Univerzita Karlova v Praze
Přírodovědecká fakulta

Doktorský studijní program: Parazitologie
Ph.D. study program: Parasitology

Mgr. Jana Szabová

Komplikovaná evoluce methionin adenosyltranferázy u eukaryot se zvláštním zaměřením na euglenidy

The complicated evolution of methionine adenosyltransferase in euglenids and eukaryotes in general

Dizertační práce / Ph.D. Thesis
Thesis supervisor: Mgr. Vladimír Hampl, Ph.D.

Praha 2014

Declaration of the author / Prohlášení autorky:

I declare that the data presented in this PhD thesis are my own work and resulted from team collaboration during projects with our partners. I also proclaim that the literary sources were cited properly and neither this work nor the particular data have been used to reach any other academic degree.

Prohlašuji, že tuto dizertační práci jsem vypracovala samostatně, že data prezentovaná v ní jsou výsledky mé vlastní práce a vznikly pomocí týmové spolupráce s partnery projektu. Taky prohlašuji, že jsem uvedla všechny použité literární zdroje a literaturu a že ani samotná práce, ani její podstatná část nebyly použity k získání jiného nebo stejného akademického titulu.

Mgr. Jana Szabová

Acknowledgements

Many thanks to my supervisor Vladimír Hampl and to all fellows for the support and help during all those years. Many thanks also to my family for their support and trust in me, because without them I wouldn´t be able to finish the studies.

Table of Contents:

Abstract

Many eukaryotic genes do not follow vertical inheritance pattern. In the present work, we have chosen as a model the gene for methionine adenosyltransferase (MAT), in which we have decided to examine in detail the evolutionary history. MAT is a ubiquitous essential enzyme that, in eukaryotes, occurs in two relatively divergent paralogs: MAT and MATX. Both paralogs have punctate distributions across the tree of eukaryotes and, except for a few cases, they are mutually exclusive. This points to the complicated evolutionary history of this gene couple, which may be caused by either differential loss of old paralogs or the spread of one of these paralogs by horizontal gene transfer (HGT). We have focused on the evolution of this enzyme particularly within one of the best-known groups of flagellates, the euglenids, because it was hypothesized that MATX evolved in photosynthetic euglenids before it spread to other lineages.

We gained 26 new sequences from 23 euglenid lineages and one prasinophyte alga *Pyramimonas parkeae*. MATX was found only in photoautotrophic euglenids. Both, mixotroph *Rapaza viridis* and the prasinophyte alga *Pyramimonas parkeae*, the closest known relative of the euglenid plastid ancestor, only displayed the MAT paralog. In contrast, both paralogues were found in two euglenid species (*Monomorphina pyrum* and *Phacus orbicularis*). However, these two MAT genes were not related to any ancestral-euglenid MATs. The distribution of MAT/MATX in euglenids can be explained by three events: a single HGT of MATX that happened after the origin of euglenid secondary plastid, and two HGTs of MAT into two photoautotrophic species.

The plausibility of processes putatively involved in MAT and MATX evolution was investigated using the *Trypanosoma brucei*/*Euglena gracilis in vivo* experimental model. This confirmed that MATX is able of both, a long-term coexistence with its MAT counterpart, and immediate replacement of MAT function. The conflict between species phylogeny and phylogeny inferred from MATX sequences suggests that MATX paralog has undergone HGT across the eukaryotic tree. Since phylogenetic analyzes do not exclude the presence of MATX in a common ancestor of eukaryotes, we assume that MATX originated in very ancient gene duplication, possibly in a common ancestor of all extant eukaryotes. This duplication was followed by more or less long period of coexistence of both paralogs until individual eukaryotic lines lost one of them. In addition, both paralogs have undergone HGTs. During one of the HGTs MATX was introduces into the lineage of photosynthetic euglenids, where it replaced the original MAT. The initial idea that euglenids was the group in which MATX originated proved to be very unlikely.

Abstrakt

Velká část eukaryotických genů se v evoluci nepřenášela výhradně vertikálně z rodičů na potomstvo. V této disertační práci jsme si vybrali jeden z takových genů, a to gen pro methionin adenosyltransferázu (MAT), a pokusili se podrobně zmapovat jeho evoluci. MAT je všudypřítomný esenciální enzym, který se u eukaryot nachází ve formě dvou paralogů: MAT a MATX. Oba paralogy jsou mezi eukaryoty nerovnoměrně rozšířeny a s výjimkou několika málo případů se u daného organismu vyskytuje jen jeden z nich. To ukazuje na komplikovanou evoluční historii tohoto genu, která může zahrnovat takové evoluční procesy jako genové duplikace a následné ztráty nebo horizontální genový přenos (HGT). My jsme se zaměřili zejména na výskyt obou forem tohoto genu u jedné z nejznámějších skupin bičíkovců, skupiny Euglenida. Předpokládalo se totiž, že by tato skupina mohla být kolébkou paralogu MATX, ze které se tento gen následně šířil do dalších eukaryotických linií.

Podařilo se nám získat 26 nových sekvencí z 23 linií euglenidů a jedné prasinofytní řasy *Pyramimonas parkeae*, která představuje nejbližšího známého příbuzného euglenidího plastidu. MATX byl zjištěn pouze u fotoautotrofních euglenidů, přičemž mixotrof *Rapaza viridis* a *P. parkeae* vykazovali přítomnost pouze paralogu MAT. Oba typy paralogů byly nalezeny u dvou druhů euglenidů – *Monomorphina pyrum* a *Phacus orbicularis*. MAT geny u těchto druhů ovšem nejsou příbuzné MAT genům heterotrofních euglenidů. Distribuci MAT/MATX u euglenidů lze vysvětlit pomocí tří HGT událostí: jednoho horizontálního přenosu MATX genu, který se odehrál až v období po vzniku sekundárního euglenidího plastidu, a dalších dvou horizontálních přenosů MAT genů do dvou fotoautotrofních euglenidů.

Uskutečnitelnost procesů potenciálně zapojených do evoluce MAT a MATX paralogů (HGT, dlouhodobá koexistence dvou paralogů po genové duplikaci) jsme zkoumali pomocí *in vivo* experimentálního modelu *Trypanosoma brucei*/*Euglena gracilis*. Ten potvrdil, že MATX je schopen dlouhodobé koexistence se svým paralogem MAT a zároveň je také schopný MAT funkčně nahradit. Statisticky prokazatelný konflikt mezi fylogenezí eukaryot a fylogenezí MATX genu naznačuje, že v evoluční minulosti MATX došlo k HGT. Jelikož fylogenetické analýzy nevylučují přítomnost MATX u společného předka eukaryot, předpokládáme, že MATX vzniknul velmi dávnou genovou duplikací, možná u společného předka všech dnešních eukaryot. Po této duplikaci následovalo více či méně dlouhé období koexistence obou paralogů, dokud nedošlo v jednotlivých eukaryotických liniích ke ztrátě jednoho z nich. Oba paralogy navíc prodělaly HGT. Jeden HGT vnesl paralog MATX do linie fotosyntetických euglenidů, kde tento nahradil původní MAT. Prvotní představa, že skupina Euglenida byla kolébkou MATX, se ukázala jako velmi nepravděpodobná.

Overview

1. Introduction

In the present work, we have focused on the evolution of methionine adenosyltransferase (MAT, also known as S-adenosylmethionine synthetase or AdoMet-synthetase) (EC 2.5.1.6). Our major interest in this enzyme lies in its complex evolutionary history that stems from the fact that in nature it occurs in two divergent paralogs that exhibit punctuate distribution among eukaryotes. This enzyme was chosen as a case study, although other genes demonstrated a similarly complicated evolutionary history and apparently not follow a simple vertical inheritance pattern. Our goal was to elucidate the relationships and distribution of these paralogs in eukaryotes and specifically among euglenids, because it was hypothesized that one paralog of this enzyme could have arisen during the secondary plastid endosymbiosis in this group.

MAT is an essential, ubiquitous enzyme that catalyzes a two-step reaction that leads to the formation of one of the most important cellular metabolites, S-adenosyl-L-methionine (also known as SAM or AdoMet, Fig. 1), which plays a major role in methylation reactions in all organisms by acting as a direct methyl group donor. Methionine is a non-polar amino acid characterized by the presence of a methyl group attached to a sulfur atom located in its side chain. To be metabolically active, methionine must be converted into SAM (by MAT, Fig. 2). With the impairment of the MAT enzyme, supplementation of organism with methionine becomes useless or even toxic as it leads to the accumulation of unused, non-activated methionine (Lieber and Packer 2002; Cantoni 1951). It follows from this the paramount importance of MAT in all living cells, as it seems to be one of those essential enzymes required for life.

MAT is involved in the specific reaction leading to SAM synthesis (scheme below), where the adenosyl group from an adenosyl triphosphate molecule (ATP) is transferred to the L-methionine. Simultaneously, the ATP is hydrolyzed to pyrophosphate (PPi) and orthophosphate (Pi), two byproducts that are released (Chiang et al. 1996; Cantoni 1951; Mudd and Cantoni 1958; Taylor and Markham 2000; Schlesier et al. 2013; Cantoni 1975). The P_i group involved in the SAM formation comes predominantly from the γ-phosphoryl group of the ATP. This indicates that motion of PPPi is restricted within the active site (Lu and Markham 2002).

L-methionine + ATP + H_2O ⟶ (SAM + PPP_i) + H_2O ⟶ SAM + PP_i + P_i

The reaction catalyzed by MAT is the only known way of SAM biosynthesis. In addition, this enzyme requires the presence of divalent cations such as Mg^{2+} as well as monovalent cation K^+ for its activation (Tabor and Tabor 1984; Kotb and Geller 1993; Mato et al. 1997; Garrido et al. 2011).

The importance of SAM lies in the fact that it is the main methyl-group biological donor to: phospholipids (thus keeping the fluidity of membranes), DNA (crucial for regulation of gene expression), RNA, hormones and neurotransmitters. It is involved also in the biosynthesis of polyamines such as spermidine and spermine, in signal-transduction system and thus, it represents an important regulatory factor in many biological processes (Fontecave et al. 2004).

Besides, SAM also contributes to other important biological reactions as it is able to donate any of the groups surrounding the sulfur atom: i) the aminopropyl group, which together with SAM-decarboxylase creates putrescine and afterwards spermidine, ii) the methyl group and also, iii) the 5´-deoxyadenosyl radicals, which are used by SAM radical proteins to produce biotine. Such is its "molecular promiscuity" that it has been estimated that SAM participates in as many reactions as ATP does (Pajares and Markham 2011; Fontecave et al. 2004; Chiang et al. 1996).

