(...), weil der Kreis die reichste, einfachste,
unerschöpflichste, leichtfaßlichste Figur ist; (...)

Jean Paul

# Chloroplast Genome Diversity in the Phototrophic Euglenoids, with Emphasis on Genome Structure, Synteny and Intron Evolution

**Dissertation**

zur Erlangung des Doktorgrades (Dr. rer. nat)

der Fakultät für Mathematik und Naturwissenschaften

der Bergischen Universität Wuppertal

angefertigt am

Lehrstuhl für Zoologie und Biologiedidaktik

vorgelegt von

Nadja Alice Faride Dabbagh

Wuppertal, 2017

BERGISCHE
UNIVERSITÄT
WUPPERTAL

Die Dissertation kann wie folgt zitiert werden:

urn:nbn:de:hbz:468-20171201-113132-6
[http://nbn-resolving.de/urn/resolver.pl?urn=urn%3Anbn%3Ade%3Ahbz%3A468-20171201-113132-6]

## Contents

# List of Tables

## 2 Material and Methods

## 3 Results

### Chapter I

### Chapter II

**Chapter III**

**Appendix**

# List of Figures

## 1 Introduction

## 3 Results

### Chapter I

### Chapter II

**Chapter III**

**4 Discussion**

# Abstract

To shed light on chloroplast genome evolution in the phototrophic euglenoids the cpGenomes of *Euglena mutabilis* (SAG 1224-9b), *Trachelomonas grandis* (SAG 204.80) and *Eutreptiella pomquetensis* (CCMP 1491) were isolated, sequenced and annotated. The chloroplast genomes were investigated intensively and compared to other cpGenomes of phototrophic euglenoids, with special focus on genome size and structure, number and localization of rRNA operons as well as introns. As a cause for genome size differences three major reasons have been identified. First, the intergenic space between the cpGenomes of different taxa varied greatly, even between closely related species. Second, the rRNA operon numbers between different taxa were not uniform. Third, the different intron numbers and intron types between different taxa led to the main reason for size differences in euglenoids cpGenomes.

Comprehensive trends of intron number and intron type have been detected in closely as well as distantly related euglenoids. These trends can be used to explain intron density and quantity as well as high or low similarities in the evolution of introns in all phototrophic euglenoids. The expansion and evolution of *psb*C introns can partly be elucidated by assumed horizontal intron transfers in the chloroplast of euglenoids after the split from Eutreptiales and Euglenales.

Findings concerning the emergence and evolution of group III introns supported the hypothesis that group III introns are degenerated group II introns. Surprisingly, the cpGenomes of the basally branching Eutreptiales are free of group III introns, although the results indicated that their evolution began in Eutreptiales as intermediate stages of group II and III introns (mini group II introns).

Furthermore, a new phylogenomic analysis of phototrophic euglenoids was performed and compared to recently published phylogenetic analyses. As a new approach genome-level characters from all known cpGenomes of euglenoids have been used as a tool to complement the phylogenomic analysis. Metacharacter analyses yielded gene arrangement, cluster arrangement and rRNA operons as viable metacharacters with partly important modifications between the taxa. Significant cluster rearrangement was identified in several clades that matched the phylogenetic reconstruction. Using the rRNA operon as a metacharacter revealed a trend of loss of one rRNA copy following the diversification of Euglenales. Basally branching Eutreptiales contained two copies, which is identical to the structure in the surmised chloroplast donor *Pyramimonas parkeae*. Only for both *Euglena gracilis* species and *Strombomonas acuminata* an independent acquirement of further rRNA operons was recognizable.

The cpGenome of *Eutreptiella pomquetensis* showed the same quadripartite cpGenome structure as *Pyramimonas parkeae*, corroborating the close relationship between these two taxa. The present work provides a sound basis for further examinations of chloroplast genome analyses to get a more thorough understanding of intron evolution within the phototrophic euglenoids. Likewise it represents a precursor for future studies concerning genome-level features in phototrophic euglenoids.

# 1 Introduction

As a rule, biological classification (taxonomy) aims at simplifying and ordering the immense diversity of life into taxa whose members share important properties and offers widely accepted names. Euglenozoa (CAVALIER-SMITH 1981) emend. SIMPSON 1997 are among the few exceptions, which hold an ambivalent status in phylogenetic systematics, caused by this are the diverse nutritional modes of euglenozoans and especially of Euglenida (BÜTSCHLI 1884) emend. SIMPSON 1997, which embrace heterotrophic as well as phototrophic organisms. Phototrophic euglenoids, which are of main interest in this work, are considered as algae and are listed in the International Code of Nomenclature for algae, fungi and plants (ICN 2012). Conversely, most heterotrophic forms of euglenoids are recognized as protozoa and listed in the International Zoological Code of Nomenclature (ICZN 1999, Kadereit et al. 2014, Wehner & Gehring 2013).

## 1.1 Phylogenetic position of euglenoids within Euglenozoa

Euglenozoa (CAVALIER-SMITH 1981) emend. SIMPSON 1997 form a well supported monophylum within the supergroup Excavata (CAVALIER-SMITH 2002) emend. Simpson 2003, which can cytologically be defined by a suspension-feeding groove (Patterson 1999, Simpson & Patterson 1999 & 2001) and comprise a huge range of free-living and parasitic single-celled flagellates. Hence, their classification is very controversial and incomplete, as there are phagotrophic, osmotrophic, symbiotrophic and parasitic life forms mixed with phototrophic forms.

Originally the taxon Euglenozoa was established as a single phylum by Cavalier-Smith (1981) to group the euglenoids and kinetoplastids together. Later on, Kivic & Walne (1984) postulated that Kinetoplastida are close relatives to Euglenida because of extensive morphological homologies like the paraxonemal rod, the flagellar apparatus and a single large mitochondrion. Subsequently, the Euglenida (BÜTSCHLI 1884) emend. SIMPSON 1997, Kinetoplastida HONIGBERG 1963, Diplonemida (CAVALIER-SMITH 1993) emend. SIMPSON 1997 and the single-species taxon *Postgaardi* with *Postgaardi mariagerensis* FENCHEL et al. 1995 were grouped together as Euglenozoa by Simpson (1997). This was later on supported by phylogenetic molecular and ultrastructural analyses (Adl et al. 2005, Busse & Preisfeld 2002a, Maslov et al. 1999, Simpson & Roger 2004). The identification of *Calkinsea aureus* and *Bihospites bacati,* euglenoid-like cells hosting symbiotic bacteria led to the inclusion of Symbiontida to Euglenozoa (Breglia et al. 2010, Yubuki et al. 2009). Initially, Symbiontida were treated as a monophylum within Euglenozoa, but other studies suggested that they are probably derived phagotrophic euglenoids (Adl et al. 2012, Walker et al. 2011). This was confirmed by Yubuki et al. (2013) based on morphological data. They classified Symbiontida as a clade nested within the Euglenida. Nevertheless, the inner sister-group relationship still remained unclear and showed incongruity in regard to relatedness of the three major clades Euglenida, Diplonemida and Kinetoplastida. Some investigations suggest that Kinetoplastida and Diplonemida are sister groups, whereas others associate Diplonemida and Euglenida (Breglia et al. 2007, Busse & Preisfeld 2002a, Triemer & Farmer 1991, von der Heyden et al. 2004, Yamaguchi et al. 2012). Another study by Cavalier-Smith (2016) enlarged the taxon sampling and analyzed the taxa ultrastructurally and molecularly and divided Euglenozoa into the three subphyla Euglenoida, Postgaardia and Glycomonada, the latter with Kinetoplastea and Diplonemea as sister classes. A recent study by Paerschke et al. (2017) used SSU rDNA data, secondary structure elements and investigated the absence or presence of the storage carbohydrate paramylon to infer the phylogeny of Eugelnozoans. They could identify robust clades for Diplonemida, Kinetoplastida and Euglenida with Petalomonadida and Symbiontida as a basal sister clade (Fig. 1.1).

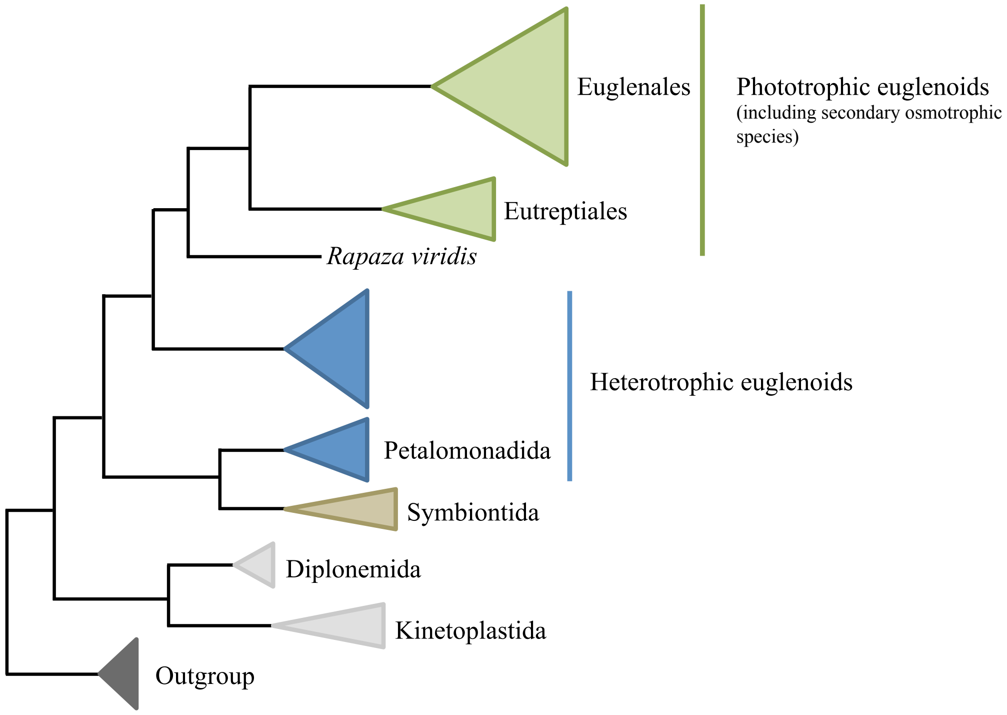


**Fig. 1.1:** Schematic phylogeny of Euglenozoa demonstrating positions of diplonemids, kinetoplastids and euglenoids. Phototrophic euglenoids formed a monophyletic clade with the mixotrophic species *Rapaza viridis* as sister lineage and clearly derived from heterotrophic euglenoids (modified from Paerschke et al. (2017) and Yamaguchi et al. (2012)).

A closer look at the internal phylogenies of the predominantly free-living euglenoids obtained by morphological and molecular data revealed three major groups: phototrophics, heterotrophics (including phagotrophics and osmotrophics) and mixotrophics (Bicudo & Menezes 2016, Yamaguchi et al. 2012). Phototrophic euglenoids formed a monophyletic clade with the mixotrophic species *Rapaza viridis* as sister lineage and clearly derived from heterotrophic euglenoids (Fig. 1.1) (Bicudo & Menezes 2016).

Phylogenetic analyses of phototrophic euglenoids made up the two orders Eutreptiales, with predominantly marine representatives and Euglenales consisting of freshwater flagellates (Adl et al. 2012). The Euglenales are very well sampled and can be separated in two families: The more basal Phacaceae with the three genera *Lepocinclis*, *Discoplastis* and *Phacus,* and the Euglenaceae with the polyphyletic genus *Euglena* and the monophyletic genera *Euglenaformis*, *Euglenaria*, *Colacium*, *Strombomonas*, *Trachelomonas*, *Monomorphina* and *Cryptoglena* (Bicudo & Menezes 2016, Kim et al. 2010, Schwartzbach & Shigeoka 2017).

## 1.2 Euglenoids' nutrition

Euglenoids as a large group of predominantly free-living aquatic microbes, colonize marine and freshwater sediments, but are also present in brackish waters and soil, thus playing an important role in these ecosystems (Ekelund & Patterson 1997, Leedale 1967). Since the first euglenoid was detected in 1674 by van Leeuwenhoek extensive research led to almost 3,000 described species (Bicudo & Menezes 2016), with diverse modes of nutrition, including phototrophic, heterotrophic and mixotrophic species (Bicudo & Menezes 2016, Triemer & Farmer 2007, Yamaguchi et al. 2012). Colorless forms of euglenoids are heterotrophic and consist of phagotrophic euglenoids with simple feeding apparatuses to ingest small prey bacteria, sometimes referred to as 'bacteriotrophs' or with highly elaborated feeding apparatuses to feed on small eukaryotes, sometimes referred to as 'eukaryotrophs' (Leander et al. 2001, Leander et al. 2007). Osmotrophic euglenoids stem like phototrophic euglenoids themselves from phagotrophs, which have lost their ingestion apparatus and absorb dissolved nutrients directly from environments. Thus, they are called primary osmotrophs. A minority of osmotrophs mostly from the lineage Euglenaceae derived from phototrophs by loss of chloroplasts and photosynthetic ability and are consequently termed secondary osmotrophs, like *Euglena longa* and *Euglena quartana* (Müllner et al. 2001, Preisfeld et al. 2000). Also phototrophic euglenoids themselves evolved from phagotrophic ancestors by way of secondary endosymbiosis. This process occurred when a phagotrophic ancestor engulfed green algal prey cells and the chloroplast was retained, which after a long time of gene transfer and rearrangement, enabled the host cell to carry out photosynthesis beside pinocytotic uptake of nutrients (Gibbs 1978, Leander et al. 2001 & 2007).

## 1.3 The evolution of plastids - Endosymbiotic Theory

Plastid origin was first noticed by Schimper (1883), when he observed that plastids of plants resembled free-living cyanobacteria. These thoughts were taken on by Mereschkowsky (1905) who formulated the endosymbiotic theory and is nowadays seen as the founding father of that theory (Archibald 2015, McFadden 2001). He designated plastids as ‘little green slaves’ that had once been free-living organisms. Rediscovered and recognized was the endosymbiotic hypothesis 1967 by Lynn Margulis. She hypothesized the presence of extranuclear DNA in eukaryotes and presumed that the three organelles mitochondria, chloroplasts and basal bodies of flagella were once free-living bacteria and evolved as symbiotic bacteria, leading to a eukaryotic cell as the result of symbioses. Since that time, numerous analyses have proposed various endosymbiotic scenarios and models. From the mid-1980s on, with the previous success story of molecular sequencing, the evolutionary scenario of endosymbiosis is unquestioned and the only appropriate explanation for eukaryotic phototrophs (Archibald 2015, Gray 2012, McFadden 2001). Nowadays it is clear that mitochondria and plastids arose from independent endosymbiotic events. Both symbionts were reduced over a long period of time and finally integrated in their host, but retained the modified genome. The ancestor for plastids are cyanobacteria, which are photosynthetically active bacteria endowed with chlorophyll a (Archibald 2009). Mitochondria originated earlier in the eukaryotic evolution when an alpha-proteobacterium was engulfed by a primitive, heterotrophic eukaryotic ancestor giving rise to the respiratory organelles. As stable characteristic of eukaryotic cells mitochondria evolved stringently vertically since they first arose (Archibald 2015, Gray 2012, Muñoz-Gómez et al. 2017).

In contrast, analyses on gene sequences encoded in plastids and nuclei of algae and higher plants makes it clear that plastid evolution happened more than once, vertically and horizontally from one eukaryote to another or led to loss of plastids (Archibald 2009 & 2015, Keeling 2010). The primary endosymbiosis of plastids describes the original uptake, when a mitochondrion containing eukaryote phagocytized and ‘enslaved’ a cyanobacterium and gave rise to the common ancestor of the supergroup Archaeplastida of the domain Eukaryota (Archibald 2015, Baldauf 2008, Keeling 2010). The dating of the primary endosymbiosis remains uncertain and contentious, outcomes of molecular clock analysis proposed that this key event happened before 1.6 billion years ago (Yoon et al. 2004). In contrast, cross-calibrated phylogenetic techniques of ATPase proteins suggested that the primary endosymbiosis of plastids occurred about 900 million years ago (Shih & Matzke 2013). Furthermore the results of fossil dates are open to varied interpretations and yielded dissenting results, with eukaryotic fossils that indicated to be 2.1 billion years old (McFadden 2014). Archaeplastida are composed of the three major photosynthetic lineages red algae, glaucophyte algae and green algae including land plants (Adl et al. 2005, Archibald 2015, Baldauf 2008). Plastids of algae and plants from primary endosymbiosis all contain necessarily chlorophyll a and differing amounts of phycobiliproteins, which are accountable for reddish or blueish color of the plastids. They possess two chloroplast membranes, which are thought to stem from the inner and outer membranes of the gram-negative cyanobacterium based on the presence of certain lipids and membrane proteins. Following this scenario, the phagosomal membrane of the host got lost (Adl et al. 2005, Cavalier-Smith 1982, Keeling 2013). Over a long time genes were transferred from the cyanobacterium into the host nucleus (EGT, endosymbiotic gene transfer) until the cyanobacterium became genetically integrated and finally a semiautonomic plastid arised (Keeling 2010).

Plastid genomes usually encode less than 200 proteins, but to establish a fully functional plastid more than a thousand genes are needed. Most of them are translated on cytoplasmic ribosomes as a consequence of EGT and targeted post-translationally into the plastid by a very complex import apparatus (Archibald 2007). This horizontal endosymbiotic gene transfer required the establishment of an efficient translocation system to transfer nucleus-encoded targeted pre-proteins back into the organelle in which they functioned. TIC (translocon inner membran complex) and TOC proteins (translocon outer membran complex) of chloroplasts are an example of such a translocation system for re-targeting of the nucleus-encoded proteins to the organelle and is nowadays the best known plastid protein import machinery (Strittmatter et al. 2010). A massive transport of genes to the nucleus happened parallel, so that the endosymbiont lost its independency and became dependent on the host nucleus. This genetic integration between the endosymbiont and host was an important step in plastid establishment and paved the way for organelle evolution (Bhattacharya et al. 2007, Gentil et al. 2017, Keeling 2010, McFadden 2014).

The monophyletic origin of primary plastids in the Archaeplastida was, among others, supported by characteristics of plastid genome structures and tested by various different phylogenetic studies, which based on mitochondrial genes, plastid genes and/ or nuclear genes encoding for plastid proteins and is nowadays unambiguously accepted. New investigations on the amoeba *Paulinella chromatophora* identified a second independent primary endosymbiosis that does not belong to the plastid clade of Archaeplastida (Chapter 1.4) (Lauterborn 1895, Melkonian & Mollenhauer 2005). This means that the ancestry of all (but one) plastids can be traced back directly or indirectly to the event of primary endosymbiosis (Baldauf et al. 2000, Keeling 2009, Mackiewicz & Gagat 2014, Palmer 2003, Rodríguez-Ezpeleta et al. 2005). The question of monophyly of primary photosynthetic organisms also raised the question of branching order and which lineage diverged first. Molecular analyses have been widely used to infer this relationship, but supported contradictory evolutionary reconstructions, with each of the three lineages as the first algal group to have emerge (Keeling 2004, Mackiewicz & Gagat 2014). Reasons for varying phylogenies might be found in inconclusive taxon sampling and differing choice of genes. Currently, the early emergence of glaucophytes is supported by the peptidoglycan cell wall as a characteristic of the ancestral cyanobacterium, although evidence on the primary branching algae remains scarce. The peptidoglycan wall has been lost in all other plastids (Archibald & Keeling 2002, Reyes-Prieto & Bhattacharya 2007), but depending on outgroup and sampling of sequences there are still studies, which observed different taxa as first emergence and a irrefutable result could not be provided until today (Deschamps & Moreira 2009, Jackson & Reyes-Prieto 2014, Rodríguez-Ezpeleta et al. 2005). Another support for the early emergence of glaucophytes is that only they retained the cyanobacterial fructose bisphosphate aldolase, while green and red algae, which appeared in a second and maybe third lineage had it replaced by a nuclear encoded cytosolic enzyme (Keeling 2010).

With a look at modern mega-phylogenetics, it becomes quite clear, that not all photosynthetic lineages bear offsprings of the primary endosymbiosis event. The so called primary plastids are solely found in Archaeplastida and only account for a fraction of eukaryotes capable of performing photosynthesis (Fig. 1.2, *).

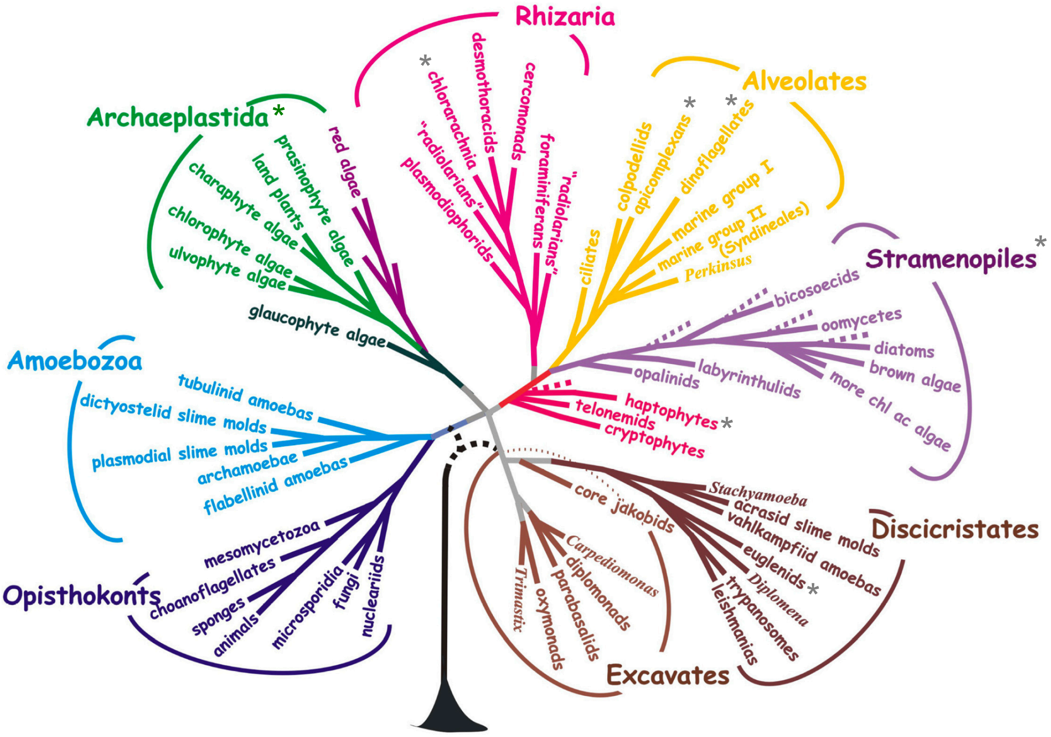


**Fig. 1.2:** Eukaryotic tree by molecular data. Primary plastids marked with a green asterisk. Plastid containing lineages are marked with grey asterisk. (modified from Baldauf 2008).

The question arises, how these many photosynthetically active protists attained their plastids. Nowadays, this can be explained by an ongoing process of endosymbiotic events, by which an algal cell was taken up and retained by another single celled eukaryote (Delwiche 1999, Gibbs 1978, Gould et al. 2008, Keeling 2010). This process of secondary endosymbiosis between two eukaryotes allowed the former heterotrophic host to be photosynthetic with a solar-powered plastid, called complex plastid. With the proliferation of secondary endosymbiosis many ecologically important organisms developed rapidly and are now generally termed phytoplankton. This diverse paraphyletic group consists of single celled photosynthetic organisms (protists) (Adl et al. 2012), which colonized marine and freshwater ecosystems and are key drivers for climate and ecology. Astonishingly, they only account for 1 % of the photosynthetic biomass on earth, but nevertheless they are responsible for 45 % of our planet's annual net primary production (Falkowski et al. 2004, Keeling 2010). This event resulted in the rise of plastid diversity and therewith the spread of photosynthetic organisms from the equator to the poles in (at least) two distinct lineages descending from green or red algae as endosymbionts (Gould et al. 2015).

First investigations on the chloroplast architecture of *Euglena* led to the speculation that organisms with complex plastids, which are surrounded by more than two membranes, evolved by a eukaryote-eukaryote endosymbiosis and that the membranes are a consequence of the phagotrophic mechanism by which they were ingested (Archibald 2007, Gibbs 1978, Keeling 2004 & 2010). In addition to the varying number of chloroplast membranes a significant difference between primary and secondary (complex) plastids is their localization inside the host. Primary plastids are located within the cytosol compartmentalized by two surrounding membranes, whereas secondary plastids are sometimes situated within the lumen of the endomembrane system. The secondary endosymbiosis happened more recently and not just once, but several times in different eukaryotic lineages. The engulfed primary algae degenerated over the time so that only the plastid remained (in most cases). Predominantly the nucleus and all other organelles of the engulfed cell are absent (Archibald 2007, Gibbs 1978, Keeling 2010). Only in the chlorarachniophyte and cryptophyte lineages the nucleus stayed as a nucleomorph (Curtis et al. 2012).

The secondary endosymbiont integration called for an even more massive gene transfer from the symbiont nucleus - as long as it stayed within the host - into the host nucleus (Archibald & Keeling 2002, Gould et al. 2015, Keeling 2010). One important question that arose was how pre-proteins synthesized in the host nucleus are transferred to the outer one or two membranes to reach the TIC/ TOC system in the inner two membranes. Another one was, whether all algal arose from one ancestor (Gould et al. 2015). These key problems resulted in many discussions and investigations about the number and nature of symbiotic events and the origin of complex plastids in organisms (Gould et al. 2015). Today it is recognized that most of the known algae acquired their plastids through secondary endosymbiosis by lateral spread between distantly related eukaryotic lineages, resulting in two important lineages distinguished by pigmentation: One that encompasses all the red algal secondary plastids. These include haptophytes, cryptomonads, stramenopiles (heterokonts), dinoflagellates and apicomplexans and they alone represent half of the presently described protist species.
Another one consisting of chlorarachinophytes and euglenoids, which acquired their plastids from green algae. To this day no secondary plastids are known that derived from glaucophytes (Fig. 1.3). Although these two paths of evolution are accepted nowadays the question as to how many times the plastids moved between eukaryotes has not been satisfactorily answered (Archibald 2007, Archibald & Keeling 2002, Keeling 2004 & 2010).

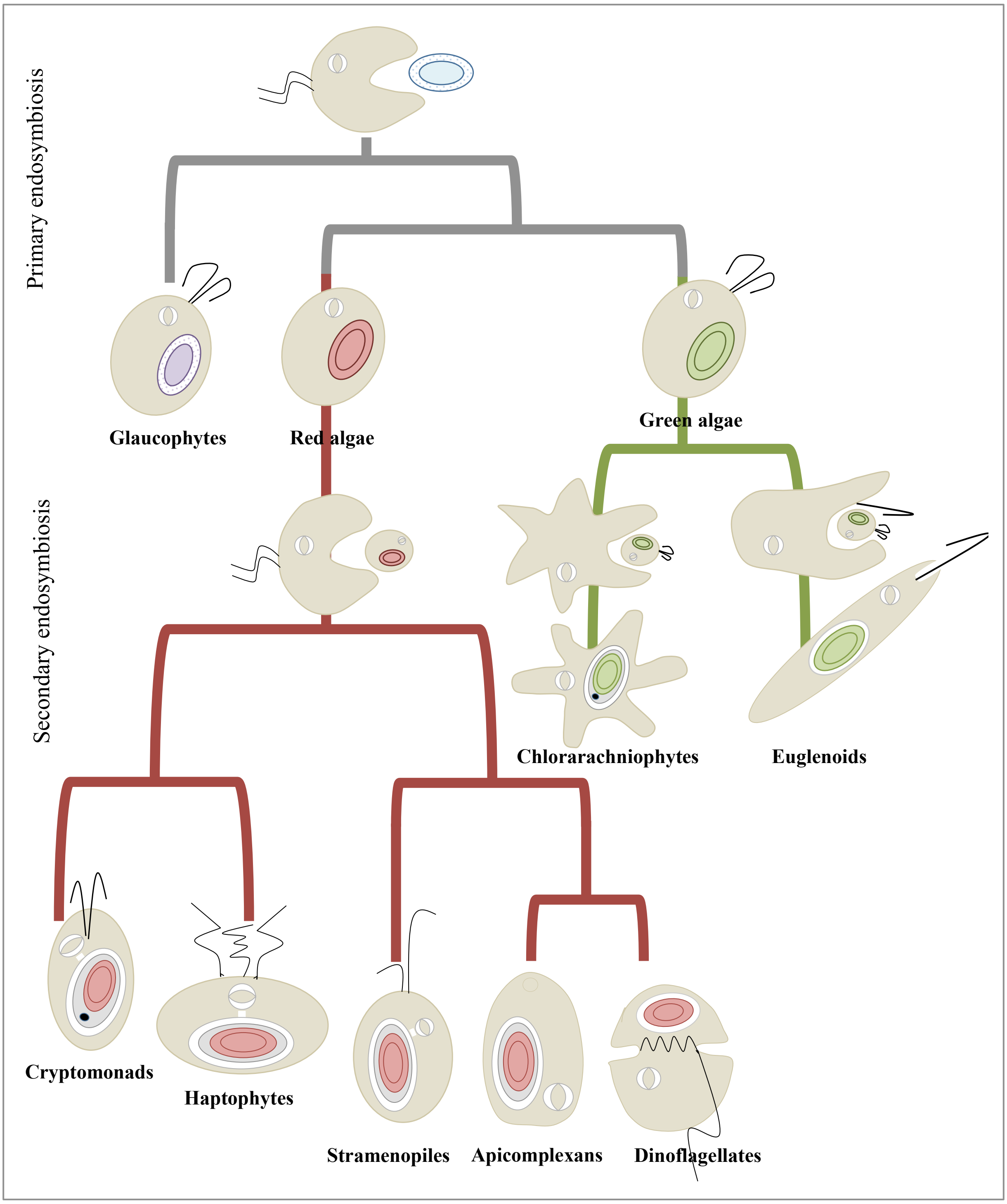


**Fig. 1.3:** Scheme of primary and secondary endosymbiosis that gave rise to phototsynthetic eukaryotes.
The single primary endosymbiosis of an ancestral cyanobacterium on top is leading to the monophyletic Archaeplastida with glaucophytes (purple chloroplast), red algae (red chloroplast) and green algae (green chloroplast), which possess plastids with two membranes. The lines for the event of primary endosymbiosis are coloured grey. Secondary endosymbiosis events are represented by red algal lineages (red) and green algal lineages (green) resulted in complex plastids surrounded by three (euglenoids, dinoflagellates) or four membranes (chlorarachniophtes, apicomplexans, stramenopiles, haptophytes, cryptomonads). Within representatives of the red lineage the outer plastid membrane is related with the ER (cryptomonads, haptophytes, stramenopiles). Chlorarachniophytes and cryptomonads both retain the nucleomorph between the inner and outer pairs of the plastid membranes, within the periplastidal-space (black circle) (presentation form modified from Keeling 2010).

The hypothesis that chlorarachinophytes and euglenoids arose from green algae is supported among other facts by comprising both chlorophyll a and chlorophyll b, the possession of which they have in common with primary plastids of green algae and land plants. Chlorophyll a is possessed by all photosynthetic eukaryotes organisms as their main light-harvesting pigment. Together with chlorophyll b, which is a synapomorphy of green algae, it can also be found in the plastids of euglenoids and chlorarachniophytes as well as in primary plastids of green algae and land plants (Archibald & Keeling 2002). The advances of molecular and phylogenetic data, like chloroplast genome sequencing of euglenoids and comparison with green algae, supported a green-algal origin (Hallick et al. 1993, Turmel et al. 2009). Early investigations assumed that chlorarachniophytes and euglenoids arose from the same secondary green algal endosymbiosis event. The so called *cabozoa* hypothesis (Cavalier-Smith 1999) is based on the principle of parsimony and assumed that both taxa obtained their plastids in a single endosymbiotic event. Their protein-targeting systems are closely related and underpinned the hypothesis that a complex process like a protein targeting system should considerably limit the number of occurrence (Archibald & Keeling 2002, Cavalier-Smith 1999, Keeling 2010). Currently, there is no evidence supporting the *cabozoa* hypothesis. Indeed, nowadays there is wide agreement that the chloroplasts of the two groups of the green lineage have no strong similarity and acquired their plastids independently. This is supported on the plastid side by the occurrence of three chloroplast membranes, paramylon granules (complex beta 1-3-glucan) surrounded by a membrane in the cytosol and group II/ group III intron proliferation in the chloroplast genomes of euglenoids. In chlorarachniophytes four membranes surround the plastids and - most importantly - they are equipped with a nucleomorph and beta-1-3-glucan in the cytosol without an enclosing membrane (Keeling 2010).

The nucleomorph is the nucleus genome of the endosymbiont nucleus (the engulfed green alga) and nested between the inner and outer pairs of the plastid membranes within the periplastidal-space, which represents the cytosol of the primary alga (van Dooren et al. 2001). Additionally, both phylogenetic and -genomic analyses approved that the plastids and the hosts of euglenoids and chlorarachniophytes are not closely related to each other. Chlorarachniophytes belong to the Rhizaria, whereas euglenoids are members of the Excavata (Archibald & Keeling 2002, Keeling 2010 & 2013).

Red line endosymbiont evolution and the question as to how many secondary endosymbioses took place in this lineage is more complicated and insufficiently answered, due to still unexplained host phylogenies and the number of groups which are involved (Archibald & Keeling 2002, Keeling 2013). One main hypothesis is the economical *chromalveolate* hypothesis (Cavalier-Smith 1999), which proposed that all red algal plastids with chlorophyll a and c, Chromista *sensu* Cavalier-Smith (1986) and Alveolata, can be traced back to a single endosymbiotic event. It also includes all non-photosynthetic relatives (Adl et al. 2005, Archibald 2015, Keeling 2010, McFadden & Waller 1997). In Chromista cryptophytes, haptophytes and stramenopiles are grouped together and can be characterized as organisms whose plastids are located in the lumen of the endoplasmic reticulum (ER) (Cavalier-Smith 1986, Yoon et al. 2002). Cavalier-Smith (1999) proposed that the number of evolutionary schemes should be limited, because it is more parsimonious to limit the number of complex events such as the gene transfer and the development of a plastid-targeting system (Cavalier-Smith 1999). With the establishment of this hypothesis over 50 % of all described protists are chromalveolates (Keeling 2009).

Over time, various extensive phylogenetic analyses with conflicting results of nuclear mitochondrial and/ or plastid genes between the major subgroups tried to substantiate or refute a single origin for the red plastids of these organisms. No result seemed to be satisfactory for all the data, because a single data set that specifically combines all chromalveolates is still missing. Phylogenetic support that strongly related the major subgroups of Chromalveolata to one another is always based on datasets which united different subsets of the super-group (Keeling 2009; Archibald & Keeling 2002, Gould et al. 2015).

For scientists one of the most relevant plastid-related evolutionary events during secondary endosymbiosis was the development of a protein targeting system for complex plastids. This led to investigations of the two outer membranes of red complex plastids, their origin and how nucleus encoded pre-proteins return back from their new site of synthesis into the plastid across four membranes. The outermost membranes of cryptophytes, haptophytes and stramenopiles (membrane 1) are identical. They carry ribosomes and are related with the hosts ER, and hence termed chloroplast endoplasmatic reticulum (CER). Through the Sec61 translocation complex pre-proteins, which are equipped with an N-terminal signal peptide, reach the first intermembrane space. While in alveolates the outermost membrane is not attached to the ER (lack of CER), they use vesicle trafficking for transport from the secretory to the outer plastid membrane (Gould et al. 2015, van Dooren et al. 2001). From outside viewed membrane number 2 derived from the plasma membrane of the former symbiont and is referred to as periplastidal membrane, in which a symbiont specific ERAD-like machinery is located. Phylogenetic and molecular studies assumed that the symbionts' endoplasmic-reticulum-associated protein degradation (ERAD) machinery was recycled, relocalized from the symbionts' ER to the second plastid membrane and was used to transport pre-proteins across the second outermost plastid membrane (SELMA = symbiont specific ER-like machinery). Investigations on the origin of the SELMA-machinery in all Chromalveolata appear to point at homologous origin. This monophyly then again emphasizes strong evidence to the *chromalveolate* hypothesis (Gould et al. 2015, Peschke et al. 2013). The two innermost membranes (membrane 3 and 4) built the chloroplast envelope membranes homologous to the primary plastid of cyanobacterial origin. Transport across these membranes is presumably conducted by a TIC/ TOC- like translocation machinery (Gould et al. 2015, Peschke et al. 2013).

Although further investigations of the metabolic enzyme gyceraldehyde-3-phosphate dehydrogenase (GAPDH) among red algae gave molecular support to the *chromalveolate* hypothesis and additional analyses of nuclear rRNA between heterokonts and alveolates as well concretize a monophyletic origin of red algae, there are still alternative views that exist (Archibald & Keeling 2002, Gould et al. 2015, Keeling 2009). One alternative assumption is that inside the secondary endosymbiosis of red algae variants of tertiary endosymbiosis occurred. The latter imply the incorporation of algae with secondary plastids followed by reduction of membrane and cell compartments to get back to the four membranes which enclose most complex plastids. The question that still remained is which way is more parsimonious, plastid loss or plastid gain (Archibald 2015)? Examples of tertiary endosymbiosis already exist within dinoflagellates and plastid losses have been described within apicomplexan parasites, for instance (McFadden & Waller 1997).

## 1.4 A second primary endosymbiosis event - an exception to the rule

The thecate (oval shaped lucid shell), filose amoeba *Paulinella chromatophora*, a member of the supergroup Rhizaria, changed the view that the primary endosymbiosis was a unique event and that the supergroup Archaeplastida obtained their chloroplast from an ultimate singular acquisition (Adl et al. 2005, Gentil et al. 2017, Nowack 2014). Most members of the genus *Paulinella* are marine heterotrophs, but the unicellular eukaryote species *P. chromatophora* Lauterborn (1895) lost its feeding apparatus after phagocytosis and instead engaged photosynthesis with two photosynthetically active blue-green chromatophores obtained from a *Synechococcus* α- cyanobacteria. Investigating its evolution by modern analysis makes *P. chromatophora* an autotrophic species that has undergone an independent and more recent primary endosymbiosis about 60 million years ago and does not belong to the plastid clade of Archaeplastida (Gentil et al. 2017, Marin et al. 2005, Melkonian & Mollenhauer 2005, Nowack & Grossman 2012, Nowack 2014). Comparative analyses between the chromatophore and cyanobacteria resulted in a similar double membrane architecture with a peptidoglycan wall in between. Investigations of EGT in *P. chromatophora* revealed a minimum of over 30 nuclear genes of possible chromatophore origin (Nowack et al. 2016). Thus, the chromatophores of *P. chromatophora* are the only known cyanobacterial descendants, besides plastids of Archaeplastida that enable photosynthesis to their eukaryotic host. (Nakayama & Ishida 2009, Nowack & Grossman 2012, Nowack et al. 2011, Nowack 2014).

## 1.5 Detailed view on the plastid origin in euglenoids

For a long time, *Euglena* has been classified within the Chlorophyta due to chlorophyll a and b as primary pigments in the chloroplasts, although nearly all its ultrastructural characteristics show striking difference from green algae. One important difference is the fact that chloroplasts of *Euglena* are surrounded by three membranes (Gibbs 1978). Gibbs proposed quite early that the capacity for photosynthesis in euglenoids originated from the acquisition of chloroplasts by a phagotrophic euglenoid via secondary endocytobiosis of a green alga (Gibbs 1978). Today, engulfed green alga cn be seen as a close relative to the genus *Pyramimonas* (Turmel et al. 2009). Still, the existence of only three membranes surrounding the chloroplast of euglenoids is not easily explained, because it is quite contrarily to the typically found four membranes in algae that underwent secondary endocytobiosis.

The latter is consistent with the simplest model of secondary endosymbiosis and a typical eukaryote-eukaryote phagocytosis result (Archibald & Keeling 2002). The two inner membranes correspond to the chloroplast of the primary endocytobiosis, the third membrane is the plasma membrane of the primary host (green alga) and the outer membrane remains as the food vacuole membrane (phagosome) of the secondary host (Archibald & Keeling 2002, Keeling 2013).

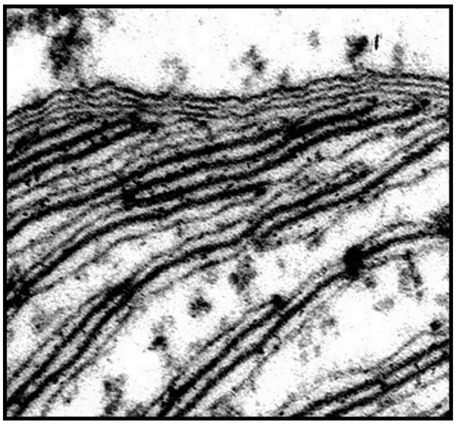


**Fig. 1.4:** Transmission electron micrograph of Euglena gracilis. Three chloroplast membranes, thylakoids and stoma are visible. (1cm = 0.15µm) (courtesy of Dr. Uwe Kahmann).

To explain the three membranes (Fig. 1.4) found in euglenoids (and dinoflagellates) several scenarios have been proposed (Archibald & Keeling 2002). One explanation was that the three membranes originated from a different feeding mechanism known as myzocytosis, a process in which only the cytoplasm and organells of the prey cell are ingested by a predator, rather than engulfing the whole cell (Archibald & Keeling 2002, Gibbs 1978, Keeling 2010, Schnepf & Deichgräber 1984). A much easier explanation is, that the process of plastid origin is the same for four and three membrane plastids and that loss of a membrane occurred (Archibald & Keeling 2002, Gibbs 1978). Nowadays, it is assumed that the plasma membrane of the engulfed alga (membrane number 2 from outside) was digested in euglenoid and dinoflagellates chloroplasts resulting in three membranes with the former phagosome membrane as the outermost (Archibald & Keeling 2002, Keeling 2010).

## 1.6 Main characteristic attributes of euglenoids

The aquatic single-celled euglenoids are a diverse group of protists, which are characterized morphologically by ultrastructural features, which offer the opportunity for investigations on cell character evolution. The cell shapes show a wide variety from round to elongate to spiral, but despite this diversity they all share distinct structural features like a pellicle, as a specifically structured cell membrane with an underlying protein layer, a paraxonemal rod of crystalline proteins in flagella and in most euglenoids paramylon or special feeding apparatuses (Hubert-Pestalozzi 1955, Leedale 1967, Pringsheim 1956).

The **pellicle** as external boundary is a morphological character that can be of rigid or flexible form and therewith determine cell plasticity. The complex structure consists of a plasma membrane, with parallel proteinaceous strips underneath and several rows of microtubules (Fig. 1.5). The strips extend along the entire length of the cell, arranged longitudinally or helically and are closely associated with tubular cisternae of the endoplasmatic reticulum (Dragoş et al. 1997, Farmer 2009, Leander et al. 2007, Leander & Farmer 2000). Each strip is connected with the neighbour strip, which enable euglenoids with flexible pellicle to change their shape dynamically, which is termed euglenoid or metabolic movement (Farmer 2009, Schwartzbach & Shigeoka 2017). Euglenoid movement is thought to facilitate the ingestion of large food particles in phagotrophic euglenids and pave the way for the secondary endosymbiosis that enable photosynthesis. Many phototrophic euglenoids are still able of euglenoid movement as a relic of their phagotrophic ancestors (Farmer 2009, Leander & Farmer 2000, Leander 2004, Leander et al. 2007, Schwartzbach & Shigeoka 2017). Below the pellicle there are muciferous bodies that secrete mucus through the surface of the cell. The mucilage is for example involved in the lorica development of Trachelomonas and Strombomonas where it then mineralized or for the gliding movement on surface. At the anterior front of the cell lies an invagination with a reservoir out of which the flagellum/ a emerge (Buetow 1982, Ciugulea & Triemer 2010).

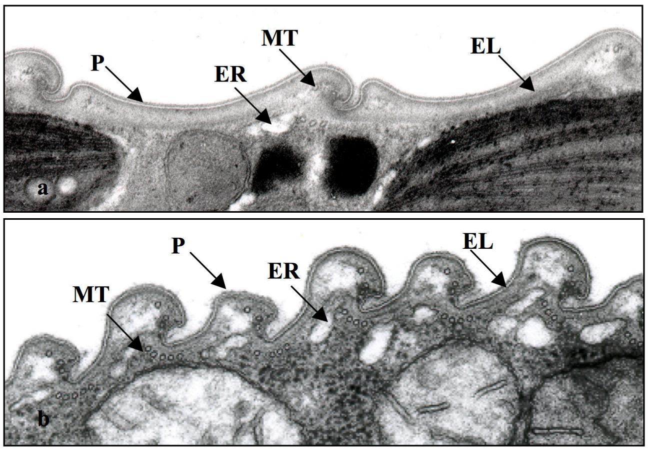


**Fig. 1.5:** Transmission electron micrograph of the euglenoid pellicle. a) Rigid pellicle of *Phacus similis*. b) Flexible pellicle of *Euglena quartana*. P: Pellicle, EL: Epiplasmatic layer, MT: Microtubules, ER: Endoplasmic reticulum, 1cm = 0.2 µm (© Angelika Preisfeld).

The euglenoid **flagellar apparatus** arises from a basal body complex in the reservoir at the front of the cell. This complex comprises short cylinders of three microtubular roots and two basal bodies, each of which giving rise to a flagellum, named dorsal or ventral flagellum, according to Leedale´s terminology (1967) and depending on their position in the cell. As in most other eukaryotes, the flagellum is composed of a cylinder of nine doublet microtubules surrounding two central microtubules (Farmer 2009, Leander 2004, Mitchell 2007, Rosati et al. 1991). Parallel to the highly conserved axoneme in euglenoids run paraxonemal rods of highly organized proteins with so far unconfirmed function. Most euglenoids possess two dynamic flagella, adorned on the surface with flagellar hairs (mastigonemes), and emerge from the base of the anterior opening of the flagellar pocket (Fig. 1.6). Some euglenoids have reduced one flagellum in length, with the emergent one leaving the reservoir, or simply have just one emergent flagellum like *Petalomonas cantuscygni*. The majority of phototrophs have one reduced flagellum, for instance *Euglena gracilis* swims using one emergent locomotory flagellum, the ventral flagellum of *E. gracilis* is shortened and hidden in the reservoir. Others have more than two flagellar, for example *Eutreptiella pomquetensis* holds four emergent flagella (Farmer 2009, Farmer & Triemer 1988, McLachlan et al. 1994, Mitchell 2007, Rosati et al. 1991, Shin et al. 2001).

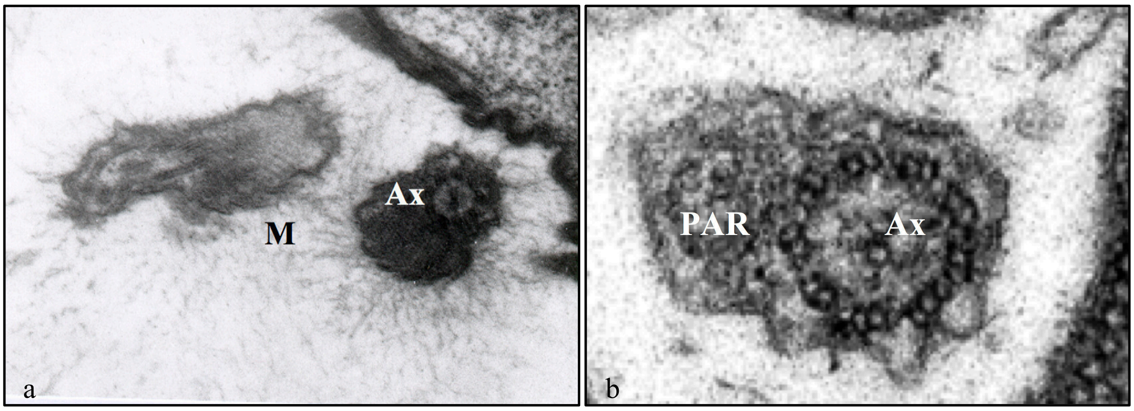


**Fig. 1.6:** The euglenoid flagellum with mastigonemes and associated paraxonemal rod. a) Cross section of a flagellum with axoneme (Ax) and mastigonemes (M) on the surface of *Distigma proteus*, 1 cm = 0.25 µm. b) Ultrastructure of transverse section through the flagellum of *Euglena quartana* showing the axoneme (Ax) and the paraflagellar rod (PAR), 1 cm = 73,6 nm (© Angelika Preisfeld).

Associated with the flagella of phototrophic euglenoids and some secondary osmotrophic euglenoids is the light perception flagellar apparatus consisting of a stigma (eyespot) and a paraflagellar body (Fig. 1.7). The red eyespot is located in the cytoplasma, adjacent to a particular portion of the flagellar reservoir and entails carotenoid pigment granules. The paracrystalline paraflagellar body is attached to the emergent dorsal flagellum and is responsible for light detection. Euglenoids with an eyespot can orientate themselves by use of the stigma shading the paraflagellar body. The stigma/ flagellar apparatus enables cells to positive phototaxis allowing them to respond to intensity and direction of light (Farmer 2009, Kivic & Vesk 1972, Rosati et al. 1991).

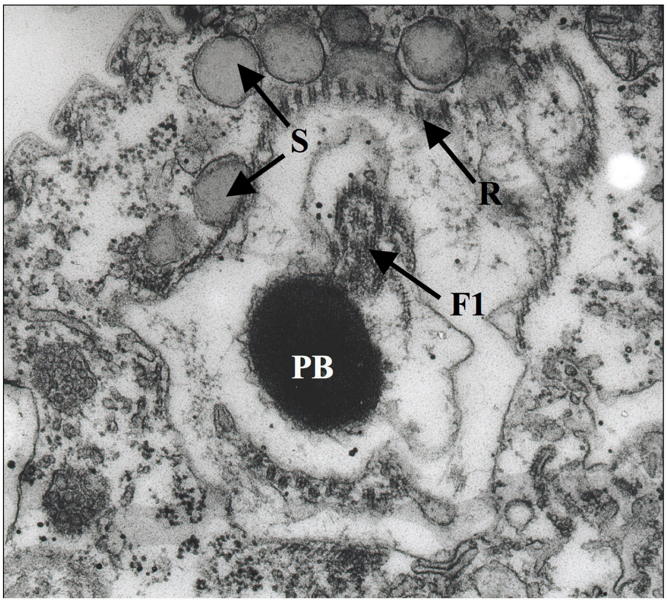


**Fig. 1.7:** Transmission electron micrograph showing a section of the light perception flagellar apparatus consisting of the flagellar reservoir with attached paraflagellar body and stigma of *E. gracilis*. F1: Dorsal flagellum, PB: Paraflagellar body, R: Reservoir, S: Carotenoid-stained lipid globuli of the stigma. 1cm = 0.24 µm (© Angelika Preisfeld).

The **Paramylon** is the food storage product and energy reserve of euglenoids is a ß-1,3-glucan and distinct to the starch (α-1,4-glucan) found in green algae (Bäumer et al. 2001). Gottlieb (1850) first extracted these energy reserve granules from a phototrophic *Euglena* and proposed to name them paramylon (Fig. 1.8). Membrane bound crystalline paramylon grains appear in the cytoplasm of all photosynthetic euglenoids and/ or cap the pyrenoids on one or both sides outside of the chloroplast and can be small, large or dimorphic (presence of two size classes of grains within a cell). Their occurrence is not correlated with the presence of chloroplasts, because paramylon is also located in heterotrophic euglenoids and is frequently used as a morphological classification system to infer generic relationship. Therefore diagnostic features like amount, shape, location and external morphology are used to support major clades on generic level (Bäumer et al. 2001, Ciugulea & Triemer 2010, Gojdics 1953, Monfils et al. 2011, Paerschke et al. 2017).

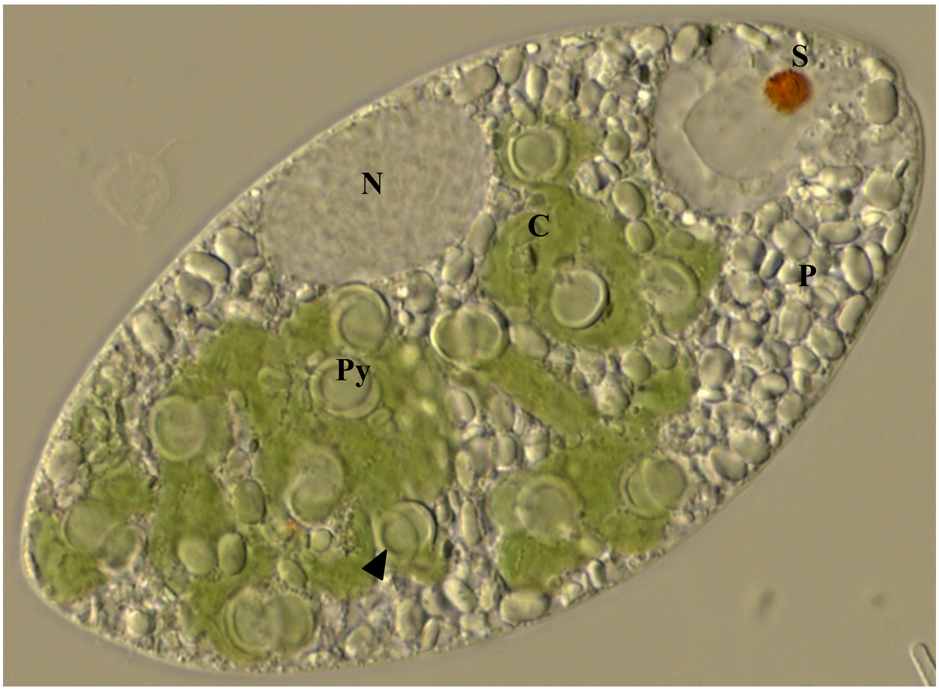


**Fig. 1.8:** Freshwater phototrophic euglenoid *Euglena velata*. C: Chloroplast, N: Nucleus, P: Paramylon, Py: Pyrenoid, S: Stigma, Black arrow: Paramylon cap (© Angelika Preisfeld & David J. Patterson).

**Chloroplasts** of phototrophic euglenoids are most similar to the chloroplasts of green algae and the result of secondary endosymbiosis. They are the organelles in which photosynthesis takes place, which contain chlorophyll a and b pigments. Their thylakoids are arranged in stacks of three. Between taxa and even between species the number, location and morphology of chloroplasts is highly diverse, they can be of discoidal, lobed, spherical or stellate shape, with or without pyrenoid. These variable morphological characteristics played a major role in the taxonomy of euglenoids. The pyrenoid is an area densely packed with ribulose -1,5-bisphospahte carboxylase oxygenase (RuBisCO), the $CO_2$ fixation enzyme. RuBisCo is composed of a small nuclear encoded subunit and a large chloroplast encoded subunit giving thereby evidence of endosymbiotic gene transfer during the process of secondary endosymbiosis (Ciugulea & Triemer 2010, Farmer 2009, Schwartzbach & Shigeoka 2017). To provide further information about this process and to understand the biology of phototrophic euglenoids in the last years, the chloroplast genomes (cpGenomes) of several euglenoids have been sequenced, annotated and published (Bennett et al. 2012, 2014 & 2017, Bennett & Triemer 2015, Dabbagh et al. 2017, Dabbagh & Preisfeld 2017, Gockel & Hachtel 2000, Hallick et al. 1993, Hrdá et al. 2012, Kasiborski et al. 2016, Pombert et al. 2012). The first cpGenome was the circular chromosome of the model organism *Euglena gracilis* strain Z by Hallick et al. (1993) with a size of 143 bp. Phylogenetic investigations indicated that the plastids of phototrophic euglenoids, surrounded by three membranes related to the members of the genus *Pyramimonas* (Turmel et al. 2009). Although the genome of *E. gracilis* is larger than that of the presumable closest relative *Pyramimonas parkeae*, during the process from a green algae chloroplast over an endosymbiont to the plastid organelle of another host, the chloroplast genome underwent distinct loss of genes, which were transferred to the nuclear genome (Schwartzbach & Shigeoka 2017). This means that although fewer genes exist in the genome of *E. gracilis,* the genome is much larger than the one of *P. parkeae*. It took some time to understand the contradiction, but nowadays it is known that the chloroplast genomes of phototrophic euglenoids have an unusual high number of introns in comparison to green algae. The genome is littered with so-called group II and group III introns. These introns can be located intergenic, within the coding region of protein-coding genes, within the genes that comprise the ribosomal RNA (rRNA) operon or even within other introns, resulting in twintrons or complex twintrons (Copertino & Hallick 1991, Copertino et al. 1991 & 1992, Doetsch et al. 1998 & 2001, Michel et al. 1989, Thompson et al. 1995 & 1997).

Now the phylogenetic relationships of euglenoids in the eukaryotic tree of life become increasingly understandably and the relationship of green algae and phototrophic euglenoids characterized by the acquisition of chloroplasts seemed unequivocal. Though little is known about the evolution of the chloroplast genome within the phototrophic euglenoids and specifically which changes occurred during the process of acquisition. All information gathered in the the last decades relied primary on the chloroplast sequence of *E. gracilis* (Hallick et al. 1993). But the diversity of the phototrophic euglenoids justifies further research to infer deep evolutionary relationships among photosynthetic euglenoids. The chloroplast genome exploration provided the opportunity to infer phylogenomic assessments and to support consisting phylogenetic relationships, which are exclusively based on single or multigene analyses mostly of nuclear and chloroplast rDNA sequences and occasionally accompanied by other genes. Furthermore, comparative analyses of these genomes to the closest living relative should clarify important questions regarding the genetic relationship and provide fundamental insights into genomic changes after endosymbiosis of a complex plastid.

## 1.7 Scope of this thesis

Regarding the abovementioned desiderata in euglenoid phylogenomic and phylogeny three carefully selected taxa of phototrophic euglenoids have been used to intensively study differences and similarities in chloroplast genome evolution in this highly diverse group of eukaryotes. Accordingly, all three cpGenomes sequenced stem from different phylogenetic positions to compare them with close relatives as well as distantly related species.

The chloroplast genome of *Euglena mutabilis* SCHMITZ 1884 will be investigated, because the genus *Euglena* possesses the highest morphological diversity in body shape as well as position and type of chloroplasts, as can be seen in the cells of *Euglena gracilis* and *Euglena viridis* which are quite diverse and belong to two different subclades. Hence, it was intended to explore another species of a third subclade and to compare it with the other two. The genome size seemed to be relevant, as well as the structure and number of rRNA operons and the intron number, since the chloroplast genomes of the previously published cpGenomes could not be more oppositional in regard of these features. Nevertheless, the genome alignments show large conserved segments, too. Additionally, another main reason to choose *E. mutabilis* was the fact that it phylogenetically showed the longest individual branch resulting in divergent positions in the Euglenaceaen lineage (Linton et al. 2010, Marin et al. 2003).

Even though the euglenoids have obtained their chloroplasts by the acquisition of chloroplasts from a green alga and in all probability the donor was a relative of the partly obligatory psychrophilic genus *Pyramimonas,* the euglenoid chloroplasts showed overwhelmingly high differences in genome structure compared to the one of *Pyramimonas parkeae* as the closest living relative up to date. Therefore, *Eutreptiella pomquetensis* (MCLACHLAN, SEGUEL & FRITZ 1994) MARIN & MELKONIAN 3003*,* a basal taxon in the phototrophic euglenoids, was another key item of this study as it displays some significant differences in morphology and habitat. *Etl. pomquetensis* is the only known phototrophic euglenoid with four flagella and is characterized as a psychrophilic representative isolated from shallow cold marine habitats, a so far unusual characteristic for euglenoids. Keeping in mind that also the green algal genus *Pyramimonas* contains psychrophilic species, *Etl. pomquetensis* should be compared to the genome structure of *P. parkeae* and the other two Eutreptiales and yield insight into chloroplast and intron evolution.

*Trachelomonas grandis* SINGH 1956 was selected for intrageneric and intergeneric comparisons. It was chosen to compare its cpGenome with those of the genera *Euglena* and *Monomorphina,* both of which displaying conserved segments and appearing to be internally free from genome rearrangements (with the exception of *E. archaeoplastidiata*), and with *Eutreptiella*, whose synteny is quite low. This study wanted to determine whether the results on the selected genera would allow to detect and hypothesize a comprehensible trend concerning intrageneric variability. A phylogenomic approach of all taxa available should give additional information on the position of *T. grandis* (among others), which is known as a species with extremely long nuclear SSU rRNA genes.

One of our objectives concerning all euglenoid cpGenomes available was to discover, which features in the cpGenomes stayed stable and which are changed during the evolution of chloroplasts in the euglenoid lineage. The outcome of these investigations should be exploited to determine changes on a generic and species level in order to ascertain whether these changes follow an evolutionary pathway or if they are indeed single and unrelated events.

