monogalactosyl and digalactosyl diglycerides are probably synthesized by two different enzymes in *E. gracilis* (Blee & Schantz, 1978). Additionally, the conversion from monogalactosyl to digalactosyl diglycerides is not possible, and the ratio of these galactolipids is significantly skewed towards digalactosyl diglycerides which are synthesized preferentially (Matson et al., 1970). These lipids are essential for the thylakoid assembly and growth, and some have been shown to be miscible with chlorophyll and phytol, major nonprotein components of the thylakoid membranes (Liljenberg & Selstam, 1980).

The major storage compound of euglenids is paramylon, a starch-like polysaccharide, which forms granules of various morphologies in the cytosol and can take up to 90% of the dry cell weight. The paramylon consists of an unbranched, water-insoluble β-1,3-glucan which is rather unusual since most eukaryotes generally synthesize α-glucans (or branched β-1,4- or β-1,6-glucans in case of fungi and some grains) (Barras & Stone, 1968; Barsanti, Vismara, Passarelli, & Gualtieri, 2001; Dwyer & Smillie, 1971; Monfils, Triemer, & Bellairs, 2011; Šantek, Felski, Friehs, Lotz, & Flaschel, 2010). The paramylon biogenesis occurs in all euglenids including nonphotosynthetic species suggesting that it is not evolutionarily related to plastid acquisition. In heterotrophic euglenids, the paramylon is synthesized in mitochondrial prominences and mitochondrion-derived vesicles by the gluconeogenesis pathway (Calvayrac & Briand, 1978). In photosynthetic euglenophytes, the paramylon can be synthesized by both the original mitochondrial pathway and during carbon fixation in pyrenoids of plastids. The close association of plastid and mitochondrion and the eccentrical flow of vesicles derived presumably from both organelles have been observed during the paramylon biogenesis (Calvayrac, Laval-Martin, Briand, & Farineau, 1981). This functional connection of the two organelles is remarkable and can be a potential source of unusual molecular phenomena regarding exchange of metabolites and possibly other compounds. The other major storage compounds of euglenids are wax esters which are not synthesized in plastid (Inui, Miyatake, Nakano, & Kitaoka, 1982; Koritala, 1989; Schneider & Betz, 1985) but whose production is linked to the production of paramylon, since interconversions between these two storage molecules occur regularly and their balance shifts in reaction to light and oxygen availability: in aerobic conditions, the cell preferentially accumulates paramylon; in anaerobic conditions, the wax ester accumulation is more significant (Barras & Stone, 1968; Inui et al., 1982). Both of these compounds are potentially usable in biotechnology, agriculture, nutrition, and even cancer prevention (Karaca et al., 2014; Krajčovič et al., 2015; Kuda, Enomoto, & Yano, 2009; Rodríguez-Zavala, Ortiz-Cruz, Mendoza-Hernández, & Moreno-Sánchez, 2010; Sugiyama et al., 2009; Watanabe et al., 2013).

7. SECONDARY OSMOTROPHY AND PLASTID BLEACHING: PLASTIDS THAT FORGOT HOW TO PLASTID

Secondarily osmotrophic euglenids represent euglenophytes which have lost their photosynthetic pigments and the ability to perform

photosynthesis. There are at least five species of euglenophytes that have lost the photosynthesis independently (Marin, 2004; Marin, Palm, Klingberg, & Melkonian, 2003). It is a matter of a debate whether some of these euglenophytes such as *Euglena quartana* (previously *Khawkinea quartana*) or *Phacus ocellatus* (previously *Hyalophacus ocellatus*) have lost the plastid compartments completely or whether they still contain residual plastids with genomes (Marin, 2004). However, the best-studied secondary osmotroph is *E. longa*, a close relative of photosynthetic *E. gracilis* (Mullner, Angeler, Samuel, Linton, & Triemer, 2001), which contains a plastid with a genome (Gockel & Hachtel, 2000).

The circular 73 kb plastid genome of *E. longa* is about half the size of a plastid genome of *E. gracilis* and it carries 57 protein-coding genes (Gockel & Hachtel, 2000)—housekeeping genes responsible for transcription and translation (*rrn*, *rpl*, *rps*, *rpo*, and *tuf* genes), tRNA genes, and several ORFs and genes with unknown function (*orf* and *ycf* genes). Transcripts and proteins of various genes, including Rubisco large subunit gene (*rbc*L), have been found suggesting that the *E. longa* plastid possesses functional transcription and translation machinery (Gockel & Hachtel, 2000; Sheveleva & Hallick, 2004; Záhonová et al., 2016). It was recently shown that an intact plastid genome is essential for *E. longa* growth (Hadariová, Vesteg, Birčák, Schwartzbach, & Krajčovič, 2017) but it remains a mystery which gene(s) is/are indispensable for its survival (Gockel & Hachtel, 2000; Hadariová et al., 2017). The photosynthesis-related genes (photosystems I and II, cytochrome *b6f* complex, ATP-synthase) have disappeared from the *E. longa* plastid genome with the exception of *rbc*L (Gockel & Hachtel, 2000). Both *rbc*L and nucleus-encoded *Rbc*S genes are translated but their abundance in cells is very low. Protein sequences of *E. longa rbc*L and *Rbc*S are highly divergent compared to their homologues in the photosynthetic relatives, suggesting that the Rubisco enzyme of *E. longa* probably has an unusual function, if any (Záhonová et al., 2016). Some other nonphotosynthetic organisms retain *rbcL* genes in their plastid genomes and it is hypothesized that these RbcL proteins may either act as oxygenases, be involved in glycine and serine biosynthesis, be required for an alternative lipid biosynthesis pathway (Schwender, Goffman, Ohlrogge, & Shachar-Hill, 2004), or perform another unidentified function (Sanchez-Puerta, Lippmeier, Apt, & Delwiche, 2007; Wolfe & dePamphilis, 1998; Záhonová et al., 2016).

E. longa lacks the eyespot and the paraflagellar swelling, but apart from that, the cell is indistinguishable from bleached mutants of *E. gracilis* under light microscopy, and thus it was formerly viewed as a naturally bleached

form *of E. gracilis* (Bodył, 1996). Bleaching of *E. gracilis* is an irreversible process inducible by antibiotics, UV light, high pressure, heat, mutagens, or carcinogens (Krajčovič, Ebringer, & Schwartzbach, 2001). In *E. gracilis*, various morphological and ultrastructural changes can be observed during bleaching. During the switching from autotrophy to heterotrophy, the plastids and plastid DNA are degraded and the colour of the culture changes from green to white (or slightly orange or pink depending on the presence of carotenoids). The bleaching level depends on several factors—pH, content of phosphate in the medium, or the age of the culture. It is possible to reach 100% plastid bleaching in some cases (Cook, Harris, & Nachtwey, 1974; Krajčovič, Ebringer, & Polónyi, 1989). Spontaneous bleaching in *E. gracilis* was observed as early as in 1912 by Ternetz and streptomycin was the first defined agent for controlled bleaching of all cells in a population. In 1961, Ebringer widened a list of bleaching agents by adding erythromycin and some other macrolide antibiotics (Krajčovič et al., 2001).

There are various types of antibacterial drugs with different modes of action. Some of them act as inhibitors of bacterial DNA synthesis (e.g. mitomycin or anthramycin) while the ones with aminohexose molecular structure function as inhibitors of bacterial as well as plastidal protein synthesis (e.g. kanamycin, pactamycin, or neomycin). Their irreversible effect on *E. gracilis* was observed in the past (Ebringer, 1972). Perhaps the most effective eliminators of *E. gracilis* plastids are quinolone antibiotics (inhibitors of bacterial DNA gyrase), especially their new derivatives—fluoroquinolones (Krajčovič et al., 1989). Various ultrastructural and plastid DNA changes occur when different bleaching agents are used (Krajčovič et al., 2001).

For example, the quinolones, but also nitrosoguanidine and furylfuramide, reduce the number of thylakoids and destroy plastids (Polónyi, Ebringer, Krajcovic, & Kapeller, 1990). The plastid DNA degradation has been observed in all bleached *E. gracilis* cultures. Conkling, Thomas, and Ortiz (1993) noticed a gradual loss of cpDNA and the study of Krajčovič et al. (1999) has demonstrated a structural rearrangement of cpDNA in *E. gracilis* cells bleached by elevated temperature. Recently, it was shown that the plastid gene *rpl*16 invariably persists in *E. gracilis* strain *Z* bleached in the long-term by ofloxacin (DNA gyrase inhibitor) or streptomycin (bacterial protein synthesis inhibitor) (Hadariová et al., 2017), which suggested that this gene was retained presumably due to its position on the cpDNA being close to the replication origin (Hadariová et al., 2017). A different gene, 16S rRNA, has been retained in *E. gracilis* FACHB47 strain

treated with the same antibiotics (Wang, Shi, & Xu, 2004). It has been reported that plastid transcripts are reduced after treating by various bleaching agents—UV light, temperature, and antibiotics (Geimer et al., 2009).

In the past decades, several stable white mutants derived from *E. gracilis* have been characterized (Polónyi, Ebringer, Dobias, & Krajčovič, 1998; Schiff, Lyman, & Russell, 1971). *W$_3$BUL* mutant was induced with UV irradiation of *E. gracilis* var. *bacillaris* strain (Schiff et al., 1971). It contains plastid remnants (Heizmann, Salvador, & Nigon, 1976; Osafune & Schiff, 1980a; Osafune, Schiff, & Hase, 1987; Parthier & Neumann, 1977) and a specific type of sulpholipid in its thylakoid membranes (Saidha & Schiff, 1989). The treatment of *E. gracilis* var. *bacillaris* with streptomycin produced another mutant strain, *W$_{10}$BSmL*. In contrast to *W$_3$BUL*, *W$_{10}$BSmL* contains neither the eyespot (Osafune & Schiff, 1980b), carotenoids (Fong & Schiff, 1979), and sulpholipids (Saidha & Schiff, 1989), nor the plastid residues (Osafune & Schiff, 1983). The *W$_{gm}$ZOflL* mutant was derived from *E. gracilis* strain *Z* by treatment with an ofloxacin derivative (Polónyi et al., 1998). The common features of *W$_{gm}$ZOflL* and *W$_3$BUL* mutants include the presence of the eyespot and carotenoids. The giant mitochondria were described in *W$_{gm}$ZOflL* in the past (Polónyi et al., 1998) but current microscopic observations suggest that *W$_{gm}$ZOflL* mutant does not possess mitochondria of such form anymore (personal experiences). The level of the plastid genome degradation, i.e., plastid genes presence/absence of the nonphotosynthetic mutants of *E. gracilis*, is still unknown but they probably lack most of the, if not all, plastid genes (unpublished data).

The bleaching of *E. gracilis* induced by fluoroquinolone ofloxacin has been shown to be useful in the investigation of potential antimutagens such as flavonoids (Križková, Nagy, Polónyi, & Ebringer, 1998) and antioxidants as ascorbic acid, sodium selenite, and many others (Ebringer et al., 1996; Kogan et al., 2004; Križková, Ďuračková, Šandula, Sasinková, & Krajčovič, 2001; Križková, Mučaji, Nagy, & Krajčovič, 2004; Križková et al., 2006). Ciprofloxacin, besides the effective bleaching of *E. gracilis*, also causes a suppression of the rudimentary plastid (apicoplast) genome replication in a parasite *Toxoplasma gondii* (Roos & Fichera, 1997). Various other quinolones have been demonstrated as useful inhibitors of a malaria-causing apicomplexan *Plasmodium falciparum* (Mahmoudi et al., 2003). Thus, *E. gracilis* plastid bleaching may also have practical implications. The research of the plastid genome degradation in bleached *E. gracilis* may represent a suitable tool for the development of new drugs harmless to humans and

successful in fighting apicomplexan parasites that cause millions of fatal disease cases in the third-world countries annually (Krajčovič et al., 2001).

8. CONCLUSIONS

Euglenophytes are a group of phototrophic protists which have been discovered as early as at the end of 17th century. During the last century, their biochemistry and ultrastructure were inspected to a great detail. However, there are still many missing pieces in the puzzle of their molecular biology, genetics, and phylogenetics as some of the methods crucial for these disciplines became widely available only recently and others are still quite limited because of the missing genomic data and insufficient means for genetic transformation. These hurdles are, however, very much worth overcoming because photosynthetic euglenids are potentially very useful models for both applied and basic research. From the applied point of view, they represent easily and environment-friendly cultivated microalgae capable of synthesizing various compounds usable in the production of biofuels and other lipid-based technological materials as well as nutritional and pharmaceutical products. From the basic point of view, they are one of the key groups to understand secondary endosymbioses, establishment and reduction of organelles and evolutionary processes taking part therein. Findings concerning euglenid molecular biology, genetics, and genomics can help to illuminate the evolution of excavates and, by extension, of eukaryotes as whole.

ACKNOWLEDGEMENTS

The authors wish to acknowledge professor Bozena Zakryś who is the author of the drawings used in Fig. 1 and who kindly approved their usage in this publication.

Support for the authors' salaries and stipend came from the project of the Ministry of Education, Youth and Sports of CR within the National Sustainability Program II (Project BIOCEV-FAR) LQ1604 and by the project "BIOCEV" (CZ.1.05/1.1.00/02.0109).

REFERENCES

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., et al. (2012). The revised classification of eukaryotes. *The Journal of Eukaryotic Microbiology*, *59*(5), 429–493. http://dx.doi.org/10.1111/j.1550-7408.2012.00644.x.

Andersson, J. O., & Roger, A. J. (2002). A cyanobacterial gene in nonphotosynthetic protists—An early chloroplast acquisition in eukaryotes? *Current Biology: CB*, *12*(2), 115–119. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11818061.

Barras, D. R., & Stone, B. A. (1968). Carbohydrate composition and metabolism in Euglena. In D. E. Buetow (Ed.), *The biology of Euglena: Vol. II. Biochemistry* (149–191). NewYork: Academic Press.

Barsanti, L., Vismara, R., Passarelli, V., & Gualtieri, P. (2001). Paramylon (β-1,3-glucan) content in wild type and WZSL mutant of Euglena gracilis. Effects of growth conditions. *Journal of Applied Phycology*, *13*(1), 59–65. http://dx.doi.org/10.1023/A:1008105416065.

Benedetti, A. P., & Checcucci, A. (1975). Paraflagellar body (PFB) pigments studied by fluorescence microscopy in Euglena gracilis. *Plant Science Letters*, *4*, 47–51.

Bennet, M., Wiegert, K., & Triemer, R. (2012). Comparative chloroplast genomics between Euglena viridis and Euglena gracilis (Euglenophyta). *Phycologia*, *51*(6), 711–718. http://dx.doi.org/10.2216/12-017.1.

Bennett, M. S., & Triemer, R. E. (2015). Chloroplast genome evolution in the Euglenaceae. *Journal of Eukaryotic Microbiology*, *62*(6), 773–785. http://dx.doi.org/10.1111/jeu.12235.

Bennett, M. S., Wiegert, K. E., & Triemer, R. E. (2014). Characterization of Euglenaformis gen. nov. and the chloroplast genome of Euglenaformis [Euglena] proxima (Euglenophyta). *Phycologia*, *53*(1), 66–73. http://dx.doi.org/10.2216/13-198.1.

Ben-Shaul, Y., Schiff, J. A., & Epstein, H. T. (1966). Studies of chloroplast development in euglena: VII. Fine structure of the developing plastid 1. *Biophysical Journal*, *6*(4), 373–383. http://dx.doi.org/10.1038/185825a0.

Bingham, S., & Schiff, J. A. (1979). Events surrounding the early development of Euglena chloroplasts. 15. Origin of plastid thylakoid polypeptides in wild-type and mutant cells. *Biochimica et Biophysica Acta (BBA)—Bioenergetics*, *547*(3), 512–530.

Bingham, S., & Shiff, J. A. (1979). Events surrounding the early development of Euglena chloroplasts. 16. Plastid thylakoid polypeptides during greening. *Biochimica et Biophysica Acta (BBA)—Bioenergetics*, *547*(3), 531–543.

Blee, E., & Schantz, R. (1978). Biosynthesis of galactolipids in Euglena gracilis: I, incorporation of UDP galactose into galactosyldiglycerides. *Plant Science Letters*, *13*, 247–255.

Bodył, A. (1996). Is the origin of Astasia longa an example of the inheritance of acquired characteristics? *Acta Protozoologica*, *35*, 87–94, Nencki Institute of Experimental Biology.

Bodył, A., Mackiewicz, P., & Milanowski, R. (2010). Did trypanosomatid parasites contain a eukaryotic alga-derived plastid in their evolutionary past? *Journal of Parasitology*, *96*(2), 465–475. http://dx.doi.org/10.1645/Ge-1810.1.

Bolte, K., Bullmann, L., Hempel, F., Bozarth, A., Zauner, S., & Maier, U.-G. (2009). Protein targeting into secondary plastids. *The Journal of Eukaryotic Microbiology*, *56*(1), 9–15. http://dx.doi.org/10.1111/j.1550-7408.2008.00370.x.

Bonen, L., & Vogel, J. (2001). The ins and outs of group II introns. *Trends in Genetics*, *17*(6), 322–331. http://dx.doi.org/10.1016/S0168-9525(01)02324-1.

Calvayrac, R., & Briand, J. (1978). Paramylon synthesis and the chondriome of Euglena gracilis Z. In G. Ducet & C. Lance (Eds.), *Plant mitochondria* (pp. 435–443). Amsterdam: Elsevier/North Holland.

Calvayrac, R., Laval-Martin, D., Briand, J., & Farineau, J. (1981). Paramylon synthesis by Euglena gracilis photoheterotrophically grown under low O2 pressure. *Planta*, *153*(1), 6–13. http://dx.doi.org/10.1007/BF00385311.

Cavalier-Smith, T. (1981). Eukaryote kingdoms: Seven or nine? *Biosystems*, *14*(3–4), 461–481. http://dx.doi.org/10.1016/0303-2647(81)90050-2.

Cavalier-Smith, T. (2016). Higher classification and phylogeny of Euglenozoa. *European Journal of Protistology*, *56*, 250–276. http://dx.doi.org/10.1016/j.ejop.2016.09.003.

Ciugulea, I., & Triemer, R. E. (2010). *A color atlas of photosynthetic euglenoids*. East Lansing, MI: Michigan State University Press. Retrieved from http://library.wur.nl/WebQuery/clc/1953618.

Conkling, B. A., Thomas, E. J., & Ortiz, W. (1993). Delayed but complete loss of chloroplast DNA in heat-bleaching cultures of Euglena gracilis. *Journal of Plant Physiology*, *142*(3), 307–311. http://dx.doi.org/10.1016/S0176-1617(11)80427-X.

Cook, J. R., Harris, P., & Nachtwey, D. S. (1974). Irreversible plastid loss in Euglena gracilis under physiological conditions. *Plant Physiology*, *53*(2), 284–290. http://dx.doi.org/10.1104/PP.53.2.284.

Copertino, D. W., Christopher, D. A., & Hallick, R. B. (1991). A mixed group-II/group-III twintron in the Euglena gracilis chloroplast ribosomal protein S3-gene—Evidence for intron insertion during gene evolution. *Nucleic Acids Research*, *19*(23), 6491–6497.

Copertino, D. W., & Hallick, R. B. (1991). Group II twintron: An intron within an intron in a chloroplast cytochrome b-559 gene. *The EMBO Journal*, *10*(2), 433–442. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=452664&tool=pmcentrez&rendertype=abstract.

Copertino, D. W., & Hallick, R. B. (1993). Group II and group III introns of twintrons: Potential relationships with nuclear pre-mRNA introns. *Trends in Biochemical Sciences*, *18*(12), 467–471. http://dx.doi.org/10.1016/0968-0004(93)90008-B.

Dabbagh, N., & Preisfeld, A. (2016). The chloroplast genome of Euglena mutabilis-cluster arrangement, intron analysis, and intrageneric trends. *Journal of Eukaryotic Microbiology*, *1993*, 31–44. http://dx.doi.org/10.1111/jeu.12334.

Davies, W. H., Mercer, E. I., & Goodwin, T. W. (1966). Some observations on the biosynthesis of the plant sulpholipid by Euglena gracilis. *The Biochemical Journal*, *98*(2), 369–373. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1264853&tool=pmcentrez&rendertype=abstract.

de M. Bicudo, C. E., & Menezes, M. (2016). Phylogeny and classification of Euglenophyceae: A brief review. *Frontiers in Ecology and Evolution*, *4*(March), 1–15. http://dx.doi.org/10.3389/fevo.2016.00017.

Doetsch, N. A. (2000). *Group III intron structure and evolutionary analysis in euglenoid chloroplast genomes (doctoral dissertation)*. The University of Arizona. http://hdl.handle.net/10150/284139.

Doetsch, N. A., Favreau, M. R., Kuscuoglu, N., Thompson, M. D., & Hallick, R. B. (2001). Chloroplast transformation in Euglena gracilis: Splicing of a group III twintron transcribed from a transgenic psbK operon. *Current Genetics*, *39*(1), 49–60. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11318107.

Doolittle, W. F. (1998). You are what you eat: A gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics: TIG*, *14*(8), 307–311. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9724962.

Douglas, S. E., & Turner, S. (1991). Molecular evidence for the origin of plastids from a cyanobacterium-like ancestor. *Journal of Molecular Evolution*, *33*(3), 267–273. http://dx.doi.org/10.1007/BF02100678.

Drager, R. G., & Hallick, R. B. (1993). A complex twintron is excised as four individual introns. *Nucleic Acids Research*, *21*(10), 2389–2394. http://dx.doi.org/10.1093/nar/21.10.2389.

Durnford, D. G., & Gray, M. W. (2006). Analysis of Euglena gracilis plastid-targeted proteins reveals different classes of transit sequences. *Eukaryotic Cell*, *5*(12), 2079–2091. http://dx.doi.org/10.1128/EC.00222-06.

Dwyer, M. R., & Smillie, R. M. (1971). β-1,3 glucan: A source of carbon and energy for chloroplast development in Euglena Gracilis. *Australian Journal of Biological Sciences*, *24*(1), 15–22.

Ebringer, L. (1972). Are plastids derived from prokaryotic micro-organisms? Action of antibiotics on chloroplasts of Euglena gracilis. *Journal of General Microbiology*, *71*(1), 35–52. http://dx.doi.org/10.1099/00221287-71-1-35.

Ebringer, L., Dobias, J., Krajčovič, J., Polónyi, J., Križková, L., & Lahitová, N. (1996). Antimutagens reduce ofloxacin-induced bleaching in Euglena gracilis. *Mutation Research/Environmental Mutagenesis and Related Subjects*, *359*(2), 85–93. http://dx.doi.org/10.1016/S0165-1161(96)90255-1.

Egan, J. M., & Carell, E. F. (1972). Studies on chloroplast development and replication in euglena: III. A study of the site of synthesis of alkaline deoxyribonuclease induced during chloroplast development in Euglena gracilis. *Plant Physiology*, *50*(3), 391–395.

Egan, J. M., Dorsky, D., & Schiff, J. A. (1975). Events surrounding the early development of euglena chloroplasts: VI. Action spectra for the formation of chlorophyll, lag elimination

in chlorophyll synthesis, and appearance of TPN-dependent triose phosphate dehydrogenase and alkaline DNase activities. *Plant Physiology*, *56*(2), 318–323.

Ehrenberg, C. G. (1830). *Organisation, Systematik und geographisches Verhältniss der Infusionsthierchen*. Berlin: Druckerei der Königlichen Akademie der Wissenschaften. Retrieved from http://www.biodiversitylibrary.org/item/18341.

Ferroni, L., Baldisserotto, C., Fasulo, M. P., Pagnoni, A., & Pancaldi, S. (2004). Adaptive modifications of the photosynthetic apparatus in Euglena gracilis Klebs exposed to manganese excess. *Protoplasma*, *224*(3–4), 167–177. http://dx.doi.org/10.1007/s00709-004-0072-4.

Fong, F., & Schiff, J. A. (1979). Blue-light-induced absorbance changes associated with carotenoids in Euglena. *Planta*, *146*(2), 119–127. http://dx.doi.org/10.1007/BF00388221.

Fox, L., Erion, J., Tarnowski, J., Spremulli, L., Brot, N., & Weissbach, H. (1980). Communication Eglena gracilis Chloroplast EF-Ts. *The Journal of Biological Chemistry*, *255*(13), 6018–6019. Retrieved from http://www.jbc.org/content/255/13/6018.full.pdf.

Fujita, T., Aoyagi, H., Ogbonna, J. C., & Tanaka, H. (2008). Effect of mixed organic substrate on α-tocopherol production by Euglena gracilis in photoheterotrophic culture. *Applied Microbiology and Biotechnology*, *79*(3), 371–378. http://dx.doi.org/10.1007/s00253-008-1443-0.

Fujita, T., Ogbonna, J. C., Tanaka, H., & Aoyagi, H. (2009). Effects of reactive oxygen species on α-tocopherol production in mitochondria and chloroplasts of Euglena gracilis. *Journal of Applied Phycology*, *21*(2), 185–191. http://dx.doi.org/10.1007/s10811-008-9349-x.

Geimer, S., Belicová, A., Legen, J., Sláviková, S., Herrmann, R. G., & Krajcovic, J. (2009). Transcriptome analysis of the Euglena gracilis plastid chromosome. *Current Genetics*, *55*(4), 425–438. http://dx.doi.org/10.1007/s00294-009-0256-8.

Gibbs, S. P. (1970). The comparative ultrastructure of the algal chloroplast. *Annals of the New York Academy of Sciences*, *175*(1), 454–473. http://dx.doi.org/10.1111/j.1749-6632.1970.tb45167.x.

Gibbs, S. P. (1978). The chloroplasts of Euglena may have evolved from symbiotic green algae. *Canadian Journal of Botany*, *56*(22), 2883–2889. http://dx.doi.org/10.1139/b78-345.

Gibbs, S. P. (1981). The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. *Annals of the New York Academy of Sciences*, *361*, 193–208. [1 Origins and E] http://dx.doi.org/10.1111/j.1749-6632.1981.tb54365.x.

Gibor, A., & Granick, S. (1962). The plastid system of normal and bleached Euglena gracilis. *The Journal of Protozoology*, *9*(3), 327–334.

Gockel, G., & Hachtel, W. (2000). Complete gene map of the plastid genome of the nonphotosynthetic Euglenoid flagellate Astasia longa. *Protist*, *151*(4), 347–351. http://dx.doi.org/10.1078/S1434-4610(04)70033-4.

Gomez-Silva, B., Timko, M. P., & Schiff, J. A. (1985). Chlorophyll biosynthesis from glutamate or 5-aminolevulinate in intact Euglena chloroplasts. *Planta*, *165*(1), 12–22. http://dx.doi.org/10.1007/BF00392206.

Gonzalez-Moreno, S., Gomez-Barrera, J., Perales, H., & Moreno-Sanchez, R. (1997). Multiple effects of salinity on photosynthesis of the protist Euglena gracilis. *Physiologia Plantarum*, *101*(4), 777–786. http://dx.doi.org/10.1111/j.1399-3054.1997.tb01063.x.

Griffiths, W. T., Threlfall, D. R., & Goodwin, T. W. (1967). Nature, intracellular distribution and formation of terpenoid quinones in maize and barley shoots. *The Biochemical Journal*, *103*(2), 589–600. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1270445&tool=pmcentrez&rendertype=abstract.

Guiry, M. D., & Guiry, G. M. (2017). *AlgaeBase*. National University of Ireland, Galway: World-ide Electronic Publication.

Hadariová, L., Vesteg, M., Birčák, E., Schwartzbach, S. D., & Krajčovič, J. (2017). An intact plastid genome is essential for the survival of colorless Euglena longa but not Euglena gracilis. *Current Genetics*, *63*(2), 331–334. http://dx.doi.org/10.1007/s00294-016-0641-z.

Hallick, R. B., Hong, L., Drager, R. G., Favreau, M. R., Monfort, A., Orsat, B., et al. (1993). Complete sequence of Euglena gracilis chloroplast DNA. *Nucleic Acids Research*, *21*(15), 3537–3544. http://dx.doi.org/10.1093/nar/21.15.3537.

Hannaert, V., Saavedra, E., Duffieux, F., Szikora, J.-P., Rigden, D. J., Michels, P. A. M., et al. (2003). Plant-like traits associated with metabolism of Trypanosoma parasites. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 1067–1071. http://dx.doi.org/10.1073/pnas.0335769100.

Harris, J. (1695). Some microscopical observations of vast numbers of animalcula seen in water by John Harris, M. A. Rector of Winchelsea in Sussex, and F.R.S. *Philosophical Transactions of the Royal Society of London*, *19*(215–235), 254–259. http://dx.doi.org/10.1098/rstl.1695.0036.

Heizmann, P., Salvador, G. F., & Nigon, V. (1976). Occurrence of plastidial rRNAs and plastidial structures in bleached mutants of Euglena gracilis. *Experimental Cell Research*, *99*(2), 253–260. http://dx.doi.org/10.1016/0014-4827(76)90581-4.

Horrum, M. A., & Schwartzbach, S. D. (1980). Nutritional regulation of organelle biogenesis in euglena: Repression of chlorophyll and NADP-glyceraldehyde-3-phosphate dehydrogenase synthesis. *Plant Physiology*, *65*(2), 382–386. http://dx.doi.org/10.1104/PP.65.2.382.

Howe, C. J., Barbrook, A. C., Nisbet, R. E. R., Lockhart, P. J., & Larkum, A. W. D. (2008). The origin of plastids. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1504), 2675–2685. http://dx.doi.org/10.1098/rstb.2008.0050.

Hrdá, Š., Fousek, J., Szabová, J., Hampl, V., & Vlček, Č. (2012). The plastid genome of eutreptiella provides a window into the process of secondary endosymbiosis of plastid in euglenids. *PloS One*, 7(3), e33746. http://dx.doi.org/10.1371/journal.pone.0033746.

Inagaki, J., Fujita, Y., Hase, T., & Yamamoto, Y. (2000). Protein translocation within chloroplast is similar in Euglena and higher plants. *Biochemical and Biophysical Research Communications*, *277*(2), 436–442. http://dx.doi.org/10.1006/bbrc.2000.3702.

Inui, H., Miyatake, K., Nakano, Y., & Kitaoka, S. (1982). Wax ester fermentation in Euglena gracilis. *FEBS Letters*, *150*(1), 89–93. http://dx.doi.org/10.1016/0014-5793(82)81310-0.

Iseki, M., Matsunaga, S., Murakami, A., Ohno, K., Shiga, K., Yoshida, K., et al. (2002). A blue-light-activated adenylyl cyclase mediates photoavoidance in Euglena gracilis. *Nature*, *415*(6875), 1047–1051.

Jarvis, P., & Robinson, C. (2004). Mechanisms of protein import and routing in chloroplasts. *Current Biology: CB*, *14*(24), R1064–R1077. http://dx.doi.org/10.1016/j.cub.2004.11.049.

Jenkins, K. P., Hong, L., & Hallick, R. B. (1995). Alternative splicing of the Euglena gracilis chloroplast roaA transcript. *RNA (New York, N.Y.)*, *1*(6), 624–633.

Karaca, H., Bozkurt, O., Ozaslan, E., Baldane, S., Berk, V., Inanc, M., et al. (2014). Positive effects of oral β-glucan on mucositis and leukopenia in colorectal cancer patients receiving adjuvant FOLFOX-4 combination chemotherapy. *Asian Pacific Journal of Cancer Prevention*, *15*, 3641–3644. http://dx.doi.org/10.7314/APJCP.2014.15.8.3641.

Karnkowska, A., Bennett, M. S., Watza, D., Kim, J. I., Zakryś, B., & Triemer, R. E. (2015). Phylogenetic relationships and morphological character evolution of photosynthetic Euglenids (Excavata) inferred from taxon-rich analyses of five genes. *Journal of Eukaryotic Microbiology*, *62*(3), 362–373. http://dx.doi.org/10.1111/jeu.12192.

Kasiborski, B. A., Bennett, M. S., Linton, E. W., & Lane, C. (2016). The chloroplast genome of Phacus orbicularis (Euglenophyceae): An initial datum point for the phacaceae. *Journal of Phycology*, *52*(3), 404–411. http://dx.doi.org/10.1111/jpy.12403.

Kim, D., Filtz, M. R., & Proteau, P. J. (2004). The methylerythritol phosphate pathway contributes to carotenoid but not phytol biosynthesis in Euglena gracilis. *Journal of Natural Products*, *67*(6), 1067–1069. http://dx.doi.org/10.1021/np049892x.

Kim, J. I., Linton, E. W., & Shin, W. (2015). Taxon-rich multigene phylogeny of the photosynthetic euglenoids (Euglenophyceae). *Frontiers in Ecology and Evolution*, *3*(August), 1–11. http://dx.doi.org/10.3389/fevo.2015.00098.

Kim, J. I., & Shin, W. (2008). Phylogeny of the euglenales inferred from plastid LSU rDNA sequences. *Journal of Phycology*, *44*(4), 994–1000. http://dx.doi.org/10.1111/j.1529-8817.2008.00536.x.

Kirk, J. T. O. (1970). Biochemical aspects of chloroplast development. *Annual Review of Plant Physiology*, *21*(1), 11–42. http://dx.doi.org/10.1146/annurev.pp.21.060170.000303.

Kogan, G., Skorik, Y., Zitnanová, I., Krizkova, L., Durackova, Z., Gomes, C., et al. (2004). Antioxidant and antimutagenic activity of N-(2-carboxyethyl)chitosan. *Toxicology and Applied Pharmacology*, *201*(3), 303–310. http://dx.doi.org/10.1016/j.taap.2004.05.009.

Koller, B., & Delius, H. (1982). Origin of replication in chloroplast DNA of Euglena gracilis located close to the region of variable size. *The EMBO Journal*, *1*(8), 995–998.

Kořený, L., & Oborník, M. (2011). Sequence evidence for the presence of two tetrapyrrole pathways in Euglena gracilis. *Genome Biology and Evolution*, *3*(1), 359–364. http://dx.doi.org/10.1093/gbe/evr029.

Koritala, S. (1989). Microbiological synthesis of wax esters by Euglena gracilis. *Journal of the American Oil Chemists Society*, *66*(1), 133–134. http://dx.doi.org/10.1007/BF02661801.

Krajčovič, J., Ebringer, L., & Polónyi, J. (1989). Quinolones and coumarins eliminate chloroplasts from Euglena gracilis. *Antimicrobial Agents and Chemotherapy*, *33*(11), 1883–1889. http://dx.doi.org/10.1128/AAC.33.11.1883.

Krajčovič, J., Ebringer, L., & Schwartzbach, S. D. (2001). Reversion of endosymbiosis? In *Symbiosis* (pp. 185–206). Dordrecht: Kluwer Academic Publishers. http://dx.doi.org/10.1007/0-306-48173-1_11.

Krajčovič, J., Vacula, R., Steiner, J. M., Löffelhardt, W., Belicová, A., Sláviková, S., et al. (1999). Molecular effects of some stress factors on the chloroplast genetic apparatus of the flagellate Euglena gracilis. In *The chloroplast: From molecular biology to biotechnology* (pp. 121–128). Dordrecht: Springer Netherlands. http://dx.doi.org/10.1007/978-94-011-4788-0_19.

Krajčovič, J., Vesteg, M., & Schwartzbach, S. D. (2015). Euglenoid flagellates: A multifaceted biotechnology platform. *Journal of Biotechnology*, *202*, 135–145. http://dx.doi.org/10.1016/j.jbiotec.2014.11.035.

Kraus, B. L., & Spremulli, L. L. (1988). Evidence for the nuclear location of the genes for chloroplast IF-2 and IF-3 in Euglena. *Plant Physiology*, *88*(4), 993–995. http://dx.doi.org/10.1104/PP.88.4.993.

Krinsky, N. I., & Goldsmith, T. H. (1960). The carotenoids of the flagellated alga, Euglena gracilis. *Archives of Biochemistry and Biophysics*, *91*(12), 271–279. http://dx.doi.org/10.1016/0003-9861(60)90501-4.

Križková, L., Ďuračková, Z., Šandula, J., Sasinková, V., & Krajčovič, J. (2001). Antioxidative and antimutagenic activity of yeast cell wall mannans in vitro. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, *497*(1–2), 213–222. http://dx.doi.org/10.1016/S1383-5718(01)00257-1.

Križková, L., Mučaji, P., Nagy, M., & Krajčovič, J. (2004). Triterpenoid cynarasaponins from Cynara cardunculus L. reduce chemically induced mutagenesis in vitro. *Phytomedicine*, *11*(7–8), 673–678. http://dx.doi.org/10.1016/j.phymed.2003.09.001.

Križková, L., Nagy, M., Polónyi, J., & Ebringer, L. (1998). The effect of flavonoids on ofloxacin-induced mutagenicity in Euglena gracilis. *Mutation Research/Genetic Toxicology*

and Environmental Mutagenesis, *416*(1–2), 85–92. http://dx.doi.org/10.1016/S1383-5718(98)00080-1.

Križková, L., Žitňanová, I., Mislovičová, D., Masárová, J., Sasinková, V., Ďuračková, Z., et al. (2006). Antioxidant and antimutagenic activity of mannan neoglycoconjugates: Mannan–human serum albumin and mannan–penicillin G acylase. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, *606*(1–2), 72–79. http://dx.doi.org/10.1016/j.mrgentox.2006.03.003.

Kuda, T., Enomoto, T., & Yano, T. (2009). Effects of two storage β-1,3-glucans, laminaran from Eicenia bicyclis and paramylon from Euglena gracili, on cecal environment and plasma lipid levels in rats. *Journal of Functional Foods*, *1*, 399–404. http://dx.doi.org/10.1016/j.jff.2009.08.003.

Lakey, B., & Triemer, R. (2016). The tetrapyrrole synthesis pathway as a model of horizontal gene transfer in euglenoids. *Journal of Phycology*, *53*, 198–217. http://dx.doi.org/10.1111/jpy.12491.

Larkum, A. W. D., Lockhart, P. J., & Howe, C. J. (2007). Shopping for plastids. *Trends in Plant Science*, *12*(5), 189–195. http://dx.doi.org/10.1016/j.tplants.2007.03.011.