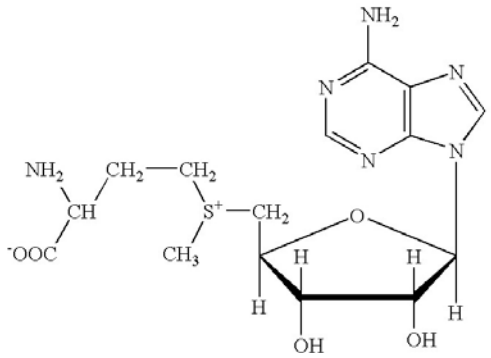

Figure 1: S-adenosyl-L-methionine

The common product of methylation reactions is S-adenosylhomocysteine (AdoHcy), which also serves as a regulator/inhibitor of the methionine adenosyltransferase at high concentration rates. The increasing amount of AdoHcy

simultaneously decreases the amount of SAM, being known this AdoMet/AdoHcy ratio as the methylation index.

Due to the complexity of the MAT reaction Cantoni (1975) assumed that this enzyme could actually be a multiprotein complex consisting of several associated polypeptide chains. This suspicion was firstly confirmed in work by Chiang and Cantoni (1977), where the authors showed that, in yeast, the MAT enzyme is made up by two subunits with different molecular weight. They also found that MAT appears in yeast in two isozymes.

From that time, MAT was found in all organisms studied to this day, except for some parasites like *Pneumocystis*, which acquire SAM from its host. In all organisms, MAT preferentially adopts an oligomeric structure, mostly tetramer made up of four identical subunits, which are encoded by a single gene. These four subunits form two tight dimers whose active sites are located between their subunits. Combining of these dimers leads to the final homotetramer enzymatic structure. In mammals, one of the isozymes adopts an heterotetrameric layout and consists of subunits encoded by two different genes (Takusagawa et al. 1996; LeGros et al. 2000).

The different conformational structures (from primary to tertiary) of this cytosolic enzyme (Lu and Markham 2002; Garrido et al. 2011) have been extensively studied. It was revealed that its primary structure (approx. 400 amino acids) and tertiary structure of monomer is highly conserved among organisms.

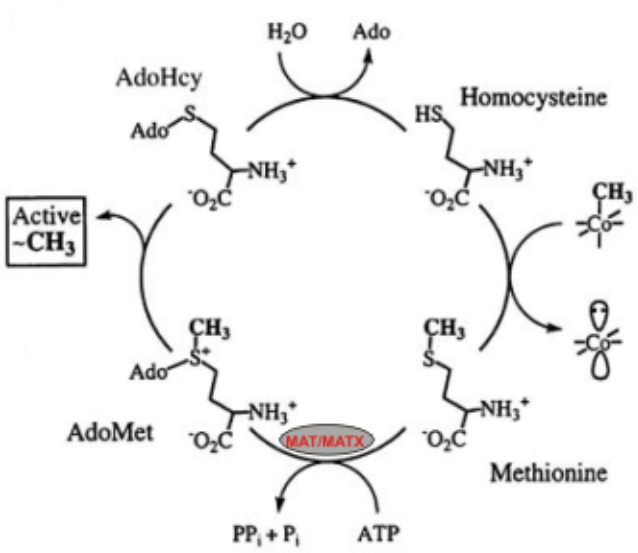

Figure 2: Schematic illustration of the methionine cycle, MAT is highlighted in red color. Adopted from Dixon et al. (1996).

2. MAT in archaea

Despite they preserve some essential MAT-family similarities, sequences of archaeal MATs are highly divergent and distinct from those found in bacteria and eukaryotes.

The archaeal MATs share with bacterial and eukaryotic MATs only about 20% sequence identity whereas bacterial and eukaryotic MATs are about 60% identical (Sánchez-Pérez et al. 2004). Although archeal MATs appear in solutions as dimers, the recently described MAT from hyperthermophile *Thermococcus kodakarensis* was found to form a tetramer in crystal packing (Schlesier et al. 2013), showing high level of thermostability. There are some indications that the sequence divergence observed between archaeal MAT and the rest of MATs could be due to the adaptation of these organisms to high temperature conditions. Increased stability could be caused by the presence of extended β-sheets in the core domain of the protein, as well as the establishment of tighter binding forces between both subunits within the dimer (Garrido et al. 2009; Schlesier et al. 2013).

The archaeal type of MAT was subsequently found in the sequences of all the completed archaeal genomes, as well as in three bacterial genomes (*Aquifex aeolicus*, *Chlorobium tepidum* and *Streptococcus pyogenes*). In contrast, this form of MAT was not found in any eukaryotic genome. Regarding to the presence in three bacterial species encoding archael type of MAT, this is probably the result of HGT events through which this type of MAT was inserted into these three bacterial genomes. Inasmuch as all of them share a recent common ancestor, it only implies the need for a single transfer from archaea. On the other hand, all of these bacteria also contain the original bacterial MAT type. Despite this fact has not been fully addressed, these circumstances point to the possibility that the acquired xenologous MAT is responsible for the regulation of their ancestral MAT (Graham et al. 2000). Furthermore, these examples clearly demonstrate that MAT's evolutionary history has involved lateral gene transfer events.

3. MAT in bacteria

Most of the MAT enzymes from bacteria appear as single copy genes and commonly are labeled as *metK* gene. *Escherichia coli* MAT is the best-studied MAT enzyme among bacteria (Yocum et al. 1996). In the high-quality finished genome of *E. coli* it was found only a single *metK* gene (Markham et al. 1984; Newman et al. 1998). The essentiality of *metK* enzyme was proved in an experiment where an *E. coli* strain carrying a deletion for its *metK* gene was not able to grow in the absence of an episomal plasmid expressing this gene.

An earlier work by Newman et al. (1998) described the influence of MAT activity on *E. coli* cell division. They showed that, when MAT activity was lower than certain critical threshold, cells were forming filaments more than 50 times longer than normal cells and without cross walls. By expressing a plasmid coding for *metK*, the wild type phenotype of cells was restored. From this pattern it was suggested that MAT plays a role in cell division and septation and the lack of MAT probably leads to cell division defect. Many studies described also the role of both MAT and SAM in the regulation of secondary metabolism and morphological differentiation in both eukaryotes and prokaryotes. There are several reports pointing out the role of intracellular SAM in the regulation of antibiotic production, sporulation and cellular differentiation (Kim et al. 2003; Ochi and Freese 1982; Huh et al. 2004; Okamoto et al. 2003; Park et al. 2005). Over-expression of MAT (and thus the increase of SAM) or addition of SAM to the culture medium leads to overproduction of antibiotics in various *Streptomyces* species, but at the same time it results in a decrease of sporulation and differentiation. Despite similar results were found for other organisms like *Bacillus subtilis* (Ochi and Freese 1982), the mechanism by which SAM regulates antibiotic production remains still unknown.

4. MAT in eukaryotes

Most of the previous studies on eukaryotic MATs have been restricted on land plants and opisthokonts. Regarding opisthokonts, these studies have been especially focused on both human and rat MATs, where three MAT isozymes (MAT I, MAT II

and MAT III) were found. While two of them (MAT I and III) are specifically expressed in the liver, the third one (MAT II) is ubiquitous being present in all tissues. Liver MAT enzymes (I, III) consist of the same α1 subunit but have different structures. MAT I appears as a homotetramer $(\alpha1)_4$, while MAT III is present as a homodimer $(\alpha1)_2$. MAT II isoform is a heterotetramer consisting of the two α2 subunits and two β subunits. Thus, mammalian MATs are encoded by three genes: *MAT1A* gene codes for α1, *MAT2A* codes for α2 and *MAT2B* codes for β subunit (Kotb and Kredich 1985; Mato et al. 1997; Halim et al. 2001; Nordgren and Peng 2011).

It was described that SAM strongly inhibits MAT II, but at the same time minimally inhibits MAT I and stimulates MAT III. However, LeGros et al. (1997) showed that MAT II regulation could vary in function of the differential oligomerization of its α2 and β subunits. They found out that the β subunits are non-catalytic (no MAT activity was detected) and probably have a regulatory function, as it seems that without these β subunits MAT II shows, simultaneously, higher activity as well as lower sensitivity to the inhibition mediated by SAM.

Lu et al. (2003) found that *MAT1A* gene is not only expressed in liver but also in rat and mouse pancreas. Therefore, MATI and III are not liver-specific enzymes, as it was previously thought. Probably, these genes can be expressed in all tissues, although the activity of these enzymes may vary multiple times among different tissues, being maximal in the liver.

In other opisthokonts like yeasts, MAT occurs in two different isozymes, which were named AdoMet synthetase I and II, respectively (Cherest et al. 1978). The study on yeasts conducted by Rouillon et al. (1999) provided the first evidence showing that SAM is essential for these organisms, and also revealed the presence of a SAM transporter in yeasts by which they can acquire this metabolite from the medium (in contrast to the bacteria *E. coli* mentioned above).

MAT genes have been studied and isolated from a variety of plants. These MATs were found usually as gene families with different number of types varying from 2 to 4. MAT genes were isolated for example from *Arabidopsis thaliana* (LeGros 1997, web Arabidopsis.org), *Pisum sativum* (pea) (Gómez-Gómez and Carrasco 1998), *Petroselinum crispum* (parsley) (Kawalleck et al. 1992), *Oryza sativa* (rice) (Lee et al. 1997), *Actinidia chinensis* (kiwifruit) (Whittaker et al. 1995), *Catharanthus roseus*

(Madagascar periwinkle) (Schröder et al. 1997) and *Solanum lycopersicum* (tomato) (Espartero et al. 1994). In addition, it seems that the expression of these distinct MAT genes differs through different plant tissues (Peleman et al. 1989) and that it is also developmentally regulated (Gómez-Gómez and Carrasco 1998).

In general, eukaryotic MAT is a highly conserved enzyme, which exhibits close relatedness across all eukaryotes. This explains the effort to use MAT as a phylogenetic marker. According to Sánchez-Pérez et al. (2004), MAT phylogeny was able to solve relationships between close and distant relatives with high bootstrap support.