Another aspect of main interest was to scrutinize the genome structure, the general gene composition and the syntenic arrangement of gene clusters in the cpGenomes, as well as the number of rRNA operon repeats and their localization in the genome. As the first studies resulted in differing numbers of rRNA operon repeats and moreover showed different positions on the genome, our concern was to investigate further on possible trends in genome composition on an intrageneric level, in all euglenoids and in certain green algal cpGenomes as donors of the chloroplast.

Another incentive was the highly unusual intron biology in euglenoid chloroplast genomes in regard to intron number and type. The chloroplast genomes of phototrophic euglenoids hold group II and group III introns and these can be found solitarily or as twintrons (introns within introns). A still unanswered question is how the unique intron types of euglenoids developed and why only one intron was identified in the Pyramimonadales counterparts. And it is equally unclear, whether these many introns spread across euglenoid genomes horizontally or vertically. For this reason, a novel approach combining RT-PCR and secondary structure analyses should be used to investigate a possible relationship between the different intron groups in euglenoid and green algal chloroplast genomes. These approaches should then help to answer the question as to why species of the same family and even the same genus exhibit such enormous differences in regard to intron possession and position. Since previous studies hypothesized that the unique group III introns are degenerated group II introns (Copertino et al. 1991, Doetsch et al. 1998) it seems just as relevant to support or decline this assumption by a thorough and currently not existing analysis of intron data. Another purpose was to find the common ancestor of the introns in this lineage.

To achieve these goals, secondary structure of domain V and VI introns needed to be folded, examined and used to identify potential group II introns and furthermore to recognize possible similar introns in the coding region of protein-coding genes and rRNA genes in one species or between different taxa.

Since the current understanding of phototrophic euglenoids' evolution is largely based on multigene analyses containing nuclear and chloroplast SSU and LSU rRNA genes, the chloroplast genome data should be used to perform a phylogenomic analysis to determine if phylogenomics can help to clarify ambiguous and controversial positions in trees.

# 2 Material and Methods

## 2.1 Organisms – Strains of euglenoid flagellates

*Euglena mutabilis* and *Trachelomonas grandis* were acquired from the Culture Collection Algensammlung Universität Göttingen (SAG). *Eutreptiella pomquetensis* was purchased from the National Center for Marine Algae and Microbiota (NCMA) of the USA (former Culture Collection of Marine Phytoplankton, CCMP) (Table 2.1).

**Table 2.1:** Species and strains of phototrophic euglenoids from culture collections.

| Species | Order | Strain |
|---|---|---|
| *Euglena mutabilis* SCHMITZ 1884 | Euglenales | SAG 1224-9b |
| *Eutreptiella pomquetensis* MCLACHLAN, SEGUEL & FRITZ MARIN & MELKONIAN 2003 | Eutreptiales | CCMP 1491 |
| *Trachelomonas grandis* SINGH 1956 | Euglenales | SAG 204.80 |

## 2.2 Media and culture conditions for euglenoid flagellates

*E. mutabilis* strain SAG 1224-9b was cultivated in Euglena medium modified after Cramer & Myers (1952) at 20 - 23 °C under 12 : 12 light : dark cycle using fluorescent tubes delivering about 30 µmol photons m$^{-2}$ s$^{-2}$ of light. Species of *T. grandis* SAG 204.80 were grown in WEES medium (Kies 1967) under the same conditions. *Etl. pomquetensis* CCMP 1491 cells were grown in modified L1-Si Medium (Guillard & Hargraves 1993) with artificial seawater Sea-Pure (CaribSea, Inc. Fort Pierce) at 2 - 4 °C with changing 3 : 3 light:dark cycle using ExoTerra Natural Light PT2190 (Hagen).

## 2.3 Isolation of purified chloroplasts

For each isolation procedure cells of phototrophic euglenoids were harvested with approx. 1.56* 10$^{6}$ cells/ ml and concentrated from 200 ml culture suspension (*Etl. pomquetensis* 300 ml) by centrifugation in a swing-out rotor (Eppendorf) at 2,205 x g for 5 min and rinsed three times with isolation buffer modified after Aronsson & Jarvis (2002). During the isolation procedure the cell material was kept at 4 °C. Each pellet was resuspended in fresh isolation buffer and 50 µM proteinase inhibitor Pefabloc® SC (Sigma-Aldrich) to avoid protein destruction. To reduce contamination cells of each species were layered over a three-step gradient of colloidal polyvinylpyrrolidone (PVP) coated silica (Percoll®, GE Healthcare Life Science). Each layer consisted of 10 ml, with the 95 % bottom layer comprising, 9.47 ml Percoll® solution and 0.53 ml gradient mixture, a 60 % middle layer (10 ml) with 6.32 ml Percoll® solution and 3.68 ml gradient mixture and a 30 % top layer (10 ml) with 3.15 ml Percoll® solution and 6.85 ml gradient mixture (modified after Aronsson & Jarvis (2002)). If the expected separation did not occur, gradients were expanded to include a greater number of incremental layers (100 %, 80 %, 50 %, 20 %). Gradients were centrifuged in a swing-out rotor at 2,000 x g for 20 min brake off, as not to cause mixing. Individual euglenoid cells were removed above the 50 %, 60 %, 80 % or 95 % interface, sampled in a falcon tube, brought to a volume of 30 ml with isolation buffer and inverted carefully to remove the Percoll® from the sample. The sample was centrifuged for 2,205 x g for 5 min to pellet the cells and the supernatant was removed. Afterwards, washing steps were repeated 3 x in isolation buffer and then the pellet was resuspended in 5 ml isolation buffer with 50 µM proteinase inhibitor (Pefabloc® SC) (Table 2.2). *Etl. pomquetensis* was then slightly prewashed 3 x for 1 min, because some cells were just disrupted at this early stage of chloroplast isolation. Thereafter, the same proceeding was implemented for all cultures.

**Table 2.2:** Composition of buffers and solutions for chloroplast isolation.

| Buffers and solutions | Component |
|---|---|
| Isolation buffer | 300 mM Sorbitol<br>5 mM MgCl2<br>5 mM EDTA<br>20 mM Hepes/KOH pH 8.0<br>10 mM NaHCO3 |
| Percoll® solution | 95 % (w/ v) Percoll®<br>3 % (w/ v) PEG 6000<br>1 % (w/ v) Ficoll<br>1 % (w/ v) BSA |
| Gradient mixture | 25 mM Hepes/KOH, pH 8.0<br>10 mM EDTA<br>5 % (w/ v) sorbitol |

The cleaned cells were divided into 250 µl homogenate in 1.5 ml tubes. They were disrupted by ultrasonic probe Sonopuls HD 60 (Bandelin) 2 - 6 times for 3 - 10 sec with intermediate washing steps on ice. For *E. mutabilis* the amplitude was set at 60 % with a 0.1 sec pulse rate. *T. grandis* cells were disrupted with the amplitude set at 80 % and *Etl. pomquetensis* at 50 % with a 0.1 sec puls rate (Table 2.). Between each sonication procedure three washing steps were performed. Therefore samples were combined in a falcon tube, pelletized, resuspended in 2 ml isolation buffer and centrifuged in a swing-out-rotor for 1 min (brake off) (Table 2.3). Supernatant with isolated chloroplasts was decanted and the washing procedure repeated twice with the remaining pellet in 2 ml fresh isolation buffer. The sequence of disruption by ultrasonic waves and chloroplast elution was repeated two times (*Etl. pomquetensis*), five times (*E. mutabilis*) or up to six times (*T. grandis*).

**Table 2.3:** Conditions of slight washing for chloroplast elution and ultrasound settings for each phototrophic euglenoid during chloroplast isolation procedure.

| Species | Slight washing: | Ultrasound: |
|---|---|---|
| | 3x 1 min (x g) | repetition x sec amplitude/ puls rate |
| *E. mutabilis* | 259 | 3x 3sec 60 %/ 0.1 |
| *Etl. pomquetensis* | 180 | 2x 3sec 50 %/ 0.1 |
| *T. grandis* | 259 | 6x 3sec 80 %/ 0.1 |

Afterwards, the supernatant from the whole sonication-washing procedure was pelletized and resuspended in 1 - 3 ml isolation buffer, depending on pellet size. The resuspended chloroplasts were loaded onto a primed five step Percoll® gradient with approximately 80 %, 60 %, 40 %, 30 % and 10 % layers from bottom to top. Tubes were centrifuged in a swing-out rotor for 30 min at 2,000 x g brake off. The appearing chloroplast fraction was recovered from the 30 % layer. Isolation buffer was added to the chloroplasts and the tube was inverted carefully one time to wash off the Percoll®. The chloroplasts were centrifuged in a swing- out rotor at 3,645 x g for 5 min (brake on). This step was repeated twice to eliminate Percoll®. Three final washing steps were performed for 1 min to ensure purity of chloroplasts (brake off) (Table 2.3). Purification of chloroplasts was completed by pelleting for 5 min at 3,976 x g (brake on) and integrity of chloroplasts was verified with a fluorescence life-time imaging microscope (Biozero, Keyence).

## 2.4 Isolation of nucleic acid

Prior to preparation of DNA or RNA 1 - 2 ml of culture depending on culture density of phototrophic euglenoids, was centrifuged in a 2 ml aliquot. The sample was spun for 2 min at max speed and the supernatant was removed carefully.

### 2.4.1 Isolation of chloroplast DNA

After Percoll® gradient cpDNA was isolated with My-budget DNA Mini Kit (Bio-Budget Technologies) following the manufacturers protocol. Elution of cpDNA was performed with 50 µl pre-warmed elution buffer (50 °C). A NanoDrop 2000 Spectrophotometer based on A260/ A280 and A260/ A230 ratios (Thermo Fisher Scientific) was used to measure the concentration and purity of extracted DNA (Table 2.4). Afterwards next-generation sequencing was performed (Chapter 2.8.2).

**Table 2.4:** Total amount of isolated chloroplast DNA used for 454 sequencing from 200 ml cell suspension with 1.56* $10^6$ cells/ ml.

| Species | cpDNA ng/ µl | Total volume<br>for sequencing (µl) |
|---|---|---|
| *E. mutabilis* | 73.8 | 15 |
| *Etl. pomquetensis* | 21.9 | 25 |
| *T: granids* | 24.9 | 25 |

### 2.4.2 Isolation of genomic DNA

Whole genomic DNA from cultures of the phototrophic euglenoids was extracted following standard procedure for preparation of DNA. Cells were broken up chemically with lysis buffer and Proteinase K using the My-budget DNA Mini Kit (Bio-Budget). As a slight modification the elution buffer was pre-warmed to 50 °C and elution of total DNA was conducted in two centrifugation steps. Two eluates were generated each using 100 µl of elution buffer and the same column. Qualitiy and quantity of resulting DNA was proved by NanoDrop 2000 Spectrophotometer, then used as template for standard PCR or fill-in PCR experiments and stored at -20 °C.

### 2.4.3 Isolation of plasmid DNA

Preparation of plasmid DNA was carried out using the E.Z.N.A.® Plasmid Miniprep I Kit (OMEGA bio-tek). Therefore 2 ml of a transformant culture was isolated, which had been incubated at 37 °C over night in liquid LB medium (Chapter 2.7). The eluted plasmid DNA was quantified by NanoDrop 2000 Spectrophotometer and used for DNA sequencing and/ or stored at -20 °C.

### 2.4.4 Isolation of total RNA

Total RNA from phototrophic euglenoids was isolated by my-Budget RNA Mini Kit (Bio-Budget) according to manufacturer´s regulation. Qualitiy and quantity of resulting RNA was proved by NanoDrop 2000 Spectrophotometer and then used as template in RT-PCR experiments to determine exon-intron boundaries of protein-coding genes and/ or stored at - 20 °C Amplification of DNA-fragments.

### 2.4.5 Oligonucleotides

The different PCR reactions were elaborated with manually created primers by Primer3Plus (Table 7.1). They were purchased from Eurofins Genomics. Lyophilized oligonucleotides were solubilized in deionized sterile water (ds$H_2O$) to stock solution with a concentration of 100 µmol according to manufacturer's synthese report.

### 2.4.6 Polymerase chain reaction (PCR)

Standard PCR experiments for the amplification of ribosomal DNA fragments and fill-in PCR experiments were executed in thermocyclers (Eppendorf). Primers for fill - in PCR were designed within the flanking regions of sequencing gaps/contigs. PCR reactions were performed using whole genomic DNA. Standard PCR experiments were performed with Dream*Taq*™ Green buffer (Thermo Fisher Scientific), including a density allowing direct loading of PCR products on a gel (Table 2.5, Table 2.6).

Duration of denaturation, annealing and elongation cycle steps varied likewise annealing temperatures, depending on choice of primer pairs (Table 7.1). The results of the PCR experiments were verified by agarose gel electrophoresis using TAE buffer (Chapter 2.5).

**Table 2.5:** Reaction components for standard PCR.

| Component | Volume (µl) |
|---|---|
| 10 x Dream*Taq*™ Green buffer | 2.5 |
| 2 mM dNTPs | 2.5 |
| Dream*Taq* DNA polymerase | 1 |
| Forward primer | 1 - 2 |
| Reverse primer | 1 - 2 |
| Template DNA | 1 - 3 |
| Nuclease-free water | to 25 |

**Table 2.6:** Thermocycling conditions for standard PCR.

| Cycler conditions | | |
|---|---|---|
| Phase | Temp. (°C) | Time |
| Initial denaturation | 95 | 3 min |
| Denaturation | 94 | 30 - 60 sec |
| Annealing | 52-62 | 30 - 60 sec |
| Elongation | 72 | 1 - 4 min |
| Repeat cycle 20 – 35 x | | |
| Final elongation | 72 | 10 min |
| Hold | 4 | ∞ |

### 2.4.7 Colony PCR

Colony-PCR was used as a high-throughput method to examine the presence or absence of insert DNA in the vector of plasmid constructs of a bacterial clone culture. Therefore 1 µl of transformants were used directly as template for PCR experiments. Selected clones were screened for inserts using standard vector primers M13-forward and -reverse from Eurofins Genomics (Table 2.7, Table 2.8). The forward and reverse M13 regions flanked the insertion site within the pCR® 2.1vector (Thermo Fisher Scientific). PCR amplicon and size of the product were reviewed by agarose gel electrophoresis alongside a DNA size marker.

**Table 2.7:** Reaction components for Colony PCR.

| Component | Volume (µl) |
|---|---:|
| 10 x Dream*Taq*™ Green buffer | 2.5 |
| 2 mM dNTPs | 2.5 |
| Dream*Taq* DNA polymerase | 0.5 |
| M13 uni (-21) | 1 |
| M13 rev (-29) | 1 |
| Template DNA | 1 |
| Nuclease-free water | to 20 |

**Table 2.8:** Thermocycling conditions for Colony PCR.

| Cycler conditions | | |
|---|---|---|
| Phase | Temp. (°C) | Time |
| Initial denaturation | 95 | 3 min |
| Denaturation | 94 | 30 sec |
| Annealing | 52 | 30 sec |
| Elongation | 72 | 1 - 2 min |
| Repeat cycle 25 x | | |
| Final elongation | 72 | 10 min |
| Hold | 4 | ∞ |

### 2.4.8 Long-range polymerase chain reaction

Long-range PCR experiments were applied for rRNA operon repeat tests to ensure amplification of large sequences that cannot be amplified using routine standard PCR methods or reagents by different kits:

(1) The LongRange PCR Kit (Qiagen) protocol 1 was used for PCR amplicons up to 10 kb. The cycling conditions for long-range PCR were conducted according to the manufacturer´s instructions used an extension time of 1 min per kb genomic DNA.

(2) DNA was amplified with the Q5® High-Fidelity DNA Polymerase (New England BioLabs), following the manufacturers protocol for amplicons up to 10 kb. The recommended extension step is 20 - 30 sec/ kb at 72 °C.

(3) Long*Amp*$^{®}$ *Taq* DNA Polymerase (New England BioLabs) was used for the amplification of DNA fragments following manufacturers guidelines. The recommended extension temperature is 65 °C. Extension times are generally 50 sec/ kb (Table 2.9, Table 2.10).

**Table 2.9:** PCR reaction setup for Long*Amp*$^{®}$ Taq DNA Polymerase.

| Component | Volume (µl) | |
|---|---|---|
| | 25 µl reaction | 50 µl reaction |
| 5 x Long*Amp Taq* buffer | 5 | 10 |
| 10 mM dNTPs | 0.75 | 1.5 |
| Forward primer | 1 - 2 | 2 - 3 |
| Reverse primer | 1 - 2 | 2 - 3 |
| Template DNA | 1 - 3 | 1 - 3 |
| Long*Amp Taq* DNA Polymerase | 1 | 2 |
| Nuclease-free water | to 25 | to 50 |

**Table 2.10**: PCR thermocycler conditions for Long*Amp*$^{®}$ Taq DNA Polymerase.

| Cycler conditions | | |
|---|---|---|
| Phase | Temp (°C) | Time |
| Initial denaturation | 94 | 30 sec |
| Denaturation | 94 | 30 sec |
| Annealing | 48 - 62 | |
| Elongation | 65 | 5 - 10 min |
| Repeat cycle 30 x | | |
| Final elongation | 65 | 10 min |
| Hold | 4 | ∞ |

### 2.4.9 Reverse transcription polymerase chain reaction

Reverse transcription-PCR experiments (RT-PCR) were carried out to detect exact exon-intron boundaries of protein-coding genes of phototrophic euglenoids cpGenomes. With the OneStep RT-PCR Kit (Qiagen) RNA was transcribed reversely into cDNA and then used as a template for amplification (Table 2.11, Table 2.12). It is advantageous, that reverse transcription and PCR were carried out in the same tube and there was no need to add further components once the thermocycler reaction has been started. All experiments have been performed with gene-specific primers (Table 7.1).

**Table 2.11:** Reaction components for OneStep RT-PCR.

| Component | Volume (µl) | |
|---|---|---|
| | 25 µl reaction | 50 µl reaction |
| 5 x OneStep RT-PCR buffer | 5 | 10 |
| 10mM dNTPs | 1 | 2 |
| Forward primer | 1 - 2 | 2 - 3 |
| Reverse primer | 1 - 2 | 2 - 3 |
| Template RNA | 1 - 3 | 1 - 3 |
| OneStep RT-PCR enzyme | 1 | 2 |
| Nuclease-free water | to 25 | to 50 |

**Table 2.12:** Thermocycling conditions for OneStep RT-PCR.

| Cycler conditions | | |
|---|---|---|
| Phase | Temp (°C) | Time |
| Reverse transcription | 50 | 30 min |
| Initial denaturation | 95 | 15 min |
| Denaturation | 94 | 1 min |
| Annealing | 50 - 62 | 1 min |
| Elongation | 72 | 1 - 2 min |
| Repeat cycle 30 x | | |
| Final elongation | 72 | 10 min |
| Hold | 4 | ∞ |

## 2.5 Agarose gel electrophoresis

This analytic method was used for the separation and visualization of DNA fragments through a TAE agarose gel matrix in an electric field to verify the results of the performed PCR experiments. DNA fragments which were negatively charged migrate through an agarose gel matrix towards the anode. The rate of migration depends on applied voltage for the electric fields, whereby longer DNA molecules move slower through the matrix than smaller ones. All experiments were performed with 1x TAE buffer and an agarose gel with Stain Clear G (Table 2.13). Latter functioned as fluorescence dye for the visualization of DNA gel bands in UV light after electrophoresis and was applied after boiling (3 µl/ 100 ml). PeqGold Universal Agarose gel (VWR Peqlab) was utilized in concentrations of 0,5 % to 1,5 % (w/ v) in an electric field with voltages of 80 - 90 V, depending on the size of bands needed to be separated. Fragment size comparison was done with different commercial DNA markers, for accurate size assessment, containing chromatography-purified individual linear DNA fragments of known length GeneRuler$^{TM}$ DNA Ladder Mix (Thermo Fisher Scientific) and Quick-Load® 1 kb Extend DNA Ladder (New England BioLabs). High resolution agrose (BioBudget) in concentration of 3 % was used for small PCR fragments (25 - 300 bp) with GeneRuler$^{TM}$ Low Range DNA Ladder Mix (Thermo Fisher Scientific) as reference bands.

**Table 2.13:** Buffer used for agarose gel electrophoresis.

| Buffer | Components |
|---|---|
| 50 x TAE buffer stock solution | 2 M Tris<br>5,71 % (v/ v) acetic acid<br>50 mM EDTA<br>pH 8.3 – 8.5 |
| 1 x TAE buffer | 20 ml/l TAE stock solution in diH$_2$O |

For each analysis, 5 µl of the sample was mixed with 1 µl of 6x DNA loading dye (Thermo Fisher Scientific), if no PCR with DreamTaq Buffer and Polymerase was performed. The use of DreamTaq Buffer allowed direct loading of PCR product on an agarose gel, since tracking dyes and a density reagent were included. Afterwards, samples were applied on horizontally arranged agarose gels with DNA ladder. Each gel result was documented photographically. When a PCR experiment amplified multiple products of different size in one sample Ultrapure$^{TM}$ agarose (Invitrogen) was used for a preparative gel. The desired DNA bands were cut out of the agarose gel and the DNA samples were purified.

## 2.6 Purification of PCR products

Purification of PCR products was conducted to remove primers, unincorporated dNTPs, salts and other possible contaminations. The PCR products were purified using my-Budget DoublePure Kit (BioBudget) the according to the manufacturer's standard protocol with preheated elution buffer to 50 °C. PCR products from preparative gels were purified with the protocol for DNA extraction from agarose gel slices also with preheated elution buffer. A NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific) was used to measure the purity and concentration of purified samples. Afterwards, cloning or sequencing was performed.

## 2.7 Molecular cloning

After purification, PCR products were cloned with TA Cloning® Kit (Invitrogen) using the protocol provided by the manufacturer. Ligation was done at room temperature for 15 min with the pCR™ 2.1 vector and ExpressLink™ T4 DNA Ligase. Afterwards, each ligation reaction was transferred into competent *E.coli* cells (New England BioLabs) and bacterial cells were regenerated in 950 µl SOC outgrowth medium (New England Biolabs) for one hour at 37 °C (Table 2.14). The transformants were grown on LB-ampicillin (100 µg/ ml) plates with x-galactose (40 µg/ ml solubilized in Dimethylformamid) for blue-white screening at 37 °C overnight. White colonies were picked and cultured overnight at 37 °C in 4.5 ml of liquid LB medium with ampicillin (50 µg/ ml). Each transformant culture was screened for inserts by colony PCR using M13 primers (Chapter 2.4.7) and isolated by minipreparation (Miniprep) of plasmid DNA (Chapter 2.4.3). Failed cloning experiments were repeated again with double volumes of template, water and pCR™ 2.1 vector (2 µl). In case of a second unsuccessful cloning, experiments were performed with the TOPO Cloning™ Kit (Invitrogen) according to manufacturer's recommendations. *E.coli* TOP10™ being a component of the kit and functioned as competent cells.

Products from unsuccessful cloning experiments were destroyed. All other clones were deposited in the clone library with sterile glycerol at -80 °C.

**Table 2.14:** Media used for molecular cloning.

| Media | Components |
|---|---|
| LB medium | 25 g/ l lysogeny broth in diH$_2$O |
| LB agar plates | 15 g/ l agar added to LB medium |
| 1 x SOC medium (ready to use) | 2 % Vegetable Peptone<br>0.5 % Yeast Extract<br>10 mM NaCl<br>2.5 mM KCl<br>10 mM MgCl2<br>10 mM MgSO4<br>20 mM Glucose |

## 2.8 Sequencing and sequence assembly

### 2.8.1 DNA sequencing and assembly

Purified PCR products and vector DNA samples were sequenced by Eurofins Genomics (Ebersberg), each sample at least twice. For PCR products a premix sample with 2 µl of primer (working solution) was prepared. Therefore template was needed to match with the recommended concentration range list. Isolated vector DNA samples containing inserts were sequenced with standard M13 primers.

Results of DNA sequencing were quality checked with sequence reports. For colony PCR samples M13-forward and -reverse regions and vector regions were rejected. Insert nucleotide sequence and purified PCR sequence results were first searched for chloroplast sequences, with BLASTN (Altschul et al. 1990). Afterwards, sequences were aligned manually to consensus chloroplast sequences depending on experiments using the software MEGA (Tamura et al. 2011).

### 2.8.2 Next generation sequencing and sequence assembly

Chloroplast genome sequencing was completed using a Roche 454 GS FLX++ system for single reads (Roche) by Eurofins Genomics. In the sequencing of each species ¼ plates of Roche 454 were run with isolated chloroplast DNA (Table 2.15).

**Table 2.15:** Roche 454 GS FLX++ sequencing statistics.

| Species | *E. mutabilis* | *T. grandis* | *Etl. pomquetensis* |
|---|---|---|---|
| Total entries | 82.587 | 68.789 | 60.225 |
| Total length | 56.747.940 | 41.846.007 | 40.683.322 |
| Min length | 26 | 22 | 32 |
| Max. length | 1.195 | 1.152 | 1.154 |
| Mean length | 687 | 608 | 675 |
| Modal length | 880 | 803 | 875 |
| GC content | 37.8 | 37.0 | 41.1 |

For de novo DNA sequence assembly raw sequencing reads were extended into longer sequence contigs by Eurofins Genomics using Roche's 454 GS Assembler, Newbler (Table 2.16).

**Table 2.16:** Assembly statistics.

| Species | All contigs | cpContigs in total | cpContig number | Average depth | GC % |
|---|---|---|---|---|---|
| *E.mutabilis* | 585 | 1 | 1 | 37.40 | 26.70 |
| *Etl. pomquetensis* | 668 | 4 | 1 | 131.70 | 34.00 |
| | | | 2 | 129.90 | 35.53 |
| | | | 3 | 112.60 | 33.20 |
| | | | 10 | 211.30 | 48.90 |
| *T. grandis* | 368 | 9 | 1 | 140.40 | 26.48 |
| | | | 2 | 121.60 | 23.59 |
| | | | 3 | 130.70 | 27.41 |
| | | | 4 | 115.10 | 31.62 |
| | | | 6 | 152.90 | 22.23 |
| | | | 10 | 189.80 | 25.34 |
| | | | 45 | 196.80 | 23.50 |
| | | | 82 | 110.79 | 28.14 |
| | | | 171 | 206.40 | 34.62 |

## 2.9 Genome annotation

Final annotations and analyses of the chloroplast genomes were performed with Geneious Pro (version 7.1.7 or 9.1.3, Kearse et al. 2012) and further databases and online tools (Table 7.3). Finally the three genomes were deposited in GenBank.

### 2.9.1 Annotation protein-coding genes

Protein-coding genes were identified through the use of BLASTX and then manually aligned in MEGA against the nucleotide coding DNA sequences (CDSs) from other photosynthetic euglenoids and prasinophyte representatives, to determine exon-intron boundaries as well as start and stop of each gene. If necessary, equivocal exon-intron boundaries have been verified by RT-PCR experiments. A number of genes were found to contain alternative start codons. These were identified by either a lack of an ATG start codon or better correlation based on comparative alignment analysis. In all cases, a traditional methionine (ATG) start codon was preferred. Before adding to the annotation CDSs were verified by BLASTX and Emboss Sixpack, a tool for six-frame sequence translation (Rice et al. 2000). RNA secondary structure analyses of group II introns have been performed by RNA folding via Mfold web server (Zuker 2003) and optimized by eye.

### 2.9.2 Annotation ribonucleic acid

Ribosomal RNA genes were identified using RNAmmer 1.2 (Lagesen et al. 2007) and Rfam (Burge et al. 2013). If present, introns within rRNA genes have been verified by RT-PCR experiments. The number of rRNA operons and positions were confirmed using standard or long-range PCR with specific rRNA primers for each species. For the annotation of tRNAs the web server tRNAscan-SE (Schattner et al. 2005) was utilized.

### 2.9.3 Annotation Open Reading Frames and Variable Number Tandem Repeat

Open Reading Frames (ORFs) were identified with the ‘Find ORFs’ function in Geneious Pro and included in the annotation when the predicted ORFs were more than 300 nucleotides (100 AA) in length and lacking protein evidence. According to convention, ORFs were named ‘orf’ followed by the length of open reading frame in amino acid codons.

In each cpGenome the presence of a Variable Number Tandem Repeat (VNTR) area was analyzed either with REPuter (Kurtz et al. 2001), Tandem Repeats Finder (Benson 1999) or the implemented 'Find Repeats' in Geneious Pro. If a VNTR region was present than it was identified in the area between the end of the 16S rRNA gene and the next annotated gene.

## 2.10 Phylogenomic analyses

For the phylogenomic analyses two alignments were created. As taxa, each phototrophic euglenoid with sequenced cpGenome was used for the analyses, as well as 18 cpGenome sequences from selected prasinophyte and charophyte algae. Taxon sampling of prasinophyte algae based on a consensus reconstruction of green algae of Leliaert et al. (2012). If present, one or two chloroplast genomes were taken from each clade of prasinophytes and extended with two cpGenomes of charophyte species, as outgroup (Results, Chapter III, Table S1, p. 117).

Gene and protein alignments were performed from the following 84 genes: three rRNA genes: 5S, 16S, 23S, 24 tRNAs: A(UGC), C(GCA), D(GUC), E(UUC), F(GAA), G(UCC), H(GUG), I(GAU), K(UUU), L(UAA), L(UAG), M(CAU) 72bp, M(CAU) 74bp, N(GUU), P(UGG), Q(UUG), R(ACG), R(UCU), S(GCU), S(UGA), T(UGU), V(UAC), W(CCA), Y(GUA); and 57 protein- coding genes: *atp*A, *atp*B, *atp*E, *atp*F, *atp*H, *atp*I, *pet*B, *pet*G, *psa*A, *psa*B, *psa*C, *psa*I, *psa*J, *psa*M, *psb*A, *psb*B, *psb*C, *psb*D, *psb*E, *psb*F, *psb*H, *psb*I, *psb*J, *psb*K, *psb*L, *psb*N, *psb*T, *rbc*L, *rpl*2, *rpl*5, *rpl*12, *rpl*14, *rpl*16, *rpl*20, *rpl*22, *rpl*23, *rpl*32, *rpl*36, *rps*2, *rps*3, *rps*4, *rps*7, *rps*8, *rps*9, *rps*11, *rps*12, *rps*14, *rps*18, *rps*19, *rpo*B, *rpo*C1, *rpo*C2, *38ioi*, *ycf*4, *ycf*9, *ycf*12 and *chl*I. Individual genes were manually aligned in MEGA7 and only homologous sites were used in the analysis. The protein data matrix contained a total of 10,640 amino acid characters. The small fragments of tRNAs used in this analysis had a total of 1,792 nucleotide sites and the three concatenated rRNAs a total of 3,966 nucleotide sites used for phylogenetic tree reconstruction.

For Maximum Likelihood (ML) analyses each protein-coding gene was divided into a separate partition, one tRNA partition and one rRNA partition, resulting in 59 partitions. We determined the best choice of model for each partition under the Akaike Information Criteria (AIC) as recommended by Posada & Buckley (2004) using the IQ-TREE web server (Trifinopoulos et al. 2016) with the additional 'New model selection procedure'. For tRNA and rRNA genes we specified the 'Sequence type' as 'DNA'.

For the partitioned protein-coding genes the ‘Sequence type’ was specified as ‘DNA→AA’ with the ‘Genetic code 11’ for ‘Bacteria, Archaeal and Plant Plastid’. Data were analyzed for tree inference with the IQ-TREE multicore version by ML (Nguyen et al. 2015), using partitioned analysis for multi-gene alignments under the recommended models (Chernomor et al. 2016) and 1,000 ultrafast bootstrap (Minh et al. 2013).

# 3 Results

The results are presented in the following format:

Chapter I: Article 1: The Chloroplast Genome of *Euglena mutabilis* - Cluster Arrangement, Intron Analysis and Intrageneric Trends

The results are published in 'The Journal of Eukaryotic Microbiology', 2017, vol. 64 (1): 31-44. doi: 10.1111/jeu.12334.

Chapter II: Article II: Chloroplast genome expansion by intron multiplication in the basal psychrophilic euglenoid *Eutreptiella pomquetensis*

The results are published in 'PeerJ', 2017, vol. 5:e3725. doi: 10.7717/peerj.3725.

Chapter III: Manuscript: Intrageneric Variability between the Chloroplast Genomes of *Trachelomonas grandis* and *Trachelomonas volvocina* and phylogenomic analysis of phototrophic euglenoids

The manuscript has been submitted for publication to 'The Journal of Eukaryotic Microbiology' on 06. October 2017.

The Journal of Eukaryotic Microbiology

ISOP Published by the International Society of Protistologists

Journal of Eukaryotic Microbiology ISSN 1066-5234

ORIGINAL ARTICLE

# The Chloroplast Genome of *Euglena mutabilis*—Cluster Arrangement, Intron Analysis, and Intrageneric Trends

Nadja Dabbagh & Angelika Preisfeld

Bergische University Wuppertal, Faculty of Mathematics and Natural Sciences, Zoology and Didactics of Biology, Wuppertal, Germany

**Keywords**
Euglenida; introns (*mat*1/2); RNA-analysis.

**Correspondence**
N. Dabbagh, Bergische University Wuppertal,
Faculty of Mathematics and Natural
Sciences, Zoology and Didactics of Biology,
Gaussstraße 20, 42119 Wuppertal,
Germany
Telephone number: +49 202 439 2599;
FAX number: +49 202 439 2967;
e-mail: dabbagh@uni-wuppertal.de

Received: 17 December 2015; revised 6 April 2016; accepted May 25, 2016.

doi:10.1111/jeu.12334

## ABSTRACT

A comparative analysis of the chloroplast genome of *Euglena mutabilis* underlined a high diversity in the evolution of plastids in euglenids. Gene clusters in more derived Euglenales increased in complexity with only a few, but remarkable changes in the genus *Euglena*. *Euglena mutabilis* differed from other *Euglena* species in a mirror-inverted arrangement of 12 from 15 identified clusters, making it very likely that the emergence at the base of the genus *Euglena*, which has been considered a long branch artifact, is truly a probable position. This was corroborated by many similarities in gene arrangement and orientation with *Strombomonas* and *Monomorphina*, rendering the genome organization of *E. mutabilis* in certain clusters as plesiomorphic feature. By RNA analysis exact exon–intron boundaries and the type of the 77 introns identified were mostly determined unambiguously. A detailed intron study of *psb*C pointed at two important issues: First, the number of introns varied even between species, and no trend from few to many introns could be observed. Second, *mat*1 was localized in Eutreptiales exclusively in intron 1, and *mat*2 was not identified. With the emergence of Euglenaceae in most species, a new intron containing *mat*2 inserted in front of the previous intron 1 and thereby became intron 2 with *mat*1.

THE freshwater flagellate *Euglena mutabilis* Schmitz 1884 was described as a photosynthetic euglenid (Euglenida, Excavata) presenting a worm-like gliding rather than swimming behavior. It features plate-like chloroplasts with naked pyrenoids, which lie pressed against the inner face of the pellicle. The slender, nearly cylindrical cell has a size of 70–122 μm length and 4–12 μm width with a dominant eyespot, several small paramylon grains, no mucocysts (Ciugulea and Triemer 2010; Gojdics 1953; Kim et al. 2015) and shows an ability to survive in extremely acidic environments (Casiot et al. 2004). Movement occurs exclusively by euglenoid contractions, since no emerging flagellum is present (Häder and Melkonian 1983). *Euglena mutabilis* is considered to be a key taxon in the early emergence of the genus *Euglena*. The independent long branch always observed in phylogenetic analyses suggests that *E. mutabilis* is a derived species with a so far uncertain early branching position in the genus *Euglena* (Kim et al. 2010, 2015; Linton et al. 2010). Furthermore, the adaptation on extreme environments might affect mutation ratios and spread of introns. After the very recently published chloroplast genomes (cpGenomes) by Bennett and Triemer (2015), this is the sixth annotated cpGenome of the genus *Euglena*, including *E. longa* (Gockel and Hachtel 2000), with a rising possibility to gain further insight into underlying intrageneric evolutionary patterns. Particularly relevant in this investigation are the genome sizes and intron numbers and distribution patterns, because so far the genus *Euglena* comprised small and large genomes with little and many introns, respectively (Bennett and Triemer 2015; Bennett et al. 2012; Hallick et al. 1993).

The first published euglenid cpGenome of *Euglena gracilis* Klebs 1883 (Hallick et al. 1993) displays a surprisingly large genome of 143,170 base pairs (bp) due to an enormous number of introns and other noncoding DNA. The newly sequenced *E. gracilis* var. *bacillaris* Klebs Pringsheim 1956 (Bennett and Triemer 2015) is not fully closed, but ranged in the same area with at least 132,034 bp and also contained all the genes identified by Hallick et al. (1993). Beside some minor differences in the ribosomal operon and the number of open reading frame (ORFs) and introns, *E. gracilis* var. *bacillaris* was highly similar to *E. gracilis* (Bennett and Triemer 2015). The other two

**Table 1.** CpGenome features of Euglenids and depicted prasinophytes

| According to GenBank | Size (bp) | A+T % | Genes | Introns[a] | ORFs | roaA | CDS of rpoA (bp) | Largest gene (bp) | Shortest gene (bp) | Gene with most introns | psbD/psbC overlap | petB I1 bp/5′start |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *E. gracilis* strain *Z* | 143,171 | 73.9 | 116[b] | 134 | 7 | + | 651 | psbC (10,861) | psaM (96) | rpoC1 (11) | + | 909/GUGCG |
| *E. gracilis* var. *bacillaris* | 132,034* | 74.2 | 104 | 134 | 6 | + | 525 | psbC (11,446) | psaM (96) | rpoC1 (11) | + | 909/GUGCG |
| *E. viridis* | 76,156 | 73.8 | 92 | 77 | 0 | + | 483 | psbC (6165) | *psa*M (96) | rpoB (10) | + | 528/GUGUG |
| *E. viridis epitype* | 91,616 | 73.6 | 92 | 76 | 13 | + | 480 | psbC (6131) | psaM (96) | rpoC1 (8) | + | 533/GUGUG |
| *E. mutabilis* | 86,975 | 73.3 | 91 | 77[e] | 0 | + | 636 | psbC (12,192) | psaM (96) | rpoC1 (9) | + | 435/GUGCG |
| *Era. anabaena* | 88,487 | 72 | 93 | 82 | 4 | + | 624 | psbA (6998) | psaM (96) | rpoB (10) | + | 551/GUGCG |
| *M. parapyrum* | 80,147 | 72 | 93 | 80 | 1 | + | 507 | psbC (6127) | psaM (96) | rpoB/-C1 (9) | + | 441/GUGCG |
| *M. aenigmatica* | 74,746 | 70.6 | 93[c] | 53 | 1 | + | – | psbC (6229) | psaM (96) | rpoC1 (9) | + | 544/GUGCG |
| *Cr. skujae* | 106,843 | 73.7 | 93 | 84 | 4 | + | 489 | psbC (8947) | psaM (96) | rpoB (10) | + | 537/GUGUG |
| *S. acuminata* | 144,166* | 73.4 | 95[c] | 112[d] | 0 | + | 486 | psaB (11,283) | psbT (90) | rpoB/rbcL (9) | + | 510/GUGCG |
| *T. volvocina* | 85,392* | 72.7 | 93 | 94 | 1 | + | 543 | psbC (7698) | psaM (96) | rpoC1 (11) | + | 536/GUGCG |
| *C. vesiculosum* | 128,892* | 73.9 | 92[c] | 128 | 6 | + | – | psbC (11,567) | psbI (105) | rpoC1 (11) | + | 470/GUGUG |
| *Efs. proxima* | 94,185* | 73.1 | 91 | 113 | 2 | + | – | psbC[f] (6648) | psaM (96) | rpoB (10) | [h] | 478/GUGCG |
| *Et. viridis* | 65,523* | 71.4 | 83 | 24 | 3 | – | 672 | psbC (5706) | psbT (96) | rpoB (5) | – | – |
| *Etl. gymnastica* | 67,623 | 65.7 | 89 | 7 | 4° | – | 846 | psbC | psaM/psbT (96) | psbC (2) | +[g] | – |
| *Pyramomonas parkae* | 101,605 | 65.3 | 125 | 1 | 5 | – | 1,056 | atpB (4224) | psbT (96) | atpB (1) | + | – |
| *Pyconococcus provasolii* | 80,211 | 60.5 | 98 | 1 | 3 | – | 1,074 | ftsH (7944) | psbT (96) | atpB (1) | + | – |
| *Ostreococcus tauri* | 71,666 | 60.1 | 92 | 1 | 2 | – | 1,071 | atpB (5188) | psaM/psbT (96) | atpB (1) | + | – |

*= Chloroplast circle not closed; ° = ORFs without mat1, +/– = exist/does not exist.

[a] Twintrons were counted as single insertion sites.

[b] Including rRNA repeats and intermediate tRNAs.

[c] Includes the identified 5S, for *S. acuminata* the two identified 5S.

[d] Includes the two introns in *rps*18.

[e] Includes the one intercistronic intron *rps*4-*rps*11.

[f] fFirst exon could not be identified so gene length is a minimum.

[g] Realigned, but with alternative start codon.

[h] Start codon not determined, due to undetermined exon1.

*Euglena* strains published were *Euglena viridis* Ehrenberg 1830 (91,606 bp, epitype) by Bennett et al. (2012) and the recently sequenced *E. viridis* strain SAG 1224-17d Pringsheim 1941 by Bennett and Triemer (2015), which differed mostly due to an ORF-rich region that was present in the epitype, but not in the *E. viridis* SAG 1224-17d cpGenome. A comparison of *Monomorphina aenigmatica* (Drezepolski) Nudelman et Triemer emend. Kosmala et Zakrys 2007 (74,746 bp, Pombert et al. 2012) and *Monomorphina parapyrum* Kim, Triemer and Shin 2013 (80,147 bp, Bennett and Triemer 2015) on the intrageneric level disclosed a high similarity in genome characters with only a few exceptions in regard to *mat*5 and major differences in the number of introns (Bennett and Triemer 2015). With the other four species *Cryptoglena skujae* (Skuja) Marin & Melkonian 2003, *Euglenaria anabaena* (Mainx) Karnkowska & Linton 2010, *Trachelomonas volvocina* (Ehrenberg) Ehrenberg 1834 (Bennett and Triemer 2015) and *Euglenaformis proxima* (Dangeard) Bennett and Triemer 2014 (Bennett et al. 2014) added to previously analyzed cpGenomes (Table 1), all lineages in the Euglenaceae have been covered and allowed for a diagrammatic phylogeny revealing first evolutionary trends (Bennett and Triemer 2015). The cpGenome of *Phacus orbicularis* Hübner 1886 was sequenced very recently (Kasiborski et al. 2016), but is not included in this analysis. All cpGenomes present a large amount of cluster similarity and many highly conserved genes. In contrast to the prasinophyte *Pyramimonas parkeae* Norris and Pearson 1975 (Turmel et al. 2009), which is considered to be the closest relative to the chloroplasts of euglenids and has only one intron, phototrophic euglenids show an accumulation of introns.

The largest of the previously published euglenid cpGenomes are those of the two *E. gracilis* strains and of *Strombonomas acuminata* (Schmarda) Deflandre 1930 (Bennett and Triemer 2015; Hallick et al. 1993; Wiegert et al. 2013). All three have more than 132 kb and more than 100 introns, which interrupt virtually all protein-coding genes. Since the number and nature of introns in all cpGenomes of phototrophic euglenids differs significantly, we aimed at a solid foundation for further intron studies in which questions like the origin of introns or intron dispersal could be clarified. Considering *P. parkeae* (Turmel et al. 2009) with only one intron and the two early-diverging euglenids *Eutreptiella gymnastica* (Hrdá et al. 2012; Pombert et al. 2012) and *Eutreptia viridis* (Wiegert et al. 2012) with 7 and 23 introns, respectively, it seemed that the number of introns rose with increasing divergence. In that case, the number of introns at least within the Euglenaceae should nearly be the same or rising more or less steadily in one lineage, which was unfortunately not the case. Complicating the matter was the fact that introns in euglenids are extremely difficult to identify and to characterize, because they display highly unusual traits in intron structure and often show degenerated splicing site sequences (Wiegert et al. 2012), all of which are prerequisites for studying insertion sites and evolution properly.

Most of the introns described in euglenid plastid genomes until now belong to group II introns. These are relatively abundant in organellar genomes of plants and lower eukaryotes and in prokaryotic mRNA and tRNA. Many introns of *Euglena* can only be described as group II-like, because they lack convincing group II core structures, due to either loss of domains I–IV or massive divergence from related group II introns. Thus, euglenid group II introns tend to be significantly shorter than other group II introns, their range is 298–618 nt, with a mean of 463 ± 90, compared for example to group II introns of the common liverwort *Marchantia polymorpha* (Archaeplastida) with a mean of 577 ± 119 (Michel et al. 1989). But they still contain the conserved 5′-boundary motif – GUGYG. Most group III introns range between 73 and 120 bp, have an average size of 102 bp and consist of a consensus boundary sequence of 5′-NUNNG. They generally are A+U-rich, appear to be abbreviated versions of group II introns lacking the domains d2–d5 and are unique to chloroplasts (Candales et al. 2011; Christopher and Hallick 1989; Copertino and Hallick 1993; Dai et al. 2003; Doetsch et al. 2001; Hong and Hallick 1994; Khan and Archibald 2008; Michel and Ferat 1995; Sheveleva and Hallick 2004; Zimmerly et al. 2001). Further a high number of different twintrons hampered the correct identification of the investigated noncoding structures: group II twintrons, mixed group II/group III twintrons and group III twintrons. In complex twintrons, even multiple internal introns can be observed (Copertino and Hallick 1993; Copertino et al. 1992, 1994; Doetsch et al. 1998; Drager and Hallick 1993).

In this study, beside the genome sequencing, cluster analysis and annotation, we investigated the position, distribution and length of each intron observed in *E. mutabilis*. A thorough RT PCR analysis was performed to validate exact coding DNA sequences (CDSs) and to substantiate presumed exon–intron boundaries. To avoid contamination with intron-rich mitochondrial and genomic DNA, and to allow for complete annotation of the circle, we isolated the chloroplasts of *E. mutabilis* beforehand by ultrasonic waves and density gradient centrifugation using Percoll® (GE Healthcare, Solingen, Germany). DNA from purified chloroplasts was then sequenced by 454 Roche.

## MATERIALS AND METHODS

### Preparation and isolation of chloroplast DNA

*Euglena mutabilis* strain SAG 1224-9b (EPSAG, Germany) was grown in *Euglena* medium modified after Cramer and Myers (1952) at 20–23 °C under 12:12 light: dark cycle using fluorescent tubes delivering about 30 μmol photons/$m^2/s^2$ of light. After 7–9 d, cells were harvested with approx. 1.56 × $10^6$ cells/ml. Cells were concentrated from 200 ml culture suspension by centrifugation in a swing-out rotor at 2,205 *g* for 5 min at 4 °C and rinsed three times with isolation buffer (0.3 M Sorbitol, 5 mM $MgCl_2$, 5 mM EDTA, 20 mM Hepes/KOH pH 8.0, 10 mM $NaHCO_3$) modified after Aronsson and Jarvis (2002). The pellet was resuspended in fresh isolation buffer and 50 μM Pefabloc® SC to avoid protein destruction. To reduce contamination, cells were loaded on a three-step Percoll®

gradient, with a bottom layer 95%, middle layer 60% and top layer 30%, in a 50-ml falcon tubes: 10 ml bottom layer with 9.47 ml Percoll® solution (95% (w/v) Percoll®, 3% (w/v) PEG 6000, 1% (w/v) Ficoll, 1% (w/v) BSA) and 0.53 ml gradient mixture (25 mM Hepes-NaOH, pH 8.0, 10 mM EDTA, 5% (w/v) sorbitol); 10 ml middle layer with 6.32 ml Percoll® solution and 3.68 ml gradient mixture; 10 ml top layer with 3.15 ml Percoll® solution and 6.85 ml gradient mixture (modified after Aronsson and Jarvis 2002). Gradients were centrifuged in a swing-out rotor at 2,000 *g* for 20 min brake off. Purified cells accumulated at the 30%/60% Percoll® interface were sampled and washed three times in isolation buffer. The pellet was resuspended in 5 ml isolation buffer with proteinase inhibitor. Subsequently, cells were disrupted by ultrasonic probe three times for 3 s with the amplitude set at 60% with a 0.1 s pulse rate (Bandelin Sonopuls HD 60, Berlin, Germany) and intermediate washing steps on ice. The samples were combined in falcon tubes and centrifuged in a swing-out rotor at 259 *g* for 1 min. The supernatant was collected and the procedure repeated twice with the remaining pellet in 2 ml fresh isolation buffer. The sequence of disruption by ultrasonic waves and chloroplast elution was repeated five times.

The combined supernatants were centrifuged in a swing-out rotor for 5 min at 3,645 *g* to pelletize isolated chloroplasts. Chloroplasts were resuspended in fresh isolation buffer and loaded on a primed five step Percoll® gradient with 80%, 60%, 40%, 30%, and 10% Percoll® (v/v) in gradient mixture. Tubes were centrifuged in a swing-out rotor for 30 min at 2,000 *g* brake off. The chloroplast fraction was recovered from the 30% layer, resuspended in isolation buffer and centrifuged at 3,645 *g* for 5 min. Washing steps were repeated twice to eliminate Percoll®. Three final washing steps were performed to ensure purity of chloroplasts for 1 min at 259 *g*. Purification of chloroplasts was completed by pelleting for 5 min at 3,976 *g* and integrity of chloroplasts was checked microscopically. cpDNA was isolated by BioBudget Kit (Bio-Budget Technologies GmbH, Krefeld, Germany) following the manufacturer's protocol. Extracted DNA was measured with a NanoDrop 2000 Spectrophotometer (Thermo Scientific, Dreieich, Germany) for concentration and purity. The total amount of DNA used for 454 sequencing was 73.8 ng/μl from 200 ml cell suspension with $1.56 \times 10^6$ cells/ml.

## Sequencing, assembly and annotation of the plastid genome

Purified plastid DNA was sequenced by Roche 454 using the GS FLX++ chemistry Rapid Shotgun Library Preparation Method technology by Eurofins Genomics (Ebersberg, Germany). In total, 82,587 reads were produced in ¼ segment of a full run with an average size of 687 bases. The de novo assembly of reads performed by Eurofins Genomics in Newbler (Roche Diagnostics, Basel, Switzerland) resulted in 585 contigs. To search for chloroplast sequences, a Blastn homology search was performed with all contigs (Altschul et al. 1990). Contamination of mitochondrial and whole genomic DNA was reduced by Percoll density gradient, but bacterial DNA was still found in the other contigs. The first single large contig contained the almost complete cpGenome of *E. mutabilis*. The two ends of the single obtained contig were linked by fill-in PCR using whole genomic DNA with the following setting: 95 °C 3 min, 30 cycles (9 °C 1 min, 5 °C 1 min, 72 °C 6 min), 72 °C 10 min, 4 °C hold on a gradient cycler (Thermo Scientific). Primers were created manually by Primer3Plus (Untergasser et al. 2012) based on the nucleotide sequence close to the two ends of the contig: forward 5′-CAGTCCTTTGCCTTACCACT-3′ and reverse 5′-CCTTTTCTTTCCCTCTCTCTTC-3′. The PCR product was seized on 1% agarose gel for identification and subsequently cloned for storage and repeating experiments and to avoid false annotation due to primer sequences with TA Cloning® Kit (Invitrogen GmbH, Darmstadt, Germany) using the protocol provided by the manufacturer. The clones grown on LB- ampicillin (100 μg/ml) plates with x-galactose (40 mg/ml) at 37 °C overnight were again cultured overnight at 37 °C in 4.5 ml of liquid LB medium + ampicillin (50 μg/ml). Selected clones were screened for inserts using M13 primers flanking the insertion site. The plasmids were purified with the E.Z.N.A.® Plasmid Mini Kit I (OMEGA Bio-Tek, Norcross, GA) using the manufacturer's protocol. The product was then Sanger-sequenced with M13 primers flanking the insertion site by Eurofins Genomics (Eurofins Genomics, Ebersberg, Germany). The contig and the obtained linking sequence were aligned manually in MEGA 5 (Tamura et al. 2011). The genome circle was closed by additional 259 bp.

The final annotation of the completed chloroplast sequence was performed with Geneious® 7 Pro (version 7.1.7 http://www.geneious.com, Kearse et al. 2012). The cpGenome was searched for tRNAs via tRNAscan-SE (Schattner et al. 2005) at default mode, the source chosen as mixed (general tRNA model) and included in the annotation. Protein-coding genes were identified by 15 conspicuous conserved gene clusters which have been ascertained in all chloroplast genomes of phototrophic euglenids and partly in *Pyramimonas parkae* (NC_012099.1) and two other chlorophytes *Pycnococcus provasolii* (FJ493498.1) and *Ostreococcus tauri* (NC_008289.1). All protein-coding gene sequences were extracted from the genome and manually aligned against the nucleotide CDS of euglenids and green algae using MEGA 5 to detect exon–intron boundaries as well as start and stop of each gene. A traditional methionine (ATG) start codon as well as one of the three possible stop codons (TAA, TAG, TGA) was always preferred for annotations.

### *RT PCR analysis of introns*

To ensure that the exon–intron boundaries were aligned correctly and that the insertions are genuine introns, all introns identified in the protein-coding genes have been verified by RT PCR experiments using primers specific to *E. mutabilis* (Table S1). This seemed necessary, because the protein-coding CDSs for some genes were not well conserved and thus hard to align unambiguously. RNA was isolated from *E. mutabilis* by *my*-Budget RNA Mini Kit (Bio-Budget Technologies GmbH) and a subsequent RT PCR was performed by One-Step RT PCR Kit (Qiagen GmbH, Hilden, Germany) following the manufacturer's protocol with a 25 μl PCR approach. The resulting products were cloned and sequenced as described above. The CDS was finally added to the annotation, after scrutinizing the CDS of protein-coding genes by Blastx and Emboss Sixpack Sequence translation (EMBL-EBI 2015). The introns within protein-coding genes were grouped into group II and group III according to the classifications of Candales et al. (2011), Christopher and Hallick (1989), Copertino and Hallick (1993), Dai et al. (2003), Doetsch et al. (2001), Hong and Hallick (1994), Khan and Archibald (2008), Michel and Ferat (1995), Sheveleva and Hallick (2004) and Zimmerly et al. (2001). Pairwise sequence comparison of the amino acid sequence of *E. gracilis* with a putative amino acid sequence of *E. mutabilis* was performed using Exonerate 2.4.0 by Slater and Birney (2005) (https://github.com/nathanweeks/exonerate) and its extension (http://www.ebi.ac.uk/~guy/exonerate/server.html) to reveal intron boundaries and start/stop of the searched genes.

### *Cluster analysis*

A Mauve analysis with all euglenid cpGenomes available (except *Euglena longa* because of its reduced genome) was exerted with the progressiveMauve algorithm of the Geneious plugin (Darling Aaron et al. 2004), compared to the same manually recognized 15 conserved gene clusters mentioned above and used to identify the 16S, 23S, and 5S rRNA genes. The exact start/stop area of 16S and 23S was determined using RNAmmer 1.2 server (Lagesen et al. 2007), set on Bacteria as selected kingdom of input sequences. The 5S rRNA start/stop region was identified by submitting the region of the cpGenome between 23SrRNA and *psa*I to Rfam (Burge et al. 2013).

In order to identify the total number of ribosomal operons present in the cpGenome, primers were designed with Primer3Plus (forward 5′-GTACTGGAAGGTGCGGCT-3′, reverse 5′-GGTACGCTCTAACCAACTGAG-3′), which cover 86 bp at the end of the 16S rRNA and the start of the tRNA Ile (GAU). In case of a single ribosomal operon, a small product of 86 bp should be identified, whereas more than 5,000 bp large PCR products should be recognized in case of multiple copies. To ensure amplification of large sequences, a long-range PCR was performed with the Q5® High-Fidelity DNA Polymerase (New England Biolabs GmbH, Frankfurt am Main, Germany), following the manufacturer's protocol for amplicons up to 10 kb. ORFs were detected using the software ORF finder (http://www.ncbi.nlm.nih.gov/projects/gorf/) as described in Bennett et al. (2014) and integrated in the final annotation, when ORFs did not overlap with identified genes or lacked Blast evidence for being a previously identified protein-coding gene.

### *Final annotation of the genome*

The variable number of tandem repeats was scanned with REPuter (Kurtz et al. 2001). The minimal repeat size was set at 15 and the Hamming Distance at 1 and 2, respectively. Repeats were searched for in forward direction (Bennett and Triemer 2015; Gao et al. 2011; Lemieux et al. 2007) on the extracted area between the 16S gene and the t-RNA-His (GUG). The resulting circle was drawn with Genome Vx (Conant and Wolfe 2008). After completing the total annotation, the whole cpGenome was oriented like the majority of the other published cpGenomes of phototrophic euglenids with rRNA genes lying on the reverse strand.

To ensure that there was no misassemble of the genome and that the cpGenome of *E. mutabilis* in fact is a mirror image of all of the other cpGenomes of *Euglena,* two PCR analyses were performed with primers linking 16S and *psb*B as well as *rpo*B and *chl*I (Table S1). The cpGenome was deposited in GenBank under accession number KT223519.

## RESULTS

One aim of the study was to isolate chloroplasts from *E. mutabilis* prior to DNA isolation in order to avoid contamination with mitochondrial and whole genomic DNA and thereby facilitating sequencing procedure and minimizing assembly errors. Purity of isolated chloroplast DNA was ascertained by NanoDrop.

### General gene analyses

The following sequencing of the isolated chloroplast DNA and subsequent de novo assembly yielded only one single chloroplast contig with a total of 5,360 reads and an average coverage depth of 37.40 reads. Fill-in PCR spanning nucleotide sequences at each end of the contig allowed to complete the circular genome with only 259 nts to be added, resulting in a total length of 86,975 bp and an overall A + T content of 73.3% (Fig. 1). A total of 91 genes were identified including 61 protein-coding genes (including also *ycf*4, *ycf*13, and *ycf*65), 27 tRNAs and three rRNAs. No *mat*5 gene was identified in *E. mutabilis*, supporting the hypothesis by Bennett and Triemer (2015) that the gene was lost in the *Euglena* lineage. Variable number tandem repeat (VNTR) sequences have not been identified. Only when two mismatches were allowed in REPuter two discontinuous repeats of 18 bp appeared from 86,897 to 86,914 and from 86,915 to 86,932. Two repeats of 15 bp were identified with one mismatch at positions 72–86 and 86,713–86,727. Since the repeats were separated by 52 bp, the areas were not considered as VNTR (Harding et al. 1992; Nakamura et al. 1987).