Lax, G., & Simpson, A. G. B. (2013). Combining molecular data with classical morphology for uncultured phagotrophic euglenids (Excavata): A single-cell approach. *Journal of Eukaryotic Microbiology*, *60*(6), 615–625. http://dx.doi.org/10.1111/jeu.12068.

Leander, B. S. (2004). Did trypanosomatid parasites have photosynthetic ancestors? *Trends in Microbiology*, *12*(6), 251–258. http://dx.doi.org/10.1016/j.tim.2004.04.001.

Leander, B. S., Esson, H. J., & Breglia, S. A. (2007). Macroevolution of complex cytoskeletal systems in euglenids. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *29*(10), 987–1000. http://dx.doi.org/10.1002/bies.20645.

Leander, B. S., Triemer, R. E., & Farmer, M. a. (2001). Character evolution in heterotrophic euglenids. *European Journal of Protistology*, *37*, 337–356.

Leedale, G. F. (1967). *Euglenoid flagellates*. Englewood Cliffs, NJ: Prentice-Hall.

Lefort-Tran, M. (1981). The triple layered organization of the Euglena chloroplast envelope (signification and functions). *Plant Biology*, *94*(1), 463–476.

Liljenberg, C., & Selstam, E. (1980). Interactions of chlorophyll a and terpenoid alcohols with chloroplast acyl lipids in monomolecular films. *Physiologia Plantarum*, *48*(1966), 428–434.

Linton, E. W., Karnkowska-Ishikawa, A., Kim, J. I., Shin, W., Bennett, M. S., Kwiatowski, J., et al. (2010). Reconstructing Euglenoid evolutionary relationships using three genes: Nuclear SSU and LSU, and chloroplast SSU rDNA sequences and the description of Euglenaria gen.nov. (Euglenophyta). *Protist*, *161*(4), 603–619. http://dx.doi.org/10.1016/j.protis.2010.02.002.

Mahmoudi, N., Ciceron, L., Franetich, J.-F., Farhati, K., Silvie, O., Eling, W., et al. (2003). In vitro activities of 25 quinolones and fluoroquinolones against liver and blood stage Plasmodium spp. *Antimicrobial Agents and Chemotherapy*, *47*(8), 2636–2639. http://dx.doi.org/10.1128/AAC.47.8.2636-2639.2003.

Maier, U. G., Zauner, S., & Hempel, F. (2015). Protein import into complex plastids: Cellular organization of higher complexity. *European Journal of Cell Biology*, *94*(7–9), 340–348. http://dx.doi.org/10.1016/j.ejcb.2015.05.008.

Marin, B. (2004). Origin and fate of chloroplasts in the Euglenoida. *Protist*, *155*(March), 13–14. http://dx.doi.org/10.1078/1434461000159.

Marin, B., Nowack, E. C. M., & Melkonian, M. (2005). A plastid in the making: Evidence for a second primary endosymbiosis. *Protist*, *156*(4), 425–432. http://dx.doi.org/10.1016/j.protis.2005.09.001.

Marin, B., Palm, A., Klingberg, M., & Melkonian, M. (2003). 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*(1), 99–145. http://dx.doi.org/10.1078/143446103764928521.

Markunas, C. M., & Triemer, R. E. (2016). Evolutionary history of the enzymes involved in the Calvin-Benson cycle in Euglenids. *Journal of Eukaryotic Microbiology*, *63*(3), 326–339. http://dx.doi.org/10.1111/jeu.12282.

Martin, W., & Borst, P. (2003). Secondary loss of chloroplasts in trypanosomes. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(3), 765–767. http://dx.doi.org/10.1073/pnas.0437776100.

Maruyama, S., Matsuzaki, M., Misawa, K., & Nozaki, H. (2009). Cyanobacterial contribution to the genomes of the plastid-lacking protists. *BMC Evolutionary Biology*, *9*(1), 197. http://dx.doi.org/10.1093/bioinformatics/17.8.754.

Maruyama, S., Suzaki, T., Weber, A. P. M., Archibald, J. M., & Nozaki, H. (2011). Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. *BMC Evolutionary Biology*, *11*(1), 105. http://dx.doi.org/10.1186/1471-2148-11-105.

Matson, R. S., Fei, M., & Chang, S. B. (1970). Comparative studies of biosynthesis of galactolipids in Euglena-gracilis strain-Z. *Plant Physiology*, *45*(4), 531.

Monfils, A. K., Triemer, R. E., & Bellairs, E. F. (2011). Characterization of paramylon morphological diversity in photosynthetic euglenoids (Euglenales, Euglenophyta). *Phycologia*, *50*(2), 156–169. http://dx.doi.org/10.2216/09-112.1.

Monroy, A. F., & Schwartzbach, S. D. (1984). Catabolite repression of chloroplast development in Euglena. *Cell Biology*, *81*, 2786–2790. Retrieved from http://www.pnas.org/content/81/9/2786.full.pdf.

Morales, J., Hashimoto, M., Williams, T. A., Hirawake-Mogi, H., Makiuchi, T., Tsubouchi, A., et al. (2016). Differential remodelling of peroxisome function underpins the environmental and metabolic adaptability of diplonemids and kinetoplastids. *Proceedings of the Royal Society of London B: Biological Sciences*, *283*(1830). Retrieved from http://rspb.royalsocietypublishing.org/content/283/1830/20160520.long.

Morden, C. W., & Golden, S. S. (1991). Sequence analysis and phylogenetic reconstruction of the genes encoding the large and small subunits of ribulose-1,5-bisphosphate carboxylase/oxygenase from the chlorophyllb-containing prokaryoteProchlorothrix hollandica. *Journal of Molecular Evolution*, *32*(5), 379–395. http://dx.doi.org/10.1007/BF02101278.

Mullner, A. N., Angeler, D. G., Samuel, R., Linton, E. W., & Triemer, R. E. (2001). Phylogenetic analysis of phagotrophic, phototrophic and osmotrophic euglenoids by using the nuclear 18S rDNA sequence. *International Journal of Systematic and Evolutionary Microbiology*, *51*(3), 783–791. http://dx.doi.org/10.1099/00207713-51-3-783.

Nowack, E. C. M., Price, D. C., Bhattacharya, D., Singer, A., Melkonian, M., & Grossman, A. R. (2016). Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of Paulinella chromatophora. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(43), 12214–12219. http://dx.doi.org/10.1073/pnas.1608016113.

O'Neill, E. C., Trick, M., Henrissat, B., & Field, R. A. (2015). Euglena in time: Evolution, control of central metabolic processes and multi-domain proteins in carbohydrate and natural product biochemistry. *Perspectives in Science*, *6*, 84–93. http://dx.doi.org/10.1016/j.pisc.2015.07.002.

Ogawa, T., Tamoi, M., Kimura, A., Mine, A., Sakuyama, H., Yoshida, E., et al. (2015). Enhancement of photosynthetic capacity in Euglena gracilis by expression of cyanobacterial fructose-1,6-/sedoheptulose-1,7-bisphosphatase leads to increases in biomass and wax ester production. *Biotechnology for Biofuels*, *8*(1), 80. http://dx.doi.org/10.1186/s13068-015-0264-5.

Osafune, T., & Schiff, J. A. (1980a). Events surrounding the early development of Euglena chloroplasts. *Journal of Ultrastructure Research*, *73*(1), 64–76. http://dx.doi.org/10.1016/0022-5320(80)90116-1.

Osafune, T., & Schiff, J. A. (1980b). Stigma and flagellar swelling in relation to light and carotenoids in Euglena gracilis var. bacillaris. *Journal of Ultrasructure Research*, *73*(3), 336–349. http://dx.doi.org/10.1016/S0022-5320(80)90093-3.

Osafune, T., & Schiff, J. A. (1983). W10BSmL, a mutant of Euglena gracilis var. bacillaris lacking plastids. *Experimental Cell Research*, *148*(2), 530–535. http://dx.doi.org/10.1016/0014-4827(83)90176-3.

Osafune, T., Schiff, J. A., & Hase, E. (1987). Light-independent and dependent phases of proplastid development in Euglena gracilis W3BUL. *Cell Structure and Function*, *12*(5), 453–461. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3119230.

Parthier, B., & Neumann, D. (1977). Structural and functional analysis of some plastid mutants of Euglena gracilis. *Biochemie und Physiologie der Pflanzen*, *171*(6), 547–562. http://dx.doi.org/10.1016/S0015-3796(17)30349-9.

Polónyi, J., Ebringer, L., Dobias, J., & Krajčovič, J. (1998). Giant mitochondria in chloroplast-deprived Euglena gracilis late after N-succinimidylofloxacin treatment. *Folia Microbiologica*, *43*(6), 661–666. http://dx.doi.org/10.1007/BF02816386.

Polónyi, J., Ebringer, L., Krajcovic, J., & Kapeller, K. (1990). Injured mitochondria in cells of Euglena gracilis after DNA gyrase inhibitors treatment. *Zeitschrift fur Mikroskopisch-Anatomische Forschung*, *104*(1), 61–78. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2161591.

Pombert, J. F., James, E. R., Janouškovec, J., & Keeling, P. J. (2012). Evidence for transitional stages in the evolution of Euglenid group II introns and twintrons in the Monomorphina aenigmatica plastid genome. *PloS One*, 7(12). http://dx.doi.org/10.1371/journal.pone.0053433.

Ravel-Chapuis, P., Heizmann, P., & Nigon, V. (1982). Electron microscopic localization of the replication origin of Euglena gracilis chloroplast DNA. *Nature*, *300*(5887), 78–81. http://dx.doi.org/10.1038/300078a0.

Richter, P., Helbling, W., Streb, C., & Häder, D.-P. (2007). PAR and UV effects on vertical migration and photosynthesis in Euglena gracilis? *Photochemistry and Photobiology*, *83*(4), 818–823. http://dx.doi.org/10.1111/j.1751-1097.2007.00134.x.

Rodríguez-Zavala, J. S., Ortiz-Cruz, M. A., Mendoza-Hernández, G., & Moreno-Sánchez, R. (2010). Increased synthesis of α-tocopherol, paramylon and tyrosine by Euglena gracilis under conditions of high biomass production. *Journal of Applied Microbiology*, *109*(6), 2160–2172. http://dx.doi.org/10.1111/j.1365-2672.2010.04848.x.

Roos, D. S., & Fichera, M. E. (1997). A plastid organelle as a drug target in apicomplexan parasites. *Nature*, *390*(6658), 407–409. http://dx.doi.org/10.1038/37132.

Rosenberg, A., & Gouaux, J. (1967). Quantitative and compositional changes in monogalactosyl and digalactosyl diglycerides during light-induced formation of chloroplasts in Euglena gracilis. *Journal of Lipid Research*, *8*, 80–83.

Rosenberg, A., Gouaux, J., & Milch, P. (1966). Monogalactosyl and digalactosyl diglycerides from heterotrophic, hetero-autotrophic, and photobiotic Euglena gracilis. 7, 733–738.

Rybicka, K. K. (1996). Glycosomes—The organelles of glycogen metabolism. *Tissue and Cell*, *28*(3), 253–265. http://dx.doi.org/10.1016/S0040-8166(96)80013-9.

Sagan, L. (1993). On the origin of mitosing cells. 1967. *The Journal of NIH Research: Life Sciences Research and News about the National Institutes of Health and the Alcohol, Drug Abuse, and Mental Health Administration*, *5*(3), 65–72. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11541390.

Saidha, T., & Schiff, J. A. (1989). The role of mitochondria in sulfolipid biosynthesis by Euglena chloroplasts. *Biochimica et Biophysica Acta (BBA)—Lipids and Lipid Metabolism*, *1001*(3), 268–273. http://dx.doi.org/10.1016/0005-2760(89)90110-0.

Sanchez-Puerta, M. V., Lippmeier, J. C., Apt, K. E., & Delwiche, C. F. (2007). Plastid genes in a non-photosynthetic dinoflagellate. *Protist*, *158*(1), 105–117. http://dx.doi.org/10.1016/j.protis.2006.09.004.

Šantek, B., Felski, M., Friehs, K., Lotz, M., & Flaschel, E. (2010). Production of paramylon, a β-1,3-glucan, by heterotrophic cultivation of Euglena gracilis on potato liquor. *Engineering in Life Sciences*, *10*(2), 165–170. http://dx.doi.org/10.1002/elsc.200900077.

Schiff, J. A., Lyman, H., & Russell, G. K. (1971). Isolation of mutants from Euglena gracilis. *Methods in Enzymology*, *23*, 143–162. http://dx.doi.org/10.1016/S0076-6879(71)23088-3.

Schneider, T., & Betz, A. (1985). Waxmonoester fermentation in Euglena gracilis T. Factors favouring the synthesis of odd-numbered fatty acids and alcohols. *Planta*, *166*(1), 67–73. http://dx.doi.org/10.1007/BF00397387.

Schwartzbach, S. D. (2017). Photo and nutritional regulation of Euglena organelle development. In S. D. Schwartzbach & S. Shigeoka (Eds.), *Euglena: Biochemistry, cell and molecular biology* (pp. 159–182). Cham, Switzerland: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-54910-1_9.

Schwartzbach, S. D., & Schiff, J. A. (1974). Chloroplast and cytoplasmic ribosomes of Euglena: Selective binding of dihydrostreptomycin to chloroplast ribosomes. *Journal of Bacteriology*, *120*(1), 334–341.

Schwartzbach, S. D., Schiff, J. A., & Goldstein, N. H. (1975). Events surrounding the early development of euglena chloroplasts: V. Control of paramylum degradation. *Plant Physiology*, *56*(2), 313–317. http://dx.doi.org/10.1111/j.1751-1097.1976.tb06873.x.

Schwender, J., Goffman, F., Ohlrogge, J. B., & Shachar-Hill, Y. (2004). Rubisco without the Calvin cycle improves the carbon efficiency of developing green seeds. *Nature*, *432*(7018), 779–782. http://dx.doi.org/10.1038/nature03145.

Shashidhara, L. S. S., Limsli, S. H., Shackletonii, J. B., Robinsonii, C., & Smith, A. G. (1992). Protein targeting across the three membranes of the Euglena chloroplast envelope. *The Journal of Biological Chemistry*, *267*, 12885–12891.

Sheiner, L., & Striepen, B. (2013). Protein sorting in complex plastids. *Biochimica et Biophysica Acta—Molecular Cell Research*, *1833*(2), 352–359. http://dx.doi.org/10.1016/j.bbamcr.2012.05.030.

Sheveleva, E. V., & Hallick, R. B. (2004). Recent horizontal intron transfer to a chloroplast genome. *Nucleic Acids Research*, *32*(2), 803–810. http://dx.doi.org/10.1093/nar/gkh225.

Shi, L.-X., & Theg, S. M. (2013). The chloroplast protein import system: From algae to trees. *Biochimica et Biophysica Acta*, *1833*(2), 314–331. http://dx.doi.org/10.1016/j.bbamcr.2012.10.002.

Shigeoka, S., Onishi, T., Nakano, Y., & Kitaoka, S. (1986). The contents and subcellular distribution of tocopherols in Euglena gracilis. *Agricultural and Biological Chemistry*, *50*(4), 1063–1065. http://dx.doi.org/10.1271/bbb1961.50.1063.

Sláviková, S., Vacula, R., Fang, Z., Ehara, T., Osafune, T., & Schwartzbach, S. D. (2005). Homologous and heterologous reconstitution of Golgi to chloroplast transport and protein import into the complex chloroplasts of Euglena. *Journal of Cell Science*, *118*(Pt. 8), 1651–1661. http://dx.doi.org/10.1242/jcs.02277.

Stern, A. I., Epstein, H. T., & Schiff, J. A. (1964a). Studies of chloroplast development in Euglena. VI. Light intensity as a controlling factor in development. *Plant Physiology*, *39*(2), 226–231.

Stern, A. I., Schiff, J. A., & Epstein, H. T. (1964b). Studies of chloroplast development in Euglena. V. Pigment biosynthesis, photosynthetic oxygen evolution and carbon dioxide fixation during chloroplast development. *Plant Physiology*, *39*(2), 220–226. http://dx.doi.org/10.1104/pp.39.2.220.

Sugiyama, A., Suzuki, K., Mitra, S., Arashida, R., Yoshida, E., Nakano, R., et al. (2009). Hepatoprotective effects of paramylon, a beta-1, 3-D-glucan isolated from Euglena gracilis Z, on acute liver injury induced by carbon tetrachloride in rats. *The Journal of Veterinary Medical Science the Japanese Society of Veterinary Science*, *71*(7), 885–890. http://dx.doi.org/10.1292/jvms.71.885.

Sulli, C., Fang, Z. W., Muchhal, U., & Schwartzbach, S. D. (1999). Topology of Euglena chloroplast protein precursors within endoplasmic reticulum to Golgi to chloroplast transport vesicles. *Journal of Biological Chemistry*, *274*(1), 457–463. http://dx.doi.org/10.1074/jbc.274.1.457.

Sulli, C., & Schwartzbach, S. D. (1995). The polyprotein precursor to the Euglena light-harvesting chlorophyll a/b-binding protein is transported to the Golgi apparatus prior to chloroplast import and polyprotein processing. *Journal of Biological Chemistry*, *270*, 13084–13090. http://dx.doi.org/10.1074/jbc.270.22.13084.

Sulli, C., & Schwartzbach, S. D. (1996). A soluble protein is imported into Euglena chloroplasts as a membrane-bound precursor. *The Plant Cell*, *8*(1), 43–53.

Teerawanichpan, P., & Qiu, X. (2010). Fatty acyl-CoA reductase and wax synthase from Euglena gracilis in the biosynthesis of medium-chain wax esters. *Lipids*, *45*(3), 263–273. http://dx.doi.org/10.1007/s11745-010-3395-2.

Thomas, B. G., & Threlfall, D. R. (1974). Synthesis of polyprenyltoluquinols from homogentisate and polyprenyl pyrophosphates in particulate fractions of Euglena and sugar beet. *Biochemical Journal*, *142*, 437–440.

Thompson, M. D., Copertino, D. W., Thompson, E., Favreau, M. R., & Hallick, R. B. (1995). Evidence for the late origin of introns in chloroplast genes from an evolutionary analysis of the genus Euglena. *Nucleic Acids Research*, *23*(23), 4745–4752. http://dx.doi.org/10.1093/nar/23.23.4745.

Triemer, R. E., Linton, E., Shin, W., Nudelman, A., Monfils, A., Bennett, M., et al. (2006). Phylogeny of the euglenales based upon combined SSU and LSU rDNA sequence comparisons and description of Discoplastis gen. nov. (Euglenophyta). *Journal of Phycology*, *42*(3), 731–740. http://dx.doi.org/10.1111/j.1529-8817.2006.00219.x.

Turmel, M., Gagnon, M.-C., O'Kelly, C. J., Otis, C., & Lemieux, C. (2009). 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. *Molecular Biology and Evolution*, *26*(3), 631–648. http://dx.doi.org/10.1093/molbev/msn285.

Vesteg, M., Vacula, R., Steiner, J. M., Mateásiková, B., Löffelhardt, W., Brejová, B., et al. (2010). A possible role for short introns in the acquisition of stroma-targeting peptides in the flagellate Euglena gracilis. *DNA Research: An International Journal for Rapid Publication of Reports on Genes and Genomes*, *17*(4), 223–231. http://dx.doi.org/10.1093/dnares/dsq015.

Walne, P. L., & Arnott, H. J. (1967). The comparative ultrastructure and possible function of eyespots: Euglena granulata and Chlamydomonas eugametos. *Planta*, 77(4), 325–353. http://dx.doi.org/10.1007/BF00389319.

Wang, J., Shi, Z., & Xu, X. (2004). Residual plastids of bleached mutants of Euglena gracilis and their effects on the expression of nucleus-encoded genes. *Progress in Natural Science*, *14*(3), 213–217. http://dx.doi.org/10.1080/10020070412331343371.

Watanabe, T., Shimada, R., Matsuyama, A., Yuasa, M., Sawamura, H., Yoshida, E., et al. (2013). Antitumor activity of the β-glucan paramylon from Euglena against preneoplastic colonic aberrant crypt foci in mice. *Food & Function*, *4*(11), 1685. http://dx.doi.org/10.1039/c3fo60256g.

Wheeler, G., Ishikawa, T., Pornsaksit, V., & Smirnoff, N. (2015). Evolution of alternative biosynthetic pathways for vitamin C following plastid acquisition in photosynthetic eukaryotes. *eLife*, *4*, 1–25. http://dx.doi.org/10.7554/eLife.06369.

Wiegert, K. E., Bennett, M. S., & Triemer, R. E. (2012). Evolution of the chloroplast genome in photosynthetic euglenoids: A comparison of Eutreptia viridis and Euglena gracilis (Euglenophyta). *Protist*, *163*(6), 832–843. http://dx.doi.org/10.1016/j.protis.2012.01.002.

Wiegert, K. E., Bennett, M. S., & Triemer, R. E. (2013). Tracing patterns of chloroplast evolution in euglenoids: Contributions from colacium vesiculosum and strombomonas acuminata (euglenophyta). *Journal of Eukaryotic Microbiology*, *60*(2), 214–221. http://dx.doi.org/10.1111/jeu.12025.

Wolfe, A. D., & dePamphilis, C. W. (1998). The effect of relaxed functional constraints on the photosynthetic gene rbcL in photosynthetic and nonphotosynthetic parasitic plants. *Molecular Biology and Evolution*, *15*(10), 1243–1258. http://dx.doi.org/10.1093/oxfordjournals.molbev.a025853.

Wolken, J. J., & Mellon, A. D. (1956). The relationship between chlorophyll and the carotenoids in the algal flagellate, Euglena. *The Journal of General Physiology*, *39*(5), 675–685.

Yamaguchi, A., Yubuki, N., & Leander, B. S. (2012). Morphostasis in a novel eukaryote illuminates the evolutionary transition from phagotrophy to phototrophy: Description of Rapaza viridis n. gen. et sp. (Euglenozoa, Euglenida). *BMC Evolutionary Biology*, *12*(1), 29. http://dx.doi.org/10.1186/1471-2148-12-29.

Yoshida, Y., Tomiyama, T., Maruta, T., Tomita, M., Ishikawa, T., & Arakawa, K. (2016). De novo assembly and comparative transcriptome analysis of Euglena gracilis in response to anaerobic conditions. *BMC Genomics*, *17*(1), 182. http://dx.doi.org/10.1186/s12864-016-2540-6.

Záhonová, K., Füssy, Z., Oborník, M., Eliáš, M., Yurchenko, V., & Stern, D. (2016). RuBisCO in non-photosynthetic alga Euglena Longa: Divergent features, transcriptomic analysis and regulation of complex formation. *PloS One*, *11*(7), e0158790. http://dx.doi.org/10.1371/journal.pone.0158790.

Záhonová, K., Hadariová, L., Vacula, R., Yurchenko, V., Eliáš, M., Krajčovič, J., et al. (2014). A small portion of plastid transcripts is polyadenylated in the flagellate Euglena gracilis. *FEBS Letters*, *588*(5), 783–788. http://dx.doi.org/10.1016/j.febslet.2014.01.034.

Zakryś, B., Milanowski, R., & Karnkowska, A. (2017). Evolutionary origin of Euglena. In S. D. Schwartzbach & S. Shigeoka (Eds.), *Euglena: Biochemistry, cell and molecular biology* (pp. 3–17). Cham, Switzerland: Springer International Publishing. http://dx.doi.org/10.1007/978-3-319-54910-1_1.

SCIENTIFIC REPORTS

OPEN

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

Received: 26 July 2018
Accepted: 2 November 2018
Published online: 19 November 2018

Kristína Záhonová[1], Zoltán Füssy[2], Erik Birčák[3], Anna M. G. Novák Vanclová[4], Vladimír Klimeš[1], Matej Vesteg[5], Juraj Krajčovič[6], Miroslav Oborník[2,7] & Marek Eliáš[1]

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 Euglenozoa[1]. 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 (ACTTTCTGAGTGTCTATTTTTTTTCG 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 paper[8]. 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 project[11], but no specific analyses of this data resource have been reported. Hence, further

[1]Life Science Research Centre, Department of Biology and Ecology and Institute of Environmental Technologies, Faculty of Science, University of Ostrava, 701 00, Ostrava, Czech Republic. [2]Institute of Parasitology, Biology Centre CAS, 370 05, České Budějovice, Czech Republic. [3]Department of Genetics, Faculty of Natural Sciences, Comenius University, 842 15, Bratislava, Slovakia. [4]Department of Parasitology, Faculty of Science, Charles University, BIOCEV, Prague, Czech Republic. [5]Department of Biology and Ecology, Faculty of Natural Sciences, Matej Bel University, 974 01, Banská Bystrica, Slovakia. [6]Department of Biology, Faculty of Natural Sciences, University of ss. Cyril and Methodius in Trnava, 917 01, Trnava, Slovakia. [7]University 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 variants[16]. 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 proteins[16]. 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 heterotrophic[20]. 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 ago[21]. 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 plastids[24], but this identification is uncertain given that euglenophyte plastids studied in detail possess three bounding membranes[1]. 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 assemble[8]. 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 species[23] and enabled identification of a novel Euglenozoa-specific form of the Rheb GTPase[26], 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 fibrillarin[27]. 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.

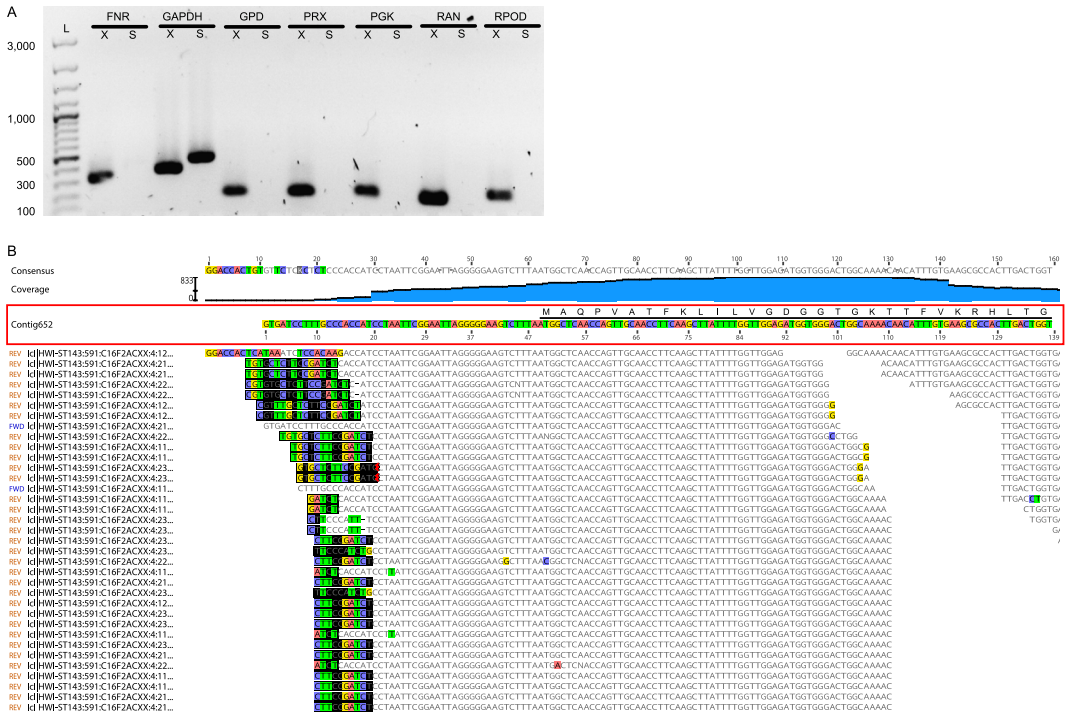

Figure 1. Presence/absence of the spliced leader (SL) in selected transcripts in *E. longa*. (**A**) PCR (with cDNA as the template) with an SL-specific forward primer was done to assay whether the lack of SL in exemplar assembled contigs (listed in Supplementary Table S2) mirrors real SL absence in 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 sequence[21] 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 b_6f 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 eukaryotes[30], 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 version[34]), 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 detail[23]. 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 glutaminyl-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 proteins[38]) 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

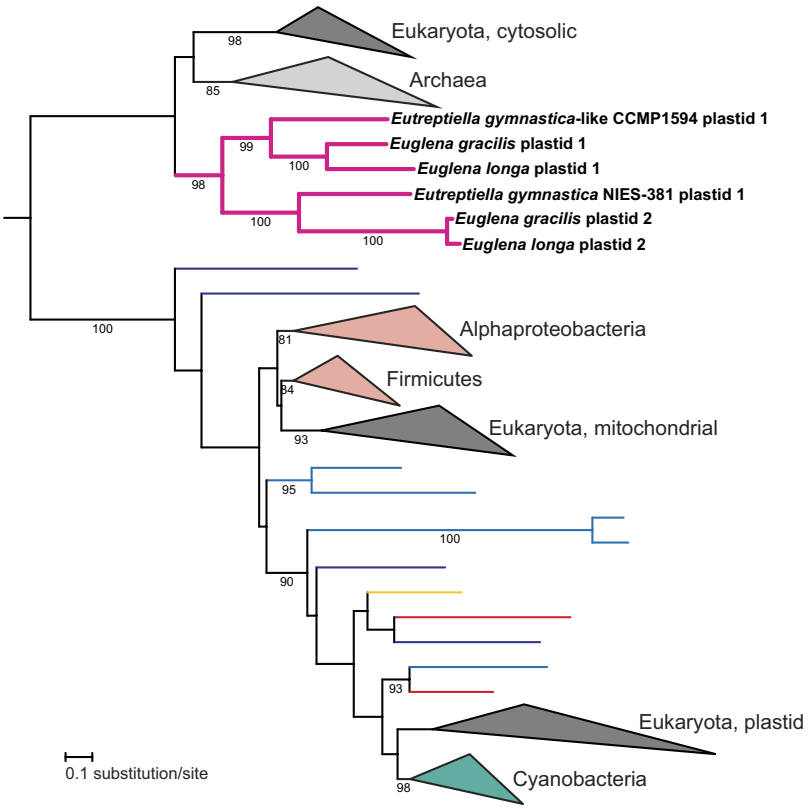

Figure 2. Phylogenetic analysis of the uL24 family of ribosomal proteins. The tree shows the phylogenetic position of the presumably plastid-localised L26-related proteins in euglenophytes. Bootstrap support values are given when ≥80.

eubacteria-type S8 protein in angiosperm mitochondria replaced by the eukaryotic protein S15A[39] 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 study[23] 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 rule[16], 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 linker[23]. 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 plastid[45], 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

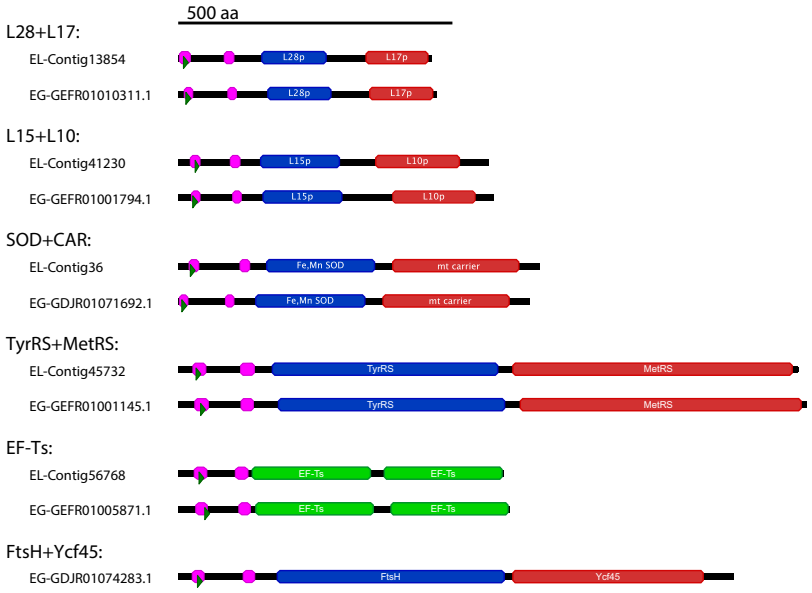

Figure 3. Domain structure of fusion proteins in *E. longa* and *E. gracilis*. Regions corresponding to separate mature proteins presumably released by processing of the fusion protein are shown as boxes in blue and red (if different) or in green (if the fusion comprises a repeat of the same monomer). 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 plastids[15]. 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 manner[50]. 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 plastids[51]. TIC55 was recently shown to serve as phyllobilin hydroxylase in the chlorophyll breakdown pathway and its role in plastid protein import was questioned[52]. 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

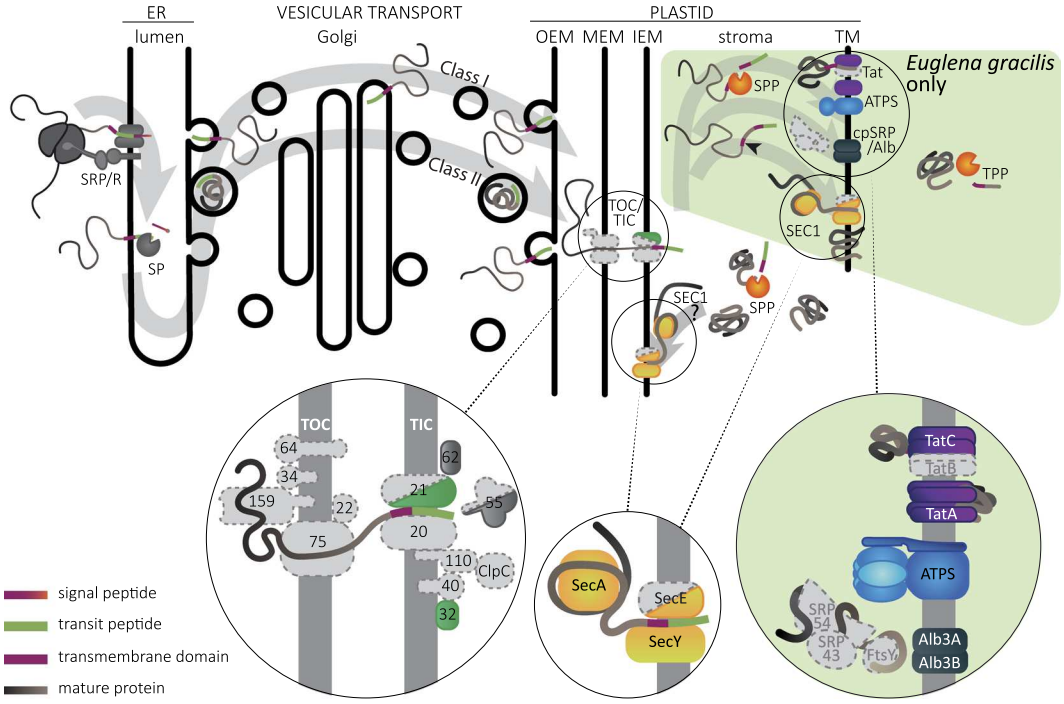

Figure 4. A hypothetical scheme of translocation of plastidial proteins in *Euglena*. The route of plastidial proteins in the *Euglena* cell is schematically depicted, with a zoom-in on individual translocon complexes and protein subunits found in the transcriptomic data analysed. Note that transport 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 SELMA[53].