Nowadays, thanks to whole genome and transcriptome sequencing experiments, sequences of MAT genes from almost all higher eukaryotic groups are available in open access databases. Before the sequencing 'boom' of the 1980s and 1990s, it was thought that MAT occurs only in two different versions that have almost strict kingdom-specific distribution: one form in Bacteria and Eukaryota and the other one in Archaea (Graham et al. 2000). Currently, it becomes clear that in eukaryotes exists another form of MAT enzyme, which occurs in few remotely related lineages of organisms. This form, named as MATX (Sánchez-Pérez et al. 2008), is apparently derived from eukaryotic MAT, from which it can be easily distinguished on phylogenetic tree by a long stem.

5. MATX

MATX differs from MAT in four unique insertions and many unique substitutions. As a result of these insertions the length of the enzyme increased to 471 residues (in *Euglena gracilis*) in comparison to the about 400 residues present in most MATs (Garrido et al. 2011; Sánchez-Pérez et al. 2008). The insertions are located within the surface loops, not in the internal parts of the protein, nor in the interface where the active sites are located (Fig. 3). This leaves the rest of the protein without significant structural changes and probably without significant effect on the enzymatic activity (Sánchez-Pérez et al. 2008). Functional equivalency of MAT/MATX was confirmed in biochemical assay by Garrido et al. (2011).

MATX was found so far in four unrelated groups of photosynthetic eukaryotes: haptophytes, diatoms, photosynthetic euglenids and dinoflagellates, but also in the data sets of a few isolated representatives of other groups. These involve the pelagophyte

alga *Aureococcus anophagefferens*, which harbors two copies of MAT in addition to the MATX; the prickly lettuce *Lactuca serriolla* and the southern pine beetle *Dendroctonus frontalis*. *L. serriolla* and *D. frontalis* seem to be result of a contamination as their MATXs are the only known occurrence of this enzyme in plants and animals, respectively. In most cases, MAT and MATX are mutually exclusive in their distribution and do not co-occur in the same organism. The hypothesis postulated by Sánchez-Pérez et al. (2008) explains this fact by the formation of unstable heterodimers or heterotetramers in cells able to express both enzymes. This could result in cell stress and damage. However, few organisms expressing both paralogs have been identified. Among these, belong five diatom species, *A. anophagefferens* and two species of photosynthetic euglenids (Kamikawa et al. 2009; Sánchez-Pérez et al. 2008; Szabová et al. 2014). The patchy distribution of MATX across the eukaryotic tree suggests a complicated evolutionary history for these two paralogous genes, which could involve events like gene duplications and subsequent gene losses (deep paralogy scenario) and/or horizontal gene transfers (horizontal gene transfer scenario).

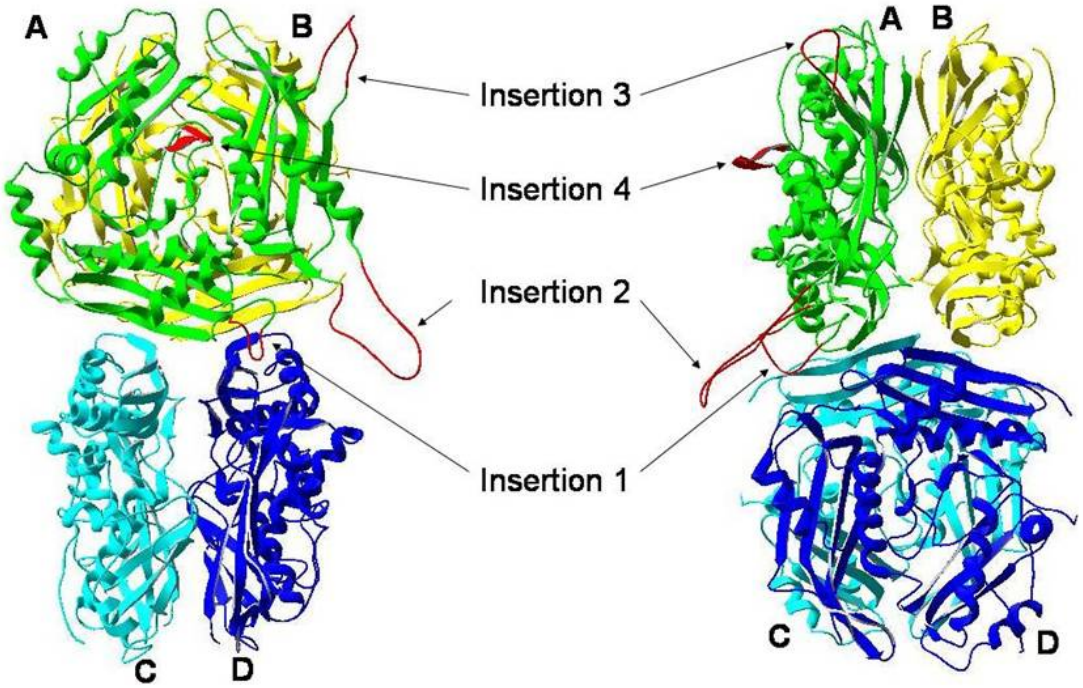

Figure 3: Schematic 3D model showing two different views of the tertiary structure of the MATX enzyme from *Euglena gracilis*. It is composed from four monomers (A, B, C and D) and the MATX insertions are marked by the arrows (Sánchez-Pérez et al. 2008).

6. Horizontal gene transfer

Horizontal gene transfer (HGT or also known as a lateral gene transfer, LGT) refers to transmission of genetic information between individuals belonging to the same, related or even unrelated species, in a manner that is not connected to reproduction. This is in contrast to vertical gene transfer, where the genetic information is transferred from parents to their offspring (transfer through generations).

Nowadays interpretation of the phylogenetic incongruities on gene trees as results of horizontal gene transfer is becoming very popular. Multiple forms of HGT have been described, like the intra-domain HGT (e.g. between prokaryotes) or the inter-domain HGT (e.g. from prokaryotes to eukaryotes). HGT events are not always easy-to-follow since sometimes due to the complexity of this processes, the direction of the transfer is very difficult to infer (Andersson 2009; Almeida et al. 2008).

HGT events have been frequently described for prokaryotes (Koonin et al., 2001), but less cases have been reported in eukaryotes. This is caused partly by to the big amount of prokaryotic data available (at both genomic and sequence levels) and partly historically because of the studies regarding the acquisition of antibiotic resistance (Huh et al. 2004; Kim et al. 2003; Dzidic and Bedeković 2003; Juhas et al. 2009). However, it mainly reflects general strategy of bacterial populations for adaptation to changing environment (Koonin and Wolf 2008). In addition to the frequent prokaryote-to-prokaryote gene transfer, there is also evidence for prokaryote-to-eukaryote gene transfers. The amount of eukaryote-to-eukaryote HGTs has gone underestimated for a long time, but this point of view is rapidly changing with the increasing sequencing data available (Andersson 2005; Huang 2013).

For a successful transfer to take place, the xenologous (foreign) gene must enter the recipient cell, integrate into the resident genome and, finally, be transmitted to the offspring (Huang 2013). This last step could be very difficult, especially in multicellular organisms like vertebrates, where the xenologous DNA must reach the germ cells. For unicellular eukaryotic organisms, HGT is easier, at least in principle (Andersson et al., 2001; Huang 2013).

Two types of HGT can be differentiated in eukaryotes depending on the source from which the genetic information is obtained. If the origin of the transferred genes is a genome, which is or was present inside the recipient cell, this process is referred as an

endosymbiontic gene transfer (EGT). Potential sources of EGT include organelles with an endosymbiotic origin, like mitochondrion, plastids or endosymbionts, which are present in the cell (Doolittle 1998; Andersson 2005; Palenik 2002; Martin et al. 2002; Karlberg et al. 2000). The second kind of gene transfer does not involve any kind of endosymbiotic associations.

After acquisition of the transferred gene the native homolog loses its essentiality and may be displaced from the genome by the foreign gene, which is taking over its function completely. In another case, the new gene may bring an advantageous function, maybe new for the organism, so the loss of any native gene is not necessary in principle. To summarize the whole process, any transferred gene must be able to: i) enter the recipient cell, which is easier in unicellular eukaryotes than in multicellular ones, ii) become expressed, and iii) completely substitute the function of the native gene or to bring a new one soon after the gene transfer.

It is often complicated to prove a case of horizontal gene transfer. The most reliable indication is the incongruence between the phylogeny of the studied gene and the organismal phylogeny. Such incongruence means that the gene probably has different evolutionary past than the rest of the genome and one possible explanation of this fact is HGT. For the successful application of the phylogenetic congruency tests, few conditions must be fulfilled. Firstly, the studied genes must carry enough phylogenetic information, secondly, the compared homologs must be in orthologous and not in paralogous relationship, finally the substitution rate of the compared genes should not be radically different (Syvanen 1994). The increased rate of substitution or contamination may cause artifacts in phylogenetic reconstruction and, consequently, misleading results. Other options to detect HGT are the calculation of GC content or the codon and amino acids usage. In this case, the foreign gene is expected to have different values of these parameters from the rest of the genome (Lawrence and Ochman 1997).

The existence of HGT often complicates uncovering the phylogenetic history of organism. In certain special cases, however, the determination of an ancient HGT could be beneficial for the reconstruction of organismal phylogeny. The presence of a gene acquired by HGT could serve as a valuable phylogenetic marker of phylogenetic relationships between lineages. In these cases, organisms that share a derived character (unique insertion, deletion, gene fusion, other genetic rearrangements or a HGT from

the same source) are expected to form a monophyletic group (Williams et al. 2010). However, the more ancient is the transfer, the harder it is to detect, because traces of its relationship to donor-lineage deteriorate with age (Huang and Gogarten 2009).

7. Deep paralogy

The existence of gene-tree/species-tree incongruities is not explained only as result of horizontal gene transfer events. Another valid explanation is the ancient (deep) unrecognized paralogies and subsequent, independent gene losses. While phylogenetic trees of genes that underwent HGT are always in conflict with expected organismal phylogenies, deep paralogy and differential gene losses produce trees that are not in conflict with the organismal phylogeny, if the paralogs are correctly identified and treated separately (Andersson and Roger 2003).