### Protein-coding genes, introns, and their classification

Of the 61 protein-coding genes, only 24 genes were without any intron. Nineteen genes contained one and 18 genes two or more introns, resulting in a total of 76 introns within protein-coding genes (77 introns in total including one intercistronic intron *rps*4-*rps*11; likely twintrons measured as one insertion site, Table 1). Since exact localization of boundaries was sometimes infeasible by merely aligning the sequences and to avoid imprecise position and number of introns, RNA

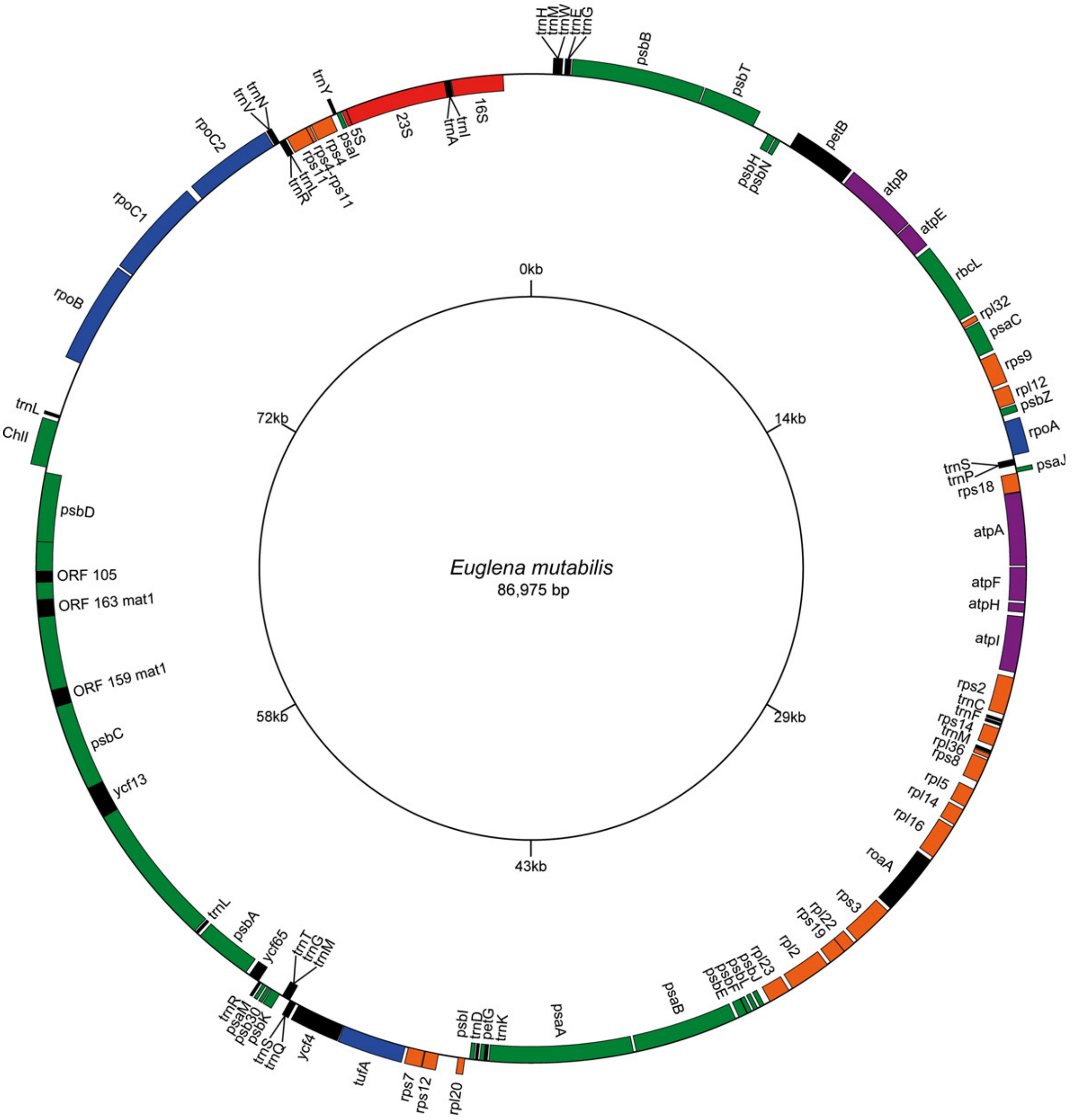


**Figure 1** Circular gene map of the *Euglena mutabilis* chloroplast genome. Boxes of different colors represent genes of similar functional groups: red = ribosomal rRNAs; green = photosystem/photosynthesis genes; orange = ribosomal proteins (*rpl*, *rps*); violet = *atp* genes; blue: transcription/translation-related genes (*rpo*, *tuf*A); black = conserved hypothetical proteins (*ycf*), open reading frames (ORF), tRNAs. Boxes are proportional to their sequence length. Outer ring: Genes on the outside of the circle are considered on the positive strand, genes inside the circle on the negative strand.

for the 37 protein-coding genes with one or more introns was scrutinized by RT PCR. Sixty-eight of the 76 introns could be confirmed and several starting points have been disambiguated by the acquired CDSs, which were further used to define exact boundaries and localization. The only CDS available from Genbank was for *psb*K (AF241280.1). For five genes (*rps*19, *rpl*2, *chl*I, *rpo*A, and *rpl*20), it was not possible to produce a reliable cDNA. Consequently, the exon–intron boundaries of *rps*19, *rpl*2, and *chl*I were set by alignment of genomic DNA only. For *rpo*A, no exact exon–intron boundaries and no start/stop region were identified. Hence, and because the different CDSs were highly variable and differed in lengths (from *E. viridis* epitype CDS 480 bp to *E. gymnastica* CDS 846 bp, see Table 1), a convincing alignment failed. To detect the exact start/stop of the gene and of alleged introns, further examinations were performed by exonerate 2.4.0 by comparing *rpo*A of *E. mutabilis* with the amino acid sequence of *E. gracilis*. In contrast to *E. gracilis* strain Z (one intron) and *E. gracilis* var. *bacillaris* (two introns), the gene in *E. mutabilis* contained three introns and a resulting CDS of 636 bp. For *rpl*20 exonerate or a manual alignment did not yield clear results, notwithstanding thorough analysis of the sequence. The protein-coding gene *rpo*B of *E. mutabilis* was intensely examined by RT PCR with three different primer pairs comprising the entire range from the tRNA-Leu (CAA) to the protein-coding gene *rpo*C1, unveiling one intron. Despite an extensive search from the beginning of the tRNA-Leu (CAA) to the only possible ATG-start codon, no other ATG-start codon was identified farther upstream. All ATG-start alternatives also failed the verification with Emboss Six-pack Sequence translation (EMBL-EBI 2015). Additionally compounded by the significance of *rpo*B for the RNA polymerase beta subunit, an alternative start has been avoided, so that *rpo*B is considered a very short, but complete gene in *E. mutabilis*.

For the protein-coding gene *rps*11, the RT PCR forward primer was located in the protein-coding gene *rps*4 to detect every possible intron of the region between *rps*4 and tRNA-Leu (UAG). The results of RT PCR showed three introns in the examined area, although the coding region of the *rps*11 gene is only interrupted by two introns of 110 and 108 bp (positions 67 and 196). The other 109 bp long intron was identified by alignment as an intron in the rps4-rps11 intercistronic region.

General gene size ranged from 12,192 bp in *psb*C (including six introns with a total length of 10,770 bp) to 96 bp in *psa*M (Table 1). The gene with the highest intron number found in the cpGenome of *E. mutabilis* was *rpo*C1 with nine introns. Traditional methionine start codons were not found for the three genes *rpo*A, *rps*18, and *rps*11, so alternative start codons were accepted (Table S2). The last exon of *psb*D overlapped with the first exon of *psb*C, like in almost all other published phototrophic euglenids and prasinophytes (Table 1).

## Intron classification

*Group II introns.* In the 19 protein-coding genes with one intron, we identified nine group II introns with typical group II 5′-start regions (GUGYG), appropriate sizes and an intron encoded ORF (Table S1). Among these, we detected two group II twintrons. One of which was the same complex group II twintron as the one found in *E. gracilis* strain Z (*psb*T), whereas another one was identified as group II twintron in *ycf*4 (800 bp) with a second GUGCG and two ORFs, where *E. gracilis* strain Z only has a group II intron (Hallick et al. 1993; Thompson et al. 1995).

In the 18 protein-coding genes with more than one intron, 15 of the 57 introns, with a typical 5′-GUGYG- start and one intronic ORF have been detected and analyzed (Table S2). Eleven of these introns ranged in the same size as other euglenid group II introns. Interestingly, four group II introns were significantly larger than typical *Euglena* group II introns (marked dark gray in Table S2). Whether these introns are group II twintrons or mixed group II/group III twintrons can only be ascertained by secondary structural analysis, because differences in intron features in *E. gracilis* did not allow for safe assertions (Hallick et al. 1993).

*Group III introns and uncertain introns/twintrons.* Thirty-three group III introns have been undoubtedly identified by RNA-analysis in the cpGenome of *E. mutabilis*. They ranged from 91 bp (I8 of *rpo*C1) to 119 bp (I4 of *rpo*C1, Table S2). Eleven introns were larger than typical group III introns, but still displayed the typical group III 5′-boundaries. However, group II and III introns can have very similar and sometimes equal sequences at the 5′-splice site and are thus difficult to distinguish (Doetsch et al. 1998). Furthermore, group III introns contain a 3′-end stem–loop motif that is functionally analogous to group II intron domain VI. We compared these larger introns with the ones of *E. gracilis* strain Z and assumed that intron 2 of the gene *rps*18 in *E. mutabilis* is indeed a complex group III intron, and intron 1 of *rps*3 is a mixed twintron that shares identical insertion sites with *E. gracilis* (Copertino and Hallick 1993; Copertino et al. 1991; Hallick et al. 1993).

For the other nine large group III introns (Table S2, marked gray), it still remains to be seen, whether these introns are group III twintrons, complex group III twintrons, or just large group II introns with an untypical 5′-boundary. The latter can be assumed for intron 2 of *psb*D and intron 6 of *psb*C in *E. mutabilis* because of a typical ORF often found in group II introns. The group II introns of *psb*D (I2) and *psb*C (I6) in *E. gracilis* strain Z were inserted at different sites. The remaining six introns of protein-coding genes (Table S2 noted with X) did not possess unique conserved boundaries of group II or group III introns and therefore have not been classified so far.

## Ribosomal operons

The long-range PCR approach to measure the copies of the RNA operon with primers flanking both, the 16S rRNA and the start of the tRNA Ile (GAU), yielded only one product of 86 bp, which was more in line with both *E. viridis* cpGenomes than with both *E. gracilis* genomes (Bennett and Triemer 2015; Bennett et al. 2012; Hallick et al. 1993).

## Open reading frames

Three ORFs were identified in the cpGenome, which were named according to the length of the coding sequence in amino acid residues as ORF 105, ORF 163, and ORF 159. All three occurred in the first intron of *psb*C. A Blastp analysis was performed against the NCBI nonredundant protein sequences (nr) database to determine whether any of the ORFs have functional similarity to previously sequenced genes. For ORF 105, nonsignificant matches returned. For ORF 163, the Blastp analysis returned moderate matches to maturase [*E. viridis*] (*e*-value 1e-08), putative reverse transcriptase/maturase [*M. aenigmatica*] (*e*-value 4e-04) and maturases of many other organisms with relatively low *e*-values. For ORF 159, the Blastp search returned matches to putative reverse transcriptase in many different phototrophic euglenids and type II intron maturase. The ORFs can possibly be considered as remnants of the once mobile elements and have been ascertained as *mat*2. Results have been included in the GenBank annotation.

## Re-analysis of published cpGenomic data

In a re-analysis of *M. aenigmatica* (NC_020018.1, Pombert et al. 2012), we identified the hitherto undetected *ycf*12 (gene synonym *psb*30) in cluster 2, where it was located in all other published cpGenomes of phototropic euglenids with the exception of *Et. viridis* (Wiegert et al. 2012) and *Colacium vesiculosum* (Wiegert et al. 2013). One Intron (12,696–12,595) inside the *ycf*12 (12,706–12,503) was revealed and the CDS validated with Blastx and Emboss Sixpack.

To discover the still missing 5S rRNA start/stop region for *M. aenigmatica* (NC_020018.1, Pombert et al. 2012), *S. acuminata* (JN674637), and *C. vesiculosum* (JN674636, Wiegert et al. 2013), the regions next to the 23S rRNA were submitted to Rfam (Burge et al. 2013). For *M. aenigmatica* (Pombert et al. 2012), the 5S rRNA was detected between the 23S rRNA and the tRNA-Ser at position from 69,307 to 69,425. In *C. vesiculosum*, it reached from base 4,639–4,757 and in *S. acuminata* (Wiegert et al. 2013) it was discovered two times on the genome. The first gene was detected next to the clockwise 23S rRNA at position 478–595 and the second with exactly the same nucleotide sequence between the tRNA-Ser and the counterclockwise 23S rRNA at the position 139,252–139,369.

Furthermore, it appeared doubtful, whether the protein-coding gene *rps*18 of *S. acuminata* (JN674637, Wiegert et al. 2013) was completely annotated, because re-analysis pointed to differing results: a MAFFT alignment with all other phototrophic euglenids uncovered an rps18 protein-coding gene (84,580–83,936) with a methionine (ATG) start codon, two introns (I1: 84,544–84,439; I2: 84,310–83,991) and a larger CDS (219 instead of 138 bp) for *S. acuminata*. The newly identified CDS is more in line with the protein-coding gene of all other euglenids and *P. parkeae* (Turmel et al. 2009) with CDSs between 186 bp in *M. aenigmatica* (Pombert et al. 2012) and 225 bp in *Efs. proxima* (Bennett et al. 2014). Validation with Blastx and Emboss Sixpack Sequence translation (EMBL-EBI 2015) substantiated the assumption that the protein-coding gene of *S. acuminata* was not completely annotated.

During intron analyses for both *E. viridis* strains (Bennett and Triemer 2015; Bennett et al. 2012), a slight annotation oversight concerning the position and start of *psb*K I2 led the authors to a small misinterpretation: The annotated –TT (*E. viridis* epitype) or – AT (*E. viridis*) endings of exon1 in *psb*K needed to be transferred to the start of exon2. By doing so, the untypical 5′-GAGAA start of intron 2 from *psb*K will become a typical group III 5′-TTGAG for the *E. viridis* epitype and 5′-ATGAG for *E. viridis* SAG 224-17d with a retained U at the second position and a G at the fifth position. The same oversight prevented recognition of a conserved 5′-boundary motif of group II (GUGYG) or group III (NUNNG) introns for *C. skujae* intron 1 in *psb*T and intron 2 in *atp*I, for *E. viridis* epitype intron 2 of *rpl*14 and intron 2 of *atp*I and for *E. viridis* SAG 1224-17d of *roa*A I1 (Bennett and Triemer 2015; Bennett et al. 2012).

## DISCUSSION

The AT content is remarkably similar to the genomes of the Euglenacea (Table 1). The genome size, however, differs significantly from the *E. gracilis* genomes and is comparable with the cpGenome size of both *E. viridis* species (Bennett and Triemer 2015; Bennett et al. 2012). The latter and the *E. mutabilis* cpGenome contained only one copy of the ribosomal operon, whereas three copies of a tandemly repeated ribosomal RNA operon, a fourth partial operon encoding a complete 16S rRNA gene and a pseudo-5S rRNA gene were found in the cpGenome of *E. gracilis* (Hallick et al. 1993), as well as three complete copies in *E. gracilis* var. *bacillaris* (Bennett and Triemer 2015). The identified single operon together with similar genome lengths and almost identical numbers of introns support the assumption that although diverging in different subclades in multiple gene analyses (Kim et al. 2015), *E. viridis* and *E. mutabilis* are more alike to each other than to *E. gracilis* (Bennett and Triemer 2015; Bennett et al. 2012; Hallick et al. 1993).

The number of genes present and the gene content of *E. mutabilis* were also similar to other euglenid cpGenomes. *Euglena mutabilis* possessed three alternative start codons for the genes *rps*11, *rpo*A, and *rps*18. For *rpo*A, it had the same alternative start codon (TTG) as the two *E. viridis* species and *E. gymnastica* (Bennett and Triemer 2015; Bennett et al. 2012; Hrdá et al. 2012). For *rps*18, the alternative start codon GAA differed from the start codon ATC in *Strombomonas acuminata* rps18 (Wiegert et al. 2013) and for *rps*11 the start codon ATA differed from the start codon ATT in *Cr. skujae rps*11. Taken together, genome length, structure and introns of *E. mutabilis* showed features linking it closely to both *E. viridis* species and *M. aenigmatica* (Bennett and Triemer 2015; Bennett et al. 2012; Pombert et al. 2012).

The gene *rpl*20 was significantly shorter in *E. mutabilis* than in all other euglenids. A possible explanation might be that a still unidentified intron is situated within the gene and that another unidentified stop codon is located downstream on the genome, so that the true length was not determined. We did not consider *rpl*20 to be a pseudogene, because its product constitutes a structural element of the ribosome and it did not occur at other sites in the genome.

The intercistronic *rps*11-*rps*4 intron of *E. mutabilis* was found at the same position as in *E. gracilis* and also shared the same consensus boundary sequence of 5′-TTGTG (Stevenson et al. 1991).

## Introns and twintrons

RT PCR analyses allowed us to identify precisely determined exon–intron boundaries of 77 introns in *E. mutabilis* (76 introns within protein-coding genes and one intercistronic intron *rps*4-*rps*11). The number corresponds to those of *E. viridis* and *E. viridis* epitype (77/76), but differed significantly from the number of introns detected in *E. gracilis* (134), *E. gracilis* var. *bacillaris* (134), *S. acuminata* (112) and *C. vesiculosum* (128), with twintrons

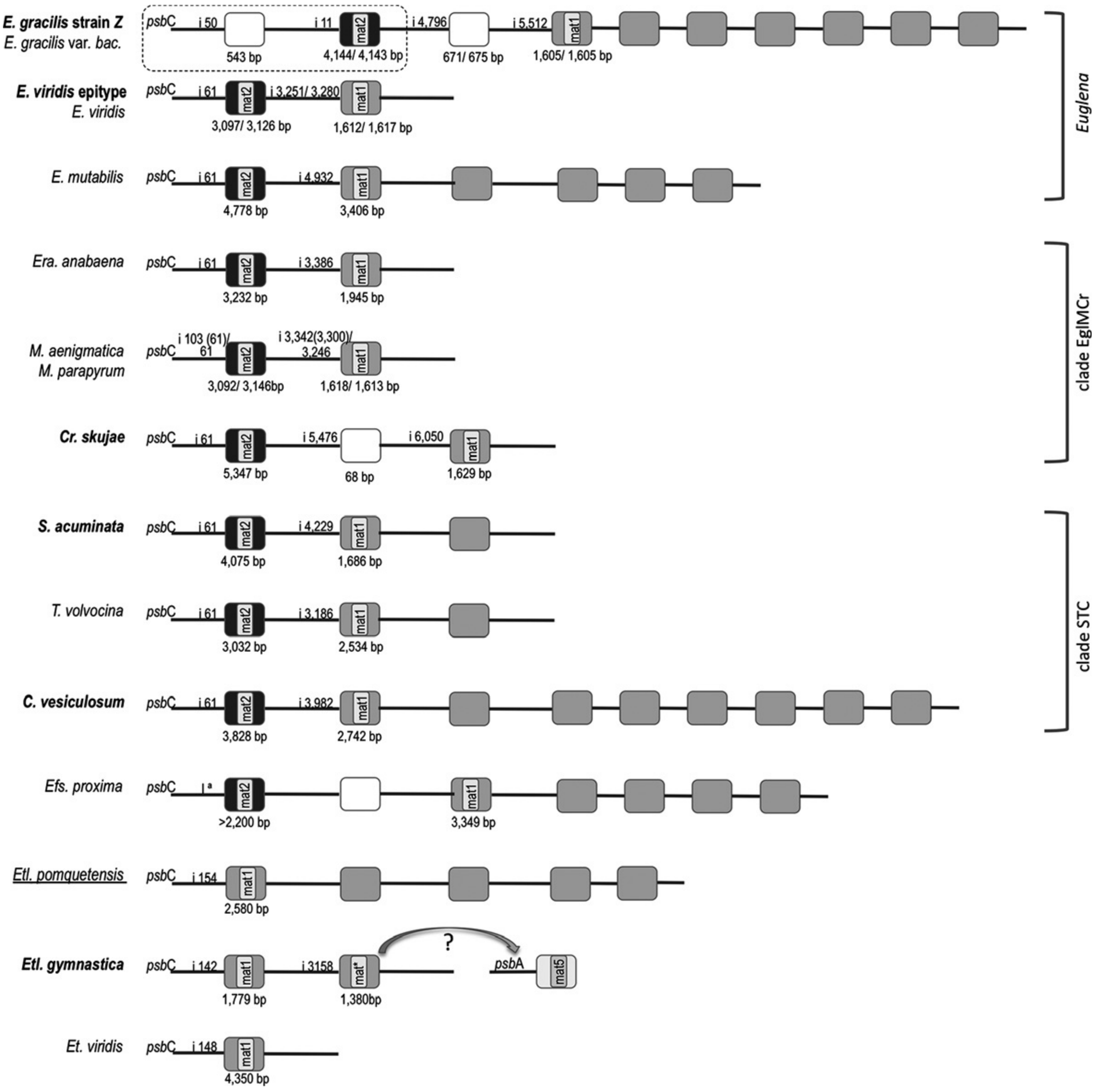


**Figure 2** Distribution of *mat*1 and *mat*2 in *psb*C-introns of euglenid cpGenomes. Clades after Kim et al. (2015). Boxes depict introns from 1 to 10. Dark gray intron boxes: newly inserted intron with *mat*2. White intron boxes: additionally inserted intron without *mat*. Dashed box: Intron insertion in *Euglena gracilis*. *Putative homology to *mat*5 (Blastp). i = insertion site. Species marked bold a re-alignment allowed for improved intron positions. Species underscored unpubl. data. [a]Insertion site of intron 1 not determined due to undetermined exon1.

counted as single insertion sites (Bennett and Triemer 2015; Bennett et al. 2012; Hallick et al. 1993; Pombert et al. 2012; Wiegert et al. 2013).

Our data support the three trends of twintrons found by Bennett and Triemer (2015). First, the twintrons identified in *E. gracilis psb*F and *psb*D were not present in any other Euglenacean cpGenome. Also in *E. mutabilis*, no intron or twintron was found in *psb*F and the introns and twintrons in *psb*D of *E. mutabilis* were not homologous with those of *E. gracilis*.

Second, the assumed ancestral euglenid intron 1 of *psb*C containing the intron encoded *mat*2, was also present in *E. mutabilis*, as in all other cpGenomes of Euglenaceae.

Third, all Euglenaceae, but no Eutreptiales so far contained an intron or twintron in *pet*B (intron1). In *E. mutabilis*, this intron was defined as a group II intron with a characteristic group II 5′-start region (GUGYG) and a length of 435 bp that corresponded to the typical length of 463 ± 90 bp of group II introns. No traits for twintrons have been identified. Since all other Euglenaceae, except for the two *E. gracilis* strains, also range in the same size of group II introns of euglenids (Table 1), we consider this a trend for intron acquisition, but not necessarily for twintrons.

## Expansion of *pbs*C introns and *mat*1/*mat*2

We examined the dispersal of maturase-like proteins *mat*1 (syn. *ycf* 13) and *mat*2 inside the introns of *psb*C of euglenids based on the phylogenetic position after Kim et al. (2015) using Geneious® 7 Pro. Additionally, we re-analyzed all introns and insertion sites of *psb*C genes of previously described euglenid cpGenomes. Closer examination of *psb*C revealed some remarkable molecular differences between Eutreptiales and Euglenales. All Eutreptiales contained *mat*1 in the first intron of *psb*C and always lacked *mat*2 (Fig. 2), whereas all Euglenales possessed both, *mat*1 and *mat*2 (Fig. 2). In Eutreptiales, we detected nearly the same insertion site for intron 1 in all Eutreptiales (Hrdá et al. 2012; Wiegert et al. 2012; Dabbagh et al. unpublished data). In Euglenales, intron 1 contained *mat*2 instead of *mat*1 with the exception of both *E. gracilis* strains (Bennett and Triemer 2015; Hallick et al. 1993), where *mat*2 was located in intron 2, which is probably due to a newly acquired intron 1 at position 50. *Mat*1 on the other hand was almost always uncovered in intron 2 of the *psb*C gene with only few exceptions, like in both *E. gracilis* strains (Bennett and Triemer 2015; Hallick et al. 1993), where it was found in intron 4 (Fig. 2, dark gray intron boxes).

A re-alignment allowed for improved intron positions for six species marked bold in Figure 2 (alignments are available upon request from the authors). Interestingly, in each *psb*C gene of Euglenales, the first intron inserted at position 61, except for both *E. gracilis* strains, *Efs. proxima* and *M. aenigmatica* (Bennett and Triemer 2015; Hallick et al. 1993; Pombert et al. 2012). Only when an alternative start codon (ACG) was accepted in *M. aenigmatica*, intron 1 also inserted at position 61. Our data suggest a probable scenario in which an intron containing *mat*2 inserted at site 61 causing the original intron with *mat*1 to an intron position 2 > 3 kbp downstream. For *Efs. proxima* (Bennett et al. 2014), the number of introns and location of maturases did not seem to fit into any (sub)clade, but were nevertheless most similar to *Cr. skujae* (Bennett and Triemer 2015). A possible explanation for both is, that an intron inserted between intron1 (with mat2) and intron 2 (with mat1) and consequently changed the position of intron 2 to intron position number three (Fig. 2, white intron boxes). Therefore, we placed *Efs. proxima* in agreement with Bennett et al. (2014) at the base of Euglenales.

A likewise scenario could apply for both *E. gracilis* strains (Bennett and Triemer 2015; Hallick et al. 1993) where a new intron was inserted in the first exon of the *psb*C, so that the former intron 1 with *mat*2 became intron 2. By this, exon 1 with usually 61 bp was split into two smaller exons with 50 and 11 bp, respectively (Fig. 2, insert dashed box). Probably, the same insertion process of a new intron happened between intron 2 and 3 (Fig. 2, white intron boxes). On the intrageneric level, *E. mutabilis* was more similar to the two *E. viridis* strains than to the *E. gracilis* cpGenomes. But again, the number of introns of *E. mutabilis* is closer to those of both *E. gracilis* strains (Bennett and Triemer 2015; Bennett et al. 2012; Hallick et al. 1993).

Surprisingly, we identified a maturase-like protein in the second intron in *psb*C of *Etl. gymnastica* (Hrdá et al. 2012) that was neither *mat*1 nor *mat*2. Blastp showed a low similarity to *mat*5 that is usually sited in the neighboring *psb*A gene in Euglenales (Fig. 2). It is conceivable that this maturase-like protein is a dysfunctional pseudogene of *mat*5. Closer examinations will be needed.

## Gene arrangement in clusters

During investigations of clusters with progressive Mauve, the cluster arrangement naturally changed by including more sequences and consequently, no large clusters were predicted, when mismatches occurred. Additionally, many single gene clusters were offered, which did not help to monitor genome-wide modifications. To avoid this and to compare clusters of all species involved to each other and to *Pyramimonas* (Turmel et al. 2009), we used a rather hands-on approach on cluster examination with a presentation form modified after Turmel et al. (2009). Whereas

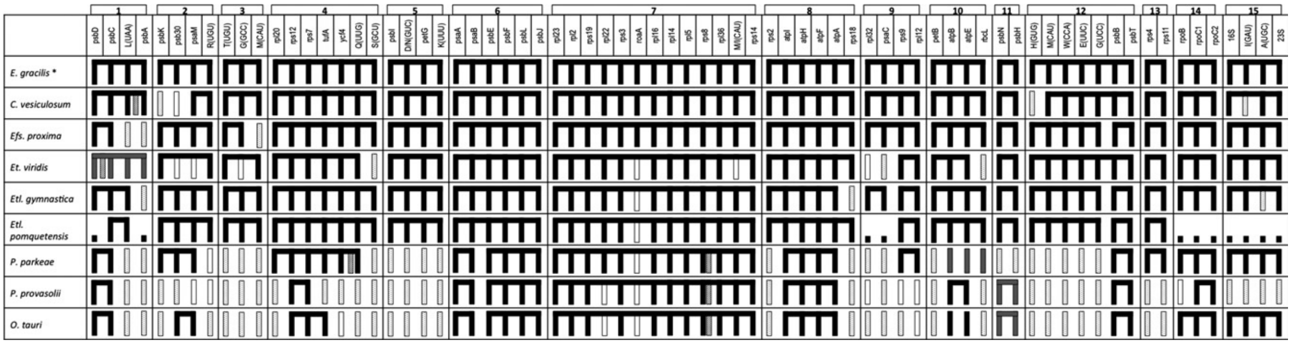


**Figure 3** Conserved gene clusters in euglenid and prasinophyte (*Pyramimonas parkeae*, *Pycnococcus provasolii*, *Ostreococcus tauri*) chloroplast genomes. *Representative for *Euglena gracilis strain Z*, *E. gracilis var. bacillaris*, *E. viridis epitype*, *E. viridis*, *E. mutabilis*, *Era. anabaena*, *M. aenigmatica*, *M. parapyrum*, *S. acuminata*, *Cr. skujae*. *Etl. pomquetensis* unpublished data. Black bars connected by a horizontal line contain genes in the same order and polarity; gray bars connected by a horizontal line contain genes of the same polarity but different order; gray bars contain genes of different order and polarity; white bars denote individual genes that are missing on the cpGenome; black dotted bars characterize genes that are relocated elsewhere on the chloroplast genome; dashed bars stand for additional genes located in the gene clusters. Half bars contain genes that are not identified yet. The relative polarities of the clusters are not represented in this figure; for this information consult Fig. 4.

Mauve is set to identify Multiple Maximal Unique Matches (multi-MUMs), which are exactly matching subsequences shared by two or more genomes (Darling Aaron et al. 2004), our manual approach detected matching regions by comparing all involved sequences in our data base as well as nonmatches. This allowed for the identification of variations within the clusters and also for inclusion of the divergent Eutreptiales and even of some prasinophytes, which in Mauve led only to fragmentation of clusters. In total, 15 conserved gene clusters, numbered 1–15, have been identified in an analysis with all taxa involved, which included 73 of the 94 genes altogether (Fig. 3). Exclusion of the prasinophytes and/or Eutreptiales led to more genes in the clusters, in fact to one very large cluster and two small clusters in the genus *Euglena* (data not shown), but also to loss of evolutionary traits. Therefore, we accepted the inclusion of fewer genes for the additional information of incomplete clusters to follow the path of synteny. The clusters comprised two rRNAs, 17 tRNAs, and 54 protein-coding genes. Only such groupings of genes were labeled as clusters, where the genes shared the same orientation of clockwise or counterclockwise arrangement. *Euglena mutabilis* shared exactly the same clusters with all other Euglenales (Fig. 3) excluding *C. vesiculosum* and *Efs. proxima* (Bennett et al. 2014; Wiegert et al. 2013).

When compared to Eutreptiales, the gene clustering was more similar among Euglenales (Fig. 3). A gene evidently evolving in the lineage of the Euglenales was *roa*A (ribosomal operon associate gene), which was not present in Eutreptiales or in prasinophytes. Additionally, similar gene arrangement and clusters of Eutreptiales were found in the not yet fully annotated cpGenome of *Eutreptiella pomquetensis* (Dabbagh et al., unpubl. data). As expected, the green alga *P. parkeae* (Turmel et al. 2009) shared more characteristics with the three members of Eutreptiales than with Euglenales (Fig. 3). Cluster 9 containing *rpl*32, *psa*C, *rps*9, and *rpl*12, as well as cluster 7 are highly conserved in Euglenales, less in Eutreptiales. Our approach also allowed us to mark genes, which were not included in the same clusters (Fig. 3 white bars), as well as such genes, which were present in the genome, but at another position (black dotted bars). To gain more information about typical gene clustering in prasinophytes as recent ancestors of euglenid chloroplasts, we additionally analyzed *P. provasolii* and *O. tauri* (Robbens et al. 2007; Turmel et al. 2009), the clustering of which are similar (clusters 1, 5, 6, 8 and 12), but not identical to *Pyramimonas*. For example, differences in clusters 4 and 13 again pointed to the closer relationship of Eutreptiales to *Pyramimonas* than to *Ostreococcus* and *Pycnococcus*, which had also been deduced by Turmel et al. (2009). Cluster 1 was the least conserved cluster with *psb*D, *psb*C, *trna*L (UAA) and *psb*A, but became more stable with rising diversification in Euglenales.

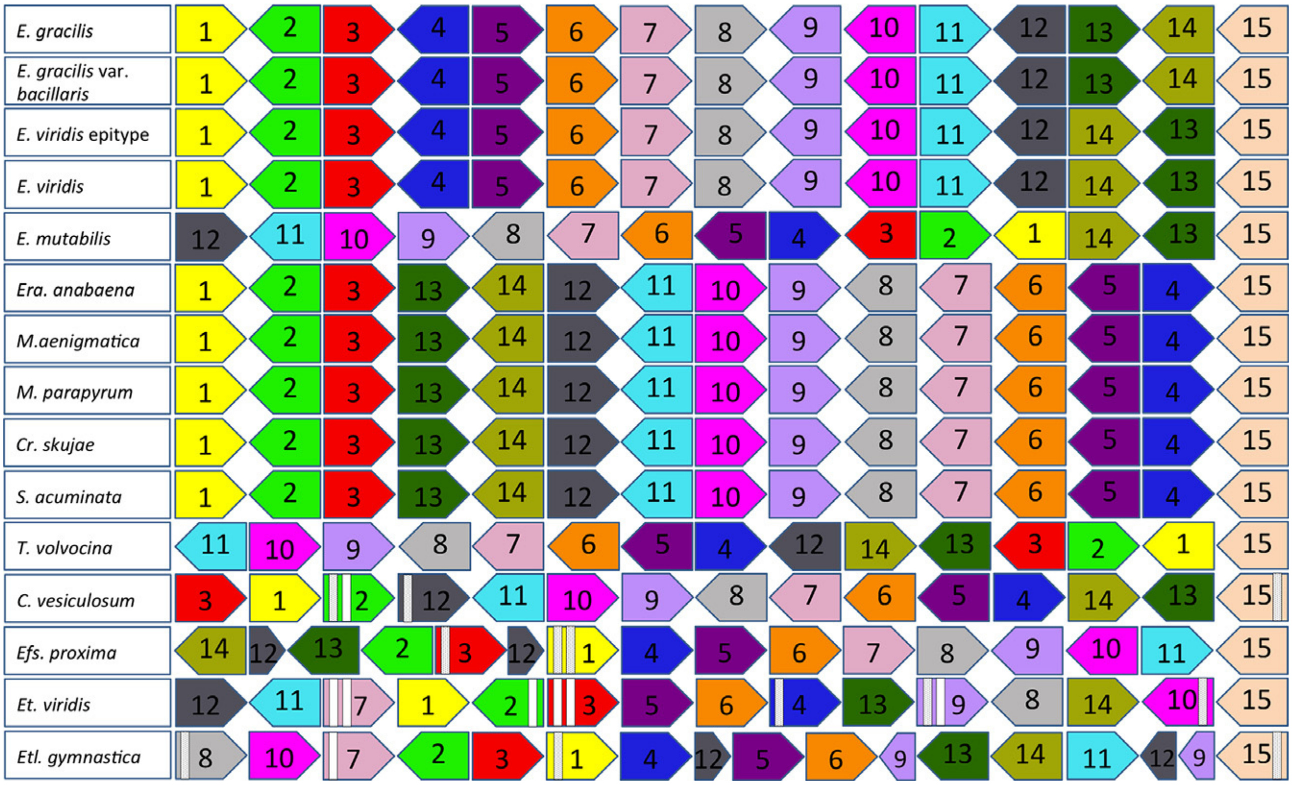

E. gracilis 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
E. gracilis var. bacillaris 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
E. viridis epitype 1 2 3 4 5 6 7 8 9 10 11 12 14 13 15
E. viridis 1 2 3 4 5 6 7 8 9 10 11 12 14 13 15
E. mutabilis 12 11 10 9 8 7 6 5 4 3 2 1 14 13 15
Era. anabaena 1 2 3 13 14 12 11 10 9 8 7 6 5 4 15
M.aenigmatica 1 2 3 13 14 12 11 10 9 8 7 6 5 4 15
M. parapyrum 1 2 3 13 14 12 11 10 9 8 7 6 5 4 15
Cr. skujae 1 2 3 13 14 12 11 10 9 8 7 6 5 4 15
S. acuminata 1 2 3 13 14 12 11 10 9 8 7 6 5 4 15
T. volvocina 11 10 9 8 7 6 5 4 12 14 13 3 2 1 15
C. vesiculosum 3 1 2 12 11 10 9 8 7 6 5 4 14 13 15
Efs. proxima 14 12 13 2 3 12 1 4 5 6 7 8 9 10 11 15
Et. viridis 12 11 7 1 2 3 5 6 4 13 9 8 14 10 15
Etl. gymnastica 8 10 7 2 3 1 4 12 5 6 9 13 14 11 12 9 15

**Figure 4** Polarity and orientation of gene clusters in euglenid cpGenomes. Black dotted bars characterize genes that are relocated elsewhere on the chloroplast genome; white bars denote individual genes that are missing on the cpGenome. ORFs between clusters were not included.

### Cluster orientation

To facilitate comparisons throughout our analyses, all genomes in this study were arranged like *E. gracilis* strain Z (Hallick et al. 1993) with rRNA genes on the reverse strand and left to base 1. Evidently, both *E. gracilis* and both *E. viridis* (Bennett and Triemer 2015; Bennett et al. 2012; Hallick et al. 1993) shared an almost identical orientation of gene clusters (Fig. 4). They only switched position and orientation in clusters 13 and 14 (Bennett and Triemer 2015). *Euglena mutabilis*, on the other hand, showed the same clusters 1–15, but an identical succession of clusters only from cluster 1–12 and these were additionally laterally reversed. These results underlined common features of the Euglenaceae, but also great alterations in regard to *E. mutabilis*. Interestingly, the reversed pattern of clusters 4–12 in *E. mutabilis* could also be found in *C. vesiculosum*, *Cr. skujae*, *Era. anabaena*, *S. acuminata* and both *Monomorphina* species (Bennett and Triemer 2015; Pombert et al. 2012; Wiegert et al. 2012), but at a different position on the genome. Almost all of these shared the same synteny and the same counterclockwise orientation (Bennett and Triemer 2015). *Trachelomonas volvocina* showed the same cluster arrangement from cluster 4–11, but at a different position on the genome (Fig. 4). The genus *Euglena* could be differentiated from *E. proxima* (Bennett et al. 2014) by a splitted cluster 12 and different gene arrangement in clusters 1 and 3. Additionally, the genus *Euglena* contained four tRNAs (tRNA-Leu, tRNA-Arg, tRNA-Asn, tRNA-Val) in a row between cluster 13 and 14, which were dispersed in the *Efs. proxima* genome (data not shown). Except for clusters four to 11, *Efs. proxima* showed no further similarity to clusters of the clade *Euglena*. These findings support the exclusion from the genus *Euglena* and re-description as *E. proxima* by Bennett et al. (2014) and Kim et al. (2015).

We can also support the two large clusters within the Euglenales identified by Bennett and Triemer (2015). One was found within the genus *Euglena* and the other as a sister grouping to the *Euglena* clade. The laterally mirrored orientation of clusters in the *E. mutabilis* cpGenome probably could be regarded as transition from the cpGenomes of the sister group to other members of the genus *Euglena.* In the same way, the cluster orientation of *C. vesiculosum* (Wiegert et al. 2012) and *T. volvocina* (Bennett and Triemer 2015), which also differed from those of *Cr. skujae, S. acuminata, Era. anabaena, M. aenigmatica*, and *M. parapyrum* (Bennett and Triemer 2015; Pombert et al. 2012; Wiegert et al. 2012), could perhaps come out as a link to the genomes of Phacaceae, which can only be ascertained after investigation of more phacacean cpGenomes.

The early-diverging lineages of Eutreptiales still have a larger variation in gene arrangement with several clusters in a different formation. The arrangements such as those in clusters 1 and 9 were probably inherited from a green alga (Fig. 3) during the establishment of the euglenid chloroplasts via secondary endocytobiosis (Gibbs 1978) and changed during the diversification of Eutreptiales and Euglenales. This became evident by looking into *roa*A, which was present in all members of the Euglenales but not in prasinophytes and Eutreptiales (Fig. 3). Besides conformity in gene arrangements, even closely related species differed extremely in cpGenome size (Table 1), though the number of conserved genes in all three *Euglena* species was not remarkably different. The unequal genome sizes were for one part caused by varying numbers of introns. The difference in cpGenome size between *E. mutabilis* and *E. viridis* epitype (Bennett et al. 2012) can be ascribed to the large ORF region of 13,773 bp in *E. viridis* epitype. A factor for size variation between both *E. gracilis* strains Bennett and Triemer 2015; Hallick et al. 1993) and the other *Euglena* species was the number of ribosomal operons. While the cpGenomes of *E. mutabilis* and *E. viridis* (Bennett and Triemer 2015; Bennett et al. 2012) only possessed one complete ribosomal operon, the *E. gracilis* genomes (Bennett and Triemer 2015; Hallick et al. 1993) contained three tandemly repeated complete copies of the ribosomal operon plus one additional 16S rRNA on the cpGenome of *E. gracilis* strain Z (Hallick et al. 1993). We found no evolutionary trend concerning the ribosomal operon in euglenids.

## CONCLUSIONS

The cpGenome of *E. mutabilis* being almost 87 kbp is in the middle length of the genomes sequences so far. Gene orientation and number of introns showed a closer placement to both *E. viridis* strains than to *E. gracilis* strains (Bennett and Triemer 2015; Bennett et al. 2012; Hallick et al. 1993) and had some similarities in cluster orientation with *Monomorphina* and relatives (Bennett and Triemer 2015; Pombert et al. 2012). Still unanswered remained the question how the large variation in intron numbers within evolutionary lineages is possible. It does not appear to be parsimonious, that intron invasion or intron loss happened several times independently in euglenid plastid evolution. Data available so far indicate different distributional patterns of introns across genera, and especially within the genus *Euglena*. An example on how this could have happened can be found in the *psb*C genes. Data available so far slightly hint on multiple uptake or at least at different distributional patterns even in one genus. It becomes very obvious that more plastid genomes need to be sequenced and secondary structure analyses of introns will have to follow, especially for the Eutreptiales, which are the closest to the acquisition of chloroplasts by secondary endocytobiosis.

## ACKNOWLEDGMENTS

The authors would like to thank Marisa Bartling, Margret Buse, Cora Berger and Stefan Paerschke for critical reading and Sabine Stratmann-Lettner and Lukas Hykes for lab assistance. Additional thank is due to A. Donath (MITOS) for consultations regarding exonerate. This work was supported by a postgraduate scholarship of Bergische University Wuppertal to ND.

## LITERATURE CITED

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410.

Aronsson, H. & Jarvis, P. 2002. A simple method for isolating import-competent *Arabidopsis* chloroplasts. *FEBS Lett.*, 2–3:215–220.

Bennett, M. S. & Triemer, R. E. 2015. Chloroplast genome evolution in the Euglenaceae. *J. Eukaryot. Microbiol.* doi:10.1111/jeu.12235.

Bennett, M. S., Wiegert, K. E. & Triemer, R. E. 2012. Comparative chloroplast genomics between *Euglena viridis* and *Euglena gracilis* (Euglenophyta). *Phycologia*, 6:711–718.

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*, 1:66–73.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P. & Bateman, A. 2013. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.* 41:D226-D232.

Candales, M. A., Duong, A., Hood, K. S., Li, T., Neufeld, R. A. E., Sun, R., McNeil, B. A., Wu, L., Jarding, A. M. & Zimmerly, S. 2011. Database for bacterial group II introns. *Nucleic Acids Res.*, D1:D187–D190.

Casiot, C., Bruneel, O., Personné, J.-C., Leblanc, M. & Elbaz-Poulichet, F. 2004. Arsenic oxidation and bioaccumulation by the acidophilic protozoan, *Euglena mutabilis*, in acid mine drainage (Carnoulès, France). *Sci. Total Environ.*, 2–3:259–267.

Christopher, D. A. & Hallick, R. B. 1989. *Euglena gracilis* chloroplast ribosomal protein operon: a new chloroplast gene for ribosomal protein L5 and description of a novel organelle intron category designated group III. *Nucleic Acids Res.*, 19:7591–7608.

Ciugulea, I. & Triemer, R. E. 2010. A color atlas of photosynthetic euglenoids. Michigan State University Press, East Lansing, MI. 204 p.

Conant, G. C. & Wolfe, K. H. 2008. GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics*, 6:861–862.

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 Res.*, 23:6491–6497.

Copertino, D. W., Hall, E. T., Van Hook, F. W., Jenkins, K. P. & Hallick, R. B. 1994. A group III twintron encoding a maturase-like gene excises through lariat intermediates. *Nucleic Acids Res.*, 6:1029–1036.

Copertino, D. W. & Hallick, R. B. 1993. Group II and group III introns of twintrons: potential relationships with nuclear pre-mRNA introns. *Trends Biochem. Sci.*, 12:467–471.

Copertino, D. W., Shigeoka, S. & Hallick, R. B. 1992. Chloroplast group III twintron excision utilizing multiple 5′- and 3′-splice sites. *EMBO J.*, 13:5041–5050.

Cramer, M. & Myers, J. 1952. Growth and photosynthetic characteristics of *Euglena gracilis*. *Arch. Mikrobiol.*, 1–4:384–402.

Dai, L., Toor, N., Olson, R., Keeping, A. & Zimmerly, S. 2003. Database for mobile group II introns. *Nucleic Acids Res.*, 1:424–426.

Darling Aaron, C. E., Mau, B., Blattner, F. R. & Perna, N. T. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, 7:1394–1403.

Doetsch, N. A., Thompson, M. D., Favreau, M. R. & Hallick, R. B. 2001. Comparison of *psb*K operon organization and group III intron content in chloroplast genomes of 12 Euglenoid species. *Mol. Gen. Genet.*, 5:682–690.

Doetsch, N. A., Thompson, M. D. & Hallick, R. B. 1998. A maturase-encoding group III twintron is conserved in deeply rooted euglenoid species: are group III introns the chicken or the egg? *Mol. Biol. Evol.*, 1:76–86.

Drager, R. G. & Hallick, R. B. 1993. A complex twintron is excised as four individual introns. *Nucleic Acids Res.*, 10:2389–2394.

Gao, L., Zhou, Y., Wang, Z.-W., Su, Y.-J. & Wang, T. 2011. Evolution of the *rpo*B-*psb*Z region in fern plastid genomes: notable structural rearrangements and highly variable intergenic spacers. *BMC Plant Biol.*, 64. doi:10.1186/1471-2229-11-64.

Gibbs, S. P. 1978. The chloroplasts of *Euglena* may have evolved from symbiotic green algae. *Can. J. Bot.*, 22:2883–2889.

Gockel, G. & Hachtel, W. 2000. Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate *Astasia longa*. *Protist*, 4:347–351.

Gojdics, M. 1953. The genus Euglena. University of Wisconsin Press, *Madison*. 268 p.

Häder, D.-P. & Melkonian, M. 1983. Phototaxis in the gliding flagellate, *Euglena mutabilis*. *Arch. Microbiol.*, 1:25–29.

Hallick, R. B., Hong, L., Drager, R. G., Favreau, M. R., Monfort, A., Orsat, B., Spielmann, A. & Stutz, E. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Res.*, 15:3537–3544.

Harding, R. M., Boyce, A. J. & Clegg, J. B. 1992. The evolution of tandemly repetitive DNA: recombination rules. *Genetics*, 3:847–859.

Hong, L. & Hallick, R. B. 1994. A group III intron is formed from domains of two individual group II introns. *Genes Dev.*, 13:1589–1599.

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*, 3:e33746. doi:10.1371/journal.pone.0033746.

Kasiborski, B. A., Bennett, M. S. & Linton, E. W. 2016. The chloroplast genome of *Phacus orbicularis* (Euglenophyceae): an initial datum point for the phacaceae. *J. Phycol.* doi:10.1111/jpy.12403

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 12:1647–1649.

Khan, H. & Archibald, J. M. 2008. Lateral transfer of introns in the cryptophyte plastid genome. *Nucleic Acids Res.*, 9:3043–3053.

Kim, J. I., Linton, E. W. & Shin, W. 2015. Taxon-rich multigene phylogeny of the photosynthetic euglenoids (Euglenophyceae). *Front. Ecol. Evol.*, 3:98. doi:10.3389/fevo.2015.00098.

Kim, J. I., Shin, W. & Triemer, R. E. 2010. Multigene analyses of photosynthetic euglenoids and new family, Phacaceae (Euglenales). *J. Phycol.*, 6:1278–1287.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J. & Giegerich, R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, 22:4633–4642.

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T. & Ussery, D. W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, 9:3100–3108.

Lemieux, C., Otis, C. & Turmel, M. 2007. A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC Biol.*, 2: doi:10.1186/1741-7007-5-2.

Linton, E. W., Karnkowska-Ishikawa, A., Kim, J. I., Shin, W., Bennett, M. S., Kwiatowski, J., Zakryś, B., Triemer, R. E. & Zakryś, B. 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*, 4:603–619.

Michel, F. & Ferat, J.-L. 1995. Structure and activities of group II introns. *Ann. Rev. Biochem.*, 1:435–461.

Michel, F., Kazuhiko, U. & Haruo, O. 1989. Comparative and functional anatomy of group II catalytic introns — a review. *Gene*, 1:5–30.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M. & Kumlin, E. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science (New York, NY)*, 4796:1616–1622.

Pombert, J.-F., James, E. R., Janouškovec, J., Keeling, P. J. & McCutcheon, J. 2012. Evidence for transitional stages in the evolution of Euglenid group II introns and twintrons in the *Monomorphina aenigmatica* plastid genome. *PLoS ONE*, 12: e53433. doi:10.1371/journal.pone.0053433.

Robbens, S., Derelle, E., Ferraz, C., Wuyts, J., Moreau, H. & de Van 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.*, 4:956–968.

Schattner, P., Brooks, A. N. & Lowe, T. M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucl. Acids Res. Web Server*: W686. doi:10.1093/nar/gki366.

Sheveleva, E. V. & Hallick, R. B. 2004. Recent horizontal intron transfer to a chloroplast genome. *Nucleic Acids Res.*, 2:803–810.

Slater, Guy St C & Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 1: 31. doi: 10.1186/1471-2105-6-31.

Stevenson, J. K., Drager, R. G., Copertino, D. W., Christopher, D. A., Jenkins, K. P., Yepiz-Plascencia, G. & Hallick, R. B. 1991. Intercistronic group III introns in polycistronic ribosomal protein operons of chloroplasts. *Mol. Gen. Genet.*, 1–2:183–192.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.*, 10:2731–2739.

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 Res.*, 23:4745–4752.

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. *Mol. Biol. Evol.*, 3:631–648.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. & Rozen, S. G. 2012. Primer3-new capabilities and interfaces. *Nucl. Acids Res.*, 15:e115.

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*, 6:832–843.

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). *J. Eukaryot. Microbiol.*, 2:214–221.

Zimmerly, S., Hausner, G. & Xc-Chu, Wu 2001. Phylogenetic relationships among group II intron ORFs. *Nucleic Acids Res.*, 5:1238–1250.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article:

**Table S1.** Primer list of protein-coding genes for RT PCR analysis.

**Table S2.** Genes, coding sequences and introns of the chloroplast genome of *Euglena mutabilis*. T: twintrons, T*: identical Intron to E.gracilis strain Z, X: not identified because of untypical 5′-boundaries. Gray: uncertain group III twintron type. Dark gray: uncertain group II twintron type.

## The Journal of Eukaryotic Microbiology

The Chloroplast Genome of *Euglena mutabilis* - Cluster Arrangement, Intron Analysis and Intrageneric Trends

Nadja Dabbagh, Angelika Preisfeld

## SUPPORTING INFORMATION

**Table S1.** Primer list of protein-coding genes for RT-PCR analysis.

| Protein coding gene | forward (5´→ 3´) | reverse (5´→ 3´) |
|---|---|---|
| *rps*2 | GCTTCAAGTGTGCATTTAGGAC | GATAAAGAGTCATCATTAGCTGG |
| *rps*3 | GGGACAAAAAGTACATCCAGAAG | GACCCATATTTTAATTCCAATTATACC |
| *rps*8 | GCTTACACGCATAAAAAATGCAAG | GCAATCCTATCTGTTACAATACC |
| *rps*9 | TTAACGCTTAGAAAATTGAGAAGC | GTAGTTAAATTAATACGAGGAGAAG |
| *rps*11 | GAAGCAAAAGTCATAAGTTTAG | TACATTCACTTCATTAAATTCTTC |
| *rps*14 | GGAAAAACGTCTAATTCTTGTTGA | ACTCCAGGTAATAATCCATAGTG |
| *rps*18 | GCTATAAATGAATATAAAAAGTCTCGC | GGTAATAGTCCCAAAATTCTAG |
| *rps*19 | GTCTCGTTCTTCAATAAAAGGTC | CATGACCCTTAAAATTTCTTGTTGG |
| *rpl*12 | CTTCCTCTGCTTTTTCTTTTGG | ATGTCAACAAAAAATAGAAGAAATTGTAG |
| *rpl*14 | CAGACAATACAGGCGCACAAAAAC | CTCTTAATTCTCTTGCAACAGGTCC |
| *rpl*16 | CATCATAGAGGAAGAATGAATGG | GGCATTTTATGTCCCGCAGT |
| *rpl*20 | TGTGGACAAAATTTTAGAGCTGG | GTATTTTTAAAGTTAACATAATGCGGCT |
| *rpl*23 | CTGAAAGTACATAATTTATGATTGATTTAA | GCAGAAAAGAAAGGTATTTTTTCAC |
| *psb*A | GATGGTATCCGTGAGCCTGTTTC | GCTAAGTCTAATGGGAAATTGTGTGC |
| *psb*B | CAATACCTGCAAAAACATCTCGG | GGGATTACCTTGGTATCGTG |
| *psb*C | CAACTTGTATTCCCGGAGGAGG | CATTTGGACCACGTAACGGTTC |
| *psb*D | GGTTGGTTCTTTGCTCCTAG | CATCCAAGCACGAATACCTTC |
| *psb*T | CACTTTTACAATACGAGGTGGTTC | GATGTTACTACACGAAAACAGCTTG |
| *rpo*C1 | CTGTGTTACCCATACATAGTTTTG | GCTTCACCACAAAAGATCTTAG |
| *rpo*C2 | CCCTCTAATATTTGTTCTATTTTTGG | GCACGAAAAGGTTTAGTTGATAC |
| *rpo*B left | GCCCAAACTTCCATTTCTCC | GGAAGGATACAATTTTGAAGATGC |
| *rpo*B middle | GCATCTTCAAAATTGTATCCTTC | GGATGAAAACAAATACGATCTAAC |
| *rpo*B right | GTTAGATCGTATTTGTTTTCATCC | GTATAAAAAGAGAGAAAATAGAATCGC |
| *atp*A | GTAATGGCAGGAGAACTTGTTG | CAACGTCTACATCATCTAAATATCC |
| *atp*B | CGAACCTGTAAACACTTCAGCTAC | CCTGTAATGGATATTTCGTTTCCTGC |
| *atp*E | GAAGGCCTTTTTTGTGATGATAG | GGCCTATTTTTCTAATATGTTTATGTC |
| *atp*I | GTTTCACTGATCTTAAATGCACTTAATG | CCTGAATACCACTAGTAAATAGACC |
| *pet*B | CAACGCTAACACTTCCTCTC | ATGTTTAGATTTTTTTACTTTTATTTATATGAG |
| *psa*A | CGGTGATCTTGGTGGTAATTTTC | CGTCAGACTGCATTTTCCAAC |
| *psa*B | GCCAAGGTTTAGCTCAAGATC | CACCCGTCATTATAAATCCTGCT |
| *psa*C | CTGTCGGACAAGCAGATTCA | GTGGATTAATTTTATATTGTCTATCGC |
| *rbc*L | CGAGCAAGATCACGACCTTC | GCTGGTGTAAAAGACTATCGTC |
| *roa*A | GTTAAACAGATTTGAAATTC | GTCTTTCTTGACATAATAAA |
| *rpo*A | CTCCTTCTCCAAGATGCTTT | CTTGTAATTGGTGTATTGAATTATTC |
| *ycf*4 | GGTATTTTTAGAAATGATGCAAT | AGATAACTTAAACCAACTAAATG |
| *ycf*13 | GCTCGTCTTACAAATCTTTCAG | CGTAAAAACCCATACCACAATG |

| Check for misassemble of the genome | | |
|---|---|---|
| | forward (5´→ 3´) | reverse (5´→ 3´) |
| PCR between 16S and *psb*B | GGCTGTACATCCTCTAAGAC | GTTAAAGTCACAACACGATACCA |
| PCR between *rpo*B and *chl*I | GATTCTGGACAAACTATTTTACAGT | TGTTAGATCGTATTTGTTTTCATCC |

**Table S2:** Genes, coding sequences and introns of the chloroplast genome of *Euglena mutabilis*. T: twintrons, T*: identical Intron to *E.gracilis* strain Z, X: not identified because of untypical 5'-boundaries. Grey: uncertain group III twintron type. Dark grey: uncertain group II twintron type.

| Gene | Gene length | CDS length | Intron no. | Intron insertion site | Intron start 5\` | Intron stop 3\` | Intron length | Intron GC% | Group II or III Intron |
|---|---|---|---|---|---|---|---|---|---|
| *atp*A | 2052 | 1512 | 1 | 753 | GTGCG | AA | 540 | 21,9 | II |
| *atp*B | 2021 | 1443 | 1 | 684 | GTGCG | AC | 578 | 18,5 | II |
| *atp*E | 733 | 414 | 1 | 22 | GTTTG | AC | 319 | 16 | II/TIII |
| *atp*F | 939 | 591 | 1 | 177 | TTTAT | TA | 348 | 19 | X |
| *chl*I | 1298 | 1050 | 1 | 65 | GTGCG | AT | 248 | 16,9 | II |
| *psa*B | 2929 | 2205 | 1 | 169 | GTGCG | AT | 724 | 17,8 | II |
| *psb*A | 1634 | 1035 | 1 | 993 | GTGCG | AT | 599 | 23 | II |
| *psb*K | 262 | 165 | 1 | 94 | TTGTG | AA | 97 | 20,6 | III |
| *psb*T | 1605 | 114 | 1 | 24 | GTTTG | AT | 1491 | 20,3 | TII* |
| *ycf*4 | 1373 | 573 | 1 | 390 | GTGCG | AT | 800 | 19,9 | TII |
| *tuf*A | 1801 | 1230 | 1 | 445 | GTGCG | AT | 571 | 19,4 | II |
| *rpl*2 | 1176 | 834 | 1 | 28 | GTGCG | AT | 342 | 17 | II |
| *rpl*12 | 504 | 396 | 1 | 80 | TTTTG | TA | 108 | 12 | III |
| *rpl*14 | 485 | 366 | 1 | 129 | TTTTG | TA | 119 | 9,2 | III |
| *rps*8 | 669 | 396 | 1 | 313 | TTGCG | AG | 273 | 23,1 | TIII |
| *rps*14 | 506 | 303 | 1 | 177 | TTATG | TG | 203 | 10,8 | TIII |
| *rbc*L | 2122 | 1428 | 1 | 73 | GTGCG | AT | 694 | 20,9 | II |
| *roa*A | 1745 | 1638 | 1 | 63 | ATTTG | AT | 107 | 15 | III |
| *rpo*B | 2799 | 2529 | 1 | 1.473 | GTACG | AT | 270 | 20 | II/TIII |
| *rps*11 | 575 | 357 | 1 | 66 | TTTTG | TT | 110 | 7,3 | III |
|  |  |  | 2 | 195 | TTGTG | TT | 108 | 14,8 | III |
| *psb*D | 1954 | 1059 | 1 | 622 | CTCTT | TA | 343 | 19,5 | X |
|  |  |  | 2 | 1.273 | ATGCG | AT | 552 | 19,9 | II/TIII |
| rps3 | 1188 | 636 | 1 | 46 | TTTTG | TT | 452 | 16,4 | TIII* |
|  |  |  | 2 | 1.035 | GTTTG | AA | 100 | 11 | III |
| *rps*18 | 523 | 192 | 1 | 11 | TTGAG | AA | 107 | 15,9 | III |
|  |  |  | 2 | 242 | TTGTG | CT | 224 | 13,8 | TIII* |

**Table S2:** Continued.

| Gene | Gene length | CDS length | Intron no. | Intron insertion site | Intron start 5` | Intron stop 3` | Intron length | Intron GC% | Group II or III Intron |
|---|---|---|---|---|---|---|---|---|---|
| *psa*C | 844 | 246 | 1 | 42 | GTGCG | TT | 267 | 16,5 | II |
| | | | 2 | 340 | GTATT | TT | 331 | 20,2 | X |
| *pet*B | 1748 | 648 | 1 | 23 | GTGCG | CC | 435 | 20,9 | II |
| | | | 2 | 505 | GTGCG | AT | 665 | 22,7 | II |
| *ycf*13 | 3149 | 1362 | 1 | 64 | GTGCG | CC | 1086 | 23,3 | TII |
| | | | 2 | 1.744 | GTGCG | AT | 701 | 19,1 | II |
| *rps*19 | 488 | 279 | 1 | 78 | ATTTG | TA | 107 | 14 | III |
| | | | 2 | 312 | ATTTG | AA | 102 | 8,8 | III |
| *psa*A | 4071 | 2256 | 1 | 458 | GTGCG | AT | 724 | 21,1 | II |
| | | | 2 | 1.267 | GTGCG | AT | 493 | 19,9 | II |
| | | | 3 | 3.014 | GTGCG | AC | 598 | 18,2 | II |
| *rpl*23 | 613 | 279 | 1 | 13 | TTTTG | GA | 116 | 17,2 | III |
| | | | 2 | 209 | TTGTG | AA | 111 | 19,8 | III |
| | | | 3 | 467 | TTGTG | TA | 107 | 15 | III |
| *rpl*16 | 1002 | 402 | 1 | 66 | TTGAG | GC | 118 | 15,3 | III |
| | | | 2 | 212 | GAGCG | AT | 383 | 19,1 | X |
| | | | 3 | 645 | TTTTG | TA | 99 | 13,1 | III |
| *rps*2 | 1062 | 663 | 1 | 191 | TTGTG | TT | 94 | 16 | III |
| | | | 2 | 438 | TTTTG | TA | 203 | 13,8 | TIII |
| | | | 3 | 718 | TTTTG | AA | 102 | 11,8 | III |
| *rps*9 | 838 | 390 | 1 | 122 | TTGAG | CA | 100 | 20 | III |
| | | | 2 | 239 | TTTTG | AA | 99 | 13,1 | III |
| | | | 3 | 539 | GTGCG | AG | 249 | 20,9 | II |
| *psb*B | 3650 | 1527 | 1 | 22 | TTGTG | TA | 107 | 17,8 | III |
| | | | 2 | 463 | GTGCG | CT | 1027 | 30,4 | TII |
| | | | 3 | 1.707 | GTGCG | AT | 989 | 23,4 | TII |
| *rpo*A | 959 | 636 | 1 | 141 | TTGAG | TA | 105 | 13,3 | III |
| | | | 2 | 573 | TTTTG | GA | 112 | 18,8 | III |
| | | | 3 | 754 | AAGTT | AA | 106 | 17,9 | X |

**Table S2 :** Continued.

| Gene | Gene length | CDS length | Intron no. | Intron insertion site | Intron start 5` | Intron stop 3` | Intron length | Intron GC% | Group II or III Intron |
|---|---|---|---|---|---|---|---|---|---|
| *atp*I | 1579 | 726 | 1 | 25 | TTGTG | AA | 117 | 12 | III |
| | | | 2 | 156 | GTTTG | AA | 216 | 16,7 | TIII |
| | | | 3 | 716 | GTACG | AC | 307 | 17,6 | TIII |
| | | | 4 | 1.094 | TTGAG | AA | 103 | 10,7 | III |
| | | | 5 | 1.334 | TTGTG | AA | 110 | 17,3 | III |
| *psb*C | 12192 | 1422 | 1 | 61 | GTGTG | CC | 4778 | 19,4 | TII |
| | | | 2 | 143 | GTGAG | AT | 3406 | 22,7 | TIII |
| | | | 3 | 3.957 | GTGCG | AC | 572 | 21,7 | II |
| | | | 4 | 4.562 | GTGCG | AT | 693 | 18,8 | II |
| | | | 5 | 5.481 | GTGCG | AT | 724 | 19,6 | II |
| | | | 6 | 6.461 | TTGCG | AT | 597 | 19,8 | II/TIII |
| *rpo*C1 | 2726 | 1698 | 1 | 105 | AAGAA | AA | 200 | 18 | X |
| | | | 2 | 511 | GTGTG | TT | 111 | 17,1 | III |
| | | | 3 | 687 | TTGAG | AA | 98 | 18,4 | III |
| | | | 4 | 927 | GTGAG | AA | 119 | 18,5 | III |
| | | | 5 | 1.071 | TTGAG | AC | 92 | 19,6 | III |
| | | | 6 | 1.532 | TTAAG | AA | 107 | 14 | III |
| | | | 7 | 1.793 | TTATG | AA | 105 | 12,4 | III |
| | | | 8 | 2.035 | GTGTG | AA | 91 | 13,2 | III |
| | | | 9 | 2.451 | TTTTG | AC | 105 | 14,3 | III |
| *rps*4-11 | intercistronic | | | | TTGTG | TT | 109 | 11,9 | III |

# Chloroplast genome expansion by intron multiplication in the basal psychrophilic euglenoid *Eutreptiella pomquetensis*

Nadja Dabbagh[1], Matthew S. Bennett[2], Richard E. Triemer[2] and Angelika Preisfeld[1]

[1] Faculty of Mathematics and Natural Sciences, Zoology and Didactics of Biology, Bergische Universität Wuppertal, Wuppertal, Germany

[2] Department of Plant Biology, Michigan State University, East Lansing, MI, United States of America

## ABSTRACT

**Background**. Over the last few years multiple studies have been published showing a great diversity in size of chloroplast genomes (cpGenomes), and in the arrangement of gene clusters, in the Euglenales. However, while these genomes provided important insights into the evolution of cpGenomes across the Euglenales and within their genera, only two genomes were analyzed in regard to genomic variability between and within Euglenales and Eutreptiales. To better understand the dynamics of chloroplast genome evolution in early evolving Eutreptiales, this study focused on the cpGenome of *Eutreptiella pomquetensis*, and the spread and peculiarities of introns.