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 translocase[54]. 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 membrane[55], 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

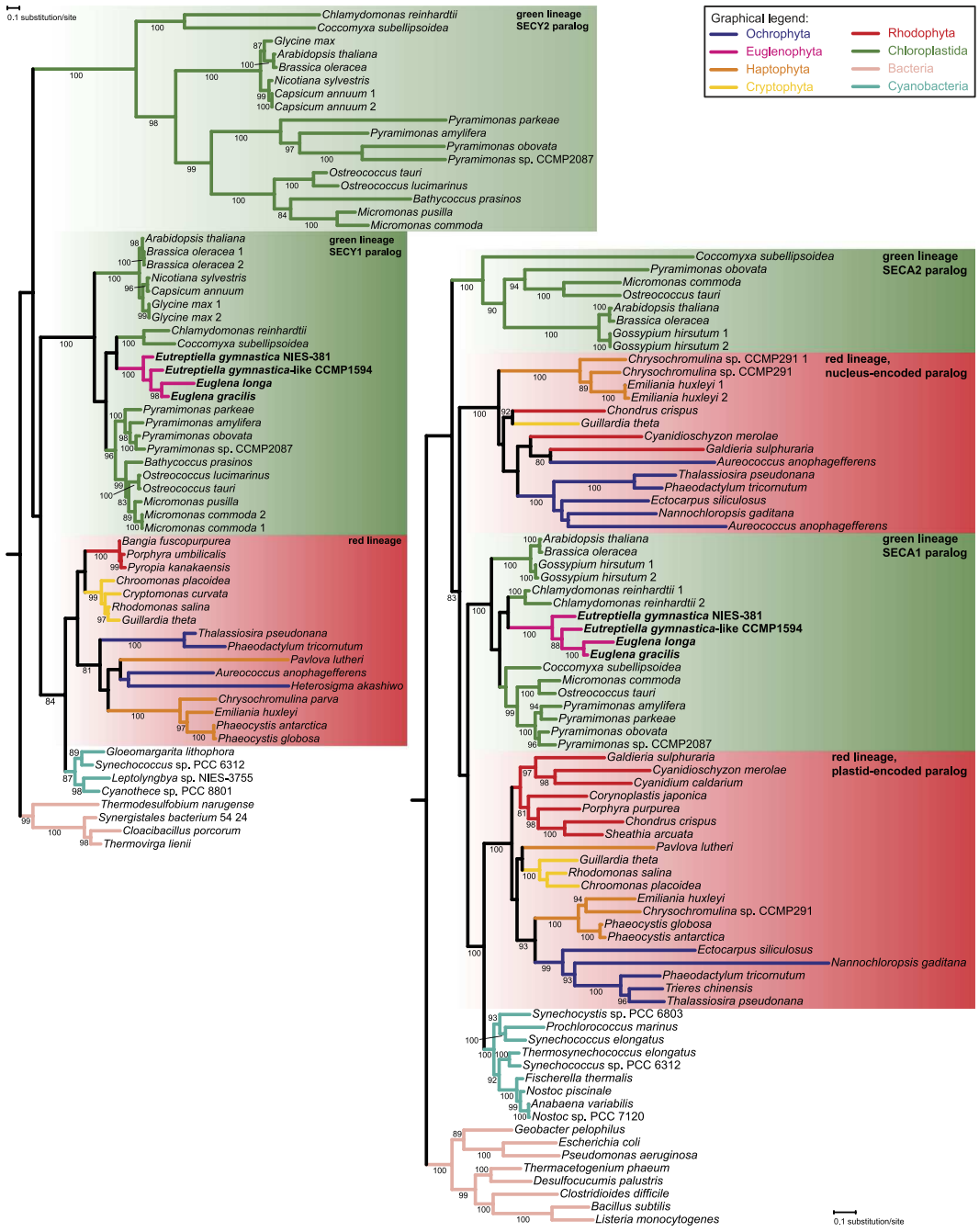

Figure 5. Phylogenetic analysis of the SecA and SecY subunits of the SEC translocon. The maximum likelihood tree of SecA and SecY proteins documents that the euglenophyte proteins are orthologs of the chlorophyte SECA1 and SECY1, respectively. Bootstrap support values are given when ≥80.

synthase at the expense of ATP. Indeed, Tat is missing from non-photosynthetic plastids that lost ATP synthase[32]. 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 insertase[54]. 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 proteases[59]. 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* homologs[60], 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 cyanobacteria[61] and glaucophytes[62]. 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 protein[54]. 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 plastid[63]. 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 myzozoans[64]. The absence of FtsZ in *E. longa* would not be particularly surprising either, since myzozoans 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 paralogs[64]. 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 general[65]. 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 machinery[66] 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 bacteria[66]. 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

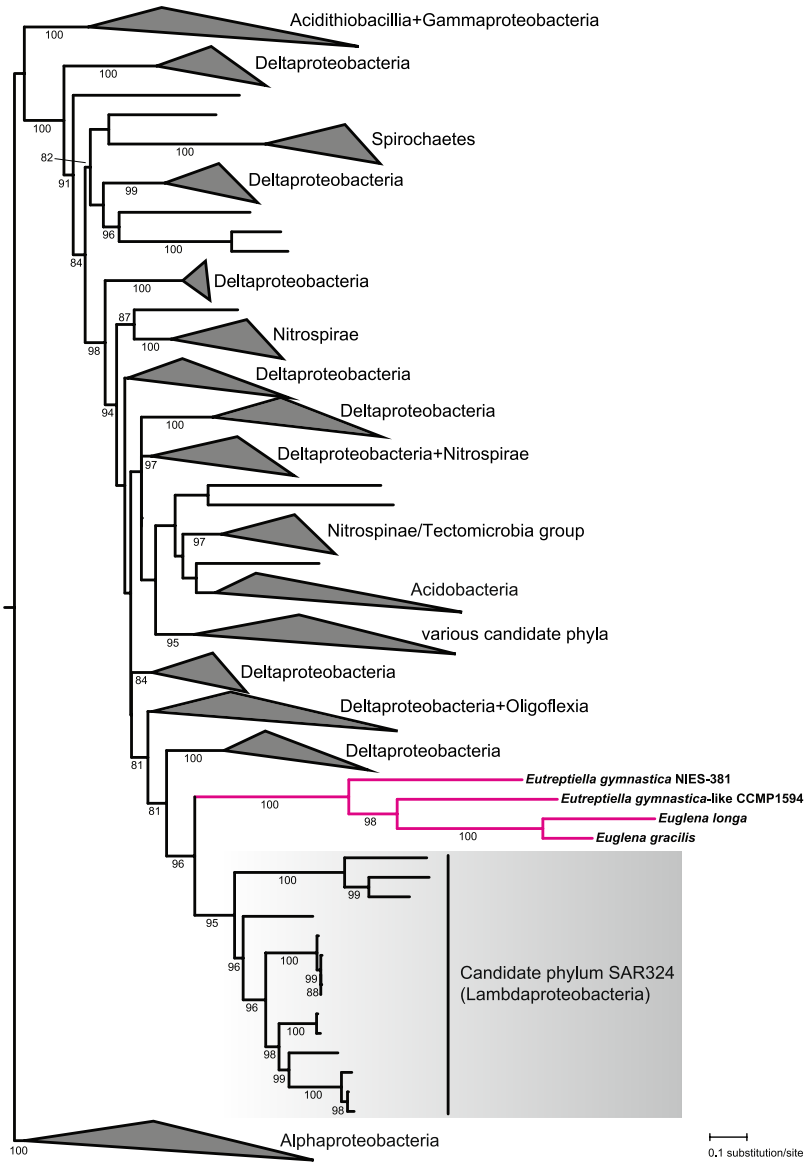

Figure 6. The phylogenetic position of the euglenophyte transcription-termination factor Rho among bacterial homologs. The tree was inferred by using the maximum likelihood method. Bootstrap support values are given when ≥80. For simplicity, various broader clades were collapsed and the taxonomic provenance 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 SAR324[67] (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 found[66]. 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 protists[72]. 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 medium[76] 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 0.43[78]. The remaining reads were adapter- and quality-trimmed by Trimmomatic 0.33[79]. The final read assembly was performed using the ABySS software 1.52[80] (k-mers 31–51), then fused with Trans-ABySS 1.48[81], Trinity r20140717[82] (k-mers 31 and 25), and SOAPdenovo-Trans 1.04[83] (k-mers 31 and 33), followed by merging the contigs (>99% sequence identity over 150 nt) using CAP3 12/21/07[84]. 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 orthologs[85].

Sequence searches and phylogenetic analyses. Homologs of proteins of interest were searched in the final transcriptome assembly using local tBLASTn[86]. 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 models[87]. 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 Database[88] (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 project[11]), 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 ClustalW[89], 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 PrediSi[90] or PredSL[91] programs; (2) one or two transmembrane domains at the N-terminus of the protein were predicted by the TMHMM program[92] available online or implemented in the Geneious 10.1.3 software[93]. The resulting set of sequences was further filtered by checking for the presence of a plastid transit peptide, which was predicted by MultiLoc2[94] 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 Project[11]; 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 tool[95] and poorly aligned positions were eliminated with the trimAL tool[96]. The alignments were manually refined using AliView[97] and ambiguously aligned positions were removed. For presentation purposes, alignments were processed using the program CHROMA[98]. Maximum likelihood (ML) trees were inferred from the alignments using the best-fitting substitution model as determined by the IQ-TREE software[99] 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

1. 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).
2. Campbell, D. A., Thomas, S. & Sturm, N. R. Transcription in kinetoplastid protozoa: why be normal? *Microbes Infect.* **5**, 1231–1240 (2003).
3. Clayton, C. E. Gene expression in kinetoplastids. *Curr Opin Microbiol.* **32**, 46–51 (2016).
4. 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).
5. 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).
6. Frantz, C., Ebel, C., Paulus, F. & Imbault, P. Characterization of *trans*-splicing in Euglenoids. *Curr Genet.* **37**, 349–355 (2000).
7. 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).
8. 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).
9. 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).
10. Yoshida, Y. *et al*. *De novo* assembly and comparative transcriptome analysis of *Euglena gracilis* in response to anaerobic conditions. *BMC Genomics.* **17**, 182 (2016).
11. 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).
12. 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).
13. 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).
14. 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).
15. Durnford, D. G. & Schwartzbach, S. D. Protein targeting to the plastid of Euglena in Euglena: Biochemistry, cell and molecular biology Vol. 979 (eds Steven D. Schwartzbach & Shigeru Shigeoka), 183–205 (Springer International Publishing, 2017).
16. 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).
17. Kořený, L. & Oborník, M. Sequence evidence for the presence of two tetrapyrrole pathways in *Euglena gracilis*. *Genome Biol Evol.* **3**, 359–364 (2011).
18. Lakey, B. & Triemer, R. The tetrapyrrole synthesis pathway as a model of horizontal gene transfer in euglenoids. *J Phycol.* **53**, 198–217 (2017).
19. 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-Martinez, 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-tRNAGln 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. Skalitzky, 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. & Moszczynski, 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. Kearse, 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.

Ebenezer et al. BMC Biology (2019) 17:11
https://doi.org/10.1186/s12915-019-0626-8

BMC Biology

RESEARCH ARTICLE Open Access

Transcriptome, proteome and draft genome of *Euglena gracilis*

ThankGod E. Ebenezer[1,2], Martin Zoltner[1], Alana Burrell[3], Anna Nenarokova[4], Anna M. G. Novák Vanclová[5], Binod Prasad[6], Petr Soukal[5], Carlos Santana-Molina[7], Ellis O'Neill[8], Nerissa N. Nankissoor[9], Nithya Vadakedath[6], Viktor Daiker[6], Samson Obado[10], Sara Silva-Pereira[11], Andrew P. Jackson[11], Damien P. Devos[7], Julius Lukeš[4], Michael Lebert[6], Sue Vaughan[3], Vladimír Hampl[5], Mark Carrington[2], Michael L. Ginger[12], Joel B. Dacks[9,13*], Steven Kelly[8*] and Mark C. Field[1,4*]

Abstract

Background: Photosynthetic euglenids are major contributors to fresh water ecosystems. *Euglena gracilis* in particular has noted metabolic flexibility, reflected by an ability to thrive in a range of harsh environments. *E. gracilis* has been a popular model organism and of considerable biotechnological interest, but the absence of a gene catalogue has hampered both basic research and translational efforts.

Results: We report a detailed transcriptome and partial genome for *E. gracilis* Z1. The nuclear genome is estimated to be around 500 Mb in size, and the transcriptome encodes over 36,000 proteins and the genome possesses less than 1% coding sequence. Annotation of coding sequences indicates a highly sophisticated endomembrane system, RNA processing mechanisms and nuclear genome contributions from several photosynthetic lineages. Multiple gene families, including likely signal transduction components, have been massively expanded. Alterations in protein abundance are controlled post-transcriptionally between light and dark conditions, surprisingly similar to trypanosomatids.

Conclusions: Our data provide evidence that a range of photosynthetic eukaryotes contributed to the *Euglena* nuclear genome, evidence in support of the 'shopping bag' hypothesis for plastid acquisition. We also suggest that euglenids possess unique regulatory mechanisms for achieving extreme adaptability, through mechanisms of paralog expansion and gene acquisition.

Keywords: *Euglena gracilis*, Transcriptome, Cellular evolution, Plastid, Horizontal gene transfer, Gene architecture, Splicing, Secondary endosymbiosis, Excavata

Introduction

Euglena gracilis, a photosynthetic flagellate, was first described by van Leeuwenhoek in 1684 [1]. There are over 250 known species in the genus *Euglena*, with around 20 predominantly cosmopolitan, including *E. gracilis* [2–5]. *Euglena* spp. are facultative mixotrophs in aquatic environments [6] and many possess a green secondary plastid derived by endosymbiosis of a chlorophyte algae [7]. Amongst the many unusual features of euglenids are a proteinaceous cell surface pellicle [8] and an eyespot [9–14]. Euglenids, together with kinetoplastids, diplonemids and symbiotids, form the Euglenozoa subgroup of the Discoba phylum [15]. Kinetoplastids are best known for the *Trypanosoma* and *Leishmania* lineages [15], important unicellular parasites, while diplonemids have been little studied, yet represent one of the most abundant and diverse eukaryotic lineages in the oceans [16].

E. gracilis is thus of importance due to evolutionary history, divergent cellular architecture, complex metabolism and biology, together with considerable potential for biotechnological exploitation [17]. However, the full complexity of euglenid biology remains to be revealed, and the absence of a complete genome sequence or annotated transcriptome has greatly hampered efforts to

* Correspondence: dacks@ualberta.ca; steven.kelly@plants.ox.ac.uk; mfield@mac.com
[9]Division of Infectious Disease, Department of Medicine, University of Alberta, Edmonton, Alberta T6G, Canada
[8]Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK
[1]School of Life Sciences, University of Dundee, Dundee DD1 5EH, UK
Full list of author information is available at the end of the article

study *E. gracilis* or to develop genetic tools [17, 18]. Two transcriptomes have been published, one derived from cells grown in light and dark conditions plus rich versus minimal media [17] and a second examining the impact of anaerobic conditions on gene expression [19]. For the most part, these studies focused on the biosynthetic properties of *E. gracilis* and not cellular systems or aspects of protein family evolution. Most recently, a study of low molecular weight RNA populations identified over 200 snoRNAs [20].

Comparisons between euglenozoans such as the free-living bodonids, early-branching trypanosomatids (*Paratrypanosoma confusum*), and parasitic forms have uncovered many genetic changes associated with parasitism [21–24]. Both the cell surface and flagellum of euglenoids are of significant importance to life cycle development, interaction with the environment and, for parasitic trypanosomes, pathogenesis and immune evasion [25, 26]. The surface macromolecules of trypanosomatids are highly lineage-specific with roles in life cycle progression [23, 27–31], but it remains to be determined to what extent *E. gracilis* shares surface proteins or other aspects of biology with the trypanosomatids or how cellular features diverge. Such information is invaluable for determining how parasitism arose in the kinetoplastids.

E. gracilis produces a wide range of secondary metabolites, and many of which are of potential commercial value [17]. Furthermore, *E. gracilis* is of considerable promise for biofuel production [32–34], and extremely resistant to conditions such as low pH and high metal ion concentrations, fueling interest as possible sentinel species or bioremediation agents [19, 35–37]. In parts of Asia, *E. gracilis* is cultivated as an important food supplement [38].

E. gracilis possesses a complex genome, with nuclear, plastid and mitochondrial components, an overall architecture known for decades. The coding potential of the mitochondrial genome is surprisingly small [39, 40], while the plastid is of more conventional structure [41]. The plastid is the result of a secondary endosymbiotic event, which is likely one of several such events occurring across eukaryotes [42]. Uncertainties concerning the origins of the plastid have remained, and not least of which has been the presence of genes from both red and green algae in the *E. gracilis* nuclear genome [19, 43]. Such a promiscuous origin for photosynthetic genes is not restricted to the euglenids and has been proposed as a general mechanism, colloquially the 'shopping bag' hypothesis, whereby multiple endosymbiotic events are proposed and responsible for the range of genes remaining in the nuclear genome, providing a record of such events and collecting of genes, but where earlier symbionts have been completely lost from the modern host [44].

The *E. gracilis* nuclear genome size has been estimated as in the gigabyte range [45–48] and organization and intron/exon boundaries of very few genes described [49–54]. In the kinetoplastids, unusual transcriptional mechanisms, involving the use of *trans*-splicing as a near universal mechanism for maturation of protein-coding transcripts and polycistronic transcription units, have been well described. As *E. gracilis* supports multiple splicing pathways, including conventional and non-conventional *cis*- [52, 53] and *trans*-splicing [55], there is scope for highly complex mechanisms for controlling expression, transcription and mRNA maturation [56], but how these are related to kinetoplastids is unclear.

We undertook a polyomic analysis of the Z1 strain of *E. gracilis* to provide a platform for improved understanding of the evolution and functional capabilities of euglenids. Using a combination of genome sequencing, together with pre-existing [17] and new RNA-seq analysis, proteomics and expert annotation, we provide an improved view of *E. gracilis* coding potential and gene expression for greater understanding of the biology of this organism.

Results and discussion

Genome sequencing of *Euglena gracilis*

We initiated sequencing of the *E. gracilis* genome using Roche 454 technology. The early assemblies from these data indicated a large genome in excess of 250 Mb and that data coverage was low. We turned to the Illumina platform and generated data from multiple-sized libraries, as well as a full lane of 150 bp paired-end sequences. These data were assembled as described in methods and as previously [48] and latterly supplemented with PacBio data generously donated by colleagues (Purificatión Lopéz-García, David Moreira and Peter Myler, with thanks). The PacBio data however failed to improve the assembly quality significantly, presumably due to low coverage.

Our final draft genome assembly has 2,066,288 sequences with N_{50} of 955 (Table 1), indicating significant fragmentation. The estimated size of the single-copy proportion of the genome is 140–160 mb and the estimated size of the whole haploid genome is 332–500 mb. This is consistent with several estimates from earlier work (e.g. [57]), *albeit* based here on molecular sequence data rather than estimates of total DNA content. Using the core eukaryotic genes mapping approach (CEGMA) [58], we estimate that the genome assembly, or at least the coding sequence proportion, is ~ 20% complete. Hence, this assembly could only support an initial analysis of genome structure and is unable to provide a full or near full open reading frame catalog (Table 2). The heterozygosity, size and frequency of low complexity

Table 1 Statistics of genome assembly

Parameter	
Number of sequences	2,066,288
Median sequence length	457
Mean sequence length	694
Max sequence length	166,587
Min sequence length	106
No. sequence > 1kbp	373,610
No. sequence > 10kbp	1459
No. sequence > 100kbp	2
No. gaps	0
Bases in gaps	0
N50	955
Combined sequence length	1,435,499,417

Following the assembly process, over two million sequences were retained, with a median sequence length of 457 bp

sequence hampered our ability to assemble this dataset (see the "Materials and Methods" section for more details). The size and frequency of low-complexity sequence clearly precluded assembly of our dataset from Illumina reads, and significantly, PacBio data had no significant impact on assembly quality. Due to the large proportion of low-complexity sequence, any estimate for the size of the genome is very much an approximation.

Restricting analysis to contigs > 10 kb, where some features of overall gene architecture could be inferred, we identified several unusual aspects of genome structure (Table 3, Fig. 1, Additional file 1: Figure S1). These contigs encompassed about 22 Mb of sequence, but with

Table 3 Characteristics of contigs assembled with length exceeding 10 kb

Contigs	Total contigs analysed > 10 kb	1459
	Total nucs in contigs analysed	22 Mb
	Contigs with CDS	53
	Percent contigs with CDS	3.6
CDS	Number analysed	135
	Average length	3790
	Total length	481,369
Exons	Number of exons analysed	421
	Average Length	174.54
	Median Length	112
	Total Length	73,482
	Average per predicted CDS	3.85
Introns	Total introns analysed	271
	Average length	1027.14
	Median length	598
	Total length	278,354
	Introns per predicted CDS	2.01
	Number/percent conventional	218/80.1
	Number/percent intermediate	30/11.1
	Number/percent non-conventional	23/8.5
	Percent nucleotides in CDS (exon)	0

The contigs were ranked by size and those exceeding 10 kbp extracted and analyzed for length, coding sequence, exon structure and other features

Table 2 CEGMA analysis of selected datasets

Assembly	Organism	Gene status	Prots	%Completeness	Total	Average	%Ortho
Genome	*E. gracilis*	Complete	22	8.87	37	1.68	54.55
		Partial	50	20.16	89	1.78	56
	T. brucei	Complete	196	79.03	259	1.32	24.49
		Partial	205	82.66	282	1.38	28.29
	L. major	Complete	194	78.23	220	1.13	11.34
		Partial	204	82.26	245	1.2	15.69
Transcriptome	*E. gracilis*	Complete	187	75.4	390	2.09	65.78
		Partial	218	87.9	506	2.32	69.72
	T. brucei	Complete	190	76.61	393	2.07	60
		Partial	205	82.66	448	2.19	63.41
	L. major	Complete	133	53.63	275	2.07	64.66
		Partial	194	78.23	405	2.1	64.43

Comparisons for CEGMA scores between *E. gracilis, T. brucei* and *L. major* as an estimate of 'completeness' based on 248 CEGs. *Prots* number of 248 ultra-conserved CEGs present in genome, *%Completeness* percentage of 248 ultra-conserved CEGs present, *Total* total number of CEGs present including putative orthologs, *Average* average number of orthologs per CEG, *%Ortho* percentage of detected CEGs that have more than 1 ortholog, *Complete* those predicted proteins in the set of 248 CEGs that when aligned to the HMM for the KOG for that protein family, give an alignment length that is 70% of the protein length. i.e. if CEGMA produces a 100 amino acid protein, and the alignment length to the HMM to which that protein should belong is 110, then we would say that the protein is "complete" (91% aligned), *Partial* those predicted proteins in the 248 sets that are incomplete, but still exceeds a pre-computed minimum alignment score. Keys are as described [58]

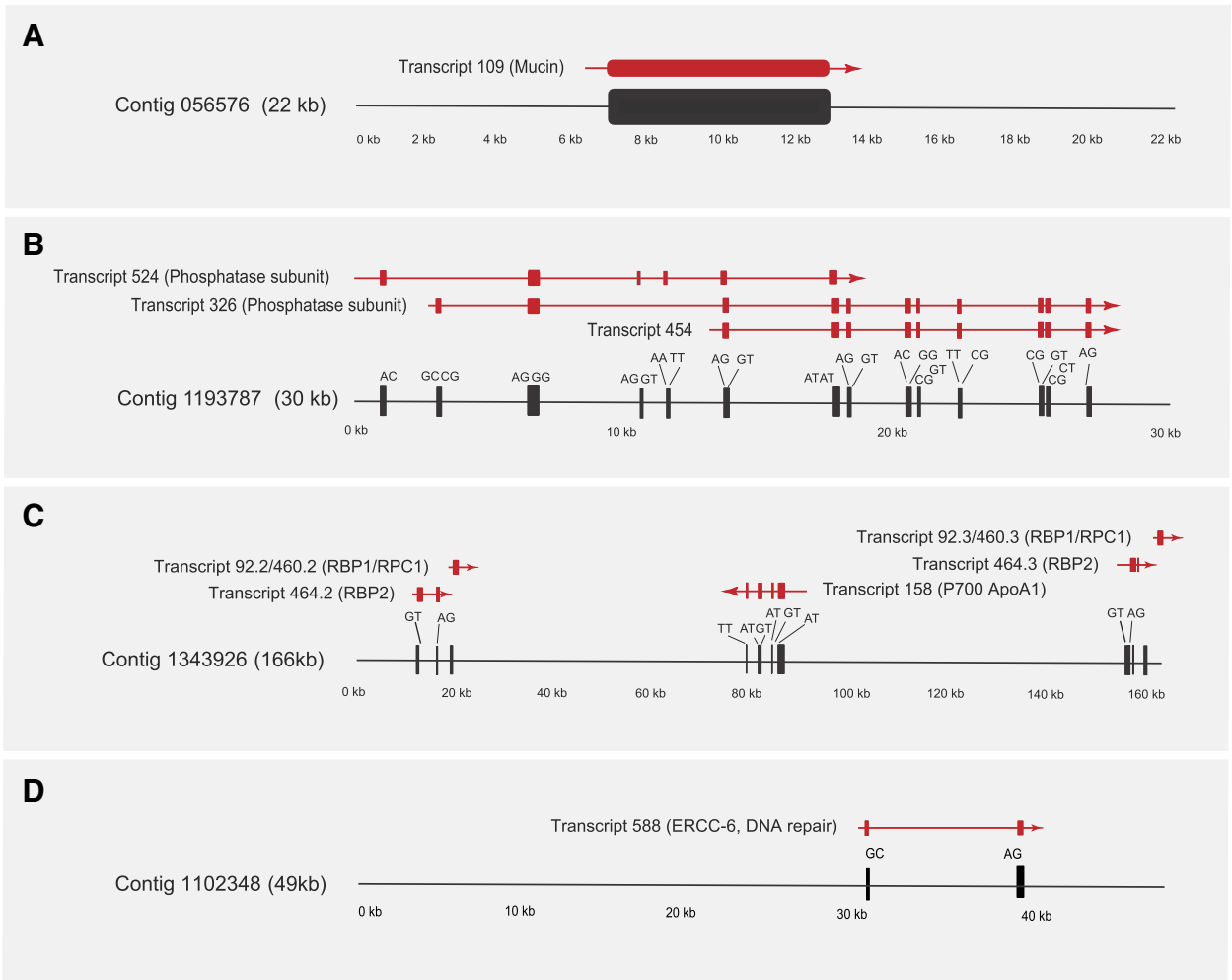

Fig. 1 a–**d** *Euglena gracilis* exon structure. The predicted gene structure of several selected contigs is shown, including the mapped transcripts (red), predicted splice sites and intergenic regions. Note that transcripts 524 and 326, (panel **b**) which encompass essentially the same portions of the genome, demonstrate possible differential exon inclusion, indicating differential open reading frame organisation and possible alternate splicing. Black boxes indicate exons, with predicted splice site dinucleotides indicated above. Transcripts are shown as arrows with the arrowhead indicating the predicted direction of transcription. Protein product annotations are indicated in parentheses. Contig sizes are shown in kilobase; note that each contig is not drawn to the same scale. Further examples of predicted contig gene organisation are given in Additional file 1: Figure S1

only 135 genes predicted based on Exonerate [59], this suggests an extremely low gene density of < 1%, similar to that in *Homo sapiens*. In those contigs that possess predicted coding sequence, there was frequently more than one open reading frame (ORF), suggesting gene clusters present within large expanses of non-coding sequence (e.g. Contig11343926, Fig. 1c), but with the caveat that we have sampled a very small proportion of total ORFs (Table 3). It is also possible that some genes were not predicted due to absence of expression under the conditions we used for RNA-seq, though we consider this likely a minor contribution as multiple culturing conditions were included within the final RNA-seq dataset (see below). Most identified genes are predicted to be *cis*-spliced and most introns are conventional, with a smaller proportion of intermediate and non-conventional splice sites (consistent with [57]). Some introns appear very large compared to the coding sequence contained between them (Contig 1102348, Transcript 588, Fig. 1d). Furthermore, some genes are apparently unspliced (Fig. 1a; Contig 056576, Transcript 109) and there is evidence for alternate splicing (Fig. 1b; Contig 1193787, Transcripts 326, 454 and 524). Evidence for alternate spicing was described earlier [19], but it was based on RNA-seq data without a genomic context, unlike here. The near complete absence of *cis*-splicing from bodonids and trypanosomatids clearly reflects loss post-speciation of these lineages from euglenids and removed a considerable mechanism for generation of proteome diversity [60]. The biological basis for the

extreme genome streamlining in the trypanosomatids versus *Euglena* is unclear.

We also sequenced and assembled an *E. gracilis* transcriptome using a combination of in-house generated sequence and publicly available data [17]. This strategy had the advantage of focusing on coding sequence, as well as including data from multiple environmental conditions (see [17], which used dark, light conditions and rich or minimal media and data from here that used distinct media and also light and dark conditions), to increase the likelihood of capturing transcripts, and represents a third analysis, *albeit* incorporating raw reads from previous work [17].

Over 32,000 unique coding transcripts were predicted by [17], which compares well with this new assembly and which accounted for 14 Mb of sequence overall. Of these transcripts, approximately 50% were annotatable using UniRef, and over 12,000 were associated with a GO term. In a second report, Yoshida et al. [19], assembled 22 Mb of coding sequence within 26,479 likely unique components, with about 40% having assignable function based on sequence similarity to Swiss-Prot.

The total number of coding sequence nucleotides in our new assembly was > 38 Mb, with a mean length of 869 bases and 36,526 unique coding sequences (Table 4). This is a significant improvement over 391 bases reported by [17], and comparable to [19], albeit with a significant increase in total sequence assembled. Transcriptome coverage of ORFs was, as expected, significantly superior to the genome, and CEGMA indicated 87.9% recovery (the *Trypanosoma brucei* genome is 82.66%) (Tables 2 and 4).

We also compared the completeness of our transcriptome with the two published transcriptomes of *E. gracilis* [17, 19]. We used TransDecoder (v2.0.1) [61] to translate nucleotide transcripts to proteins and then excluded duplicated proteins with CD-HIT utility (v4.6) with standard parameters [62]. The final comparison, made by BUSCO (v2.0.1) [63] with the eukaryotic database, is shown as Additional file 1: Figure S12. Note that all three studies report similar statistics, including concordance in the cohort of BUSCOs not found; these may have failed to be detected or genuinely be absent. Given that 19 BUSCOs were not found in concatenated data (i.e. all three assemblies), with between four to eight missing BUSCOs specific to individual assemblies, it is highly likely that these datasets are robust while also indicating saturation in terms of achieving 'completeness', together with possible limitations with BUSCO for divergent species such as *E. gracilis*.

Comparisons between genome and transcriptome assembly sizes confirmed the very small coding component, with genome contigs containing significantly less than 1% coding sequence, despite the total number of *E. gracilis* ORFs (36526) being two to three times greater than *Bodo saltans* (18963), *T. brucei* (9068) or *Naegleria gruberi* (15727) [64–66]. This is in full agreement with earlier estimates of genome versus transcriptome size [17] as well as estimates of the proportion of coding and total genomic sequence discussed above. This is also similar to other large genomes and, specifically, *Homo sapiens*. Blast2GO and InterProScan annotated over 19,000 sequences with GO terms, a proportion similar to previous reports (Additional file 1: Figure S2, [17, 19]).

In addition to the formal analysis and calculation of the numbers of unique sequences, our annotation of the transcriptome adds additional confidence that the dataset is a good resource:

(i) Most expected metabolic pathways could be reconstructed, with very few exceptions,
(ii) Major known differences between kinetoplastids and *Euglena* were identified, supporting sampling to a deep level,

Table 4 Assembly statistics for the transcriptome

Transcripts		Coding sequence (CDS)		Proteins	
Number of sequences	72,509	Number of sequences	36,526	Number of proteins	36,526
Median sequence length	540	Median sequence length	765	Median protein length	254
Mean sequence length	869	Mean sequence length	1041	Mean protein length	346
Max sequence length	25,763	Max sequence length	25,218	Max protein length	8406
Min sequence length	202	Min sequence length	297	Min protein length	98
No. sequence > 1kbp	19,765	No. sequence > 1kbp	13,991	No. proteins > 1kaa	1290
No. sequence > 10kbp	25	No. sequence > 10kbp	24	N50	471
No. sequence > 100kbp	0	N50	1413		
No. gaps	0	Combined sequence length	38,030,668		
Bases in gaps	0				
N50	1242				
Combined sequence length	63,050,794				

(iii) For most analyzed protein complexes, all subunits or none were identified, indicating that partial coverage of components is likely rare.

Overall, we conclude that the transcriptome is of sufficient quality for robust annotation and prediction and encompasses more than previous datasets.

Post-transcriptional control of protein expression

Trypanosomatids exploit post-transcriptional mechanisms for control of protein abundance, where essentially all genes are produced from polycistronic transcripts via *trans*-splicing. To improve annotation and investigate gene expression in *E. gracilis*, we conducted comparative proteomic analysis between light and dark-adapted *E. gracilis* but retained in the same media and temperature. Previous work suggested that control of protein abundance may be post-transcriptional [67, 68], but analysis was limited and did not consider the entire proteome, while a separate study identified some changes to mRNA abundance under low oxygen tension [19]. Under these well-controlled conditions, however, significant changes to the proteome were expected. We confirmed by UV/VIS spectroscopy and SDS-PAGE that photosynthetic pigments were lost following dark adaptation and that ensuing ultrastructural changes, i.e. loss of plastid contents, were as expected (Additional file 1: Figure S3). Total protein extracts were separated by SDS-PAGE with 8661 distinct protein groups (representing peptides mapping to distinct predicted ORFs, but which may not distinguish closely related paralogs) identified. Ratios for 4681 protein groups were quantified (Additional file 2: Table S1) including 384 that were observed in only one state (232 in light and 152 in dark). In parallel, we extracted RNA for RNA-seq analysis; comparing transcript hits with protein groups identified 4287 gene products with robust information for both protein and RNA abundance.

Correlations between changes to transcript and protein abundance were remarkably poor (Fig. 2, Additional file 1: Figure S3, Additional file 2: Table S1), consistent with

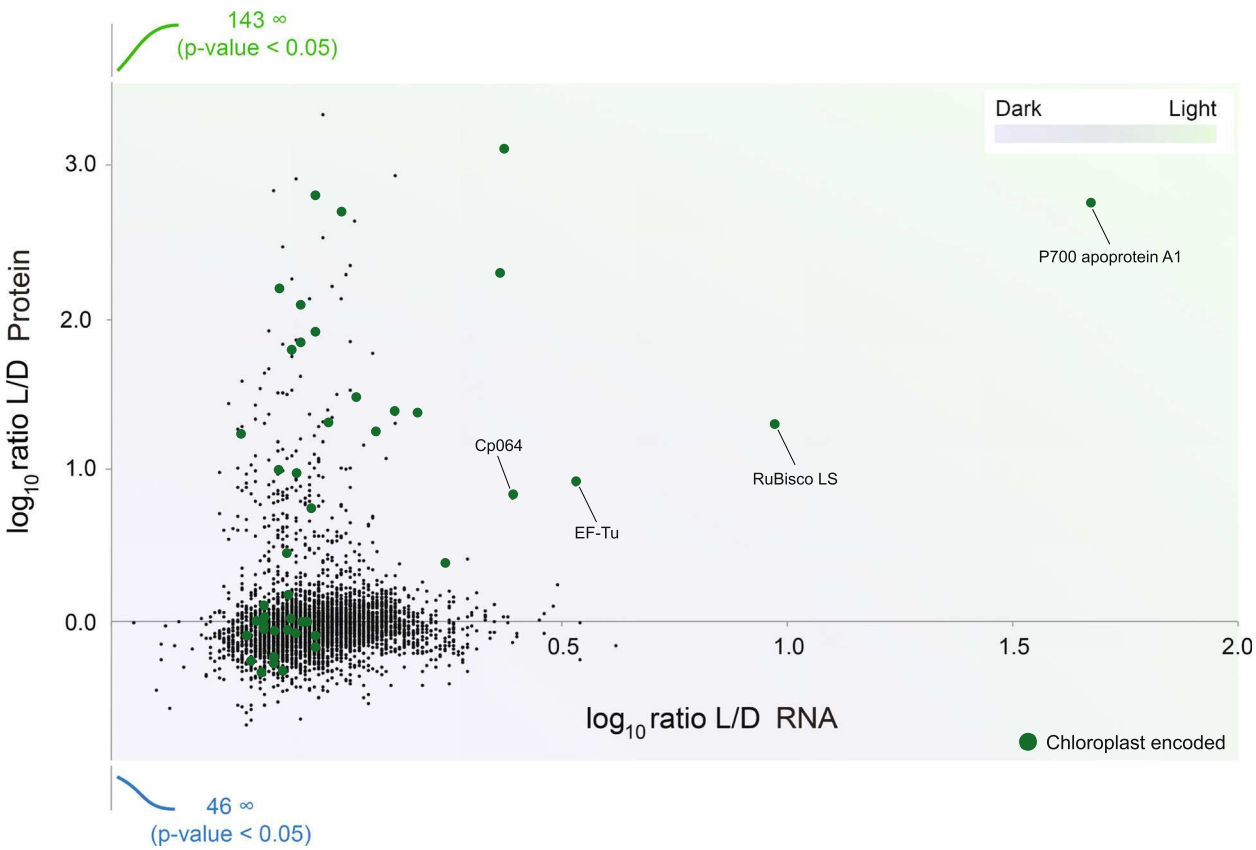

Fig. 2 Expression level changes induced by light are mainly post-transcriptional. Alterations to the transcriptome and proteome in response to ambient light or complete darkness were analysed using RNA-seq and SILAC/LCMS^2 proteomics respectively. Data are plotted for individual transcripts/polypeptides as the $\log_{10}$ ratio between the two conditions, light (L) and dark (D), with protein on the *y*-axis and RNA on the *x*-axis. The presence of a number of proteins that were detected exclusively under one or other condition (hence infinite ratio) are indicated in green (for light) and blue (for dark). With the exception of a few transcripts, which are plastid encoded (green dots), there is little alteration to RNA abundance, but considerable changes to protein levels. Raw data for transcriptome/proteome analysis are provided in Additional file 3

some much smaller earlier studies [67, 68] and broadly with the more extensive study reported in [19]. BLAST analysis revealed that those transcripts where differential abundance did correlate with protein abundance are encoded by the chloroplast genome, including several photosystem I proteins, i.e. P_{700} chlorophyll apoprotein A_1, the large subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (RuBisCO) and chloroplast encoded EF-Tu. Nuclear elongation factors are not influenced by switching growth conditions from dark to light [69], consistent with our finding of no differential expression of nuclear EF-1α, while both the chloroplast EF-Tu protein and corresponding transcript (EG_transcript_1495) are highly upregulated by light. This absence of transcriptional control for proteome changes between these two conditions is highly similar to that reported for the kinetoplastids, despite the presence of widespread *cis*-splicing and a sparse genome that likely precludes extensive polycistronic transcription. It remains to be determined if this is a general feature for *E. gracilis* or only for certain environmental cues; a cohort of genes are strongly impacted at the RNA level when comparing aerobic to anaerobic transcripts for example, but in that instance none of these transcripts were plastid-encoded nor was a protein analysis performed [19].

Ancestry of *Euglena gracilis* genes

We used two different approaches to analyze the evolutionary origin of genes predicted from the *E. gracilis* transcriptome. Firstly, we used OrthoFinder [70] to identify *E. gracilis* ortholog gene families shared across eukaryotes and those restricted to specific taxonomic groupings (Fig. 3a, Additional file 1: Figure S4). As expected, the largest proportion was represented by all supergroups and dominated by core metabolic, structural and informational processes, consistent with previous work [19]. A second cohort is shared between *E. gracilis* and other excavates. These classes are broadly within the relative frequencies of previous analyses of excavate genomes [19, 71]. A third cohort represents nuclear transfer of endosymbiotic genes from acquisition of the plastid, and consequently, the genome is a complex mosaic as all eukaryotic genomes also harbour genes driven from the mitochondrial endosymbiont. GO terms associated with orthogroups indicated increased frequency of regulatory function genes in green/secondary plastid orthogroups (Additional file 1: Figure S2). Previous transcriptome studies reported the presence of pan-eukaryotic genes and cohorts shared with kinetoplastids and plants [17, 19], but these were not analyzed in detail, and specifically did not determine which plant taxa were acting as potential gene donors. This is important in terms of understanding the origins of the *Euglena* plastid and where earlier data suggested the presence of a diverse set of genes from at least green, red and brown algae ([43, 72]). Particularly relevant here is that plastid acquisition in euglenoids is relatively recent [73].