Paralogs are genes that are derived from a single ancestral gene by duplication within a genome, in contrast with orthologs, which are genes that have diverged from a common ancestral gene along with speciation of the organisms. Orthologs retain the same function, while paralogs often evolve to new functions. Gene duplications can occur in various ways like: unequal crossing over, retroposition or chromosomal (genome) duplication. Gene duplication could have both long- and short-term benefits. The short-term benefit is normally a higher level of gene expression. Long-term benefit is the acquirement of evolutionary innovations through the neofunctionalization or subfuntionalization of one copy of the duplicated gene. This is possible because the newborn paralog copy becomes free from forces of purifying selection and evolves rapidly while the other copy continues to fulfill the original function (Behe and Snoke 2004). On the other hand, the gene duplication increases energetic and material costs, because the cell has to replicate higher amount of DNA. Another fact about the gene duplication, which has been fully addressed by several authors, is that these duplicated genes protect the organism from harmful mutations by increasing the number of copies that are available for a specific function (Taylor and Raes 2004). Undoubtedly, the gene duplication is an important aspect of the genome evolution because new genes and innovations could originate in this way (Wagner 2010; Makarova et al. 2005).

If a gene was duplicated in a common ancestor and then only one randomly chosen copy was preserved in the lineage by its descendants (every lineage chooses independently from the others), the tree resulting from mixture of both paralogs will be incongruent with the organismal phylogeny. Such result can be often wrongly interpreted as a case of HGT (Glansdorff 2000).

Indication of a deep paralogy is the dual appearance of two paralogs in one organism, because this could be the remnant of the state after the gene duplication. Deep paralogy also assumes that the two paralogs should be able of long-term coexistence in one organism, i.e. should split the work rather than compete for the same function. This is an inherent assumption of every scenario that involves gene duplication and differential loss. Methods of the tree reconciliation that embed the gene tree into the species tree can be used to estimate the number of gene duplications and losses (Page and Charleston 1997; Guigó 1996).

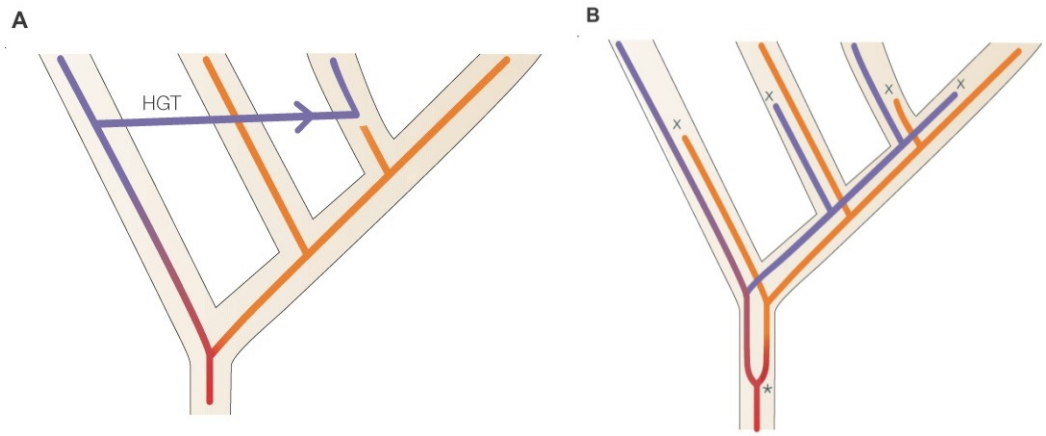

Figure 4. Schematic trees illustrating two evolutionary scenarios explaining the patchy distribution of a gene in the species tree (A) horizontal gene transfer scenario, (B) deep paralogy scenario followed by differential loss (Gogarten and Townsend 2005). Notice long periods of co-occurrence of two paralogs in B.

Usually it is hard to distinguish between HGT and deep paralogy followed by differential losses.

8. MAT/MATX *vs.* EF-1α/EFL

Several cases of genes patchy distributed among organisms have been reported so far (Andersson et al. 2006; Andersson and Roger 2003; Gile et al. 2009). Whilst

some of them were explained by gene duplication and subsequent loss in unrelated lineages, a HGT scenario was able to elucidate the other cases. Two genes received more attention because their evolutionary history seems particularly complicated. Besides MAT and MATX it was the case of translation elongation factors in eukaryotes. Both cases share several common features.

Eukaryotes and archaea possess the elongation factor 1α (EF-1α), while bacteria have its ortholog named elongation factor Tu (EF-Tu). Both are highly conserved GTPases that are involved in the process of translation elongation by delivering of aminoacyl tRNAs to the ribosome and, as well as MAT, they are essential housekeeping genes. Eukaryotic EF-1α is also involved in other cellular processes like nuclear export, cytoskeletal organization or negative regulation of genes involved in apoptosis (Gross and Kinzy 2005; Blanch et al. 2013; Duttaroy et al. 1998; Khacho et al. 2008).

Similarly to MATX, a paralog of EF-1α was found in eukaryotes; it was named elongation factor-like (EFL) (Keeling and Inagaki 2004). EFL is distributed frequently but discontinuously among eukaryotes and this distribution is usually mutually exclusive with EF-1α, like it is observed for the MAT/MATX case. The differences between EF-1α and EFL consist in six different insertions with variable length and number of unique substitutions (Dreher et al. 1999; Keeling and Inagaki 2004; Noble et al. 2007; Atkinson et al. 2014).

Firstly, the distribution of EFL was explained as the result of HGT because EFL was found in distantly related organisms, whose relatives possessed only the EF-1α form. At that time, no genome carrying both genes simultaneously was known (Keeling and Inagaki 2004). Since then, more organisms containing EFL and several containing both EF-1α and EFL genes have been discovered, so the HGT explanation for the evolution EFL became less convincing (Kamikawa et al. 2008; Sakaguchi et al. 2009; Keeling and Inagaki 2004). Existence of such “dual” organisms points to the presence of both paralogs in the common ancestor, demonstrating that this EF-1α/EFL duality is possible and may reflect the ancestral state. Nowadays, the evolution of this gene couple is explained purely by deep paralogy and subsequent losses, as no robust case of HGT has been demonstrated (Noble et al. 2007; Gile et al. 2009; Kamikawa et al. 2013; Keeling and Palmer 2008).

Until now, EFL was found in eight unrelated eukaryotic groups: dinoflagellates, haptophytes, cercozoa, green algae, choanoflagellates, fungi, diatoms and radiolarians (Keeling and Inagaki 2004; Kamikawa et al. 2011; Noble et al. 2007; Gile et al. 2009; Kamikawa et al. 2013; Mikhailov et al. 2014; Henk and Fisher 2012; Ishitani et al. 2012). Dual-EF1α/EFL-containing species were found in five distantly related lineages: goniomonadida, apusomonadida, diatoms, oomycetes and fungi. In all of these cases, the EFL gene is likely used as the principal elongation factor, since it has higher transcriptional level than EF-1α. Therefore, it was hypothesized that the EF-1α has been re-modeled to be functional only in auxiliary roles (Kamikawa et al. 2008; Kamikawa et al. 2013).

When we compare EF1α/EFL and MAT/MATX distributions, it is noticeable that the distribution of EFL is much patchier than of MATX, and also the dual EF1α/EFL-state was found in more organisms than the MAT/MATX-gene couple.

9. Euglenids and MATX evolution

On the MAT/MATX phylogenetic tree, MATX is forming a long branch clearly separated from MAT sequences. The position of the MATX clade is not resolved and the bootstrap supports for every placement were very low. Thus it is very hard (or even not possible) to reveal the affinities of MATX to the exact MAT lineage (Sánchez-Pérez et al. 2008).

Because MATX occurs only in taxa that possess a secondary plastid, one hypothesis postulates that MATX could arise during the secondary endosymbiosis of plastid from the endosymbiont MAT gene coded in the nucleomorph (Fig. 5) (Sánchez-Pérez et al. 2008). As it was reported before (Patron et al. 2006), genes encoded in nucleomorphs are released from the purifying selection and may undergo accelerated sequence evolution resulting in a divergent form of the gene. The hypothesis of Sánchez-Pérez et al. (2008) postulates that after the origin of MATX in one group of algae, MATX spread to other eukaryotic lineages *via* HGT (Fig. 4A). There were suggested three algal groups in which MATX could originate – dinoflagellates, haptophytes and euglenids (Sánchez-Pérez et al. 2008). Diatoms were not considered because until that time the MATX was found only in one centric diatom *T. pseudonana*,

which bears also a MAT copy of the gene. It was suggested that *T. pseudonana* gained the MATX paralog by HGT from a haptophyte (Sánchez-Pérez et al. 2008). In later analyses, it came out that also other diatoms and pelagophyte alga *Aureococcus anophagefferens* possess MATX, and in some of them appeared dual MAT/MATX state (Kamikawa et al. 2009). Kamikawa et al. (2009) concluded that the ancestral diatom cell already possessed both paralogs. Then the diatoms like *Achnanthes kuwaitensis*, *Fragilariopsis cylindrus* and *Phaeodactylum tricornutum*, which harbor only MAT, have secondarily lost their MATX copy, while some others like *Thalassionema* and *Skeletonema* have secondarily lost their MAT. Finally, several diatoms (*Thalassiosira pseudonana*, *Detonula confervacea*, *Ditylum brightwellii*, *Asterionella glacialis* and *Cylindrotheca closterium*) preserved both.

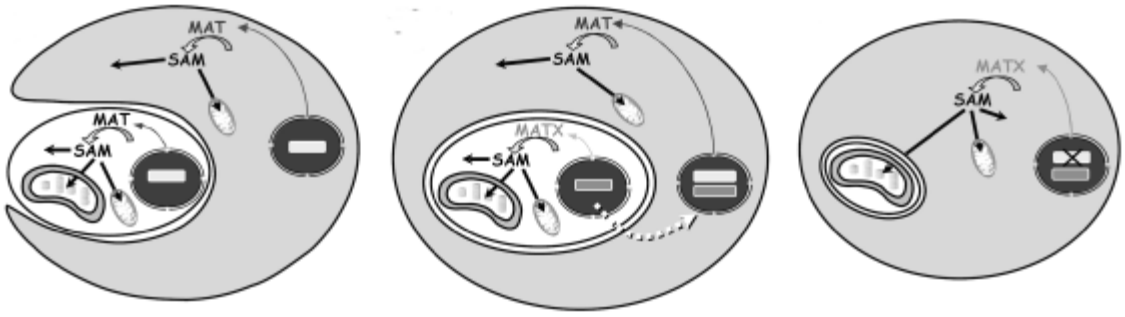

Figure 5: Illustration of a possible MATX origin in a photosynthetic organism during a secondary plastid endosymbiosis. A eukaryovore engulfed a plastid containing organism, both of which have MAT gene. After the plastid endosymbiosis, there was a rapid evolution of the MAT gene encoded in the endosymbionts nucleomorph. This led to the origin of MATX, which afterwards replaced the ancestral MAT gene in the host (Sánchez-Pérez et al. 2008)**.**

Regarding to the origin of MATX in the rest of the candidates, dinoflagellates were the main suspects, because of their predatory lifestyle and the ability to acquire different types of plastids by endosymbiosis. Therefore, there is a theoretical possibility that MATX evolved in one of these plastid endosymbionts. Some euglenids and haptophytes are also known to be able to engulf eukaryotic cells, contain complex plastids and theoretically could be the first hosts of MATX.