**Methods**. The *Etl. pomquetensis* cpGenome was sequenced, annotated and afterwards examined in structure, size, gene order and intron content. These features were compared with other euglenoid cpGenomes as well as those of prasinophyte green algae, including *Pyramimonas parkeae*.

**Results and Discussion**. With about 130,561 bp the chloroplast genome of *Etl. pomquetensis*, a basal taxon in the phototrophic euglenoids, was considerably larger than the two other Eutreptiales cpGenomes sequenced so far. Although the detected quadripartite structure resembled most green algae and plant chloroplast genomes, the gene content of the single copy regions in *Etl. pomquetensis* was completely different from those observed in green algae and plants. The gene composition of *Etl. pomquetensis* was extensively changed and turned out to be almost identical to other Eutreptiales and Euglenales, and not to *P. parkeae*. Furthermore, the cpGenome of *Etl. pomquetensis* was unexpectedly permeated by a high number of introns, which led to a substantially larger genome. The 51 identified introns of *Etl. pomquetensis* showed two major unique features: (i) more than half of the introns displayed a high level of pairwise identities; (ii) no group III introns could be identified in the protein coding genes. These findings support the hypothesis that group III introns are degenerated group II introns and evolved later.

**Subjects** Cell Biology, Evolutionary Studies, Genomics

**Keywords** *Eutreptiella pomquetensis*, Introns, Twintrons, Genome structure, Chloroplast genome

Submitted 10 March 2017
Accepted 1 August 2017
Published 25 August 2017

Corresponding author
Nadja Dabbagh,
dabbagh@uni-wuppertal.de

Academic editor
Mikhail Gelfand

Additional Information and Declarations can be found on page 17

DOI 10.7717/peerj.3725

**How to cite this article** Dabbagh et al. (2017), Chloroplast genome expansion by intron multiplication in the basal psychrophilic euglenoid *Eutreptiella pomquetensis*. **PeerJ 5:e3725; DOI 10.7717/peerj.3725**

## INTRODUCTION

Recent analyses of chloroplast genomes (cpGenomes) have been largely used to retrace evolutionary steps of phototrophic euglenoids. Members of the genera *Euglena*, *Monomorphina*, *Euglenaformis*, *Colacium*, *Strombomonas* and recently *Phacus* (*Bennett, Wiegert & Triemer, 2012*; *Bennett, Wiegert & Triemer, 2014*; *Bennett & Triemer, 2015*; *Bennett, Shiu & Triemer, 2017*; *Dabbagh & Preisfeld, 2016*; *Gockel & Hachtel, 2000*; *Hallick et al., 1993*; *Kasiborski, Bennett & Linton, 2016*; *Pombert et al., 2012*; *Wiegert, Bennett & Triemer, 2013*) cpGenomes of the 'crown group' Euglenales have been studied intensely. Overall aims were to tackle questions of relatedness, gene arrangement, synteny and genome size as well as possession and dispersal of introns. However, the knowledge on cpGenomes of the basal lineage, Eutreptiales, is comparatively low. The two known genomes were reported to show a smaller genome size and display only seven and 27 introns in *Eutreptiella gymnastica* and *Eutreptia viridis*, respectively (*Hrdá et al., 2012*; *Wiegert, Bennett & Triemer, 2012*). Fitting into a scheme of increasing intron quantity and genome size, the invasion of introns in euglenoids was assumed to have started with very low intron numbers and as a consequence small cpGenomes in Eutreptiales, which both increased during diversification of photosynthetic euglenoids (*Bennett & Triemer, 2015*; *Hrdá et al., 2012*; *Thompson et al., 1995*; *Wiegert, Bennett & Triemer, 2012*). This hypothesis was initially corroborated by the fact that in *Pyramimonas parkeae*, as the closest living relative of the euglenoid plastid, only one intron was detected (*Turmel et al., 2009*). However, this was later refuted by analysis of different lineages in the Euglenales, all of which presented species with large cpGenomes and more than 110 introns (both *E. gracilis* strains, *S. acuminata*, *C. vesiculosum*) in addition to small cpGenomes with low intron numbers like *M. aenigmatica* (*Bennett & Triemer, 2015*; *Hallick et al., 1993*; *Pombert et al., 2012*; *Wiegert, Bennett & Triemer, 2013*). Although it could be assumed that introns spread independently within the lineages, it was unknown whether a small or a large cpGenome was present when phototrophic euglenoids emerged and how (un)evenly these early introns were distributed in the Eutreptiales.

In the present study *Eutreptiella pomquetensis* (McLachlan, Seguel & Fritz) Marin & Melkonian in *Marin et al. (2003)* was analyzed as a member of the scarcely investigated Eutreptiales. It was originally isolated from shallow, cold, marine habitats and is the only known phototrophic euglenoid with four flagella (*McLachlan, Seguel & Fritz, 1994*). It was classified as an obligate psychrophilic species, which is an unusual characteristic for euglenoids, and worthy of investigation.

The Eutreptiales only consist of two genera, *Eutreptiella* da Cunha, with ten described species, and *Eutreptia* Perty, with eight known species. They are regarded as basal phototrophic euglenoids in aspects of morphology (*Leander, Witek & Farmer, 2001*; *Leedale, 1967*) as well as in molecular analyses and molecular studies combined with morphological characters (*Linton et al., 1999*; *Linton et al., 2000*; *Marin et al., 2003*; *Preisfeld et al., 2001*; *Yamaguchi, Yubuki & Leander, 2012*) and hence of particular interest, where the evolution of euglenoid chloroplasts is reflected upon. The capacity for photosynthesis in euglenoids was found to have originated with the acquisition of chloroplasts by a phagotrophic euglenoid via secondary endocytobiosis of a green alga in a marine environment, which is still unknown (*Gibbs, 1978*, *Gibbs, 1981*). Presumably, the donor was a relative of the partly obligatory psychrophilic genus *Pyramimonas* (*Marin, 2004*; *Turmel et al., 2009*).

Thus, it was our interest to investigate the psychrophilic *Eutreptiella pomquetensis* for two reasons: First, to compare this cpGenome with that of *P. parkeae* (*Turmel et al., 2009*), and the other two Eutreptiales (*Hrdá et al., 2012*; *Wiegert, Bennett & Triemer, 2012*) with regard to genome structure and size, intron number and propagation, and gene content as well as arrangement; second, to diminish the bias in taxon sampling in euglenoid cpGenomic analyses.

# MATERIALS AND METHODS

## Growth, isolation, sequencing and assembly

*Eutreptiella pomquetensis* (McLachlan, Seguel & Fritz) Marin & Melkonian in *Marin et al. (2003)* strain CCMP 1491 cells were grown in modified L1-Si Medium (*Guillard & Hargraves, 1993*) with artificial seawater Sea-Pure (CaribSea, Inc. Fort Pierce, USA) at 2–4 °C with changing 3:3 light:dark cycle using ExoTerra Natural Light PT2190 (Hagen, Holm Germany).

Three-hundred mL of cell culture were harvested by centrifugation and submitted to cell cleaning and chloroplast isolation protocol as described in *Dabbagh & Preisfeld (2016)* with a slight change during sonication of cells. Purified cells were subjected to sonication twice for three seconds with the amplitude set at 50% and a pulse rate of 0.1 s (Bandelin Sonopuls HD 60; Bandelin, Berlin, Germany). The DNA was sequenced with 454 sequencing according to the GS FLX ++ chemistry Rapid Shotgun Library Preparation Method technology (Eurofins Genomics Ebersberg, Germany). In total 60,225 reads were produced in $\frac{1}{4}$ segment of a full run with an average size of 608 bases. Automatic assembly of reads by Eurofins Genomics in Newbler (Roche, Basel, Switzerland) resulted in 668 contigs (N50 contig size was 1,157 bases).

## Annotation of the plastid genome

Using a BLASTN homology search (*Altschul et al., 1990*) the four largest contigs were identified as major parts of the plastid genome, and were subsequently linked by fill-in PCR from the end of each contig using whole genomic DNA. Contig 1 consists of 10,450 number of reads with an average coverage depth of 131.70, contig 2 of 3,918 number of reads with an average coverage depth of 129.90, contig 3 of 3,398 number of reads with an average coverage depth of 112.60. Contig 4 was identified as major part of the chloroplast rRNA operon and showed an average coverage depth of 211.30. The average depth is the mean read coverage and helps to identify repetitive parts of the chloroplast genome. Based on coverage depth of the ribosomal operon components (5S, 16S, 23S) compared to single copy protein coding genes, it appears that at least two copies of the operon are present on the genome. Closing the circle failed in spite of many different approaches using PCR experiments from *rpl*32 to *psa*C and further from each rRNA gene to *psa*C with specifically designed primers. Experiments to close the circle were performed with both a Long Range

PCR Kit (Qiagen GmbH, Hilden, Germany) and a Long Amp Taq DNA Polymerase (New England BioLabs GmbH, Frankfurt am Main, Germany).

The final annotation of the chloroplast sequence was performed with Geneious 9 Pro (version 9.1.3, *Kearse et al., 2012*) with the option to translate the nucleotide sequence in all frames selected, and the "Genetic Code" was identified as "Bacterial".

Protein coding genes were manually aligned in MEGA 7 (*Tamura et al., 2011*) against the nucleotide coding DNA sequences (CDS) from other photosynthetic euglenoids and prasinophyte representatives to determine exon-intron boundaries as well as start and stop of each gene. In all cases, a traditional methionine (ATG) start codon was preferred. CDS was verified by BLASTx, "Genetic code" set at "Bacteria and Archaea (11)" and Emboss Sixpack Sequence translation (EMBL- EBI 2015) "Codon Table" set at "Bacterial" and added to the annotation. The introns within protein coding genes were analyzed for the presence of potential twintrons as described in *Bennett & Triemer (2015)*. This analysis was modified such that the 3′ motifs were established using a Python script instead of a manual search. The script browsed the homologous external introns for the conserved 3′ motifs (*abcdef* (3–8 nucleotides) $f'e'd'A^*c'b'a'$ (four nucleotides)). Afterwards, all 51 introns were searched for the conserved 5′ insertion sequence GUGYG. RNA secondary structure for group II introns was created by RNA folding via Mfold web server using default settings (*Zuker, 2003*), manually optimized and illustrated with the PseudoViewer web application (*Byun & Han, 2006*). For *roa*A a pairwise sequence comparison of the amino acid sequence of *E. gracilis* with a putative amino acid sequence of *Etl. pomquetensis* was performed using Exonerate 2.4.0 by *Slater & Birney (2005)* to reveal intron boundaries and start/ stop of the searched gene.

tRNAscan-SE 1.21 (*Schattner, Brooks & Lowe, 2005*), with the default settings and the source given as mito/chloroplast, was used to identify tRNAs. Uncharacterized open reading frames (ORFs) were identified with ORF finder within Geneious, with the genetic code set to bacterial. Only ORFs which were at least 300 bp, did not overlap with the coding region of another gene, and lacked BLASTP evidence (default settings) for being a previously identified chloroplast protein-coding gene were included in the annotation. ORFs were named according to the number of amino acids in the coding region. To evaluate the proportion of short repeated sequences the variable number of tandem repeats was scanned with the online version of REPuter (*Kurtz et al., 2001*) under the same settings as described in *Bennett & Triemer (2015)* and with Tandem Repeats Finder, with the option "Basic", using default parameters (*Benson, 1999*).

The start/stop areas of the 16S and 23S rRNA genes were identified using RNAmmer 1.2 (*Lagesen et al., 2007*), with "Bacteria" chosen as the sequence kingdom of origin. The 5S rRNA start/stop regions were identified using Rfam 12.1 Sequence Search (*Burge et al., 2013*). The number of rRNA operons flanked by the protein-coding genes *rpo*C2 and *psb*A were confirmed using PCR. One further rRNA operon was identified by PCR experiments next to the protein coding gene *rpl*32 by long range PCR. To verify the exact sequence a Long Range PCR was performed with primers (forward 5′ -AGAGTTTGATCCTGGCTCAG- 3′; reverse 5′-TGCTTCCATACACTTTTACGCATA- 3′) from the beginning of the 16S to the *rpl*32 gene. Primers were created manually by Primer3Plus (*Untergasser et al., 2012*) based on the nucleotide sequence. The PCR product (5,080 bp) was purified and used as DNA matrix for further PCRs to determine the sequence of the rRNA genes and the noncoding regions in between. The number of rRNA operons next to the *rpl*32 gene was performed using long-range PCR. The long-range PCR approach to measure the copies of the RNA operon yielded only one product.

Synteny between the cpGenomes of all three sequenced Eutreptiales was determined using Mauve (*Darling et al., 2004*), as a plugin for Geneious, with the alignment algorithm set as progressive Mauve. Each genome was displayed as a linear sequence with blocks representing a homologous gene cluster. In the Mauve alignment the repeat regions of rRNA were not included because Mauve will not align repeat regions which have multiple matches on both genomes. The circular genome map was created using GenomeVx (*Conant & Wolfe, 2008*).

# RESULTS AND DISCUSSION

## General genome analyses

The cpGenome of *Etl. pomquetensis* is presented as an incomplete circle, because attempts to close the gap between the 16S rRNA gene and the protein coding gene *psa*C were unsuccessful, even with long range approaches. Thus, the cpGenome contained at least 130,561 bp, which is twice the size of *Eutreptiella gymnastica* with 67,622 bp (*Hrdá et al., 2012*) and *Eutreptia viridis* with 65,523 bp (*Wiegert, Bennett & Triemer, 2012*). The new cpGenome resembled the members of the Euglenales *E. gracilis var. bacillaris* (132,034 bp) and *C. vesiculosum* (128,892 bp, Table S1) in size. The content of genes was similar to those of other phototrophic euglenoids and reduced as compared to *P. parkeae* (*Turmel et al., 2009*) or *Ostreococcus tauri* (*Robbens et al., 2007*). The organization of the whole genome, however, resembled those of higher plants and algae (*Cattolico et al., 2008*; *Lemieux, Otis & Turmel, 2007*; *Ravi et al., 2008*; *Robbens et al., 2007*; *Turmel et al., 2009*) more than other euglenoids. The genome was composed of a large single copy region (LSC 80,941 bp), a small single copy region (SSC 39,856 bp) and two inverted repeats (IR) containing the rRNA genes in a way similar to *O. tauri,* but different in gene content (Figs. 1A & 1B). In the cpGenome of *P. parkeae*, the putative chloroplast donor for euglenoids, the organization is very much alike, but lacks the 5S rRNA in both inverted repeats. However, the possibility of non-recognition of the sequence as described by *Turmel et al. (2009)* still has to be considered. The fact that one operon was localized on the positive and one on the negative strand points at another similarity between the green algae *P. parkeae, O.tauri* and *Etl. pomquetensi* s. In the close relative *Etl. gymnastica,* the rRNA operon consisted of two incomplete copies, without a 5S rRNA, as in *P. parkeae*, but additionally one operon was divided into two parts separated by parts of the LSC (*Hrdá et al., 2012*, Fig. 1C). The G+C base composition of 35.1% again resembled that of *Etl. gymnastica* and *P. parkeae* and was higher than that of *Et. viridis* with 28.6% (Table S1).

## Analysis of gene content and arrangement

In total, 94 genes were identified and annotated in the cpGenome of *Etl. pomquetensis*, including 60 protein coding genes, two complete copies of the rRNA operon and 28 tRNAs

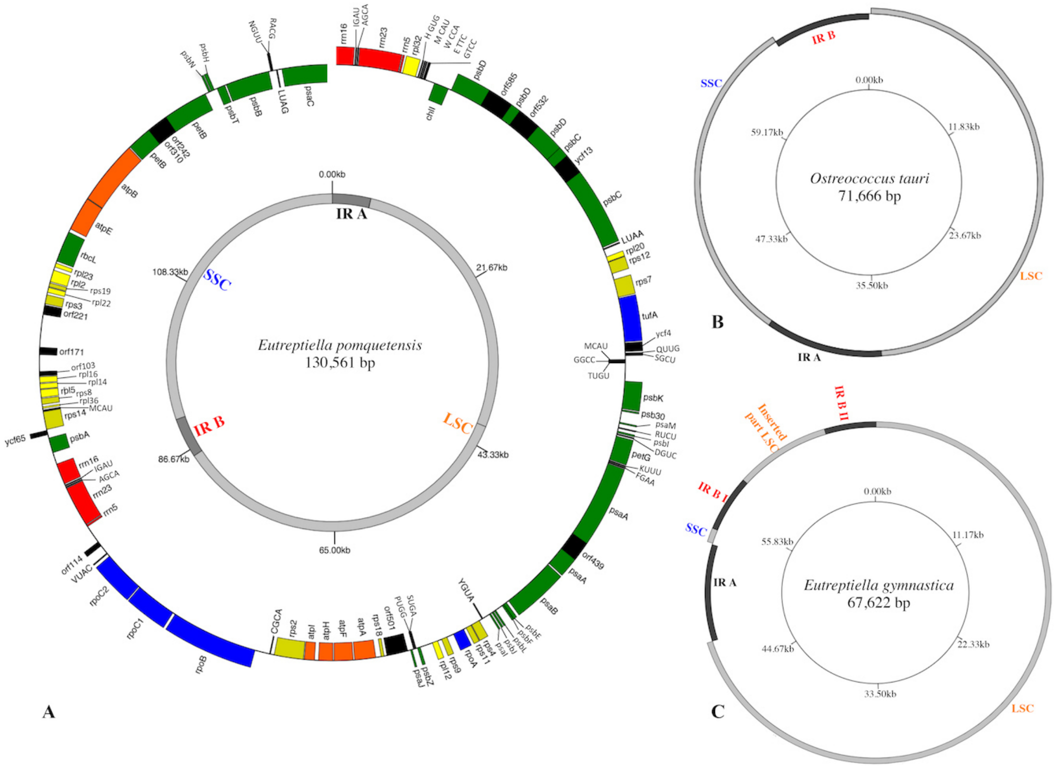


**Figure 1** **Gene maps of chloroplast genomes.** (A) Map of the plastid genome of *Eutreptiella pomquetensis*. Boxes of different colors represent genes of similar functional groups: red, ribosomal rRNAs; green, photosystem/photosynthesis genes; yellow, ribosomal proteins (*rpl*, *rps*); orange, *atp* genes; blue: transcription/translation- related genes (*rpo*, *tuf*A); black, conserved hypothetical proteins (*ycf* ), open reading frames (ORF), tRNAs. Boxes are proportional to their sequence length. Outer ring: Genes on the outside of the circle are considered on the positive strand, genes inside the circle on the negative strand. Inner circle shows the large single copy region (LSC) and short single copy region (SSC) in light grey and inverted repeats (IR) in dark grey. (B–C) Simplified maps of the plastid genomes of *Ostreococcus tauri* (B) and *Eutreptiella gymnastica* (C) to demonstrate similarities and differences of genome structures. Copies of the IR sequences are represented in dark grey and LSC and SSC in light grey.

(Fig. 1A). Alignments and analysis of protein coding genes indicated that the coding regions were more similar to those of *P. parkeae* than to those of the *Euglena* clade. For example, a pairwise comparison between *psb*D coding regions from *P. parkeae* and *Etl. pomquetensis* pointed out an 84.4% identity at the nucleotide level, whereas the same region from *E. gracilis* and *Etl. pomquetensis* showed only an 80.5% identity making the resemblance of *Etl. pomquetensis* to *P. parkeae* more apparent. Traditional methionine start codons (ATG) were found for each protein coding gene, except *rpo*A, where an alternative start codon (ATA) was accepted. Four protein coding genes were annotated with alternative start codons in *Etl. gymnastica* and *Et. viridis* (Table 1; *Hrdá et al., 2012*; *Wiegert, Bennett & Triemer, 2012*) and three in *P. parkeae* (*Turmel et al., 2009*). The number of the protein coding genes (60) was most similar to *Etl. gymnastica* (59), where *psa*I was missing. *Et. viridis* lacked *psa*M, *ycf*12 (*psb*30) and *ycf*65 and hence counted only 57 protein coding genes. No *mat*5 or *mat*2 genes have been identified in *Etl. pomquetensis*, but *mat*1 (*ycf*13) was detected, as was

**Table 1 Alternative start codons usage in protein coding genes of cpGenomes of Eutreptiales and *Pyramimonas parkeae* (Chlorophyta).**

| | Alternative start codons: | | | |
|---|---|---|---|---|
| | ***Etl.pomquetensis*** | ***Etl. gymnastica*** | ***Et. viridis*** | ***P. parkeae*** |
| Total number | 1 | 4 | 4 | 3 |
| Gene/start | *rpo*A (ATA) | *rpo*A (TTG)<br>*psb*C (TAT)<br>*ycf*13 (GTG)<br>*atp*F (TTG) | *psa*I (ATT)<br>*rps*11 (ATT)<br>*atp*E (ATT)<br>*pet*B (GTG) | rps11 (GTG)<br>*rpo*A (GTG)<br>*rps*18 (GTG) |

expected from results in other Eutreptiales (*Hrdá et al., 2012*; *Wiegert, Bennett & Triemer, 2012*). Just like *Et. viridis* (*Wiegert, Bennett & Triemer, 2012*), *Etl. pomquetensis* also lacked the common land plant chloroplast genes *rpl*33, *inf* A, *clp*P, *frx*B, *ndh*A-K, *pet*A, *pet*D, *psb*M, *rps*15 and *rps*16.

Progressive Mauve was used to analyze related chloroplast genomes (*Darling et al., 2004*). A comparison of *Etl. pomquetensis* and *Etl. gymnastica* gene content and arrangement identified 10 conserved gene clusters (Fig. 2, Table 2). Although gene content was similar in the two studied *Eutreptiella* species, the gene clusters showed significant rearrangements in position and strand orientation between *Etl. gymnastica* and *Etl. pomquetensis*. Block I was the largest in *Etl. pomquetensis*, included 18 genes, and was more than 19 kb long. The clusters themselves showed that extensive rearrangements occurred between *Etl. gymnastica* and *Etl. pomquetensis*. This lack of synteny was surprising, because high intrageneric variability between other taxa had not been noted so far. For example, a comparison between *M. aenigmatica* and *M. parapyrum* or *E. gracilis* and *E. viridis* cpGenomes revealed only one and two blocks, respectively. But, although *Etl. gymnastica* and *Etl. pomquetensis* are described as belonging to one genus, the evolutionary distance between euglenoid taxa is usually relatively high and makes differences probable. On the other hand, *Etl. pomquetensis* lives under psychrophilic conditions, whereas *Etl. gymnastica* lives under moderate marine conditions, which means that the environmental pressure is varying.

The noted difference in gene density between *Etl. pomquetensis* and *Etl. gymnastica* was not only due to an increase of introns from seven introns in *Etl. gymnastica* (total amount of intron space 6,893 bp) to 51 introns in *Etl. pomquetensis* (total amount of intron space 52,999 bp), but additionally to an increased intergenic space in *Etl. pomquetensis*. The intergenic space of *Etl. pomquetensis* comprised more than 23 kb, which was more than twice in that of *Etl. gymnastica*. While most of the blocks in *Etl. gymnastica* were quite compact with little intergenic or intron space in blocks C, E and G, all of the identified clusters showed heavily fragmented blocks in *Etl. pomquetensis*, except A and B (Fig. 2).

A second Mauve analysis of *Etl. pomquetensis* and the two other basal phototrophic Eutreptiales *Et. viridis* and *Etl. gymnastica* identified 14 conserved gene clusters (Fig. S1). The gene order within the clusters was mostly conserved and equal to the ten clusters found in the previous analysis. However, four gene clusters were further divided into two clusters each (Table 2, bar in blocks C, H, I, J).

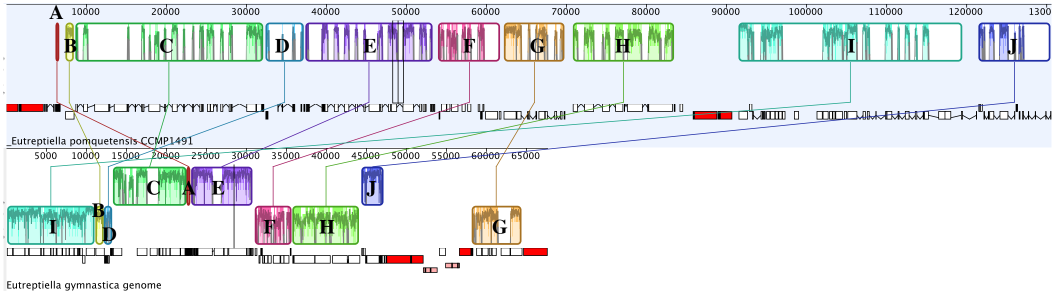


**Figure 2 Progressive Mauve analysis comparing the cpGenomes of *Etl. pomquetensis* and *Etl. gymnastica*.** Each box represents a cluster of homologous genes with *Eutreptiella pomquetensis* as the reference genome. Like blocks are labelled by letters A–J. See Table 4 for a list of genes contained in each block. In the Mauve alignment the repeat regions of rRNA were not included, because Mauve will not align repeat regions, which have multiple matches on both genomes.

**Table 2 Gene clusters resulting from Progressive Mauve cpGenome analysis of the two *Eutreptiella* species.** Gene clusters (blocks) are labelled with letters (A–J) and relevant genes listed. Bars in blocks C, H, I, J mark positions of a second Progressive Mauve analysis of the three Eutreptiales, where blocks are divided (see Fig. S1).

| Block | Gene Clusters |
|---|---|
| A | tRNA-His, tRNA-Met, tRNA-Trp, tRNA-Glu, tRNA-Gly |
| B | *chl*I |
| C | *psb*D *psb*C tRNA-Leu / *rpl*20 *rps*12 *rps*7 *tuf*A ycf4 tRNA-Gln tRNA-Ser |
| D | tRNA-Arg *psa*M *psb*30 (syn. *ycf*12) psbK tRNA-Thr tRNA-Gly tRNA-Met |
| E | *psb* I tRNA-Asp *pet* G tRNA-Lys tRNA-Phe *psa*A *psa*B *psb*E *psb*F *psb*L *psbJ* |
| F | *rps*18 *psa*J tRNA-Pro tRNA-Ser *psb*Z *rpl*12 *rps*9 *rpo*A *rps*11 *rps*4 tRNA-Tyr |
| G | tRNA-Cys *rps*2 *atp*I *atp*H *atp*F *atp*A |
| H | tRNA-Val / *rpo*C2 *rpo*C1 *rpo*B |
| I | *pet*B *atp*B *atp*E / *rbc*L *rpl*23 *rpl*2 *rps*19 *rpl*22 *rps*3 *rpl*16 *rpl*14 *rpl*5 *rps*8 *rpl*36 tRNA-Met *rps*14 ycf65 *psb*A |
| J | *psb*N *psb*H *psb*T *psb*B tRNA-Asn tRNA-Arg tRNA-Leu / *psa*C |

Three additional Mauve analyses using *Etl. pomquetensis* identified 31 clusters with *P. parkeae*, 26 with *P. provasolii*, and 21 with *O. tauri* (Fig. S2). A comparison of the Mauve analyses found more homologous regions between *Etl. pomquetensis* and the other Eutreptiales than with the prasinophytes (the group containing the putative chloroplast donor). As the phototrophic euglenoids have a reduced amount of protein coding genes in contrast to the green algae, this high number of clusters was expected.

## Open reading frames

Ten uncharacterized open reading frames (ORFs) were found in *Etl. pomquetensis*. A BLASTp analysis was performed against the NCBI nonredundant protein sequences (nr) database to determine whether any of the ORFs had functional similarity to previously sequenced genes (Table 3). The *psb*D gene of *Etl. pomquetensis* contained two ORFs (*orf*585 and *orf*532). The intron encoded *orf*585 of *Etl. pomquetensis psb*D I2 shared strong similarity with the *orf*583 of *atp*B I1 in the chloroplast genome of *Pycnococcus provasolii* (Turmel et

**Table 3 Open Reading Frames.** BLASTp analysis of ten uncharacterized ORFs in the *Eutreptiella pomquetensis* cpGenome against NCBI nonredundant protein sequences (nr) database. For each ORF the best match is reported.

| ORF | Accession number | Best BLASTp match | | |
|---|---|---|---|---|
| | | Organism | Product | *E*-value |
| 585 | YP_002600812.1 | *Pycnococcus provasolii* | putative reverse transcriptase and intron maturase | 0.0 |
| 532 | WP_041039849.1 | *Tolypothrix campylonemoides* | group II intron reverse transcriptase/maturase | 3*e* − 57 |
| 439 | YP_009306333.1 | *Caulerpa cliftonii* | hypothetical protein | 2*e* − 39 |
| 501 | WP_050045085.1 | *Tolypothrix bouteillei* | group II intron reverse transcriptase/maturase | 4*e* − 65 |
| 114 | – | – | no significant similarity found | – |
| 310 | BAM65725.1 | *Helminthostachys zeylanica* | maturase K | 9*e* − 14 |
| 242 | WP_061793822.1 | *Bacillus firmus* | hypothetical protein | 0.14 |
| 221 | AOC61650.1 | *Gloeotilopsis planctonica* | putative reverse transcriptase and intron maturase | 5*e* − 35 |
| 171[a] | AOC61481.1 | *Gloeotilopsis sarcinoidea* | putative reverse transcriptase and intron maturase | 3*e* − 12 |
| 103[a] | AOC61650.1 | *Gloeotilopsis planctonica* | putative reverse transcriptase and intron maturase | 7*e* − 09 |

**Notes.**
[a] maybe *roa*A.

*al., 2009*), with an *e*-value of 0.0. *Turmel et al. (2009)* determined that the *Pycnococcus* and *Ostreococcus* intron ORFs share strong similarity with each other, and for example, also with *mat*4 in *Euglena myxocylindracea* (*Turmel et al., 2009*). The open reading frames *orf*171 and *orf*103 next to the *rpl*16 gene showed weak similarity to the *roa*A gene annotated in some Euglenales chloroplast genomes. However, in either case the best match is reported for putative reverse transcriptase and intron maturase. Further, exonerate 2.4.0 (*Slater & Birney, 2005*) and a manual alignment were performed to evaluate if the two ORFs were part of the *roa*A gene. Neither of these methods yielded clear results, and no exact exon-intron boundaries or start/ stop regions could be identified. Additionally, RT-PCR experiments for detecting a putative intron between *orf*103 and *orf*171 failed, indicating that these ORFs may not have a true function *in vivo*.

There is no evidence of a VNTR (variable number of tandem repeat) sequence, though this could be a result of our inability to circularize the genome.

## Intron sequence similarity

Twenty-three out of the 60 protein-coding genes contained one or more introns, resulting in a total of 51 introns with likely twintrons measured as one insertion site. *psa*A contained the highest count with six introns (Table S1). The number of introns revealed, is twice as high as in *Et. viridis* (27), nearly eight times higher than found in *Etl. gymnastica* (7), and consequently constitutes the highest intron number known in the Eutreptiales (*Hrdá et al., 2012*; *Pombert et al., 2012*; *Wiegert, Bennett & Triemer, 2012*). Upon closer inspection of the intron sequences, we discovered 90% pairwise identities in introns of different genes in *Etl. pomquetensis*.

Therefore, and to gather information on the relatedness of the introns in basal euglenoids, we aligned all intron sequences and detected 28 introns (773–1,578 bp, Table S2 marked bold) in *Etl. pomquetensis* with pairwise identities of 87.4% and identical 5′-GTGCG boundaries typical for group II introns. Since group II introns in euglenoids are short for

**Table 4 Features of the presumed ancestral *psb*C twintron in all cpGenomes of phototrophic euglenoids.**

| | Intron containing *mat* 1 | *psb*C total Intron length (bp) | length *mat* 1 (bp) | *psb*C intron length without *mat* 1 (bp) |
|---|---|---|---|---|
| *E. gracilis* | I4 | 1,605 | 1,377 | 228 |
| *E. gracilis var. bacillaris* | I4 | 1,605 | 1,377 | 228 |
| *E. viridis* | I2 | 1,612 | 1,359 | 258 |
| *E. viridis epitype* | I2 | 1,617 | 1,359 | 258 |
| *E. mutabilis* | I2 | 3,406 | 3,149 | 257 |
| *Era. anabaena* | I2 | 1,945 | 1,683 | 262 |
| *M. parapyrum* | I2 | 1,613 | 1,338 | 275 |
| *M. aenigmatica* | I2 | 1,618 | 1,389 | 229 |
| *Cr. skujae* | I3 | 1,629 | 1,362 | 267 |
| *S. acuminata* | I2 | 1,686 | 1,371 | 315 |
| *T. volvocina* | I2 | 2,534 | 1,672 | 862 |
| *C. vesiculosum* | [a] | 2,742 | | |
| *Efs. proxima* | I3 | 3,349 | 2,669 | 680 |
| *P. orbicularis* | I1 | 1,716 | 1,533 | 183 |
| *Et. viridis* | I1 | 4,350 | 3,609 | 741 |
| *Etl. gymnastica* | I1 | 1,778 | 1,137 | 641 |
| *Etl. pomquetensis* | I1 | 2,580 | 1,389 | 1,191 |

**Notes.**

[a] annotation mistake.

group II intron membership and usually do not show high sequence similarities, except in bounding regions, the strongly conserved GAAA terminal loop and portions of the domain V stem and, if present, in maturases (*Michel & Ferat, 1995*; *Thompson et al., 1997*) it was surprising to discover pairwise identities of about 90% in introns of different genes in *Etl. pomquetensis*. Moreover, 3′ boundaries always showed matching ACGTTCAT motifs (except for *pet*G I1 and *psa*C I2) with the presumed "branch-point" *A for splicing at position eight in domain VI, where the first transesterification takes place (*Lambowitz & Belfort, 2015*). The last two nucleotides AY represent the typical conserved ending for group II-introns (*Lambowitz & Belfort, 2015*). As expected, domain V, known to play a catalytic role in intron excision, showed a highly conserved secondary structure (*Kelchner, 2002*; *Michel & Ferat, 1995*; *Thompson et al., 1997*; *Toor, Hausner & Zimmerly, 2001*). The 28 introns scrutinized, except for *pet*G I1 and *psa*C I2 (Table S2 marked bold), showed a highly conserved domain V with 24 out of 34 nucleotides identical. Beside the fact that three base pairs (5′- …AGC …GUU…-3′) near the base of the stem were completely identical (Fig. S3), the secondary structure was unambiguously the same as the secondary structure of group IIB introns predicted by *Kelchner (2002)*. Also of interest was that more than half out of the 51 nucleotides forming the stem and loop of domain VI were identical and resulted in the same secondary structure (Fig. S3).

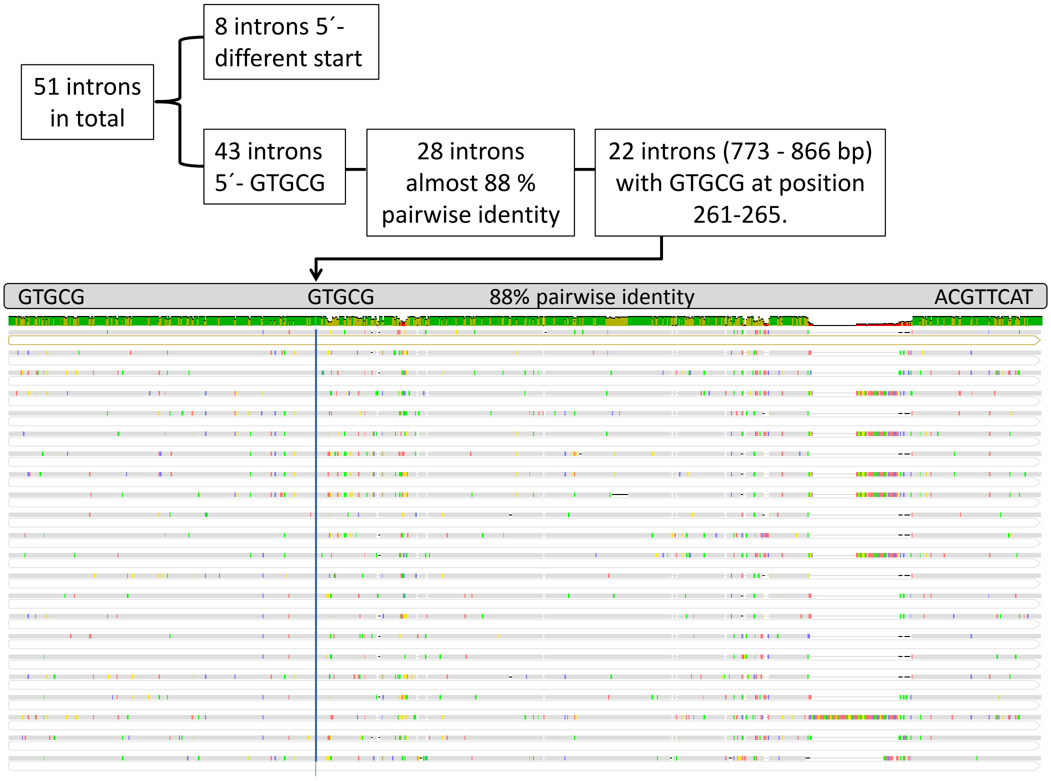


**Figure 3 Alignment of group II introns.** Intron identity according to boundaries and position of additional GTGCG (blue line). Geneious nucleotide alignment with absolute pairwise identities (green in first line, grey in alignment, different nucleotides in colored bars) of 21 introns in *Etl. pomquetensis* in various genes. Introns top down: *atp*B I2, *atp*B I3, *atp*B I4, *atp*E I2, *atp*H I1, *psa*A I1, *psa*A I2, *psa*A I3, *psa*A I4, *psa*A I6, *psa*B I2, *psa*C I3, *psb*B I2, *psb*C I3, *psb*C I5, *psb*D I1, *psb*D I5, *rbc*L I1, *rpl*32 I1, *rpo*B I1, *rps*7 I1, *rps*12 I1.

Twenty- two of these introns (773–866 bp) in *Etl. pomquetensis* additionally showed the same GTGCG motif at positions nt 261 to 265 upstream from base one of the intron, with pairwise identities of 88% (Fig. 3, Table S2 highlighted in gray).

We assume that all of these 28 introns of *Etl. pomquetensis* with high pairwise identity were closely related and arose from a single ancestor proliferating via retrotransposition and moved horizontally into DNA target sequences, which resembled the homing site. According to *Lambowitz & Zimmerly (2011)* and *Lambowitz & Belfort (2015)*, retrotransposition to ectopic sites plays a major role in intron dissemination to novel locations, so that the many and very similar introns in *Etl. pomquetensis* could be explained.

## Possible proliferation of group II introns

Still the question remained of how these introns could be spliced without an ORF including maturase activity in domain IV. One possibility was that they rely on trans-acting RNAs or proteins with two feasible splicing mechanisms: (1) The introns of *Etl. pomquetensis* used host encoded proteins to promote splicing, reverse splicing and mobility, which is typical for most mitochondrial and plant chloroplast group II introns (*Lambowitz & Zimmerly, 2011*).

*Chlamydomonas reinhardtii* even utilized nuclear-encoded maturases for splicing of the trans-spliced group II introns (Merendino et al., 2006).

(2) All these introns could be spliced by a single IEP (Intron-Encoded Protein) that could either be free-standing or located in a functional intron. This would provide an accessible splicing apparatus and allow all but one intron to lose its own IEP (Dai & Zimmerly, 2003; Lambowitz & Belfort, 2015; Lambowitz & Zimmerly, 2011). Brouard et al. (2016) assumed that the freestanding *orf1311* in *Oedocladium* (Chlorophyceae), with an intron encoded maturase, could function as promoter for splicing the ORF-less group II introns. Turmel, Otis & Lemieux (2016) detected introns in *G. planctonica* without ORFs, which may reflect an evolutionary pressure for a smaller and more compact intron structure enabling increased efficiency of splicing and mobility, when maturase activity is provided from elsewhere. Furthermore, it might be assumed that an early event in the *Etl. pomquetensis* cpGenome was the deletion of an intron encoded ORF, which appeared to have occurred prior to the spreading of introns across the genome and that other group II introns with encoded IEPs or freestanding ORFs acted *in trans* to promote splicing and mobility of ORF-free introns (Doetsch, Thompson & Hallick, 1998). To gain information about DNA target sites, which the introns of *Etl. pomquetensis* use for retrotransposition, we checked the insertion sites and the sequences of flanking exons. The exon nucleotides at the 5′-insertion site of the intron did not show any similarity, which might be due to a not strictly controlled transposition/ retrotransposition processes (Pombert et al., 2012), thus helping with random intron invasion all over the genome, and on both strands. The only conspicuous DNA target site the 28 homologous introns with high sequence similarity used for reverse splicing was a pyrimidine base, which represented the first nucleotide of the following exon (except for *atp*E exon 3, I2). The gene *psa*A contained the most of these introns, and five of six introns contained high similarity.

A search for related introns in *Etl. gymnastica*, *Et. viridis* and *P. parkeae* and all other euglenoid cpGenomes did not reveal any sequential or positional homology. Insertion sites found in *Etl. pomquetensis* were unique to that taxon.

The highest pairwise identity of introns was found in *E. gracilis var. bacillaris* with 56.7%, but only for three small 97 bp long group III introns (*rps*16 I1, *rpo*C1 I7 and *rps19* I2). Also, outside of euglenoid chloroplast introns, very few species showed high pairwise similarity. For instance, Brouard et al. (2016) found six group IIA introns in the chlorophyte *Oedocladium carolinianum* with high levels of nucleotide identities, which displayed over 80% pairwise identity. As well Turmel, Otis & Lemieux (2016) found several group II introns with high nucleotide identities also at various insertion sites, but only in small numbers. The introns of *ycf*3 and *psb*H in *Gloeotilopsis sarcinoides* were 85.6% identical. To our knowledge, *Etl. pomquetensis* is the first organism with more than 50 introns within protein coding genes and half of those sharing a pairwise identity of 90%.

## Lack of group III introns in the genome

The second peculiarity in *Etl. pomquetensis* is the absence of group III introns. Group III introns are believed to be the descendants of group II introns which only retained domains DI and DVI (Christopher & Hallick, 1989). The 5′ -boundaries are more variable than in group II introns, but most group III introns have a U at position 2 and a G at position 5. Most of them are of dyad symmetry near the 3′ -end similar to domain VI of group II introns. The motif driving the symmetry follows *abcdef* (3–8)$f'e'd'A^*c'b'a'$ (*Drager & Hallick, 1993*). The 3′ sequence of group II and III introns are variable, although the branch-point $A^*$ is usually at position eight, sometimes at seven, and occasionally at position nine. Interestingly, none of the 51 identified introns of *Etl. pomquetensis* complied with the typically confined group III intron size of 91–120 nucleotides (*Christopher & Hallick, 1989*; *Copertino & Hallick, 1993*; *Doetsch, Thompson & Hallick, 1998*; *Drager & Hallick, 1993*). Underpinning these findings, 43 of 51 introns started with a typical group II 5′-GTGCG (Table S2, start marked bold) and even the smallest intron was over 300 bp long (*rpo*C1 I1 356 bp). Furthermore, the intron size was even larger than group III twintrons (group III introns within group III introns), which were found in the chloroplast genome of *E. gracilis* (*Copertino, Shigeoka & Hallick, 1992*; *Copertino et al., 1994*). The smallest introns of *Etl. gymnastica* and *Et. viridis* were *rpo*B I1 with 179 bp (re-analyses of data from (*Hrdá et al., 2012*; *Wiegert, Bennett & Triemer, 2012*) and 156 bp, respectively, and these were larger than group III introns. Hence, we assumed that group III introns probably evolved after *Etl. pomquetensis* diverged. Secondary the structure of domain V and VI of *rpo*B I1 in *Etl. gymnastica* and *Et. viridis* was recognizable when some mismatches were allowed in the analyses (Fig. S4). Degeneration and mutation of group II introns in euglenoids have been described before and are known to impact secondary structure elements. Even domain V tolerates a surprising number of mismatches (*Michel & Ferat, 1995*). To our present knowledge, such introns best resemble mini-group II introns, which lack different domains (*Doetsch, Thompson & Hallick, 1998*). Under the presumption that *rpo*B I1 of *Etl. gymnastica* and *Et. viridis* are mini group II introns and not group III introns, we assume that group III introns evolved probably within the Euglenales after fresh-water and brackish environments became accessible together with warmer temperatures. The impact of the environmental medium could have been a driving force on degenerating group II introns. The change from group II intron to group III introns was observed in the *psb*C intron containing *mat*1 (*ycf*13). It is clearly a group II intron/ twintron in all Eutreptiales, but a group III twintron in *E. gracilis* with an open reading frame (*ycf*13, *mat*1) within the internal group III intron (*Copertino et al., 1994*; Table 4). *Copertino et al. (1994)* proposed that *mat*1 may be involved in group III intron metabolism and is required for group III intron excision and/or mobility in *Euglena* and *Astasia*. The ORF of *Euglena gracilis psb*C I4 has detectable similarity to the RT domain of group II intron ORFs, although it lacks characteristics of functional RT activity (*Copertino et al., 1994*; *Doetsch, Thompson & Hallick, 1998*; *Mohr, Perlman & Lambowitz, 1993*).

Based on the greater length of the *psb*C intron in *Etl. pomquetensis*, and a typical group II intron 5′ -boundary, it seems likely that the *psb*C intron is instead a group II intron/twintron. All three Eutreptiales have a *psbC* intron including *mat*1 (*ycf*13) that is at least three times larger than the group III twintron (I4) including *mat*1 of *E. gracilis* (Table 4). These findings, and the fact that *E. gracilis* contained a group II -type maturase in a group III twintron (*Doetsch, Thompson & Hallick, 1998*; *Mohr, Perlman & Lambowitz, 1993*), underpin the possibility that group II introns evolved first in basally branching euglenoid species. Subsequently, they degenerated by loss of different domains (in more derived species) to group III introns, containing only DI-like and DVI-like structures (*Doetsch, Thompson & Hallick, 1998*; *Lambowitz & Belfort, 2015*). This finding is also supported by identification of two maturase encoded introns and their predicted secondary structure models in *Lepocinclis buetschlii* by *Doetsch, Thompson & Hallick (1998)*. The authors interpreted these introns as group II/group III intermediates just in the process of losing group II intron domains and they were designated as mini-group-II introns.

Summarizing, we presume that group II introns appeared first in an intron-less ancestral genome and gave rise to group III introns and from there on degeneration went on independently in different lineages. Further on, either the *Etl. pomquetensis* group II intron *mat* 1 or another intron encoded protein (IEP) act *in trans* to promote splicing and mobility of ORF-less introns.

## Intron trends in Euglenoids

In their characterization of Euglenaceae, *Bennett & Triemer (2015)* noted that all Euglenaceae, but no Eutreptiales, contained an intron or twintron in *pet*B (I1) and that this intron/twintron may be a synapomorphy for at least the Euglenaceae. *Kasiborski, Bennett & Linton (2016)* identified a homologous intron/twintron within *pet*B I1 of *P. orbicularis* and discussed this intron/ twintron as a putative synapomorphy for the order Euglenales. However, in the cpGenome of *Etl. pomquetensis* two introns were detected in *pet*B. The first was found at the identical insertion site, but nearly two times larger than that of *E. gracilis* strain Z and five times larger than that of *P. orbicularis*. All *pet*B I1 introns started with a typical group II 5′ -GUGYG (*P. orbicularis* re-analysis, Table S1). This means, a group II intron in *pet*B could neither be a synapomorphy for the Euglenales, nor for the Euglenaceae, but evidently evolved at least in *Eutreptiella*.

## Twintron analysis

All 51 external introns were investigated for the presence of potential twintrons using a Python script, which searched for the conserved 3′ motif of group II and group III introns reported in *Copertino & Hallick (1993)*. The search resulted in 28 external introns which contained at least one 3′ motif (see GenBank accession). Sixteen of the 28 introns contained four kinds of repeated 3′ motifs (Table S2, indicated by number of asterisks). Additionally, four potential group II twintrons were found (*rpo*B I1, *rps*2 I2, *psb*C I2, *psb*D I4, added to annotation) with only one 3′ motif and only one 5′ -GUGYG prior to the identified 3′ motif. Two of these potential group II twintrons (*rpo*B I1 and *psb*D I4) were those which share strong nucleotide identity with half of the introns detected in *Etl. pomquetensis*. We assume that all 28 introns (Table S2 marked bold except for *pet*G I1 and *psa*C I2) with equal intron organization (5′ motif GTGCG, 3′ motif ACGTTCAT and further GTGCG at nt 261-265) are potential twintrons with an external and internal group II intron (Fig. 4A). Secondary structure analysis of domain V and VI of the potential internal introns of *rpo*B I1and *psb*D I4 in *Etl. pomquetensis* showed recognizable counterparts, when mismatches were allowed in the analyses (Fig. S5). For the potential internal introns in *rpo*B I1 and *psb*D I4 the conserved three base pairs (5′- ... AGC ... −3′) near the base of the stem of domain

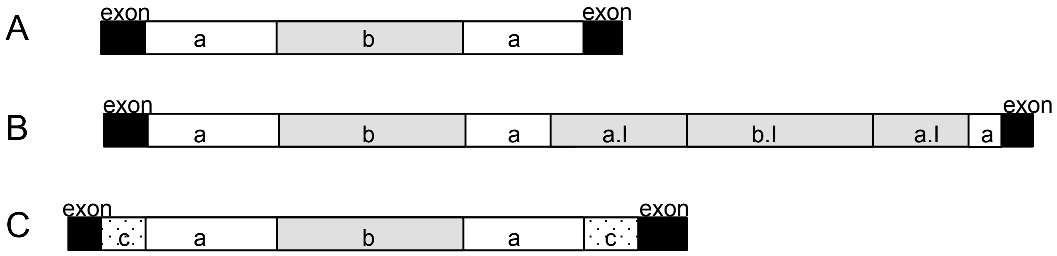


**Figure 4 Analysis of potential twintrons with high sequence similarity.** (A) Highly conserved introns are shown. (B) Structure of the *pet*G I1 complex twintron. (C) Structure of *psa*C I2. Black boxes represent exons. White boxes (a) are external introns of twintrons, white dotted boxes (c) are external introns of complex twintrons. Grey boxes (b, a.I, b.I) represent internal introns, whereby a.I showed high sequence similarity to external intron a and b.I to internal intron b.

V were detectable, but the secondary structure showed a slightly altered terminal loop and no branch-point *A** was detectable in domain VI (*Michel & Ferat, 1995*; *Thompson et al., 1997*). Since the Phyton script only detects an unaltered conserved 3′ motif, only two of the close related introns have been detected as potential twintrons. This underpins several statements, that group II introns of phototrophic euglenoids are highly degenerated and persistent to detailed analysis (*Michel & Ferat, 1995*; *Mohr, Ghanem & Lambowitz, 2010*). Two introns, *psa*C I2 and *petG* I1, out of the 28 potential twintrons with high sequence similarity were significantly larger and thus investigated for the presence of potential complex twintrons.

*psa*C I2 analysis: The intron *psa*C I2 of *Etl. pomquetensis* was 1,294 bp long and by this more than 400 bp longer than the average. The nucleotide sequence alignment of all 28 introns (Table S2) was remarkably well conserved. It showed that *psa*C I2 is a complex twintron with an external intron interrupted by the same potential internal twintron as all the others (Figs. 4A and 4C).

The potential internal twintron (825 bp) shared 88% pairwise identity with the other 27 potential twintrons (Table S2). It is located 281 bp downstream of the external 5′ splice site. Comparing the secondary structure of domain V of the external intron a of the internal twintron (Fig. 4C) with the other highly conserved twintrons (Fig. 4A) resulted in identical stems and loops with only two out of 34 nucleotides differing (Fig. S3).

A BLASTN search for the external intron of *psa*C I2 (Fig. 4C dotted intron c) revealed weak similarity with *psb*C I2 (containing a still unspecified maturase) of *Etl. gymnastica*. Secondary structure analysis of domains V and VI of *psb*C I2 from *Etl. gymnastica*, realigned by *Dabbagh & Preisfeld (2016)*, and the external intron of *psa*C I2 in *Etl. pomquetensis* revealed highly conserved structures of domains V (Fig. S6). They only differed in six nucleotides and contained the AGC motif near the base of the stem from the 5′-boundary (*Thompson et al., 1997*). We presume that the external intron of *psa*C I2 in *Etl. pomquetensis* (Fig. 4C dotted intron c) is closely related to and arose from the same ancestral intron as *psb*C I2 in *Etl.gymnastica* and that the intron degeneration and loss of the maturase in *Etl. pomquetensis* took place afterwards.

*pet*G I1 analysis: We were also interested in closely investigating *pet*G I1, because it was more than twice the size of all other highly conserved potential twintrons, but shared pairwise identities of 87.4%. This resulted in the identification of *pet*G I1 as a complex twintron with high pairwise identities of internal and external twintrons (Fig. 4B). The two twintrons in *pet*G I1 were the same and showed 90% pairwise identity. Both started with a 5′-GTGCG boundary, a 3′ -boundary ACGTTCAT motif and an additional GTGCG at insertion site 261. A comparison of the secondary structure of the introns (Fig. 4B intron a/ intron a.I) with the consensus domain V from the other highly conserved potential twintrons (Fig. S3) showed that 33 out of the 34 nucleotides were identical. The internal twintron comprised 799 bp and was located three nucleotides upstream from the 3′ splice site of the external twintron. It seems reasonable that the internal twintron proliferated into the external twintron and that both originated from the same twintron as the other ones (Figs. 4A and 4B).

## CONCLUSION

Analysis of the genome of all euglenoids sequenced so far in regard to sequence and structural levels makes it apparent that the green algae origin is most visible in the cpGenome of *Etl. pomquetensis*. This can be seen by high pairwise identities in coding regions with the putative chloroplast ancestor *P. parkeae* and a typical green algae and land plant quadripartite genome structure. Still, independent evolution of the genomes since secondary endosymbiosis can also be observed in *Etl. pomquetensis* by decreased protein coding gene content and increased intron numbers compared to *P. parkeae*.

The cpGenome size of *Etl. pomquetensis* was substantially larger than those of other Eutreptiales published so far due to an increased number of introns and intergenic space, and was closest in size to the largest known euglenoid cpGenomes. This contradicts earlier assumptions that introns invaded cpGenomes massively in Euglenales. Interestingly, and unique within the phototrophic euglenoids, we detected a high similarity between more than half of the 51 introns. Another singularity was that no group III introns, or group III twintrons could be identified. This underlines the hypothesis that group II introns arrived first in basally branching euglenoid species and group III introns emerged from group II introns.

Finally, we speculate that future investigations could explore the possibility of a psychrophilic member of the *Pyramimonas* genus as a putative chloroplast donor to the euglenoid lineage and that *Etl. pomquetensis* may very well be the nearest relative up to date.

## ACKNOWLEDGEMENTS

The authors thank Onur Baltaci and Sabine Stratmann-Lettner for lab assistance. Additional thank is due to A. Donath (MITOS) for consultations regarding exonerate.

## ADDITIONAL INFORMATION AND DECLARATIONS

### Funding

ND received a doctoral scholarship from the Bergische University of Wuppertal, Germany. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

### Grant Disclosures

The following grant information was disclosed by the authors:
Bergische University of Wuppertal, Germany.

### Competing Interests

The authors declare there are no competing interests.

### Author Contributions

- Nadja Dabbagh conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables.
- Matthew S. Bennett performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, reviewed drafts of the paper.
- Richard E. Triemer contributed reagents/materials/analysis tools.
- Angelika Preisfeld conceived and designed the experiments, contributed reagents/materials/analysis tools, prepared figures and/or tables, reviewed drafts of the paper.

### DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The sequences have been uploaded as Supplemental Files. The data is also available under GenBank accession number KY706202.

### Data Availability

The following information was supplied regarding data availability:

The raw data has been supplied as a Supplementary File.

### Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.3725#supplemental-information.

## REFERENCES

**Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990.** Basic local alignment search tool. *Journal of Molecular Biology* **215**:403–410 DOI 10.1016/S0022-2836(05)80360-2.

**Bennett MS, Shiu SH, Triemer RE. 2017.** A rare case of plastid protein-coding gene duplication in the chloroplast genome of *Euglena archaeoplastidiata* (Euglenophyta). *Journal of Phycology* **53**:493–502 DOI 10.1111/jpy.12531.

**Bennett MS, Triemer RE. 2015.** Chloroplast genome evolution in the euglenaceae. *Journal of Eukaryotic Microbiology* **62**:773–785 DOI 10.1111/jeu.12235.

**Bennett MS, Wiegert KE, Triemer RE. 2012.** Comparative chloroplast genomics between *Euglena viridis* and *Euglena gracilis* (Euglenophyta). *Phycologia* **51**:711–718 DOI 10.2216/12-017.1.

**Bennett MS, Wiegert KE, Triemer RE. 2014.** Characterization of Euglenaformis gen. nov. and the chloroplast genome of *Euglenaformis* [*Euglena*] *proxima* (Euglenophyta). *Phycologia* **53**:66–73 DOI 10.2216/13-198.1.