To address this question, we employed a second approach, in which we performed exhaustive analysis to establish phylogenetic ancestry of individual proteins from the predicted *Euglena* proteome by generating single-protein phylogenies. Unlike the analyses of orthogroup sharing, this second approach can be used only for a subset of proteins with a sufficiently robust phylogenetic signal, but also allows determination of the gene ancestry; moreover, this is applicable for members of complex gene families. From all predicted *E. gracilis* proteins only 18,108 formed reliable alignment (> 75 positions) with more than two sequences from our custom database, which comprised 207 taxa in total (Additional file 3 Table S2) and was used for tree construction. In 4087 trees, *E. gracilis* formed a robust (bootstrap support > = 75%) sister relationship with a taxonomically homogeneous clade (Fig. 3b). Of these, 1816 (44%) were related to one of the lineages of Excavata and 1420 (35%) were related specifically to kinetoplastids. This major fraction represents mostly the vertically inherited component of the genome. The largest non-vertical component forms a group of 572 (14%) proteins related to green plants and green algae, likely representing genes acquired by endosymbiotic gene transfer from the *Euglena* secondary chloroplast, but it should be noted that the direction of transfer cannot be objectively determined. This category is followed by four groups related to the algal groups: haptophytes, cryptophytes, ochrophytes and chlorarachniophytes. While many proteins within the chlorarachniophyte group may represent mis-assigned genes related to green algae, these relatively large numbers related to the three brown-algal groups (723 in total) suggests that these algae contributed considerably to the *E. gracilis* genome and that the process of chloroplast endosymbiosis was complex (see below). On the other hand, the number of proteins related to red algae and glaucophytes (50 and 53) is near negligible. Proteins in groups shared with prokaryotes (220) and non-photosynthetic eukaryotes, e.g. Metazoa (149) and Amoebozoa (145), are most probably the result of horizontal gene transfers, differential gene losses or artifacts caused by biased phylogenetic reconstructions or contaminations in the data sets used to construct the custom database. The robust nature of our analysis, being restricted to phylogenetically well-resolved trees, provides an additional level of confidence to the concept of multiple origins for LGT genes.

It was initially thought that plastid-possessing organisms would overwhelmingly possess nuclear genes derived by transfer from the endosymbiont corresponding

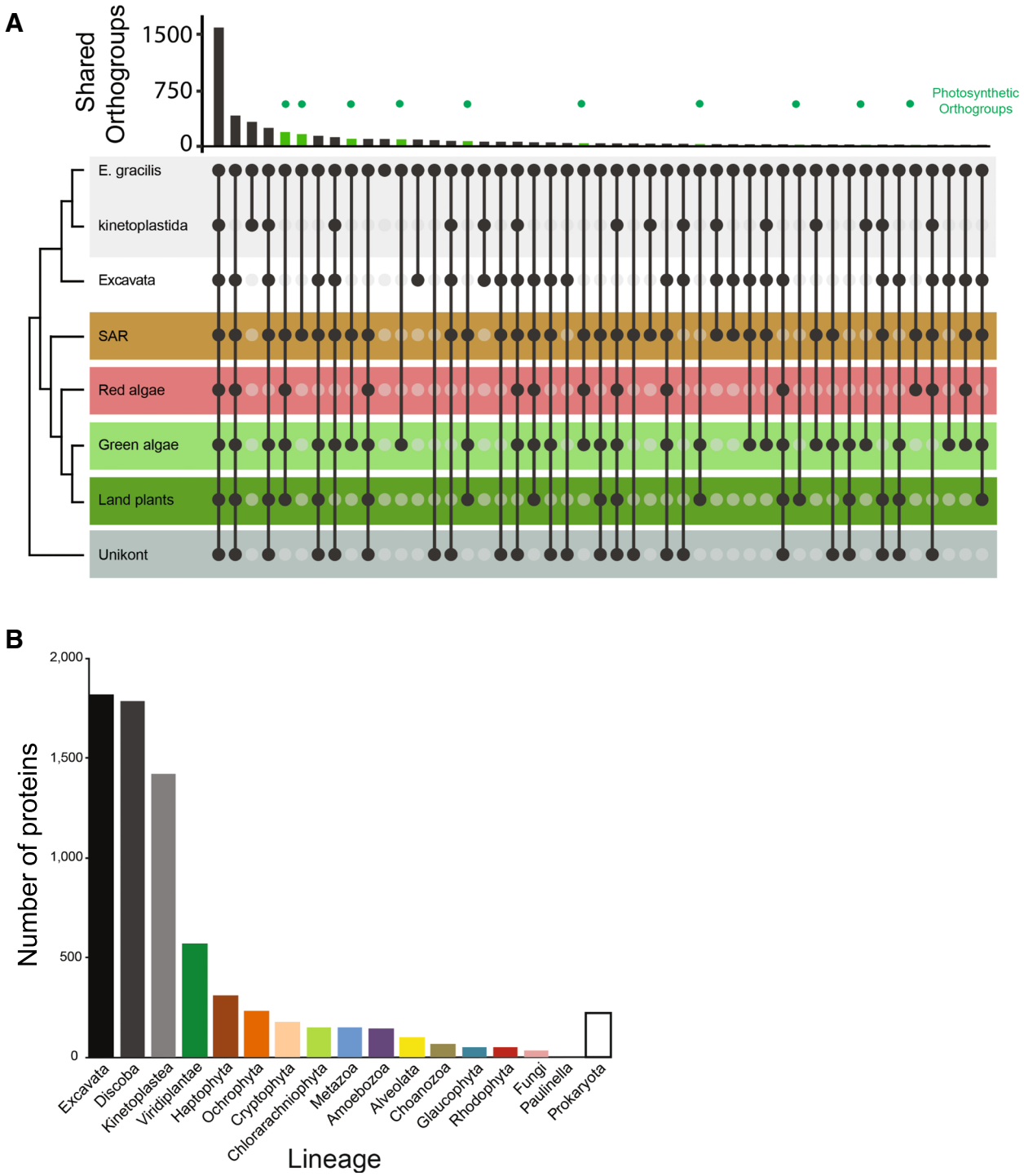

Fig. 3 *Euglena gracilis* shares orthologs with a diverse array of lineages. Panel **a** (top): Histogram of *E. gracilis* orthologous groups clustering with selected eukaryotic lineages as determined with OrthoFinder. The *x*-axis shows the number of orthogroups and *y*-axis shows the taxon groupings representative of selected eukaryotic groups. Histogram bars highlighted in green indicate orthogroups shared with photosynthetic organisms. Panel **a** (lower): taxa sharing orthogroups with *E. gracilis*, where black circles correspond to the presence of orthogroup members while light gray circles correspond to the absence of orthogroup members in the genome. Black tie bars linking black circles are for clarity only. Eukaryotic taxon groupings are colored accordingly: gray, *Euglena* and kinetoplastida; white, other members of the Excavata excluding Euglenozoa; brown, SAR, pink, red algae; light green, green algae; dark green, land (vascular) plants and dark gray, Unikona. An expanded version of this figure, broken down by species is given as Additional file 1: Figure S4. Panel **b**: The number of *E. gracilis* proteins that clustered (BS > 75%) in their single-protein phylogenetic tree with taxonomic group are indicated on the x-axis

to the plastid currently present, but this has been challenged [74, 75]. While contributions from multiple algal lineages could be explained by incomplete phylogenetic sampling, this is also consistent with the 'shopping bag' hypothesis, which proposes an extended process of transient endosymbiosis and gene acquisition by the host prior to the present configuration [44, 75] and which is likely a quite general phenomenon and occurs in many lineages. Our analysis strongly supports the concept of sequential endosymbiotic events.

Expansive paralog families

Several orthogroups consist of an expansive cohort of *E. gracilis* sequences, and a selected few were analyzed phylogenetically and annotated for protein architectural/domain features (Additional file 1: Figure S5, Additional file 4: Table S3). Firstly, highly significant in terms of size and evolutionary history is a family of nucleotidylcyclase III (NCIII)-domain-containing proteins widely distributed across eukaryotes. In African trypanosomes, adenylate cyclases are mediators of immune modulation in the mammalian host [71]. One nucleotidylcyclase subfamily is restricted to kinetoplastids and organisms with secondary plastids and contains photosensor adenylate cyclases [12] that possess one or two BLUF domains (blue light sensor) with a double NCIII domain (Fig. 4). These nucleotidylcyclases are phylogenetically similar to the NCIII-family of *N. gruberi* [66]. A second subfamily is pan-eukaryotic and possesses one NCIII domain and several *trans*-membrane domains, a HAMP (histidine kinases, adenylate cyclases, methyl-accepting proteins and phosphatases) domain as well as cache 1 (calcium channel and chemotaxis receptor) domains. These domains are associated with proteins involved, as their name implies, in signal transduction, particularly chemotaxis [76, 77]. Again, this subfamily is closely related to NCIII-family genes from *N. gruberi*. The third subfamily represents a kinetoplastid cluster with *trans*-membrane proteins and frequently also HAMP and cache1 domains. This complexity indicates considerable flexibility in nucleotidylcyclase evolution and that many lineage-specific paralogs have arisen, with implications for signal transduction, suggesting an extensive regulatory and sensory capacity in *E. gracilis*.

A second example is a large protein kinase C-domain containing a group of protein kinases, which also exhibit extensive lineage-specific expansions in *E. gracilis* (several orthogroups contained a very large number of *E. gracilis* sequences, and a few selected were analysed phylogenetically and annotated for architecture (Additional file 1: Figure S5)). A third orthogroup possess a signal receiver domain (REC) with clear lineage-specific *E. gracilis* paralogs present (Additional file 1: Figure S5). The *E. gracilis* members possess an H-ATPase domain, which is distinct from the Per-Arnt-Sim (PAS) domain present in many orthologs from other lineages. The presence of independently expanded signaling protein families in *E. gracilis* suggests both highly complex and divergent pathways. These very large families likely partly explain the expanded coding potential in *E. gracilis*, as well as provide some indication of how sensing and adaptation to diverse environments is achieved.

Conservation and divergence of systems between *E. gracilis* and kinetoplastids

To better understand the evolution of *Euglena* and its relationship to free living and parasitic relatives, we selected multiple cellular systems for detailed annotation. These were selected based on documented divergence between kinetoplastids and other eukaryotic lineages and encompass features of metabolism, the cytoskeleton, the endomembrane system and others (Additional file 5: Table S4). Additional annotations of systems not discussed here are available in Additional file 5: Table S4 and provided in Additional file 6: Supplementary analysis.

A unique feature of energy metabolism in kinetoplastids is compartmentalisation of several glycolytic enzymes within peroxisome-derived glycosomes and the presence of additional enzymes for metabolism of the glycolytic intermediate phospho-enolpyruvate to succinate [78]. Glycosomes have been recently reported in diplonemids, the second major euglenozoan group, suggesting an origin predating kinetoplastida [79]. Using 159 query protein sequences for experimentally supported glycosomal *T. brucei* proteins [80], we found candidate orthologs for the majority, but based on the absence of detectable PTS-1 or PTS-2 targeting signals, no evidence that enzymes linked to carbohydrate metabolism are (glyco)peroxisomal. Of the 159 queries, 49 are annotated as hypothetical or trypanosomatid-specific and none had a detectable ortholog in *E. gracilis* (Additional file 5: Table S4). Collectively, this suggests that peroxisomes in *E. gracilis* most likely function in diverse aspects of lipid metabolism rather than glycolysis or other aspects of carbohydrate metabolism and distinct from kinetoplastids.

The surface membrane of *E. gracilis* is in close association with a microtubule corset, and with some structural similarity to the subpellicular array of trypanosomatids, but with very unique architecture [81]. While the plasma membrane composition of kinetoplastids is lineage-specific, in terms of many major surface proteins and a major contributor to host-parasite interactions [82], transporters and some additional surface protein families are more conserved. To compare with *E. gracilis*, we predicted membrane proteins using the signal peptide together with orthogroup clustering, which will encompass both surface and endomembrane compartment constituents. Many genes have significant similarity to

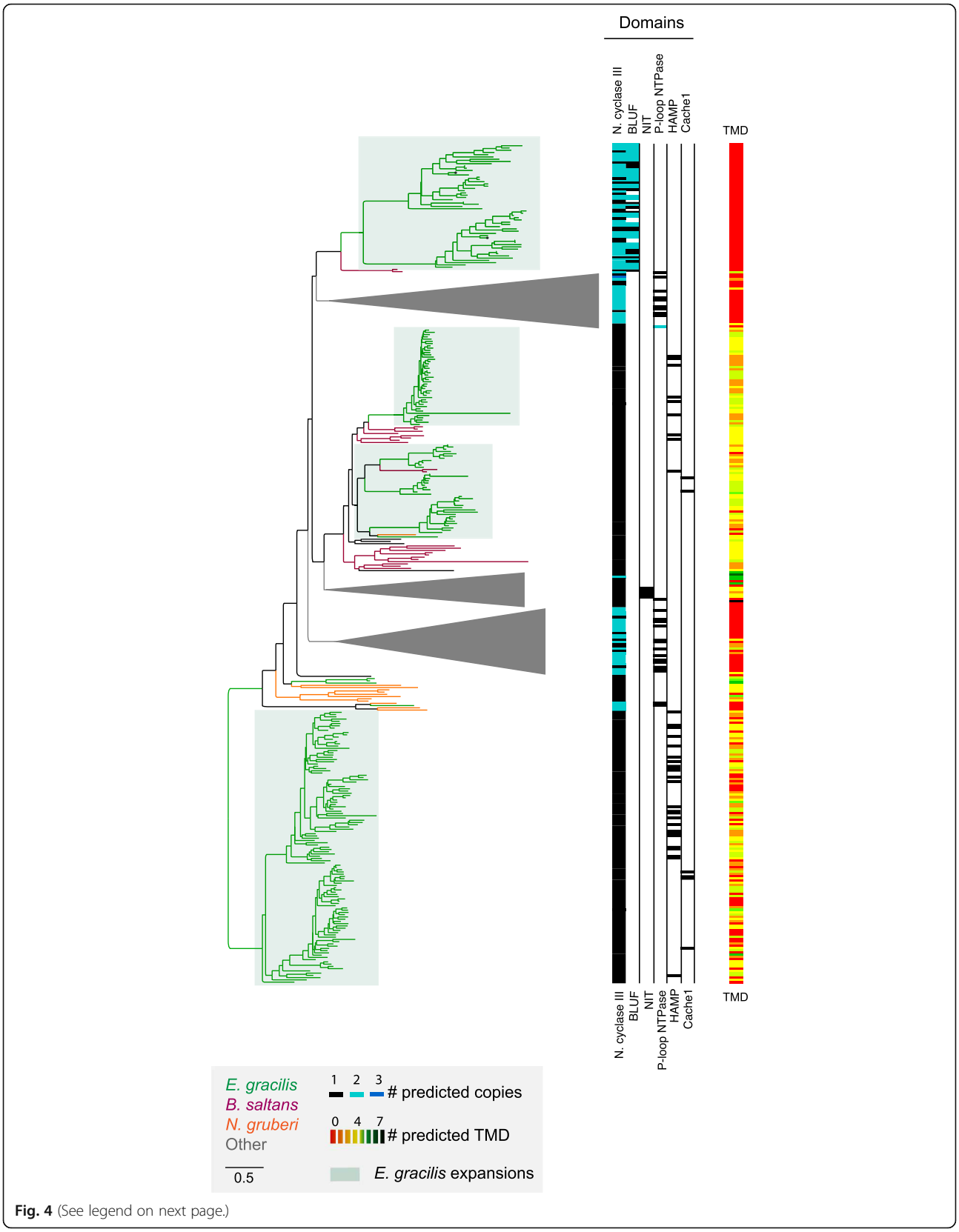

Fig. 4 (See legend on next page.)

(See figure on previous page.)
Fig. 4 Large paralog gene families are present in the *Euglena gracilis* genome. Several orthogroups contain many *E. gracilis* paralogs. The phylogenetic distribution of one large orthogroup, the nucleotidylcyclase III domain-containing proteins, is shown**.** Lineage groupings are colour coded: gray, all eukaryotes (and collapsed for clarity); red, *N. gruberi*; amber, *B. saltans*; and green, *E. gracilis*. Clades containing only *Euglena* sequences are boxed in green. Each sequence has been assigned a domain composition (colour gradient black to teal to blue), number of predicted trans-membrane domains (colour coded red to orange to black gradient). To obtain this phylogenetic tree, sequences with likely low coverage (less than 30% of the length of the overall alignment) were removed during alignment to avoid conflicting homology or artefact generation. Domain compositions identified are nucleotidylcyclase III, BLUF, NIT, P-loopNTPase, HAMP and Cache1

kinetoplastids (1103), *B. saltans* (32) or non-kinetoplastida (487) (Additional file 7: Table S5). About 698 proteins with a signal peptide appear to be *E. gracilis* specific, and most of these are a single copy (87.5%), while there are clear large families that possess conserved features (see above). Notably, we were unable to identify a rhodopsin homolog, in contrast to several biochemical analyses suggesting the presence of retinal, the rhodopsin cofactor, which has been interpreted as evidence for a rhodopsin-like light sensor. It remains possible that the euglenid rhodopsin was not represented in the transcriptome or is too divergent to detect [83].

In common with *B. saltans*, *E. gracilis* has a distinct class of amastin, a major kinetoplastid surface protein and which arose from a single ancestor shared with the last euglenozoan common ancestor (Additional file 1: Figure S6). *E. gracilis* also possesses enzymes for the synthesis of lipophosphoglycan (LPG), a glycoconjugate first described in *Leishmania* and implicated in defense and disease mechanisms, together with the pathways for synthesis of GPI protein anchors and free lipids. These data suggest that LPG predates the evolution of parasitism and that the ancestral role was possibly more general, for example, a defense against proteases or predation, or in cell-cell/cell-substrate interactions. Significantly, gp63, a major surface protein present in the vast majority of eukaryotes and also involved in *Leishmania* pathogenesis, is absent and represents a secondary loss following separation from the kinetoplastid lineage.

The endomembrane system is responsible for biosynthesis, degradation and targeting of proteins and lipids and can be considered as a proxy for intracellular complexity. Compartments and transport routes can be predicted with accuracy based on the presence of genes encoding proteins mediating these routes. Using such an analysis, it has been predicted that the complexity of endomembrane compartments in trypanosomatids is decreased compared with free-living bodonids [23, 84]. *E. gracilis* possesses a relatively complete set of membrane-trafficking proteins, extending this trend further (Additional file 1: Figure S7). Two key adaptin family complexes involved in vesicle coat formation and post-Golgi transport, AP5 and TSET, are absent from kinetoplastids, and while AP5 is also absent from *E. gracilis*, a near complete TSET is present. Significantly, endosomal pathways are predicted as more complex than kinetoplastids, with multiple Rab7 (late endosome/lysosome) and Rab11 (recycling endosome) paralogs, together with ER-associated paralogs for Rab1 (early anterograde transport) and Rab32, respectively. Rab32 may also be associated with the contractile vacuole, an endolysosomal organelle responsible for osmoregulation in many freshwater protists, but these aspects of *E. gracilis* biology remain to be explored.

In kinetoplastids, an unusual cytoskeletal element, the bilobe, plays a central role in Golgi, flagellar pocket collar and flagellum attachment zone biogenesis [74]. All of the structural proteins (MORN1, RRP1, BILBO1, Centrin-2 and Centrin-4) were found [85–90] (Additional file 5: Table S4). Therefore, the potential for the synthesis of a bilobe-like structure in *E. gracilis* is supported, although clearly experimental evidence is needed for the presence of such a structure, but which suggests an origin predating the kinetoplastids.

The considerable size of the *E. gracilis* genome and complex splicing patterns suggests the presence of sophisticated mechanisms for organizing chromatin, mRNA processing and transcription [53, 57]. Furthermore, the *E. gracilis* nucleus has somewhat unusual heterochromatin morphology, with electron-dense regions appearing as numerous foci throughout the nucleoplasm (Additional file 1: Figure S8). Nucleoskeletal proteins related to lamins, NMCPs of plants or kinetoplastid-specific NUP-1/2 are all absent from *E. gracilis*, suggesting that anchoring of chromatin to the nuclear envelope exploits a distinct mechanism [91]. Further, while much of the nuclear pore complex (NPC) is well conserved across most lineages, orthologs for DBP5 and Gle1, two proteins involved in mRNA export in mammalian, yeast and plant NPCs, but absent from trypanosomes, are present. This is consistent with an earlier proposal that the absence of DBP5/Gle1 is connected to the loss of *cis*-splicing in kinetoplastids, but indicates that this is not due to the presence of *trans*-splicing per se as this is common to *E. gracilis* and the kinetoplastids [92]. Finally, kinetochores, required for engagement of chromosomes with the mitotic spindle, are also highly divergent in trypanosomes (Additional file 1: Figure S8) [93, 94]. Of the trypanosomatid kinetochore proteins, only KKT19 and KKT10 are obviously present in *E. gracilis*; as these are a kinase and phosphatase,

respectively, they may not be bona fide kinetochore proteins in *E. gracilis*. Further, very few canonical kinetochore proteins were found, suggesting possible divergence from both higher eukaryote and trypanosome configurations. Overall, these observations suggest unique mechanisms operate in the *E. gracilis* nucleus, which may reflect transitions between conventional kinetochores, lamins and nuclear pores into the more radical configuration present in kinetoplastids. Additional systems are discussed in supplementary material (Additional file 6).

The *Euglena* mitochondrion

In kinetoplastids, unique mitochondrial genome structures are present [95]. Typically, kinetoplastid mitochondrial genomes comprise ~40 copies of a maxicircle encoding several mitochondrial proteins and several thousand minicircles encoding guide RNAs for editing maxicircle transcripts [40, 95]. In trypanosomatids, this structure is attached to the flagellum basal body via a complex cytoskeletal element, the tri-partite attachment complex (TAC) [95]. We find no evidence for RNA editing in *E. gracilis*, nor for the TAC, both of which are consistent with the presence of a mitochondrial genome composed of only short linear DNA molecules and a conventional mitochondrial mRNA transcription system [39]. Specifically, only 16 of 51 proteins involved in RNA editing in *T. brucei* [96] had reciprocal best BLAST hits, and only one predicted protein contained a mitochondrial targeting signal. No homologs to TAC proteins were found (Additional file 5: Table S4).

The *E. gracilis* mitochondrial proteome is predicted to exceed 1000 proteins and encompasses 16 functional categories (Additional file 1: Figure S9A). The kinetoplastid mitochondrion possesses a non-canonical outer mitochondrial membrane translocase (A)TOM (archaic translocase of the outer membrane). The major component is (A)TOM40, a conserved beta-barrel protein that forms the conducting pore, but which is highly diverged in kinetoplastids [97–99]. We identified homologs of two specific receptor subunits of (A)TOM, namely ATOM46 and ATOM69 [100], and two TOM40-like proteins; both these latter are highly divergent and could not be assigned unequivocally as TOM40 orthologs.

We also identified canonical subunits of respiratory chain complexes I–V and 27 homologs of kinetoplastid-specific proteins, together with the widely represented alternative oxidase, consistent with earlier work [101]. Moreover, an ortholog of *T. brucei* alternative type II NADH dehydrogenase (NDH2) was detected. We found only 38 of 133 canonical and only three of 56 kinetoplastid-specific mitoribosomal proteins, which suggests considerable divergence. Hence, the *E. gracilis* mitochondrion has unique features, representing an intermediate between the mitochondria familiar from yeast or mammals and the atypical organelle present in kinetoplastids (Fig. 5).

The *Euglena* plastid

The *Euglena* chloroplast, as a secondary acquisition, represents a near unique configuration for studying fundamental aspects of organelle origins and evolution. The predicted *E. gracilis* plastid proteome contains 1902 proteins (Fig. 6, Additional file 1: Figure S9B; Additional file 8: Table S6). Typical plastid metabolic pathways and enzymes are present, including 70 proteins involved in the chloroplast electron transport chain and light harvesting antennae. A few expected genes were absent, such as glycolytic glucose-6-phosphate isomerase and carotenoid synthesis 15-*cis*-phytoene desaturase; as both pathways are known to be present, these likely arise from incomplete sequence data [41]. The C_5 tetrapyrrole pathway was completely reconstructed, while the C_4 pathway for aminolevulinate synthesis is absent, consistent with previous findings [102]. Enzymes connecting the cytosolic/mitochondrial mevalonate and plastid methyl-D-erythritol pathway (MEP/DOXP) pathways of terpenoid synthesis were not found, in accordance with separate plastid and cytosolic pools of geranylgeranyl pyrophosphate. Carotenoid and non-plastid isoprenoid (e.g. sterols, dolichols) biosynthetic pathways appear unconnected [103]. Significantly, over 50% of the predicted plastid proteome represent proteins with no homology in the databases, suggesting considerable novel metabolic potential.

Protein targeting to the *E. gracilis* plastid involves trafficking via the Golgi complex. Since the plastid was newly established in the euglenoid lineage, this implies that at least two novel membrane-trafficking pathways should be present, one anterograde *trans*-Golgi to plastid and a retrograde pathway operating in reverse. The relevant machinery for such pathways could be produced via either gene transfer from the green algal host or duplication of host membrane-trafficking machinery. We found no reliable evidence for contributions to the endomembrane protein complement by endosymbiotic gene transfer, but there are extensive gene duplications within the endomembrane machinery. Specifically, additional paralogs of key factors involved in post-Golgi to endosome transport, e.g. AP1 and Rab14, are present, as are expansions in retromer and syntaxin16 that specifically serve to retrieve material from endosomes to the *trans*-Golgi network. Overall, we suggest both a period of kleptoplasty prior to stable establishment of the secondary green plastid and a model whereby novel transport pathways were established by gene duplication, as proposed by the organelle paralogy hypothesis [44].

Conclusions

We present here a detailed analysis of the protein-coding complement of *E. gracilis*, together with insights into genome organization. The genome is very

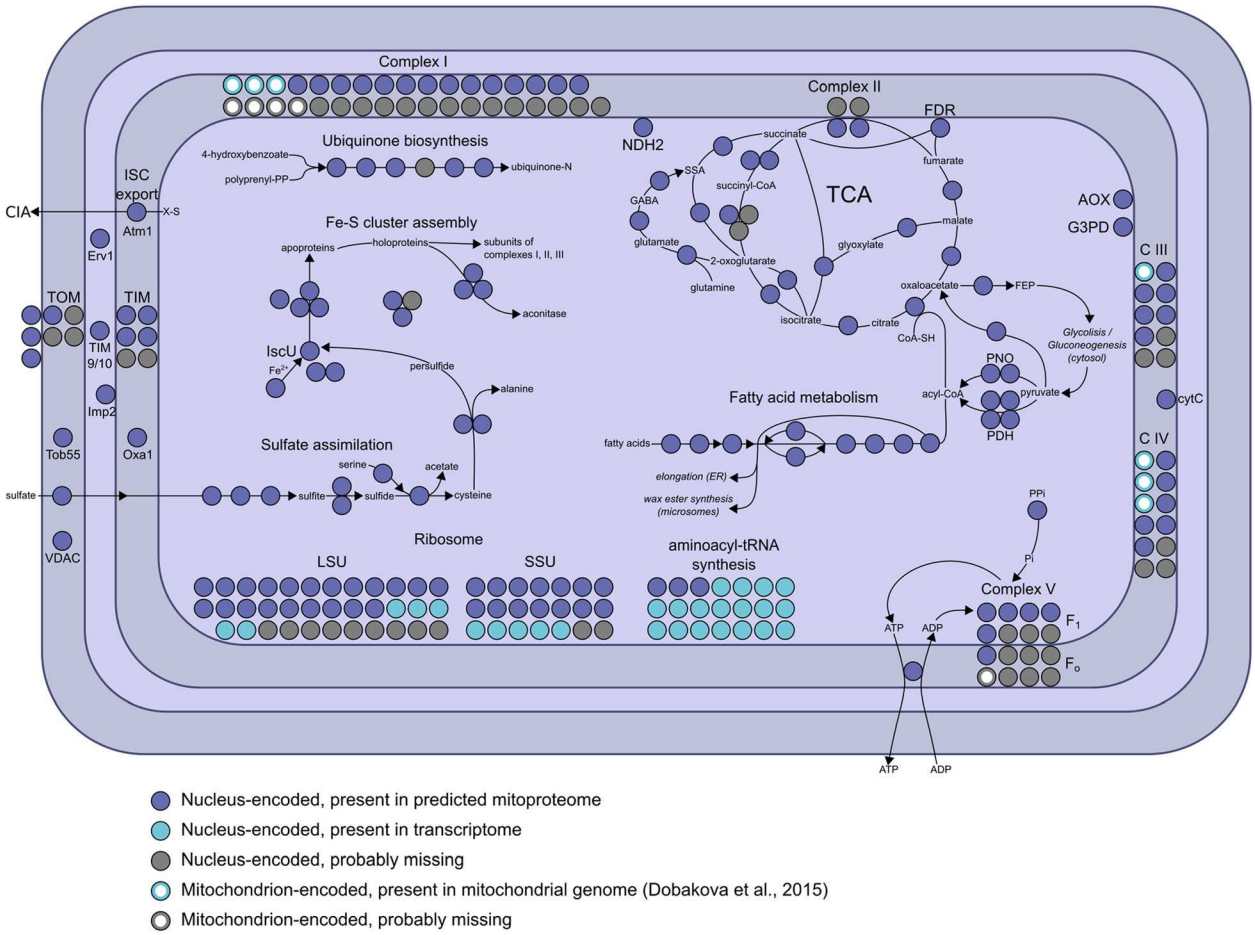

Fig. 5 *Euglena gracilis* has flexible and fault-tolerant mitochondrial metabolism. Proteins involved in mitochondrial pathways and complexes are shown, including: tricarboxylic acid (TCA) cycle, pyruvate dehydrogenase, fatty acid metabolism, complexes I-V of respiratory chain, ubiquinone biosynthesis, sulfate assimilation pathway, Fe-S cluster assembly and export, TIM/TOM complex and mitochondrial import machinery. Colour codes: dark blue, nucleus encoded, present in predicted mitochondrial proteome; light blue, present in transcriptome without evidence for mitochondrial localization; light blue/white, mitochondrion-encoded proteins identified previously [39]; grey, expected in nuclear transcriptome and not found; grey/white, expected in mitochondrial genome and not found. The *E. gracilis* mitochondrion can produce energy under both aerobic and anaerobic conditions and has workarounds for the main mitochondrial pathways, such as TCA cycle and respiratory chain, which may in part explain the outstanding adaptability of this organism

large for a unicellular organism, consistent with many earlier estimates and has exceptionally low coding content, similar to large metazoan genomes. BUSCO, CEGMA and also annotation of many metabolic pathways, complexes and systems indicate that both our data and that from previous work attained very high coverage of the transcriptome. Significantly concatenation of all three datasets resulted in essentially negligible improvement to BUSCO scores, suggesting that the data approach a complete sampling.

We predict a highly divergent surface proteome with expanded signal transduction capabilities likely present at the plasma membrane. *E. gracilis* possesses machinery for synthesis of lipophosphoglycan, suggesting the presence of a defensive phosphoglycan sheath [104]. Significantly, we find evidence for gradual loss of conventional kinetochores, *cis*-splicing and complex RNA processing at the NPC during Euglenozoa evolution. Unexpectedly, there is little evidence for transcriptional control, highly similar to kinetoplastids. Reliance on post-transcriptional processes has been recognized as a feature of *E. gracilis* [105] with mounting evidence that translational and degradative processes are crucial determinants of protein abundance and in agreement with this work [106]. An extensive endomembrane system indicates complex internal organization and multiple endosomal routes representing mechanisms for the sorting, uptake and digestion of material from a range of sources. We also find evidence for novel trafficking pathways between the endomembrane system and the chloroplast; this, together with analysis of the nuclear genome and likely origins of many genes, provides insights into the processes by which secondary

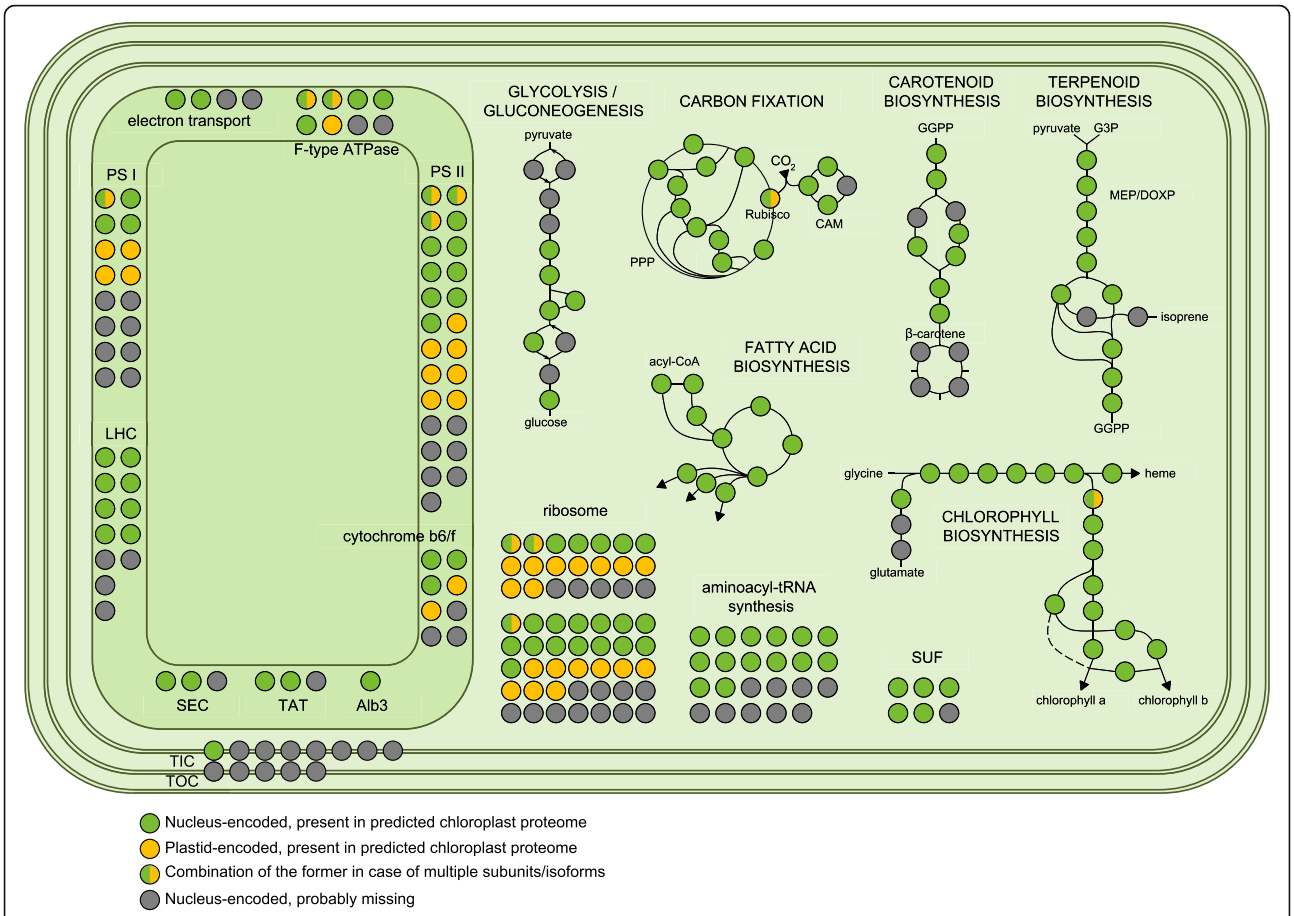

Fig. 6 The *Euglena gracilis* plastid possesses broad metabolic potential. Proteins involved in core plastid metabolic pathways were identified and include glycolysis/gluconeogenesis, carbon fixation, fatty acid biosynthesis, carotenoid biosynthesis, isoprenoid biosynthesis, and chlorophyll biosynthesis. Colour codes: green, nucleus encoded, present in predicted chloroplast proteome; amber, plastid encoded, present in predicted chloroplast proteome; light green/white, combination of green and amber in case of multiple subunits/isoforms; and gray, expected but not found

plastids become enslaved, and is consistent with a protracted period of plastid acquisition.

Materials and methods

Cultivation

E. gracilis strain Z1 was provided by William Martin (Düsseldorf). Cells were cultivated at ambient temperature under continuous illumination from a 60-W tungsten filament bulb at 20 cm from the culture vessel, in Hutner's media [107]. Cells were collected in exponential growth phase at $\sim 9 \times 10^5$ cells/ml, measured using a haemocytometer. For light and dark adaptation, cells were adapted to Hutner heterotrophic medium [107] for 16 days prior to the initiation of a light or dark growth period. Cultures were subcultured and dark-adapted cultures transferred to a light proof box adjacent to the light cultures. Subculturing was done under low light conditions periodically and cultures maintained for up to 2 weeks prior to harvesting. The impact of a prolonged period under dark conditions was assessed by microscopy (Zeiss LSM 700 confocal microscope; × 40 Plan-Neofuar NA1.3 lens under phase contrast, by UV/VIS spectroscopy using a Shimadzu UV-2450, wavelength scan of 190–800 nm and SDS-PAGE).

Isolation of RNA and proteins for gene expression studies

Equivalent numbers (1×10^7 cells) of dark or light cultured cells were harvested by centrifugation at 25 °C, 1000*g* for 10 mins. RNA extraction was performed using the Qiagen RNeasy Mini Kit (Cat. No. 74104). Genomic DNA contamination was eliminated by performing on-column DNase digestion. Extracted RNA was preserved at − 80 °C for RNA sequencing. For proteomics, cells were washed with PBS containing complete protease inhibitors (Roche), extracted with NuPAGE sample buffer (3X), sonicated and lysates containing 1×10^7 cells fractionated on a NuPAGE Bis-Tris 4–12% gradient polyacrylamide gel (Thermo Scientific, Waltham, MA, USA) under reducing conditions. The sample lane was

divided into eight slices that were subjected to tryptic digestion and reductive alkylation.