We had the opportunity to investigate evolution of both MAT and MATX in one of this candidate groups, euglenids. In order to do so, it was necessary to expand the sampling of euglenid taxa in the MAT/MATX tree, because in previous studies it was very low.

Euglenids are free-living flagellates that belong to the kingdom Excavata. It is a large group of organisms, which live in marine but predominantly fresh water environments. They have different modes of nutrition, including autotrophy, heterotrophy and mixotrophy. According to Leander (2004), the euglenid common ancestors were bacterivorous, from which later evolved eukaryovory. One lineage of eukaryovores gave rise to primary osmotrophs and another, after the acquisition of plastid, to photoautotrophic euglenids (Fig. 6).

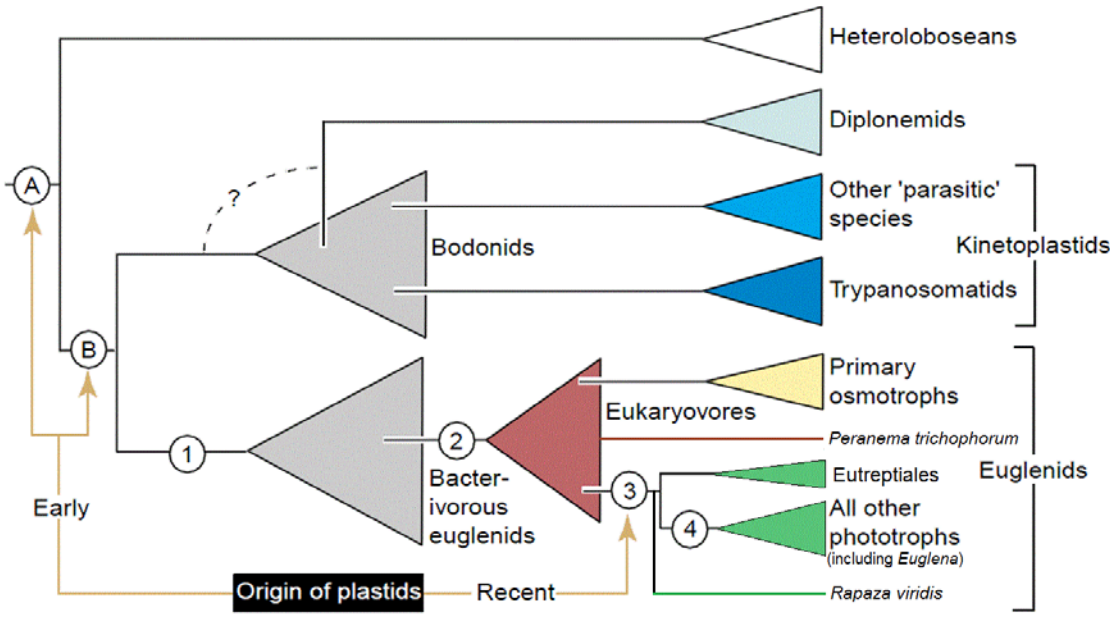

Figure 6: Schematic illustration of euglenid relationships. Adopted from Leander (2004)**.**

The photosynthetic euglenids together with secondarily osmotrophic euglenids (colorless euglenids, which secondarily lost plastids) form a robust clade, located within the rest of euglenids. The only known mixotroph today, *Rapaza viridis*, branches as the sister of photoautotrophs (Yamaguchi et al. 2012). It was previously known that autotroph *Euglena gracilis* possess MATX gene while eukaryovore *Peranema trichophorum* possess MAT gene. However, for the rest of the group we had no knowledge concerning the presence of MAT or MATX.

As it was mentioned above, photosynthetic euglenids possess a secondary green plastid. It was probably gained in the time period after the common ancestor of autotrophs + *Peranema* but before the common ancestor of autotrophs + *Rapaza*. This scenario is so called the “plastid-late hypothesis” or “plastid-recent hypothesis” (Fig. 6,

circle 3). In contrast, the “plastid-early hypothesis” (Fig. 6, circles A, B) assumes that the plastid was gained earlier in the euglenid evolution – already in the common ancestor of Euglenozoa. This hypothesis was suggested by Hannaert et al. (2003) when they found few genes encoding plant-like enzymes in kinetoplastids. However, according to Nozaki et al. (2003), these plant-like genes, which were found also in heterolobosea (Andersson and Roger 2002), could be derived from the ancient primary plastid endosymbiosis. Moreover, the re-analyses of the phylogenies of these genes by using recent wider taxon sampling did not support this hypothesis. Therefore, these plant-like genes were probably acquired by HGT from red and green algae (Soukal 2013).

The closest known relative to euglenid plastid is the prasinophyte marine alga *Pyramimonas parkeae* (Turmel et al. 2009). If the hypothesis that MATX came to euglenid lineage with the plastid endosymbiont is correct, the closest representative to MATX of euglenids would be *Pyramimonas parkeae*.

Besides analysing the distribution of MAT and MATX across euglenids, we decided to investigate in general changes, which occurred after plastid endosymbiosis in euglenids and thereby understand in more detail the changes in gene content, transfers of genes from plastid to nucleus or loss of genes in this particular case. For this purpose, we sequenced the plastid genome of *Eutreptiella gymnastica*, which represented, before the discovery of *R. viridis*, the lineage of photoautotrophic euglenids most distantly related to *Euglena*, and compared it with the plastid genome of *Euglena gracilis*.

The main aims of the thesis

- To experimentally test, how the two paralogs MAT and MATX satisfy assumptions of the two evolutionary scenarios: deep paralogy and horizontal gene transfer.

- To map the distribution of MAT/MATX within the group of euglenids.

- To characterize the possible evolutionary history of MATX paralog in eukaryotes with the particular focus on euglenids and to support of reject the hypothesis about the origin of MATX in this eukaryotic group.

- To analyze the changes in gene content of the secondary plastid of euglenids after the event of secondary endosymbiosis.

Publications

Szabová J., Růžička P., Verner Z., Hampl V., and Lukeš J. (2011): Experimental Examination of EFL and MATX Eukaryotic Horizontal Gene Transfers: Coexistence of Mutually Exclusive Transcripts Predates Functional Rescue. *Mol Biol Evol.* 28(8):2371-8.

Hrdá Š., Fousek J., **Szabová J.**, Hampl V., Vlček Č. (2012): The plastid genome of Eutreptiella provides a window into the process of secondary endosymbiosis of plastid in euglenids. *PLoS One* 7(3):e33746.

Szabová J., Yubuki N., Leander B. S., Triemer R. E., Hampl V. (2014): The evolution of paralogous enzymes MAT and MATX within the Euglenida and beyond. *BMC Evol Biol.* 14(25)

Examination of EFL and MATX Eukaryotic Horizontal Gene Transfers: Coexistence of Mutually Exclusive Transcripts Predates Functional Rescue

Szabová J., Růžička P., Verner Z., Hampl V., and Lukeš J. (2011).

Molecular Biology and Evolution 28(8):2371-8.

Experimental Examination of EFL and MATX Eukaryotic Horizontal Gene Transfers: Coexistence of Mutually Exclusive Transcripts Predates Functional Rescue

Jana Szabová,§,[1] Petr Růžička,†§,[2] Zdeněk Verner,‡,[2] Vladimír Hampl,§,[1] and Julius Lukeš*§,[2]

[1]Charles University in Prague, Faculty of Science, Department of Parasitology, Prague, Czech Republic

[2]Biology Centre, Institute of Parasitology, Czech Academy of Sciences, and Faculty of Sciences, University of South Bohemia, České Budějovice (Budweis), Czech Republic

†Present address: Institute of Experimental Botany, Czech Academy of Sciences, Prague, Czech Republic.

‡Present address: Sanatorium Helios s.r.o., Brno, Czech Republic.

§These authors contributed equally to this work.

***Corresponding author:** E-mail: jula@paru.cas.cz.

Associate editor: Hervé Philippe

Abstract

Many eukaryotic genes do not follow simple vertical inheritance. Elongation factor 1α (EF-1α) and methionine adenosyl transferase (MAT) are enzymes with complicated evolutionary histories and, interestingly, the two cases have several features in common. These essential enzymes occur as two relatively divergent paralogs (EF-1α/EFL, MAT/MATX) that have patchy distributions in eukaryotic lineages that are nearly mutually exclusive. To explain such distributions, we must invoke either multiple eukaryote-to-eukaryote horizontal gene transfers (HGTs) followed by functional replacement or presence of both paralogs in the common ancestor followed by long-term coexistence and differential losses in various eukaryotic lineages. To understand the evolution of these paralogs, we have performed in vivo experiments in *Trypanosoma brucei* addressing the consequences of long-term coexpression and functional replacement. In the first experiment of its kind, we have demonstrated that EF-1α and MAT can be simultaneously expressed with EFL and MATX, respectively, without affecting the growth of the flagellates. After the endogenous MAT or EF-1α was downregulated by RNA interference, MATX immediately substituted for its paralog, whereas EFL was not able to substitute for EF-1α, leading to mortality. We conclude that MATX is naturally capable of evolving patchy paralog distribution via HGTs and/or long-term coexpression and differential losses. The capability of EFL to spread by HGT is lower and so the patchy distribution of EF-1α/EFL paralogs was probably shaped mainly by deep paralogy followed by long-term coexistence and differential losses.

Key words: EFL, MATX, horizontal gene transfer, functional rescue, RNAi, *Trypanosoma*.