**Benson G. 1999.** Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**:573–580 DOI 10.1093/nar/27.2.573.

**Brouard J, Turmel M, Otis C, Lemieux C. 2016.** Proliferation of group II introns in the chloroplast genome of the green alga *Oedocladium carolinianum* (Chlorophyceae). *PeerJ* **4**:e2627 DOI 10.7717/peerj.2627.

**Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A. 2013.** Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research Database* **41**:D226–D232 DOI 10.1093/nar/gks1005.

**Byun Y, Han K. 2006.** PseudoViewer: web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucleic Acids Research* **34**:W416–W422 DOI 10.1093/nar/gkl210.

**Cattolico R, Jacobs MA, Zhou Y, Chang J, Duplessis M, Lybrand T, McKay J, Ong H, Sims E, Rocap G. 2008.** Chloroplast genome sequencing analysis of *Heterosigma akashiwo* CCMP452 (West Atlantic) and NIES293 (West Pacific) strains. *BMC Genomics* **9**:211 DOI 10.1186/1471-2164-9-211.

**Christopher DA, Hallick RB. 1989.** Euglena gracilis chloroplast ribosomal protein operon: a new chloroplast gene for ribosomal protein L5 and description of a novel organelle intron category designated group III. *Nucleic Acids Research* **17**:7591–7608 DOI 10.1093/nar/17.19.7591.

**Conant GC, Wolfe KH. 2008.** GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics* **24**:861–862 DOI 10.1093/bioinformatics/btm598.

**Copertino DW, Hall ET, Van Hook FW, Jenkins KP, Hallick RB. 1994.** A group III twintron encoding a maturase-like gene excises through lariat intermediates. *Nucleic Acids Research* **22**:1029–1036 DOI 10.1093/nar/22.6.1029.

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

**Copertino DW, Shigeoka S, Hallick RB. 1992.** Chloroplast group III twintron excision utilizing multiple 5′- and 3′-splice sites. *The EMBO Journal* **11**:5041–5050.

**Dabbagh N, Preisfeld A. 2016.** The chloroplast genome of *Euglena mutabilis* —Cluster arrangement, intron analysis, and intrageneric trends. *Journal of Eukaryotic Microbiology* **64**:31–44 DOI 10.1111/jeu.12334.

**Dai L, Zimmerly S. 2003.** ORF-less and reverse-transcriptase-encoding group II introns in archaebacteria, with a pattern of homing into related group II intron ORFs. *RNA* **9**:14–19 DOI 10.1261/rna.2126203.

**Darling AC, Mau B, Blattner FR, Perna NT. 2004.** Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Research* **14**:1394–1403 DOI 10.1101/gr.2289704.

**Doetsch NA, Thompson MD, Hallick RB. 1998.** A maturase-encoding group III twintron is conserved in deeply rooted euglenoid species: are group III introns the chicken or the egg? *Molecular Biology and Evolution* **15**:76–86 DOI 10.1093/oxfordjournals.molbev.a025850.

**Drager RG, Hallick RB. 1993.** A complex twintron is excised as four individual introns. *Nucleic Acids Research* **21**:2389–2394 DOI 10.1093/nar/21.10.2389.

**Gibbs SP. 1978.** The chloroplasts of *Euglena* may have evolved from symbiotic green algae. *Canadian Journal of Botany* **56**:2883–2889 DOI 10.1139/b78-345.

**Gibbs SP. 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 DOI 10.1111/j.1749-6632.1981.tb54365.x.

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

**Guillard RRL, Hargraves PE. 1993.** Stichochrysis immobilis is a diatom, not a chrysophyte. *Phycologia* **32**:234–236 DOI 10.2216/i0031-8884-32-3-234.1.

**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 DOI 10.1093/nar/21.15.3537.

**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**:e33746 DOI 10.1371/journal.pone.0033746.

**Kasiborski BA, Bennett MS, Linton EW. 2016.** The chloroplast genome of *Phacus orbicularis* (Euglenophyceae): an initial datum point for the phacaceae. *Journal of Phycology* **52**:404–411 DOI 10.1111/jpy.12403.

**Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012.** Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**:1647–1649 DOI 10.1093/bioinformatics/bts199.

**Kelchner SA. 2002.** Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. *American Journal of Botany* **89**:1651–1669 DOI 10.3732/ajb.89.10.1651.

**Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R. 2001.** REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research* **29**:4633–4642 DOI 10.1093/nar/29.22.4633.

**Lagesen K, Hallin P, Rødland EA, Staerfeldt H, Rognes T, Ussery DW. 2007.** RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research* **35**:3100–3108 DOI 10.1093/nar/gkm160.

**Lambowitz AM, Belfort M. 2015.** Mobile bacterial group II introns at the crux of eukaryotic evolution. *Microbiology Spectrum* **3**:MDNA3-0050-2014 DOI 10.1128/microbiolspec.MDNA3-0050-2014.

**Lambowitz AM, Zimmerly S. 2011.** Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harbor Perspectives in Biology* **3**:a003616 DOI 10.1101/cshperspect.a003616.

**Leander BS, Witek RP, Farmer MA. 2001.** Trends in the evolution of the euglenid pellicle. *Evolution; International Journal of Organic Evolution* **55**:2215–2235.

**Leedale GF. 1967.** *Euglenoid flagellates*. New Jersey: Prentice Hall, 242.

**Lemieux C, Otis C, Turmel M. 2007.** A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. *BMC Biology* **5**:2–10 DOI 10.1186/1741-7007-5-2.

**Linton EW, Hittner D, Lewandowski C, Auld T, Triemer RE. 1999.** A molecular study of euglenoid phylogeny using small subunit rDNA. *The Journal of Eukaryotic Microbiology* **46**:217–223.

**Linton EW, Nudelman MA, Conforti V, Triemer RE. 2000.** A molecular analysis of the Euglenophytes using SSU rDNA. *Journal of Phycology* **36**:740–746 DOI 10.1046/j.1529-8817.2000.99226.x.

**Marin B. 2004.** Origin and fate of chloroplasts in the euglenoida. *Protist* **155**:13–14 DOI 10.1078/1434461000159.

**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**:99–145.

**McLachlan JL, Seguel MR, Fritz L. 1994.** *Tetreutreptia pomquetensis* gen. et sp. nov. (Euglenophyceae): a quadriflagellate, phototrophic marine euglenoid. *Journal of Phycology* **30**:538–544 DOI 10.1111/j.0022-3646.1994.00538.x.

**Merendino L, Perron K, Rahire M, Howald I, Rochaix JD, Goldschmidt-Clermont M. 2006.** A novel multifunctional factor involved in trans-splicing of chloroplast introns in Chlamydomonas. *Nucleic Acids Research* **34**:262–274 DOI 10.1093/nar/gkj429.

**Michel F, Ferat JL. 1995.** Structure and activities of group II introns. *Annual Review of Biochemistry* **64**:435–461 DOI 10.1146/annurev.bi.64.070195.002251.

**Mohr G, Ghanem E, Lambowitz AM. 2010.** Mechanisms used for genomic proliferation by thermophilic group II introns. *PLOS Biology* **8**:e1000391 DOI 10.1371/journal.pbio.1000391.

**Mohr G, Perlman PS, Lambowitz AM. 1993.** Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Research* **21**:4991–4997 DOI 10.1093/nar/21.22.4991.

**Pombert J, James ER, Janouškovec J, Keeling PJ, McCutcheon J. 2012.** Evidence for transitional stages in the evolution of euglenid group II introns and twintrons in the *Monomorphina aenigmatica* plastid genome. *PLOS ONE* **12**:e53433 DOI 10.1371/journal.pone.0053433.

**Preisfeld A, Busse I, Klingberg M, Talke S, Ruppel HG. 2001.** Phylogenetic position and inter-relationships of the osmotrophic euglenids based on SSU rDNA data, with emphasis on the Rhabdomonadales (Euglenozoa). *International Journal of Systematic and Evolutionary Microbiology* **51**:751–758 DOI 10.1099/00207713-51-3-751.

**Ravi V, Khurana JP, Tyagi AK, Khurana P. 2008.** An update on chloroplast genomes. *Plant Systematics and Evolution* **271**:101–122 DOI 10.1007/s00606-007-0608-0.

**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. *Molecular Biology and Evolution* **24**:956–968 DOI 10.1093/molbev/msm012.

**Schattner P, Brooks AN, Lowe TM. 2005.** The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research Web Server* **33**:W686–W689 DOI 10.1093/nar/gki366.

**Slater GStC, Birney E. 2005.** Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **1**:31 DOI 10.1186/1471-2105-6-31.

**Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011.** MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28**:2731–2739 DOI 10.1093/molbev/msr121.

**Thompson MD, Copertino DW, Thompson E, Favreau MR, Hallick RB. 1995.** Evidence for the late origin of introns in chloroplast genes from an evolutionary analysis of the genus *Euglena*. *Nucleic Acids Research* **23**:4745–4752 DOI 10.1093/nar/23.23.4745.

**Thompson MD, Zhang L, Hong L, Hallick RB. 1997.** Extensive structural conservation exists among several homologs of two Euglena chloroplast group II introns. *Molecular and General Genetics* **257**:45–54 DOI 10.1007/s004380050622.

**Toor N, Hausner G, Zimmerly S. 2001.** Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA* **7**:1142–1152 DOI 10.1017/S1355838201010251.

**Turmel M, Gagnon M, 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**:631–648 DOI 10.1093/molbev/msn285.

**Turmel M, Otis C, Lemieux C. 2016.** Mitochondrion-to-chloroplast DNA transfers and intragenomic proliferation of chloroplast group II introns in Gloeotilopsis green algae (Ulotrichales, Ulvophyceae). *Genome Biology and Evolution* **8**:2789–2805 DOI 10.1093/gbe/evw190.

**Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012.** Primer3—new capabilities and interfaces. *Nucleic Acids Research* **40**:e115 DOI 10.1093/nar/gks596.

**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**:832–843 DOI 10.1016/j.protis.2012.01.002.

**Wiegert KE, Bennett MS, Triemer RE. 2013.** Tracing patterns of chloroplast evolution in euglenoids: contributions from *Colacium vesiculosum* and *Strombomonas acuminata* (Euglenophyta). *Journal of Eukaryotic Microbiology* **60**:214–221 DOI 10.1111/jeu.12025.

**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 DOI 10.1186/1471-2148-12-29.

**Zuker M. 2003.** Mfold web server for nuclei acid folding and hybridization prediction. *Nucleic Acids Research* **31**:3406–3415 DOI 10.1093/nar/gkg595.

## PeerJ

Chloroplast genome expansion by intron multiplication in the basal psychrophilic euglenoid *Eutreptiella pomquetensis*

Nadja Dabbagh, Matthew S. Bennett, Richard E. Triemer, Angelika Preisfeld

## SUPPORTING INFORMATION

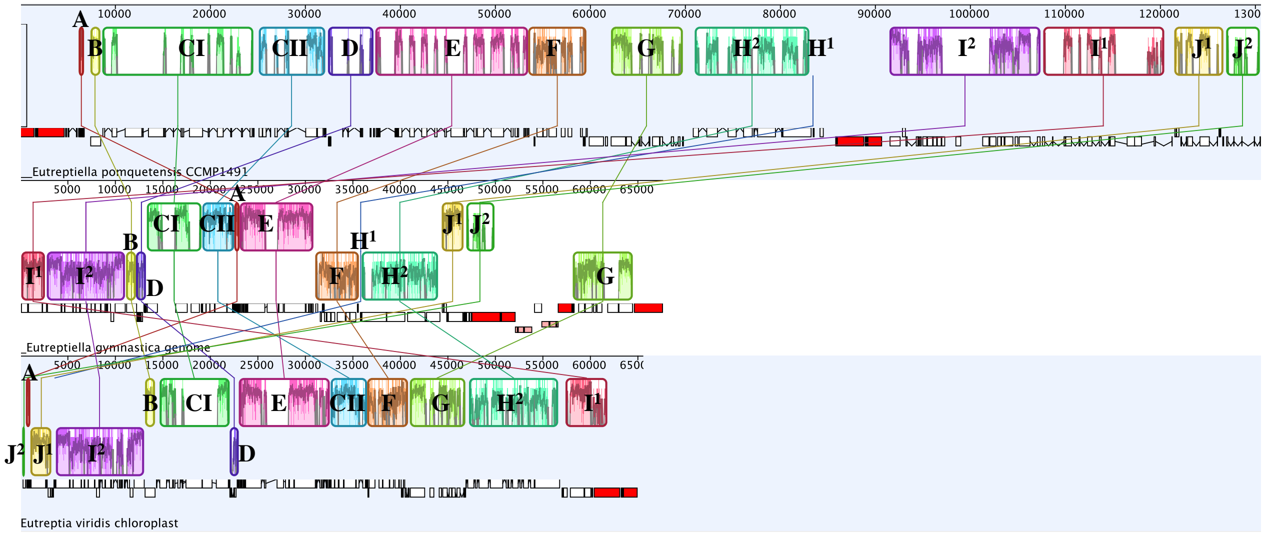

Eutreptiella pomquetensis CCMP1491
Eutreptiella gymnastica genome
Eutreptia viridis chloroplast
A
B
CI
CII
D
E
F
G
H¹
H²
I¹
I²
J¹
J²

**Fig. S1: Progressive Mauve analysis of Eutreptiales.** Each box represents a cluster of homologous genes with *Eutreptiella pomquetensis* as the reference genome. Like blocks are labelled by letters A-J. See Table 4 for a list of genes contained in each block. In the Mauve alignment the repeat regions of rRNA were not included, because Mauve will not align repeat regions, which have multiple matches on both genomes.

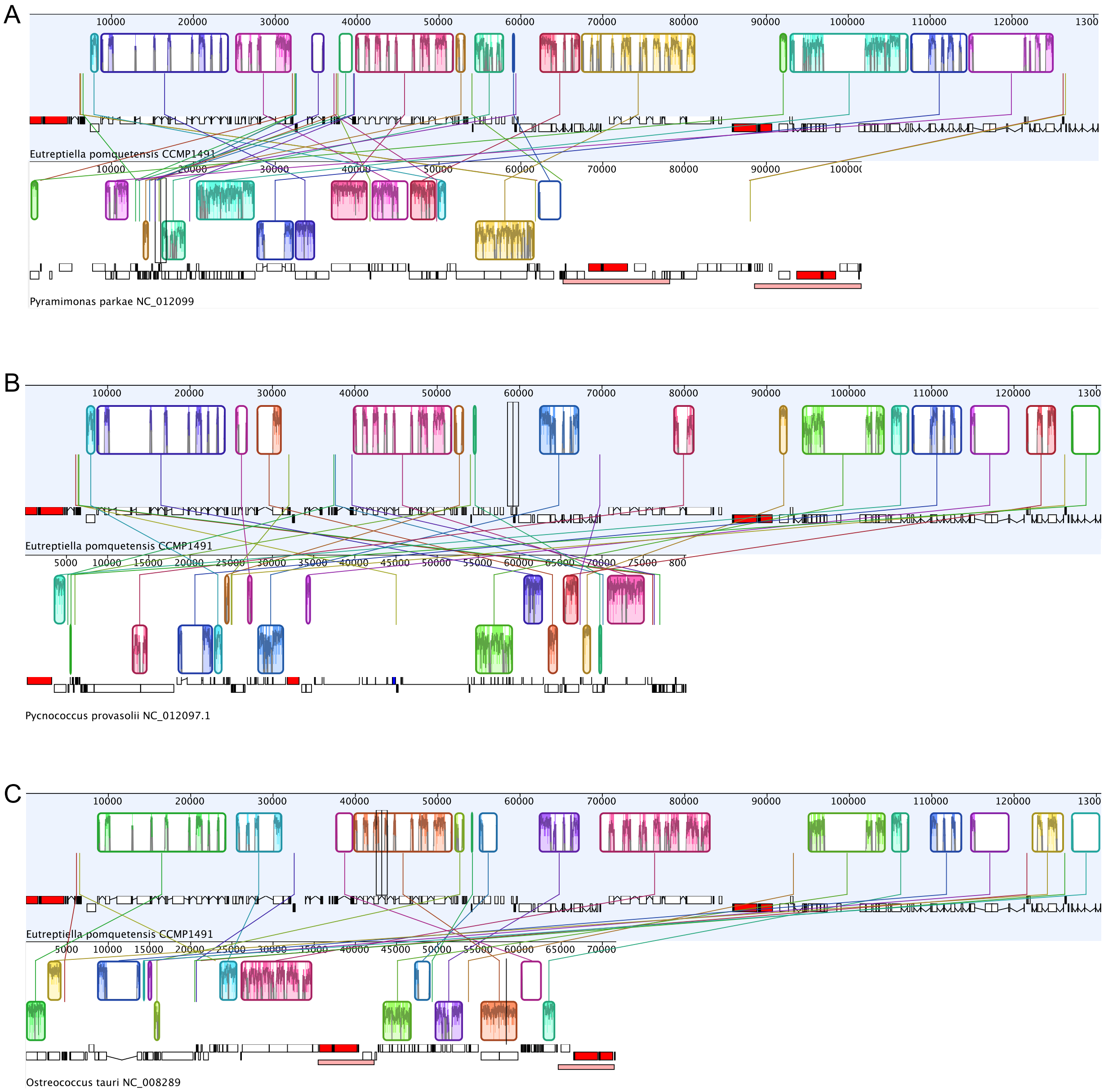


**Fig. S2: Progressive Mauve analysis comparing the cpGenomes of *Etl. pomquetensis* and three Chlorophyta.** Each box represents a cluster of homologous genes between *Eutreptiella pomquetensis* as the reference genome and *Pyramimonas parkeae* (A), *Pycnococcus provasolii* (B) and *Ostreococcus tauri* (C). In the Mauve alignment the repeat regions of rRNA were not included, because Mauve will not align repeat regions, which have multiple matches on both genomes.

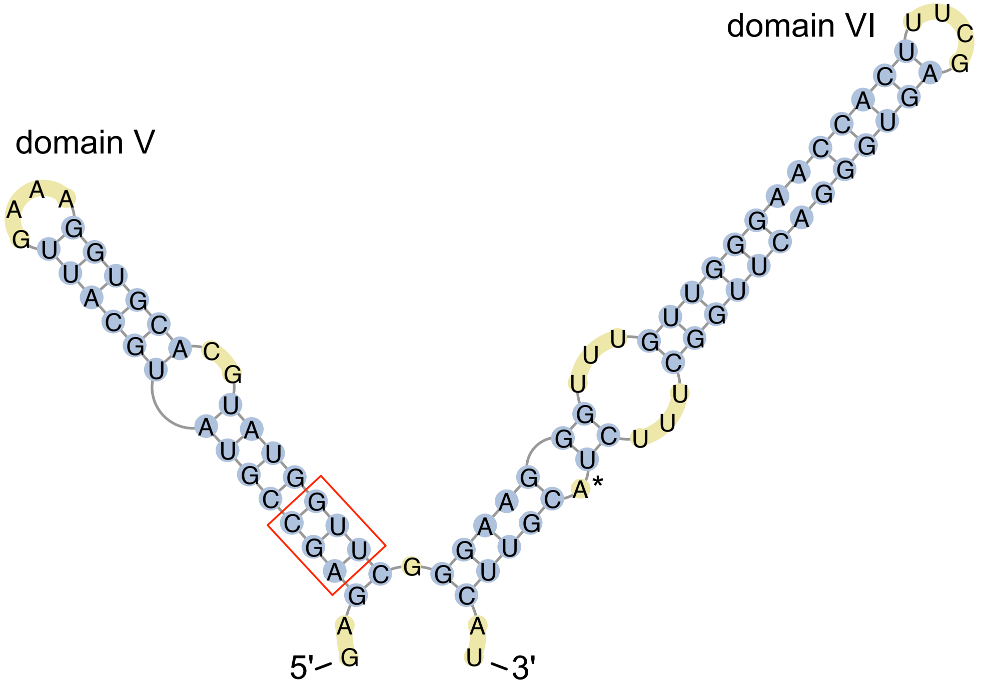


**Fig. S3: Consensus secondary structure model of domain V and VI of highly conserved introns.** Consensus secondary structure model of domain V and VI of the highly conserved introns of *Etl. pomquetensis* based on the model proposed by Michel et al. (1989) and on comparative analysis of other euglenoid group II introns (Thompson et al. 1997). The three base pairs (5’- ...AGC ... GUU…-3’) near the base of stem V were invariant (red box). Introns that form consensus sequence: *atp*B I1- I4; *atp*E I2; *atp*H I1; *psa*A I1- I4 & I6; *psa*B I1-I2; *psa*C I3; *psb*B I2; *psb*C I3 & I5; *psb*D I1& I4-I5; *rbc*L I1; *rpl*32 I1; *rpo*B I1& I3; *rps*7 I1; *rps*12 I1.

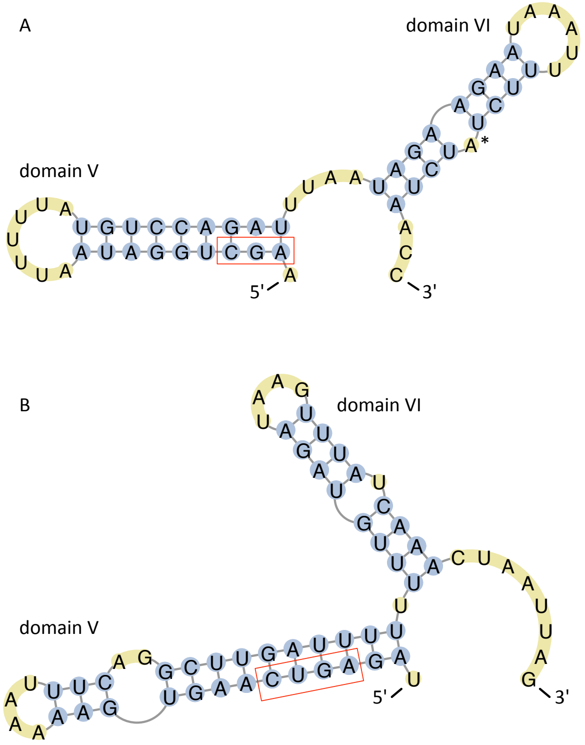


**Fig. S4: Putative secondary structure model of domain V and VI of supposed mini-group II intron.** Domain V and VI of *rpo*B I1 of *Et. viridis* (A) with branch-point A* at position 8 of domain VI and conserved three base pairs (5’- …AGC …-3’) near the base of the stem of domain V (red box). Domain V and VI of *rpo*B I1 of *Etl. gymnastica* (B) with slightly altered base pairs (5’- …AGUC …-3’) of domain V (red box), but without branch-point A* at position 8 of domain VI.

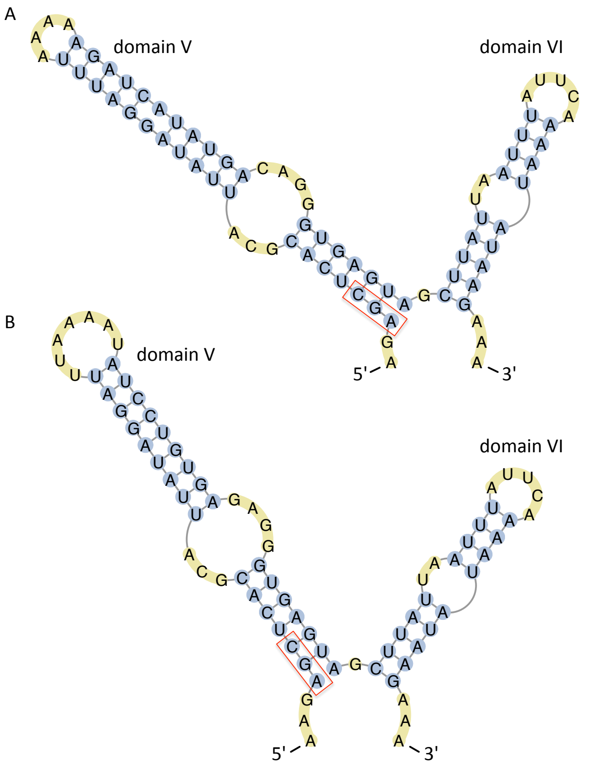


**Fig. S5: Secondary structure model of potential internal introns of twintrons.** Internal introns of *rpo*B I1 of *Etl. pomquetensis* (A) with conserved three base pairs (5'- …AGC …-3') near the base of the stem of domain V (red box). Internal intron of *psb*D I4 of *Etl. pomquetensis* (B) with conserved three base pairs (5'- …AGC …-3') near the base of the stem of domain V (red box).

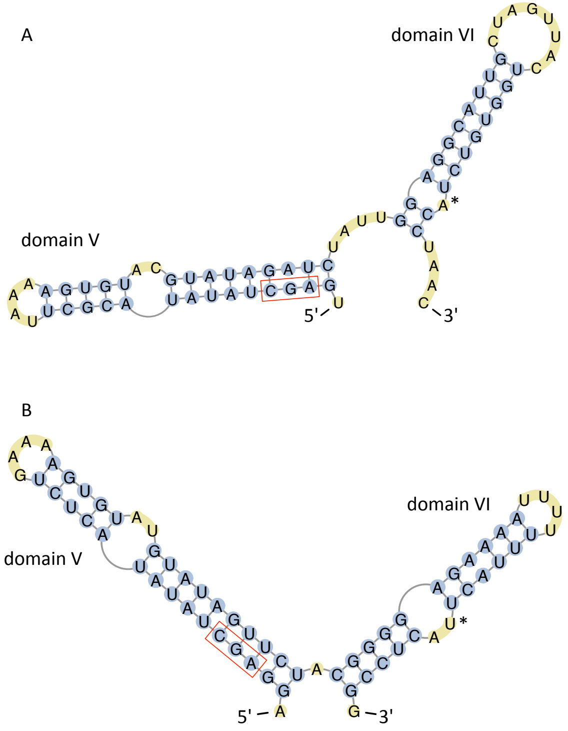


**Fig. S6: Comparative secondary structure analysis.** Secondary structure model of putative domain V and VI of *psa*C I2 external intron of *Etl. pomquetensis* (A) with branch-point A* at position 7 of domain VI and conserved three base pairs (5'- …AGC …-3') near the base of the stem of domain V (red box). Secondary structure model of putative domain V and VI of *psb*C I2 of *Etl. gymnastica* (B) with slightly altered branch-point AU* at positions 7 and 8 of domain VI and conserved three base pairs (5'- …AGC …-3') near the base of the stem of domain V (red box).

**Table S1:** CpGenome features of euglenoids and depicted prasinophytes according to NCBI annotation.
* Chloroplast circle not closed. a:Twintrons were counted as single insertion sites, b: Including rRNA repeats and intermediate tRNAs, c: Includes the identified 5S, for *S. acuminata* the two identified 5S, d: Includes the two introns in rps18, e: Includes the one intercistronic intron *rps*4-*rps*11, f: First exon could not be identified, so gene length is a minimum, g: Realigned, but with alternative start codon, h: Start codon not determined, due to undetermined exon1, i: New intron start after re-analyses.

| Taxon | Size (bp) | A+T % | Genes | Introns[a] | ORFs | *roa*A | CDS of *rpo*A (bp) | Largest gene (bp) | Shortest gene (bp) | Gene with most introns | *psb*D/*psb*C overlap | *pet*B I1 bp/5´start |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *E. gracilis* strain *Z* | 143,171 | 73.9 | 116[b] | 134 | 7 | + | 651 | *psb*C (10861) | *psa*M (96) | *rpo*C1 (11) | + | 909/GUGCG |
| *E. gracilis* var. *bacillaris* | 132,034* | 74.2 | 104 | 134 | 6 | + | 525 | *psb*C (11446) | *psa*M (96) | *rpo*C1 (11) | + | 909/GUGCG |
| *E. viridis* | 76,156 | 73.8 | 92 | 77 | 0 | + | 483 | *psb*C (6165) | *psa*M (96) | *rpo*B (10) | + | 528/GUGUG |
| *E. viridis epitype* | 91,616 | 73.6 | 92 | 76 | 13 | + | 480 | *psb*C (6131) | *psa*M (96) | *rpo*C1 (8) | + | 533/GUGUG |
| *E. mutabilis* | 86,975 | 73.3 | 91 | 77[e] | 0 | + | 636 | *psb*C (12192) | *psa*M (96) | *rpo*C1 (9) | + | 435/GUGCG |
| *Era. anabaena* | 88,487 | 72 | 93 | 82 | 4 | + | 624 | *psb*A (6998) | *psa*M (96) | *rpo*B (10) | + | 551/GUGCG |
| *M. parapyrum* | 80,147 | 72 | 93 | 80 | 1 | + | 507 | *psb*C (6127) | *psa*M (96) | *rpo*B/-C1 (9) | + | 441/GUGCG |
| *M. aenigmatica* | 74,746 | 70.6 | 93[c] | 53 | 1 | + | - | *psb*C (6229) | *psa*M (96) | *rpo*C1 (9) | + | 544/GUGCG |
| *Cr. skujae* | 106,843 | 73.7 | 93 | 84 | 4 | + | 489 | *psb*C (8947) | *psa*M (96) | *rpo*B (10) | + | 537/GUGUG |
| *S. acuminata* | 144,166* | 73.4 | 95[c] | 112[d] | 0 | + | 486 | *psa*B (11283) | *psb*T (90) | *rpo*B/ *rbc*L (9) | + | 510/GUGCG |
| *T. volvocina* | 85,392* | 72.7 | 93 | 94 | 1 | + | 543 | *psb*C (7698) | *psa*M (96) | *rpo*C1 (11) | + | 536/GUGCG |

**Table S1:** Continued.

| Taxon | Size (bp) | A+T % | Genes | Introns[a] | ORFs | roaA | CDS of *rpo*A (bp) | Largest gene (bp) | Shortest gene (bp) | Gene with most introns | *psb*D/*psb*C overlap | *pet*B I1 bp/5´start |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *C. vesiculosum* | 128,892* | 73.9 | 92[c] | 128 | 6 | + | - | *psb*C (11567) | *psb*I (105) | *rpo*C1 (11) | + | 470/GUGUG |
| *Efs. proxima* | 94,185* | 73.1 | 91 | 113 | 2 | + | - | *psb*C[f](6648) | *psa*M (96) | *rpo*B (10) | [h] | 478/GUGCG |
| *P. orbicularis* | 65,992 | 72,8 | 91 | 66 | 1 | + | - | *rpo*B (3993) | *psa*M (96) | *rpo*B (10) | + | 357/GUGCG[i] |
| *Et. viridis* | 65,523* | 71.4 | 83 | 24 | 3 | - | 672 | *psb*C (5706) | *psb*T (96) | *rpo*B (5) | - | - |
| *Etl. gymnastica* | 67,623 | 65.7 | 89 | 7 | 4 | - | 846 | *psb*C | *psa*M/*psb*T (96) | *psb*C (2) | +[g] | - |
| ***Etl. pomquetensis*** | **130,561*** | **64.9** | **94** | **51** | **10** | | | ***psb*D (8412)** | ***psa*M (96)** | ***psa*A (6)** | + | **1736/GUGUG** |
| *Pyramomonas parkae* | 101,605 | 65.3 | 125 | 1 | 5 | - | 1056 | *atp*B (4224) | *psb*T (96) | *atp*B (1) | + | - |
| *Pyconococcus provasolii* | 80,211 | 60.5 | 98 | 1 | 3 | - | 1074 | *fts*H (7944) | *psb*T (96) | *atp*B (1) | + | - |
| *Ostreococcus tauri* | 71,666 | 60.1 | 92 | 1 | 2 | - | 1071 | *atp*B (5188) | *psa*M/psbT (96) | *atp*B (1) | + | - |

**Table S2:** Features of introns in protein-coding genes of *Etl. pomquetensis*

| Gene | Intron | Start | Stop | Length | GC content (%) | Insertion site | Additional GTGCG | Insertion additional GTGCG (nt) | Repeated 3' motifs |
|---|---|---|---|---|---|---|---|---|---|
| ***atp*B** | **1** | **GTGCG** | **ACGTTCAT** | **803** | **37,5** | **70** | | | TTAATTTATTCAAAATATAAGAAA * |
| | **2** | **GTGCG** | **ACGTTCAT** | **776** | **36,2** | **1061** | **1** | **261** | ATAATTCGGCAAAATATATTAGA *** |
| | **3** | **GTGCG** | **ACGTTCAT** | **787** | **37,7** | **2309** | **1** | **261** | TTCGGTTGTATACTACCAGAAGAAA ** |
| | **4** | **GTGCG** | **ACGTTCAT** | **791** | **35,5** | **3209** | **1** | **261** | |
| *atp*E | 1 | GTGCG | AGTTTAAC | 1173 | 34,3 | 136 | | | AAGAAATGTTTTACTTTAGA **** |
| | **2** | **GTGCG** | **ACGTTCAT** | **829** | **36,9** | **1347** | **1** | **261** | |
| *atp*F | 1 | GTGCG | CCTAACCA | 814 | 29,4 | 21 | 2 | 29/79 | |
| ***atp*H** | **1** | **GTGCG** | **ACGTTCAT** | **775** | **36,8** | **88** | **1** | **261** | |
| *pet*B | 1 | GTGTG | TACCTGAC | 1736 | 32,5 | 23 | 2 | 671/ 931 | |
| | 2 | GTGCG | ACTTTCAT | 4128 | 30,6 | 1898 | 2 | 784/ 2937 | AAGAAATGTTTTACTTTAGA **** |
| ***pet*G** | **1** | **GTGCG** | **ACGTTCATCAT** | **1578** | **36,4** | **49** | **4** | **261/777/1057** | |
| ***psa*A** | **1** | **GTGCG** | **ACGTTCAT** | **827** | **36,9** | **343** | **1** | **261** | |
| | **2** | **GTGCG** | **ACGTTCAT** | **774** | **37,6** | **1654** | **1** | **261** | |
| | **3** | **GTGCG** | **ACGTTCAT** | **823** | **36,8** | **2977** | **1** | **261** | |
| | **4** | **GTGCG** | **ACGTTCAT** | **809** | **36,6** | **4048** | **1** | **261** | |
| | 5 | GTGCG | AATTTAAC | 1908 | 28 | 5030 | | | |
| | **6** | **GTGCG** | **ACGTTCAT** | **773** | **37** | **7147** | **1** | **261** | |
| ***psa*B** | **1** | **GTGCG** | **ACGTTCAT** | **786** | **34,6** | **669** | | | |
| | **2** | **GTGCG** | **ACGTTCAT** | **781** | **36,2** | **2772** | **1** | **261** | |

Table S2 Continued.

| Gene | Intron | Start | Stop | Length | GC content (%) | Insertion site | Additional GTGCG | Insertion additional GTGCG (nt) | Repeated 3' motifs |
|---|---|---|---|---|---|---|---|---|---|
| *psa*C | 1 | GTGCG | CCGAGTTT | 576 | 26,7 | 42 | 2 | 3/334 | |
| | **2** | **GTGCG** | **TACCTAAC** | **1294** | **34,2** | **649** | **2** | **282/ 540** | |
| | **3** | **GTGCG** | **ACGTTCAT** | **827** | **35,6** | **1957** | **1** | **261** | |
| | 4 | GTGCG | AGTTTAAT | 619 | 28,3 | 2809 | | | |
| *psb*B | 1 | GTGCG | ATCCATAT | 837 | 25,8 | 20 | | | |
| | **2** | **GTGCG** | **ACGTTCAT** | **775** | **38,2** | **1849** | **1** | **261** | |
| *psb*C | 1 | GTGCG | AGTTTAAG | 2580 | 29,1 | 154 | 3 | 45/ 2420/ 2554 | |
| | 2 | GTGCG | TTCACCTA | 844 | 36,1 | 2847 | 1 | 259 | TTAATTTATTCAAAATATAAGAAA * |
| | **3** | **GTGCG** | **ACGTTCAT** | **788** | **36,7** | **4310** | **1** | **261** | ATAATTCGGCAAAATATATTAGA *** |
| | 4 | TTGCG | ACGCTCAT | 1029 | 38,7 | 5199 | 1 | 955 | |
| | **5** | **GTGCG** | **ACGTTCAT** | **787** | **37,1** | **6347** | **1** | **261** | ATAATTCGGCAAAATATATTAGA *** |
| ***psb*D** | **1** | **GTGCG** | **ACGTTCAT** | **778** | **36,5** | **192** | **1** | **261** | ATAATTCGGCAAAATATATTAGA *** |
| | 2 | GTGCG | AGTTTAAC | 2744 | 35,5 | 970 | 1 | 1176 | AAGAAATGTTTTACTTTAGA **** |
| | 3 | GTGCG | ACTTTCAT | 2235 | 29,1 | 4248 | 1 | 421 | |
| | **4** | **GTGCG** | **ACGTTCAT** | **818** | **36,1** | **6604** | **1** | **259** | TTAATTTATTCAAAATATAAGAAA * |
| | **5** | **GTGCG** | **ACGTTCAT** | **778** | **36,5** | **7445** | **1** | **261** | ATAATTCGGCAAAATATATTAGA *** |
| *psb*K | 1 | GTATG | ATAACGAT | 504 | 22,8 | 33 | | | |
| | 2 | GTGTG | AGTTTACG | 1255 | 33,2 | 592 | 2 | 115/ 347 | |
| *psb*T | 1 | GTGCG | TACTCAAT | 497 | 29 | 25 | | | |
| ***rbc*L** | **1** | **GTGCG** | **ACGTTCAT** | **774** | **37,5** | **290** | **1** | **261** | |
| ***rpl*32** | **1** | **GTGCG** | **ACGTTCAT** | **791** | **36,5** | **92** | **1** | **261** | ATAATTCGGCAAAATATATTAGA ***<br>TTCGGTTGTATACTACCAGAAGAAA ** |

Table S2 Continued.

| Gene | Intron | Start | Stop | Length | GC content (%) | Insertion site | Additional GTGCG | Insertion additional GTGCG (nt) | Repeated 3' motifs |
|---|---|---|---|---|---|---|---|---|---|
| ***rpo*B** | **1** | **GTGCG** | **ACGTTCAT** | **866** | **36,3** | **555** | **1** | **261** | TTAATTTATTCAAAATATAAGAAA * |
| | 2 | GTGCG | AGTTTAAC | 1112 | 35,1 | 2550 | | | |
| | **3** | **GTGCG** | **ACGTTCAT** | **774** | **36,6** | **4523** | | | |
| *rpo*C1 | 1 | GTAAA | CAGAAGCC | 356 | 28,9 | 12 | | | |
| | 2 | GTGTG | GACCACTA | 654 | 31 | 2245 | | | |
| ***rps*12** | **1** | **GTGCG** | **ACGTTCAT** | **793** | **36,9** | **362** | **1** | **261** | |
| *rps*14 | 1 | GTAAG | ATTTAACC | 962 | 26,8 | 177 | | | |
| rps2 | 1 | GTGCG | ATCCAGGG | 491 | 28,3 | 112 | | | |
| | 2 | GTGTA | ACCTACCA | 921 | 24,6 | 840 | | | AAAATAATATTTTTATATTTTGTT |
| ***rps*7** | **1** | **GTGCG** | **ACGTTCAT** | **789** | **36,4** | **261** | **1** | **261** | ATAATTCGGCAAAATATATTAGA *** |
| *tuf*A | 1 | GTGCG | AGTTTCAC | 1829 | 33,1 | 75 | 1 | 626 | |

## 3.3 Chapter III: Intrageneric Variability between the Chloroplast Genomes of *Trachelomonas grandis* and *Trachelomonas volvocina* and phylogenomic analysis of phototrophic euglenoids

Running Head: *Trachelomonas grandis* Chloroplast Genome

Title: Intrageneric Variability between the Chloroplast Genomes of *Trachelomonas grandis* and *Trachelomonas volvocina* and phylogenomic analysis of phototrophic euglenoids

Authors: Nadja Dabbagh and Angelika Preisfeld

Institute: Bergische University Wuppertal
Faculty of Mathematics and Natural Sciences
Zoology and Didactics of Biology
Gaußstraße 20, 42115 Wuppertal, Germany

Date of submission: 06. October 2017

Under minor revision: 16. November 2017

**Keywords: Chloroplast genome; Euglenaceae; Phylogenomic analysis; *Trachelomonas grandis*.**

## ABSTRACT

The latest studies of chloroplast genomes of phototrophic euglenoids yielded different results according to intrageneric variability such as cluster arrangement or diversity of introns. Whereas the genera *Euglena* and *Monomorphina* show high syntenic arrangements at the intrageneric level, the *Eutreptiella* species comprise low synteny. Previous phylogenetic and secondary structure studies also detected a high intrageneric diversity of nuclear SSU genes in the genus *Trachelomonas*. Consequently, this study aims at the analysis of the chloroplast genome of *Trachelomonas grandis* and a comparative examination of *Trachelomonas volvocina* to better understand if the genomic intrageneric diversity followed a general trend. Although these analyses resulted in almost identical gene content to other Euglenaceae, the chloroplast genome showed significant novelties: In the rRNA operon we detected group II introns, not yet found in any other cpGenome of Euglenaceae and a substantially heterogeneous cluster arrangement in the genus *Trachelomonas.* The high intrageneric variety between the two *Trachelomonas* cpGenomes is mirrored in the diversity identified between nuclear genes of this genus. The phylogenomic analysis with 84 genes of 19 phototrophic euglenoids and 18 cpGenome sequences from Chlorophyta and Streptophyta resulted in a well supported cpGenome phylogeny, which is in accordance to former phylogenetic analyses.

## INTRODUCTION

The euglenoids represent a diverse, ancient eukaryotic lineage with different nutrition modes, in which osmotrophics and phototrophics have independently evolved from phagotrophic ancestors (Busse & Preisfeld 2002b). The systematics and taxonomic identification of phototrophic euglenoids find their roots with Ehrenberg (1830), who described several of the major taxa. As a distinctive character these euglenoids contain chloroplasts acquired via secondary endosymbiosis (Gibbs 1978, 1981) of a prasinophyte green alga (Turmel et al. 2009) by an unknown phagotrophic ancestor. Currently phototrophic euglenoids comprise 14 genera with distinct morphological features such as the number and shape of chloroplasts and paramylon granules or pellicle plasticity (Adl et al. 2012, Bicudo & Menezes 2016, Marin et al. 2003). Some of these quite heterogeneous genera are only covered by a pellicle, whereas others, like *Trachelomonas*, are surrounded by an envelope. Originally, all taxa with an envelope, also known as lorica, were classified in the genus *Trachelomonas* Ehrenberg (1833). It was Deflandre (1930) who separated the subgroup Saccatae from *Trachelomonas* and established the new genus *Strombomonas,* solely on characteristics of the lorica. Then again, early phylogenetic studies considered the loricate genera *Strombomonas* and *Trachelomonas* as ambiguous (Brosnan et al. 2003, Nudelman et al. 2003) and Marin et al.(2003), who revised euglenophyte taxonomy based on comprehensive SSU rDNA analysis, subsumed *Strombomonas* under *Trachelomonas* again. However, the detailed morphological study of Brosnan et al. (2005) based on lorica development and posterior strip reduction, supported the separation of *Strombomonas* and *Trachelomonas* as distinct genera. Later investigations based on combined molecular and morphological data strengthened the separation between *Trachelomonas* and *Strombomonas* (Ciugulea et al. 2008, Poniewozik 2017, Triemer et al. 2006). Subsequent molecular studies using different nuclear and/ or chloroplast encoded genes confirmed *Strombomonas* and *Trachelomonas* as two distinct genera (Karnkowska et al. 2015, Kim et al. 2015, Linton et al. 2010).

*Trachelomonas grandis* is one of over 350 *Trachelomonas* species globally distributed in fresh waters. It possesses a large ellipsoidal lorica with a short narrow collar and is of 30 – 43 µm length and 25 - 32µm width. The outer surface shows short irregularly arranged granules, which probably stem from epibiotic bacteria attached to the surface of the lorica (Rosowski & Langenberg 1994, Rosowski & Coute 1996). The locomotory flagellum is four to five times the body length and moves apically like a spinning lasso (Ciugulea & Triemer 2010, Leedale 1967).

Since Hallick et al. (1993) published the first chloroplast genomes of phototrophic euglenoids and Turmel et al. (2009) provided unequivocal evidence of euglenoid chloroplasts being inherited from a member of the Pyramimonadales, quite a number of cpGenomes of phototrophic euglenoids and prasinophytes have been analyzed. The aim of this study is to compare the intrageneric variations in the cpGenome of *Trachelomonas grandis* to those of *T. volvocina* and furthermore, to perform phylogenomic analyses of published cpGenomes of phototrophic euglenoids and fourteen representatives of Chlorophyta and two Streptophyta as outgroup.

The investigation of the chloroplast genome of *Trachelomonas grandis* Singh 1956 seems promising because of two important reasons: Only one annotated chloroplast genome of *Trachelomonas volvocina* is available (Bennett & Triemer 2015) and more are needed for a phylogenomic reconstruction. Moreover, Marin et al. (2003) identified *T. grandis* as a Euglenophyceae with extremely long nuclear SSU rRNA genes. Beyond that they recognized a high intrageneric diversity and long individual branches in nuclear SSU rRNA genes in the genus *Trachelomonas*. This study aims to explore intrageneric diversity and genomic changes between the two *Trachelomonas* plastid genomes and to examine whether the diversity in nuclear genes is reflected in the chloroplast genes and genomes.

The phylogenomic analysis of all sequenced euglenoid cpGenomes up to date expands the taxon sampling of further studies (Bennett & Triemer 2015, Bennett et al. 2017, Dabbagh & Preisfeld 2017, Dabbagh et al. 2017, Kasiborski et al. 2016). It was also carried out, to examine the position of *Colacium vesiculosum*, since the only phylogenomic analysis performed previously by Bennett et al. (2014) resulted in trees, which weakly supported *Colacium vesiculosum* as sister to *Euglena gracilis* and *Euglena viridis*, a position which seems doubtful in regard to morphological characters. For example, *C. vesiculosum* contains parietal, discoid chloroplasts with haplopyrenoids whereas *E. gracilis* and *E. viridis* have disc shaped or lobed chloroplasts with naked or diplopyrenoids (Kim et al. 2015). In the light of these morphological characters and the ability of *C. vesiculosum* to produce large amounts of mucilage, a closer relationship to the clade combining the loricate *Strombomonas* and *Trachelomonas* would seem more convincing (Brown et al. 2003). However, molecular data at the moment are quite contradictory (Karnkowska et al. 2015, Kim et al. 2015, Marin et al. 2003, Milanowski et al. 2006, Triemer et al. 2006).

## MATERIAL AND METHODS

### Sequencing and analysis

*Trachelomonas grandis* strain SAG 204.80 (EPSAG, Germany) was grown in WEES medium (Kies 1967) at 20 - 23 °C under 12 : 12 light : dark cycle using fluorescent tubes, which supplied approximately 30μmol photons $m^{-2}$ $s^{-2}$ of light. Cells were concentrated and washed, chloroplasts isolated and DNA extracted as described previously (Dabbagh & Preisfeld 2017), with the following modification: To isolate the chloroplast, cells were disrupted by ultrasonic probe 6 times for 3 sec with the amplitude set at 80 % and a 0.1 sec puls rate (Bandelin Sonoplus HD 60, Berlin, Germany) with intermediate washing steps on ice. After the Percoll gradient was centrifuged, the chloroplast fraction was recovered from the 30 % layer.

The DNA was sequenced with 454 sequencing using a Roche GS FLX ++ system for single reads (Eurofins Genomics Ebersberg, Germany). Raw sequence reads were then automatically assembled into contigs by Eurofins Genomics using Roche´s 454 GS Assembler, Newbler. The assembled contigs were searched for chloroplast genome sequences using $\text{Blast}_N$ homology search, and subsequently chloroplast contigs were circularized. Geneious Pro 9 (version 9.1.3, Kearse et al. 2012) was used for final annotation of the chloroplast sequence as described before (Dabbagh et al. 2017). Protein-coding genes and their introns were identified and manually aligned in MEGA 7 (Kumar et al. 2016) against the nucleotide coding DNA sequences (CDSs) from other photosynthetic euglenoids and prasinophyte representatives, to determine exon-intron boundaries as well as the start and stop of each gene. Methionine start codons (ATG) were preferred as start codon. After review with $\text{BLAST}_X$ and Emboss Sixpack Sequence translation (EMBL- EBI 2015) CDSs were added to the annotation. The annotated introns within protein-coding genes were manually compared with the introns of *T. volvocina*. In addition, each intron of *T. grandis* was searched for typical domain V motifs of group II introns, to determine group II or group III introns.

The start/stop area of rRNA gene sequences for 16S and 23S rRNA were identified with RNAmmer 1.2 Server (Lagesen et al. 2007) using bacteria as the sequence kingdom of origin. The 5S rRNA start/ stop regions were identified using Rfam 12.1 Sequence Search (Burge et al. 2013). All three genes were manually realigned and all introns identified in the 16S and 23S rRNA genes have been verified by RT-PCR experiments. Primers were created manually by Primer3Plus (Untergasser et al. 2012) based on the nucleotide sequence. RNA was isolated by my-Budget RNA Mini Kit (Bio-Budget Technologies GmbH) following the manufacturer’s recommendations and afterwards RT-PCR was performed by OneStep RT-PCR Kit (Qiagen GmbH, Hilden Germany) using specific primers to *T. grandis*. In addition the rRNA operon was tested for potentially repeats in PCR reactions using specific *T. grandis* primers. Furthermore, the average depth of the contig containing the rRNA genes was compared with the average depth of single copy protein-coding genes without finding any differences. To identify tRNA-encoding genes, the cpGenome was submitted to tRNAscan-SE 2.0 server. The sequence was searched with the default settings and the source given as mito/ chloroplast (Schattner et al. 2005).

The final annotation file was checked for open reading frames (ORFs) using the ORF Finder option within Geneious Pro 9 (Kearse et al. 2012) with genetic code 11. The predicted ORFs were checked manually and corresponding ORFs were added to the final annotation as described previously (Dabbagh et al. 2017). The Variable Number of Tandem Repeat (VNTR) region was detected with the plugin find repeats from Geneious Pro 9. Repeats were searched with minimum repeat length set at 20 and maximum mismatches set at 0 %. Final annotated repeats were those repeats, which did not overlap with larger repeats. The cpGenome was arranged in a way that the rRNA genes were counterclockwise; this arrangement allowed for easier comparison with other cpGenomes of phototrophic euglenoids and was established by Hallick et al. (1993). The circular genome map was generated using GenomeVx and manually edited. The cpGenome sequence was deposited in the NCBI GenBank database.

For structural and synteny comparison between *T. grandis*, *T. volvocina* and *S. acuminata* the cpGenomes were aligned using progressive Mauve with default settings (Darling et al. 2004) as a plugin from Geneious Pro 9.

### Phylogenomic analyses

To understand the evolution of the coding regions of phototrophic euglenoids better, phylogenetic analyses were carried out on datasets created by 84 genes of 37 plastid genomes including all nineteen phototrophic euglenoids sequenced so far as well as 18 cpGenome sequences from Chlorophyta and two Streptophyta based on GenBank sequences (Table S1). The sequences of the 84 genes were individually aligned with MAFFT, implemented as a plugin of Geneious Pro 9 and afterwards manually adjusted in MEGA 7 (Kumar et al. 2016). In order to avoid inconsistencies or errors in the published annotations some coding regions have been reannotated (alignments are available upon request from the authors). Each single protein-coding gene was aligned according to their amino acid sequence, sites that could not be unambiguously aligned were removed and only homologous sites were used for further analysis. The protein data matrix contained a total of 10,640 amino acid characters. The small fragments of tRNAs used in this analysis had a total of 1,792 bp and the three concatenated rRNAs a total of 3,966 nucleotide sites used for phylogenetic tree reconstruction. For Maximum Likelihood (ML) analyses each protein-coding gene was divided into a separate partition, one tRNA partition and one rRNA partition, resulting in 59 partitions. We determined the best choice of model for each partition under the Akaike Information Criteria (AIC) as recommended by Posada & Buckley (2004) using the IQ-TREE web server (Trifinopoulos et al. 2016) with the additional 'New model selection procedure'. For tRNA and rRNA genes we specified the 'Sequence type' as 'DNA'. For the partitioned protein-coding genes the 'Sequence type' was specified as 'DNA → AA' with the 'Genetic code 11' for 'Bacteria, Archaeal and Plant Plastid'. Data were analyzed for tree inference with the IQ-TREE multicore version by ML (Nguyen et al. 2015), using partitioned analysis for multi-gene alignments under the recommended models (Chernomor et al. 2016) and 1,000 ultrafast bootstrap (Minh et al. 2013).

## RESULTS AND DISCUSSION

### The cpGenome of *T. grandis*

The chloroplast genome of *Trachelomonas grandis* (Fig.1) was circularized, annotated and resulted in a cpGenome with 113,311 bp. The size is approximately 28 kb larger than the chloroplast genome of *T. volvocina*, but in the range of other published euglenoid cpGenomes, for example such as the one of *Cryptoglena skujae* (Bennett & Triemer 2015). One reason for the size difference between both *Trachelomonas* species originated in the intergenic space, which was found to be twice the size in *T. grandis* (18.8 kb). With 26.5 % the GC content lies in the same range as for *T. volvocina* and all other phototrophic euglenoids (Table S2).

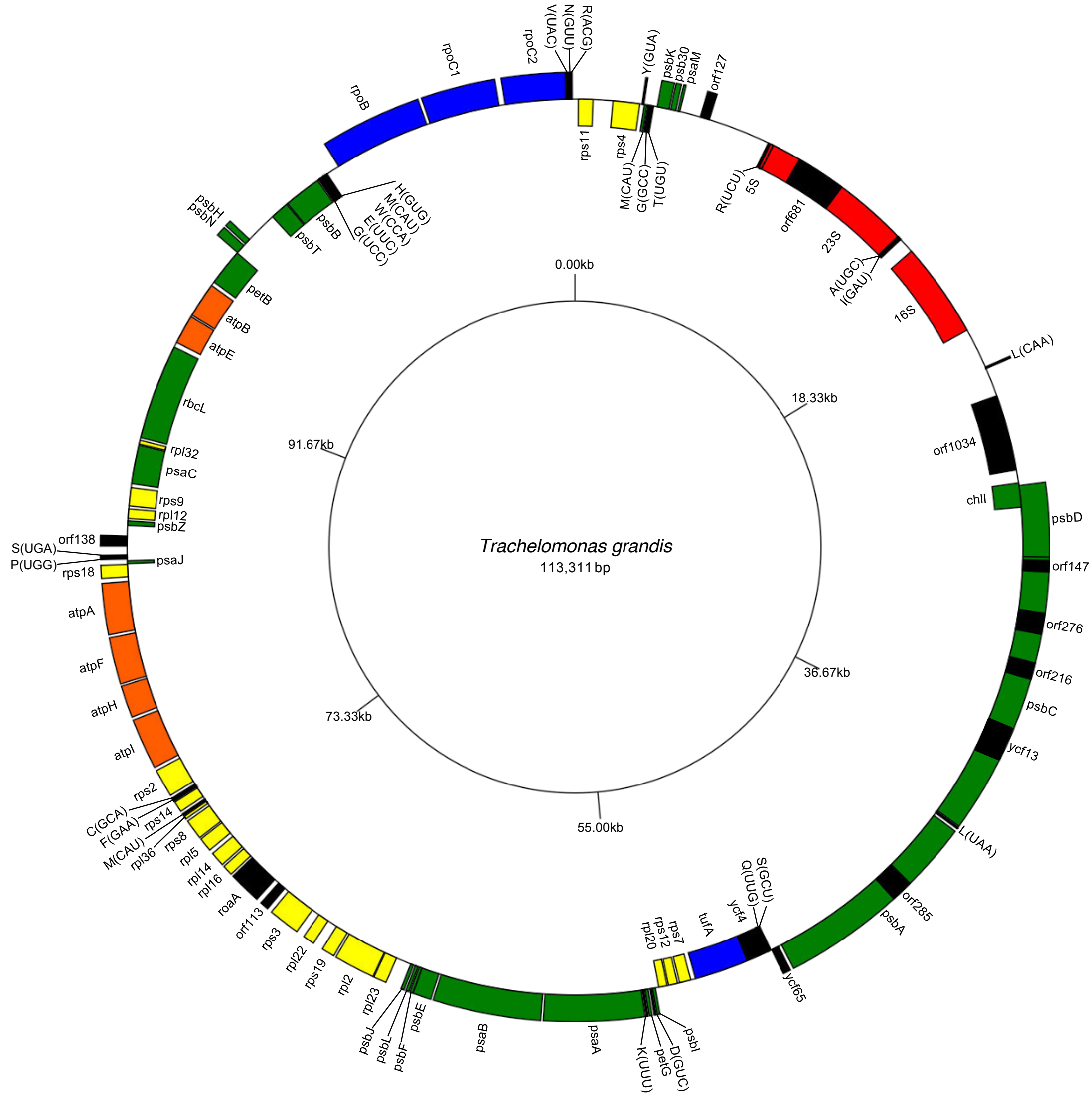


**Fig. 1:** Circular chloroplast genome map of *Trachelomonas grandis*. Boxes are proportional to the gene sequence length. Genes on the outside of the circle are considered on the positive strand, genes inside the circle on the negative strand. Genes are colored according to their function: red= ribosomal rRNAs; green= photosystem/ photosynthesis genes; yellow = ribosomal proteins (*rpl*, *rps*); orange = *atp* genes; blue = transcription/ translation- related genes (*rpo*, *tuf*A); black = conserved hypothetical proteins (*ycf*), open reading frames (*orf*), tRNAs.

## Phylogenomic analyses

Thirty-seven taxa belonging to the outgroup (18) and phototrophic euglenoids (19) including newly sequenced *Trachelomonas grandis* were analyzed using a combination of 57 protein-coding genes, three rRNAs and 24 tRNAs. For each protein-coding gene an individual model was selected, taking into account that not all genes evolved under the same rate (Wolf et al. 2009), thus resulting in 59 partitions. Analyses of data set with (data not shown) and without tRNAs (Fig. 2) recovered trees of identical topology, but greater support was resolved without the highly conserved tRNAs. In a second analysis, we excluded *Colacium vesiculosum* (Wiegert et al. 2013) from the dataset because the ML analysis provided no significant support for the position of *C. vesiculosum* (<70, Fig. 2, position of *C. vesiculosum* schematically depicted with dotted lines). The elimination of *C. vesiculosum* did not lead to a different topology for all other phototrophic euglenoids, but it reinforced the support for sister group relationships in clade B (Fig. 2).