Proteomics analysis for gene expression studies

Liquid chromatography tandem mass spectrometry ($LC\text{-}MS^2$) was performed in house at the University of Dundee, UK. Samples were analyzed on a Dionex UltiMate 3000 RSLCnano System (Thermo Scientific, Waltham, MA, USA) coupled to an Orbitrap Q-exactive mass spectrometer (Thermo Scientific) at the University of Dundee proteomics facility. Protein mass spectra were analyzed using MaxQuant version 1.5 [108] searching the predicted *E. gracilis* proteome from the de novo transcriptome assembly reported here. Minimum peptide length was set at six amino acids, isoleucine and leucine were considered indistinguishable and false discovery rates (FDR) of 0.01 were calculated at the levels of peptides, proteins and modification sites based on the number of hits against the reversed sequence database. Ratios were calculated from label-free quantification intensities using only peptides that could be uniquely mapped to a given protein. If the identified peptide sequence set of one protein contained the peptide set of another protein, these two proteins were assigned to the same protein group. *P* values were calculated applying *t* test-based statistics using Perseus [109]. There were 8661 distinct protein groups identified by MaxQuant analysis. For further analyses, data were reduced to 4297 protein groups by rejecting those groups not identified at the peptide level in each of the three replicates for one state. Additionally, a cohort of 384 protein groups was extracted that were observed in only one state (232 light and 152 dark).

Ultrastructure of *E. gracilis* cells in light and dark conditions

Two populations of *E. gracilis* cells cultured in either light or dark conditions were initially fixed using 2.5% glutaraldehyde and 2% paraformaldehyde in 0.1 M sodium cacodylate buffer pH 7.2. Both samples were post-fixed for an hour in buffered 1% (*w*/*v*) OsO_4 and embedded in molten agarose prior to incubating overnight in 2% (*w*/*v*) uranyl acetate. Agarose pellets were dehydrated through a graded acetone series and slowly embedded in Low Viscosity resin (TAAB Ltd.) over 4 days. Following polymerization, 70–90-nm-thin sections were cut by ultramicrotome, post-stained using 2% (*w*/*v*) uranyl acetate and Reynolds lead citrate [110] and imaged with a Hitachi H-7650 transmission electron microscope. Image resolution varied between 20 and 0.3 nm per pixel, depending on the magnification.

Transcriptome analysis for gene expression studies

Extracted RNA was sequenced at the Beijing Genomics Institute (https://www.bgi.com/global/). Analysis and comparisons of the data were performed using standard pipelines. An estimated 62 M clean reads were generated which were subject to quality filtering using Trimmomatic [111], to remove low-quality bases and read pairs as well as contaminating adaptor sequences, prior to assembly. Sequences were searched for all common Illumina adaptors and settings for read processing by Trimmomatic were LEADING:10 TRAILING:10 SLIDINGWINDOW:5:15 MINLEN:50. The trimmed filtered reads were then used to quantify the de novo-assembled transcriptome using Salmon [112] with the bias-correction option operating. Expected counts were integerised before being subject to differential expression testing using DESeq2 [113] using default parameters. In the transcriptomics analysis, 66,542 distinct sequence classes were detected and the data was reduced to 41,045 applying the same rejection criteria as the proteome (minimum three replicates).

Nucleic acid isolation and purification for genomic and transcriptomic studies

E. gracilis genomic DNA was isolated using the Qiagen DNA purification system to obtain low and high molecular weight DNA for Illumina paired-end and mate-pair read libraries (100-bp paired-end libraries with insert sizes of 170 bp, 500 bp and 800 bp, and mate-pair libraries with insert sizes of 2 kbp, 5 kbp and 40 kbp). For the shorter length libraries (≤ 5 kbp), cells were harvested by centrifugation for 10 mins at 1000 g and DNA extracted using the Qiagen DNAeasy blood and tissue kit (Qiagen Inc., Cat.No. 69504). The cultured animal cell protocol was modified and involved firstly, using 1×10^7 cells, and secondly, prior to adding Buffer AL, 200 μl of RNase A was added to eliminate RNA contamination. Immediately after the washing step with Buffer AW2, centrifugation was performed for 1 min at 20,000*g* to eliminate traces of ethanol. To obtain high molecular weight DNA fragments for the ≥ 40 kb insert size library, the Qiagen Genomic-DNA isolation kit (blood and cell culture DNA kit - Maxi, Cat. No. 13362) was used. In this case, 1×10^8 cells were harvested. Prior to adding Buffer C1, samples were ground in liquid nitrogen using a planetary ball mill (Retsch) [114] at 300 rpm for 3 min (the grinding was limited to two cycles to minimize DNA shearing). Four wash steps were performed to remove contaminants including traces of RNA. To determine molecular weight, 400 ng of DNA was loaded onto a 0.45% agarose gel in TAE buffer, stained with Thermo Scientific 6X Orange Loading Dye, and electrophoresed at 80 V for 2 h. A NanoDrop spectrophotometer (DeNovix DS-11+) was used to determine concentration and purity. Total RNA from *E. gracilis* was isolated using the Qiagen RNeasy Mini kit (Cat. No. 74104), and the protocol for the purification of total RNA from animal cells using spin technology was employed as above.

Library preparation and sequencing for genomic and transcriptomic studies

Genome and transcriptome library preparation and sequencing were performed at the Beijing Genomic Institute, using Illumina Genome Analyzer HiSeq2000 and MiSeq. In the former case, paired-end genomic sequence of multiple read lengths (49 bp and 100 bp) corresponding to eight insert size libraries (170 bp, 250 bp, 500 bp, 540 bp, 800 bp, 2 kbp, 5 kbp, and 40 kbp) were generated with a combined length of ~ 57 Gbp. Additional PacBio libraries were generated at the University of Seattle (5.5 Gbp combined length) and Université Paris-Sud (3.3 Gbp combined length), and the data were kind gifts. A combined total of 305,447 PacBio circular consensus reads (CCS) were generated with estimated average length of 8870 bases and estimated coverage of ~ 1X.

Genome and transcriptome assembly

Multiple routes were explored for the generation of an acceptable assembly [48]. The most successful strategy, as assessed by core eukaryotic gene mapping analysis (CEGMA) and the proportion of RNAseq reads that mapped to the genome assembly [115, 116], utilised Platanus [117], SSPACE [118] and String Graph Assembler (SGA) [119]. Here, the two MiSeq paired-end read libraries (150 bp paired-end and 300 bp paired-end libraries) and 100 bp (170 bp insert size) paired-end HiSeq read libraries were used for the Platanus assembly. Each of the paired-end read libraries was subject to overlapping paired-end read joining using the ErrorCorrectReads.pl algorithm of the ALLPATHS assembly package [120]. This step in ALLPATHS reduces the complexity of the input data by combining overlapping paired-end reads into single larger reads and performs well on independent benchmark tests of real and simulated data [120]. No other steps in the ALLPATHS assembly algorithm were used. These joined paired-end reads were provided to Platanus as single-end reads. The 500 bp and 800 bp insert size read libraries, which could not be subject to read joining as their insert sizes were too large, were included as single-end reads. This collective set of reads was provided to Platanus, and the method was run using its default parameters. The combined Illumina read data provided an estimated 25x coverage of the single-copy component of the genome by k-mer spectrum analysis using ALLPATHS (Additional file 1: Fig. S11). The resulting contigs from the Platanus [117] assembly were subject to six rounds of scaffolding and gap filling using the SSPACE [118] and SGA [119] algorithms. SSPACE was run with the following settings –a 0.7 –m 30 –n 50 –o 20 using the 500 bp and 800 bp insert size paired-end read libraries and the 2000 bp, 5000 bp and 40,000 bp insert size mate pair read libraries. Following each round of scaffolding, SGA was run on the scaffolds in gap filling mode ("-gapfill") using the same combined input read library as Platanus above. This resulted in a de novo assembly with an N_{50} of 955 bp, comprising 2,066,288 scaffolds (Table S1).

A k-mer spectrum for the genome was calculated from the highest coverage read library (150 bp paired-end read library). It generated a single peak at 8.8× coverage, corresponding to the homozygous single-copy portion of the genome (Additional file 1: Figure S11A). Assuming a Poisson distribution that would be observed if all regions of the genome were single copy and homozygous, the estimated genome size of the single-copy proportion of genome is 487.2 Mb and the estimated size of the whole genome 2.33 Gb. The discrepancy between the Poisson model and the observed corresponds to multi-copy sequences, with a large proportion of low to medium copy number sequences represented at high frequency. There are more than 80,000 unique k-mers of length 31 that appear more than 10,000 times. These high copy number repeat sequences are those we refer to in the results and are most likely responsible for the difficulty with progressing an assembly further than we have been able to achieve.

To estimate the genome size and the proportion of the genome that is comprised of repetitive unique sequence a k-mer spectrum analysis was conducted (Additional file 1: Figure S11A). The largest Illumina paired-end read library (150-bp paired-end) was used for this analysis. Canonical k-mers were counted using jellyfish (Marçais et al. Bioinformatics 27(6): 764–770) at a range of different k-mer sizes (19, 21, 27 and 31). The resulting k-mer count histograms were analysed using GenomeScope [121]. Using these methods the haploid genome size was estimated to be between 330 mb and 500 mb (Additional file 1: Figure S11A). The repetitive component of the genome was estimated to be between 191 and 339 mb, and the unique component of the genome was estimated to be 141 mb to 160 mb (Additional file 1: Figure S11A). Heterozygosity was estimated to be between 2.2 and 2.6%.

The transcriptome assembly was generated by combining multiple different read libraries into a single transcriptome assembly. These included two 100 bp paired-end read libraries generated on an Illumina HiSeq2500 (200 bp insert size) that were previously published in [17]. *Euglena* transcriptome (PRJEB10085, 17) and the six 100-bp paired-end read libraries (200 bp insert size) were generated on an Illumina HiSeq2000 generated in this study (Additional file 2: Table S1, PRJNA310762). These read libraries were combined to give a total of 2.05×10^8 paired-end reads that were provided as input for transcriptome assembly. Illumina adaptors and low-quality bases were trimmed from the reads using Trimmomatic. Ribosomal RNA sequence was removed using SortMeRNA [122] using default

settings, before read error correction using BayesHammer [123] with default settings. Reads were normalized using khmer [124] with settings –C 20 –k 21 –M 8e9, and overlapping paired-end reads joined using ALLPATHS-LG [120] and all reads subject to de novo assembly using SGA, minimum overlap size of 80 nucleotides, no mismatches. These filtered, normalized, and joined reads were then mapped to this assembly using Bowtie2 [125]. Reads that were absent from the assembly were identified and placed with the assembled contigs into a new input file. This file containing the unassembled reads and assembled contigs was subject to assembly using SGA with an overlap size of 70. This process of identifying unmapped reads and reassembling with SGA was repeated each time, decreasing the overlap size by 10 nucleotides until a minimum overlap size of 40 was reached. This strategy was taken to minimize the occurrence of assembly errors that are commonly obtained when a default small k-mer size is used in de Bruijn graph assembly. Contigs were then subject to scaffolding using SSPACE and the full set of non-ribosomal, corrected, normalized paired-end reads using the settings –k 10, –a 0.7, –n 50, –o 20. Scaffolds were subject to gap filling using the SGA gap filling function. Finally, the assembled contigs were subject to base-error correction using Pilon [126] with the default settings. CEGMA [58] suggests ~88% completeness in terms of representation of coding sequence.

Genome and transcriptome structural and functional automatic annotation

In silico analysis such as open reading frame (ORF) determination, gene predictions, gene ontology (GO) and KEGG (biological pathways) and taxa distribution were performed as part of an automatic functional annotation previously described [127] with minor modifications. Six frame translation and ORF determination of assembled transcriptome sequences were predicted using TransDecoder prediction tool [61] and Gene MarkS-T [128], and the longest ORF with coding characteristics, BLAST homology, and PFAM domain information extracted [129]. The predicted ORF was queried against the NCBI non-redundant protein database using BLASTp homology searches, and the top hit for each protein with an *E* value cutoff $< 1e^{-10}$ retained. Using the Blast2GO automatic functional annotation tool [130], the GO annotations of the best BLAST results with an *E* value cutoff $< 1e^{-10}$ were generated from the GO database. The protein domain, biological pathway analyses, and top species distributions were determined using InterPro, BLAST, enzyme code and KEGG [131]. To greatly reduce run times, BLASTp and Interpro scans were processed locally prior to uploading to Blast2GO in .xml file formats.

Assembling sequence data, data mining and phylogenetic inference

Homology searches for orthologs and paralogs of specific biological annotations were performed against the predicted proteome for *E. gracilis* using BLASTp. Clustering at 100% identity was performed for the predicted *E. gracilis* proteins using the Cluster Database at High Identity (CD-HIT) [62] algorithm to remove gapped/incomplete and redundant sequences. Sequences with significant BLASTp top hit search (E value = $1e^{-10}$) were subjected to both Reversed Position Specific BLAST RPS-BLAST and InterProScan [132]. The annotated sequences with domain and/or protein signature matches were extracted using a combination of custom UNIX commands and BioPerl scripts and clustered to 99% identity using CD-HIT. CD-HIT outputs a set of 'non-redundant' (nr) protein representative sequences which were aligned to known eukaryotic protein reference sequences using ClustalX2 [133] and MAFFT [134]. Poorly aligned positions or gaps were removed using the gap deletion command prior to alignment, and the final alignments processed locally for phylogenetic inference with the PhyML Command Line Interface (CLI) using default settings [135], RAxML [136], FastTree [137] and MrBayes [138]. Annotations of the trees were performed using TreeGraph2 [139] and Adobe Illustrator (Adobe Inc.).

Contigs > 10 kbp in the *E. gracilis* genome

For an initial insight into the architecture of the genome contigs > 10 kbp were analyzed. These contigs were interrogated using tBLASTn with the *E. gracilis* proteome predicted from the transcriptome. Sequences with hits were further interrogated using the Exonerate algorithm [59] for insights into splicing mechanisms and coding regions using the --protein2genome and --showquerygff and --showtargetgff options. Sequences, and their respective splicing coordinates in gff3, were uploaded to the Artemis genome viewer [140] for visualization. Coding regions in gff formats were extracted and translated using a combination of BEDtools getfasta [141] and the EMBOSS getorf [142] tools.

Orthologous group clustering

To identify orthologous genes in *E. gracilis* shared across eukaryotic taxa, we clustered the *E. gracilis* predicted proteome with 30 selected eukaryotic taxa using OrthoFinder [70] with taxa distribution including kinetoplastids, other members of the excavates, unikonts, bikonts, green algae, land plants and red algae.

Phylogenetic analyses of ancestry of *Euglena* genes

All 36,526 predicted nucleus-encoded proteins were searched (BLASTp 2.2.29) against a custom database containing 207 organisms (Additional file 3: Table S2).

Homologues with E value $< 10^{-2}$ were retrieved. Since an unrooted phylogenetic tree can be calculated only for three or more organisms, all proteins with less than three recovered homologues (16,636 proteins) were excluded. The remaining (19,890 proteins) were aligned (MAFFT 7.273; default parameters) and trimmed (trimAl 1.2 [143], default parameters). Alignments longer than 74 amino acid residues and with all sequences determined, i.e. there was no sequence containing only undetermined characters, (18,108 alignments) were used for tree reconstruction. The trees were calculated with RAxML [136] (v8.1.17; 100 rapid bootstraps) in Metacentrum (The National Grid Infrastructure in the Czech Republic). Custom scripts (Python 3.4) were used to sort the trees into bins based on the taxonomic affiliation of the clan in which *E. gracilis* branched. The tree was included in a bin if a bipartition supported by bootstrap 75% and higher comprised of *E. gracilis* and members of one defined taxonomic group only. In 34 cases, in which *E. gracilis* was contained in two such bipartitions containing taxa from different defined group, the tree was assigned to the two respective bins.

Mitochondrial proteome prediction

The predicted proteins were subjected to Blast2GO [130] and KEGG automatic annotation server (KAAS [144]) automatic annotation, BLASTp searches against the *T. brucei*, *Homo sapiens*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* reference mitoproteomes and, finally, targeting signal prediction using TargetP [145]. *E. gracilis* protein was predicted as mitochondrial if (i) TargetP mitochondrial score was higher than 0.9 (607 proteins), or (ii) there was an ortholog in at least one reference mitoproteome, not associated with non-mitochondrial functions (343 proteins), or (iii) assigned mitochondrial by Blast2GO (with the exception of the MTERF family) (62 proteins). The missing members of the found mitochondrial pathways and modules were identified by a manual search (81 proteins). To streamline the final annotated output and to ensure retention of only the most reliable predictions, we chose the most confident annotation between Blast2GO, BLASTp and KAAS for each protein. The final mitochondrial dataset includes 1093 proteins.

Plastid proteome prediction

The translated *E. gracilis* transcriptome (predicted proteome) was subjected to signal prediction pipeline using a combination of SignalP [146] and PrediSI [147] while chloroplast transit peptide prediction was performed using ChloroP [148]. The sequences which scored positive by either SignalP (2551 sequences) or PrediSI (4857 sequences) were cut at the predicted signal peptide cleavage site. The sequences were then truncated to maximum length of 200 amino acid residues for faster calculation and analyzed by ChloroP. The preliminary dataset of *E. gracilis* plastid targeted proteins (1679 sequences) consisted of transcripts which scored positive in SignalP + ChloroP (59 sequences), PrediSI + ChloroP (1002 sequences) and SignalP + PrediSI + ChloroP (618 sequences) analysis. In the second step, model dataset of 920 sequences of *Arabidopsis thaliana* proteins localized to the plastid envelope, stroma, thylakoid, grana and lamellae obtained from the public AT_CHLORO proteomic database [149] were searched by BLAST against the whole translated *E. gracilis* transcriptome and the identified orthologs were then combined with the results of orthogroup clustering performed by OrthoFinder (see above). Based on these searches, an additional 144 sequences representing orthologs of *A. thaliana* chloroplast proteins were added to the dataset of *E. gracilis*-predicted plastid proteome regardless of their targeting sequences. This enriched dataset of 1823 proteins was annotated automatically using BLAST at NCBI, KOBAS [150] and KAAS [144] independently. All automatic annotations including KO and EC numbers were then revised and edited or corrected manually and used for metabolic map reconstruction. The missing enzymes and subunits of otherwise chloroplast pathways and complexes were investigated and eventually added manually to the set regardless of their targeting sequences during the manual annotation and pathway reconstruction. This approach resulted in inclusion of another 79 sequences. The final set of predicted *E. gracilis* chloroplast proteins consisted of 1902 entries.

Additional files

Additional file 1: Figure S1. Organisation of open reading frames in the *E. gracilis* genome. **Figure S2.** Functional analysis of *E. gracilis* coding capacity by Gene Ontology. Figure S3. Dark adapted cells have altered proteomes and transcriptomes. **Figure S4.** Orthogroup clusters in *E. gracilis* and selected eukaryotes. **Figure S5.** Phylogeny of selected shared large paralog families. **Figure S6.** Surface families of *E. gracilis*. Figure S7. The *E. gracilis* endomembrane system. **Figure S8.** The *E. gracilis* nuclear pore and kinetochore complexes. **Figure S9.** The predicted proteomes of *E. gracilis* organelles. **Figure S10.** Metabolism in *E. gracilis*. **Figure S11.** Additional assembly features. **Figure S12.** BUSCO comparisons between the present work and prior transcriptomes. (PDF 10993 kb)

Additional file 2: Table S1. Raw data for proteomics and transcriptomics of *E. gracilis* under adaptive conditions. Cells were grown under dark or light conditions as described in methods and subjected to protein or RNA extraction and analysed by mass spectrometry or RNAseq. Each condition was analysed in triplicate ($n = 3$) and data for individual samples together with the merged data are provided (Transcripts, Proteome), together with BLAST annotation of altered transcripts (additional tabs). (XLSX 19876 kb)

Additional file 3: Table S2. Analysis of phylogenetic relationships of *E. gracilis* proteins. The sheet contains three tables. First table summarizes the taxon composition of the custom database used for the search of homologues of *E. gracilis* proteins. Second table summarizes the number of items in each step and the pipeline. The third table gives exact numbers of trees that fell into defined taxonomic bins. (XLSX 16396 kb)

Additional file 4: Table S3. Analysis of GO term frequency, domains and large orthogroup architecture. Sheet 1: GO terms in orthogroups. The sheet has two subtables. In one the GO terms represented above 5% in each orthogroup are shown - all other GO terms with less than 5% frequency have been omitted as the numbers of sequences included are very small. The second shows the number of annotated and non-annotated sequences of each taxonomic group selected. Yellow highlight shows the GO terms of interest belonging to *molecular process* that are analyzed in this study. Sheet 2: Conserved domains from NCBI database (CDD) detected in those sequences with the GO terms of interest highlighted in sheet 1. Output provided by CDD searches. For the sequence identifiers, note that first field separated with "_", represents the taxonomic group to which it belongs. Sheet 3: Incidence of conserved domains detected in CDD searches and orthogroups. This table summarizes the output of the CDD searches. Gray highlight represents the conserved domains in parallel with the respective orthogroup (OG number) of the sequences for which we provide phylogenetic analyses. Sheet 4: Data for annotation of NCIII tree. *Trans*-membrane domains and conserved domains. Sheet 5: Data for annotation of REC tree. *Trans*-membrane domains and conserved domains. (XLSX 127 kb)

Additional file 5: Table S4. Accessions of genes associated with specific cellular functions. Each worksheet contains details of the orthologs and their accession numbers for a specific subset of predicted ORFs associated with an indicated cellular function, metabolic process or organelle. The first two sheets show the overall predictions (all annotated transcripts) and a summary graphic (Distributions). (XLSX 870 kb)

Additional file 6: Supplementary analyses. (DOCX 17 kb)

Additional file 7: Table S5. Surface/endomembrane proteome predictions. Panel A: Predicted numbers of ORFs encoded in the *E. gracilis* predicted proteome that contain a signal sequence (SS) together with additional determinants for stable membrane attachment (i.e. a glycosylphosphatidylinositol anchor (GPI) or trans-membrane domain (TMD)). Panel B: Frequency distribution of predicted *Euglena*-specific surface gene families, shown as number of families according to size. 608 (87.5%). *Euglena*-specific surface genes are predicted to be single-copy, whereas five families are predicted to have more than seven members. Panel C: PHYRE 2.0 summary results for an element of each multi-copy family ($n > 4$) of *E. gracilis*, including family size, residues matching the model and correspondent coverage of the sequence, percentage identity, confidence of prediction, and description of top template model. (XLXS 44 kb)

Additional file 8: Table S6. Predicted proteomes for the *E. gracilis* plastid and the mitochondrion. Panels include summaries for each organelle for numbers of genes in functional categories found, annotations for transcripts predicted as mitochondrial or chloroplastic and finally a reconstruction of major mitochondrial complexes and pathways. (DOCX 141 kb)

Acknowledgements
We are most grateful to Purificatión Lopéz-García, David Moreira and Peter Myler for the most generous donation of PacBio sequence data and also to Robert Field for permission to reutilize transcriptome data. We thank Douglas Lamont and the Fingerprints proteomics facility at the University of Dundee for excellent mass spectrometric analysis. Some computational resources were provided by the CESNET LM2015042 and the CERIT Scientific Cloud LM2015085, provided under the program "Projects of large research, development, and innovations infrastructures".

Funding
This work was supported by the Yousef Jameel Academic Program (through the Yousef Jameel PhD Scholarship), the Cambridge Commonwealth, European and International Trust, the Cambridge University Student Registry, the Cambridge Philosophical Society (all to TEE), the Medical Research Council (Grant #: P009018/1 to MCF), and German Aerospace Center - DLR, Cologne, on the behalf of Federal Ministry of Education and Research (BMBF), Germany (Grant no: 50WB1128 and 50WB1528 to ML), the European Research Council CZ LL1601 BFU2013-40866-P (to DPD) and the Czech Ministry of Education, Youth and Sports - National Sustainability Program II (Project BIOCEV-FAR) LQ 1604, by the project BIOCEV (CZ.1.05/1.1.00/02.0109), by the Centre for research of pathogenicity and virulence of parasites CZ.02.1.01/0.0/0.0/16_019/0000759 and by the Czech Science Foundation project nr. 16-25280S (to VH, AV and PS).

Availability of data and materials
Assembled transcripts and predicted proteome (PRJNA298469), light/dark adapted transcriptomes (PRJNA310762). Genome data and assembly data are available from the European Nucleotide Archive under the project accession ERP109500. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD009998. Supporting analyses of several annotated systems are available in Additional file 5: Table S4 and Additional file 6: Supplementary analysis.

Authors' contributions
MCF, MLG, SK and ML conceived the study. TEE, MZ and AB carried out the experimental. TEE, MZ, AB, AN, AMGNV, MG, BP, PS, CS-M, EO'N, NNN, SSP, NV, VD, SO and MCF analyzed the data. MCF, MG, SK, ML, MZ, APJ, DD, JL, JBD, ML, SV and VH supervised the research. TEE, MZ, AB, AN, AMGNV, BP, PS, CS-M, EO'N, NNN, SSP, JBD, APJ and MCF drafted the manuscript. MCF, VH, TEE and SK edited the final draft. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]School of Life Sciences, University of Dundee, Dundee DD1 5EH, UK. [2]Department of Biochemistry, University of Cambridge, Cambridge CB2 1QW, UK. [3]Department of Biological and Medical Sciences, Faculty of Health and Life Sciences, Oxford Brookes University, Oxford OX3 0BP, UK. [4]Biology Centre, Institute of Parasitology, Czech Academy of Sciences, and Faculty of Sciences, University of South Bohemia, 37005 České Budějovice, Czech Republic. [5]Department of Parasitology, Faculty of Science,, Charles University, BIOCEV, 252 50 Vestec, Czech Republic. [6]Cell Biology Division, Department of Biology, University of Erlangen-Nuremberg, 91058 Erlangen, Germany. [7]Centro Andaluz de Biología del Desarrollo (CABD)-CSIC, Pablo de Olavide University, Seville, Spain. [8]Department of Plant Sciences, University of Oxford, Oxford OX1 3RB, UK. [9]Division of Infectious Disease, Department of Medicine, University of Alberta, Edmonton, Alberta T6G, Canada. [10]Laboratory of Cellular and Structural Biology, The Rockefeller University, New York, NY 10065, USA. [11]Department of Infection Biology, Institute of Infection and Global Health, University of Liverpool, Liverpool, UK. [12]Department of Biological and Geographical Sciences, School of Applied Sciences, University of Huddersfield, Queensgate, Huddersfield HD1 3DH, UK. [13]Department of Life Sciences, The Natural History Museum, Cromwell Road, London SW7 5BD, UK.

Received: 9 February 2018 Accepted: 8 January 2019
Published online: 07 February 2019

References
1. Dobell C. Antony van Leeuwenhoek and his "Little Animals." 1932. doi: https://doi.org/10.1038/130679a0.
2. Kim JT, Boo SM, Zakryś B. Floristic and taxonomic accounts of the genus *Euglena* (Euglenophyceae) from Korean fresh waters. Algae. 1998;13:173–97.
3. Gojdics M. The genus *Euglena*. American Association for the Advancement of Science; 1953. doi:https://doi.org/10.1126/science.120.3124.799-a.
4. Zakryś B, Walne PL. Floristic, taxonomic and phytogeographic studies of green Euglenophyta from the Southeastern United States, with emphasis

on new and rare species. Algol Stud für Hydrobiol Suppl Vol. 1994;72:71–114.
5. Zakryś B. The nuclear DNA level as a potential taxonomic character in *Euglena* Ehr. (Euglenophyceae). Algol Stud für Hydrobiol Suppl Vol. 1988;: 483–504.
6. Buetow DE. The biology of *Euglena*: Academic Press; 1968;49.
7. McFadden GI. Primary and secondary endosymbiosis and the origin of plastids. J Phycol. 2001;37:951–9. https://doi.org/10.1046/j.1529-8817.2001.01126.x.
8. Dragoş N, Péterfi LŞ, Popescu C. Comparative fine structure of pellicular cytoskeleton in *Euglena* Ehrenberg. Arch Protistenkd. 1997;148:277–85. https://doi.org/10.1016/S0003-9365(97)80008-5.
9. Daiker V, Lebert M, Richter P, Häder D-P. Molecular characterization of a calmodulin involved in the signal transduction chain of gravitaxis in *Euglena gracilis*. Planta. 2010;231:1229–36. https://doi.org/10.1007/s00425-010-1126-9.
10. van der Horst MA, Hellingwerf KJ. Photoreceptor proteins, "star actors of modern times": a review of the functional dynamics in the structure of representative members of six different photoreceptor families. Acc Chem Res. 2004;37:13–20. https://doi.org/10.1021/ar020219d.
11. Heijde M, Ulm R. UV-B photoreceptor-mediated signalling in plants. Trends Plant Sci. 2012;17:230–7. https://doi.org/10.1016/j.tplants.2012.01.007.
12. Iseki M, Matsunaga S, Murakami A, Ohno K, Shiga K, Yoshida K, et al. A blue-light-activated adenylyl cyclase mediates photoavoidance in *Euglena gracilis*. Nature. 2002;415:1047–51. https://doi.org/10.1038/4151047a.
13. Masuda S. Light detection and signal transduction in the BLUF photoreceptors. Plant Cell Physiol. 2013;54:171–9. https://doi.org/10.1093/pcp/pcs173.
14. Richter PR, Schuster M, Lebert M, Streb C, Häder D-P. Gravitaxis of *Euglena gracilis* depends only partially on passive buoyancy. Adv Sp Res. 2007;39: 1218–24. https://doi.org/10.1016/J.ASR.2006.11.024.
15. Adl SM, Simpson AGB, Lane CE, Lukeš J, Bass D, Bowser SS, et al. The revised classification of eukaryotes. J Eukaryot Microbiol. 2012;59:429–93. https://doi.org/10.1111/j.1550-7408.2012.00644.x.
16. Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, et al. Extreme diversity of diplonemid eukaryotes in the ocean. Curr Biol. 2016;26: 3060–5.
17. O'Neill EC, Trick M, Hill L, Rejzek M, Dusi RG, Hamilton CJ, et al. The transcriptome of *Euglena gracilis* reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. Mol BioSyst. 2015;11: 2808–20. https://doi.org/10.1039/C5MB00319A.
18. O'Neill EC, Trick M, Henrissat B, Field RA. *Euglena* in time: evolution, control of central metabolic processes and multi-domain proteins in carbohydrate and natural product biochemistry. Perspect Sci. 2015;6:84–93. https://doi.org/10.1016/J.PISC.2015.07.002.
19. Yoshida Y, Tomiyama T, Maruta T, Tomita M, Ishikawa T, Arakawa K. De novo assembly and comparative transcriptome analysis of *Euglena gracilis* in response to anaerobic conditions. BMC Genomics. 2016;17:182. https://doi.org/10.1186/s12864-016-2540-6.
20. Moore AN, McWatters DC, Hudson AJ, Russell AG. RNA-Seq employing a novel rRNA depletion strategy reveals a rich repertoire of snoRNAs in *Euglena gracilis* including box C/D and Ψ-guide RNAs targeting the modification of rRNA extremities. RNA Biol. 2018;15:1309–18. https://doi.org/10.1080/15476286.2018.1526561.
21. Lukeš J, Skalický T, Týč J, Votýpka J, Yurchenko V. Evolution of parasitism in kinetoplastid flagellates. Mol Biochem Parasitol. 2014;195:115–22. https://doi.org/10.1016/j.molbiopara.2014.05.007.
22. Flegontov P, Votýpka J, Skalický T, Logacheva MDD, Penin AAA, Tanifuji G, et al. Paratrypanosoma is a novel early-branching trypanosomatid. Curr Biol. 2013;23:1787–93.
23. Jackson AP, Otto TD, Aslett M, Armstrong SD, Bringaud F, Schlacht A, et al. Kinetoplastid phylogenomics reveals the evolutionary innovations associated with the origins of parasitism. Curr Biol. 2016;26:161–72. https://doi.org/10.1016/j.cub.2015.11.055.
24. Jackson AP. Gene family phylogeny and the evolution of parasite cell surfaces. Mol Biochem Parasitol. 2016;209:64–75. https://doi.org/10.1016/j.molbiopara.2016.03.007.
25. Langousis G, Hill KL. Motility and more: the flagellum of Trypanosoma brucei. Nat Rev Microbiol. 2014;12:505–18.
26. Perdomo D, Bonhivers M, Robinson D. The trypanosome flagellar pocket collar and its ring forming protein—TbBILBO1. Cell. 2016;5:9. https://doi.org/10.3390/cells5010009.
27. Kalb LC, Frederico YCA, Boehm C, Moreira CM do N, Soares MJ, Field MC. Conservation and divergence within the clathrin interactome of Trypanosoma cruzi. Sci Rep. 2016;6:31212. https://doi.org/10.1038/srep31212.
28. Zoltner M, Horn D, de Koning HP, Field MC. Exploiting the Achilles' heel of membrane trafficking in trypanosomes. Curr Opin Microbiol. 2016;34:97–103. https://doi.org/10.1016/j.mib.2016.08.005.
29. Hovel-Miner G, Mugnier MR, Goldwater B, Cross GAM, Papavasiliou FN. A conserved DNA repeat promotes selection of a diverse repertoire of Trypanosoma brucei surface antigens from the genomic archive. PLoS Genet. 2016;12:e1005994. https://doi.org/10.1371/journal.pgen.1005994.
30. Devault A, Bañuls A-L. The promastigote surface antigen gene family of the Leishmania parasite: differential evolution by positive selection and recombination. BMC Evol Biol. 2008;8:292. https://doi.org/10.1186/1471-2148-8-292.
31. Chamakh-Ayari R, Bras-Gonçalves R, Bahi-Jaber N, Petitdidier E, Markikou-Ouni W, Aoun K, et al. In vitro evaluation of a soluble Leishmania promastigote surface antigen as a potential vaccine candidate against human leishmaniasis. PLoS One. 2014;9:e92708. https://doi.org/10.1371/journal.pone.0092708.
32. Mahapatra DM, Chanakya HN, Ramachandra TV. *Euglena* sp. as a suitable source of lipids for potential use as biofuel and sustainable wastewater treatment. J Appl Phycol. 2013;25:855–65. https://doi.org/10.1007/s10811-013-9979-5.
33. Furuhashi T, Ogawa T, Nakai R, Nakazawa M, Okazawa A, Padermschoke A, et al. Wax ester and lipophilic compound profiling of *Euglena gracilis* by gas chromatography-mass spectrometry: toward understanding of wax ester fermentation under hypoxia. Metabolomics. 2015;11:175–83. https://doi.org/10.1007/s11306-014-0687-1.
34. Yamada K, Suzuki H, Takeuchi T, Kazama Y, Mitra S, Abe T, et al. Efficient selective breeding of live oil-rich *Euglena gracilis* with fluorescence-activated cell sorting. Sci Rep. 2016;6:26327. https://doi.org/10.1038/srep26327.
35. Miazek K, Iwanek W, Remacle C, Richel A, Goffin D. Effect of metals, metalloids and metallic nanoparticles on microalgae growth and industrial product biosynthesis: a review. Int J Mol Sci. 2015;16:23929–69. https://doi.org/10.3390/ijms161023929.
36. Rodríguez-Zavala JS, García-García JD, Ortiz-Cruz MA, Moreno-Sánchez R. Molecular mechanisms of resistance to heavy metals in the protist *Euglena gracilis*. J Environ Sci Heal Part A. 2007;42:1365–78. https://doi.org/10.1080/10934520701480326.
37. dos Santos Ferreira V, Rocchetta I, Conforti V, Bench S, Feldman R, Levin MJ, et al. Gene expression patterns in *Euglena gracilis*: insights into the cellular response to environmental stress. Gene. 2007;389:136–45.
38. Zeng M, Hao W, Zou Y, Shi M, Jiang Y, Xiao P, et al. Fatty acid and metabolomic profiling approaches differentiate heterotrophic and mixotrophic culture conditions in a microalgal food supplement "*Euglena*". BMC Biotechnol. 2016;16:49. https://doi.org/10.1186/s12896-016-0279-4.
39. Dobáková E, Flegontov P, Skalický T, Lukeš J. Unexpectedly streamlined mitochondrial genome of the euglenozoan *Euglena gracilis*. Genome Biol Evol. 2015;7:3358–67. https://doi.org/10.1093/gbe/evv229.
40. Faktorová D, Dobáková E, Peña-Diaz P, Lukeš J. From simple to supercomplex: mitochondrial genomes of euglenozoan protists. F1000Research. 2016;5:392. doi:https://doi.org/10.12688/f1000research.8040.1.
41. Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, et al. Complete sequence of *Euglena gracilis* chloroplast DNA. Nucleic Acids Res. 1993;21:3537–44.
42. Rogers MB, Gilson PR, Su V, McFadden GI, Keeling PJ. The complete chloroplast genome of the chlorarachniophyte Bigelowiella natans: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. Mol Biol Evol. 2007;24:54–62. https://doi.org/10.1093/molbev/msl129.
43. Maruyama S, Suzaki T, Weber AP, Archibald JM, Nozaki H. Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. BMC Evol Biol. 2011;11:105. https://doi.org/10.1186/1471-2148-11-105.
44. Howe CJ, Barbrook AC, Nisbet RER, Lockhart PJ, Larkum AWD. The origin of plastids. Philos Trans R Soc Lond Ser B Biol Sci. 2008;363:2675–85. https://doi.org/10.1098/rstb.2008.0050.
45. Dooijes D, Chaves I, Kieft R, Dirks-Mulder A, Martin W, Borst P. Base J originally found in kinetoplastida is also a minor constituent of nuclear DNA of *Euglena gracilis*. Nucleic Acids Res. 2000;28:3017–21.
46. Stankiewicz AJ, Falchuk KH, Vallee BL. Composition and structure of zinc-deficient *Euglena gracilis* chromatin. Biochemistry. 1983;22:5150–6.