Introduction

The transfer of genetic information among distantly related organisms called horizontal (= lateral) gene transfer (HGT) represents one of the major driving forces of evolution (Keeling and Palmer 2008). The pervasive occurrence of HGT among prokaryotic organisms is apparent in their genomes and can be easily experimentally demonstrated, thus disputing the actual existence of stable bacterial species (Welch et al. 2002; Doolittle and Papke 2006; Doolittle and Bapteste 2007; Papke 2009). Recently, the role of HGT is becoming recognized also as an important force in the evolution of eukaryotes and an increasing number of examples are being reported (Andersson 2005, 2009; Richards et al. 2006; Watkins and Gray 2006; Andersson et al. 2007; Keeling 2009; Whitaker et al. 2009; Stairs et al. 2011). Even though several mechanisms of HGT have been proposed (Gogarten 2003), two are believed to be prevalent: endosymbiotic gene transfer (Martin and Schnarrenberger 1997) and "you-are-what-you-eat" (Doolittle 1998). The former occurred upon endosymbiosis of the bacterial ancestors of mitochondria and plastids and represents HGT on the largest scale because more than 90% of the endosymbiosed genomes were subsequently lost or transferred to the host cell nucleus and thus form a substantial part of the coding capacity of the nuclei in extant eukaryotes (Esser et al. 2004; Bock and Timmis 2008). Most of these organelle-derived proteins remain functionally associated with the organelle of their evolutionary origin (Kurland and Andersson 2000). However, some of these proteins have eventually found their way to other cellular compartments, being responsible for the mosaic pattern of most metabolic pathways in a typical eukaryotic cell (Gabaldón and Huynen 2004; Oborník and Green 2005).

Independently of these massive endosymbiosis-driven HGT events, most eukaryotes were subject to intermittent acquisitions of genomic material from prokaryotes or other eukaryotes. According to the "you-are-what-you-eat" concept, the digested prey is the pervasive source of these transferred genes (Doolittle 1998). Various eukaryotic groups differ in the extent of HGT from prokaryotes, and their life style and environment seem to be an important factor—HGTs occurred frequently in the evolutionary

history of rumen ciliates (Ricard et al. 2006), anaerobic protists (Andersson 2006; Andersson et al. 2007), or diatoms (Bowler et al. 2008), whereas very few such events have so far been documented in yeasts (Dujon et al. 2004) or animals (Kondrashov et al. 2006).

Although the eukaryote-to-eukaryote HGTs are likely to be underestimated (Keeling and Palmer 2008), this process can hardly be considered a frequent one. Particularly interesting are highly conserved essential genes that have a surprisingly complex evolutionary history—elongation factor 1α (EF-1α) and methionine adenosyltransferase (MAT) (Keeling and Inagaki 2004; Gile et al. 2006; Ruiz-Trillo et al. 2006; Noble et al. 2007; Kamikawa et al. 2008, 2009, 2010, 2011; Sanchez-Perez et al. 2008; Cocquyt et al. 2009; Gile, Faktorová, et al. 2009; Gile, Novis, et al. 2009; Sakaguchi et al. 2009), being the best known examples. The main function of EF-1α is to bring an aminoacyl-transfer RNA into the A site of the ribosome (Andersen et al. 2003). This extremely abundant protein has also been implicated in ubiquitin-dependent protein degradation (Chuang et al. 2005) and localization of selected transcripts via simultaneous binding of EF-1α to actin (Liu et al. 2002). MAT is the only enzyme synthesizing S-adenosyl-L-methionine, which is one of the key metabolites, as it donates the methyl group for most methylation reactions in prokaryotic and eukaryotic cells (Chiang et al. 1996).

Although elongation factor like (EFL) and MATX were initially considered to have evolved by vertical descent, the respective phylogenetic trees were inconsistent with such a simple scenario. In both cases, the analyses revealed that subsets of unrelated organisms possess a divergent version of the gene—EFL and MATX (Keeling and Inagaki 2004; Sanchez-Perez et al. 2008). In general, the patchy distribution of two paralogs can be explained by two outermost scenarios: 1) deep paralogy—presence of both paralogs in the common ancestor—followed by differential loss of one variant in individual lineages or 2) more recent origin of one paralog in one lineage of eukaryotes followed by its spread by eukaryote-to-eukaryote HGT. It is difficult to distinguish between these alternatives purely on the basis of phylogenetic analyses of protein sequences and their distributions. Theoretically, because the scenarios differ in the types of events they invoke, we might use the principle of parsimony and prefer the one that requires fewer improbable events. The first scenario minimizes the events of HGT replacements and expects long-term coexistence of both paralogs. The assumption of this scenario therefore is that the two paralogs are (or were in the past) capable of long-term coexpression in a single cell compartment without negative effect on the fitness of the organism. The second scenario expects that one paralog (probably the less frequent and less diversified one, namely MATX and EFL) spreads among eukaryotes via HGT. The assumption of this scenario is that this paralog is in a very short time able to substitute the function of its counterpart. Unfortunately, in the cases of EFL/EF-1α and MATX/MAT, we have no information how they fulfill one assumption or the other.

The coexistence of these paralogs under natural conditions is rare if not totally lacking. The distribution of EFL/EF-1α paralogs and MATX/MAT paralogs is almost strictly mutually exclusive, that is, organisms have either EF-1α or EFL but not both, the same applying to MAT and MATX. Strangely enough, the exceptional group in both cases is the diatoms. The genome of *Thalassiosira pseudonana* harbors both variants of these genes (EFL, EF-1α, MAT, and MATX) (Armbrust et al. 2004), transcripts of both EFL and EF-1α have been detected in *Th. pseudonana* and in five other diatom species (Kamikawa et al. 2008), and finally transcripts of both MAT and MATX were revealed in another four diatom species (Kamikawa et al. 2009). This almost strict mutually exclusive distribution led to a proposal that the long-term coexistence of both paralogs in one compartment is detrimental for the cell, probably due to problems with regulation, competition for substrate, or in the case of MAT/MATX, formation of less functional heteromers (Sanchez-Perez et al. 2008).

The process of HGT replacement of these essential proteins by their paralogs is, however, potentially problematic as well. It is difficult to envisage a smooth switch, during which these essential proteins are replaced with their horizontally acquired paralogs that take instantly over their functions. Moreover, the replacement is inevitably preceded by the potentially hazardous period of coexpression of both variants.

We have decided to experimentally test on the model of *Trypanosoma brucei* how the two paralog couples (EF-1α/EFL and MAT/MATX) satisfy assumptions of the two evolutionary scenarios. For the first time, we have simulated step-by-step the process of HGT under laboratory conditions. We have shown that EFL and MATX can coexist with EF-1α and MAT, respectively. Moreover, the MATX gene from *Euglena gracilis* was able to rescue the RNA interference (RNAi) knockdown for MAT in *Tr. brucei*, but in the same organism, the EFL gene from *Diplonema papillatum* failed to rescue the knockdown of EF-1α. Although MAT/MATX fulfills assumptions of both scenarios, EF-1α/EFL apparently fulfills just one of them.

Materials and Methods

EF-1α and EFL Constructs

Oligonucleotides for generation of gene fragments suitable for generation of RNAi knockdown cell lines were designed using the RNAit online tool available on the TrypanoFAN web site (http://trypanofan.path.cam.ac.uk/software/RNAit.html). The 453 bp-long 5′ region of the *Tr. brucei* EF-1α gene was amplified using oligonucleotides EF-1-F (5′-GGATCCTGGAGGCACTAGACATGCTG-3′) and EF-1-R (5′-AAGCTTCGATCTTCGACTCGATCTCC-3′) (added BamH*I* and Hind*III* restriction sites are underlined) and cloned via these restriction sites into the p2T7-177 RNAi vector carrying phleomycin resistance.

For constitutive expression in *Tr. brucei* of the full-size exogenous EFL, genes from *D. papillatum* or *Isochrysis galbana* were used. EFL from *D. papillatum* was expressed using

either a pABPURO vector containing an HA_3-tag and puromycin resistance (Long et al. 2008) or a pHD1344tub vector with TAP-tag and puromycin resistance (Carnes et al. 2008). For the expression of EFL from *I. galbana*, the pABPURO vectors with or without HA_3-tag were used. The entire open reading frame of the EFL gene (accession number ACO50119) was polymerase chain reaction (PCR) amplified from the cDNA of the diplonemid *D. papillatum* using oligonucleotides Dp-HA-F (5′-TCACATCGATATGGCTAACGCTACCGA-3′) and Dp-HA-R (5′-AGTGGCTAGCCTTCTTCTTGCCCTTGC-3′) (added ClaI and NheI restriction sites are underlined) for pABPURO, and oligonucleotides Dp-TAP-F (5′-TCACAAGCTTATGGCTAACGCTACCGA-3′) and Dp-TAP-R (5′-AGTGGGATCCCTTCTTCTTGCCCTTGC-3′) (added BamHI and HindIII restriction sites are underlined) for pHD1344tub.

In the case of the *I. galbana* EFL gene (accession number AAV34146), its entire open reading frame was PCR amplified from the total DNA using oligonucleotides Ig1-F (5′-AAGCTTATGGCCTCCGAGAAA-3′) and Ig1-R (5′-GGATCCCTACTTCTTCTTCTT-3′) (added HindIII and BamHI restriction sites are underlined), the amplicon was cloned into pCRII TOPO (Invitrogen) and subsequently recloned into the puromycin resistance–carrying pABPURO vectors with or without the HA_3-tag (Long et al. 2008). Proper integration of each construct was confirmed by sequencing. Comparison of the *I. galbana* EFL sequence with other EFL sequences has not revealed any introns that could potentially preclude the proper expression in trypanosomes.

MAT and MATX Constructs

To generate the RNAi knockdown cells, a 438 bp-long 5′ fragment of the *Tr. brucei* MAT gene was amplified using oligonucleotides IF-F (5′-TCACTCTAGAACGACGGTGTGTCAAATGAA-3′) and IF-R (5′-AGTGAAGCTTGCAGTCGGAAGTTTTTCTGC-3′) (added XbaI and HindIII restriction sites are underlined) and cloned into the p2T7-177 RNAi vector. Furthermore, the full-size MATX gene from the euglenid *E. gracilis* (accession number GU989640) was amplified from a cDNA clone using oligonucleotides RE-F (5′-TCACATCGATATGGCTGAATCTGCTTC-3′) and RE-R (5′-AGTGGCTAGCGTCCA CCCACTTCTGCA-3′) (added NheI and ClaI restriction sites are underlined). The amplicon was cloned into the pABPURO vector containing HA_3-tag and puromycin resistance as described above.