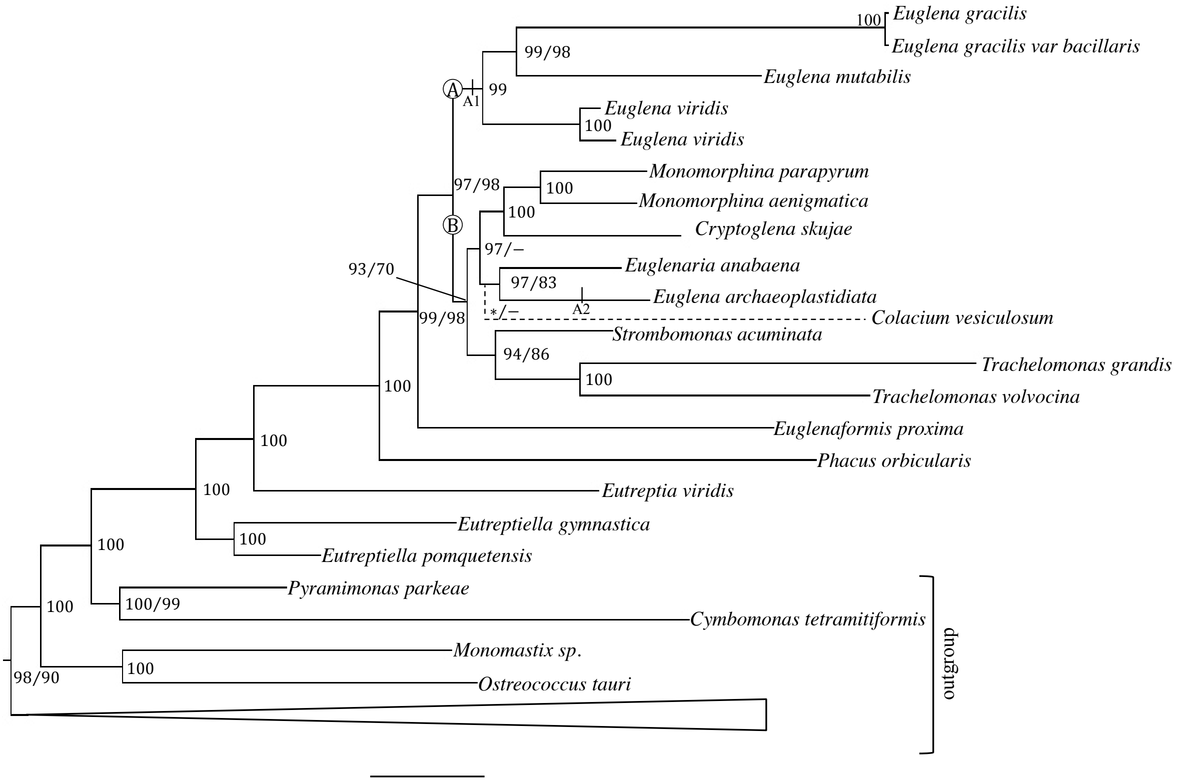
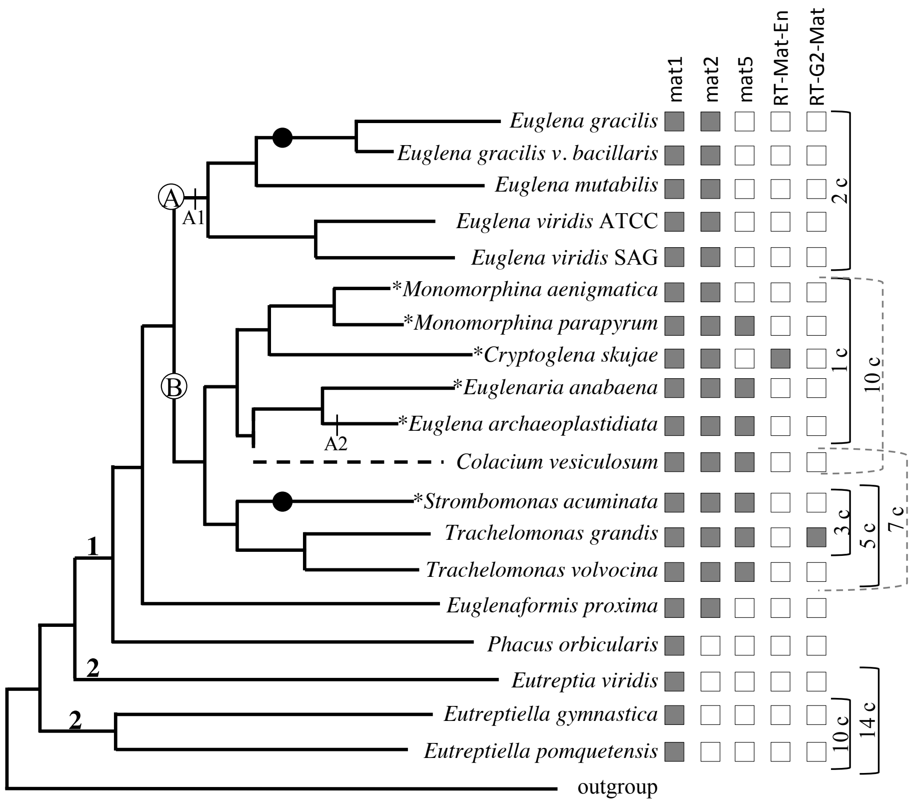


**Fig. 2:** Maximum likelihood (ML) tree obtained from 18 ingroup taxa of phototrophic euglenoids, sequences of Chlorophyta and Streptophyta were used as outgroup. For taxon sampling see Table S1. 14,606 aligned sites were partitioned into 57 datasets of protein coding genes (in total 10,640 amino acid characters) and one datasets of rRNA genes (3,966 nucleotide sites). The ML tree including *Colacium vesiculosum* as ingroup taxa resulted in the same tree topology. The position of *C. vesiculosum* of this analysis is marked with dotted line. The numbers on each node represents ML bootstrap support (bs) without *Colacium vesiculosum* (left) or with *C. vesiculosum* (right), only one number on the node displayed consistency between the analyses. Bs values below 70 are marked with dashes. * = no bs value on the left side, because *C. vesiculosum* was excluded of this analysis. Taxon names are given. Clade A is further divided into A1 and A2. The scale bar represents 0.07 substitutions/ site.

The inconsistent position of *Colacium* is not new to euglenoid researchers. For instance, the only other phylogenomic analysis performed by Bennett et al. (2014) included 79 chloroplast genes and *C. vesiculosum* branched as sister to *Euglena gracilis* and *Euglena viridis*. Also the multigene analyses of Kim et al. (2015) showed that ten species of the genus *Colacium* formed a strongly-supported monophyletic clade (named G) that was sister to six strains of *Euglena* (named A3). These two clades together formed a sister relationship with the two loricate genera *Strombomonas* and *Trachelomonas*. Further proof for the ambiguous position of *Colacium* brought the phylogenetic tree of combined 16S and 18S rDNA by Milanowski et al. (2006), where *Colacium vesiculosum* formed a clade together with *Cryptoglena pigra*, *Euglena anabaena* and the genus *Monomorphina.* This clade was sister to the genera *Strombomonas* and *Trachelomonas*. Then again in some analyses (Karnkowska et al. 2015, Kim et al. 2010) the position presumed on behalf of morphological characters close to the loricate genera was found. Overall, the position of *Colacium* is still inconclusive, obviously depending on the applied sequence data and taxon sampling (Karnkowska et al. 2015, Kim et al. 2010 & 2015, Marin et al. 2003, Milanowski et al. 2006). A solution to this problem may be further analyses of more chloroplast genomes of the genus *Colacium*.

The backbone of the tree (Fig. 2) was highly supported as were the clades of the ingroup as well as sister group relationships. At the base the two marine *Eutreptiella* species formed a highly supported clade next to the Pyramimonadales as the donor of the euglenoid chloroplasts. *Eutreptia viridis* was positioned paraphyletically in regard to Eutreptiales as sister to all freshwater photosynthetic euglenoids. This node was highly supported and the topology was consistent with phylogenomic analyses of Bennett et al. (2014) and phylogenetic ML analyses of Marin et al. (2003). But also contradictory data exist (Yamaguchi et al. 2012), where Eutreptiales were displayed weakly supported (bs 51) in a monophyletic clade and *Eutreptia viridis* was sister to the two *Eutreptiella* taxa.

Within the freshwater photosynthetic euglenoids two well supported main clades were recovered: The Phacaceae in this analysis only represented by *Phacus orbicularis* (Kasiborski et al. 2016), and the Euglenaceae. Inside the family Euglenaceae *Euglenaformis proxima* branched at the base (Bennett et al. 2014). The genus *Euglena* was not monophyletic and split into one major subclade A1, highly supported, and the newly sequenced *Euglena archaeoplastidiata* A2 (Bennett et al. 2017) next to *Euglenaria anabaena* (Bennett & Triemer 2015) (Fig. 2). In clade B *Trachelomonas* and *Strombomonas*, both known as species enclosed in a mineralized lorica, were sisters to each other. Also, *Monomorphina* and *Cryptoglena* shared a common ancestor and formed a well-supported sister clade and were likewise part of clade B (Fig. 2). These sistergroup relationships were also detected in former multigene analyses (Karnkowska et al. 2015, Kim et al. 2015, Linton et al. 2010) and can be regarded as stable. Sisters to the latter were *Euglenaria anabaena* and *Euglena archeoplastidiata*, as mentioned previously by Karnkowska et al. (2015) and Linton et al. (2010). In addition to our phylogenomic results, Bennett et al. (2017) detected features in the cpGenome, which would indicate that *E. archaeoplastidiata* is erroneously allocated to the genus *Euglena*. A particularity of *E. archaeoplastidiata* segregating it from clade A, is the presence of *mat*5 within *psb*A intron 1 found by Bennett et al. (2017), which is absent in the other members of the genus *Euglena* examined (Fig. 3).



| | mat1 | mat2 | mat5 | RT-Mat-En | RT-G2-Mat |
|---|---|---|---|---|---|
| *Euglena gracilis* | ■ | ■ | □ | □ | □ |
| *Euglena gracilis v. bacillaris* | ■ | ■ | □ | □ | □ |
| *Euglena mutabilis* | ■ | ■ | □ | □ | □ |
| *Euglena viridis* ATCC | ■ | ■ | □ | □ | □ |
| *Euglena viridis* SAG | ■ | ■ | □ | □ | □ |
| **Monomorphina aenigmatica* | ■ | ■ | □ | □ | □ |
| **Monomorphina parapyrum* | ■ | ■ | ■ | □ | □ |
| **Cryptoglena skujae* | ■ | ■ | □ | ■ | □ |
| **Euglenaria anabaena* | ■ | ■ | ■ | □ | □ |
| **Euglena archaeoplastidiata* | ■ | ■ | ■ | □ | □ |
| *Colacium vesiculosum* | ■ | ■ | ■ | □ | □ |
| **Strombomonas acuminata* | ■ | ■ | ■ | □ | □ |
| *Trachelomonas grandis* | ■ | ■ | ■ | □ | ■ |
| *Trachelomonas volvocina* | ■ | ■ | ■ | □ | □ |
| *Euglenaformis proxima* | ■ | ■ | □ | □ | □ |
| *Phacus orbicularis* | ■ | □ | □ | □ | □ |
| *Eutreptia viridis* | ■ | □ | □ | □ | □ |
| *Eutreptiella gymnastica* | ■ | □ | □ | □ | □ |
| *Eutreptiella pomquetensis* | ■ | □ | □ | □ | □ |
| outgroup | | | | | |

**Fig. 3:** Schematic phylogeny of phototrophic euglenoids with *Colacium vesiculosum* as sister to *Euglenaria anabaena* and *Euglena archaeoplastidiata* demonstrating progressive Mauve synteny analysis comparing the chloroplast genomes of each clade and between different taxa (bracket with number of resulting clusters (c)). * represent taxa that can be included in one analysis resulting in one cluster. 2 on the diagrammatic phylogeny indicates the presence of two rRNA operons, 1 shows loss of one rRNA operon, and the dots indicate independent achievement of further rRNA operons. Existence of maturases within the chloroplast genomes is illustrated with squares: gray squares code presence and blanks absence according to GenBank annotation of each Genome.

Moreover, the syntenic genome arrangement of *Euglena archaeoplastidita* was equal to *Monomorphina*, *Cryptoglena*, *Euglenaria* and *Strombomonas* and in contrast differs highly to the cluster arrangement of the other *Euglena* cpGenomes (Bennett et al. 2017) (Fig. 3 asteriks). A substantial re-analysis of *E. archaeoplastidiata* seems appropriate, because comparative morphological features like diplopyrenoid, small paramylon bodies and metaboly resemble *Euglena*, but the single parietal chloroplast is found in *Monomorphina* and *Cryptoglena* and not in *Euglena* (Ciugulea & Triemer 2010, Karnkowska et al. 2015, Kosmala et al. 2007).

## Intrageneric variability of *Trachelomonas*

As it is known that intrageneric comparisons of the genus *Euglena* (with the exception of *E. archaeoplastidiata*) and the genus *Monomorphina* on one side show high synteny, whereas the *Eutreptiella* species on the other side comprise low synteny (Bennett et al. 2012 & 2017, Bennett & Triemer 2015, Dabbagh & Preisfeld 2017, Dabbagh et al. 2017, Hrdá et al. 2012, Pombert et al. 2012). In light of a high intrageneric diversity of nuclear genes in *Trachelomonas* (Marin et al. 2003), we chose another *Trachelomonas* species to explore a possible trend in intrageneric variation across the tree.

An arrangement comparison between the cpGenomes of *T. grandis* and *T. volvocina* with progressiveMauve resulted in five conserved gene clusters relative to the *T. grandis* cpGenome (Fig. 4, Table 1). Block C was the largest, including 52 genes, which is more than half of the total gene content. The five gene clusters showed significant rearrangement in position and strand orientation between *T. grandis* and *T. volvocina*. For example, apart from blocks B and C shifting position, the orientation was also changed. A further Mauve analysis of the two *Trachelomonas* species and *Strombomonas acuminata* (Wiegert et al. 2013) identified the same five conserved gene clusters as mentioned previously (Fig. 4, Table 1).

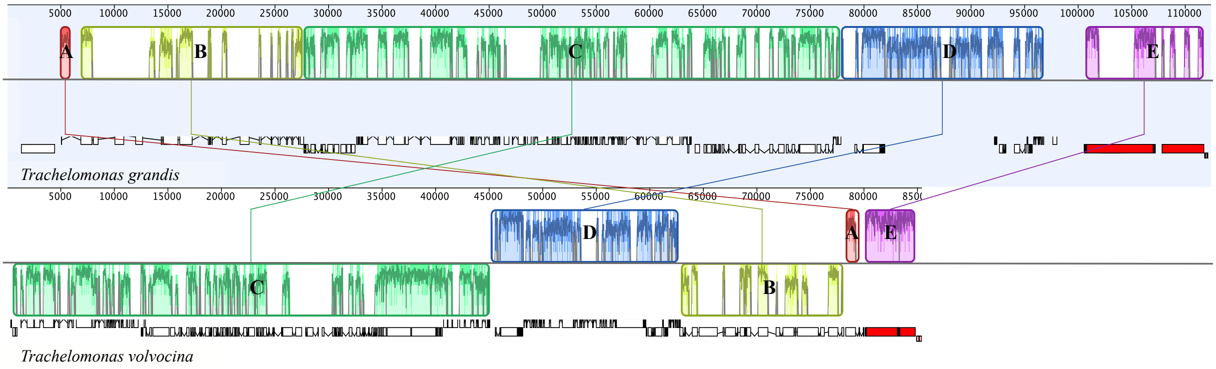


**Fig. 4:** Progressive Mauve analysis comparing the chloroplast genomes of *Trachelomonas grandis* and *Trachelomonas volvocina*. Linearized cpGenomes are shown with lettered boxes representing homologous gene clusters with *T. grandis* as the reference genome. Boxes oriented in the same direction as *T. grandis* lie above the horizontal line, while those below the line represent gene clusters that are oriented on the opposite strand relative to *T. grandis*. For a list of genes contained in each block see Table 1.

**Table 1:** Gene clusters of *Trachelomonas grandis* and *Trachelomonas volvocina* cpGenomes identified in progressive Mauve analysis (Fig. 4). Blocks (gene clusters) are labeled with letters (A-E) and genes are listed.

| Cluster | Conserved Gene Clusters |
|---|---|
| A | *chl*I |
| B | *ycf*65, *psb*A, tRNA-Leu, *psb*C, *psb*D |
| C | *psb*H, *psb*N, *pet*B, *atp*B, *atp*E, *rbc*L, *rpl*32, *psa*C, *rps*9, *rpl*12, *psb*Z, *rpo*A, tRNA-Ser, tRNA-Pro, *psa*J, *rps*18, *atp*A, *atp*F, *atp*H, *atp*I, *rps*2, tRNA-Cys, *rps*14, tRNA-Met, *rpl*36, *rps*8, *rpl*5, *rpl*14, *rpl*16, *roa*A, *rps*3, *rpl*22, *rps*19, *rpl*2, *rpl*23, *psb*J, *psb*L , *psb*F , *psb*E, *psa*B, *psa*A, tRNA-Lys, *pet*G, tRNA-Asp, *psb*I, *rpl*20, *rps*12, *rps*7, *tuf*A, *ycf*4, tRNA-Gln, tRNA-Ser |
| D | *psb*T, *psb*B, tRNA-Gly, tRNA-Glu, tRNA-Trp, tRNA-Met, tRNA-His, *rpo*B, *rpo*C1, *rpo*C2, tRNA-Val, tRNA-Asn, tRNA-Arg tRNA-Leu*, *rps*11, *rps*4, tRNA-Tyr, *psa*I, tRNA-Met, tRNA-Gly, tRNA-Thr, *psb*K, *psb*30, *psa*M, tRNA-Arg* |
| E | 5S rRNA, 23S rRNA, tRNA-Ala, tRNA-Ile, 16S rRNA |

Remarkably, when the gene arrangement of only one of the *Trachelomonas* species was compared to *S. acuminata,* only three or four conserved gene clusters arose. Comparing *S. acuminata* with the cpGenome of *T. grandis* reveals that the five blocks shown in Fig. 4 partly merged, resulting in three blocks, of which the largest block B contains 76 genes (Fig. 5). Then again, the arrangement comparison of *T. volvocina* and *S. acum*inata resulted in four blocks (data not shown). Interestingly this means, that the cluster arrangement between the two *Trachelomonas* species differed more than that between *S. acuminata* and each *Trachelomonas*. This indicates large gene rearrangements between *T. grandis* and *T. volvocina*.

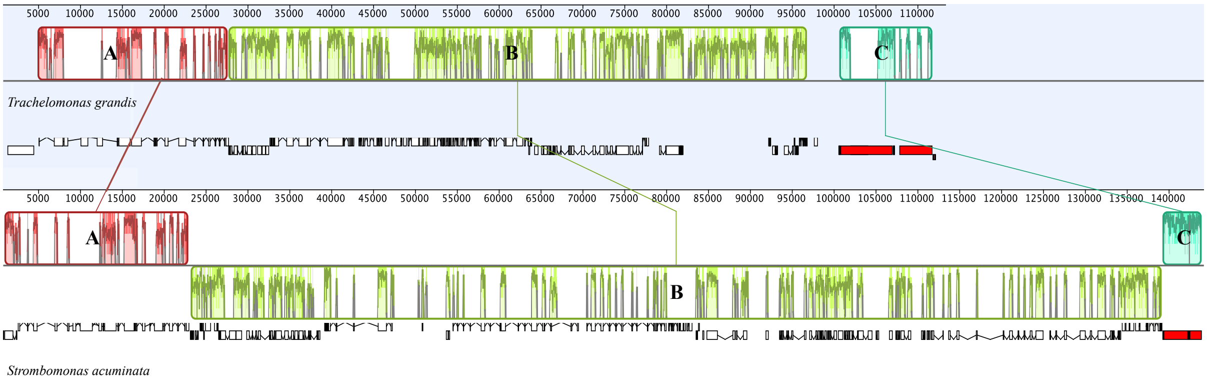


**Fig. 5:** Progressive Mauve synteny analysis comparing the chloroplast genomes of *Trachelomonas grandis* and *Stombomonas acuminata*. Gene clusters that are oriented in the same direction as *T. grandis* are shown above the horizontal line, whereas gene clusters that are oriented in the opposite direction lie below the line. Lettered boxes represent homologous gene clusters of the linearized cpGenomes relative to *T. grandis*. Box A: *chl*I – *ycf*65. Box B: *psa*M – tRNA-Ser. Box C: 5S rRNA – 16S rRNA.

Cluster analysis of *T. volvocina* with species of clade B resulted in more clusters than analysis of the latter with *T. grandis*. This may indicate that *T. grandis* is the taxon, in which rearrangements in form of merging of clusters occurred. Previous investigations of intrageneric variability in each of the genera *Monomorphina* and *Euglena* resulted in one and two clusters, respectively. (Bennett et al. 2012, Bennett & Triemer 2015, Dabbagh & Preisfeld 2017, Hallick et al. 1993, Pombert et al. 2012). These and other results imply that cluster analysis of the same genus of the Euglenales mostly yielded a low number of clusters. Our results show that the genus *Trachelomonas* contradicts this intrageneric trend.

An approach to bring cluster arrangements together with phylogenomic analysis (Fig. 3) did not reveal minimal cluster arrangements in clade B, when *C. vesiculosum* (dotted lines) was included. Only, when excluded, the same syntenic arrangement was found in clade B, as well as in *Strombomonas*, which is not located in the same clade (Bennett et al. 2015) (Fig. 3 asteriks). In an attempt to interpret these confusing results, it could implicate that the more the euglenoid taxa are derived, the more consistent the cluster arrangements became.

## Gene composition

Eighty-nine unique genes have been identified, including 60 protein-coding genes, 26 tRNAs and 3 rRNAs, matching the general genome features of other cpGenomes. Annotation revealed similarities that had been previously seen in other phototrophic euglenoid cpGenomes (Table S2). Unique for *T. grandis* is that there are two overlapping genes (*psb*D - *chl*I), in which *psb*D was oriented in clockwise direction and *chl*I counterclockwise. The cpGenome contains at least 68 introns (Table S2). For two genes, *rpo*B and *rpo*C1, the correct coding region could not be identified satisfactorily and RT-PCR experiments to detect the exon-intron boundaries of the introns failed. Therefore, we only took into account 68 introns and are aware that this number is considered a minimum. Although the intron amount among the *Trachelomonas* species was highly similar (*rpo*B and *rpo*C1 of both species were excluded from this analysis, resulting in 73 introns for *T. volvocina*) the average intron length of the cpGenome of *T. grandis* was almost twice the size (637 bp) in contrast to *T. volvocina* (372 bp). This is the second reason for size differences between both *Trachelomonas* cpGenomes.

## Maturases and ORFs

The *T. grandis* chloroplast genome contained three additional putative reverse transcriptase/maturases to *mat*1 (*ycf*13) (Fig. 3, Table 2). One was located within *psb*C I1 (*orf*147, *orf*276 and *orf*216) and BLASTP analyses revealed moderate matches for putative reverse transcriptase/maturase (*mat*2). The other was situated within *psb*A I2 (*orf*285). Again, BLASTP revealed moderate matches for putative reverse transcriptase/ maturase (*mat*5). All three were also identified in other phototrophic euglenoids (Bennett & Triemer 2015). The third additional maturase (*orf*681) was located in the intron of the 23S rRNA gene and is so far unique in euglenoid cpGenomes. Functional analysis showed moderate matches to group II intron reverse transcriptase/ maturase (Table 2). Secondary structure analyses confirmed that this maturase is likely located within domain IV of the group II intron (Fig. 6).

**Table 2:** BLASTP analysis of the reverse transcriptases/maturases and ORFs detected in the cpGenome of *Trachelomonas grandis* against the NCBI nonredundant protein sequences (nr) database. For each maturase and ORF the best match is reported.

| ORF | Accession no. | Best BLASTP Match | | |
|---|---|---|---|---|
| | | Organism | Product | E-value[1] |
| 285 | YP_009145553.1 | *Trachelomonas granids* | maturase-like protein | 1e-65 |
| 147 | YP_009144869.1 | *Euglenaria anabaena* | maturase-like protein | 0.005 |
| 276 | YP_009032719.2 | *Euglenaformis proxima* | maturase-like protein | 3e-15 |
| 216 | YP_009389097.1 | *Euglena archaeoplastidiata* | maturase-like protein | 1e-37 |
| 681 | WP_072717411.1 | *Planktothrix tepida* | group II intron reverse transcriptase/maturase | 2e-106 |
| 1034 | WP_023274716.1 | *Acinetobacter tjernbergiae* | lipid A hydroxylase LpxO | 6.9 |
| 127 | - | - | - | - |
| 113 | WP_066394260.1 | *Bacillus mesonae* | ROK family transcriptional regulator | 6.0 |
| 138* | OBA25748.1 | *Hanseniaspora valbyensis* | hypothetical protein HANVADRAFT_7761 | 6.6 |

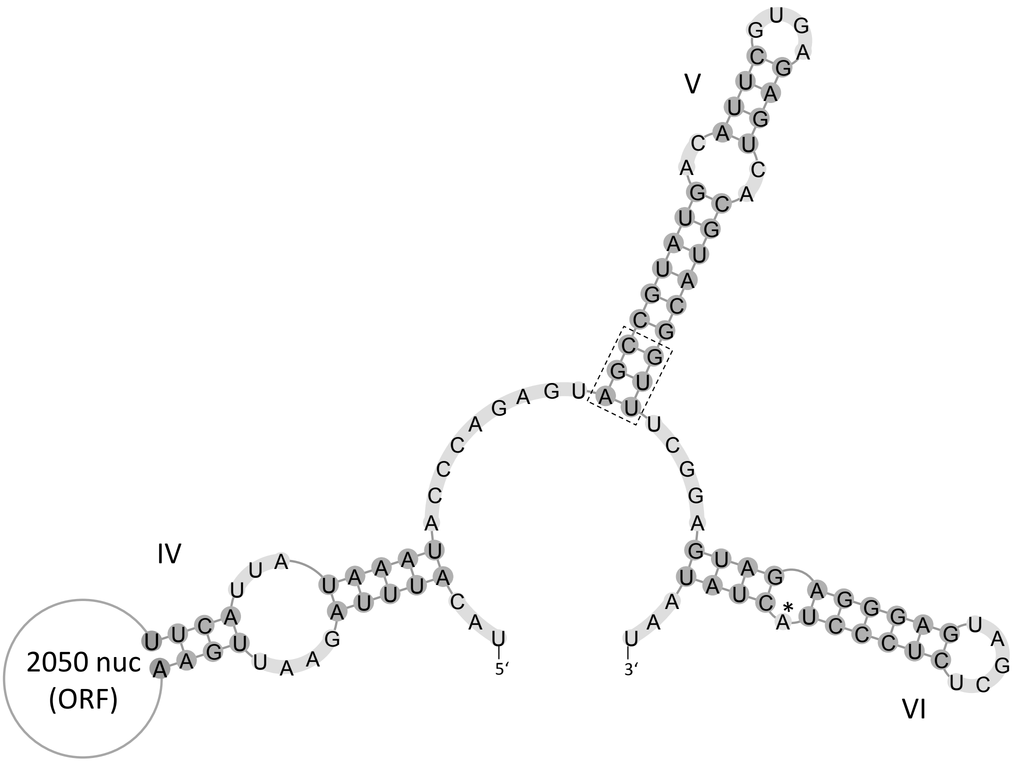


**Fig. 6:** Putative secondary structure model of domain IV - VI of the group II intron within the 23S rRNA gene. Domain IV contains the ORF region, the AGC triad is located near the base of the stem of domain V (dashed box) and domain VI contains the branch-point A (*).

In addition, four freestanding open reading frames (ORFs) have been identified and BLASTP analysis (NCBI) was performed (Table 2). The identified VNTR region next to the 16S rRNA was 332 bp long and composed of two repeat units, each 124 bp long, interrupted by a nonrepeating unit (Fig. 1).

## rRNA operon

The *Trachelomonas grandis* cpGenome contained only a single rRNA operon, like *T. volvocina* and the majority of phototrophic euglenoids investigated so far. In regard of the rRNA operon copies an evolutionary trend can be observed regarding the phylogeny of euglenoids. The most basally branching euglenoids of the genus *Eutreptiella* both contain two rRNA copies (Dabbagh et al. 2017, Hrdá et al. 2012) (Fig. 3). For *Eutreptia viridis* only one copy was detected, though sequencing coverage hints at rather more than one copy, which is corroborated by a failed circularization in this region (Wiegert et al. 2012). We decided to treat the number of rRNA operons at least as two with the consequence that presumably one rRNA operon got lost after the split between Euglenales and Eutreptiales occurred. Only for both *Euglena gracilis* species and for *Strombomonas acuminata* further copies of the rRNA operons were detected (Bennett & Triemer 2015, Hallick et al. 1993, Wiegert et al. 2013). The operons of Eutreptiales species largely resemble the operons and their strand orientation in *Pyramimonas parkeae* and *Ostreococcus tauri* (Dabbagh et al. 2017, Robbens et al. 2007, Turmel et al. 2009). We assume that one copy was lost during divergence of Euglenales, which is supported by the fact that only *S. acuminata* and both *E. gracilis* strains achieved, probably independently, two or more copies, whereas all other taxa show the plesiomorphic condition of a single copy operon, accept in *Phacus orbicularis* where the single copy rRNA genes are not arranged in an operon (Kasiborski et al. 2016). The three tandemly repeated rRNA copies of *E. gracilis* remain exceptional. Therefore we suppose that the second operon first disappeared and further operons were then added as tandemly structured operons in both *E. gracilis* strains and one operon (missing the 16S) in *S. acuminata* (Fig. 3, dot).

The rRNA operon of *T. grandis* was 5,814 bp larger than that of *T. volvocina* and it exceeds the range of each rRNA operon detected in sequenced cpGenomes of phototrophic euglenoids so far. This difference in size between the rRNA operons can be attributed to four introns localized within the 16S rRNA gene (3 introns) and one in the 23S rRNA gene of *T. grandis*. To date, *T. grandis* is the only species of sequenced cpGenomes of phototrophic euglenoids that contains introns within the rRNA operon and in different rRNA genes.

All four introns located in the genes of the rRNA operon were group II introns with a characteristic 5' - GUGYG boundary, a conserved ending AY - 3' and the presumed ‘branch-point’ *A for splicing in domain VI, where the first transesterification takes place (Lambowitz & Belfort 2015). Also, domain V, known to play a catalytic role in intron excision, showed a highly conserved stem near the base of domain V (5' - AGC...GUU…- 3'), as illustrated in a secondary structure model for DIV-DVI of the intron in the gene for 23S rRNA in Fig. 6 (Kelchner 2002, Lambowitz & Belfort 2015, Michel & Ferat 1995, Toor et al. 2001).

A comparison of 309 16S rRNA sequences from GenBank revealed only one group II intron in *Euglena cantabrica* SAG 1224-25, which was detected by Milanowski et al. (2006). Since the 16S rRNA sequence of *E. cantabrica* was incomplete, only an assumption about the insertion site could be made. The 5' exon boundary (GTGCCAGCAGC) next to the group II intron of *E. cantabrica* was highly conserved and detectable in *T. grandis* 23 basepairs upstream from the insertion site of the first intron. Apart from different insertion sites, the two introns varied in size and intron start. Although both introns showed a conserved 5'-..AGC-GUU… -3' stem near the base of domain V, the rest of domain V and VI was not related to each other. A comparison between the group II intron of *E. cantabrica* and intron 2 and intron 3 of the 16S rRNA gene of *T. grandis* did not reveal further sequence similarity. Thus, we regard the insertions in both taxa as unrelated events.

One hundred and thirteen 23S rRNA sequences of phototrophic euglenoids from GenBank have been examined and compared to the 23S rRNA gene of *T. grandis*. Only in the incomplete sequence of *Trachelomonas bernadiniensis* ACOI1103 was a group II intron detected (Kim et al. 2010). With 945 bp this intron was more than 2,400 bp smaller than the group II intron of the 23S gene of *T. grandis* and did not share the same insertion site. The intron of *T. bernadiniensis* was located next to the highly conserved 5' -GATAAAAGTT sequence, which was also detected in exon 2 of the 23S of *T. grandis*.

In conclusion, the two other identified introns in rRNA genes of *E. cantabrica* and *T. bernadiniensis* were not related to the ones found in *T. grandis*. The exhibition of four introns in 16S and 23S genes is unique to phototrophic euglenoids. In contrast to some introns, which spread across the tree vertically and are thus to be found in several taxa, these introns seem to originate late in the stem line leading to *T. grandis*.

**Table 3:** Intron insertion sites that are shared between the two *Trachelomonas* species.
$^{a}$ slight insertion site difference indicated by the number of base pairs shown in brackets.

| Genes | Shared insertion site $^{(a)}$ |
|---|---|
| *atp*E | I1 |
| *atp*H | I1 $^{(2)}$ |
| *pet*B | I1 |
| *psa*C | I1 |
| *psb*T | I1 |
| *psb*30 | I1 |
| *rpl*23 | I1 $^{(2)}$ |
| *rps*2 | I1 $^{(5)}$, I2 $^{(2)}$, I3 $^{(3)}$ |
| *rps*3 | I1 |
| *rps*11 | I1 $^{(3)}$ |
| *rps*14 | I1 $^{(3)}$ |
| *rps*19 | I2 |
| *tuf*A | I1 |
| *ycf*4 | I1 $^{(3)}$ |

## Introns within protein-coding genes

All introns of protein-coding genes of *T. grandis* (apart from *rpo*B and *rpo*C1) have been examined according to the size and the conserved catalytic triad AGC of domain V and manually divided into putative group II, group III or mini group II introns (Table S3). Comparisons and investigations of insertion sites between the introns of protein-coding genes of *T. grandis* and *T. volvocina* revealed overall 16 introns with identical insertion sites (Table 3, not included in the analyses were *rpo*B, *rpo*C1, *mat*2 *mat*5) after some small corrections to the annotation of *T. volvocina* (alignments are available upon request from the authors). That means under 15 % of the introns located in *Trachelomonas grandis* had a homologous position to *T. volvocina*. This supports the intrageneric variability of the genus *Trachelomonas* in contrast to the genus *Monomorphina*, where over 60 % of the introns in *M. parapyrum* shared the same position with *M. aenigmatica* (Bennett et al. 2015).

Moreover, the insertion of two new introns in one of the two species, respectively, supported the high degree of variability in *Trachelomonas* (Fig. 7). The protein-coding gene *psb*A of *T. volvocina* consisted of three introns (Bennett & Triemer 2015). The first intron contained *mat*5. In the cpGenome of *T. grandis* the *psb*A gene contained seven introns with *mat*5 located in intron 2. A comparison of both introns, in which *mat*5 is located, revealed two conspicuities: First, the insertion site of *psb*A I2 of *T. grandis* was at position 1,348 upstream of base one. This position is equivalent to the summation of the insertion site of *T. volvocina* (i397) with the length of the first intron of *T. grandis* (951 bp). A scenario explaining this could be the insertion of a new intron in the first exon of *psb*A of *T. grandis*, so that the former intron 1 with *mat*5 became intron 2. Hence, the former insertion site of 397 bp became 1,348 bp (Fig. 7a).

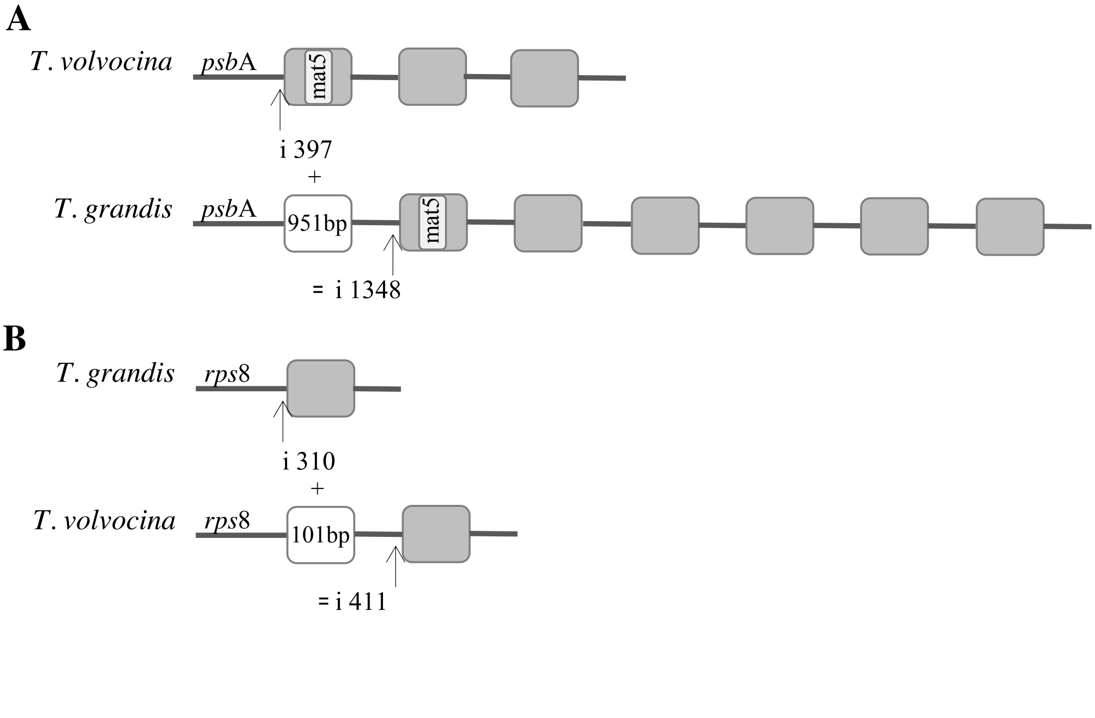


**Fig. 7:** Visualization of putative new introns within the coding region of protein coding genes. A) The white box shows the newly inserted Intron within the coding region of *psb*A of *Trachelomonas grandis*, so that the former intron 1 with *mat*5 became intron 2. B) The white box shows the newly inserted Intron within the coding region of *rps*8 of *Trachelomonas volvocina* so that the former intron (gray) became intron 2.

Second, we identified that the 3' end of intron 7 of *psb*A in *T. grandis* possessed the same bases as the 3' end of intron 3 of *psb*A in *T. volvocina* (TTAGTTTAAC- 3'), assuming that further introns proliferated between intron 2 containing *mat*5 and the intron with the TTAGTTTAAC- 3' of *T. grandis*. The same proliferation process was recognized in *rps*8 of *T. volvocina* (Fig. 7b). Another presumption could be that intron loss instead of intron gain occurred within the two genes, and then the question remained open, which is more likely: Intron loss or intron gain?

## Observed twintron trends

As expected, the first two twintron trends established by Bennett & Triemer (2015) in their characterization of Euglenaceae could be confirmed. We also did not detect a *psb*F intron and the intron in *psb*D of *T. grandis* was not homologous to that of *E. gracilis* and *psb*C I1 was identified.

The third trend after Bennett & Triemer (2015) that the *pet*B intron/ twintron may be a synapomorphy for just the Euglenaceae was already refuted by investigations on the cpGenomes of *Phacus orbicularis* (Kasiborski et al. 2016) and *Eutreptiella pomquetensis* (Dabbagh et al. 2017) because both *pet*B coding regions were interrupted by introns. The cpGenome of *T. grandis* comprised two introns in the *pet*B gene. Interestingly, all phototrophic euglenoids sequenced so far, except the two Eutreptiales *Et. viridis* and *Etl. gymnastica* (Hrdá et al. 2012, Pombert et al. 2012, Wiegert et al. 2012), contained at least one intron in the *pet*B gene. Comparing all *pet*B genes of these phototrophic euglenoids revealed that the insertion size is always the same (i23, Table S2), although intron size varied extremely due to interruption of some *pet*B coding regions by introns or twintrons. The hypothesis of a possible trend should be changed as follows: If a *pet*B intron/ twintron is present in the coding region, it is inserted consistently in the gene after the 23rd base of the coding region in all phototrophic euglenoids. These introns seem to be closely related and spread presumably vertically by diversification of species and not horizontally as most euglenoid introns appeared.

## Conclusion

Although the general genome characters correspond well with those of all other Euglenaceae, several differences can be found in the two species of *Trachelomonas*: Only *T. grandis* contained introns in the 16S and 23S rRNA genes and an additional maturase (RT-G2-Mat). Since these were not detectable in any other euglenoid species, the presence seems to be a single event having occurred exclusively in the cpGenomes of *T. grandis*. A multiple loss in other taxa seems to be less parsimonious. The larger cpGenome of *T. grandis* can be attributed to more intergenic space and larger introns accompanied by low resemblance of insertion sites to *T. volvocina*.

The cluster arrangement (Mauve) was more consistent with *Strombomonas acuminata* than with *T. volvocina,* a member of the same genus, and stands in contrast to intrageneric synteny in the genera *Monomorphina* or *Euglena*. It reflected more the intrageneric variability of the genus *Eutreptiella*. At the moment, with only 19 genomes sequenced, no cluster arrangement trend fully corresponds to the results of the presented phylogenomic analysis, although this supported the considerations of Bennett et al. (2017) to place *E. archaeoplastidiata* outside of the genus *Euglena*. The most appropriate statement today is that the more the euglenoid taxa are derived, the more consistent the genome arrangements become.

## ACKNOWLEDGMENTS

The authors would like to thank Margret Buse and Melanie Beudels for critical reading and Sabine Stratmann-Lettner and Heiko Helmich for lab assistance. This work was supported by a postgraduate scholarship of Bergische University Wuppertal to ND.

## References

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., Le Gall, L., Lynn, D. H., McManus, H., Mitchell, E. A. D., Mozley-Stanridge, S. E., Parfrey, L. W., Pawlowski, J., Rueckert, S., Shadwick, L., Shadwick, L., Schoch, C. L., Smirnov, A. & Spiegel, F. W. 2012. The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology*, 59:429–493. doi: 10.1111/j.1550-7408.2012.00644.x.

Bennett, M. S. & Triemer, R. E. 2015. Chloroplast Genome Evolution in the Euglenaceae. *J. Eukaryot. Microbiol.*:n/a. doi: 10.1111/jeu.12235.

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

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*, 53:66–73. doi: 10.2216/13-198.1.

Bennett, M. S., Shiu, S.-H. & Triemer, R. E. 2017. A rare case of plastid protein-coding gene duplication in the chloroplast genome of *Euglena archaeoplastidiata* (Euglenophyta). *J Phycol*, 53:493–502. doi: 10.1111/jpy.12531.

Bicudo, C. E. d. M. & Menezes, M. 2016. Phylogeny and Classification of Euglenophyceae: A Brief Review. *Front. Ecol. Evol.*, 4:429. doi: 10.3389/fevo.2016.00017.

Brosnan, S., Shin, W., Kjer, K. M. & Triemer, R. E. 2003. Phylogeny of the photosynthetic euglenophytes inferred from the nuclear SSU and partial LSU rDNA. *International Journal of Systematic and Evolutionary Microbiology*, 53:1175–1186. doi: 10.1099/ijs.0.02518-0.

Brosnan, S., Brown, P. J. P., Farmer, M. A. & Triemer, R. E. 2005. Morphological separation of the euglenoid genera *Trachelomonas* and *Strombomonas* (Euglenophyta) based on lorica development and posterior strip reduction. *Journal of Phycology*, 41:590–605. doi: 10.1111/j.1529-8817.2005.00068.x.

Brown, P. J. P., Zakry, B.e. & Farmer, M. A. 2003. Plastid morphology, ultrastructure, and development in *Colacium* and the loricate Euglenophytes (Euglenophyceae). *Journal of Phycology*, 39:115–121. doi: 10.1046/j.1529-8817.2003.01244.x.

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P. & Bateman, A. 2013. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*, 41:D226-32. doi: 10.1093/nar/gks1005.

Busse, I. & Preisfeld, A. 2002. Unusually expanded SSU ribosomal DNA of primary osmotrophic euglenids: molecular evolution and phylogenetic inference. *Journal of Molecular Evolution*, 55:757–767. doi: 10.1007/s00239-002-2371-8.

Chernomor, O., Haeseler, A. von & Minh, B. Q. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, 65:997–1008. doi: 10.1093/sysbio/syw037.

Ciugulea, I. & Triemer, R. E. 2010. A color atlas of photosynthetic euglenoids. Michigan State University Press, East Lansing. xx, 204.

Ciugulea, I., Nudelman, M. A., Brosnan, S. & Triemer, R. E. 2008. Phylogeny of the euglenoid loricate genera *Trachelomonas* and *Strombomonas* (Euglenophyta) inferred from nuclear SSU and LSU rDNA. *Journal of Phycology*, 44:406–418. doi: 10.1111/j.1529-8817.2008.00472.x.

Dabbagh, N. & Preisfeld, A. 2017. The Chloroplast Genome of *Euglena mutabilis* -Cluster Arrangement, Intron Analysis, and Intrageneric Trends. *Journal of Eukaryotic Microbiology*, 64:31–44. doi: 10.1111/jeu.12334.

Dabbagh, N., Bennett, M. S., Triemer, R. E. & Preisfeld, A. 2017. Chloroplast genome expansion by intron multiplication in the basal psychrophilic euglenoid *Eutreptiella pomquetensis*. *PeerJ*, 5:e3725. doi: 10.7717/peerj.3725.

Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14:1394–1403. doi: 10.1101/gr.2289704.

Deflandre, G. 1930. *Strombomonas* nouveau genere d'euglénacées (Trachelomonas Ehr. pro parte). *Archiv für Protistenkunde*, 69:551–614.

Ehrenberg, C. G. 1830. Organisation, systematik und geographisches verhältniss der infusionsthierchen. Zwei vorträge, in der Akademie der wissenschaften zu Berlin gehalten in den jahren 1828 und 1830, von C.G. Ehrenberg. Mit 8 kupfertafeln in folio. Druckerei der Königlichen akademie der wissenschaften, Berlin.

Ehrenberg, C. G. 1833. Dritter Beitrag zur Erkenntnis grosser Organisation in der Richtung des kleinsten Raumes. Königlichen Akademie der Wissenschaften, Berlin.

Gibbs, S. P. 1978. The chloroplasts of *Euglena* may have evolved from symbiotic green algae. *Can. J. Bot.*, 56:2883–2889. doi: 10.1139/b78-345.

Gibbs, S. P. 1981. The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. *Ann N Y Acad Sci*, 361:193–208. doi: 10.1111/j.1749-6632.1981.tb54365.x.

Hallick, R. B., Hong, L., Drager, R. G., Favreau, M. R., Monfort, A., Orsat, B., Spielmann, A. & Stutz, E. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Research*, 21:3537–3544. doi: 10.1093/nar/21.15.3537.

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:e33746. doi: 10.1371/journal.pone.0033746.

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*, 62:362–373. doi: 10.1111/jeu.12192.

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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28:1647–1649. doi: 10.1093/bioinformatics/bts199.

Kelchner, S. A. 2002. Group II introns as phylogenetic tools: Structure, function, and evolutionary constraints. *American Journal of Botany*, 89:1651–1669. doi: 10.3732/ajb.89.10.1651.

Kies, L. 1967. Oogamie bei *Eremosphaera viridis* De Bary. *Flora oder Allgemeine botanische Zeitung. Abt. B, Morphologie und Geobotanik*, 157:1–12.

Kim, J. I., Shin, W. & Triemer, R. E. 2010. Multigene analyses of photosynthetic euglenoids and new family Phacaceae (Euglenales). *Journal of Phycology*, 46:1278–1287. doi: 10.1111/j.1529-8817.2010.00910.x.

Kim, J. I., Linton, E. W. & Shin, W. 2015. Taxon-rich multigene phylogeny of the photosynthetic euglenoids (Euglenophyceae). *Front. Ecol. Evol.*, 3:254. doi: 10.3389/fevo.2015.00098.

Kosmala, S., Milanowski, R., Brzóska, K., Pękala, M., Kwiatowski, J. & Zakryś, B. 2007. Phylogeny and systematics of the genus *Monomorphina* (Euglenaceae) based on morphological and molecular data. *Journal of Phycology*, 43:171–185. doi: 10.1111/j.1529-8817.2006.00298.x.

Kumar, S., Stecher, G. & Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, 33:1870–1874. doi: 10.1093/molbev/msw054.

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T. & Ussery, D. W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35:3100–3108. doi: 10.1093/nar/gkm160.

Lambowitz, A. M. & Belfort, M. 2015. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiology Spectrum*, 3doi: 10.1128/microbiolspec.MDNA3-0050-2014.

Leedale, G. F. 1967. Euglenoid Flagellates. Prentice-Hall, Engelwood-Cliffs, New Jersey. 242.

Linton, E. W., Karnkowska-Ishikawa, A., Im Kim, J., Shin, W., Bennett, M. S., Kwiatowski, J., Zakryś, B. & Triemer, R. E. 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*, 161:603–619. doi: 10.1016/j.protis.2010.02.002.

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*, 154:99–145.

Michel, F. & Ferat, J. L. 1995. Structure and activities of group II introns. *Annual Reviews of Biochemistry*, 64:435–461. doi: 10.1146/annurev.bi.64.070195.002251.

Milanowski, R., Kosmala, S., Zakryś, B. & Kwiatowski, J. 2006. Phylogeny of photosyntetic Euglenophytes based on combined chloroplast and cytoplasmic SSU rDNA sequence analysis. *Journal of Phycology*, 42:721–730. doi: 10.1111/j.1529-8817.2006.00216.x.

Minh, B. Q., Nguyen, M. A. T. & Haeseler, A. von 2013. Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and. Evolution*, 30:1188–1195. doi: 10.1093/molbev/mst024.

Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32:268–274. doi: 10.1093/molbev/msu300.

Nudelman, M. A., Rossi, M. S., Conforti, V. & Triemer, R. E. 2003. Phylogeny of Euglenophyceae based on small subunit rDNA sequences: Taxonomic implications. *Journal of Phycology*, 39:226–235. doi: 10.1046/j.1529-8817.2003.02075.x.

Pombert, J.-F., James, E. R., Janouškovec, J., Keeling, P. J. & McCutcheon, 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:e53433. doi: 10.1371/journal.pone.0053433.

Poniewozik, M. 2017. Element Composition of *Trachelomonas* Envelopes (Euglenophyta). *Polish Botanical Journal*, 62:1. doi: 10.1515/pbj-2017-0007.

Posada, D. & Buckley, T. R. 2004. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53:793–808. doi: 10.1080/10635150490522304.

Robbens, S., Derelle, E., Ferraz, C., Wuyts, J., Moreau, H. & Van de Peer, Yves 2007. The complete chloroplast and mitochondrial DNA sequence of *Ostreococcus tauri*: Organelle genomes of the smallest eukaryote are examples of compaction. *Molecular Biology and Evolution*, 24:956–968. doi: 10.1093/molbev/msm012.

Rosowski, J. R. & Coute, A. 1996. Bacteria of the lorica of *Trachelomonas* occur in nature, not just in culture. *Journal of Phycology*, 32:697–698. doi: 10.1111/j.0022-3646.1996.00697.x.

Rosowski, J. R. & Langenberg, W. G. 1994. The near-spineless *Trachelomonas grandis* (Euglenophyceae) superficially appears spiny by attracting bacteria to its surface. *Journal of Phycology*, 30:1012–1022. doi: 10.1111/j.0022-3646.1994.01012.x.

Schattner, P., Brooks, A. N. & Lowe, T. M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research*, 33:W686-W689. doi: 10.1093/nar/gki366.

Singh, K. P. 1956. Studies in the genus *Trachelomonas*. I. Description of six organisms in cultivation. *American Journal of Botany*, 43:258–266.

Toor, N., Hausner, G. & Zimmerly, S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA*, 7:1142–1152.

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

Trifinopoulos, J., Nguyen, L.-T., Haeseler, A. von & Minh, B. Q. 2016. W-IQ-TREE: A fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, 44:W232-5. doi: 10.1093/nar/gkw256.

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*, 26:631–648. doi: 10.1093/molbev/msn285.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. & Rozen, S. G. 2012. Primer3--new capabilities and interfaces. *Nucleic Acids Research*, 40:e115. doi: 10.1093/nar/gks596.

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*, 163:832–843. doi: 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*, 60:214–221. doi: 10.1111/jeu.12025.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V. & Lipman, D. J. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A*, 106:7273–7280. doi: 10.1073/pnas.0901808106.

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 Evol Biol*, 12:29. doi: 10.1186/1471-2148-12-29.

## SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article:

**Table S1:** Sampling of phototrophic euglenoids and green algae outgroup taxa used for alignments of phylogenomic analyses and belonging accession numbers, sorted alphabetically.

| No. | Name/ Taxon | Accession number |
|---|---|---|
| 1 | *Chlorokybus atmophyticus* | DQ422812 |
| 2 | *Colacium vesiculosum* | JN674636 |
| 3 | *Cryptoglena skujae* | KP410781 |
| 4 | *Cymbomonas tetramitiformis* | KX013545 |
| 5 | *Euglenaformis proxima* | KC684276 |
| 6 | *Euglena gracilis* | X70810 |
| 7 | *Euglena gracilis var. bacillaris* | KP686076 |
| 8 | *Euglena mutabilis* | KT223519 |
| 9 | *Euglenaria anabaena* | KP453743 |
| 10 | *Euglena viridis* | JQ237893 |
| 11 | *Euglena viridis* | KP686075 |
| 12 | *Eutreptia viridis* | JN643723 |
| 13 | *Eutreptiella gymnastica* | NC_017754 |
| 14 | *Eutreptiella pomquetensis* | KY706202 |
| 15 | *Mesostigma viride* | AF166114 |
| 16 | *Monomastix sp.* | FJ493497 |
| 17 | *Monomorphina aenigmatica* | JX457480 |
| 18 | *Monomorphina parapyrum* | KP455987 |
| 19 | *Nephroselmis astigmatica* | KJ746600 |
| 20 | *Nephroselmis olivacea* | AF137379 |
| 21 | *Ostreococcus tauri* | CR954199 |
| 22 | *Palmophyllum crassum* | AP017927 |
| 23 | *Phacus orbicularis* | KR921747 |
| 24 | *Picocystis salinarum* | KJ746599 |
| 25 | *Prasinococcus sp.* | KJ746597 |
| 26 | *Prasinoderma coloniale* | KJ746598 |
| 27 | *Prasinophyceae sp.* | KJ746601 |

| | | |
|---|---|---|
| 28 | *Prasinophyceae sp.* | KJ746602 |
| 29 | *Pycnococcus provasolii* | FJ493498 |
| 30 | *Pyramimonas parkeae* | FJ493499 |
| 31 | *Scherffelia dubia* | KU167098 |
| 32 | *Strombomonas acuminata* | JN674637 |
| 33 | *Tetraselmis sp.* | KU167097 |
| 34 | *Trachelomonas grandis* | Accession pending |
| 35 | *Trachelomonas volvocina* | KP686077 |
| 36 | *Verdigellas peltata* | LT174527 |

**Table S2:** CpGenome features of euglenoids and depicted prasinophytes according to NCBI annotation.

* Chloroplast circle not closed. a:Twintrons were counted as single insertion sites, b: Including rRNA repeats and intermediate tRNAs, c: Includes the identified 5S, for *S. acuminata* the two identified 5S, d: Includes the two introns in rps18, e: Includes the one intercistronic intron *rps*4-*rps*11, f: First exon could not be identified, so gene length is a minimum, g: Realigned, but with alternative start codon, h: Start codon not determined, due to undetermined exon1, i: New intron start after re-analyses. k: aware that this number of introns is a minimum, not included are introns within *rpo*B and *rpo*C1.

| Taxon | Size (bp) | A+T % | Genes | Introns[a] | ORFs | *roa*A | CDS of *rpo*A (bp) | Largest gene (bp) | Shortest gene (bp) | Gene with most introns | *psb*D/*psb*C overlap | *pet*B I1 bp/5´start |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *E. gracilis* strain *Z* | 143,171 | 73.9 | 116[b] | 134 | 7 | + | 651 | *psb*C (10861) | *psa*M (96) | *rpo*C1 (11) | + | 909/GUGCG |
| *E. gracilis* var. *bacillaris* | 132,034* | 74.2 | 104 | 134 | 6 | + | 525 | *psb*C (11446) | *psa*M (96) | *rpo*C1 (11) | + | 909/GUGCG |
| *E. viridis* | 76,156 | 73.8 | 92 | 77 | 0 | + | 483 | *psb*C (6165) | *psa*M (96) | *rpo*B (10) | + | 528/GUGUG |
| *E. viridis epitype* | 91,616 | 73.6 | 92 | 76 | 13 | + | 480 | *psb*C (6131) | *psa*M (96) | *rpo*C1(8) | + | 533/GUGUG |
| *E. mutabilis* | 86,975 | 73.3 | 91 | 77[e] | 0 | + | 636 | *psb*C (12192) | *psa*M (96) | *rpo*C1 (9) | + | 435/GUGCG |
| *Era. anabaena* | 88,487 | 72 | 93 | 82 | 4 | + | 624 | *psb*A (6998) | *psa*M (96) | *rpo*B (10) | + | 551/GUGCG |
| *M. parapyrum* | 80,147 | 72 | 93 | 80 | 1 | + | 507 | *psb*C (6127) | *psa*M (96) | *rpo*B/-C1 (9) | + | 441/GUGCG |
| *M. aenigmatica* | 74,746 | 70.6 | 93[c] | 53 | 1 | + | - | *psb*C (6229) | *psa*M (96) | *rpo*C1 (9) | + | 544/GUGCG |
| *Cr. skujae* | 106,843 | 73.7 | 93 | 84 | 4 | + | 489 | *psb*C (8947) | *psa*M (96) | *rpo*B (10) | + | 537/GUGUG |
| *S. acuminata* | 144,166* | 73.4 | 95[c] | 112[d] | 0 | + | 486 | *psa*B (11283) | *psb*T (90) | *rpo*B/ *rbc*L (9) | + | 510/GUGCG |
| *T. volvocina* | 85,392* | 72.7 | 93 | 94 | 1 | + | 543 | *psb*C (7698) | *psa*M (96) | *rpo*C1 (11) | + | 536/GUGCG |
| ***T. grandis*** | **113,311** | **73.5** | **89** | **68[k]** | **4** | + | **-** | ***psb*C (11055)** | ***psa*M (96)** | | + | **23/410/GUGCG** |

**Table S2:** Continued.

| Taxon | Size (bp) | A+T % | Genes | Introns[a] | ORFs | *roa*A | CDS of *rpo*A (bp) | Largest gene (bp) | Shortest gene (bp) | Gene with most introns | *psb*D/*psb*C overlap | *pet*B I1 bp/5´start |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *C. vesiculosum* | 128,892* | 73.9 | 92[c] | 128 | 6 | + | - | *psb*C (11567) | *psb*I (105) | *rpo*C1 (11) | + | 470/GUGUG |
| *Efs. proxima* | 94,185* | 73.1 | 91 | 113 | 2 | + | - | *psb*C[f](6648) | *psa*M (96) | *rpo*B (10) | [h] | 478/GUGCG |
| *P. orbicularis* | 65,992 | 72,8 | 91 | 66 | 1 | + | - | *rpo*B (3993) | *psa*M (96) | *rpo*B (10) | + | 357/GUGCG[i] |
| *Et. viridis* | 65,523* | 71.4 | 83 | 24 | 3 | - | 672 | *psb*C (5706) | *psb*T (96) | *rpo*B (5) | - | - |
| *Etl. gymnastica* | 67,623 | 65.7 | 89 | 7 | 4 | - | 846 | *psb*C | *psa*M/*psb*T (96) | *psb*C (2) | +[g] | - |
| *Etl. pomquetensis* | 130,561* | 64.9 | 94 | 51 | 10 | | | *psb*D (8412) | *psa*M (96) | *psa*A (6) | + | 1736/GUGUG |
| *Pyramomonas parkae* | 101,605 | 65.3 | 125 | 1 | 5 | - | 1056 | *atp*B (4224) | *psb*T (96) | *atp*B (1) | + | - |
| *Pyconococcus provasolii* | 80,211 | 60.5 | 98 | 1 | 3 | - | 1074 | *fts*H (7944) | *psb*T (96) | *atp*B (1) | + | - |
| *Ostreococcus tauri* | 71,666 | 60.1 | 92 | 1 | 2 | - | 1071 | *atp*B (5188) | *psa*M/psbT (96) | *atp*B (1) | + | - |

**Table S3:** Features of introns within the cpGenome of *Trachelomonas grandis*. gII = possible group II introns or group II twintrons, mini gII = possible mini group II introns, gIII = possible group III introns. tgIII possible group III introns, * = modified AGC → AAC, [1] = modified AGC → AGT.

| Intron | Length | GC content (%) | ISS | Intron start | Putative mini group II, group II or group III intron |
|---|---|---|---|---|---|
| *atp*A intron 1 | 467 | 21.6 | 792 | ATATG | |
| *atp*E intron 1 | 393 | 19.8 | 22 | GTTTG | |
| *atp*E intron 2 | 364 | 20.6 | 655 | TATGG | |
| *atp*F intron 1 | 369 | 17.6 | 104 | ATTTA | gII |
| *atp*F intron 2 | 913 | 18.1 | 749 | GTTTA | |
| *atp*H intron 1 | 999 | 19.4 | 214 | GTGCG | gII |
| *atp*I intron 1 | 165 | 14.5 | 101 | CAATT | gIII |
| *atp*I intron 2 | 318 | 14.8 | 610 | ATATT | |
| *atp*I intron 3 | 102 | 14.7 | 1003 | GTGTG | gIII |
| *atp*I intron 4 | 349 | 20.3 | 1147 | TTATA | gII |
| *atp*I intron 5 | 100 | 21.0 | 1591 | ATGTG | gIII |
| *atp*I intron 6 | 118 | 18.6 | 1793 | TTGTG | gIII |
| *psa*A intron 1* | 961 | 20,2 | 464 | GTTTT | gII |
| *psa*A intron2 | 665 | 20,5 | 2207 | TTTAC | |
| psaB intron 1 | 1117 | 20,1 | 281 | GTTGG | gII |
| *psa*B intron 2 | 852 | 17,5 | 1494 | GAGGG | gII |
| *psa*C intron 1 | 913 | 22.1 | 42 | TTGTG | |
| *psa*C intron 2 | 435 | 18.6 | 986 | TTGTT | gII |
| *psb*A intron 1 | 951 | 36 | 25 | TTGCG | gII |
| *psb*A intron 2 | 3080 | 23,9 | 1348 | GTTTG | |
| *psb*A intron 3 | 943 | 28,3 | 4584 | GTGTG | gII |
| *psb*A intron 4 | 589 | 22,8 | 5597 | GTGTG | gII |
| *psb*A intron 5[1] | 592 | 20,4 | 6367 | GTTTG | gII |
| *psb*A intron 6 | 412 | 21,4 | 7049 | GTGTG | gII |
| *psb*A intron 7 | 522 | 23,8 | 7560 | GTGCA | gII |
| *psb*C intron 1 | 6460 | 21,5 | 100 | TCCGA | gII |
| *psb*C intron 2 | 1627 | 25,7 | 6656 | GTGTG | |
| *psb*C intron 3 | 1504 | 20,8 | 9450 | GAGCG | |
| *psb*D intron1 | 1794 | 26,9 | 12 | GTAAC | |
| *psb*E intron 1 | 465 | 23.9 | 78 | GTGTG | gII |

**Table S3:** Continued.

| Intron | Length | GC content (%) | ISS | Intron start | Putative mini group II, group II or group III intron |
|---|---|---|---|---|---|
| *psb*K intron 1 | 123 | 26 | 42 | GTGCG | gIII |
| *psb*K intron 2 | 211 | 18,5 | 205 | GTATG | tgIII |
| *psb*N intron 1 | 253 | 21.3 | 13 | TTTAG | tgIII |
| *psb*T intron1 | 551 | 20.1 | 25 | TTGCG | gII |
| *psb*30 intron 1 | 114 | 17,5 | 10 | TAGTG | gIII |
| *rps*2 intron1 | 105 | 11,4 | 192 | TTATT | gIII |
| *rps*2 intron2 | 212 | 12,7 | 449 | TTTTG | tgIII |
| *rps*2 intron 3 | 102 | 12.7 | 738 | ATGTG | gIII |
| *rps*3 intron 1 | 491 | 17.7 | 46 | GTTTG | |
| *rps*3 intron 2 | 98 | 12.2 | 1077 | TTTTG | gIII |
| *rps*4 intron 1 | 481 | 18,5 | 117 | GTGTG | gII |
| *rps*8 intron 1 | 387 | 18,9 | 310 | TTTCG | gII |
| *rps*9 intron 1 | 101 | 13,9 | 98 | TTGTG | gIII |
| *rps*9 intron 2 | 111 | 16,2 | 216 | TTTAG | gIII |
| *rps*9 intron 3 | 97 | 14,4 | 419 | TTGTG | gIII |
| *rps*11 intron1 | 111 | 17.1 | 69 | TTGTG | gIII |
| *rps*11 intron 2 | 98 | 16.3 | 197 | TTATG | gIII |
| *rps*14 intron 1 | 116 | 12.1 | 174 | TCCTT | gIII |
| *rps*18 intron 1 | 314 | 14.0 | 132 | TTATG | |
| *rps*19 intron1 | 103 | 20.4 | 89 | GTATG | gIII |
| *rps*19 intron 2 | 122 | 22.1 | 313 | TTGTG | gIII |
| *rpl*2 intron 1 | 338 | 20.4 | 142 | GTTAA | |
| *rpl*2 intron 2 | 432 | 18.8 | 779 | TTAGC | gII |
| *rpl*14 intron 1 | 104 | 15.4 | 131 | GTGAT | gIII |
| *rpl*14 intron 2 | 95 | 22.1 | 389 | ATGTG | gIII |
| *rpl*23 intron 1 | 112 | 17.0 | 13 | GTGTG | gIII |
| *rpl*23 intron2 | 120 | 12.5 | 205 | TTTTG | gIII |
| *rbc*L intron 1 | 532 | 21.8 | 516 | GTGTG | gII |

**Table S3:** Continued.

| Intron | Length | GC content (%) | ISS | Intron start | Putative mini group II, group II or group III intron |
|---|---|---|---|---|---|
| *rbc*L intron 1 | 532 | 21.8 | 516 | GTGTG | gII |
| *rbc*L intron 2 | 1032 | 26.6 | 1335 | GGGCG | gII |
| *rbc*L intron 3 | 895 | 29.3 | 2714 | GTGTG | gII |
| *roa*Aintron 1 | 271 | 20,7 | 132 | GTGCG | mini gII |
| *pet*B intron 1 | 410 | 22.9 | 23 | GTGCG | gII |
| *pet*B intron 2 | 448 | 16.7 | 701 | GTATG | |
| *tuf*A intron 1 | 113 | 21,2 | 425 | TTACG | gIII |
| *tuf*A intron 2 | 243 | 17,7 | 1033 | GTAAG | mini gII |
| *tuf*A intron 3 | 106 | 16 | 1511 | CTGCG | gIII |
| *tuf*A intron 4 | 516 | 18,8 | 1668 | GAGCG | gII |
| *ycf*4 intron 1 | 273 | 18,7 | 366 | GTAAG | mini gII |
| *16*S intron 1 | 874 | 33,1 | 499 | GTGCG | gII |
| *16*S intron 2 | 920 | 29,5 | 1783 | GTGCG | gII |
| *16*S intron 3 | 641 | 23,7 | 3115 | GTGTG | gII |
| *23*S intron1 | 3354 | 34,3 | 1779 | GTGCG | gII |

# 4 Discussion

## 4.1 The chloroplast genomes of phototrophic euglenoids

Since only little is known about the chloroplast genome evolution of phototrophic euglenoids, one aim of this study was to compare the cpGenomes of three exemplarily selected taxa with those of other published chloroplast genomes of euglenoids. It has been a key objective to examine differences and similarities to gain an overview and idea of chloroplast evolution of this unusual divers group of euglenoids.