47. Mazus B, Falchuk KH, Vallee BL. Histone formation, gene expression, and zinc deficiency in *Euglena gracilis*. Biochemistry. 1984;23:42–7.
48. Ebenezer TE, Carrington M, Lebert M, Kelly S, Field MC. *Euglena gracilis* genome and transcriptome: organelles, nuclear genome assembly strategies and initial features. In: Advances in experimental medicine and biology; 2017. p. 125–40. https://doi.org/10.1007/978-3-319-54910-1_7.
49. Schantz ML, Schantz R. Sequence of a cDNA clone encoding beta tubulin from *Euglena gracilis*. Nucleic Acids Res. 1989;17:6727.
50. Jackson AP, Vaughan S, Gull K. Evolution of tubulin gene arrays in trypanosomatid parasites: genomic restructuring in Leishmania. BMC Genomics. 2006;7:261. https://doi.org/10.1186/1471-2164-7-261.
51. Levasseur PJ, Meng Q, Bouck GB. Tubulin genes in the algal protist *Euglena gracilis*. J Eukaryot Microbiol. 1994;41:468–77.
52. Milanowski R, Karnkowska A, Ishikawa T, Zakryś B. Distribution of conventional and nonconventional introns in tubulin (α and β) genes of euglenids. Mol Biol Evol. 2014;31:584–93. https://doi.org/10.1093/molbev/mst227.
53. Milanowski R, Gumińska N, Karnkowska A, Ishikawa T, Zakryś B. Intermediate introns in nuclear genes of euglenids – are they a distinct type? BMC Evol Biol. 2016;16:49.
54. Canaday J, Tessier LH, Imbault P, Paulus F. Analysis of *Euglena gracilis* alpha-, beta- and gamma-tubulin genes: introns and pre-mRNA maturation. Mol Gen Genomics. 2001;265:153–60.
55. Tessier L, Keller M, Chan RL, Fournier R, Weil J. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in *Euglena*. EMBO J. 1991;10:2621–5.
56. Keller M, Chan RL, Tessier L-H, Weil J-H, Imbault P. Post-transcriptional regulation by light of the biosynthesis of *Euglena* ribulose-1,5-bisphosphate carboxylase/oxygenase small subunit. Plant Mol Biol. 1991;17:73–82. https://doi.org/10.1007/BF00036807.
57. Rawson JR. The characterization of *Euglena gracilis* DNA by its reassociation kinetics. Biochim Biophys Acta. 1975;402:171–8.
58. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23:1061–7. https://doi.org/10.1093/bioinformatics/btm071.
59. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6:31. https://doi.org/10.1186/1471-2105-6-31.
60. Mair G, Shi H, Li H, Djikeng A, Aviles HO, Bishop JR, et al. A new twist in trypanosome RNA metabolism: cis-splicing of pre-mRNA. RNA. 2000;6:163–9.
61. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
62. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
63. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–12.
64. Jackson AP, Quail MA, Berriman M. Insights into the genome sequence of a free-living kinetoplastid: Bodo saltans (Kinetoplastida: Euglenozoa). BMC Genomics. 2008;9:594. https://doi.org/10.1186/1471-2164-9-594.
65. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. The genome of the African trypanosome Trypanosoma brucei. Science. 2005;309:416–22.
66. Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, Field MC, et al. The genome of Naegleria gruberi illuminates early eukaryotic versatility. Cell. 2010;140:631–42. https://doi.org/10.1016/j.cell.2010.01.032.
67. Van Assche E, Van Puyvelde S, Vanderleyden J, Steenackers HP. RNA-binding proteins involved in post-transcriptional regulation in bacteria. Front Microbiol. 2015;6:141. https://doi.org/10.3389/fmicb.2015.00141.
68. Araújo PR, Teixeira SM. Regulatory elements involved in the post-transcriptional control of stage-specific gene expression in Trypanosoma cruzi: a review. Mem Inst Oswaldo Cruz. 2011;106:257–66.
69. Montandon PE, Stutz E. Structure and expression of the *Euglena* nuclear gene coding for the translation elongation factor EF-1 alpha. Nucleic Acids Res. 1990;18:75–82.
70. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16:157. https://doi.org/10.1186/s13059-015-0721-2.
71. Salmon D, Vanwalleghem G, Morias Y, Denoeud J, Krumbholz C, Lhomme F, et al. Adenylate cyclases of Trypanosoma brucei inhibit the innate immune response of the host. Science. 2012;337:463–6. https://doi.org/10.1126/science.1222753.
72. Ponce-Toledo RI, Moreira D, López-García P, Deschamps P. Secondary plastids of euglenids and chlorarachniophytes function with a mix of genes of red and green algal ancestry. Mol Biol Evol. 2018;35:2198–204. https://doi.org/10.1093/molbev/msy121.
73. Jackson C, Knoll AH, Chan CX, Verbruggen H. Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green plastids. Sci Rep. 2018;8:1523. https://doi.org/10.1038/s41598-017-18805-w.
74. Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, et al. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. Nature. 2012;492:59–65. https://doi.org/10.1038/nature11681.
75. Dorrell RG, Gile G, McCallum G, Méheust R, Bapteste EP, Klinger CM, et al. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. elife. 2017;6. https://doi.org/10.7554/eLife.23717.
76. Dunin-Horkawicz S, Lupas AN. Comprehensive analysis of HAMP domains: implications for transmembrane signal transduction. J Mol Biol. 2010;397:1156–74. https://doi.org/10.1016/j.jmb.2010.02.031.
77. Anantharaman V, Aravind L. Cache – a signaling domain common to animal Ca(2+)-channel subunits and a class of prokaryotic chemotaxis receptors. Trends Biochem Sci. 2000;25:535–7.
78. Szöör B, Haanstra JR, Gualdrón-López M, Michels PA. Evolution, dynamics and specialized functions of glycosomes in metabolism and development of trypanosomatids. Curr Opin Microbiol. 2014;22:79–87. https://doi.org/10.1016/J.MIB.2014.09.006.
79. Morales J, Hashimoto M, Williams TA, Hirawake-mogi H, Makiuchi T, Tsubouchi A, et al. Differential remodelling of peroxisome function underpins the environmental and metabolic adaptability of diplonemids and kinetoplastids. Proc R Soc B. 2016;283:20160520.
80. Güther MLS, Urbaniak MD, Tavendale A, Prescott A, Ferguson MAJ. High-confidence glycosome proteome for procyclic form Trypanosoma brucei by epitope-tag organelle enrichment and SILAC proteomics. J Proteome Res. 2014;13:2796–806. https://doi.org/10.1021/pr401209w.
81. Lonergan TA. Regulation of cell shape in *Euglena*. IV. Localization of actin, myosin and calmodulin. J Cell Sci. 1985;77:197–208.
82. Gadelha C, Zhang W, Chamberlain JW, Chait BT, Wickstead B, Field MC. Architecture of a host-parasite interface: complex targeting mechanisms revealed through proteomics. Mol Cell Proteomics. 2015;14:1911–26. https://doi.org/10.1074/mcp.M114.047647.
83. Barsanti L, Passarelli V, Walne PL, Gualtieri P. The photoreceptor protein of *Euglena*. FEBS Lett. 2000;482:247–51.
84. Venkatesh D, Boehm C, Barlow LD, Nankissoor NN, O'Reilly A, Kelly S, et al. Evolution of the endomembrane systems of trypanosomatids – conservation and specialisation. J Cell Sci. 2017;130:1421–34. https://doi.org/10.1242/jcs.197640.
85. Zhou Q, Gheiratmand L, Chen Y, Lim TK, Zhang J, Li S, et al. A comparative proteomic analysis reveals a new bi-lobe protein required for bi-lobe duplication and cell division in Trypanosoma brucei. PLoS One. 2010;5:e9660. https://doi.org/10.1371/journal.pone.0009660.
86. Esson HJ, Morriswood B, Yavuz S, Vidilaseris K, Dong G, Warren G. Morphology of the trypanosome bilobe, a novel cytoskeletal structure. Eukaryot Cell. 2012;11:761–72. https://doi.org/10.1128/EC.05287-11.
87. Morriswood B, Havlicek K, Demmel L, Yavuz S, Sealey-Cardona M, Vidilaseris K, et al. Novel bilobe components in Trypanosoma brucei identified using proximity-dependent biotinylation. Eukaryot Cell. 2013;12:356–67. https://doi.org/10.1128/EC.00326-12.
88. McAllaster MR, Ikeda KN, Lozano-Núñez A, Anrather D, Unterwurzacher V, Gossenreiter T, et al. Proteomic identification of novel cytoskeletal proteins associated with TbPLK, an essential regulator of cell morphogenesis in Trypanosoma brucei. Mol Biol Cell. 2015;26:3013–29. https://doi.org/10.1091/mbc.E15-04-0219.
89. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 2010;38(Database issue):D457–62. https://doi.org/10.1093/nar/gkp851.
90. Bugreev DV, Pezza RJ, Mazina OM, Voloshin ON, Camerini-Otero RD, Mazin AV. The resistance of DMC1 D-loops to dissociation may account for the DMC1 requirement in meiosis. Nat Struct Mol Biol. 2011;18:56–60. https://doi.org/10.1038/nsmb.1946.
91. Koreny L, Field MC. Ancient eukaryotic origin and evolutionary plasticity of nuclear lamina. Genome Biol Evol. 2016;8:2663–71.

92. Obado SO, Brillantes M, Uryu K, Zhang W, Ketaren NE, Chait BT, et al. Interactome mapping reveals the evolutionary history of the nuclear pore complex. PLoS Biol. 2016;14:e1002365. https://doi.org/10.1371/journal.pbio.1002365.
93. Akiyoshi B, Gull K. Discovery of unconventional kinetochores in kinetoplastids. Cell. 2014;156:1247–58. https://doi.org/10.1016/j.cell.2014.01.049.
94. D'Archivio S, Wickstead B. Trypanosome outer kinetochore proteins suggest conservation of chromosome segregation machinery across eukaryotes. J Cell Biol. 2017;216:379–91. https://doi.org/10.1083/jcb.201608043.
95. Lukeš J, Guilbride DL, Votýpka J, Zíková A, Benne R, Englund PT. Kinetoplast DNA network: evolution of an improbable structure. Eukaryot Cell. 2002;1:495–502.
96. David V, Flegontov P, Gerasimov E, Tanifuji G, Hashimi H, Logacheva MD, et al. Gene loss and error-prone RNA editing in the mitochondrion of Perkinsela, an endosymbiotic kinetoplastid. MBio. 2015;6:1–12.
97. Pusnik M, Schmidt O, Perry AJJ, Oeljeklaus S, Niemann M, Warscheid B, et al. Mitochondrial preprotein translocase of trypanosomatids has a bacterial origin. Curr Biol. 2011;21:1738–43.
98. Zarsky V, Tachezy J, Dolezal P. Tom40 is likely common to all mitochondria. Curr Biol. 2012;22:R479–81.
99. Pusnik M, Schmidt O, Perry AJ, Oeljeklaus S, Niemann M, Warscheid B, et al. Response to Zarsky et al. Curr Biol. 2012;22:R481–2.
100. Mani J, Meisinger C, Schneider A. Peeping at TOMs — diverse entry gates to mitochondria provide insights into the evolution of eukaryotes. Mol Biol Evol. 2016;33:337–51.
101. Perez E, Lapaille M, Degand H, Cilibrasi L, Villavicencio-Queijeiro A, Morsomme P, et al. The mitochondrial respiratory chain of the secondary green alga *Euglena* shares many additional subunits with parasitic Trypanosomatidae. Mitochondrion. 2014;19:338–49.
102. Gomez-Silva B, Timko MP, Schiff JA. Chlorophyll biosynthesis from glutamate or 5-aminolevulinate in intact *Euglena* chloroplasts. Planta. 1985; 165:12–22. https://doi.org/10.1007/BF00392206.
103. Kim D, Filtz MR, Proteau PJ. The methylerythritol phosphate pathway contributes to carotenoid but not phytol biosynthesis in *Euglena*. J Nat Prod. 2004;67:1067–9. https://doi.org/10.1021/np049892x.
104. Eggimann G, Sweeney K, Bolt H, Rozatian N, Cobb S, Denny P. The role of phosphoglycans in the susceptibility of Leishmania mexicana to the temporin family of anti-microbial peptides. Molecules. 2015;20:2775–85. https://doi.org/10.3390/molecules20022775.
105. Saint-Guily A, Schantz ML, Schantz R. Structure and expression of a cDNA encoding a histone H2A from *Euglena*. Plant Mol Biol. 1994;24:941–8.
106. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012;13:227–32. https://doi.org/10.1038/nrg3185.
107. Hutner SH, Zahalsky AC, Aaronson S, Baker H, Frank O. Culture media for *Euglena*. In: Methods in Cell Biology. Academic Press; 1966. p. 217–28. https://doi.org/10.1016/S0091-679X(08)62140-8.
108. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol. 2008;26:1367–72. https://doi.org/10.1038/nbt.1511.
109. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13:731–40. https://doi.org/10.1038/nmeth.3901.
110. Reynolds ES. The use of lead citrate at high pH as an electron-opaque stain in electron microscopy. J Cell Biol. 1963;17:208–12 http://www.ncbi.nlm.nih.gov/pubmed/13986422.
111. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20. https://doi.org/10.1093/bioinformatics/btu170.
112. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017; 14:417–9. https://doi.org/10.1038/nmeth.4197.
113. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550. https://doi.org/10.1186/s13059-014-0550-8.
114. Obado S, Field MC, Chait BT, Rout MP. 'High-efficiency isolation of nuclear envelope protein complexes from trypanosomes.' Methods Mol Biol. 2016; 1411:67–80.
115. Hornett EA, Wheat CW. Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. BMC Genomics. 2012;13: 361. https://doi.org/10.1186/1471-2164-13-361.
116. O'Neil ST, Emrich SJ. Assessing de novo transcriptome assembly metrics for consistency and utility. BMC Genomics. 2013;14:465. https://doi.org/10.1186/1471-2164-14-465.
117. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014;24:1384–95. https://doi.org/10.1101/gr.170720.113.
118. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27:578–9. https://doi.org/10.1093/bioinformatics/btq683.
119. Simpson JT, Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res. 2012;22:549–56. https://doi.org/10.1101/gr.126953.111.
120. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A. 2011;108:1513–8. https://doi.org/10.1073/pnas.1017351108.
121. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33:2202–4. https://doi.org/10.1093/bioinformatics/btx153.
122. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28:3211–7. https://doi.org/10.1093/bioinformatics/bts611.
123. Nikolenko SI, Korobeynikov AI, Alekseyev MA. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics. 2013;14(Suppl 1):S7. https://doi.org/10.1186/1471-2164-14-S1-S7.
124. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, et al. The khmer software package: enabling efficient nucleotide sequence analysis. F1000Research. 2015;4. doi:https://doi.org/10.12688/f1000research.6924.1.
125. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.
126. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963. https://doi.org/10.1371/journal.pone.0112963.
127. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–62.
128. Tang S, Lomsadze A, Borodovsky M. Identification of protein coding regions in RNA transcripts. Nucleic Acids Res. 2015;43:e78. https://doi.org/10.1093/nar/gkv227.
129. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2018. https://doi.org/10.1093/nar/gky995.
130. Conesa A, Götz S, García-gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21:3674–6. https://doi.org/10.1093/bioinformatics/bti610.
131. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45:D353–61. https://doi.org/10.1093/nar/gkw1092.
132. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30: 1236–40. https://doi.org/10.1093/bioinformatics/btu031.
133. Larkin MA, Blackshields G, Brown NP, Chenna R, Mcgettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947–8. https://doi.org/10.1093/bioinformatics/btm404.
134. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30:772–80. https://doi.org/10.1093/molbev/mst010.
135. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21. https://doi.org/10.1093/sysbio/syq010.
136. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3. https://doi.org/10.1093/bioinformatics/btu033.
137. Price MN, Dehal PS, Arkin AP. FastTree 2 - approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5:e9490. https://doi.org/10.1371/journal.pone.0009490.
138. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001;17:754–5.

139. Stöver BC, Müller KF. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics. 2010;11:7. https://doi.org/10.1186/1471-2105-11-7.
140. Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. Bioinformatics. 2012;28:464–9. https://doi.org/10.1093/bioinformatics/btr703.
141. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2. https://doi.org/10.1093/bioinformatics/btq033.
142. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000;16:276–7 http://www.ncbi.nlm.nih.gov/pubmed/10827456.
143. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinforma Appl NOTE. 2009;25:1972–3. https://doi.org/10.1093/bioinformatics/btp348.
144. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 2007;35(Web Server issue):W182–5. https://doi.org/10.1093/nar/gkm321.
145. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2007;2:953–71. https://doi.org/10.1038/nprot.2007.131.
146. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8:785–6. https://doi.org/10.1038/nmeth.1701.
147. Hiller K, Grote A, Scheer M, Münch R, Jahn D. PrediSi: prediction of signal peptides and their cleavage positions. Nucleic Acids Res. 2004;32(WEB SERVER ISS):W375–9. https://doi.org/10.1093/nar/gkh378.
148. Emanuelsson O, Nielsen H, von Heijne G. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci. 1999;8:978–84. https://doi.org/10.1110/ps.8.5.978.
149. Bruley C, Dupierris V, Salvi D, Rolland N, Ferro M. AT_CHLORO: a chloroplast protein database dedicated to sub-plastidial localization. Front Plant Sci. 2012;3:205. https://doi.org/10.3389/fpls.2012.00205.
150. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Res. 2011;39(Web Server issue):W316–22. https://doi.org/10.1093/nar/gkr483.

Metabolic quirks and the colourful history of the *Euglena gracilis* secondary plastid

Anna M. G. Novák Vanclová[1], Martin Zoltner[1,2], Steven Kelly[3], Petr Soukal[1], Kristína Záhonová[1,4,5], Zoltán Füssy[5], ThankGod E. Ebenezer[2,6], Eva Lacová Dobáková[5], Marek Eliáš[4], Julius Lukeš[5,7], Mark C. Field[2,5*] and Vladimír Hampl[1*].

1. Faculty of Science, Charles University, BIOCEV, Vestec, Czechia
2. School of Life Sciences, University of Dundee, Dundee, UK
3. Department of Plant Sciences, University of Oxford, Oxford, UK
4. Faculty of Science, University of Ostrava, Ostrava, Czechia
5. Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice, Czechia
6. Department of Biochemistry, University of Cambridge, Cambridge, UK
7. Faculty of Science, University of South Bohemia, České Budějovice, Czechia

* Correspondence to mfield@mac.com (MCF, proteomics) and vlada@natur.cuni.cz (VH, plastid evolution)

Keywords

Euglena gracilis, lateral gene transfer, metabolic reconstruction, plastid, proteome, protein import, shopping bag hypothesis, SUF pathway

Summary

1. *Euglena* spp. are phototrophic flagellates with considerable ecological presence and impact. *E. gracilis* harbours secondary green plastids, but an incompletely characterized proteome precludes accurate understanding of both plastid function and evolutionary history.
2. Using subcellular fractionation, an improved sequence database and mass spectrometry we determined the composition, evolutionary relationships, and hence predicted functions of the *E. gracilis* plastid proteome.
3. We confidently identified 1,345 distinct plastid protein groups and find that at least 100 proteins represent horizontal acquisitions from organisms other than green algae

or prokaryotes. Metabolic reconstruction confirms previously studied/predicted enzymes/pathways and provides evidence for multiple unusual features, including uncoupling of carotenoid and phytol metabolism, a limited role in amino acid metabolism, and dual sets of the SUF pathway for FeS cluster assembly, one of which was acquired by lateral gene transfer from Chlamydiae. Plastid paralogs of trafficking-associated proteins potentially mediating fusion of transport vesicles with the outermost plastid membrane were identified, together with derlin-related proteins, potential translocases across the middle membrane, and an extremely simplified TIC complex.

4. The *Euglena* plastid as the product of many genomes combines novel and conserved features of metabolism and transport.

Introduction

Euglenids are a diverse group of flagellates belonging to the phylum Euglenozoa and have a significant role in the biosphere, as well as being an important model organism and of biotechnological value (Leander *et al.*, 2017). Complex nutritional strategies and an ability to adapt to various environments are major features of the euglenids (Leander *et al.*, 2001; Leander, 2004). The biology of heterotrophic euglenids remains relatively unexplored, but the ease of collection and cultivation of photosynthetic members has made them one of the most widely studied protist groups, despite an absence of a reliable genetic manipulation system. Recent advances in defining the transcriptome and proteome of *Euglena gracilis* promises to improve understanding of the molecular mechanisms employed by this organism and its relatives (Ebenezer *et al.*, 2019).

Euglenophytes harbour green, triple membrane-bound plastids that evolved ~500 Mya from an endosymbiont related to Pyramimonadales (Turmel *et al.*, 2009; Jackson *et al.*, 2018). However, it is possible that other symbionts shared both environment and genes with the ancestors of euglenophytes, as a significant proportion of euglenophyte genes is potentially derived from other algal groups (Maruyama *et al.* 2011; Markunas & Triemer 2016; Lakey & Triemer 2017; Ponce-Toledo *et al.* 2018, Ebenezer *et al.*, 2019), possibly originating during multiple rounds of endosymbiotic gene transfer (Larkum *et al.*, 2007). Higher-order endosymbiotic relationships are the result of complex events that can improve our understanding of organellogenesis. Transfer of genes from endosymbiont to host and establishment of new routes for targeting proteins to the organelle are essential

for endosymbiont-to-organelle transformation. Protein targeting systems of secondary plastids are notably similar in otherwise unrelated lineages (Durnford & Gray, 2006; Sommer *et al.*, 2007; Hempel *et al.*, 2009; Spork *et al.*, 2009; Minge *et al.*, 2010; Felsner *et al.*, 2011; Lau *et al.*, 2016), suggesting general mechanistic constraints rather than common descent. In brief, components of the secretory pathway have been recruited for protein translocation across the outermost plastid membrane: the process starts as co-translational and signal peptide-dependent import into endoplasmic reticulum (ER) and continues via vesicular transport, either directly or through the Golgi complex (Tonkin *et al.*, 2006; Stiller *et al.*, 2014; Maier *et al.*, 2015), while the major TIC/TOC complex translocases of primary plastids retain their functions at the inner two membranes (van Dooren & Striepen 2013; Sheiner & Striepen 2013; Archibald 2015; Gould *et al.* 2015; Maier *et al.* 2015; Bölter & Soll 2016). In plastids of "chromalveolates", a duplicated version of the ER-associated protein degradation (ERAD) pathway mediates transport across the second-outermost membranes (Spork *et al.*, 2009; Stork *et al.*, 2012; Maier *et al.*, 2015).

In euglenophytes, vesicular transport pathways between the ER, Golgi and plastids are present (Sulli *et al.*, 1999) and have been reconstituted *in vitro* and biochemically suggest the involvement of GTPases but not SNARE proteins (Sláviková *et al.*, 2005), however, the molecular machinery is uncharacterized. A plant-like plastid targeting signal (transit peptide) is essential for plastid import, implying the presence of a TIC/TOC-like pathway (Sláviková *et al.*, 2005). However, *in silico* analysis of the *E. gracilis* and *Euglena longa* transcriptomes identified few TIC subunit homologs, with all TOC subunits failing to be identified even by sensitive HMMER-based approach using multiple strategies for profile building (Záhonová *et al.* 2018, Ebenezer *et al.*, 2019), implying a highly divergent or alternative translocation mechanism.

The ultrastructure, metabolism, plastid and mitochondrial genomes of *E. gracilis* have been studied in great detail (Tessier *et al.*, 1991; Hallick *et al.*, 1993; Muchhal & Schwartzbach, 1994; Jenkins *et al.*, 1995; Doetsch *et al.*, 2001; Geimer *et al.*, 2009; Mateášiková-Kováčová *et al.*, 2012; Kuo *et al.*, 2013; Dobáková *et al.*, 2015; Watanabe *et al.*, 2017; Gumińska *et al.*, 2018) Moreover, three transcriptome datasets have been generated recently (O'Neill *et al.*, 2015; Yoshida *et al.*, 2016; Ebenezer *et al.*, 2019), the latter coupled with a draft genome and proteomics analysis of whole cell lysates. Features specific to *E. gracilis* were uncovered by these studies, which also enabled predictions of plastid protein composition (Záhonová *et al.*, 2018; Ebenezer *et al.*, 2019). However, such predictions rely

on both complete understanding of targeting signals, as well as availability of full-length sequence to encompass N-terminal (and possibly other) plastid targeting peptides. As neither of these criteria is currently met, additional subcellular fractionation evidence is essential for validating predictions and to define a robust proteome. Significantly, the sizes of plastid proteomes vary greatly, from over 1,400 proteins in *Arabidopsis thaliana* to under 400 for the *Plasmodium falciparum* apicoplast (Huang *et al.*, 2013; Boucher *et al.*, 2018).

Here we isolated *E. gracilis* plastids and analysed their composition using unbiased mass spectrometry to identify over 1,300 protein groups. We validate many previous predictions, but also uncovered novel metabolic pathways, illuminated targeting mechanisms and identified many lateral gene transfer events that support the sequential endosymbiosis model, or "shopping bag hypothesis", for endosymbiotic evolution.

Materials and methods

Isolation of plastid fraction: Plastidial, mitochondrial and peroxisomal fractions of *E. gracilis* were prepared by gradient ultracentrifugation based on protocols used previously (Davis & Merrett, 1973; Dobáková *et al.*, 2015) and described in detail in the supplementary methods. The resulting gradient and assessment of fraction quality are documented in the supplementary figures S1.2-3.

Mass spectrometry-based identification and quantification of proteins: Plastid and mitochondrial fractions were sonicated in NuPAGE LDS sample buffer (Thermo Scientific) containing 2 mM dithiothreitol and separated on a NuPAGE Bis-Tris 4–12% gradient polyacrylamide gel (Thermo Scientific) under reducing conditions. Each lane was divided into eight slices that were excised, destained and subjected to tryptic digestion and reductive alkylation. These fractions were subjected to liquid chromatography tandem mass spectrometry (LC-MS/MS) on an UltiMate 3000 RSLCnano System (Thermo Scientific) coupled to a Q-exactive mass spectrometer (Thermo Scientific) at the Fingerprints Facility of the University of Dundee. Mass spectra were analysed using MaxQuant (version 1.5, Cox & Mann, 2008), using the predicted translated transcriptome (GEFR00000000.1; Ebenezer *et al.*, 2019). Minimum peptide length was set to six amino acids, isoleucine and leucine were considered indistinguishable and false discovery rates (FDR) of 0.01 were calculated at the levels of peptides, proteins, and modification sites based on the number of hits against a reversed sequence database. The mass spectrometry proteomics data have been deposited to

the ProteomeXchange Consortium via PRIDE partner repository (Perez-Riverol *et al.*, 2019) with the dataset identifier PXD014767. Ratios were calculated from label-free quantification (LFQ) intensities using only peptides uniquely mapped to a given protein. If the identified peptide sequence set of one protein overlapped another protein, these two proteins were assigned to the same protein group. *P* values were calculated applying t-test based statistics in Perseus (Cox & Mann, 2008; Tyanova *et al.*, 2016). 3,736 distinct protein groups were identified and reduced to 2,544 protein groups by rejecting those groups not identified at the peptide level in at least two of three replicates for one organellar fraction. This filtered set includes a cohort of 774 protein groups that were observed in only one organellar fraction (606 in the plastid fraction) and in order to form ratios, a constant small value (0.01) was added to the average LFQ intensities of each fraction (to avoid division by zero).

The resulting CP/MT ratio reflects the enrichment of protein groups in the plastid compared to the mitochondrial fraction and is the main indicator for confidence in plastid localization of a given protein in this study. For clarity, CP/MT ratios were $\log_{10}$ transformed and values >3 (CP/MT ratio > 1000) indicated as "3+" in all tables and figures, indicating extremely high, or "infinite" enrichment. The purity of the plastid and mitochondrial fractions was assessed based on distribution and relative abundance of marker proteins (Figs. S1 and S2). LFQ analysis of the two organellar fractions and whole cell lysate detected 8,216 protein groups and revealed moderate contamination of both mitochondrial and plastid fractions by other cell compartments (Fig. S2.2 and Table S2.1). Comparison of the plastid and mitochondrial fractions suggests a modest level of cross-contamination between these organelles (Fig. S2.3).

As proteins encoded by the *E. gracilis* plastid genome (Hallick *et al.*, 1993) were absent from the translated transcriptome, the MaxQuant LFQ analysis was repeated using the plastid gene set and the resulting quantifications of an additional 32 plastid encoded proteins included in the plastid candidate dataset (supplementary-dataset-1.xlsx; seqid prefix "NP"). Additionally, we searched the translated transcriptome of Yoshida *et al.* (2016; GDJR00000000.1); non-redundant identifications are provided in a separate sheet in supplementary-dataset-2.xlsx. While these proteins were not included in the final plastid candidate dataset, we did consider some of them in the reconstructions presented.

Proteome parsing and annotation: Proteins enriched in the plastid fraction were sorted into four categories based on $\log_{10}$CP/MT ratio (0-1, 1-2, 2-3, and 3+), reflecting

the robustness/discrimination of plastid targeting and association. All candidates were annotated using BLAST (Altschul *et al.*, 1997) against the NCBI non-redundant protein database (https://www.ncbi.nlm.nih.gov/protein, 08/16 version), and assigned a KO number by KAAS (http://www.genome.jp/tools/kaas/; Moriya *et al.*, 2007). Annotations were validated manually using UniProt (http://www.uniprot.org/) and OrthoFinder (Emms & Kelly, 2015) and corrected as necessary. Proteins with no, very few (less than five), or very low-scoring (E-value > 1×10^{-5}) homologs identifiable by BLAST were additionally probed using HHpred against PDB, COG, ECOD, and Pfam databases (https://toolkit.tuebingen.mpg.de/#/tools/hhpred; Soding *et al.*, 2005) and assigned a tentative annotation. Proteins with KO and/or EC numbers assigned, proteins of definite functions in certain metabolic pathways or molecular complexes, as well as proteins with less precise but clear predicted function were sorted manually into 18 custom-defined categories based on the KEGG pathway classification (Table S4). 32 proteins were discarded from the preliminary dataset of 1,377 candidates based on a clearly non-plastid function combined with low enrichment ($\log_{10}$ CP/MT ratio in the 0-1 range; supplementary-dataset-1.xlsx), resulting in a proteome of 1,345 protein groups (supplementary-dataset-1.xlsx).

Bipartite plastid targeting sequences were predicted using a combination of SignalP (version 4.1, http://www.cbs.dtu.dk/services/SignalP/; Petersen *et al.*, 2011) and PrediSI (http://www.predisi.de/; Hiller *et al.*, 2004) for prediction of signal peptides, with ChloroP (http://www.cbs.dtu.dk/services/ChloroP/; Emanuelsson *et al.*, 1999) for prediction of chloroplast transit peptides after *in silico* removal of predicted signal peptides at their putative cleavage sites. Proteins were then binned as having a full bipartite signal, signal peptide only, transit peptide only or no plastid targeting signal sequence.

Major metabolic pathways were reconstructed using the KEGG Mapper web tool (https://www.genome.jp/kegg/tool/map_pathway.html) combined with manual curation. Minor pathways with few members present and/or most members with weak evidence for plastid localization as judged by enrichment ($\log_{10}$ CP/MT ratio close to zero), as well as pathways of known non-plastid localization were excluded.

Determination of the evolutionary origins of plastid proteome members: Each putative plastid protein group was used as query for BLAST searches against 208 transcriptome and genome assemblies from eukaryotes, eubacteria and archaebacteria. Phylogenetic trees containing *E. gracilis* and close homologs were constructed and sorted according to *E. gracilis* protein

affiliation supported by bootstrap >= 75% using a semi-automatic pipeline. The description of the bioinformatic pipeline, involving existing bioinformatic software (Stamatakis, 2006; Capella-Gutierrez *et al.*, 2009; Katoh & Standley, 2013) and custom scripts, and access to all trees is given in supplementary material.

N-terminal signal sequence prediction: The hydrophobicity and amino acid composition of the N-termini of plastid-targeted proteins were investigated as described in detail in supplementary methods. Analysis was restricted to sequences likely not truncated at the N-terminus, that are translated in the cytoplasm and imported into the plastid, i.e. meeting the following criteria: robust plastid localization ($\log_{10}$ CP/MT >1.0) and/or predicted photosynthetic function and encoded by complete transcripts containing the spliced leader (ATTTTTTTTCG; Tessier *et al.*, 1991). This cohort consisted of 375 sequences, and a second sequence set of the same size but disregarding plastid localization, was randomly selected from the transcriptome as a control. The differences in amino acid composition were analyzed using $\chi 2$ test (R-Core-Team, 2013).

Results

A plastid proteome for E. gracilis*:* The *E. gracilis* plastid proteome determined here includes 1,345 proteins, 48 of which are encoded by the plastid genome (Hallick *et al.*, 1993); 774 (57.2%) could be functionally annotated while 571 (42.5%) have unknown or hypothetical function and 109 of these (8%) have no identifiable homologs in the nr database at NCBI (supplementary-dataset-1.xlsx). To assess the completeness of the current dataset we found that 48/67 (72%) plastid encoded proteins were present. We also identified 74/83 (89%) and 113/131 (86%) high-confidence plastid-targeted candidates from two earlier studies (Durnford & Gray, 2006; Záhonová *et al.*, 2018; supplementary-dataset-2.xlsx), indicating concordance approaching 90%. By contrast, only 474/1902 (25%) proteins *in silico* predicted as plastid-targeted from transcriptome data (Ebenezer *et al.*, 2019) were detected by proteomics, and 871 proteins identified by proteomics were absent from this *in silico* set, suggesting that available prediction tools are inaccurate in the case of *Euglena*. This could arise partially as a result of N-terminal sequence truncation but could also signify our incomplete understanding of euglenid plastid-targeting signals. It is also relevant that the plastids analysed here were from light grown cells only, and previous work has demonstrated that significant changes to protein composition are not accompanied by transcript changes:

some proteins are therefore likely absent from the present plastid material (Ebenezer *et al.*, 2019). It is also possible, that some plastidial proteins are expressed but their physiological abundance or number of ionizable peptides they produce is too low.

Significantly, this new proteome is of similar size to the *Arabidopsis thaliana* plastid proteome (1,462 protein groups; Huang *et al.*, 2013) which has itself been determined using a combination of *in silico* prediction and proteomic analysis, but is substantially larger than proteomes of plastid or plastid-derived organelles from several unicellular organisms, including *Chlamydomonas reinhardtii* (996; Terashima *et al.*, 2011), *Bigelowiella natans* (324; Hopkins *et al.*, 2012) and the secondarily non-photosynthetic parasite *Plasmodium falciparum* (345; Boucher *et al.*, 2018).

Biosynthetic pathways embedded in the Euglena *plastid:* Current state of knowledge regarding metabolic pathway localization in Euglena is largely based on biochemical experiments and *in silico* prediction (recently reviewed in Inwongwan *et al.*, 2019). In this study, we rely on comprehensive proteomic dataset to reconstruct metabolic pathways localized in *E. gracilis* plastid. The 774 functionally annotated proteins were sorted into 18 categories (Fig. 1), and a map of major metabolic pathways and complexes was reconstructed (see Fig. 2 for overall schematic and Figs. S5.1-11 for detailed pathways), and while this includes core conserved pathways, we also found features indicative of considerable novelty. As expected, the photosynthetic apparatus is well represented. It should be noted that the precise identification of light-harvesting proteins is not straightforward as these are encoded and translated as polyproteins and are difficult to distinguish using standard proteomic approach without a specific focus on these proteins (Koziol & Durnford, 2008). In accordance with previous findings (Wildner & Hauska, 1974), no gene or protein for plastocyanin was identified (white circle in e^- transport section of Fig. 2, Fig. S5.2), so electron transfer between the cytochrome b_6f complex and the photosystem I relies solely on two isoforms of cytochrome c_6. Interestingly, *E. gracilis* also encodes a homolog of cytochrome c_{6A} (seqid 17930), previously only known from land plants and green algae (Howe *et al.*, 2006). Cytochrome c_{6A} is thought to function as a redox-sensing regulator within the thylakoid lumen and the structure of its N-terminus in *E. gracilis* is in accord with this localization, but was not detected here due to low abundance and difficulty to detect even in plants (Howe *et al.*, 2006). Notable is also the presence of two homologs of the alpha subunit of the F-type ATPase in the proteome, a canonical copy encoded by the plastid genome and a divergent

copy (seqid 3018) encoded in the nucleus. Finally, the *E. gracilis* plastid contains two distinct terminal oxidases (seqids 2887 and 18315), i.e. components of the chloro-respiration pathway (Nawrocki *et al.*, 2015). One represents the conventional enzyme (phylogenetically affiliated with dinoflagellates), whereas the other falls among mitochondrial alternative oxidases (Fig. S6). The latter is apparently not a contaminant as it is significantly enriched in the plastid fraction and has a predicted plastid-targeting signal, while at least three different isoforms of the mitochondrial alternative oxidase bearing mitochondrial transit peptides are also present in *E. gracilis*.