Transfection, Cloning, and RNAi Induction

The HA_3-tagged *E. gracilis* MATX in pABPURO was digested with MluI. Digestion with NotI was used to linearize all the other constructs. After digestion, 10 μg of each linearized vector was individually transfected into exponentially growing (at 27 °C in SDM-79 medium) procyclic *Tr. brucei* 29-13 strain or cell lines derived from thereof, using 2-mm cuvettes and a BTX electroporator with the settings of 1600 V, 25 μfarads, and 500 ohms. The clones were obtained after about 2-week cultivation by limiting dilution in 24-well plates at 27 °C in the presence of 5% CO_2, with 1 μg/ml puromycin or 1 μg/ml phleomycin as the selectable agent depending on the type of construct.

The following clonal cell lines derived from the 29-13 strain were prepared for the EF-1α/EFL experiments: 1) RNAi knockdowns containing *Tr. brucei* EF-1α in p2T7-177; 2) cells constitutively overexpressing HA_3-tagged *D. papillatum* EFL in pABPURO; 3) cells constitutively overexpressing HA_3-tagged *D. papillatum* EFL cotransfected with p2T7-177 containing EF-1α; 5) cells constitutively overexpressing TAP-tagged *D. papillatum* EFL in pHD1344tub; 5) cells constitutively overexpressing TAP-tagged *D. papillatum* EFL cotransfected with p2T7-177 containing EF-1α; 6) cells constitutively overexpressing HA_3-tagged *I. galbana* EFL in pABPURO; 7) cells constitutively overexpressing HA_3-tagged *I. galbana* EFL cotransfected with p2T7-177 containing EF-1α; 8) cells constitutively overexpressing nontagged *I. galbana* EFL in pABPURO; and 9) cells constitutively overexpressing nontagged *I. galbana* EFL cotransfected with p2T7-177 containing EF-1α.

The following clonal cell lines derived from the 29-13 strain were made for the MAT/MATX experiments: 1) RNAi knockdowns containing *Tr. brucei* MAT in p2T7-177; 2) cells constitutively overexpressing HA_3-tagged *E. gracilis* MATX in pABPURO; and 3) cells constitutively overexpressing HA_3-tagged *E. gracilis* MATX cotransfected with p2T7-177 containing MAT construct.

From each cell line containing the RNAi p2T7-177 vector, always a single clone was selected for further experiments based on the tightness of tetracycline-inducible expression of target double-stranded (ds) RNA and the corresponding robust elimination of target mRNA, as determined by Northern blot analysis using the *Tr. brucei* EF-1α or MAT gene as a probe. Synthesis of dsRNA was induced by the addition of 1 μg/ml tetracycline to the medium.

Northern Blot Analysis

Approximately 5 μg of total RNA/lane was loaded on a 1% formaldehyde agarose gel, blotted, and cross-linked following standard protocols. After prehybridization in NaPi solution (0.25 M Na_2HPO_4 and 0.25 M NaH_2PO_4, pH 7.2, 1 mM ethylenediaminetetraacetic acid, and 7% sodium dodecyl sulfate [SDS]) for 30 min at 60 °C, hybridization was performed overnight in the same solution at the same temperature. A wash in 2× saline sodium citrate (SSC) + 0.1% SDS at room temperature for 20 min was followed by three washes in 0.2× SSC + 0.1% SDS for 20 min each at 55 °C.

Western Blot Analysis

Cell lysates corresponding to 2.5×10^6 cells/lane were separated on a 15% sodium dodecyl sulfate–polyacrylamide gel electrophoresis, blotted, and for HA_3-tagged constructs, the membranes were treated with an anti-HA_3-tag mouse monoclonal antibody, followed by chicken anti-mouse antibody coupled to horseradish peroxidase. For TAP-tagged construct, the membranes were treated with anti-TAP-tag mouse monoclonal antibody, followed by rabbit anti-mouse antibody. Signal in Western blots was quantified with the Bio-Rad quantity one software.

Growth Analysis

Growth curves of selected *Tr. brucei* clones representing the (non)-induced RNAi knockdowns and the other genetically manipulated cells, obtained over a period of 12 days after RNAi induction were established using the Beckman Z2 Cell Counter.

Results

EFL and EF-1α mRNAs Are Fully Compatible

We have first tested whether EFL from *D. papillatum*, a diplonemid flagellate closely related to kinetoplastids, will be compatible with the *Tr. brucei* proteome, which contains the EF-1α. For that purpose, the full-size *D. papillatum* EFL gene was cloned into a vector that allows its constitutive expression under the procyclin promoter. The Not*I*-linearized vector was transfected into the 29-13 *Tr. brucei* strain where it was integrated into the tubulin locus. Total RNA was isolated from a puromycin-resistant clonal cell line and analyzed by Northern blotting using the full-size *D. papillatum* EFL gene as a probe. The analysis showed that the introduced gene was strongly transcribed in the transfectants, whereas no signal was detected in the wild type 29-13 cells (fig. 1*A*). At stringent hybridization conditions (60 °C), no cross-hybridization with the EF-1α mRNA, transcribed from three endogenous copies of the EF-1α gene, was observed. Western blot analysis using specific anti-HA_3 or anti-TAP tag monoclonal antibodies detected a tagged protein translated from the heterologous gene (fig. 1*B*). The morphology of *Tr. brucei* cells containing both the endogenous EF-1α and exogenous EFL appeared normal by light microscopy (data not shown), and their growth was similar in comparison with the wild-type cells (fig. 2*B* and *C*) indicating that EF-1α and EFL are fully compatible.

Silencing of EF-1α Inhibits Growth

The addition of tetracycline into the medium triggers synthesis of dsRNA in trypanosomes transfected with the EF-1α-containing inducible RNAi construct. The extent of EF-1α mRNA silencing and the tightness of its inducible ablation was determined by Northern blotting, which revealed lack of leaky dsRNA transcription (fig. 1*C*, lanes 3 and 7) and virtually complete ablation of EF-1α mRNA upon the induction of RNAi (fig. 1*C*, lanes 4 and 8). Next, we have followed the growth of cells constitutively expressing exogenous *D. papillatum* EFL in which RNAi against EF-1α was induced. The elimination of EF-1α mRNA triggers an almost instant cessation of growth, eventually causing death regardless of the absence (fig. 2*A*) or presence (fig. 2*B* and *C*) of exogenous EFL that is efficiently translated (fig. 1*B*), yet still fails to rescue the growth phenotype.

To confirm these findings, similar experiments were performed with cells with constitutive expression of the EFL gene from a different organism, the haptophyte *I. galbana*. In this series of experiments, we also wanted to establish whether the attachment of a tag to the C-terminus of expressed exogenous EFL protein does (not) interfere with its enzymatic function. Therefore, two parallel experiments were performed; one using construct with HA_3 tagged EFL and the other using nontagged EFL. Because results of both experiments were very similar, only the experiments without tag are presented below. Successful constitutive expression of exogenous EFL and inducible downregulation of endogenous EF-1α was confirmed using Northern blotting (supplementary fig. S1, Supplementary Material online). As in the case of EFL from *D. papillatum*, trypanosomes in which both EF-1α and *I. galbana* EFL were expressed grew at the same rate as the wild-type cells, with the exogenous EFL failing to rescue the growth of the cells with RNAi-ablated EF-1α (fig. 2*D*).

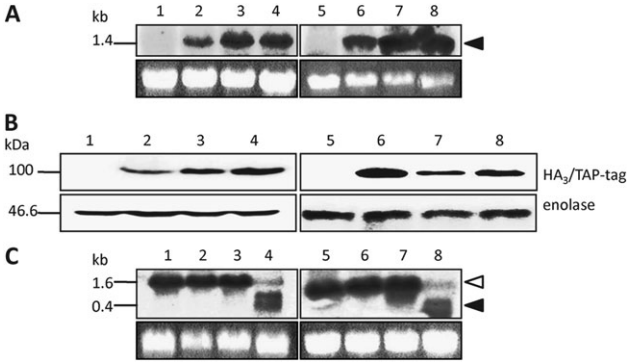

Fig. 1. Expression of exogenous EFL from *Diplonema papillatum* and (parallel) RNAi silencing of EF-1α in *Trypanosoma brucei*. (*A*) The EFL mRNA is expressed in *Trypanosoma brucei* cells. Upper panels: Levels of EFL mRNA with HA3-tag (lanes 1–4) and EFL mRNA with TAP-tag (lanes 5–8) were analyzed by blotting 10 μg of total RNA/lane extracted from 29-13 wild-type cells (lanes 1 and 5), cells constitutively expressing exogenous EFL (lanes 2, 6), noninduced cells expressing EFL and also containing RNAi vector against EF-1α (lanes 3 and 7), and same cells as in lanes 3 and 7 in which RNAi was induced (lanes 4 and 8). The full-length EFL gene was used as a probe, and hybridization was performed at 60 °C, at which no cross-hybridization with EF-1α mRNA occurs. Lower panels: As a loading control, the gel was stained with ethidium bromide to visualize ribosomal RNA (rRNA) bands. (*B*) The EFL protein is expressed in *Tr. brucei* cells. Upper panels: The levels of the HA3-tagged (lanes 1–4) and TAP-tagged (lanes 5–8) exogenous EFL protein were followed using specific mouse monoclonal antibodies. The levels were analyzed in total lysates (from 5×10^6cells) from 29-13 wild-type cells (lanes 1 and 5), cells constitutively expressing EFL (lanes 2 and 6), noninduced cells constitutively expressing exogenous EFL and containing RNAi vector against EF-1α (lanes 3 and 7), and the same cells as in lanes 3 and 7 in which RNAi was induced (lanes 4 and 8). Lower panels: Enolase visualized by specific rabbit polyclonal antibodies was used as loading control. (*C*) Down regulation of EF-1α. Upper panels: Level of EF-1α mRNA in the cells with HA3-tagged (lanes 1–4) and TAP-tagged exogenous EFL (lanes 5–8) were analyzed in total RNA extracted from 29-13 wild-type cells (lanes 1 and 5), cells constitutively expressing EFL (lanes 2 and 6), noninduced cells containing RNAi vector against EF-1α and constitutively expressing EFL (lanes 3 and 7); and same cells as in lanes 3 and 7 in which RNAi was induced (lanes 4 and 8). The full-length EF-1α gene was used as a probe, and hybridization was performed at 60 °C, at which no cross-hybridization with EFL mRNA occurs. The positions of the EF-1α mRNAs and respective dsRNA are indicated with white and gray arrowheads, respectively. Lower panels: As a loading control, the gel was stained with ethidium bromide to visualize rRNA bands.