Although nowadays the secondary endosymbiosis of a green alga by a phagotrophic euglenoid has found unanimous agreement (Gibbs 1978, 1981), it took further years until the first chloroplast genome (cpGenome) of a phototrophic euglenoid (*Euglena gracilis*) was fully sequenced and annotated (Hallick et al. 1993). The cpGenome has been identified as extremely large (more than 143 kb) due to the surprising fact, that 38 % of the chloroplast DNA was represented by a mixture of different introns. The second sequenced chloroplast genome was the heavily reduced one of *Euglena longa*, a colorless euglenoid, which had lost the chloroplast and all genes for photosynthesis (except for *rbc*L) along with its ability to photosynthesize in its genome (Gockel & Hachtel 2000) and will thus not be discussed any further in this study. From 2012 on eighteen further chloroplast genomes of phototrophic euglenoids have been sequenced and compared with one another on intergeneric or intrageneric levels and between different families (Bennett et al. 2012, 2014 & 2017, Bennett & Triemer 2015, Dabbagh et al. 2017, Dabbagh & Preisfeld 2017, Hrdá et al. 2012, Kasiborski et al. 2016, Pombert et al. 2012, Wiegert et al. 2012 & 2013). Results of the three cpGenomes analysed in this study will be discussed with regard to evolutionary trends on all cpGenomes and on phylogenetic evolution.

After the investigation of all three cpGenomes and comparing them with close and distantly related representatives of the group, it became evident, that some features are stable and can be found within all cpGenomes sequenced so far. Something all cpGenomes of phototrophic euglenoids have in common is a circular DNA molecule with remarkably similar AT content, that ranges from 64.9 % in *Eutreptiella pomquetensis* (Dabbagh et al. 2017) to 73.9 % in *Euglena gracilis* (Hallick et al. 1993) and *Colacium vesiculosum* (Wiegert et al. 2013). For five of the euglenoid cpGenomes the experimental circularization of the cpGenomes failed.

Remarkably, the loose ends were always located within or next to the rRNA operon, which themselves differed significantly in structure (Results, Chapter III, rRNA operon, p. 105). In addition, all investigated plastids, except for *Eutreptia viridis* (86 genes), have an almost identical gene composition of about 90 to 94 protein-coding genes, RNA-coding genes and genes for ribosomal proteins. The protein-coding genes encode the elongation factor EF-TU, photosynthetic proteins like *rbc*L for the $CO_2$-fixing enzyme ribulose-1,5-bisphosphate carboxylase/ oxygenase, genes for the chloroplast ATP synthase complex and genes for components of the photosystem I and II. The RNA-coding genes are divided into transfer RNAs and ribosomal RNAs. The ribosomal proteins encode for the 30S small and 50S large subunit of the chloroplast genome ribosomes. Considering the protein-coding genes and genes for ribosomal proteins in the plastid genomes of phototrophic euglenoids, then almost all of them show largely homologous coding regions. Nonetheless, there are intermittently missing protein-coding genes, genes for ribosomal proteins genes or tRNAs in individual taxa. Whether and why those genes got lost in the chloroplast genomes is not reported until today and still no results have been achieved that allow a suitable explanation (Hallick et al. 1993, Schwartzbach & Shigeoka 2017).

Although the composition and content of protein-coding genes (in the following always including genes for ribosomal proteins) and tRNAs are almost identical and equal in number, the cpGenomes of phototrophic euglenoids show remarkable disparities in size and significant rearrangements of genome components.

## 4.2 General characteristics of the plastid genomes across the lineage

The cpGenome size ranged from 65,523 bp in *Eutreptia viridis* to 144,166 bp in *Strombomonas acuminata*. This noteworthy heterogeneity in size can be explained in different ways. First, there are the rRNA genes encoded in operons in each cpGenome sequenced so far. The operon codes for the 16S, 23S and 5S genes are interrupted by intergenic spaces. The operon configuration is specific for most chloroplast rRNA genes in algae and higher plants and has a characteristic prokaryotic gene order (Bogorad & Vasil 1991), supporting the insight that the primary endocytobiosis event occurred just once, followed by several secondary endocytobiosis events (Archibald 2015, Keeling 2010) and that the phototrophic euglenoids are the result of the latter with a green alga. As recent studies have shown remarkable differences in number and organization on the cpGenomes among different species, it was of main interest to figure out, if the increasing number of investigated species will offer an evolutionary pathway on rRNA operon evolution.

For instance, comparing the cpGenomes of the genus *Euglena* shows for *E. gracilis* strain Z that the operon structure is tandemly repeated three times plus a fourth 16S rRNA gene (Hallick et al. 1993). This alone resulted in over 21,000 bp, which yield a 15 % share of the full cpGenome. In contrast, the cpGenome of *E. mutabilis* (SAG 1224-9b) contains only one rRNA operon with a length of 4,640 bp, which represents only 5 % of the total cpGenomes size (Dabbagh & Preisfeld 2017). It is still unknown how and why these differences occur.

Fourteen out of the 19 cpGenomes of phototrophic euglenoids sequenced, comprise only one rRNA copy, including the uncircularized cpGenomes of *Et. viridis*. Though for the latter the sequencing coverage was more than twice that of single copy protein-coding genes, implicating strongly that at least two copies should be present (Wiegert et al. 2012). The two basally branching *Eutreptiella* species contain two rRNA operons, whereby the cpGenome of *Etl. gymnastica* contains two incomplete copies with the 5S rRNA genes missing (Hrdá et al. 2012). Another peculiarity regarding the rRNA operons of *Etl. gymnastica* is that one of the rRNA operons is divided by protein-coding genes and thus not arranged like a typical operon (Dabbagh et al. 2017, Hrdá et al. 2012). The cpGenome of *S. acuminata* shows one complete and one incomplete copy, missing the second 16S rRNA gene (Wiegert et al. 2013). Only the circularized cpGenomes of both *E. gracilis* strains contain three copies of the rRNA operon arranged in tandem repeat units. Beside that a further 16S rRNA copy is found in the cpGenome of *E. gracilis* strain Z.

Concludingly, the three tandemly repeated rRNA copies of the two *Euglena* species are unique so far and when mapping rRNA operon features onto the phylogenomic tree of the phototrophic euglenoids one can assume an evolutionary pattern (Results, Chapter III, Fig. 3, p. 100). The most basally branching euglenoids of the genus *Eutreptiella* both contain two rRNA copies, which correspond most closely to the cpGenomes of green algae (Dabbagh et al. 2017, Hrdá et al. 2012). The rRNA operons of the psychrophilic *Etl. pomquetensis* exhibit the fewest evolutionary changes towards the rRNA operons of Pyramimonadales and consequently to the resulting genome structure. Nevertheless, the two copies are identified in all three Eutreptiales as the closest relatives to green algae chloroplasts. In all other euglenoids chloroplast genomes only one copy was identified, albeit with two exceptions. A probable scenario could be, that during the diversification of the *Euglenales* the number of the rRNA copies was reduced to one. Only in *S. acuminata* and both *E. gracilis* species, a second incomplete copy and two further copies of the rRNA operon were acquired independently, because the gene arrangement in these three taxa differs significantly. The *Euglena* strains show a tandem structure and *Strombomonas*, which is not yet experimentally circularized, has two operons coded next to each other on different strands. The operons are encoded adjacently tail to tail (Bleidorn 2017).

As to the function of different numbers and organization of the RNA operon one can only speculate. Theoretically, the more copies of genes are encoded in a genome, the faster expression can generally proceed (Elowitz et al. 2002), but it is still highly dependent on regulatory dynamics, transcription rates and many other genetic factors controlling expression as Eberhard et al. (2002) ascertained for the chlorophyte *Chlamydomonas reinhardtii*. For euglenoids cpGenomes the function is hard to investigate, because euglenoids do not reproduce by sex, only by cell division (Gillott & Triemer 1978, Leedale 1967), thus mutation and altered expression of genes cannot be monitored easily. One possible explanation to understand the heterogeneous organization of euglenoids cpGenomes could be the low control of mutation in these early eukaryotes that can also be found in ribosomal genes of the nucleus, especially in osmotrophic, but also in all other euglenoids (Busse & Preisfeld 2002b, Busse et al. 2003). Yet, the advantages of varying operon structures and copy numbers in euglenoids are unknown.

A second reason for the size variance between the euglenoids cpGenomes is the heterogeneous intergenic space (IGS). For instance, the IGS between the two closely related *Eutreptiella* species is strikingly different. The IGS of *Etl. pomquetensis* occupied more than 23 kb, which was more than twice that of *Etl. gymnastica*, showing only an intergenic space of approximately 11 kb (Dabbagh et al. 2017, Hrdá et al. 2012). It is conspicuous, that the largest cpGenomes comprised the highest intergenic space, with *S. acuminata* encompassing an intergenic space of more than 25 kb, with *E. gracilis* strain Z for more than 24 kb and with *Etl. pomquetensis* 23 kb (Dabbagh et al. 2017, Hallick et al. 1993, Wiegert et al. 2013). By contrast, the two smallest cpGenomes *Et. viridis* (cpGenomes of 65,523 bp) and *P. orbicularis* (cpGenomes of 65,992 bp) also have the smallest intergenic space with approximately 7,9 kb (Kasiborski et al. 2016, Wiegert et al. 2012). In total, the average intergenic space of euglenoid cpGenomes amounts to 13,9 kb. But still the last two reasons are not alone and not the only once and not foremost accountable for the size differences among the cpGenomes.

The third and main reason for the size differences is a very unequal number of introns. Especially these introns were of remarkable interest in this study. Correspondingly questions regarding intron evolution occupy large parts of the analysis of changes on the intrageneric level as well as concerning all investigated phototrophic euglenoids. It was a matter of concern to investigate, if trends were detectable within single chloroplast genomes, between different species or even within the whole phototrophic lineage. The highest number of introns was located in the two sequenced *E. gracilis* strains. *Euglena gracilis var. bacillaris* contains 134 introns, while *E. gracilis* strain Z encompasses a total of over 150 introns (Bennett & Triemer 2015, Hrdá et al. 2012). These numbers result in over 60 % of protein-coding genes of the cpGenomes, which contain at least 1 intron. In contrast *E. mutabilis* contains ‘only’ 76 introns within protein-coding genes (Dabbagh & Preisfeld 2017). Here again, it can only be surmises that a low genetic control allows so many introns to invade the genome. It is also known that introns are of utmost importance in stimulating gene expression or regulation, called intron-mediated enhancement (IME) (Mascarenhas et al. 1990). Some of the possible functions of introns may be related to facilitating protein evolution for example by exon-shuffling (Rose et al. 2008). Additionally, it was shown that the signal for splicing can enhance the transcription by triggering the RNA polymerase (Le Hir et al. 2003). Though the advantage of a facilitated gene expression is a high price to be paid for inserting and removing the introns and to suppress mistakes. But anyway, in case of the euglenoid chloroplast, that was taken over from a green alga, it might have been necessary to support the regular genetic apparatus, for which communication with a chloroplast was a new development, by regulatory help from introns. Afterwards these introns stayed in the cpGenome, although gene transfer between the adopted plastid and the new nuclear genome had already been accomplished. The question as to whether such regulatory functions also apply to the euglenoids as a basal eukaryotic group remains to be answered.

Summarizing the reasons for noticeable size differences in euglenoid cpGenomes, it becomes clear, that the introns are the main factor, followed by organization of RNA operons and intergenic spaces.

## 4.3 Origin of introns

The discovery of introns in protein-coding genes started in 1977 (Berget et al. 1977, Chow et al. 1977). From then on, introns were detected in many species of eukaryotes. The former, but nowadays corrected (Martínez-Abarca & Toro 2000) conclusion that introns were absent in prokaryotes raised two main hypotheses about intron evolution and hence the ‘intron-early’ and ‘intron-late’ hypotheses were established and discussed (Cavalier-Smith 1985, Doolittle 1978 & 1987, Gilbert 1978, Koonin 2006, Orgel & Crick 1980). The detection of more than 150 introns in the chloroplast of *E. gracilis* was considered as important new data for this debate (Hallick et al. 1993, Hong & Hallick 1994). The following investigations led to the assumption that *E. gracilis* acquired their introns late in the evolution of euglenoids (Hallick et al. 1993, Thompson et al. 1995). Reinforced was this hypothesis by the fact, that the prasinophyte *P. parkeae* as the closest living relative of the euglenoid chloroplast donor contains only one intron in its cpGenome (Turmel et al. 2009). Additionally, this ‘intron-late’ hypothesis for phototrophic euglenoids was affirmed by the announcement of the cpGenome sequences of the two early diverging euglenoids *Et viridis* and *Etl. gymnastica*, both containing only a small number of introns (Hrdá et al. 2012, Pombert et al. 2012, Wiegert et al. 2012). In addition, the results obtained on the cpGenome of *M. aenigmatica* strengthened the intron-late hypothesis with an intermediate amount of introns (53 intron insertion sites) and a phylogenetic position between Eutreptiales and the genus *Euglena* by Pombert et al. (2012). Furthermore, they detected that the species *M. aenigmatica* and *E. gracilis* shared more insertion sites than those two with *Et viridis* (Pombert et al. 2012). Afterwards, with the increasing number of cpGenome sequences of phototrophic euglenoids, the intron-late hypothesis has been changed little by little, because cpGenomes with varying number of introns appeared within same clades. In addition, some ‘basal’ species showed a higher intron number than ‘crown’ species. *E. mutabilis* and *E. viridis*, both derived species of the crown clade *Euglena,* possess 76 and 71 introns in protein-coding genes, respectively (Bennett et al. 2012, Dabbagh & Preisfeld 2017). In contrast, *C. vesiculosum* an intermediate branching phototrophic euglenoid contained more than 120 introns. Completely new findings have been achieved with the recently published cpGenomes of *Etl. pomquetensis*, which was unexpectedly permeated by 51 introns (Dabbagh et al. 2017).

Even if it is still valid, that the chloroplast ancestor was poor in introns, it unalterably remains unclear how the heterogeneous spread of introns across the tree can be explained. What is undisputed is that an overwhelming majority of introns were acquired later, after the secondary endosymbiosis, since the closest chloroplast donor *P. parkeae* only contained a single intron. Strengthening this thought, is the discovery, that some euglenoid intron types are not known to exist outside of euglenoids (Christopher & Hallick 1989, Copertino et al. 1991, Doetsch et al. 1998, Michel et al. 1989). At least, common intron types have been essentially changed during euglenoid evolution.

## 4.4 Intron types in the cpGenome of phototrophic euglenoids

Within the phototrophic euglenoids there are two types of introns to be found. One is the well-known group II intron (subgroup IIB), a common intron class also present in prokaryotes, fungal and plant mitochondria and plant chloroplast genomes (Lambowitz & Zimmerly 2011, Michel et al. 1989). Group II introns fold into a secondary structure consisting of six domains (DI – DVI) (Fig. 4.1), which interact among each other and radiate from a central wheel (Michel et al. 1989).

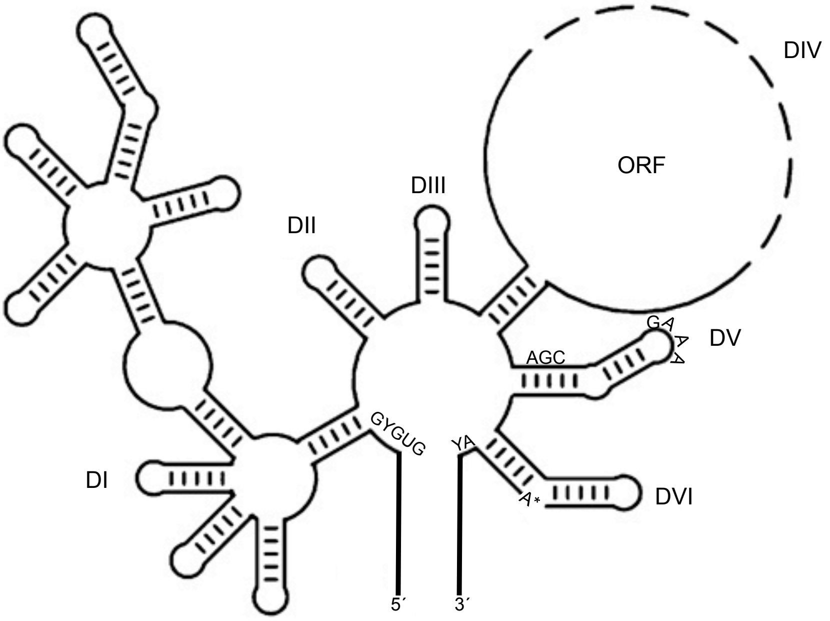


**Fig. 4.1:** Secondary structure of group II introns consisting of six domains (DI – DVI) arranged around a central wheel. The scheme shows consensus nucleotides that are common for group II introns. Group II introns usually have conserved 5′- and - 3′ splicing sites (5′ - GUGYG and AY - 3′). DIV encodes an ORF and DV is highly conserved in sequence, with an AGC – triad and a conserved loop (GAAA). A conserved branch-point adenosine (A*) in DVI plays a central role in the splicing process (modified from Michel & Ferat 1995).

Common group II introns have conserved 5′- and 3′- boundaries (5′- GUGYG and AY - 3′), an ORF, which encodes the intron-encoded protein (IEP), a multifunctional reverse transcriptase (RT) promoting intron mobility encoded in DIV and a bulged A 7-8 nt upstream from the 3′ intron- exon junction within DVI. Additionally known is the high consistency of DV with a conserved stem AGC - triad and a conserved GAAA terminal loop (Lambowitz & Belfort 2015, Lambowitz & Zimmerly 2011, Michel et al. 1989). Many introns of phototrophic euglenoids can only be described as group II-like. They lack conserved group II core structures, most structural elements became unrecognizable and show massive divergence to related group II introns. Thus, euglenoid group II introns have a tendency toward being significantly shorter than other group II introns. Comparisons demonstrated that group II introns in chloroplasts of euglenoids range from 298-618 nt, with a mean of 463 +/ - 90, whereas for example group II introns of the common liverwort *Marchantia polymorpha* (Archaeplastida) showed a mean of 577 +/ - 119 (Michel et al. 1989).

The other introns located in cpGenomes of phototrophic euglenoids are comparably short introns, which range in size from 95 - 110 nt and are designated group III introns. They are considered as degenerated group II introns only containing DI and DIV Mostly, they only hold two conserved bases in the 5′-boundary NTNNG and a bulged A within DVI, otherwise they do not show any other conserved secondary structural features (Christopher & Hallick 1989).

Various analyses detected that euglenoid chloroplast introns can be found solitarily or as twintrons. These forms of introns, first described in *Euglena* (Copertino & Hallick 1991), are introns located within introns. They occur in different variations as group II twintrons, mixed group II/ group III twintrons nested within each other in both possible ways and group III twintrons. Even complex twintrons as a consequence of multiple intron insertion events in one intron can be observed in euglenoid cpGenomes reflecting their activity as mobile elements (Copertino et al. 1992 & 1994, Copertino & Hallick 1993, Doetsch et al. 1998, Drager & Hallick 1993, Zimmerly & Semper 2015). The high intron number and different intron types in euglenoids point to a high vulnerability for introns due to low genetic control (Busse & Preisfeld 2003). An insertion of one intron inside another intron might facilitate seizing in and splicing due to an imaginable combined use of the splicing apparatus. It therefore might be advantageous for the intron to minimize the effects on the genes they have invaded (Hafez & Hausner 2015).

The plastid genomes of phototrophic euglenoids are highly diverse regarding intron number and intron type in the same gene. The quantity of introns in regard to genome size ranges from 10 % in *Etl. gymnastica* to 51 % in *S. acuminata* (Hrdá et al. 2012, Wiegert et al. 2013). As a result, only a few comprehensive trends can be presented, which but yet explain convincingly the intron density and quantity as well as high or low similarities in the evolution of introns in all phototrophic euglenoids. Although several strands of tendencies are visible, they mostly apply only to a few or even single representatives. The expansion of *psb*C introns in the chloroplast of euglenoids, which was detectable during investigations of *E. mutabilis*, is one example of intron evolution that occurred after the split from Eutreptiales and Euglenales. Then a new intron containing *mat*2 was inserted in the *psb*C gene and is detectable in almost all Euglenaceae at the same insertion site (Results, Chapter I, Fig. 2, p. 9). Additionally *pet*B intron 1 that was detected during the analyses of the chloroplast of *T. grandis* as an intron that presumably spread vertically by diversification of species, because all intron 1 of petB genes in euglenoids possess the same insertion site. Nevertheless, the size varied extremely among different species due to interruption of petB coding regions by other introns or twintrons (Results, Chapter III, Table S2, p. 119). Whereas, for example, petB of *E. mutabilis* contains a group II intron (435 bp) with a detectable domain V, the petB intron of *E. gracilis* strain Z is a complex twintron. The external intron (399 bp) is the same group II intron detected in the petB gene of *E. mutabilis*. It would be very interesting to infer a phylogenetic tree of introns to get to the source of the introns as can be done for example with group I introns (Busse & Preisfeld 2003). Unfortunately, group II and III introns in euglenoids are too heterogeneous and/ or too small to provide sufficient phylogenetic signal.

The present study yielded also evolutionary trends of introns that can only be observed inside of one genome or between single taxa, but not within all sequenced cpGenomes of phototrophic euglenoids so far. These findings are very helpful nevertheless, because they allow us to explain or at least to develop a scenario about how twintrons can come into being and how the number of introns can rise.

The introns of the cpGenomes of *Etl. pomquetensis* do not merely show the highest number of introns detected in the basally branching Eutreptiales. These introns also have the highest of sequence similarities ever shown. More than half of the introns in the cpGenome of *Etl. pomquetensis* showed pairwise identities of 87.4 %. This high number of similar introns has not been detected previously in euglenoids, nor in green algae and it might be speculated as a process of intron distribution in progress with little time for changes in the intron sequence (Dabbagh et al. 2017). The unique intron similarities in *Etl. pomquetensis* underpin the tolerated genetic variance and presumed low control in euglenoids two factors, which might well be able to facilitate the invasion of different genes by introns.

Additionally, group II twintron evolution is visible in *Etl. pomquetensis* in individual genes (Dabbagh et al. 2017). It is likely, that all 28 similar introns identified with a further GTGCG at nt 261-165 are potential twintrons, composed of an external and internal group II intron. Moreover, for two (*psa*C I2 and *pet*G I1) of these potential twintrons with high sequence similarity ongoing twintron evolution was detectable, which resulted in complex twintrons. While the *pet*G I1 twintron was again invaded by the same twintron with high sequence similarity the internal intron of *psa*C I2 was also one of these introns that were characterized by high sequence similarity, whereas the external intron of *psa*C I2 is closely related to and probably arose from the same ancestral intron as *psb*C I2 in *Etl. gymnastica* (Results, Chapter II, Fig. 4A - C, p. 15).

## 4.5 Degeneration of group II introns led to group III introns

It was of main concern to substantiate a possible relationship between the different intron groups. Since Copertino et al. (1991) ascertained that group III introns can be regarded as degenerated group II introns, this process would imply an evolutionary stage in which somehow altered group II introns are detectable as intermediate stages. The analyses of this study showed that no group III introns could be identified in the cpGenomes of basally branching Eutreptiales, which are strictly bound to marine habitats. As mentioned previously the introns detected in *Etl. pomquetensis* are exceptional. The chloroplast genome has not only held a high number of identical introns in this basal species. Moreover, investigations also showed that none of the 51 introns of *Etl. pomquetensis* complied with typical group III introns. The most striking difference thereby was the size. The smallest intron was 356 bp long, which exceeded the typical size of group III introns and even group III twintrons (Dabbagh et al. 2017).

It seems relevant to perform further extensive analysis of the other two published cpGenomes of the basal Eutreptiales. During the investigation it became apparent that also the smallest introns of *Et. viridis* (156 bp) and *Etl. gymnastica* (179 bp) were larger than typical group III introns (Dabbagh et al. 2017). The size of these introns could be related to a very small group II introns and namely the secondary structure of domain V of group II introns was recognizable in *rpo*B I1 of *Etl. gymnastica* and *Et. viridis.* Both confirmed a likeliness to group II introns, because in group III introns, the catalytic domain V is always absent, and only domains DI and DVI (Christopher & Hallick 1989) remained during the degeneration process. Hence, the introns described in this analysis are larger than group III introns, contain DV and do not show other similarities to group III introns (Dabbagh et al. 2017). One possibility is that these group II introns in Eutreptiales underwent drastic degeneration and performed as an intermediate stage, best resembled mini-group II introns. Mini-group II introns are characterized by the absence of different domains and have been detected previously in chloroplast genes of *Lepocinclis buetschlii* (Doetsch et al. 1998).

With the investigations of this study a pattern is forming that group II introns appeared first in an intron-less ancestral genome. The evolution of group III introns is thought to have evolved in *Etl. pomquetensis* by degeneration of group II introns that led to intermediate stages as mini group II introns, still detectable in the cpGenome. These then invaded the cpGenomes of the Euglenales by vertical evolution and partially degenerated further to group III introns.

Another observation in support of the mentioned hypothesis above that group III introns are absent in the Eutreptiales is that all three Eutreptiales have a *psb*C intron including *mat*1 (*ycf*13) that is at least three times larger than the group III twintron (I4) including *mat*1 of *E. gracilis*. These features, and the fact that *E. gracilis* contained a group - II - type maturase in a group III twintron (Doetsch et al. 1998, Mohr et al. 1993), underpin the possibility that group II introns evolved prior to group III introns in basally branching euglenoid species. Subsequently, they degenerated by loss of different domains (in more derived species) to group III introns, containing only DI-like and DVI-like structures (Doetsch et al. 1998, Lambowitz & Belfort 2015).

Since today it is not clear whether group III introns of phototrophic euglenoids are self-splicing or have lost this ability. Group III introns have no catalytic domain V and no intron encoded reverse transcriptases (maturases), which functions as intron-specific splicing factor (Zimmerly & Semper 2015). In many group II introns of phototrophic euglenoids the intron encoded reverse transcriptase maturase is absent and thus the question arises by which splicing mechanisms these introns are removed from the preRNA. Because of the lacking ORF with maturase activity we assume that these introns are probably completely dependent on trans-acting proteins and/ or RNA splicing factors and that intron specific maturases located somewhere in the chloroplast genome are able to splice multiple introns. A scenario could be, that *Etl. pomquetensis* group II intron *mat*1 acts *in trans* for all ORF-less introns (Dabbagh et al. 2017). A hypothesis on trans-acting generally was not new (Doetsch et al. 1998) and was supported by the findings of Pombert et al. that the psbC intron containing *mat*1 (synonym *ycf*13) is the ancestral euglenoid intron, which was detected in all species until then. This proposal was further corroborated by all three investigated cpGenomes in this study. It seems plausible that the *mat*1 discovered in the cpGenomes of phototrophic euglenoids, acts *in trans* for all ORF-less introns. But it is also conceivable that another intron encoded protein (IEP) acts *in trans* to promote splicing and mobility of some of the ORF-less introns (Dabbagh et al. 2017). Although there is no clear evidence, whether *mat*1 is the maturase that acts *in trans* to splice multiple introns in a chloroplast genome, this maturase is a highly likely candidate. Another reason for assuming mat1 to be the maturase acting *in trans* is, that the high number of introns in euglenoids otherwise would make it necessary that each intron is spliced by its own splicing factors (Copertino & Hallick 1991, Copertino et al. 1994, Dabbagh et al. 2017, Schwartzbach & Shigeoka 2017), which seems to be highly unparsimonious.

## 4.6 Chloroplast genome comparison to the closely related green algae

Comparing the genome compositions of phototrophic euglenoids and assorted green algae leads to the conclusion, that only one single secondary endosymbiotic event took place within the diverse lineage of phototrophic euglenoids. An analysis of the cpGenomes of basal Eutreptiales and *Pyramimonas parkeae* supports the evidence, that the ancestor of the euglenoid chloroplasts was a member of the Pyramimonadales. Comparative investigations on the chloroplast genomes of *Etl. pomquetensis* and the closest living relative *P. parkeae* clarified important questions regarding the evolution of the chloroplasts and which genomic changes occurred. Moreover, the results of *Etl. pomquetensis* underlined the hypothesis that the event of secondary endosymbiosis occurred in a marine environment (Gibbs 1978 & 1981, Dabbagh et al. 2017) or even cold marine waters, since *Etl. pomquetensis* only subsists in cold marine environment, *P. parkeae* admittedly does not, but in turn other species of the genus are also psychrophilic. Phylogenetic analyses and genome structure, however, indicate that *Etl. pomquetensis* and *P. parkeae* are the closest living representatives identified on both sides of the secondary endosymbiosis event.

With a size of 101,605 bp, the cpGenome of *P. parkeae* is smaller than almost half of all euglenoid cpGenomes, nevertheless at the same time encodes 110 conserved genes in contrast to approximately 90 to 93 genes in euglenoids. The cpGenomes of phototrophic euglenoids do not contain the common land plant chloroplast genes such as *rpl*33, *inf*A, *clp*P, *frx*B, *ndh*A-K, *pet*A, *pet*D, *psb*M, *rps*15, *rps*16, *psb*O and *rbc*S (Dabbagh et al. 2017, Wiegert et al. 2012).

It is known that most of the genes like *pet*A, *pet*D as well as *psb*O and *rbc*S migrated from the chloroplast to the nucleus by horizontal gene transfer, but still encode chloroplast proteins (Chan et al. 1990, Santillán Torres et al. 2003, Vesteg et al. 2009).

The intergenic space of about 18 kb (Turmel et al. 2009) in the cpGenome of *P. parkeae* resembled the IGS of *T. grandis* in size. Furthermore, there are some additional similarities the cpGenome of *P. parkeae* has in common with basal phototrophic euglenoids. First of all, with a GC content of 65.3 % it resembled the cpGenome of the two *Eutreptiella* species. Besides, *psb*T (96 bp) is the smallest gene in the genomes of *P. parkeae*, *Et. viridis* and *Etl. gymnastica*. In the genome of *P. parkeae* there are two cases of overlapping genes (*psb*C - p*sb*D and *ndh*C - *ndh*K); the first one is also detectable in euglenoids cpGenome. There is a clear distinction between the intron amounts within the cpGenome of *P. parkea*, which features only one group II intron in the *atp*B gene and the intron-perforated cpGenomes of euglenoids.

Notably conserved in the chloroplast genomes is the operon content of polycistronic transcription units between *P. parkeae* and euglenoids. Transcription units detected in *E. gracilis,* like for example, the protein operon that encodes the genes *rpo*B - *rpo*C1 - *rpo*C2 or the ribosomal protein operon that encodes the genes from *rpl*23 to *rps*14 (Copertino et al. 1992), are related to those of *P. parkeae*. While the operon organization is conserved, the chloroplast genomes of the majority of phototrophic euglenoids most remarkably do not have the same genome structure. Most of the euglenoid cpGenomes are not divided into the quadripartite structure generally found in green algae chloroplasts, which consists of a large and a small copy region (LSC and SSC), separated by inverted repeats (IR) comprising the rRNA genes. Within the phototrophic euglenoids the plant-like genome composition of a large single copy region, a small single copy region and two inverted repeats (IR) is unique to *Etl. pomquetensis* and was never detected before (Results, Chapter II, Fig. 1, p. 6). Likewise, the orientation of the rRNA operons - one operon on the positive and one on the negative strand - points out the high similarity between *Etl. pomquetensis* and green algae cpGenomes such as *P. parkeae* (Dabbagh et al. 2017, Hrdá et al. 2012, Turmel et al. 2009). It is obvious that also the coding regions of the basally branching *Eutreptiales* were more similar to those of *P. parkeae* than to those of the derived *Euglena* clade (Dabbagh et al. 2017). These data of high similarities between *P. parkeae* and *Etl. pomquetensis* implicate that the latter did not diverge much time after the secondary endocytobiosis and therewith the engulfment of the green algae and subsequent integration of its chloroplast as a euglenoids organelle took place.

## 4.7 Phylogenomic analyses and genomic metacharacters

During the last years the increasing number of sequences of phototrophic euglenoids also prompted phylogenetic studies combining different molecular markers. However, most of the understanding about evolution of phototrophic euglenoids relates primarily to nuclear and plastid LSU and SSU datasets (Kim & Shin 2008, Kim et al. 2010 & 2015, Linton et al. 2010, Marin et al. 2003). The new genomic information on phototrophic euglenoids, which has become accessible by next generation sequencing, has been used in this study to identify full gene lengths present in the chloroplast genomes and to expand the available phylogenetic information. The present phylogenomic analysis is the second phylogenomic analysis of phototrophic euglenoids and the biggest dataset available for the moment. On the one hand, it results in relationships that were consistent with previous analyses (Karnkowska et al. 2015, Kim et al. 2015, Linton et al. 2010). On the other hand, this analysis leads to positions of species that either have not been included in previous studies or not paid great attention to. To clarify these positions, the genomic analyses in phototrophic euglenoids offered a new possibility to use not only sequence-based genomic data. Rather the chloroplast genomes enabled us to use genomic metacharacters. These metacharacters also referred to as molecular morphology or genome-level features, are often used in mitochondrial genome analyses (Boore & Brown 1998, Donath & Stadler 2014, Perseke et al. 2008) or to clarify phylogeny of insects (Niehuis et al. 2012). Beyond sequence-based characters, molecular morphology can be used as indicators of divergence and allow to integrate genome-level features such as intron gain and loss, gene arrangement changes or novel genes and therewith offers enormous potential for molecular systematics (Donath & Stadler 2014). Rokas & Holland (2000) defined rare genomic changes (RGC) as almost perfect phylogenetic characters to complement phylogenetic analyses. The problems that still exist are missing robust statistical methods or standards for each molecular morphology character that describes the evolutionary dynamics. These characters are expected to have different functional impact and moreover they are expected to change in a saltatory non-clockwise way. Nevertheless these genome-level features have a huge potential to solve controversial relationships (Boore 2006, Donath & Stadler 2014), but still the question remains of how to analyze all markers and how to weight their support.

The investigations of the three selected phototrophic euglenoids and their comparison with other cpGenomes resulted in the identification of metacharacters that can potentially be used to complement phylogenetic analyses. The evolutionary pathway of these molecular morphology characters can be reasonably traced like morphological data and mapped onto the phylogenetic tree to draw a convincing line of genomic evolution (Bleidorn 2017).

One molecular morphology character that was detected in phototrophic euglenoids and allowed to follow an evolutionary pathway was the acquisition of new introns and together with that *mat*2 after the split between Eutreptiales and Euglenales (Dabbagh & Preisfeld 2017). In eukaryotes introns have been used in various studies as phylogenetic marker to analyze relationships. In Metazoa for example they have been used to infer fish phylogeny or as metacharacter in insect orders (Niehuis et al. 2012, Rokas & Holland 2000). Since phototrophic euglenoids show high and diverse intron numbers, it seems doubtful whether all of the introns can perform as genome-level features. But as shared intron positions indicate homology, it is obvious that at least these introns allow following the phylogenetic process in the chloroplast genomes. So far this trend was only shown for some introns in hitherto limited chloroplast genome analyses.

Other possible genome-level characters that have been detected in this study and that obviously follow a trend during the evolution and thus can likely help to clarify phylogenetic relationships within euglenoids is the number of rRNA operon repeats, gene order and cluster arrangement in chloroplast genomes of phototrophic euglenoids (Dabbagh & Preisfeld 2017, Dabbagh et al. 2017). They do offer the potential to be used as a phylogenetic tool in combination with sequence data.

The analyses of *T. grandis* and the intrageneric comparison with *T. volvocina* revealed intergeneric changes in the synteny of genome structure between different euglenoid genera. The genome structure rearrangement that occurred between different genera can be observed as a trend in regard to the phylogeny of euglenoids. While the two *Eutreptiella* species comprise low synteny (10 clusters) during progressive Mauve analysis, it becomes obvious that the more the euglenoid taxa derived the more consistent the genome arrangement became, when the results of the cluster arrangement are mapped onto the phylogenomic tree. This resulted in only one cluster in clade B and two clusters in clade A (Results, Chapter III, Fig. 3, p. 100). In the near future, it would be of main interest to analyze further chloroplast genomes of the genus *Colacium* and use the resulting cluster arrangement as genome-feature and therewith determine its most likely relation to other phototrophic euglenoids.

## 4.8 Gene order as a genomic metacharacter

The use of gene order changes as metacharacter is nothing new and has already proved useful in phylogenetics. Gene order changes were first used to investigate the evolution of drosophilids and were often consulted as phylogenetic markers in mitochondrial genomes, but were also applied in case of plant chloroplasts phylogeny (Bleidorn 2017, Boore 2006). Since mitochondrial genomes are very small, they are well-developed in regard of genome-level characters and their investigations have pioneered the use of metacharacters for phylogenetic inference (Boore & Brown 1998, Boore 2006). Mitochondrial genome evolution can be analyzed by CREx, a web-based tool for comparisons of gene order data that heuristically explores rearrangement between mitochondrial genomes (Bernt et al. 2007). To facilitate comparisons between chloroplast genomes of euglenoids in this study the genomes were arranged in the manner that the rRNA genes were designated to be on the reverse strand as established by Hallick et al. (1993). Although the chloroplasts in the genus *Euglena* show high synteny of cluster arrangement, the chloroplast genome of *E. mutabilis* shows a specific peculiarity regarding the genus. *E. mutabilis* is characterized by an almost complete mirror-inverted gene and cluster arrangement in contrast to the other investigated *Euglena* species (Dabbagh & Preisfeld 2017). It is interesting that large parts of the gene order (clusters 4 - 12, Results, Chapter I, Fig. 4, p. 11) resemble species within the sister clade of the genera *Monomorphina*, *Cryptoglena*, *Euglenaria* and *Strombomonas* (Dabbagh & Preisfeld 2017). Since it is known that molecular morphology features are expected to change in a saltatory non-clockwise way, the chloroplast genome of *E. mutabilis* could be regarded as transition from the cpGenomes of the sister group to other members of the genus *Euglena.* Although the results of the phylogenomic analyses do not show that *E. mutabilis* branched at the base of the *Euglena* clade, the hypothesis still seems plausible due to the saltatory mobility.

The use of genome-level features of cpGenomes of phototrophic euglenoids in comparison with phylogenomic data is totally new and it enables us to rule out questionable positions of taxa. For prospective investigations it seems indispensable to define ideal metacharacters as phylogenetic markers in the chloroplast genomes of phototrophic euglenoids and to determine how to analyze this marker and how to weight their support.

# 5 Conclusion

The three carefully selected taxa of phototrophic euglenoids in this study have been used to compare their chloroplast genomes with further chloroplast genomes of euglenoids to get an overview of chloroplast evolution in the highly diverse lineage within the Euglenozoa. Although the general gene composition was almost identical in all investigated cpGenomes, the chloroplast genomes show remarkable differences in size. The varying number of RNA repeats, IGS differences and as main factor intron and twintron content, have been identified as the three most pressing causes for size differences. The conducted intrageneric and intergeneric comparisons yielded large cluster rearrangements, which occurred between different clades and resulted in a high synteny of derived taxa due to merging clusters. Despite the approach to detect lineage encompassing trends and consistencies within the phototrophic euglenoids regarding the evolution of euglenoids and their chloroplasts, which could only be found in between species and only as an exception between genera, here molecular morphology trends have been detected in the chloroplast genomes of euglenoids for the first time. These metacharacters appear suitable to support phylogenomic and phylogenetic analyses and can be used to understand and possibly rule out questionable positions. Inter alia cluster arrangement, gene order and individual introns have been determined as significant metacharacters of the euglenoid chloroplast genome. Since plastid genomes from some euglenoid families are still not avaliable, an increasing sampling of euglenoid taxa across the tree would allow to explore and maybe confirm the described metacharacters in other taxa as well. Thereby it would be important to define more potential and appropriate molecular morphology features and to establish a standardized system for analyzing these genome-level features.

As former studies have shown that the secondary endosymbiosis within the phototrophic euglenoids occurred once as a single event and that *Pyramimonas parkeae* is the closest living relative up to date (Gibbs 1978 & 1987, Turmel et al. 2009, Wiegert et al. 2012), the investigations on the chloroplast genome of the psychrophilic *Eutreptiella pomquetensis* lead to the result, that it is the only euglenoid showing a typical quadripartite genome structure with two complete inverted repeats (Dabbagh et al. 2017), as *P. parkeae* does. From these data it can be concluded that *Etl. pomquetensis* is the nearest living relative to the Pyramimonadales up to date. To complement the understanding of secondary endosymbiosis it would be of substantial interest to investigate the chloroplast genome of a psychrophilic marine green alga like the antarctic *Pyramimonas gelidicola* sp. nov. to confirm the hypothesis, that the engulfment of a green alga originated in a marine environment by an *Etl. pomquentensis*-like ancestor.

The present work also brought substantial new insights into the intron evolution of euglenoid chloroplasts. On the basis of these results, the hypothesis by Copertino et al. (1991) that group III introns are degenerated group II introns was supported and, moreover, it was ascertained that the evolution of group III introns began in Eutreptiales as intermediate stages of group II and group III introns (mini group II introns), but actually spread even more degenerated as group III introns in Euglenales. So the absence of group III introns in Eutreptiales is not merely a coincidence, but the result of intron evolution by further degeneration after the split from Eutreptiales and Euglenales (Dabbagh et al. 2017). Thus a lot of work still needs to be done to reconstruct and understand the intron evolution and their remarkable characteristics in euglenoids as a whole. Although for instance twintron evolution was made visible in individual species or between different species, little is known about the reason for their existence and one can only speculate that they enhance other biological pathways or simply do not perform any task. Furthermore, albeit experiments have shown that maturases are intron-specific splicing factors (Lambowitz et al. 1999, Sheveleva & Hallick 2004), it is still unclear whether the majority of ORF-less introns in cpGenomes of euglenoids are self-splicing or completely dependent on *trans*-acting maturases. Indication that individual group II intron maturases assume such a generalized role by splicing multiple introns was experimentally demonstrated in bacteria and higher plants. Within land plants the *mat*K protein acquired the ability to bind and splice multiple ORF less group II intron (Lambowitz & Zimmerly 2011, Meng et al. 2005, Vogel et al. 1999). Comparably, one can assume that group II introns in euglenoids are spliced by a single maturase encoded by one of the introns (Copertino et al. 1994). The maturase *mat*1 was detected in each cpGenome of euglenoids investigated hitherto and could comply with such a generalized role in splicing multiple introns, but hard evidence is still missing and the topic needs to be further examined.

Likewise, a pending issue and thus also in need to be examined are the detected introns in the rRNA operon of *Trachelomonas grandis*. It can be assumed, that these four group II introns are the result of intron transfer, since they are not related to other introns found in the chloroplast rRNA operons of euglenoids, nor to chlorophytes leading to the euglenoid chloroplasts. A detection of closely related introns would enhance our understanding in intron evolution Moreover, it would be of special interest to investigate the large group II intron encoded maturase in the 23S intron of *T. grandis* and simultaneously examine whether it supports the splicing of the other three ORF-less group II introns located in the 16S rRNA gene.

Subsuming up to date, the cpGenomes of all sequenced euglenoids are a first step to understand intron and chloroplast evolution. To comprehend the mingled intron evolution of euglenoids, still more cpGenomes and detailed intron studies are needed.

# 6 References

Abrusan G., Grundmann N., DeMester L, Makalowski W. 2009. TEclass--a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* 10: 1329–1330. doi: 10.1093/bioinformatics/btp084.

Adl, S. M., Simpson, A. G. B., Farmer, M. A., Andersen, R. A., Anderson, O. R., Barta, J. R., Bowser, S. S., Brugerolle, G., Fensome, R. A., Fredericq, S., James, T. Y., Karpov, S., Kugrens, P., Krug, J., Lane, C. E., Lewis, L. A., Lodge, J., Lynn, D. H., Mann, D. G., McCourt, R. M., Mendoza, L., Moestrup, O., Mozley-Standridge, S. E., Nerad, T. A., Shearer, C. A., Smirnov, A. V., Spiegel, F. W. & Taylor, M. F. J. R. 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *Journal of Eukaryotic Microbiology*, 52:399–451. doi: 10.1111/j.1550-7408.2005.00053.x.

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., Le Gall, L., Lynn, D. H., McManus, H., Mitchell, E. A. D., Mozley-Stanridge, S. E., Parfrey, L. W., Pawlowski, J., Rueckert, S., Shadwick, L., Shadwick, L., Schoch, C. L., Smirnov, A. & Spiegel, F. W. 2012. The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology*, 59:429–493. doi: 10.1111/j.1550-7408.2012.00644.x.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410. doi: 10.1016/S0022-2836(05)80360-2.

Archibald, J. M. 2007. Nucleomorph genomes: Structure, function, origin and evolution. *Bioessays*, 29:392–402. doi: 10.1002/bies.20551.

Archibald, J. M. 2009. The puzzle of plastid evolution. *Current Biology*, 19:R81-8. doi: 10.1016/j.cub.2008.11.067.

Archibald, J. M. 2015. Endosymbiosis and Eukaryotic Cell Evolution. *Current Biology*, 25:R911-21. doi: 10.1016/j.cub.2015.07.055.

Archibald, J. M. & Keeling, P. J. 2002. Recycled plastids: A ‘green movement’ in eukaryotic evolution. *Trends in Genetics*, 18:577–584. doi: 10.1016/S0168-9525(02)02777-4.

Aronsson, H. & Jarvis, P. 2002. A simple method for isolating import-competent *Arabidopsis chloroplasts*. *FEBS Letters*, 529:215–220. doi: 10.1016/S0014-5793(02)03342-2.

Baldauf, S. L., Roger, A. J. & Wenk-Siefert, I., Doolittle, W.F. 2000. A Kingdom-Level Phylogeny of Eukaryotes Based on Combined Protein Data. *Science*, 290:972–977. doi: 10.1126/science.290.5493.972.

Baldauf, S. L. 2008. An overview of the phylogeny and diversity of eukaryotes. *Journal of Systematics and Evolution*:236–273.

Bäumer, D., Preisfeld, A. & Ruppel, H. G. 2001. Isolation and characterization of paramylon synthase from *Euglena gracilis* (Euglenophyceae). *Journal of Phycology*, 37:38–46. doi: 10.1046/j.1529-8817.2001.037001038.x.

Bennett, M. S. & Triemer, R. E. 2015. Chloroplast Genome Evolution in the Euglenaceae. *J. Eukaryot. Microbiol.*:n/a. doi: 10.1111/jeu.12235.

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

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*, 53:66–73. doi: 10.2216/13-198.1.

Bennett, M. S., Shiu, S.-H. & Triemer, R. E. 2017. A rare case of plastid protein-coding gene duplication in the chloroplast genome of *Euglena archaeoplastidiata* (Euglenophyta). *Journal of Phycology*, 53:493–502. doi: 10.1111/jpy.12531.

Benson, G. 1999. Tandem repeats finder: A program to analyze DNA sequences. *Nucl Acids Res*, 27:573–580.

Berget, S. M., Moore, C. & Sharp, P. A. 1977. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences*, 74:3171–3175. doi: 10.1073/pnas.74.8.3171.

Bernt, M., Merkle, D., Ramsch, K., Fritzsch, G., Perseke, M., Bernhard, D., Schlegel, M., Stadler, P. F. & Middendorf, M. 2007. CREx: Inferring genomic rearrangements based on common intervals. *Bioinformatics*, 23:2957–2958. doi: 10.1093/bioinformatics/btm468.

Bhattacharya, D., Archibald, J. M., Weber, A. P. M. & Reyes-Prieto, A. 2007. How do endosymbionts become organelles? Understanding early events in plastid evolution. *Bioessays*, 29:1239–1246. doi: 10.1002/bies.20671.

Bicudo, C. E. d. M. & Menezes, M. 2016. Phylogeny and Classification of Euglenophyceae: A Brief Review. *Front. Ecol. Evol.*, 4:429. doi: 10.3389/fevo.2016.00017.

Bleidorn, C. 2017. Phylogenomics: An Introduction. Springer International Publishing, Cham. 87.

Bogorad, L. & Vasil, I. K. 1991.: Cell Culture and Somatic Cell Genetics of Plants *In:* Bogorad, L., Vasil, I. K. (eds) The Molecular Biology of Plastids. Elsevier Science, Oxford.

Boore, J. L. & Brown, W. M. 1998. Big trees from little genomes: Mitochondrial gene order as a phylogenetic tool. *Current Opinion in Genetics & Development*, 8:668–674.

Boore, J. L. 2006. The use of genome-level characters for phylogenetic reconstruction. *Trends Ecol Evol (Amst )*, 21:439–446. doi: 10.1016/j.tree.2006.05.009.

Breglia, S. A., Slamovits, C. H. & Leander, B. S. 2007. Phylogeny of phagotrophic euglenids (Euglenozoa) as inferred from hsp90 gene sequences. *J. Eukaryot. Microbiol.*, 54:86–92. doi: 10.1111/j.1550-7408.2006.00233.x.

Breglia, S. A., Yubuki, N., Hoppenrath, M. & Leander, B. S. 2010. Ultrastructure and molecular phylogenetic position of a novel euglenozoan with extrusive episymbiotic bacteria: *Bihospites bacati* n. gen. et sp. (Symbiontida). *BMC Microbiol*, 10:145. doi: 10.1186/1471-2180-10-145.

Brosnan, S., Shin, W., Kjer, K. M. & Triemer, R. E. 2003. Phylogeny of the photosynthetic euglenophytes inferred from the nuclear SSU and partial LSU rDNA. *International Journal of Systematic and Evolutionary Microbiology*, 53:1175–1186. doi: 10.1099/ijs.0.02518-0.

Brosnan, S., Brown, P. J. P., Farmer, M. A. & Triemer, R. E. 2005. Morphological separation of the euglenoid genera *Trachelomonas* and *Strombomonas* (Euglenophyta) based on lorica development and posterior strip reduction. *Journal of Phycoloy*, 41:590–605. doi: 10.1111/j.1529-8817.2005.00068.x.

Brouard J., Turmel M., Otis C., Lemieux C. 2016. Proliferation of group II introns in the chloroplast genome of the green alga *Oedocladium carolinianum* (Chlorophyceae). PeerJ: e2627. doi: 10.7717/peerj.2627.

Brown, P. J. P., Zakry, B.e. & Farmer, M. A. 2003. Plastid morphology, ultrastructure, and development in *Colacium* and the loricate Euglenophytes (Euglenophyceae). *Journal of Phycology*, 39:115–121. doi: 10.1046/j.1529-8817.2003.01244.x.

Buetow, D. E. 1982. Biology of euglena *In:* Buetow, D. E. (ed) Physiology. Elsevier Science, Oxford. v. 3

Burge, S. W., Daub, J., Eberhardt, R., Tate, J., Barquist, L., Nawrocki, E. P., Eddy, S. R., Gardner, P. P. & Bateman, A. 2013. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Research*, 41:D226-32. doi: 10.1093/nar/gks1005.

Busse, I. & Preisfeld, A. 2002a. Phylogenetic position of *Rhynchopus* sp. and *Diplonema ambulator* as indicated by analyses of euglenozoan small subunit ribosomal DNA. *Gene*, 284:83–91.

Busse, I. & Preisfeld, A. 2002b. Unusually expanded SSU ribosomal DNA of primary osmotrophic euglenids: molecular evolution and phylogenetic inference. *Journal of Molecular Evolution*, 55:757–767. doi: 10.1007/s00239-002-2371-8.

Busse, I. & Preisfeld, A. 2003. Discovery of a group I intron in the SSU rDNA of *Ploeotia costata* (Euglenozoa). *Protist*, 154:57–69. doi: 10.1078/143446103764928495.

Busse, I., Patterson, D. J. & Preisfeld, A. 2003. Phylogeny of phagotrophic euglenids (euglenozoa): A molecular approach based on culture material and environmental samples. *Journal of Phycology*, 39:828–836. doi: 10.1046/j.1529-8817.2003.02178.x.

Bütschli, O. 1884. Protozoa. *In:* Bronn, H. G. (ed.), Dr. H.G. Bronn's Klassen und Ordnungen des Thier-Reichs, wissenschaftlich dargestellt in Wort und Bild. 2, Abt. II Mastigophora. C. F. Winter`sche Verlagsbuchhandlung, Leipzig und Heidelberg:617–1097.

Byun, Y. & Han, K. 2006. PseudoViewer: Web application and web service for visualizing RNA pseudoknots and secondary structures. *Nucl Acids Res*, 34:W416-22. doi: 10.1093/nar/gkl210.

Candales, M. A., Duong, A., Hood, K. S., Li, T., Neufeld, R. A. E., Sun, R., McNeil, B. A., Wu, L., Jarding, A. M. & Zimmerly, S. 2011. Database for bacterial group II introns. *Nucleic Acids Research* D1: D187–D190. doi: 10.1093/nar/gkr1043.

Casiot, C., Bruneel, O., Personné, J.-C., Leblanc, M. & Elbaz-Poulichet, F. 2004. Arsenic oxidation and bioaccumulation by the acidophilic protozoan, *Euglena mutabilis*, in acid mine drainage (Carnoulès, France). The Science of the total environment 2-3: 259–267. doi: 10.1016/j.scitotenv.2003.08.004.

Cattolico R., Jacobs M.A., Zhou Y., Chang J., Duplessis M., Lybrand T., McKay J., Ong H., Sims E., Rocap G. 2008. Chloroplast genome sequencing analysis of *Heterosigma akashiwo* CCMP452 (West Atlantic) and NIES293 (West Pacific) strains. *BMC Genomics* 1: 211. doi: 10.1186/1471-2164-9-211.

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

Cavalier-Smith, T. 1982. The origins of plastids. *Biological Journal of the Linnean Society*, 17:289–306. doi: 10.1111/j.1095-8312.1982.tb02023.x.

Cavalier-Smith, T. 1985. Selfish DNA and the origin of introns. *Nature*, 315:283–284.

Cavalier-Smith, T. 1986. The kingdom Chromista: origin and systematics. *In:* Round, F. E., Chapman, D. J. (eds.), Progress in phycological research. Biopress Ltd, Bristol UK. 4:309–347.

Cavalier-Smith, T. 1993. Kingdom protozoa and its 18 phyla. *Microbiological reviews*, 57:953–994.

Cavalier-Smith, T. 1999. Principles of Protein and Lipid Targeting in Secondary Symbiogenesis: Euglenoid, Dinoflagellate, and Sporozoan Plastid Origins and the Eukaryote Family Tree, 2. *Journal of Eukaryotic Microbiology*, 46:347–366. doi: 10.1111/j.1550-7408.1999.tb04614.x.

Cavalier-Smith, T. 2002. The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *International Journal of Systematic and Evolutionary Microbiology*, 52

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

Chan, R. L., Keller, M., Canaday, J., Weil, J. H. & Imbault, P. 1990. 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.

Chernomor, O., Haeseler, A. von & Minh, B. Q. 2016. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. *Systematic Biology*, 65:997–1008. doi: 10.1093/sysbio/syw037.

Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. 1977. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. *Cell*, 12:1–8. doi: 10.1016/0092-8674(77)90180-5.

Christopher, D. A. & Hallick, R. B. 1989. Euglena gracilis chloroplast ribosomal protein operon: a new chloroplast gene for ribosomal protein L5 and description of a novel organelle intron category designated group III. *Nucl Acids Res*, 17:7591–7608. doi: 10.1093/nar/17.19.7591.

Ciugulea, I. & Triemer, R. E. 2010. A color atlas of photosynthetic euglenoids. Michigan State University Press, East Lansing. xx, 204.

Ciugulea, I., Nudelman, M. A., Brosnan, S. & Triemer, R. E. 2008. Phylogeny of the euglenoid loricate genera *Trachelomonas* and *Strombomonas* (Euglenophyta) inferred from nuclear SSU and LSU rDNA. *Journal of Phycology*, 44:406–418. doi: 10.1111/j.1529-8817.2008.00472.x.

Conant, G. C. & Wolfe, K. H. 2008. GenomeVx: simple web-based creation of editable circular chromosome maps. *Bioinformatics*, 24:861–862. doi: 10.1093/bioinformatics/btm598.

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*, 18:467–471. doi: 10.1016/0968-0004(93)90008-B.

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*, 19:6491–6497. doi: 10.1093/nar/19.23.6491.

Copertino, D. W. & Hallick, R. B. 1991. Group II twintron: an intron within an intron in a chloroplast cytochrome b-559 gene. *EMBO J.*, 10:433–442.