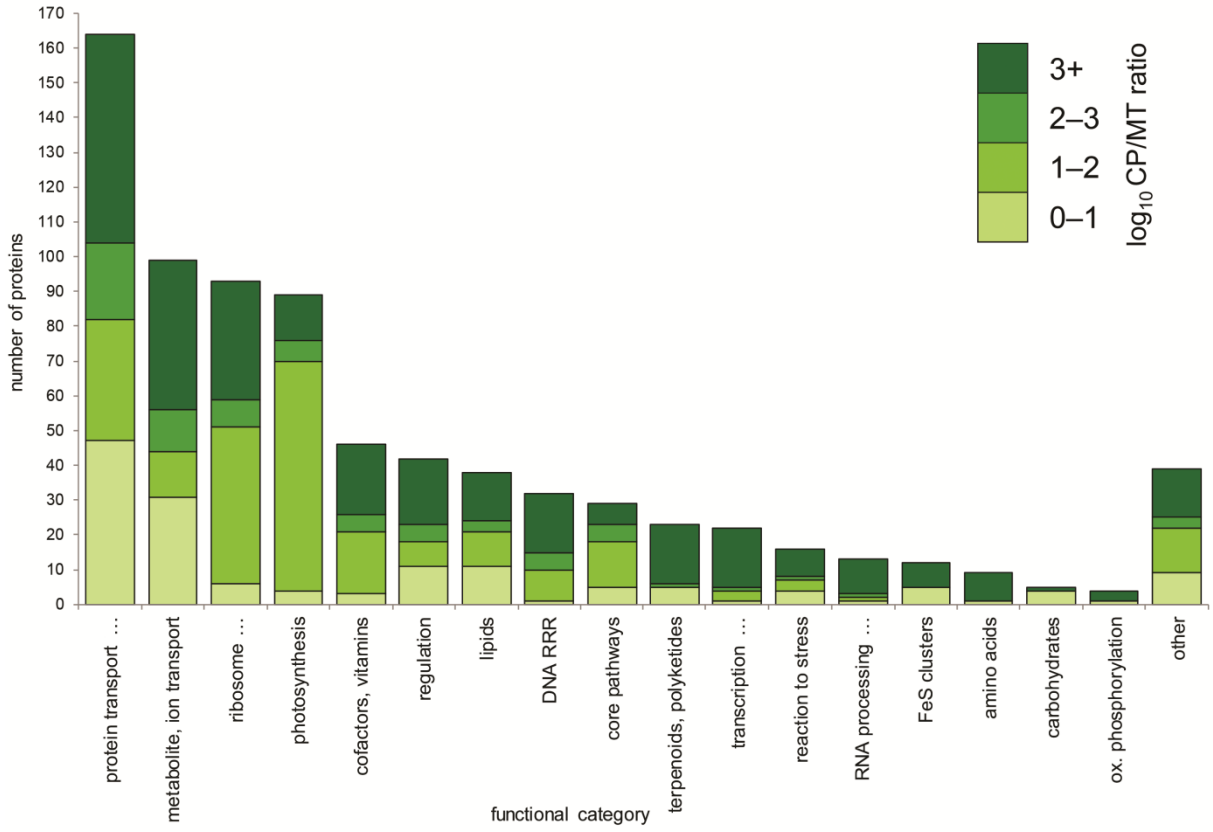

Fig. 1: Distribution of functional categories among 774 plastid proteins with predicted function (57.5% of the whole proteome) and $\log_{10}$ CP/MT ratio distribution in each category, represented by shades of green (0-1 meaning 1-10× higher amount of the protein in plastid fraction compared to the mitochondrial one, 1-2 for 10-100×, 2-3 for 100-1000×; and >3 for more than 1000× or "infinite" value in case the protein was detected in plastid fraction only; the full category names are listed in Materials and Methods, for detailed description with examples see table S3).

The Calvin cycle is fully represented, but a route to glucose metabolism is missing due to the lack of glucose-6-phosphate isomerase. This is probably related to the absence of starch synthesis and a switch to the non-plastid β-1,3-glucan paramylon polysaccharide (Barsanti *et*

al., 2001). The key enzyme for paramylon synthesis, glucan synthase-like 2 (GSL2; Tanaka *et al.*, 2017), is absent from the plastid proteome, but its paralog, GSL1 (seqid 4050), was identified with high-confidence. Unfortunately, the substrate specificity and function of GSL1 are unknown. The nearly complete representation of the C5 pathway of tetrapyrrole synthesis, using glutamate as the precursor of the key intermediate aminolevulinate, is in accord with previous evidence (Gomez-Silva *et al.*, 1985; Kořený & Oborník, 2011).

The *E. gracilis* plastid is also a site for biosynthesis of fatty acids and glycerolipids, the latter including the major plastid phospholipid phosphatidylglycerol and three glycolipids, monogalactosyl- and digalactosyl-diacylglycerol (MGDG and DGDG) and sulfoquinovosyl-diacylglycerol (SQDG) (Matson *et al.*, 1970; Blee & Schantz, 1978; Shibata *et al.*, 2018). We reconstructed pathways for synthesizing these compounds, albeit with several enzymes only predicted by the transcriptome. Interestingly, our data suggests that *E. gracilis* lacks a key enzyme for SQDG synthesis, UDP-sulfoquinovose synthase (SQD1), which catalyses formation of UDP-sulfoquinovose from UDP-glucose. However, sulfite, one of the SQD1 substrates, is readily incorporated into SQDG when incubated with isolated *E. gracilis* plastids (Saidha & Schiff, 1989), suggesting the presence of an alternative pathway to UDP-sulfoquinovose, a substrate of the conventional sulfoquinovosyltransferase (SQD2) catalyzying the ultimate step of SQDG formation (Fig. 2).

Most algal and plant plastids support production of various amino acids, some of which (such as phenylalanine, tyrosine, and tryptophan) are produced exclusively in the organelle, but the *Euglena* plastid lacks these pathways and its contribution to amino acid metabolism is very low. Only an incomplete serine synthesis pathway was reconstructed, with no identifiable plastid-targeted phosphoserine aminotransferase. In addition, only phosphoserine phosphatase was identified in the proteome, whereas the plastid localization of the enzyme catalysing the initial step of the pathway (phosphoglycerate dehydrogenase, divided into two contigs, seqids 23072 and 17279) relies on *in silico* prediction. The *E. gracilis* plastid proteome does contain an isoform of cysteine synthase A, but a plastidial version of the enzyme that converts serine to O-acetylserine, the substrate of cysteine synthase A, was not detected, suggesting that O-acetylserine needs to be imported into the organelle. The plastid further harbours serine/glycine hydroxymethyltransferase (SHMT), whose primary role likely is to provide a formyl group for synthesis of formylmethionyl-tRNA for translation initiation in the plastid rather than to make glycine from serine (Fig. S5.1). Direct BLAST search of the whole transcriptome confirmed that the enzymes of amino

303 acid biosynthesis are indeed present (*E. gracilis* does not require supplementation by any
304 amino acid when grown autotrophically on minimal medium; Hall & Schoenborn, 1939), but
305 their localization is non-plastidial, in accord with the previous *in silico* prediction (Ebenezer
306 *et al.*, 2019). Possible exceptions are shikimate kinase (seqid 11810) (Inwongwan *et al.*,
307 2019), which exhibits the typical plastidial targeting signal, and the following enzyme in the
308 shikimate pathway 5-enolpyruvylshikimate-3-phosphate synthase, which has been previously
309 shown to localise in the *E. gracilis* plastid (Reinbothe *et al.*, 1994). However, the former was
310 not detected in the plastid fraction, and the latter is absent from the transcriptomes.
311 Regardless, synthesis of aromatic amino acids in *E. gracilis* may be fully secured by a
312 complete shikimate pathway in the cytosol.

313 In eukaryotes, two pathways are responsible for synthesis of isopentenyl
314 pyrophosphate (IPP), a precursor to steroids and terpenoids; the MVA pathway, localized to
315 the mitochondria/cytosol, and the MEP (DOXP) pathway, compartmentalized into the plastid
316 (Zhao *et al.*, 2013). *E. gracilis* harbours both pathways, and there is experimental evidence for
317 a role for the MEP pathway in carotenoid synthesis (Kim *et al.*, 2004). Our data demonstrate a
318 plastid localization for the MEP pathway, and also identify the enzyme for the synthesis of the
319 carotenoid precursor geranylgeranyl pyrophosphate (GGPP) from IPP and most enzymes
320 required for synthesis of carotenoids previously reported present in *Euglena* (Krinsky &
321 Goldsmith, 1960): antheraxanthin, neoxanthin, β-, γ- and ζ-carotene, retinol, lutein and
322 cryptoxanthin. Our proteomic and transcriptomic data, however, failed to provide evidence
323 for β-carotene ketolase synthesizing two carotenoids, echinenone and canthaxanthin, that have
324 been detected (Krinsky & Goldsmith, 1960), although some of these enzymes are represented
325 in the transcriptome (marked as grey circles in Fig. 2).

326

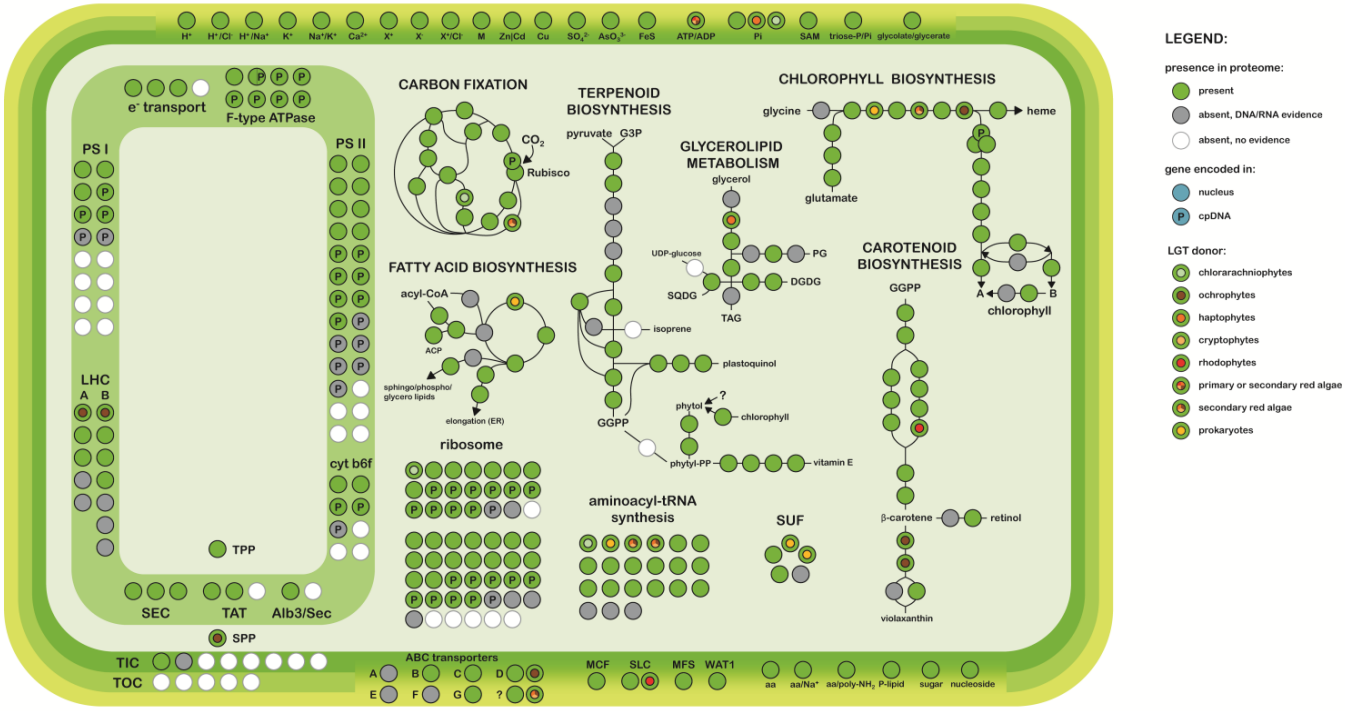

Fig. 2: Overview of the *E. gracilis* plastid metabolism as reconstructed from mass spectrometry-based proteome. Enzymes present in the plastid proteome in at least one isoform are marked as green circles, grey circles represent enzymes which were identified on the RNA or DNA level (in this study or previously) but are absent from the proteome; white circles represent genes completely absent in *Euglena*; circles marked by the letter P represent genes coded in the plastid genome while the rest of the circles represent genes coded in the nucleus. Circles with colored dots in the middle represent genes with at least one of their isoforms presumably gained via lateral transfer from a donor group other than green algae and "miscellaneous" algae: pale green for chlorarachniophytes, brown for ochrophytes, dark orange for haptophytes, pale orange for cryptophytes, red for rhodophytes, and a combination of the former in case of proteins of "mixed red" origin (see Fig. S4 for larger and colorblind-friendly version). Multiple overlapping circles represent enzymes composed of multiple subunits with different characteristics. Note that a single protein may be represented by multiple circles due to its role in multiple pathways or reactions. A more detailed schematic including additional pathways and transcript identifiers for each enzyme is available as Figs. S5.1-11.

Furthermore, the proteome suggests that IPP, made by the MEP pathway, is used for the synthesis of plastoquinone (via solanesyl-PP), as is usual for plastids in general (Lohr *et al.*, 2012). In contrast, uniquely among plastid-bearing eukaryotes, the MEP pathway of *Euglena* does not provide GGPP for synthesis of phytyl-PP, a precursor of tocopherols (vitamin E), phylloquinone (vitamin K), and chlorophyll (Kim *et al.*, 2004). Our analyses support this earlier experimental finding and provide an insight into the alternative pathway. Phytyl-PP is conventionally synthesised by ChlP, a GGPP reductase conserved between eukaryotic and prokaryotic phototrophic organisms (Heyes & Neil Hunter, 2009). The *E. gracilis* plastid proteome and transcriptome lack this enzyme, explaining why GGPP synthesised by the MEP pathway cannot serve as a phytyl-PP precursor in this organism. An

alternative route to phytyl-PP operates in higher plant chloroplasts, which recycles phytol released during chlorophyll degradation and significantly contributes to tocopherol biosynthesis (Vom Dorp *et al.*, 2015). This pathway consists of VTE5 (phytol kinase, seqid 13197) and VTE6 (phytyl phosphate kinase, seqid 14381), both of which are present in the plastid proteome. The proteome also contains a homolog of a recently characterized chlorophyll dephytylase (CLD1, seqid 12254; Fig. S5.5; Lin *et al.*, 2016), which further suggests that phytol is salvaged from degraded chlorophyll. This obviously cannot be the only means of producing phytyl-PP, as the compound itself is required for chlorophyll synthesis. Interestingly, the non-photosynthetic chlorophyll-lacking *E. longa* retains both VTE5 and VTE6 (GenBank accession numbers GGOE01004182.1 and GGOE01053790.1) and suggests that phytol is a genuine intermediate for *de novo* phytyl-PP in *Euglena*. As direct phytol synthesis has not been characterized in any organism, how and in which cellular compartment phytol is produced in *Euglena* is thus unclear.

E. gracilis can produce tocopherol (vitamin E) (Watanabe *et al.*, 2017) and homologs of several enzymes for tocopherol synthesis have been described in the transcriptome (O'Neill *et al.*, 2015). There is some uncertainty concerning the subcellular localization of *E. gracilis* tocopherol production (Watanabe *et al.*, 2017), which we now resolve by reconstructing the full pathway, from the precursors homogentisate and phytyl-PP to α-tocopherol, by identification of all four enzymes in the plastid (Fig. 2). Another terpenoid-quinone produced by *E. gracilis* is the phylloquinone derivative 5'-monohydroxyphylloquinone (Ziegler *et al.*, 1989). In plants and green algae most steps of phylloquinone synthesis occur in the plastid, except three reactions from o-succinylbenzoate to dihydroxynaphthoate, which are catalyzed by peroxisomal enzymes (Emonds-Alt *et al.*, 2017; Cenci *et al.*, 2018). *E. gracilis* encodes a homolog of the multifunctional protein PHYLLO catalyzing all four steps of the first part of the pathway (seqid 121), but the protein is absent from the plastid proteome and lacks a targeting presequence, consistent with biochemical evidence placing synthesis of o-succinylbenzoate in the cytosol (Seeger & Bentley, 1991). The same study tentatively associated the o-succinylbenzoate to dihydroxynaphthoate part of the pathway to the plastid envelope rather than peroxisomes, but the respective enzymes must be unrelated or extremely divergent to the canonical peroxisomal enzymes MenE, MenB, and MenI, as their homologs were not detected in either the plastid proteome or transcriptome. Genome analyses indicate the existence of an novel pathway of dihydroxynaphthoate synthesis in some algae (e.g. glaucophytes) and cyanobacteria (Cenci *et al.*, 2018), so it is possible that *E. gracilis* converts

o-succinylbenzoate to dihydroxynaphthoate by employing enzymes of this route. The final part of the pathway, starting with phytylation of dihydroxynaphthoate, is predicted as plastid-localized in *E. gracilis* (Fig. S5.5), although only one of the respective enzymes (MenA, seqid 7550) was identified in the proteome and the identity of the enzyme catalysing the presumably final step (phylloquinone hydroxylation) is unknown.

Protein import and vesicle targeting to the Euglena *plastid:* A major plastid import pathway is mediated by the TIC/TOC translocation complexes. Surprisingly, only two subunits were identified in the transcriptome: Tic32 and three paralogs of Tic21 (Záhonová *et al.*, 2018). Proteomics confirmed the plastid localization for all Tic21 isoforms, but not Tic32 (Fig. 3). Tic21 is structurally similar to Tic20, the main channel-forming subunit in canonical TIC, and it was sporadically isolated in complex with Tic20 and other central subunits. This suggests that Tic21 is non-essential in plant plastids (Teng *et al.*, 2006; Kikuchi *et al.*, 2009; Nakai, 2018), but its assuming main translocase function in a highly reduced complex is still conceivable. Tic32, on the other hand, is a regulatory subunit with versatile enzymatic activity (Balsera *et al.*, 2010) that could readily serve a TIC-independent purpose or was simply not detected.

Two proteins (seqids 13308 and 11030), representing a novel subgroup of rhomboid-related pseudoproteases distantly related to derlins of the ERAD pathway were also found (Fig. S11.1, supplementary alignment file derlins.txt). In chromalveolate plastids, the ERAD machinery is partially duplicated and recruited for protein transport forming a system termed SELMA, with two derlin family proteins (Der1-1 and Der1-2) presumably constituting the protein-conducting channel (Maier *et al.* 2015). The *E. gracilis* proteins, however, represent distinct and more distantly related homologs of ERAD derlins. Although the phylogeny is poorly supported it suggests that these proteins originated by duplication of a host gene, because a third paralogue (seqid 22665, absent from plastid proteome) is present in the transcriptomes of euglenids including heterotroph *Rhabdomonas costata* (Fig. S11.2, derlins.txt). It is intriguing to speculate that the detected derlin-like proteins in *E. gracilis* plastid gained a role similar to SELMA through convergent evolution.

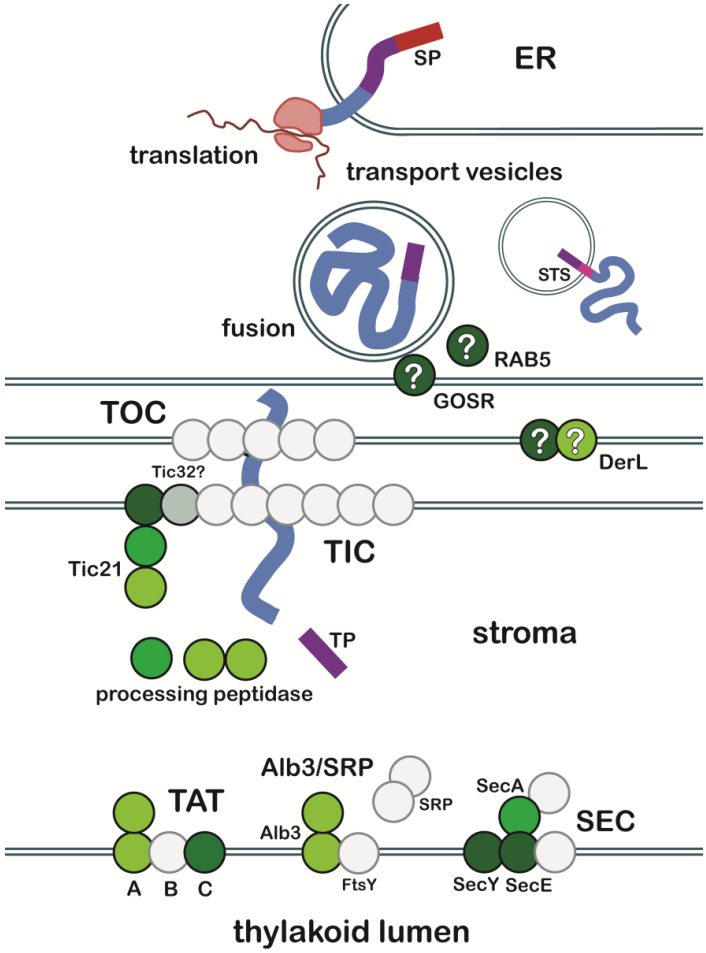

Fig. 3: Reconstruction of *E. gracilis* plastid protein import machinery: The N-terminal signal-peptide (SP, red) bearing nuclear-encoded plastid-targeted protein (blue chain) is synthesized on the rough endoplasmic reticulum (RER). If the N-terminus does not contain stop transfer signal (STS), it is co-translationally imported into the ER lumen as indicated. Otherwise, the major part of the protein remains in the cytosol while being anchored in the ER membrane. The SP is cleaved by signal peptidase in the ER lumen. The protein then passes through Golgi and is loaded into/onto a vesicle which eventually fuses with the outermost plastid membrane. Plastidial homologs of GOSR and Rab5 GTPase might mediate the fusion. The protein passes the middle membrane via unknown mechanism dependent on transit peptide (TP, purple) but not TOC complex, possibly employing derlin-like proteins (DerL). Then it passes the inner membrane via highly reduced TIC translocase, possibly consisting of multiple isoforms of a single subunit. The protein is folded and has its transit peptide cleaved in the plastid stroma and, in case additional signal is revealed, enter thylakoid membrane or lumen via TAT, SEC, or SRP/Alb3 pathway. Each circle represents a putative subunit described in model plastids, green circles represent proteins identified in the plastid proteome, different color shades code CP/MT ratio, the darkest shade depicting the most credible evidence for plastid localization, grey circles represent subunits with transcript-level evidence only. Finally, white circles represent subunits which are completely absent.

The delivery of some plastid-targeted proteins in euglenophytes involves fusion of Golgi-derived vesicles with the outermost plastid membrane and requires GTP hydrolysis, but is not N-ethylmaleimid-sensitive, suggesting the involvement of a Rab/ARF GTPase but not a requirement for conventional SNARE-mediated fusion (Sláviková *et al.*, 2005). Multiple paralogs of some membrane trafficking components are present in *E. gracilis*, and we

previously speculated that some of these may be involved in plastid transport (Ebenezer *et al.*, 2019). Interestingly, the plastid proteome contains homologs of coat complex subunits, several Rab GTPases and SNARE proteins, some of which cannot be dismissed as contaminants. For example, SNARE protein GOSR1 has two paralogs in *E. gracilis,* one of which is detected in the plastid proteome with $\log_{10}$ CP/MT ratio > 3 (seqid 18194), and the other (seqid 19152) was not detected; so the apparent insensitivity to NEM might potentially be a technical issue or that GOSR1 functions in an NSF-independent manner. A second notable candidate is Rab5, a conserved Rab GTPase associated with the endosomal system (Langemeyer *et al.*, 2018), again with convincing $\log_{10}$ CP/MT ratio > 3. These two proteins thus may play roles in targeting protein-transporting vesicles to the outermost plastid membrane (Fig. 3), although the current state of understanding of this pathway is rudimentary.

Evolutionary origins of plastid proteins: The semiautomatic phylogenetic pipeline recovered among plastidial proteins 379 (28%) potential lateral gene transfer (LGT) candidates: 346 of these (91%) are related to phototrophic eukaryotes (primary or secondary algae), 20 (5%) to other eukaryotes, and 13 (3%) to prokaryotes (Fig. 4). Most algae-affiliated proteins are related to green algae (95, 28%), as expected given the plastid ancestry. The number of genes likely derived from rhodophytes, cryptophytes, and glaucophytes is very low (around 1-2% each), arguably at the level of technical error. By contrast gene cohorts related to chlorarachniophytes, ochrophytes, and haptophytes (6-7% each, 19% in total) are sufficient to consider as genuine contributions to the *E. gracilis* plastid. Some genes were related to a clan composed of several groups of phototrophs, and therefore assigned to one of several "miscellaneous" categories (Fig. 4). The distribution of the metabolic functions of proteins affiliated to non-green algae has no clear pattern (Fig. 2). The comparison of these results to the previously published findings on the same topic (Markunas & Triemer, 2016; Ponce-Toledo *et al.*, 2018) is available in supplementary-dataset-2.xlsx.

Additionally, 37 plastidial proteins were determined as affiliated to Discoba, i.e. pre-existing proteins that might have been repurposed or duplicated and recruited for plastidial function. These include mostly membrane transporters, vesicle-associated proteins and chaperones (see supplementary-dataset-1.xlsx). Most of these have $\log_{10}$ CP/MT ratio close to zero and might represent dual-localized proteins or contaminations from other cell fractions, but the functions of the more credible candidates suggest that the host genome contributed to

plastid proteome mostly by housekeeping or regulatory functions through ER/Golgi-like proteins and transporters.

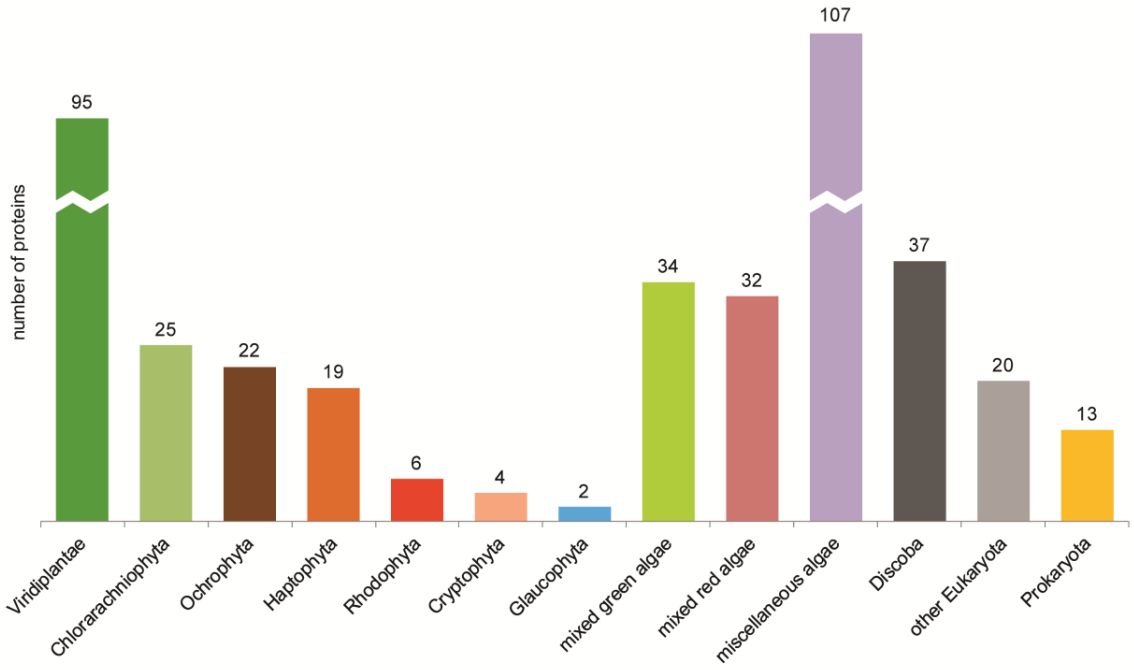

Fig. 4: Evolutionary affiliations of 416 nuclear-encoded plastid proteins presumably representing lateral gene transfer into or from *E. gracilis*. Protein trees are available at http://www.protistologie.cz/hampllab/data/Novak_Vanclova_trees.zip.

Several of the 13 prokaryote-derived proteins encode significant and unexpected functions. For example, four components (SufS, SufB, SufC and SufD) of the SUF system for iron-sulfur (FeS) cluster assembly were found. In transcriptomic data, we also recovered the fifth SUF pathway component, SufE, also of prokaryotic origin. Three of these proteins were specifically related to Chlamydiae in which the SufBCDS genes form a single operon that could have been laterally transferred as one unit from a single source. It is unknown whether these genes cluster together as they were too fragmented in the *E. gracilis* genome assembly (Ebenezer *et al.*, 2019). In general, the SUF pathway endosymbiotically derived from cyanobacteria is essential and universally present in plastids, supplying multiple core enzymes with FeS clusters (Lu, 2018). The chlamydiae-like set of proteins represents a second path for FeS cluster assembly alongside the ancestral plastidial cyanobacteria-related SUF system homologues (Fig. 5, Table S7 and Fig. S8). This apparent redundancy does not represent a transient state and/or result of neutral evolution because these two parallel SUF systems are also present in *E. longa* and *Eutreptiella* spp. (Table S7, Fig. S8). The relative protein abundance suggests that chlamydiae-like SUF proteins are generally more abundant than the

cyanobacteria-like cohort, which suggests that the chlamydiae-like pathway is active and contributes towards plastid FeS metabolism (Fig. 4). Co-occurrence of two SUF pathways in euglenophyte plastids may indicate functional diversification, for example with one in the stroma in common with other plastids (Lu, 2018), and the second restricted to one of the intermembrane spaces, serving co-localized FeS proteins. This hypothesis of spatial separation of the two SUF pathways is further supported by the fact that the chlamydiae-like proteins unlike the other cohort, possess N-terminal extensions which are generally shorter and do not conform to a typical plastid-targeting domain (supplementary alignments file sufs.txt). Similarly, a parallel CIA pathway for FeS cluster assembly was described in cryptophytes and proposed to function in the periplastidial compartment (PPC) which contains a nucleomorph and represents the remnant endosymbiont cytoplasm (Grosche *et al.*, 2018).

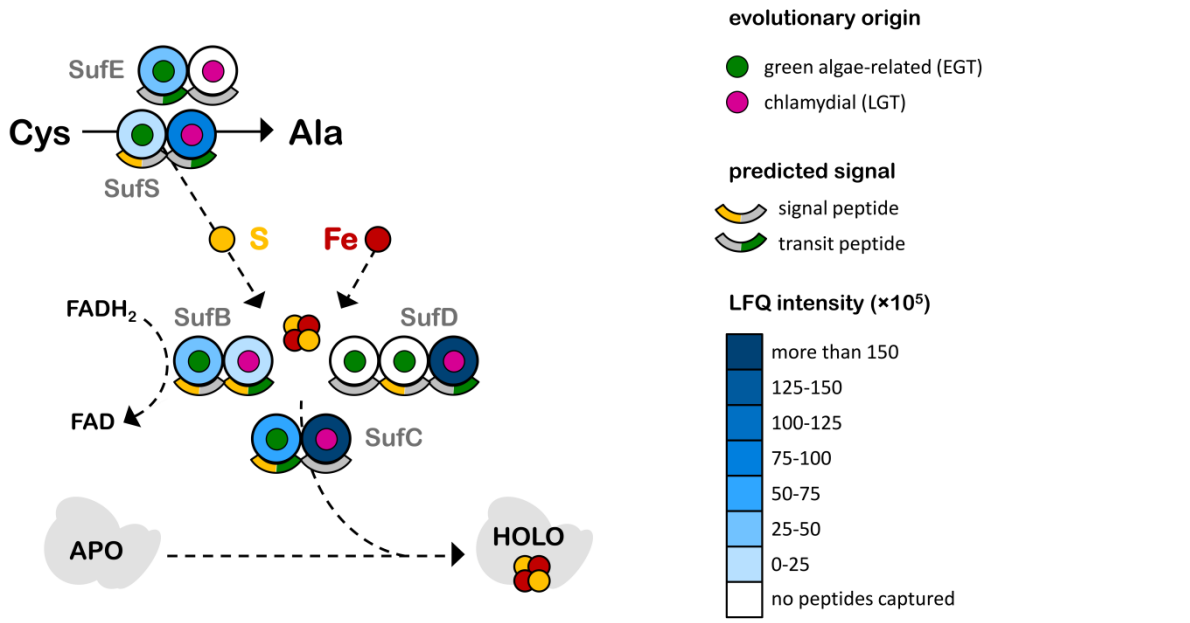

Fig. 5: Plastidal SUF pathway of *E. gracilis*. Multiple copies of each subunit are represented by circles colored in shades of blue based on their average LFQ intensity per replicate which indirectly corresponds to the relative protein abundance. Proteins captured in only one replicate are also included. The evolutionary origin of the particular subunit copy is represented by the dot in the middle: green for green algae-related, i.e. gained along with the plastid via endosymbiotic gene transfer (EGT), or magenta for chlamydial-like, i.e. gained via LGT. The presence of predicted signal and transit peptides is also indicated.

A further novel aspect of FeS cluster assembly machinery in *E. gracilis* is the presence of an ABC transporter (seqid 3116) in the plastid proteome, homologous to mitochondrial Atm1 protein required for the export of unspecified FeS cluster intermediates to the cytosol. The presence of this transporter suggests that such intermediates are transported across a

membrane, either into the plastid or between sub-compartments (supplementary-dataset-1.xlsx).

While direct experimental data are necessary to understand the function of the second FeS pathway, two additional chlamydial-like proteins were also found, both with orthologs in *E. longa* and *Eutreptiella*. One, an isoform of ferredoxin (divided into two contigs, seqids 37420 and 61063 in our data, seqid 74687 in Yoshida *et al.* (2016), supplementary-dataset-2.xlsx; Fig. S9), is itself a FeS protein, and may represent a client of the chlamydial-like SUF. The second is the alpha subunit of NADPH-dependent sulfite reductase (CysJ). Significantly, the beta subunit of this enzyme (CysI) contains a Fe_4S_4 cluster (Smith & Stroupe, 2012) and seems to be related to spirochaetes (Fig. S10). The plastid localization of sulfite reductase is itself notable, as previous biochemical studies associated sulfate assimilation, including the reaction catalyzed by sulfite reductase, with the mitochondrion in *E. gracilis* (Saidha *et al.*, 1988). The plastid localization of sulfite reduction is consistent with the presence of cysteine synthase in this organelle, as this enzyme assimilates hydrogen sulfide produced in this reaction.

N-terminal plastid-targeting sequences in E. gracilis*:* We took advantage of the confidently identified plastid-targeted proteins to re-evaluate characteristics of *E. gracilis* plastid-targeting sequences. N-terminal sequences of 375 preproteins with well-supported plastid localization ($\log_{10}$ CP/MT ratio > 1 or photosynthetic function) and non-truncated N-termini, verified by the presence of the spliced leader, were analysed. The presence of one or two hydrophobic domains was confirmed using the Kyte-Doolittle hydrophobicity score (Kyte & Doolittle, 1982) in positions 9-60 (putative signal peptide, SP) and 97-130 (putative stop-transfer signal, STS) (Fig. 6). The sequences were sorted into previously defined classes (Durnford & Gray, 2006): 47% class I (SP and STS), 37% class II (SP only), with 16% exhibiting neither of these hydrophobic domains, referred to as “unclassified”. The proportion of “unclassified” preproteins is relatively high and may suggest a significant cohort of plastid proteins with non-conventional targeting signals. No significant correlation between predicted function or preprotein classes was observed.

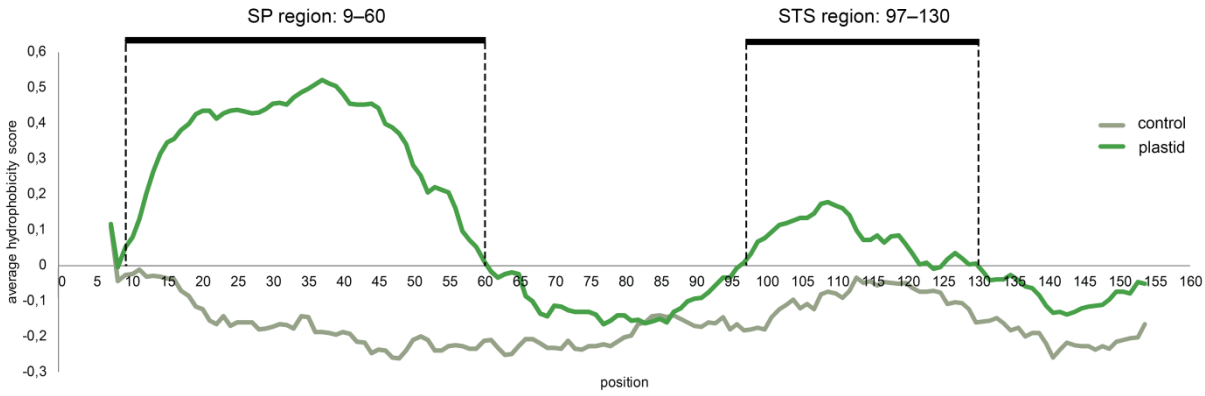

Fig. 6: The two hydrophobic domains in the N-termini of 375 well-supported plastid-targeted proteins with complete N-terminus as seen on the graph of average hydrophobicity score per position compared to the negative control (375 randomly selected proteins of *E. gracilis*). The range of the SP was estimated as positions 9-60 (note, that the first and last six positions of the 160 aa long sequence are missing as a result of window-based scoring; position 7 is the first position with score assigned), the range of the STS was estimated as positions 97-130. The hydrophobic peak representing the STS is clearly less accentuated than the one representing SP reflecting the fact that it is present in just a subset of plastid-targeted proteins termed class II pre-proteins.

The canonical euglenophyte plastid-targeting sequence includes a region directly downstream of the SP that exhibits features of a plastid transit peptide (transit peptide-like region, TPL) (Inagaki *et al.*, 2000; Slávikova *et al.*, 2005; Durnford & Gray, 2006), which is surprising, given the apparent absence of a conventional TOC complex in these organisms. To understand the TPL region we investigated its sequence characteristics more closely. The putative TPL region were selected from the 375 plastid proteins, either based on the actual positions of the hydrophobic domains or from TPL region (61-95) estimated from the overall hydrophobicity profile of all studied N-termini. Amino acid compositions of these putative TPLs differed significantly from their respective mature protein regions (Fig. 7). This contrasts to the control sets of 375 randomly selected proteins, where the only significant difference between TPL region and the rest of the protein was a relative enrichment in serine (for full results for both sets and all classes of preproteins, see Tables S12 and bar plots S13).

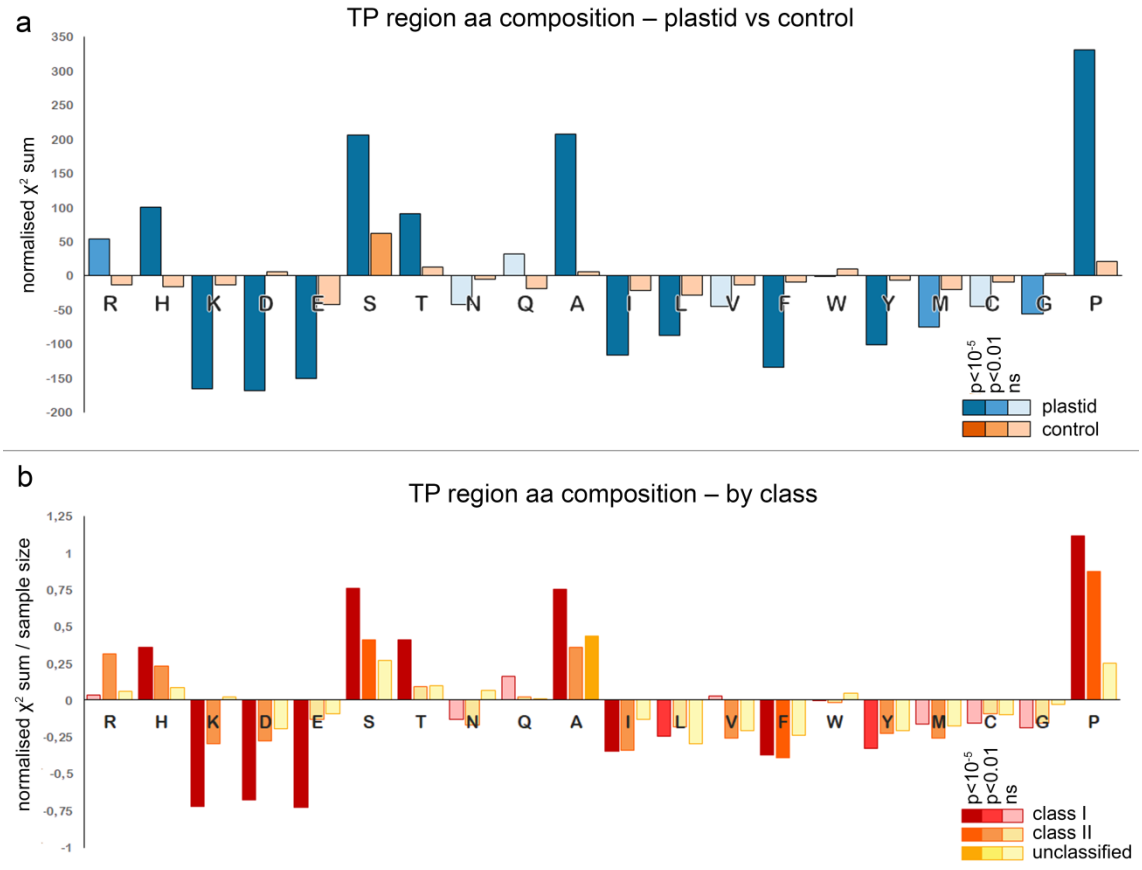

Fig. 7: Statistical comparison of the overall amino acid frequencies in the putative transit peptide region and the putative mature chain of the same protein for the set of 375 proteins regardless of their classification for the plastid protein sample in comparison to the negative control (shades of blue and orange, respectively, a). The same comparison was also performed separately for the plastid proteins of class I, class II, and "unclassified" (shades of red, orange, and yellow, respectively, b). The vertical axis represents normalized χ^2 sum reflecting whether the amino acid frequency is higher or lower than expected (positive or negative values) and the relative degree of its enrichment or depletion. The medium coloured bars represent statistically significant results with $p < 0.01$, the dark coloured bars represent those with $p < 10^{-5}$.

Discussion

The *E. gracilis* plastid proteome described here is similar in size to *Arabidopsis thaliana* (Huang *et al.*, 2013), but larger than *C. reinhardtii* (Terashima *et al.*, 2011) and *B. natans* (Hopkins *et al.*, 2012) suggesting that its functional capacity might be higher than other unicellular phototrophs, which could be connected to the metabolic versatility of euglenophytes. More proteins lacking readily identifiable homologs are present, compared to e.g. *C. reinhardtii,* and the proportion of proteins belonging to equivalent functional categories clearly differs, suggesting divergence in the relative significance of functions (Tab. 1). While the former is in part a result of experimental organism bias, this also indicates

575 the importance of analysis of organelles from divergent lineages to gain both an accurate
576 understanding of function in the organism itself as well as across eukaryotic systems.

577 Tab. 1: Comparison of the proportions of proteins of selected functional categories in the four plastid proteomes;
578 the number in brackets under the respective organism names is the total number of identified plastid proteins.

	E. gracilis (1345)	*A. thaliana* (1462)	*C. reinhardtii* (996)	*B. natans* (324)
Photosynthesis	6.4%	9.1%	11.9%	14.8%
Lipid metabolism	2.9%	4.7%	2.5%	2.2%
Amino acid metabolism	0.7%	5.6%	5.5%	2.2%
Secondary metabolism	2.4%	2.9%	2.7%	2.8%
Tetrapyrrole, cofactor and vitamin metabolism	3.5%	5.3%	3.2%	3.4%
Signalling	3.1%	2.3%	2.6%	3.7%

579

580 The euglenid plastid originated from pyramimonadalean green alga (Turmel *et al.*,
581 2009; Jackson *et al.*, 2018), but a proportion of plastid proteins show distinct phylogenetic
582 affinity. Being noted previously (Maruyama *et al.* 2011; Markunas & Triemer 2016; Lakey &
583 Triemer 2017; Ponce-Toledo *et al.* 2018), it was suggested that these genes were acquired
584 from "chromophyte" prey or symbiont by the common ancestor of eukaryovorous and/or
585 phototrophic euglenids. The present set of experimentally determined plastid proteins both
586 supports this hypothesis and allows semiquantitative evaluation of contributions from
587 phototrophic eukaryotes. Gene cohorts related to chlorarachniophytes, ochrophytes and
588 haptophytes account for over 60 *E. gracilis* plastid proteins (19% of the algal LGT
589 candidates). These genes could have originated (1) from the eukaryovorous ancestor of
590 euglenids, *sensu* the "you are what you eat" hypothesis (Doolittle, 1998), (2) from initial
591 stages of endosymbiont integration when the euglenid host was presumably obligatory
592 mixotrophic (much like *Rapaza viridis;* Yamaguchi *et al.*, 2012), such transfers could have
593 compensated for the reductive evolution of the endosymbiont genome, as proposed for the
594 chromatophore of *Paulinella* (Marin *et al.*, 2005; Nowack *et al.*, 2016), (3) gene transfers
595 from a cryptic endosymbiont, kleptoplastid, or even true plastid putatively present in the
596 euglenophyte ancestor and replaced by the extant organelle, in the "shopping bag" (Larkum *et*
597 *al.*, 2007) or "red carpet" (Ponce-Toledo *et al.*, 2019) sense and similar to the cases of serial
598 endosymbioses and overall plastid fluidity in dinoflagellates (Saldarriaga *et al.*, 2001; Yoon *et*
599 *al.*, 2005; Takano *et al.*, 2008; Matsumoto *et al.*, 2011; Xia *et al.*, 2013; Burki *et al.*, 2014;

Dorrell & Howe, 2015) or (4) horizontal transfers from euglenids to the other algal group. This latter is a possible model for proteins affiliated to chlorarachniophytes given that euglenids are estimated as > 200 MY younger than this group (Jackson *et al.* 2018).

The proportion of credible plastidial proteins affiliated to Discoba, likely representing ancestral euglenozoan proteins newly recruited for plastidial function, is relatively low. The set is noticeably biased towards proteins related to membrane and vesicular transport, which however, given the methodological constrains of this study, can be interpreted either as the result of the role of host endomembrane system in the transport of biomolecules, or sometimes more obviously as a mere contamination of both analyzed fractions by other subcellular compartments, namely ER, Golgi, and/or vesicles.

The presence of proteins of prokaryotic affiliation is noteworthy and unexpected. Many point at the Chlamydiae as the donor group, the most notable example is a complete second SUF pathway of FeS cluster assembly and hence potentially the result of a single genetic event. A surprisingly high number of chlamydial-like proteins are present in primary plastids of plants and algae (Moustafa *et al.*, 2008; Becker *et al.*, 2008), which led to the "ménage à trois" hypothesis proposing that a chlamydial endosymbiont was directly implicated in integration of a nascent plastid, and in some manner may have even been essential to the process (Facchinelli *et al.*, 2013; Cenci *et al.*, 2016). Whether this could also be the case for secondary plastid establishment remains an intriguing possibility.

While the *E. gracilis* plastid TPL region is not conserved at the sequence level, there is a characteristic amino acid composition (Figure 7). In addition to previously described positive net charge, enrichment of hydroxy residues and alanine common with TP of plant plastid proteins (Bruce, 2000; Durnford & Gray, 2006; Patron & Waller, 2007; Felsner *et al.*, 2010; Li & Teng, 2013) there is a paucity of non-polar residues and the high proline content that may confer distinctive secondary structure (MacArthur & Thornton, 1991). Interestingly, proline residues in TP were recently proposed to be important for correct import of plastid proteins with multiple transmembrane domains in plants (Lee *et al.*, 2018), but it is unclear what the significance of general proline-richness in TPs, as observed here, could be. The additional properties of the TPL are likely recognised by an as yet unknown receptor, consistent with a major reorganization to translocation systems, i.e. a reduced TIC/TOC complex and the presence of derlin-like proteins. The absence of unambiguously identifiable additional components, such as Npl4 and Ubx, leaves it unclear if a translocon analogous to

SELMA is functional in *E. gracilis*. Nevertheless, we believe it is a noteworthy candidate for components of the unknown euglenophyte plastid protein import system, specifically as translocases of the middle plastid membrane (Fig. 3), filling a gap in the protein import machinery caused by the absence of TOC. These findings also hint at the possibility of the middle membrane being derived from the host endomembrane system and not the outer cyanobacterial-like membrane of the ancestral primary plastid. Most secondary plastids keep all four membranes and it is not clear how probable or improbable the loss of one of the cyanobacterial membranes is, however, the recently described structure of the plastid of the stramenopile *Chrysoparadoxa* suggests that this membrane topology can indeed arise (Wetherbee *et al.*, 2018).

In summary, this is the first report of a proteome of a euglenid plastid, based on direct organellar isolation and proteomics. Novel import systems, metabolic pathways and the presence of significant transfers from prokaryotic genomes all contribute towards a complex and divergent organelle which has a major impact on the biosphere.

Acknowledgements

The authors thank Anna Nenarokova (Biology Centre, České Budějovice) for suggestions and consultations regarding data analysis and Dougie Lamont and colleagues at the Proteomics Facility of the University of Dundee for excellent work. This work was supported by Czech Grant Agency (15-21974S to V.H., 17-21409S to M.E.); ERC CZ (award LL1601 to J.L.); ERD Funds (project OPVVV CZ.02.1.01/0.0/0.0/16_019/0000759 to J.L. and V.H.); Yousef Jameel Academic Program (through the Yousef Jameel PhD Scholarship); the Cambridge Commonwealth; European and International Trust; the Cambridge University Student Registry; the Cambridge Philosophical Society (all to T.E.E.) and the Medical Research Council (Grant #: P009018/1 to M.C.F.). Computational resources were supplied by the Ministry of Education, Youth and Sports of the Czech Republic under the Projects CESNET (Project No. LM2015042) and CERIT-Scientific Cloud (Project No. LM2015085) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

Author contributions:

MCF, JL, and VH conceived the original research plans; JL and MCF supervised the experiments; ELD performed the cell fractionation; MZ and SK performed the mass spectrometry analysis; AMGNV and TGE performed protein annotation and sorting, AMGNV, KZ, ZF, and ME interpreted the annotation results, AMGNV performed the signal domain analysis, PS performed the phylogenetic analysis, AMGNV conceived the project and wrote the article with contributions of all the authors; VH, ME, MCF, and JL supervised and complemented the writing. VH agrees to serve as the author responsible for contact and ensuring communication.

Supplementary data:

- http://www.protistologie.cz/hampllab/data/Novak_Vanclova_supplement.zip (methods, tables, and figures)
- http://www.protistologie.cz/hampllab/data/Novak_Vanclova_trees.zip (phylogenetic trees)

References:

Altschul S, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. **1997**. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.

Archibald JM. **2015**. Genomic perspectives on the birth and spread of plastids. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 1421374112-.

Balsera M, Soll J, Buchanan BB. **2010**. Redox extends its regulatory reach to chloroplast protein import. *Trends in plant science* **15**: 515–21.

Barsanti L, Vismara R, Passarelli V, Gualtieri P. **2001**. Paramylon (β-1,3-glucan) content in wild type and WZSL mutant of Euglena gracilis. Effects of growth conditions. *Journal of Applied Phycology* **13**: 59–65.

Becker B, Hoef-Emden K, Melkonian M. **2008**. Chlamydial genes shed light on the evolution of photoautotrophic eukaryotes. *BMC Evolutionary Biology* **8**: 203.

Blee E, Schantz R. **1978**. Biosynthesis of galactolipids in Euglena gracilis: I, Incorporation of UDP galactose into galactosyldiglycerides. *Plant Science Letters* **13**: 247–255.

Bölter B, Soll J. **2016**. Once upon a Time - Chloroplast Protein Import Research from Infancy to Future Challenges. *Molecular Plant* **9**: 798–812.

Boucher MJ, Ghosh S, Zhang L, Lal A, Jang SW, Ju A, Zhang S, Wang X, Ralph SA, Zou J, *et al.* 2018. Integrative proteomics and bioinformatic prediction enable a high-confidence apicoplast proteome in malaria parasites (B Striepen, Ed.). *PLOS Biology* **16**: e2005895.

Bruce BD. **2000**. Chloroplast transit peptides: structure, function and evolution. *Trends in Cell Biology* **10**: 440–447.

Burki F, Imanian B, Hehenberger E, Hirakawa Y, Maruyama S, Keeling PJ. **2014**. Endosymbiotic gene transfer in tertiary plastid-containing dinoflagellates. *Eukaryotic Cell* **13**: 246–255.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. **2009**. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.

Cenci U, Ducatez M, Kadouche D, Colleoni C, Ball SG. **2016**. Was the Chlamydial Adaptative Strategy to Tryptophan Starvation an Early Determinant of Plastid Endosymbiosis? *Frontiers in cellular and infection microbiology* **6**: 67.

Cenci U, Qiu H, Pillonel T, Cardol P, Remacle C, Colleoni C, Kadouche D, Chabi M, Greub G, Bhattacharya D, *et al.* 2018. Host-pathogen biotic interactions shaped vitamin K metabolism in Archaeplastida. *Scientific Reports* **8**: 15243.

Cox J, Mann M. **2008**. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **26**: 1367–1372.

Davis B, Merrett MJ. **1973**. Malate Dehydrogenase Isoenzymes in Division Synchronized Cultures of Euglena. *Plant Physiology* **51**.

Dobáková E, Flegontov P, Skalický T, Lukeš J. **2015**. Unexpectedly Streamlined Mitochondrial Genome of the Euglenozoan Euglena gracilis. *Genome Biology and Evolution* **7**: 3358–3367.

Doetsch NA, Favreau MR, Kuscuoglu N, Thompson MD, Hallick RB. **2001**. Chloroplast transformation in Euglena gracilis: splicing of a group III twintron transcribed from a transgenic psbK operon. *Current genetics* **39**: 49–60.

Doolittle WF. **1998**. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in genetics : TIG* **14**: 307–11.

van Dooren GG, Striepen B. **2013**. The Algal Past and Parasite Present of the Apicoplast. *Annual Review of Microbiology* **67**: 271–289.

Vom Dorp K, Hölzl G, Plohmann C, Eisenhut M, Abraham M, Weber APM, Hanson AD, Dörmann P. **2015**. Remobilization of Phytol from Chlorophyll Degradation Is Essential for Tocopherol Synthesis and Growth of Arabidopsis. *The Plant cell* **27**: 2846–59.

Dorrell RG, Howe CJ. **2015**. Integration of plastids with their hosts: Lessons learned from dinoflagellates. *Proceedings of the National Academy of Sciences of the United States of America* **112**: 10247–54.

Durnford DG, Gray MW. **2006**. Analysis of Euglena gracilis plastid-targeted proteins reveals different classes of transit sequences. *Eukaryotic Cell* **5**: 2079–2091.

Ebenezer TE, Zoltner M, Burrell A, Nenarokova A, Novák Vanclová AMG, Prasad B, Soukal P, Santana-Molina C, O'Neill E, Nankissoor NN, *et al.* 2019. Transcriptome, proteome and draft genome of Euglena gracilis. *BMC Biology* **17**: 11.

Emanuelsson O, Nielsen H, von Heijne G. **1999**. ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Science : A Publication of the Protein Society* **8**: 978–984.

Emms DM, Kelly S. **2015**. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**: 157.

Emonds-Alt B, Coosemans N, Gerards T, Remacle C, Cardol P. **2017**. Isolation and characterization of mutants corresponding to the MENA, MENB, MENC and MENE enzymatic steps of 5'-monohydroxyphylloquinone biosynthesis in Chlamydomonas reinhardtii. *The Plant journal : for cell and molecular biology* **89**: 141–154.

Facchinelli F, Colleoni C, Ball SG, Weber APM. **2013**. Chlamydia, cyanobiont, or host: who was on top in the ménage à trois? *Trends in plant science* **18**: 673–9.

Felsner G, Sommer MS, Gruenheit N, Hempel F, Moog D, Zauner S, Martin W, Maier UG. **2011**. ERAD components in organisms with complex red plastids suggest recruitment of a preexisting protein transport pathway for the periplastid membrane. *Genome biology and evolution* **3**: 140–50.

Felsner G, Sommer MS, Maier UG. **2010**. The physical and functional borders of transit peptide-like sequences in secondary endosymbionts. *BMC plant biology* **10**: 223.

Geimer S, Belicová A, Legen J, Sláviková S, Herrmann RG, Krajcovic J. **2009**. Transcriptome analysis of the Euglena gracilis plastid chromosome. *Current genetics* **55**: 425–38.

Gomez-Silva B, Timko MP, Schiff JA. **1985**. Chlorophyll biosynthesis from glutamate or 5-aminolevulinate in intact Euglena chloroplasts. *Planta* **165**: 12–22.

Gould SB, Maier U-G, Martin WF. **2015**. Protein Import and the Origin of Red Complex Plastids. *Current biology : CB* **25**: R515–R521.

Grosche C, Diehl A, Rensing SA, Maier UG. **2018**. Iron–Sulfur Cluster Biosynthesis in Algae with Complex Plastids (M Embley, Ed.). *Genome Biology and Evolution* **10**: 2061–2071.

Gumińska N, Płecha M, Zakryś B, Milanowski R. **2018**. Order of removal of conventional and nonconventional introns from nuclear transcripts of Euglena gracilis (MC Field, Ed.). *PLOS Genetics* **14**: e1007761.

Hall RP, Schoenborn HW. **1939**. *The Question of Autotrophic Nutrition in Euglena gracilis*.

Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, Orsat B, Spielmann A, Stutz E. **1993**. Complete sequence of Euglena gracilis chloroplast DNA. *Nucleic acids research* **21**: 3537–3544.

Hempel F, Bullmann L, Lau J, Zauner S, Maier UG. **2009**. ERAD-derived preprotein transport across the second outermost plastid membrane of diatoms. *Molecular biology and evolution* **26**: 1781–90.

Heyes DJ, Neil Hunter C. **2009**. Biosynthesis of Chlorophyll and Bacteriochlorophyll. In: Tetrapyrroles. New York, NY: Springer New York, 235–249.

Hiller K, Grote A, Scheer M, Münch R, Jahn D. **2004**. PrediSi: Prediction of signal peptides and their cleavage positions. *Nucleic Acids Research* **32**: 375–379.

Hopkins JF, Spencer DF, Laboissiere S, Neilson JAD, Eveleigh RJM, Durnford DG, Gray MW, Archibald JM. **2012**. Proteomics Reveals Plastid- and Periplastid-Targeted Proteins in the Chlorarachniophyte Alga Bigelowiella natans. *Genome Biology and Evolution* **4**: 1391–1406.

Howe CJ, Schlarb-Ridley BG, Wastl J, Purton S, Bendall DS. **2006**. The novel cytochrome c6 of chloroplasts: a case of evolutionary bricolage? *Journal of Experimental Botany* **57**: 13–22.

Huang M, Friso G, Nishimura K, Qu X, Olinares PDB, Majeran W, Sun Q, van Wijk KJ. **2013**. Construction of Plastid Reference Proteomes for Maize and *Arabidopsis* and Evaluation of Their Orthologous Relationships; The Concept of Orthoproteomics. *Journal of Proteome Research* **12**: 491–504.

Inagaki J, Fujita Y, Hase T, Yamamoto Y. **2000**. Protein translocation within chloroplast is similar in Euglena and higher plants. *Biochemical and biophysical research communications* **277**: 436–442.

Inwongwan S, Kruger NJ, Ratcliffe RG, O'Neill EC. **2019**. Euglena Central Metabolic Pathways and Their Subcellular Locations.

Jackson C, Knoll AH, Chan CX, Verbruggen H. **2018**. Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green plastids. *Scientific Reports* **8**: 1523.

Jenkins KP, Hong L, Hallick RB. **1995**. Alternative splicing of the Euglena gracilis chloroplast roaA transcript. *RNA (New York, N.Y.)* **1**: 624–633.

Katoh K, Standley DM. **2013**. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **30**: 772–780.

Kikuchi S, Oishi M, Hirabayashi Y, Lee DW, Hwang I, Nakai M. **2009**. A 1-megadalton translocation complex containing Tic20 and Tic21 mediates chloroplast protein import at the inner envelope membrane. *The Plant cell* **21**: 1781–97.

Kim D, Filtz MR, Proteau PJ. **2004**. The methylerythritol phosphate pathway contributes to carotenoid but not phytol biosynthesis in Euglena gracilis. *Journal of Natural Products* **67**: 1067–1069.

Kořený L, Oborník M. **2011**. Sequence evidence for the presence of two tetrapyrrole pathways in Euglena gracilis. *Genome Biology and Evolution* **3**: 359–364.

Koziol AG, Durnford DG. **2008**. Euglena Light-Harvesting Complexes Are Encoded by Multifarious Polyprotein mRNAs that Evolve in Concert. *Molecular Biology and Evolution* **25**: 92–100.

Krinsky NI, Goldsmith TH. **1960**. The carotenoids of the flagellated alga, Euglena gracilis. *Archives of biochemistry and biophysics* **91**: 271–279.

Kuo RC, Zhang H, Zhuang Y, Hannick L, Lin S. **2013**. Transcriptomic Study Reveals Widespread Spliced Leader Trans-Splicing, Short 5'-UTRs and Potential Complex Carbon

812 Fixation Mechanisms in the Euglenoid Alga Eutreptiella sp. *PLoS ONE* **8**: e60826.

813 **Kyte J, Doolittle RF**. **1982**. A simple method for displaying the hydropathic character of a
814 protein. *Journal of Molecular Biology* **157**: 105–132.

815 **Lakey B, Triemer R**. **2017**. The tetrapyrrole synthesis pathway as a model of horizontal gene
816 transfer in euglenoids (K Müller, Ed.). *Journal of Phycology* **53**: 198–217.

817 **Langemeyer L, Fröhlich F, Ungermann C**. **2018**. Rab GTPase Function in Endosome and
818 Lysosome Biogenesis. *Trends in cell biology* **28**: 957–970.

819 **Larkum AWD, Lockhart PJ, Howe CJ**. **2007**. Shopping for plastids. *Trends in Plant*
820 *Science* **12**: 189–195.

821 **Lau JB, Stork S, Moog D, Schulz J, Maier UG**. **2016**. Protein-protein interactions indicate
822 composition of a 480 kDa SELMA complex in the second outermost membrane of diatom
823 complex plastids. *Molecular Microbiology* **100**: 76–89.

824 **Leander BS**. **2004**. Did trypanosomatid parasites have photosynthetic ancestors? *Trends in*
825 *Microbiology* **12**: 251–258.

826 **Leander BS, Lax G, Karnkowska A, Simpson AGB**. **2017**. Euglenida. In: Handbook of the
827 Protists. Cham: Springer International Publishing, 1–42.

828 **Leander BS, Triemer RE, Farmer M a**. **2001**. Character evolution in heterotrophic
829 euglenids. *European Journal of Protistology* **37**: 337–356.

830 **Lee DW, Yoo Y-J, Razzak MA, Hwang I**. **2018**. Prolines in Transit Peptides Are Crucial for
831 Efficient Preprotein Translocation into Chloroplasts. *Plant physiology* **176**: 663–677.

832 **Li H min, Teng YS**. **2013**. Transit peptide design and plastid import regulation. *Trends in*
833 *Plant Science* **18**: 360–366.

834 **Lin Y-P, Wu M-C, Charng Y-Y**. **2016**. Identification of a Chlorophyll Dephytylase
835 Involved in Chlorophyll Turnover in Arabidopsis. *The Plant cell* **28**: 2974–2990.

836 **Lohr M, Schwender J, Polle JEW**. **2012**. Isoprenoid biosynthesis in eukaryotic phototrophs:
837 A spotlight on algae. *Plant Science* **185–186**: 9–22.

838 **Lu Y**. **2018**. Assembly and Transfer of Iron–Sulfur Clusters in the Plastid. *Frontiers in Plant*
839 *Science* **9**: 336.

840 **MacArthur MW, Thornton JM**. **1991**. Influence of proline residues on protein
841 conformation. *Journal of Molecular Biology* **218**: 397–412.

842 **Maier UG, Zauner S, Hempel F**. **2015**. Protein import into complex plastids: Cellular
843 organization of higher complexity. *European journal of cell biology*.

844 **Marin B, Nowack ECM, Melkonian M**. **2005**. A plastid in the making: evidence for a
845 second primary endosymbiosis. *Protist* **156**: 425–32.

846 **Markunas CM, Triemer RE**. **2016**. Evolutionary History of the Enzymes Involved in the
847 Calvin-Benson Cycle in Euglenids. *Journal of Eukaryotic Microbiology* **63**: 326–339.

848 **Maruyama S, Suzaki T, Weber APM, Archibald JM, Nozaki H**. **2011**. Eukaryote-to-

eukaryote gene transfer gives rise to genome mosaicism in euglenids. *BMC evolutionary biology* **11**: 105.

Mateášiková-Kováčová B, Vesteg M, Drahovská H, Záhonová K, Vacula R, Krajčovič J. **2012**. Nucleus-encoded mRNAs for chloroplast proteins GapA, PetA, and PsbO are trans-spliced in the flagellate Euglena gracilis irrespective of light and plastid function. *Journal of Eukaryotic Microbiology* **59**: 651–653.

Matson RS, Meifei, Chang SB. **1970**. Comparative studies of biosynthesis of galactolipids in Euglena-gracilis strain-Z . *Plant Physiology* **45**: 531-.

Matsumoto T, Shinozaki F, Chikuni T, Yabuki A, Takishita K, Kawachi M, Nakayama T, Inouye I, Hashimoto T, Inagaki Y. **2011**. Green-colored Plastids in the Dinoflagellate Genus Lepidodinium are of Core Chlorophyte Origin. *Protist* **162**: 268–276.

Minge M a, Shalchian-Tabrizi K, Tørresen OK, Takishita K, Probert I, Inagaki Y, Klaveness D, Jakobsen KS. **2010**. A phylogenetic mosaic plastid proteome and unusual plastid-targeting signals in the green-colored dinoflagellate Lepidodinium chlorophorum. *BMC evolutionary biology* **10**: 191.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. **2007**. KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic acids research* **35**: W182-5.

Moustafa A, Reyes-Prieto A, Bhattacharya D. **2008**. Chlamydiae Has Contributed at Least 55 Genes to Plantae with Predominantly Plastid Functions (R DeSalle, Ed.). *PLoS ONE* **3**: e2205.

Muchhal US, Schwartzbach SD. **1994**. Characterization of the unique intron-exon junctions of Euglena gene(s) encoding the polyprotein precursor to the light-harvesting chlorophyll a/b binding protein of photosystem II. *Nucleic acids research* **22**: 5737–44.

Nakai M. **2018**. New Perspectives on Chloroplast Protein Import. *Plant and Cell Physiology* **59**: 1111–1119.

Nawrocki WJ, Tourasse NJ, Taly A, Rappaport F, Wollman F-A. **2015**. The Plastid Terminal Oxidase: Its Elusive Function Points to Multiple Contributions to Plastid Physiology. *Annual Review of Plant Biology* **66**: 49–74.

Nowack ECM, Price DC, Bhattacharya D, Singer A, Melkonian M, Grossman AR. **2016**. Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of Paulinella chromatophora. *Proceedings of the National Academy of Sciences of the United States of America* **113**: 12214–12219.

O'Neill EC, Trick M, Hill L, Rejzek M, Dusi RG, Hamilton CJ, Zimba P V., Henrissat B, Field RA. **2015**. The transcriptome of Euglena gracilis reveals unexpected metabolic capabilities for carbohydrate and natural product biochemistry. *Molecular BioSystems* **11**: 2808–2820.

Patron NJ, Waller RF. **2007**. Transit peptide diversity and divergence: A global analysis of plastid targeting signals. *BioEssays : news and reviews in molecular, cellular and developmental biology* **29**: 1048–58.

Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, Inuganti A, Griss J, Mayer G, Eisenacher M, *et al.* 2019. The PRIDE database and related

890 tools and resources in 2019: improving support for quantification data. *Nucleic Acids*
891 *Research* **47**: D442–D450.

892 **Petersen TN, Brunak S, von Heijne G, Nielsen H**. **2011**. SignalP 4.0: discriminating signal
893 peptides from transmembrane regions. *Nature methods* **8**: 785–6.

894 **Ponce-Toledo RI, Moreira D, López-García P, Deschamps P, Ruiz-Trillo I**. **2018**.
895 Secondary Plastids of Euglenids and Chlorarachniophytes Function with a Mix of Genes of
896 Red and Green Algal Ancestry (I Ruiz-Trillo, Ed.). *Molecular Biology and Evolution*.

897 **Ponce-Toledo RI, López-García P, Moreira D**. **2019**. Horizontal and endosymbiotic gene
898 transfer in early plastid evolution. *New Phytologist*: nph.15965.

899 **R-Core-Team**. **2013**. R: A language and environment for statistical computing.

900 **Reinbothe C, Ortel B, Parthier B, Reinbothe S**. **1994**. Cytosolic and plastid forms of 5-
901 enolpyruvylshikimate-3-phosphate synthase in Euglena gracilis are differentially expressed
902 during light-induced chloroplast development. *Molecular & general genetics : MGG* **245**:
903 616–22.

904 **Saidha T, Na SQ, Li JY, Schiff JA**. **1988**. A sulphate metabolizing centre in Euglena
905 mitochondria. *The Biochemical journal* **253**: 533–9.

906 **Saidha T, Schiff JA**. **1989**. The role of mitochondria in sulfolipid biosynthesis by Euglena
907 chloroplasts. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism* **1001**: 268–
908 273.

909 **Saldarriaga JF, Taylor FJR, Keeling PJ, Cavalier-Smith T**. **2001**. Dinoflagellate Nuclear
910 SSU rRNA Phylogeny Suggests Multiple Plastid Losses and Replacements. *Journal of*
911 *Molecular Evolution* **53**: 204–213.

912 **Seeger JW, Bentley R**. **1991**. Phylloquinone (Vitamin K1) biosynthesis inEuglena gracilis
913 strain Z. *Phytochemistry* **30**: 3585–3589.

914 **Sheiner L, Striepen B**. **2013**. Protein sorting in complex plastids. *Biochimica et Biophysica*
915 *Acta - Molecular Cell Research* **1833**: 352–359.

916 **Shibata S, Arimura S, Ishikawa T, Awai K**. **2018**. Alterations of Membrane Lipid Content
917 Correlated With Chloroplast and Mitochondria Development in Euglena gracilis. *Frontiers in*
918 *Plant Science* **9**: 370.

919 **Sláviková S, Vacula R, Fang Z, Ehara T, Osafune T, Schwartzbach SD**. **2005**.
920 Homologous and heterologous reconstitution of Golgi to chloroplast transport and protein
921 import into the complex chloroplasts of Euglena. *Journal of cell science* **118**: 1651–1661.

922 **Smith KW, Stroupe ME**. **2012**. Mutational Analysis of Sulfite Reductase Hemoprotein
923 Reveals the Mechanism for Coordinated Electron and Proton Transfer. *Biochemistry* **51**:
924 9857–9868.

925 **Soding J, Biegert A, Lupas AN**. **2005**. The HHpred interactive server for protein homology
926 detection and structure prediction. *Nucleic Acids Research* **33**: W244–W248.

927 **Sommer MS, Gould SB, Lehmann P, Gruber A, Przyborski JM, Maier U-G**. **2007**. Der1-
928 mediated Preprotein Import into the Periplastid Compartment of Chromalveolates? *Molecular*

Biology and Evolution **24**: 918–928.

Spork S, Hiss J a, Mandel K, Sommer M, Kooij TW a, Chu T, Schneider G, Maier UG, Przyborski JM. **2009**. An unusual ERAD-like complex is targeted to the apicoplast of Plasmodium falciparum. *Eukaryotic cell* **8**: 1134–45.

Stamatakis A. **2006**. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688–2690.

Stiller JW, Schreiber J, Yue J, Guo H, Ding Q, Huang J. **2014**. The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nature Communications* **5**: 5764.

Stork S, Moog D, Przyborski JM, Wilhelmi I, Zauner S, Maier UG. **2012**. Distribution of the SELMA Translocon in Secondary Plastids of Red Algal Origin and Predicted Uncoupling of Ubiquitin-Dependent Translocation from Degradation. *Eukaryotic cell* **11**: 1472–1481.

Sulli C, Fang ZW, Muchhal U, Schwartzbach SD. **1999**. Topology of Euglena chloroplast protein precursors within endoplasmic reticulum to Golgi to chloroplast transport vesicles. *Journal of Biological Chemistry* **274**: 457–463.

Takano Y, Hansen G, Fujita D, Horiguchi T. **2008**. Serial Replacement of Diatom Endosymbionts in Two Freshwater Dinoflagellates, Peridiniopsis spp. (Peridiniales, Dinophyceae). *Phycologia* **47**: 41–53.

Tanaka Y, Ogawa T, Maruta T, Yoshida Y, Arakawa K, Ishikawa T. **2017**. Glucan synthase-like 2 is indispensable for paramylon synthesis in Euglena gracilis. *FEBS Letters* **591**: 1360–1370.

Teng Y-S, Su Y, Chen L-J, Lee YJ, Hwang I, Li H. **2006**. Tic21 is an essential translocon component for protein translocation across the chloroplast inner envelope membrane. *The Plant cell* **18**: 2247–57.

Terashima M, Specht M, Hippler M. **2011**. The chloroplast proteome: a survey from the Chlamydomonas reinhardtii perspective with a focus on distinctive features. *Current genetics* **57**: 151–68.

Tessier LH, Keller M, Chan RL, Fournier R, Weil JH, Imbault P. **1991**. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in Euglena. *The EMBO journal* **10**: 2621–5.

Tonkin CJ, Struck NS, Mullin K a, Stimmler LM, McFadden GI. **2006**. Evidence for Golgi-independent transport from the early secretory pathway to the plastid in malaria parasites. *Molecular microbiology* **61**: 614–30.

Turmel M, Gagnon M-C, O'Kelly CJ, Otis C, Lemieux C. **2009**. 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. *Molecular biology and evolution* **26**: 631–48.

Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J. **2016**. The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods* **13**: 731–740.

Watanabe F, Yoshimura K, Shigeoka S. **2017**. Biochemistry and Physiology of Vitamins in Euglena. In: Schwartzbach SD, Shigeoka S, eds. Euglena: Biochemistry, Cell and Molecular Biology. Cham, Switzerland: Springer International Publishing, 65–90.

Wetherbee R, Jackson CJ, Repetti SI, Clementson LA, Costa JF, van de Meene A, Crawford S, Verbruggen H. **2018**. The golden paradox – a new heterokont lineage with chloroplasts surrounded by two membranes. *Journal of Phycology*: jpy.12822.

Wildner GF, Hauska G. **1974**. Localization of the reaction site of cytochrome 552 in chloroplasts from Euglena gracilis: Cytochrome content and photooxidation in different chloroplast preparations. *Archives of Biochemistry and Biophysics* **164**: 127–135.

Xia S, Zhang Q, Zhu H, Cheng Y, Liu G, Hu Z. **2013**. Systematics of a Kleptoplastidal Dinoflagellate, Gymnodinium eucyaneum Hu (Dinophyceae), and Its Cryptomonad Endosymbiont (I Söderhäll, Ed.). *PLoS ONE* **8**: e53820.

Yamaguchi A, Yubuki N, Leander BS. **2012**. Morphostasis in a novel eukaryote illuminates the evolutionary transition from phagotrophy to phototrophy: description of Rapaza viridis n. gen. et sp. (Euglenozoa, Euglenida). *BMC evolutionary biology* **12**: 29.

Yoon HS, Hackett JD, Van Dolah FM, Nosenko T, Lidie KL, Bhattacharya D. **2005**. Tertiary endosymbiosis driven genome evolution in dinoflagellate algae. *Molecular biology and evolution* **22**: 1299–308.

Yoshida Y, Tomiyama T, Maruta T, Tomita M, Ishikawa T, Arakawa K. **2016**. De novo assembly and comparative transcriptome analysis of Euglena gracilis in response to anaerobic conditions. *BMC genomics* **17**: 182.

Záhonová K, Füssy Z, Birčák E, Novák Vanclová AMG, Klimeš V, Vesteg M, Krajčovič J, Oborník M, Eliáš M. **2018**. Peculiar features of the plastids of the colourless alga Euglena longa and photosynthetic euglenophytes unveiled by transcriptome analyses. *Scientific Reports* **8**: 17012.

Zhao L, Chang W, Xiao Y, Liu H, Liu P. **2013**. Methylerythritol Phosphate Pathway of Isoprenoid Biosynthesis. *Annual Review of Biochemistry* **82**: 497–530.

Ziegler K, Maldener I, Lockau W. **1989**. 5'-Monohydroxyphylloquinone as a Component of Photosystem I. *Zeitschrift für Naturforschung C* **44**: 468–472.