Copertino, D. W., Shigeoka, S. & Hallick, R. B. 1992. Chloroplast group III twintron excision utilizing multiple 5'- and 3'-splice sites. *EMBO J.*, 11:5041–5050.

Copertino, D. W., Hall, E. T., Van Hook, F W, Jenkins, K. P. & Hallick, R. B. 1994. A group III twintron encoding a maturase-like gene excises through lariat intermediates. *Nucleic Acids Research*, 22:1029–1036. doi: 10.1093/nar/22.6.1029.

Cramer, M. & Myers, J. 1952. Growth and photosynthetic characteristics of *Euglena gracilis*. *Arch. Microbiol.*, 17:384–402. doi: 10.1007/BF00410835.

Curtis, B. A., Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M. C., Ball, S. G., Gile, G. H., Hirakawa, Y., Hopkins, J. F., Kuo, A., Rensing, S. A., Schmutz, J., Symeonidi, A., Elias, M., Eveleigh, R. J. M., Herman, E. K., Klute, M. J., Nakayama, T., Oborník, M., Reyes-Prieto, A., Armbrust, E. V., Aves, S. J., Beiko, R. G., Coutinho, P., Dacks, J. B., Durnford, D. G., Fast, N. M., Green, B. R., Grisdale, C. J., Hempel, F., Henrissat, B., Höppner, M. P., Ishida, K.-i., Kim, E., Kořený, L., Kroth, P. G., Liu, Y., Malik, S.-B., Maier, U. G., McRose, D., Mock, T., Neilson, J. A. D., Onodera, N. T., Poole, A. M., Pritham, E. J., Richards, T. A., Rocap, G., Roy, S. W., Sarai, C., Schaack, S., Shirato, S., Slamovits, C. H., Spencer, D. F., Suzuki, S., Worden, A. Z., Zauner, S., Barry, K., Bell, C., Bharti, A. K., Crow, J. A., Grimwood, J., Kramer, R., Lindquist, E., Lucas, S., Salamov, A., McFadden, G. I., Lane, C. E., Keeling, P. J., Gray, M. W., Grigoriev, I. V. & Archibald, J. M. 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature*, 492:59–65. doi: 10.1038/nature11681.

Dabbagh, N. & Preisfeld, A. 2017. The Chloroplast Genome of *Euglena mutabilis* -Cluster Arrangement, Intron Analysis, and Intrageneric Trends. *J. Eukaryot. Microbiol.*, 64:31–44. doi: 10.1111/jeu.12334.

Dabbagh, N., Bennett, M. S., Triemer, R. E. & Preisfeld, A. 2017. Chloroplast genome expansion by intron multiplication in the basal psychrophilic euglenoid *Eutreptiella pomquetensis*. *PeerJ*, 5:e3725. doi: 10.7717/peerj.3725.

Dai, L., Toor, N., Olson, R., Keeping, A. & Zimmerly, S. 2003. Database for mobile group II introns. *Nucleic Acids Research*, 31:424–426.

Dai L., Zimmerly S.. 2003. ORF-less and reverse-transcriptase-encoding group II introns in archaebacteria, with a pattern of homing into related group II intron ORFs. RNA (New York, N.Y.) 1: 14–19. doi: 10.1261/rna.2126203.

Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. 2004. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Research*, 14:1394–1403. doi: 10.1101/gr.2289704.

Deflandre, G. 1930. *Strombomonas* nouveau genere d'euglénacées (Trachelomonas Ehr. pro parte). *Archiv für Protistenkunde*, 69:551–614.

Delwiche, C. F. 1999. Tracing the Thread of Plastid Diversity through the Tapestry of Life. *Am Nat*, 154:S164-S177. doi: 10.1086/303291.

Deschamps, P. & Moreira, D. 2009. Signal conflicts in the phylogeny of the primary photosynthetic eukaryotes. *Molecular Biology and Evolution*, 26:2745–2753. doi: 10.1093/molbev/msp189.

Doetsch, N. A., Thompson, M. D. & Hallick, R. B. 1998. A maturase-encoding group III twintron is conserved in deeply rooted euglenoid species: are group III introns the chicken or the egg? *Mol. Biol. Evol.*, 15:76–86.

Doetsch, N. A., Thompson, M. D., Favreau, M. R. & Hallick, R. B. 2001. Comparison of *psb*K operon organization and group III intron content in chloroplast genomes of 12 Euglenoid species. *Molecular and General Genetics MGG*, 264:682–690. doi: 10.1007/s004380000355.

Donath, A. & Stadler, P. F. 2014. 25 Molecular morphology: Higher order characters derivable from sequence information. *In:* Wägele, J. W., Bartolomaeus, T. (eds.), Deep Metazoan Phylogeny: The Backbone of the Tree of Life. De Gruyter, Berlin, Boston

Doolittle, W. F. 1978. Genes in pieces: Were they ever together? *Nature*, 272:581–582. doi: 10.1038/272581a0.

Doolittle, W. F. 1987. The Origin and Function of Intervening Sequences in DNA: A Review. *Am Nat*, 130:915–928. doi: 10.1086/284755.

Drager, R. G. & Hallick, R. B. 1993. A complex twintron is excised as four individual introns. *Nucl Acids Res*, 21:2389–2394. doi: 10.1093/nar/21.10.2389.

Dragoş, N., Péterfi, L. Ş. & Popescu, C. 1997. Comparative fine structure of pellicular cytoskeleton in EuglenaEhrenberg. *Archiv für Protistenkunde*, 148:277–285. doi: 10.1016/S0003-9365(97)80008-5.

Eberhard, S., Drapier, D. & Wollman, F.-A. 2002. Searching limiting steps in the expression of chloroplast-encoded proteins: relations between gene copy number, transcription, transcript abundance and translation rate in the chloroplast of *Chlamydomonas reinhardtii*. *Plant J*, 31:149–160. doi: 10.1046/j.1365-313X.2002.01340.x.

Ehrenberg, C. G. 1830. Organisation, systematik und geographisches verhältniss der infusionsthierchen. Zwei vorträge, in der Akademie der wissenschaften zu Berlin gehalten in den jahren 1828 und 1830, von C.G. Ehrenberg. Mit 8 kupfertafeln in folio. Druckerei der Königlichen akademie der wissenschaften, Berlin.

Ehrenberg, C. G. 1833. Dritter Beitrag zur Erkenntnis grosser Organisation in der Richtung des kleinsten Raumes. Königlichen Akademie der Wissenschaften, Berlin.

Ekelund, F. & Patterson, D. J. 1997. Some Heterotrophic Flagellates from a Cultivated Garden Soil in Australia. *Archiv für Protistenkunde*, 148:461–478. doi: 10.1016/S0003-9365(97)80022-X.

Elowitz, M. B., Levine, A. J., Siggia, E. D. & Swain, P. S. 2002. Stochastic gene expression in a single cell. *Science*, 297:1183–1186. doi: 10.1126/science.1070919.

Falkowski, P. G., Katz, M. E., Knoll, A. H., Quigg, A., Raven, J. A., Schofield, O. & Taylor, F. J. R. 2004. The evolution of modern eukaryotic phytoplankton. *Science*, 305:354–360. doi: 10.1126/science.1095964.

Farmer, M. A. & Triemer, R. E. 1988. Flagellar systems in the euglenoid flagellates. *Biosystems*, 21:283–291. doi: 10.1016/0303-2647(88)90024-X.

Farmer, M. A. 2009. Euglenozoa. *In:* Schaechter, M. (ed.), Encyclopedia of microbiology, 3. ed. Elsevier, Amsterdam:634–645.

Fenchel, T., Bernard, C., Esteban, G., Finlay, B. J., Hansen, P. J. & Iversen, N. 1995. Microbial diversity and activity in a Danish Fjord with anoxic deep water. *Ophelia*, 43:45–100. doi: 10.1080/00785326.1995.10430576.

Gao, L., Zhou, Y., Wang, Z.-W., Su, Y.-J. & Wang, T. 2011. Evolution of the *rpo*B-*psb*Z region in fern plastid genomes: notable structural rearrangements and highly variable intergenic spacers. *BMC Plant Biology*: 64. doi: 10.1186/1471-2229-11-64.

Gentil, J., Hempel, F., Moog, D., Zauner, S. & Maier, U. G. 2017. Review: Origin of complex algae by secondary endosymbiosis: a journey through time. *Protoplasmadoi:* 10.1007/s00709-017-1098-8.

Gibbs, S. P. 1978. The chloroplasts of *Euglena* may have evolved from symbiotic green algae. *Can. J. Bot.*, 56:2883–2889. doi: 10.1139/b78-345.

Gibbs, S. P. 1981. The chloroplasts of some algal groups may have evolved from endosymbiotic eukaryotic algae. *Ann N Y Acad Sci*, 361:193–208. doi: 10.1111/j.1749-6632.1981.tb54365.x.

Gilbert, W. 1978. Why genes in pieces? *Nature*, 271:501. doi: 10.1038/271501a0.

Gilbert, W. 1987. The exon theory of genes. *Cold Spring Harb Symp Quant Biol*, 52:901–905.

Gillott, M. A. & Triemer, R. E. 1978. The ultrastructure of cell division in Euglena gracilis. *Journal of cell science*, 31:25–35.

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

Gojdics, M. 1953. The Genus Euglena. The University of Wisconsin Press, Madison.

Gojdics, M. & Lowndes, A. G. 1954. The Genus *Euglena*. Journal of the Marine Biological Association of the United Kingdom 01: 296. doi: 10.1017/S0025315400003611.

Gottlieb, J. 1850. Über eine neue, mit Stärkemehl isomere Substanz;. *Ann. Chem. Pharm.*, 75:51–61. doi: 10.1002/jlac.18500750105.

Gould, S. B., Waller, R. F. & McFadden, G. I. 2008. Plastid Evolution. *Annu. Rev. Plant Biol.*, 59:491–517. doi: 10.1146/annurev.arplant.59.032607.092915.

Gould, S. B., Maier, U.-G. & Martin, W. F. 2015. Protein import and the origin of red complex plastids. *Current Biology*, 25:R515-21. doi: 10.1016/j.cub.2015.04.033.

Gray, M. W. 2012. Mitochondrial evolution. *Cold Spring Harbor Perspectives in Biology*, 4:a011403. doi: 10.1101/cshperspect.a011403.

Guillard, R. R. L. & Hargraves, P. E. 1993. Stichochrysis immobilis is a diatom, not a chrysophyte. *Phycologia*, 32:234–236. doi: 10.2216/i0031-8884-32-3-234.1.

Häder, D.-P. & Melkonian, M. 1983. Phototaxis in the gliding flagellate, *Euglena mutabilis*. Archives of Microbiology 1: 25–29. doi: 10.1007/BF00419477.

Hafez, M. & Hausner, G. 2015. Convergent evolution of twintron-like configurations: One is never enough. *RNA Biol*, 12:1275–1288. doi: 10.1080/15476286.2015.1103427.

Hallick, R. B., Hong, L., Drager, R. G., Favreau, M. R., Monfort, A., Orsat, B., Spielmann, A. & Stutz, E. 1993. Complete sequence of *Euglena gracilis* chloroplast DNA. *Nucleic Acids Research*, 21:3537–3544. doi: 10.1093/nar/21.15.3537.

Harding, R. M., Boyce, A. J. & Clegg, J. B. 1992. The evolution of tandemly repetitive DNA: recombination rules. *Genetics* 3: 847–859.

Hong, L. & Hallick, R. B. 1994. A group III intron is formed from domains of two individual group II introns. *Genes Dev.*, 8:1589–1599. doi: 10.1101/gad.8.13.1589.

Honigberg, B. M. 1963. Evolutionary and Systematic Relationships in the Flagellate Order Trichomonadida Kirby*. *The Journal of Protozoology*, 10:20–63. doi: 10.1111/j.1550-7408.1963.tb01635.x.

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:e33746. doi: 10.1371/journal.pone.0033746.

Hubert-Pestalozzi, G. 1955. Das Phytoplankton des Süßwassers. Systematik und Biologie.: 4.Teil:Euglenophyceen. *In:* Thienemann, A. (ed.), Die Binnengewässer. Schweizerbart, Stuttgart. Band XVI:1–606.

International Botanical Congress 2012. International code of nomenclature for algae, fungi and plants (Melbourne code): Adopted by the Eighteenth International Botanical Congress, Melbourne, Australia, July 2011. Regnum vegetabile. 154. Koeltz, Königstein.

International Commission on Zoological Nomenclature; International Union of Biological Sciences 1999. International code of zoological nomenclature, 4. ed. International Trust for Zoological Nomenclature, London.

Jackson, C. J. & Reyes-Prieto, A. 2014. The mitochondrial genomes of the glaucophytes *Gloeochaete wittrockiana* and *Cyanoptyche gloeocystis*: Multilocus phylogenetics suggests a monophyletic archaeplastida. *Genome Biol Evol*, 6:2774–2785. doi: 10.1093/gbe/evu218.

Kadereit, J. W., Körner, C., Kost, B. & Sonnewald, U. 2014. Strasburger − Lehrbuch der Pflanzenwissenschaften. Springer Berlin Heidelberg, Berlin, Heidelberg. 919972.

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*, 62:362–373. doi: 10.1111/jeu.12192.

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

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P. & Drummond, A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, 28:1647–1649. doi: 10.1093/bioinformatics/bts199.

Keeling, P. J. 2004. Diversity and evolutionary history of plastids and their hosts. *Am J Bot*, 91:1481–1493. doi: 10.3732/ajb.91.10.1481.

Keeling, P. J. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J. Eukaryot. Microbiol.*, 56:1–8. doi: 10.1111/j.1550-7408.2008.00371.x.

Keeling, P. J. 2010. The endosymbiotic origin, diversification and fate of plastids. *Philos Trans R Soc Lond , B, Biol Sci*, 365:729–748. doi: 10.1098/rstb.2009.0103.

Keeling, P. J. 2013. The number, speed, and impact of plastid endosymbioses in eukaryotic evolution. *Annu Rev Plant Biol*, 64:583–607. doi: 10.1146/annurev-arplant-050312-120144.

Kelchner, S. A. 2002. Group II introns as phylogenetic tools: Structure, function, and evolutionary constraints. *Americn Journal of Botany*, 89:1651–1669. doi: 10.3732/ajb.89.10.1651.

Khan, H. & Archibald, J. M. 2008. Lateral transfer of introns in the cryptophyte plastid genome. Nucleic Acids Research 9: 3043–3053. doi: 10.1093/nar/gkn095.

Kies, L. 1967. Oogamie bei *Eremosphaera viridis* De Bary. *Flora oder Allgemeine botanische Zeitung. Abt. B, Morphologie und Geobotanik*, 157:1–12.

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

Kim, J. I., Shin, W. & Triemer, R. E. 2010. Multigene analyses of photosynthetic euglenoids and new family Phacaceae (Euglenales). *Journal of Phycology*, 46:1278–1287. doi: 10.1111/j.1529-8817.2010.00910.x.

Kim, J. I., Linton, E. W. & Shin, W. 2015. Taxon-rich multigene phylogeny of the photosynthetic euglenoids (Euglenophyceae). *Front. Ecol. Evol.*, 3:254. doi: 10.3389/fevo.2015.00098.

Kivic, P. A. & Vesk, M. 1972. Structure and function in the euglenoid eyespot apparatus: The fine structure, and response to environmental changes. *Planta*, 105:1–14. doi: 10.1007/BF00385158.

Kivic, P. A. & Walne, P. L. 1984. An evaluation of a possible phylogenetic relationship between the Euglenophyta and Kinetoplastida. *Origins Life Evol Biosphere*, 13:269–288. doi: 10.1007/BF00927177.

Koonin, E. V. 2006. The origin of introns and their role in eukaryogenesis: A compromise solution to the introns-early versus introns-late debate? *Biol Direct*, 1:22. doi: 10.1186/1745-6150-1-22.

Kosmala, S., Milanowski, R., Brzóska, K., Pękala, M., Kwiatowski, J. & Zakryś, B. 2007. Phylogeny and systematics of the genus *Monomorphina* (Euglenaceae) based on morphological and molecular data. *Journal of Phycology*, 43:171–185. doi: 10.1111/j.1529-8817.2006.00298.x.

Kumar, S., Stecher, G. & Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution*, 33:1870–1874. doi: 10.1093/molbev/msw054.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J. & Giegerich, R. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Research*, 29:4633–4642. doi: 10.1093/nar/29.22.4633.

Lagesen, K., Hallin, P., Rødland, E. A., Staerfeldt, H.-H., Rognes, T. & Ussery, D. W. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35:3100–3108. doi: 10.1093/nar/gkm160.

Lambowitz, A. M. & Zimmerly, S. 2011. Group II Introns: Mobile Ribozymes that Invade DNA. *Cold Spring Harbor Perspectives in Biology*, 3:a003616. doi: 10.1101/cshperspect.a003616.

Lambowitz, A. M. & Belfort, M. 2015. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. *Microbiology Spectrum*, 3doi: 10.1128/microbiolspec.MDNA3-0050-2014.

Lambowitz, A. M., Caprara, M. G., Zimmerly, S. & Perlman, P. S. 1999. Group I and Group II ribozymes as RNPs:clues to the past and guides to the future. *In:* Gesteland, R. F., Chech, T. R., Atkins J.F. (eds.), The RNA World, 2nd edn. Cold Spring Harbor Laboratory Press, *Cold Spring Harbor*, NY:451–485.

Lauterborn, R. 1895. Protozoenstudien II. *Paulinella chromatophora* nov. gen. spec., ein beschalter Rhizopode des Süßwassers mit blaugrünen chromatophorenartigen Einschlüssen. *Zeitschrift für wissenschafliche Zoologie*, 1895:537–544.

Le Hir, H., Nott, A. & Moore, M. J. 2003. How introns influence and enhance eukaryotic gene expression. *Trends in Biochemical Sciences*, 28:215–220. doi: 10.1016/S0968-0004(03)00052-5.

Leander, B. S. & Farmer, M. A. 2000. Comparative morphology of the euglenid pellicle. I. Patterns of strips and pores. *Journal of Eukaryotic Microbiology*, 47:469–479.

Leander, B. S. 2004. Did trypanosomatid parasites have photosynthetic ancestors? *Trends Microbiol*ogy, 12:251–258. doi: 10.1016/j.tim.2004.04.001.

Leander, B. S., Triemer, R. E. & Farmer, M. A. 2001. Character evolution in heterotrophic euglenids. *Eur J Protistol*, 37:337–356. doi: 10.1078/0932-4739-00842.

Leander B.S., Witek R.P., Farmer M.A. 2001. Trends in the evolution of the euglenid pellicle. Evolution; I*nternational Journal of Organic Evolution* 11: 2215–2235.

Leander, B. S., Esson, H. J. & Breglia, S. A. 2007. Macroevolution of complex cytoskeletal systems in euglenids. *Bioessays*, 29:987–1000. doi: 10.1002/bies.20645.

Leedale, G. F. 1967. Euglenoid Flagellates. Prentice-Hall, Engelwood-Cliffs, New Jersey. 242.

Leliaert, F., Smith, D. R., Moreau, H., Herron, M. D., Verbruggen, H., Delwiche, C. F. & Clerck, O. de 2012. Phylogeny and Molecular Evolution of the Green Algae. *Critical Reviews in Plant Sciences*, 31:1–46. doi: 10.1080/07352689.2011.615705.

Lemieux, C., Otis, C. & Turmel, M. 2007. A clade uniting the green algae *Mesostigma viride* and *Chlorokybus atmophyticus* represents the deepest branch of the Streptophyta in chloroplast genome-based phylogenies. BMC biology: 2. doi: 10.1186/1741-7007-5-2.

Linton E.W., Hittner D., Lewandowski C., Auld T., Triemer R.E. 1999. A molecular study of euglenoid phylogeny using small subunit rDNA. *J. Eukaryot. Microbiol* 2: 217–223.

Linton, E. W., Karnkowska-Ishikawa, A., Im Kim, J., Shin, W., Bennett, M. S., Kwiatowski, J., Zakryś, B. & Triemer, R. E. 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*, 161:603–619. doi: 10.1016/j.protis.2010.02.002.

Linton E.W., Nudelman M.A., Conforti V., Triemer R.E. 2000. A molecular analysis of the Euglenophytes using SSU rDNA. *Journal of Phycology* 4: 740–746 DOI 10.1046/j.1529-8817.2000.99226.x.

Mackiewicz, P. & Gagat, P. 2014. Monophyly of Archaeplastida supergroup and relationships among its lineages in the light of phylogenetic and phylogenomic studies. Are we close to a consensus? *Acta Soc Bot Pol*, 83:263–280. doi: 10.5586/asbp.2014.044.

Marin B. 2004. Origin and Fate of Chloroplasts in the Euglenoida. *Protist,* 1: 13–14. doi: 10.1078/1434461000159.

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*, 154:99–145.

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

Martínez-Abarca, F. & Toro, N. 2000. RecA-independent ectopic transposition in vivo of a bacterial group II intron. *Nucl Acids Res*, 28:4397–4402.

Mascarenhas, D., Mettler, I. J., Pierce, D. A. & Lowe, H. W. 1990. Intron-mediated enhancement of heterologous gene expression in maize. *Plant Mol Biol*, 15:913–920.

Maslov, D. A., Yasuhira, S. & Simpson, L. 1999. Phylogenetic Affinities of *Diplonema* within the Euglenozoa as Inferred from the SSU rRNA Gene and Partial COI Protein Sequences. *Protist*, 150:33–42. doi: 10.1016/S1434-4610(99)70007-6.

McFadden, G. I. & Waller, R. F. 1997. Plastids in parasites of humans. *BioEssays*, 19:1033–1040. doi: 10.1002/bies.950191114.

McFadden, G. I. 2014. Origin and evolution of plastids and photosynthesis in eukaryotes. *Cold Spring Harbor Perspectives in Biology*, 6:a016105. doi: 10.1101/cshperspect.a016105.

McFadden, G. I. 2001. Primary and secondary endosymbiosis and the origin of plastids. *Journal of Phycology*, 37:951–959. doi: 10.1046/j.1529-8817.2001.01126.x.

McLachlan, J. L., Seguel, M. R. & Fritz, L. 1994. *Tetreutreptia pomquetensis* gen. et sp. nov. (Euglenophyceae): a quadriflagellate, phototrophic marine euglenoid. *Journal of Phycology*, 30:538–544. doi: 10.1111/j.0022-3646.1994.00538.x.

Melkonian, M. & Mollenhauer, D. 2005. Robert Lauterborn (1869-1952) and his *Paulinella chromatophora*. *Protist*, 156:253–262. doi: 10.1016/j.protis.2005.06.001.

Meng, Q., Wang, Y. & Liu, X.-Q. 2005. An intron-encoded protein assists RNA splicing of multiple similar introns of different bacterial genes. *Journal of Biological Chemistry*, 280:35085–35088. doi: 10.1074/jbc.C500328200.

Merendino L., Perron K., Rahire M., Howald I., Rochaix J.D., Goldschmidt-Clermont M. 2006. A novel mutlifunctional factor involved in trans-splicing of chloroplast introns in *Chlamydomonas*. Nucleic Acids Research 34:262-274.

Mereschkowsky, C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. *Biologisches Centralblatt*:593–604.

Michel, F., Kazuhiko, U. & Haruo, O. 1989. Comparative and functional anatomy of group II catalytic introns - a review. *Gene*, 82:5–30. doi: 10.1016/0378-1119(89)90026-7.

Michel, F. & Ferat, J. L. 1995. Structure and activities of group II introns. *Annual Reviews of Biochemistry*, 64:435–461. doi: 10.1146/annurev.bi.64.070195.002251.

Milanowski, R., Kosmala, S., Zakryś, B. & Kwiatowski, J. 2006. Phylogeny of photosyntetic Euglenophytes based on combined chloroplast and cytoplasmic SSU rDNA sequence analysis. *Journal of Phycology*, 42:721–730. doi: 10.1111/j.1529-8817.2006.00216.x.

Minh, B. Q., Nguyen, M. A. T. & Haeseler, A. von 2013. Ultrafast approximation for phylogenetic bootstrap. *Molecular Biology and. Evolution*, 30:1188–1195. doi: 10.1093/molbev/mst024.

Mitchell, D. R. 2007. The evolution of eukaryotic cilia and flagella as motile and sensory organelles. *Adv Exp Med Biol*, 607:130–140. doi: 10.1007/978-0-387-74021-8_11.

Mohr G., Ghanem E., Lambowitz A.M. 2010. Mechanisms used for genomic proliferation by thermophilic group II introns. PLoS Biology 8: e1000391. doi: 10.1371/journal.pbio.1000391.

Mohr, G., Perlman, P. S. & Lambowitz, A. M. 1993. Evolutionary relationships among group II intron-encoded proteins and identification of a conserved domain that may be related to maturase function. *Nucleic Acids Research*, 21:4991–4997.

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

Müllner, 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*, 51:783–791. doi: 10.1099/00207713-51-3-783.

Muñoz-Gómez, S. A., Wideman, J. G., Roger, A. J. & Slamovits, C. H. 2017. The Origin of Mitochondrial Cristae from Alphaproteobacteria. *Molecular Biology and Evolution*, 34:943–956. doi: 10.1093/molbev/msw298.

Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M. & Kumlin, E. 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. Science (New York, N.Y.) 4796: 1616–1622.

Nakayama, T. & Ishida, K.-i. 2009. Another acquisition of a primary photosynthetic organelle is underway in Paulinella chromatophora. *Current Biology*, 19:R284-5. doi: 10.1016/j.cub.2009.02.043.

Nguyen, L.-T., Schmidt, H. A., Haeseler, A. von & Minh, B. Q. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32:268–274. doi: 10.1093/molbev/msu300.

Niehuis, O., Hartig, G., Grath, S., Pohl, H., Lehmann, J., Tafer, H., Donath, A., Krauss, V., Eisenhardt, C., Hertel, J., Petersen, M., Mayer, C., Meusemann, K., Peters, R. S., Stadler, P. F., Beutel, R. G., Bornberg-Bauer, E., McKenna, D. D. & Misof, B. 2012. Genomic and morphological evidence converge to resolve the enigma of Strepsiptera. *Current Biology*, 22:1309–1313. doi: 10.1016/j.cub.2012.05.018.

Nowack, E. C. M. & Grossman, A. R. 2012. Trafficking of protein into the recently established photosynthetic organelles of *Paulinella chromatophora*. *Proc Natl Acad Sci U S A*, 109:5340–5345. doi: 10.1073/pnas.1118800109.

Nowack, E. C. M., Vogel, H., Groth, M., Grossman, A. R., Melkonian, M. & Glöckner, G. 2011. Endosymbiotic gene transfer and transcriptional regulation of transferred genes in *Paulinella chromatophora*. *Molecular Biology and Evolution*, 28:407–422. doi: 10.1093/molbev/msq209.

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*. *Proc Natl Acad Sci U S A*, 113:12214–12219. doi: 10.1073/pnas.1608016113.

Nowack, E. C.M. 2014. *Paulinella chromatophora* – rethinking the transition from endosymbiont to organelle. *Acta Soc Bot Pol*, 83:387–397. doi: 10.5586/asbp.2014.049.

Nudelman, M. A., Rossi, M. S., Conforti, V. & Triemer, R. E. 2003. Phylogeny of Euglenophyceae based on small subunit rDNA sequences: Taxonomic implications. *Journal of Phycology*, 39:226–235. doi: 10.1046/j.1529-8817.2003.02075.x.

Orgel, L. E. & Crick, F. H. 1980. Selfish DNA: The ultimate parasite. *Nature*, 284:604–607.

Paerschke, S., Vollmer, A. H. & Preisfeld, A. 2017. Ultrastructural and immunocytochemical investigation of paramylon combined with new 18S rDNA-based secondary structure analysis clarifies phylogenetic affiliation of Entosiphon sulcatum (Euglenida:

Euglenozoa). *Organisms Diversity and Evolution*, 59:429. doi: 10.1007/s13127-017-0330-x.

Palmer, J. D. 2003. The symbiotic birth and spread of plastids: how many times and whodunit? *Journal of Phycology*, 39:4–12. doi: 10.1046/j.1529-8817.2003.02185.x.

Patterson 1999. The Diversity of Eukaryotes. *Am Nat*, 154:S96-S124. doi: 10.1086/303287.

Perseke, M., Fritzsch, G., Ramsch, K., Bernt, M., Merkle, D., Middendorf, M., Bernhard, D., Stadler, P. F. & Schlegel, M. 2008. Evolution of mitochondrial gene orders in echinoderms. *Molecular Phylogenetics and Evolution*, 47:855–864. doi: 10.1016/j.ympev.2007.11.034.

Peschke, M., Moog, D., Klingl, A., Maier, U. G. & Hempel, F. 2013. Evidence for glycoprotein transport into complex plastids. *Proc Natl Acad Sci U S A*, 110:10860–10865. doi: 10.1073/pnas.1301945110.

Pombert, J.-F., James, E. R., Janouškovec, J., Keeling, P. J. & McCutcheon, 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:e53433. doi: 10.1371/journal.pone.0053433.

Poniewozik, M. 2017. Element Composition of *Trachelomonas* Envelopes (Euglenophyta). *Polish Botanical Journal*, 62:1. doi: 10.1515/pbj-2017-0007.

Posada, D. & Buckley, T. R. 2004. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53:793–808. doi: 10.1080/10635150490522304.

Preisfeld, A., Berger, S., Busse, I., Liller, S. & Ruppel, H. G. 2000. Phylogenetic analyses of various euglenoid taxa (euglenozoa) based on 18s rdna sequence data. *Journal of Phycology*, 36:220–226. doi: 10.1046/j.1529-8817.2000.99091.x.

Preisfeld A., Busse I., Klingberg M., Talke S., Ruppel H.G. 2001. Phylogenetic position and inter-relationships of the osmotrophic euglenids based on SSU rDNA data, with emphasis on the Rhabdomonadales (Euglenozoa). International Journal of Systematic and Evolutionary Microbiology 3: 751–758. doi: 10.1099/00207713-51-3-751.

Pringsheim, E. G. 1956. Contributions towards a monograph of the genus *Euglena*. *Nova Acta Leopold*, 18:1–168.

Ravi V., Khurana J.P., Tyagi A.K., Khurana P. 2008. An update on chloroplast genomes. *Plant Systematics and Evolution* 1-2: 101–122. doi: 10.1007/s00606-007-0608-0.

Reyes-Prieto, A. & Bhattacharya, D. 2007. Phylogeny of nuclear-encoded plastid-targeted proteins supports an early divergence of glaucophytes within Plantae. *Molecular Biology and Evolution*, 24:2358–2361. doi: 10.1093/molbev/msm186.

Rice, P., Longden, I. & Bleasby, A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, 16:276–277.

Robbens, S., Derelle, E., Ferraz, C., Wuyts, J., Moreau, H. & Van de Peer, Yves 2007. The complete chloroplast and mitochondrial DNA sequence of *Ostreococcus tauri*: Organelle genomes of the smallest eukaryote are examples of compaction. *Molecular Biology and Evolution*, 24:956–968. doi: 10.1093/molbev/msm012.

Rodríguez-Ezpeleta, N., Brinkmann, H., Burey, S. C., Roure, B., Burger, G., Löffelhardt, W., Bohnert, H. J., Philippe, H. & Lang, B. F. 2005. Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. *Current Biology,* 15:1325–1330. doi: 10.1016/j.cub.2005.06.040.

Rokas & Holland 2000. Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol (Amst )*, 15:454–459.

Rosati, G., Verni, F., Barsanti, L., Passarelli, V. & Gualtieri, P. 1991. Ultrastructure of the apical zone of *Euglena gracilis*: Photoreceptors and motor apparatus. *Electron Microscopy Reviews*, 4:319–342. doi: 10.1016/0892-0354(91)90008-Z.

Rose, A. B., Elfersi, T., Parra, G. & Korf, I. 2008. Promoter-proximal introns in *Arabidopsis thaliana* are enriched in dispersed signals that elevate gene expression. *THE PLANT CELL ONLINE*, 20:543–551. doi: 10.1105/tpc.107.057190.

Rosowski, J. R. & Coute, A. 1996. Bacteria of the lorica of *Trachelomonas* occur in nature, not just in culture. *Journal of Phycology*, 32:697–698. doi: 10.1111/j.0022-3646.1996.00697.x.

Rosowski, J. R. & Langenberg, W. G. 1994. The near-spineless *Trachelomonas grandis* (Euglenophyceae) superficially appears spiny by attracting bacteria to its surface. *Journal of Phycology*, 30:1012–1022. doi: 10.1111/j.0022-3646.1994.01012.x.

Sagan, L. 1967. On the origin of mitosing cells. *Journal of Theoretical Biology*, 14:225-IN6. doi: 10.1016/0022-5193(67)90079-3.

Slater, Guy St C., Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 1: 31. doi: 10.1186/1471-2105-6-31.

Santillán Torres, J. L., Atteia, A., Claros, M.G. & González-Halphen, D. 2003. Cytochrome f and subunit IV, two essential components of the photosynthetic bf complex typically encoded in the chloroplast genome, are nucleus-encoded in Euglena gracilis. *Biochimica* *et Biophysica Acta (BBA) - Bioenergetics*, 1604:180–189. doi: 10.1016/S0005-2728(03)00058-6.

Schattner, P., Brooks, A. N. & Lowe, T. M. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Research*, 33:W686-W689. doi: 10.1093/nar/gki366.

Schimper, A. F. 1883. Über die Entwicklung der Chlorophyllkörner und Farbkörper. *Botanische Zeitung*:105–114, 121–131, 137–146, 153–162.

Schnepf, E. & Deichgräber, G. 1984. Myzocytosis, a kind of endocytosis with implications to compartmentation in endosymbiosis. *Naturwissenschaften*, 71:218–219. doi: 10.1007/BF00490442.

Schwartzbach, S. D. & Shigeoka, S. 2017. Advances in Experimental Medicine and Biology: Biochemistry, Cell and Molecular Biology *In:* Schwartzbach, S. D., Shigeoka, S. (eds) Euglena. Springer International Publishing, Cham.s.l. 979

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

Shih, P. M. & Matzke, N. J. 2013. Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proc Natl Acad Sci U S A*, 110:12355–12360. doi: 10.1073/pnas.1305813110.

Shin, W., Boo, S. M. & Triemer, R. E. 2001. Ultrastructure of the basal body complex and putative vestigial feeding apparatus in *Phacus pleuronectes* (Euglenophyceae). *Journal of Phycology*, 37:913–921. doi: 10.1046/j.1529-8817.2001.01041.x.

Simpson, A. G. B. 2003. Cytoskeletal organization, phylogenetic affinities and systematics in the contentious taxon Excavata (Eukaryota). *International Journal of Systematic and Evolutionary Microbiology*, 53:1759–1777. doi: 10.1099/ijs.0.02578-0.

Simpson, A. G. B. & Patterson, D. J. 2001. On Core Jakobids and Excavate Taxa: The Ultrastructure of Jakoba incarcerata. *J. Eukaryot. Microbiol.*, 48:480–492. doi: 10.1111/j.1550-7408.2001.tb00183.x.

Simpson, A. G.B. & Roger, A. J. 2004. Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. *Molecular Phylogenetics and Evolution*, 30:201–212. doi: 10.1016/S1055-7903(03)00177-5.

Simpson, A. G.B. 1997. The identity and composition of the Euglenozoa. *Archiv für Protistenkunde*, 148:318–328. doi: 10.1016/S0003-9365(97)80012-7.

Simpson, A. G.B. & Patterson, D. J. 1999. The ultrastructure of Carpediemonas membranifera (Eukaryota) with reference to the “excavate hypothesis”. *Eur J Protistol*, 35:353–370. doi: 10.1016/S0932-4739(99)80044-3.

Singh, K. P. 1956. Studies in the genus *Trachelomonas*. I. Description of six organisms in cultivation. *American Journal of Botany*, 43:258–266.

Strittmatter, P., Soll, J. & Bölter, B. 2010. The chloroplast protein import machinery: A review. *Methods Mol Biol*, 619:307–321. doi: 10.1007/978-1-60327-412-8_18.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. & Kumar, S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution*, 28:2731–2739. doi: 10.1093/molbev/msr121.

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*, 23:4745–4752. doi: 10.1093/nar/23.23.4745.

Thompson, M. D., Zhang, L., Hong, L. & Hallick, R. B. 1997. Extensive structural conservation exists among several homologs of two *Euglena* chloroplast group II introns. *Mol Gen Genet*, 257:45–54. doi: 10.1007/s004380050622.

Toor, N., Hausner, G. & Zimmerly, S. 2001. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. *RNA*, 7:1142–1152.

Triemer, R. & Farmer, M. 2007. A decade of euglenoid molecular phylogenetics. *In:* Brodie, J., Lewis, J. (eds.), Unravelling the algae: The past, present, and future of algae systematics. Taylor & Francis distributor, Boca Raton, FL, London. 20072976:315–330.

Triemer, R. E. & Farmer, M. A. 1991. An ultrastructural comparison of the mitotic apparatus, feeding apparatus, flagellar apparatus and cytoskeleton in euglenoids and kinetoplastids. *Protoplasma*, 164:91–104. doi: 10.1007/BF01320817.

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

Trifinopoulos, J., Nguyen, L.-T., Haeseler, A. von & Minh, B. Q. 2016. W-IQ-TREE: A fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research*, 44:W232-5. doi: 10.1093/nar/gkw256.

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*, 26:631–648. doi: 10.1093/molbev/msn285.

Turmel M., Otis C., Lemieux C. 2016. Mitochondrion-to-chloroplast DNA transfers and intragenomic proliferation of chloroplast group II introns in *Gloeotilopsis* green algae (Ulotrichales, Ulvophyceae) Genome Biology and Evolution 8:2789–2805. doi: 10.1093/gbe/evw190.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M. & Rozen, S. G. 2012. Primer3--new capabilities and interfaces. *Nucleic Acids Research*, 40:e115. doi: 10.1093/nar/gks596.

van Dooren, G. G., Schwartzbach, S. D., Osafune, T. & McFadden, G. I. 2001. Translocation of proteins across the multiple membranes of complex plastids. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1541:34–53. doi: 10.1016/S0167-4889(01)00154-9.

Vesteg, M., Vacula, R., Burey, S., Löffelhardt, W., Drahovská, H., Martin, W. & Krajčovič, J. 2009. Expression of Nucleus-Encoded Genes for Chloroplast Proteins in the Flagellate Euglena gracilis. *Journal of Eukaryotic Microbiology*, 56:159–166. doi: 10.1111/j.1550-7408.2008.00383.x.

Vogel, J., Börner, T. & Hess, W. R. 1999. Comparative analysis of splicing of the complete set of chloroplast group II introns in three higher plant mutants. *Nucl Acids Res*, 27:3866–3874.

von der Heyden, S., Chao, E. E., Vickerman, K. & Cavalier-Smith, T. 2004. Ribosomal RNA Phylogeny of Bodonid and Diplonemid Flagellates and the Evolution of Euglenozoa. *Journal of Eukaryotic Microbiology*, 51:402–416. doi: 10.1111/j.1550-7408.2004.tb00387.x.

Walker, G., Dorrell, R. G., Schlacht, A. & Dacks, J. B. 2011. Eukaryotic systematics: A user's guide for cell biologists and parasitologists. *Parasitology*, 138:1638–1663. doi: 10.1017/S0031182010001708.

Wehner, R. & Gehring, W. J. 2013. Zoologie. Georg Thieme Verlag KG, s.l. 792.

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*, 163:832–843. doi: 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*, 60:214–221. doi: 10.1111/jeu.12025.

Wolf, Y. I., Novichkov, P. S., Karev, G. P., Koonin, E. V. & Lipman, D. J. 2009. The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci U S A*, 106:7273–7280. doi: 10.1073/pnas.0901808106.

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 Evol Biol*, 12:29. doi: 10.1186/1471-2148-12-29.

Yoon, H. S., Hackett, J. D., Pinto, G. & Bhattacharya, D. 2002. The Single, Ancient Origin of Chromist Plastids. *Journal of Phycology*, 38:40. doi: 10.1046/j.1529-8817.38.s1.8.x.

Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G. & Bhattacharya, D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Molecular Biology and Evolution*, 21:809–818. doi: 10.1093/molbev/msh075.

Yubuki, N., Edgcomb, V. P., Bernhard, J. M. & Leander, B. S. 2009. Ultrastructure and molecular phylogeny of *Calkinsia aureus*: Cellular identity of a novel clade of deep-sea euglenozoans with epibiotic bacteria. *BMC Microbiol*, 9:16. doi: 10.1186/1471-2180-9-16.

Yubuki, N., Simpson, A. G. B. & Leander, B. S. 2013. Reconstruction of the feeding apparatus in *Postgaardi mariagerensis* provides evidence for character evolution within the Symbiontida (Euglenozoa). *Eur J Protistol*, 49:32–39. doi: 10.1016/j.ejop.2012.07.001.

Zimmerly, S., Hausner, G. & Wu Xc-Chu 2001. Phylogenetic relationships among group II intron ORFs. Nucleic Acids Research 5: 1238–1250. doi: 10.1093/nar/29.5.1238.

Zimmerly, S. & Semper, C. 2015. Evolution of group II introns. *Mobile DNA*, 6:322. doi: 10.1186/s13100-015-0037-5.

Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31:3406–3415. doi: 10.1093/nar/gkg595.

# 7 Appendix

**Table 7.1:** Oligonucleotides used as primers in PCR experiments, sorted by species names. Abbreviations in primer name hint to species and orientation: El = *Eutreptiella pomquetensis*; Tm = *Trachelomonas grandis*, for = forward, rev = reverse. Sequences are shown from 5' - 3'. Oligonucleotides used as primers in PCR experiments for *Euglena mutabilis* are listed in the publication (Results, Chapter I, Table S1, p. 55).

| Name | Sequence 5' - 3' |
|---|---|
| ElpomMC1 for | CTAAGCCACCTACAGACGCT |
| ElpomM C1 rev | GCCCAACACTTCCAGTTACC |
| Elpom C2 for | GTTGAAGTAATTCGCTCCTGTA |
| Elpom C3 rev | GTGTAAACAAAAATTCCTAGTGAGAT |
| ElpoBUWfor RING | CAATACATGTATCATAAATTTTTACAGAATG |
| El.pomBUWrev Ring | GCT ACC CAG CAT GTC CTG TT |
| El.pom16S in rev | TAA TCC CGT TCG CTA CCC TA |
| El/Tm 16S rev | CAG CGT TCA TCC TGA GCC A |
| El.pom 23S for | CAACAGGACATGCTGGGTAG |
| El.pom 23S rev | GATGTTTCAGTTCACCGGGT |
| El.pom 23Sfor/32 | ACCTTGGCACTTAGAGACGA |
| El.pom 16S for/32 | AGTCGTAACAAGGTAGCCGT |
| El.pom roaA for1 | ACGATATTTTGTTCGTTTAGGGC |
| El.pom roaA rev1 | TCGATTCGAGGACTACGTATTAC |
| El.pom roaA for2 | TCCACTAAAAGTACTGGGATGGT |
| El.pom roaA rev2 | GCTGGTAGACTAAATGGGGC |
| El.pom rpoA for | GCGTGGGTAATCTGTGTTGT |
| El.pom rpo rev | CGGCAGATTTTCTTCTTCCTGT |
| Tm C1 for 1 | GTAACCATGTCTCCTTAATCCAT |
| Tm C1 rev 1 | GAGCCGATTGTGGAAACTTCT |
| Tm C2 for | CGAACATCACCATTAATTAAACGAAA |
| Tm C2 rev | GCATCAACCCCAAAAGTCTC |
| Tm C3 for | TCCCTTTTGAGGAGTGTATTGT |
| Tm C3 rev | ACAGCACATACAAACGAAAGATC |
| Tm C4 for | TTTTCTCATTATTCATTGATTTTATGGC |
| Tm C4 rev 1 | TCAGATCAAATCGAACCAGCG |
| TmC4 rev 2 | TACTATAGCATCCCAAAAAGAAAGC |

| | |
|---|---|
| Tm C6 for | TGCTGTAAGATAGGATTCGTTCT |
| Tm C6 rev | GCTAGAAAAATACTAAATTGGAAATCAG |
| Tm C10 for | CTCTATTCGTTTGCGTGAAAAAC |
| Tm C10 rev | TTCATTAATACCTGCCATACTTGC |
| Tm C45 for | TAGGAGCCCAACCTAAACGA |
| Tm C45 rev | TTTCTCCTATTGTTAGGGTAAGGT |
| Tm C82 for | CCTCTATAAATCTTACGCTTAACAC |
| Tm C82 rev | CGAACAGGCGTCGCATTTAA |
| Tm C171 for | TTTCTTCGAAACTACTTGCAGCA |
| Tm C171 rev | CGAATCCCTTCTAGCCCGAT |
| Tm rpoA for | TGG AAA GAA AAT AGG AAA TCA CAA |
| Tm rpoA rev | TGA GAC GTT GGA ATT TTC AGA |
| Tm roaA for | TGG TTT CAA AAT CTT TGT ATT CGT |
| Tm roaA rev | CCA AAC ATC ATA ATA ATA ATC AAC G |
| Tm_16S for1 | GTG GCG TAC GGG TGA GTA AT |
| Tm_16S rev1 | CTA CGC ATT TCA CCG CTA CA |
| Tm_16S for2 | AGC GGT GGA ACA TGT GAT TTA |
| Tm_16S rev2 | AGC GGT GGA ACA TGT GAT TTA |
| Tm_23S for | AAA GGA GCG CGA GAT AAC AC |
| Tm_23S rev | ATT TCA CCG AGT CAC GTT CC |
| Tm rpoB for | ATT TCC ATG TCG TCC GGA TAG |
| Tm rpoB rev | ACG GAC TCG AAC GAT CTG AA |
| Tm rpoB1 for | TCCGATTAACATGCGTTCTG |
| Tm rpoB1 rev | TGGGTTCAAACATGCAAAAACAA |
| Tm rpoB2 for | CTATTTGCGTCGTTATGTTCAAGA |
| Tm rpoB2 rev | CACAGTCCTCTGCCTTACC |
| Tm rpoC1 for | TGGGACGTAAATTTGCTGGT |
| Tm rpoC1 rev | GTCGAACTAAAATTTCCAGTAATACA |
| Tm rpoC1.1 for | CCTGATTCCGTTGCTATTTTG |
| Tm rpoC1.1 rev | CAGAACGCATGTTAATCGGAAAA |
| Tm rpoC2 for | TTTCTTCAACTTTATTAAGTCCTTGTG |
| Tm rpoC2 rev | ACG ATG CGC ACA TTT CAT AC |

**Table 7.2:** Suppliers of laboratory equipment.

| Technical equipment | Company |
|---|---|
| Analytic balance CP124S | Sartorius AG, Göttingen, Germany |
| Autoclav systec VX-120 | Systec GmbH, Wettenberg, Germany |
| Centrifuge 5424 (rotor F45-24-11) | Eppendorf AG, Hamburg, Germany |
| Centrifuge 5804 (swing-bucket-rotos A-4-44) | Eppendorf AG, Hamburg, Germany |
| Centrifuge 5804 R (F45-30-11 or swing-bucket-rotor A-4-44) | Eppendorf AG, Hamburg, Germany |
| Electrophoresis power supply EV 231 | Consort n.v., Turnhout Belgium |
| Freezer HERA Ultra-low temperature | Thermo Fisher Scientific GmBH, Schwerte, Germany |
| Gel documentation UV- System | Intas Science Imaging Instruments, Göttingen, Germany |
| Magnetic stirrer MR Hei- Standard | Heidolph Instruments GmbH, Schwabach, Germany |
| Microliter pipettes | VWR International GmbH, Langenfeld, Germany |
| Microscope BA300 | Motic GmbH, Wetzlar, Germany |
| Microscope Fluorescence Lifetime Imaging | Keyence GmbH, Neu-Isenburg, Germany |
| MIDI 1 horizontal electrophoresis unit | Carl- Roth GmbH & Co. KG, Karlsruhe |
| Mini centrifuge MCF- 2360 | Laboratory & Medical Supplies Inc., Tokyo, Japan |
| Nanodrop Lite Spectrophotometer | |
| PCR Mastercycler® gradient | Eppendorf AG, Hamburg, Germany |
| PCR Mastercycler® personal | Eppendorf AG, Hamburg, Germany |
| pH meter HI 223 | Hanna Instruments GmbH, Kehl, Germany |
| Sonopuls HD 60 | Bandelin, Berlin, Germany |
| Rotator SB3 | VWR International GmbH, Langenfeld, Germany |
| Thermoblock TB2 | Analytik Jena AG, Jena, Germany |
| Thermomixer TS1 | Analytik Jena AG, Jena, Germany |
| Vortexter VV3 | VWR International GmbH, Langenfeld, Germany |

**Table 7.3:** Applied bioinformatics software tools and server.

| Name | Internet adress | Reference |
|---|---|---|
| BLAST | https://blast.ncbi.nlm.nih.gov/Blast.cgi | Altschul et al. 1990 |
| Database group II introns | http://webapps2.ucalgary.ca/~groupii/ | Dai et al. 2003 |
| EMBOSS Sixpack | http://www.ebi.ac.uk/Tools/st/emboss_sixpack/ | Rice et al. 2000 |
| Geneious 7/9 | http://www.geneious.com/ | Kearse et al. 2012 |
| GenomeVx | http://wolfe.ucd.ie/GenomeVx/ | Conant & Wolfe 2008 |
| IQ-Tree web server | http://iqtree.cibiv.univie.ac.at/ | Trifinopoulos et al. 2016 |
| IQ-Tree | http://www.iqtree.org/ | Nguyen et al. 2015 |
| MEGA 5/6/7 | http://megasoftware.net/ | Tamura et al. 2011 |
| Mfold web server | http://unafold.rna.albany.edu/?q=mfold/RNA-Folding-Form | Zuker 2003 |
| NCBI | https://www.ncbi.nlm.nih.gov/ | |
| Primer3Plus | http://primer3plus.com/cgi-bin/dev/primer3plus.cgi | Untergasser et al. 2012 |
| PseudoViewer | http://pseudoviewer.inha.ac.kr/ | Byun & Han 2006 |
| PubMed | https://www.ncbi.nlm.nih.gov/pmc/ | |
| Rfam | http://rfam.xfam.org/ | Burge et al. 2013 |
| RNAmmer 1.2 server | http://www.cbs.dtu.dk/services/RNAmmer/ | Lagesen et al. 2007 |
| RePuter | https://bibiserv2.cebitec.uni-bielefeld.de/reputer | Kurtz et al. 2001 |
| Tandem Repeats Finder | https://tandem.bu.edu/trf/trf.html | Benson 1999 |
| tRNAscan-SE | http://lowelab.ucsc.edu/tRNAscan-SE/ | Schattner et al. 2005 |

**Table 7.4:** Authority of species.

| Name/Taxon | Authority |
|---|---|
| *Chlorokybus atmophyticus* | Geitler 1942 |
| *Colacium vesiculosum* | Ehrenberg 1834 |
| *Cryptoglena skujae* | Marin and Melkonian 2003 |
| *Cymbomonas tetramitiformis* | Schiller 1913 |
| *Distigma proteus* | Ehrenberg 1831 |
| *Euglena archaeoplastidiata* | Chadefaud 1937 |
| *Euglena cantabrica* | Pringsheim 1956 |
| *Euglenaformis proxima* | (Dangeard) Bennett and Triemer 2014 |
| *Euglena gracilis* | Klebs 1883 |
| *Euglena gracilis var. bacillaris* | Klebs 1883 |
| *Euglena longa* | (Pringsheim) Marin and Melkonian 2003 |
| *Euglena mutabilis* | Schmitz 1884 |
| *Euglenaria anabaena* | (Mainx) Karnkowska and Linton 2010 |
| *Euglena velata* | Klebs 1883 |
| *Euglena viridis* | (Müller) Ehrenberg1830 |
| *Euglena viridis* | (Müller) Ehrenberg1830 |
| *Eutreptia viridis* | Perty 1852 |
| *Eutreptiella gymnastica* | Throndsen 1969 |
| *Eutreptiella pomquetensis* | (McLachlan, Seguel and Fritz) Marin and Melkonian 2003 |
| *Euglena quartana* | Moroff 1903 |
| *Lepocinclis buetschlii* | Lemmermann 1901 |
| *Mesostigma viride* | Lauterborn 1894 |
| *Monomastix sp.* | Scherffel 1912 |
| *Monomorphina aenigmatica* | (Drezepolski) Nudelman and Triemer 2006 |
| *Monomorphina parapyrum* | Kim, Triemer and Shin 2013 |
| *Nephroselmis astigmatica* | Inouye and Pienaar 1984 |
| *Nephroselmis olivacea* | Stein 1878 |
| *Ostreococcus tauri* | Courties and Chrétiennot-Dinet 1995 |
| *Palmophyllum crassum* | (Naccari) Rabenhorst 1868 |
| *Paulinella chromatophora* | Lauterborn 1895 |
| *Petalomonas cantuscygni* | Cann and Pennick 1986 |

| | |
|---|---|
| *Phacus orbicularis* | Hübner 1886 |
| *Phacus similis* | Christen 1962 |
| *Picocystis salinarum* | Lewin 2001 |
| *Prasinococcus sp.* | Miyashita and Chihara 1993 |
| *Prasinoderma coloniale* | Hasegawa and Chihara 1996 |
| *Prasinophyceae sp.*CCMP1205 | Christensen ex. Silva 1980 |
| *Prasinophyceae sp.* MBIC10622 | Christensen ex. Silva 1980 |
| *Pycnococcus provasolii* | Guillard 1991 |
| *Pyramimonas gelidicola* | McFadden, Moestrup and Wetherbee 1982 |
| *Pyramimonas parkeae* | Norris Pearson 1975 |
| *Scherffelia dubia* | (Perty) Pascher 1912 |
| *Strombomonas acuminata* | (Schmarda) Deflandre 1930 |
| *Tetraselmis sp.* | Stein 1878 |
| *Trachelomonas grandis* | Singh 1956 |
| *Trachelomonas volvocina* | Ehrenberg 1834 |
| *Verdigellas peltata* | Ballantine and Norris 1994 |

## List of Abbreviations

| | |
|---|---|
| A | Adenine |
| atp | ATPase |
| BLAST | Basic Local Alignment Search Tool |
| bp | Base pair |
| bs | bootstrap |
| C | Cytosine |
| °C | Degree Celsius |
| CCMP | Culture Collection of Marine Phytoplankton |
| cDNA | Complemetary DNA |
| CDS | DNA Coding sequence or region |
| chlI | chlorophyll biosynthesis |
| cp | chloroplast |
| di$H_2O$ | Purified (deionized) water |
| dNTPs | deoxynucleotide triphosphates |
| ds$H_2O$ | Highly purified (deionized and sterilized) water |
| DNA | Deoxyribonucleic acid |
| EDTA | Ethylenediamine-tetraaceticacid |
| EGT | Endosymbiotic gene transfer |
| emend. | emended |
| ER | Endoplasmatic reticulum |
| Fig. | Figure |
| g | Gram |
| G | Guanine |
| IGS | Intergenic space |
| l | Litre |
| LB | Lysogeny broth |
| LSU | Large subunit |
| m-2 s-2 | per square meter per second |
| mat | maturase |
| min | Minute |
| ML | Maximum likelihood |
| µ | Micro |

| | |
|---|---|
| m | mili |
| nt | Nucleotide |
| ORF | Open Reading Frame |
| PCR | Polymerase chain reaction |
| pH | negative decadic logarithm of H+ concentration |
| psa | Photosystem I |
| psb | Photosystem II |
| rDNA | Ribosomal DNA |
| rev | reverse |
| RNA | Ribonucleic acid |
| RNase | Ribonuclease |
| rpl | Ribosomal protein L |
| rpm | Rounds per minute |
| rps | Ribosomal protein S |
| rpo | RNA polymerase |
| rRNA | Ribosomal RNA |
| RT-PCR | reverse transcriptase PCR |
| S | Svedberg unit (sedimentation coefficient) |
| SAG | Sammlung von Algenkulturen Göttingen |
| sec | second |
| SOC | Super optimal broth with catabolite repression |
| SSU | Small subunit |
| T | Thymine |
| TAE | Tris-acetate-EDTA |
| Taq | Thermus aquaticus DNA polymerase |
| Temp | temperature |
| TIC | translocon inner membran complex |
| TOC | translocon outer membran complex |
| Tris | Tris-(hydroxymethyl)-aminomethane |
| tRNA | Transfer RNA |
| tufA | translation elongation factor EF-Tu |
| U | Unit |
| UV | Ultraviolet light |
| V | Volt |

| | |
|---|---|
| v/v | Volume per volume |
| VNTR | Variable Number Tandem Repeat |
| w/v | Weight per volume |
| x g | times gravity |
| ycf | hypothetical protein |

## Acknowledgements

Die vorliegende Arbeit wurde durch ein Grundstipendium der Graduiertenförderung der Bergischen Universität Wuppertal gefördert.

Zuerst und ganz besonders möchte ich mich bei Frau Prof´in Dr. Angelika Preisfeld bedanken, dass sie mich damals in ihre Arbeitsgruppe aufgenommen hat und mir die Möglichkeit gegeben hat diese Arbeit zu schreiben und diese faszinierenden kleinen Lebewesen kennenzulernen. Sie hat mich immer in meiner Arbeit unterstützt, mir zur Seite gestanden und mich doch selbstständig arbeiten lassen. Danke!

Für die Übernahme des Gutachtens danke ich Frau Prof´in Dr. Heike Wägele vom ZFMK.

Bedanken möchte ich mich auch bei allen ehemaligen und jetzigen Mitgliedern der Arbeitsgruppe „Zoologie und Biologiedidaktik". Ich danke allen voran Melanie Beudels, Cora Berger und Margret Buse für das Lesen der Arbeit und die hilfreichen Anmerkungen; ihr seid toll! Sabine Stratmann-Lettner gilt ebenfalls ein großer Dank. Sie hat mich damals im Labor unterstützt, eingearbeitet und mir alles Wichtige gezeigt. Sabine, ich danke dir für die vielen Diskussionen und dein offenes Ohr. Sabrina Bleidißel hat mich nicht nur mit Cappuccino versorgt, sondern auch meine Stammbaumanalyse kritisch beleuchtet; danke! Außerdem schaffst nur du es einem bewusst zu machen, dass nichts so schlimm ist, wie es einem vielleicht scheint. Karsten Damerau danke ich für die Tagungsbegleitung nach Sevilla, denn nur dadurch ist das zweite Paper so wie es jetzt ist. Außerdem ist ein Galadinner zu zweit immer schöner als alleine. Marisa Bartling danke ich für die unvergessliche Zeit unserer Zusammenarbeit und die daraus entstandene Freundschaft.

Ich danke Herrn Dr. Alexander Donath vom ZFMK für die hilfreichen und verständlichen Tipps per E-Mail bei den phylogenomischen Analysen.

Ich möchte mich weiterhin bei allen bedanken, die mir diese Arbeit ermöglicht haben: Ein ganz besonderer Dank geht an meine Eltern. Ihr habt mir immer alles ermöglicht und mich auch während der Anfertigung dieser Arbeit stets unterstützt. Danke, dass ihr immer nur das Beste für uns wollt und alles Erdenkliche dafür tut. Ein ganz großer Dank geht an Lina, Ayad und Rischdi, die immer an mich geglaubt haben und sich jeden Erfolg, aber auch Misserfolg, angehört haben!

Ein besonderer Dank gilt meinen Freunden; danke für die Geduld. Larissa Bartsch, Sina Mittmann und Denise Laskowski danke ich besonders für die Freundschaft und die Unterstützung zu jeder Zeit.

Danken möchte ich auch dem Realtime Department für 16 GB Ram; ein wirklich hilfreiches Geschenk.

Pepe mein treuer Freund, du hast wirklich jeden Moment dieser Arbeit mit mir verbracht. Ich kann mir keinen besseren Wegbegleiter vorstellen.

Torsten Hauck, ich danke dir von Herzen für alles. Du warst immer da, hast mir den Rücken freigehalten, mich motiviert und unermüdlich unterstützt. Du hast es verstanden wie kein anderer.

# Erklärung

Hiermit erkläre ich, dass ich

1. die von mir eingereichte Dissertation selbständig und ohne fremde Hilfe verfasst habe,
2. nur die in der Dissertation angegebenen Hilfsmittel benutzt und alle wörtlich oder inhaltlich übernommenen Stellen als solche unter Angabe der Quelle gekennzeichnet habe,
3. die Dissertation weder in der vorliegenden noch in ähnlicher Form bei anderen Hochschulen oder wissenschaftichen Instituten vorgelegt habe und
4. bislang keine Promotionsversuche unternommen habe.

Ich bin damit einverstanden, dass meine Dissertation wissenschaftlich interessierten Personen oder Institutionen zur Einsichtnahme zur Verfügung gestellt werden kann.

Wuppertal,

______________________________

Nadja Dabbagh