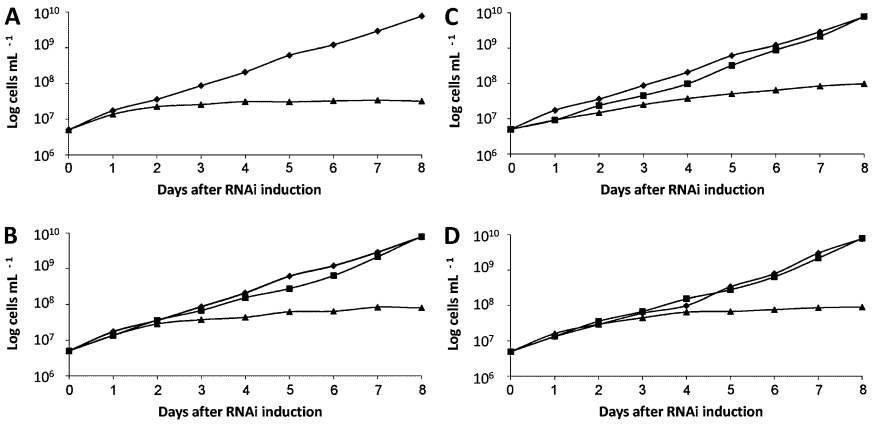

FIG. 2. Lethality RNAi-ablation of EF-1α is not rescued by exogenous EFL in *Trypanosoma brucei*. Cell numbers were measured using a Coulter Counter Z2. The *y* axis is labeled by a logarithmic scale and represents the product of cell densities measured and total dilution. Growth curves are one representative set from three independent experiments. (*A*) The growth of cells with ablated EF-1α mRNA is inhibited (triangles), as compared with the 29-13 wild-type cells (diamonds). (*B*) The growth of cells with inducibly ablated EF-1α mRNA that also constitutively express *Diplonema papilatum* EFL with HA3-tag is inhibited (triangles) in comparison with the noninduced cells constitutively expressing the same EFL (squares) and 29-13 wild-type cells (diamonds), which grow at the same rate. (*C*) The growth of cells with inducibly ablated EF-1α mRNA that also constitutively express *D. papilatum* EFL with TAP-tag is inhibited (triangles) in comparison with the noninduced cells constitutively expressing the same EFL (squares) and 29-13 wild-type cells (diamonds), which grow at the same rate. (*D*) The growth of cells with inducibly ablated EF-1α mRNA that also constitutively express *Isochrysis galbana* EFL is inhibited (triangles) in comparison with the noninduced cells constitutively expressing the same EFL (squares) and 29-13 wild-type cells (diamonds), which grow at the same rate.

MAT and MATX mRNAs Are Compatible

The same strategy as the one described above for the EFL and EF-1α genes was also used for the MAT/MATX system. First, using Northern blot analysis and the MATX gene as a probe, we have shown that in transfected *Tr. brucei*, the MATX gene from *E. gracilis* is indeed expressed (fig. 3*A*). This approach also excluded the unlikely yet possible presence of another MATX gene in the 29-13 wild-type cells, which contain nine copies of the MAT gene in their genome. As shown by Western blot analysis using specific anti-HA_3 tag monoclonal antibody, MATX is not only transcribed but also efficiently translated in *Tr. brucei* transfected with the respective construct (fig. 3*B*). Growth curve analysis of wild-type cells and those overexpressing MATX clearly demonstrated that trypanosomes fully tolerate expression of this exogenous gene (fig. 4*B*).

MATX Rescues MAT Deficiency

Next, we have downregulated endogenous MAT mRNA using RNAi. In selected clones, the ablation was very efficient, because after 48 h of RNAi induction, virtually no target transcript was detectable by Northern blot analysis (fig. 3*C*). We have then followed the consequences of such depletion on cell growth. As shown in figure 4*A*, MAT is clearly an essential protein for trypanosomes, as they were unable to propagate in its absence. The situation was, however, strikingly different with cells depleted for MAT but overexpressing *E. gracilis* MATX. The constitutive expression of MATX in the inducible *Tr. brucei* RNAi MAT knockdown fully rescued the growth, which differed neither from the 29-13 wild-type cells nor from the noninduced knockdowns (fig. 4*B*). This experiment shows that when the euglenid MATX is concurrently expressed with the endogenous MAT, following the depletion of the latter protein, MATX quickly takes over its function(s) and rescues the cells, which would otherwise die.

Discussion

Using in vivo experiments, we have mimicked the process of acquisition of an exogenous paralog of an enzyme by HGT followed by a period of simultaneous expression and eventually leading to functional replacement of the endogenous paralog. Our experiments have several important implications for the evolution of the studied paralogs.

For two gene couples (MAT/MATX and EF-1α/EFL) whose members are virtually never found simultaneously in one cell, we have demonstrated for the first time that their products can cohabitate in *Tr. brucei* in the same compartment at least for several weeks, which is a substantially long period for an organism with 8 h-long generation time. Although our results do not attach any selective advantage to this highly risky and cumbersome process, they show that it can indeed happen under experimentally controlled conditions. Growth phenotype of trypanosomes coexpressing MAT and MATX or EF-1α and EFL is similar and comparable with the wild-type cells. The results of our experiments are in agreement with the observations that MAT and MATX mRNA as well as EF-1α and EFL mRNA are present in several diatom species at the same time (Kamikawa et al. 2008, 2009). In the light of these findings, the long-term cohabitation followed by slow and more or less random differential losses of one or the other paralog is a plausible scenario for both paralog couples.

MATX and EFL differ in their ability to substitute the endogenous paralog in trypanosomes. After downregulation of MAT by RNAi, the expressed exogenous MATX was able

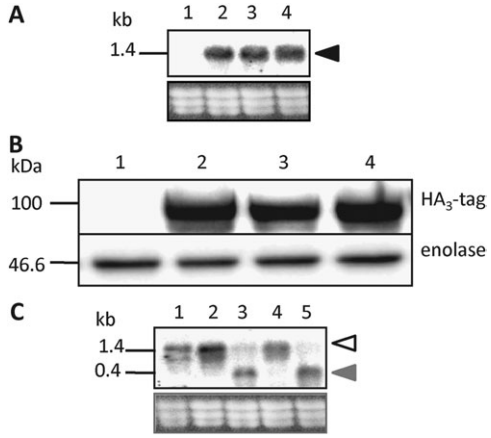

FIG. 3. Expression of exogenous MATX from *Euglena gracilis* and (parallel) RNAi silencing of MAT in *Trypanosoma brucei*. (*A*) The MATX mRNA is expressed in *T. brucei* cells. Upper panel: Level of MATX mRNA was analyzed by blotting 10 μg of total RNA extracted from 29-13 cells (lane 1), cells constitutively expressing MATX from *E. gracilis* (lane 2), noninduced cells constitutively expressing exogenous MATX and containing RNAi vector against endogenous MAT (lane 3), and the same cells as in lane 3 in which RNAi was induced (lane 4). The full-size MATX gene was used as a probe, and hybridization was performed at 60 °C. The position of the MATX mRNA is indicated with a black arrowhead. Lower panel: As a loading control, the gel was stained with ethidium bromide to visualize ribosomal RNA (rRNA) bands. (*B*) The MATX protein is expressed in *T. brucei*. Upper panel: The levels of the HA3-tagged exogenous MATX protein were followed using specific mouse monoclonal antibodies. The levels were analyzed in total lysates from 29-13 wild-type cells (lane 1), cells constitutively expressing MATX (lane 2), noninduced cells constitutively expressing exogenous MATX and containing RNAi vector against MAT (lane 3), and the same cells as in lane 3 in which RNAi was induced (lane 4). Lower panel: Enolase visualized by specific rabbit polyclonal antibodies was used as loading control. (*C*) Down regulation of MAT. Upper panel: Levels of MAT mRNA and respective dsRNA were analyzed in total RNA extracted from the following cell lines: 29-13 wild-type cells (lane 1); noninduced cells containing RNAi vector against endogenous MAT and constitutively expressing exogenous MATX (lane 2); same cells as in lane 2 in which RNAi was induced (lane 3); noninduced cells containing RNAi vector against MAT (lane 4); same cells as in lane 4 in which RNAi was induced (lane 5). The 5′ region of the *T. brucei* MAT gene was used as a probe, and hybridization was performed at 60 °C. The positions of the targeted MAT mRNA and the dsRNA are indicated with white and gray arrowheads, respectively. Lower panel: As a loading control, the gel was stained with ethidium bromide to visualize rRNA bands.

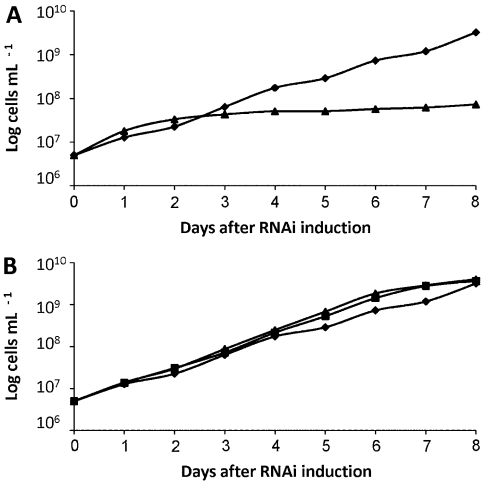

FIG. 4. Lethality of RNAi-ablated MAT is rescued by exogenous MATX in *Trypanosoma brucei*. The growth curves were performed as described in figure 2. (*A*) The growth of cells with ablated MAT mRNA is inhibited (triangles), as compared with 29-13 wild-type cells (diamonds). (*B*) The growth of cells with inducibly ablated MAT mRNA is rescued by the expression of exogenous MATX (triangles), and the cells grow at about the same rate as the noninduced cells constitutively expressing the same MATX (squares) and 29-13 wild-type cells (diamonds).

to take over the function of its counterpart. The functional replacement of MAT by MATX happened immediately with no effect on the growth phenotype. This result is in agreement with the work of Ho et al. (2007), who showed that MATX from a dinoflagellate *Crypthecodinium cohnii* rescued the MAT knockout of yeast. On the other hand, cells expressing EFL but depleted for EF-1α died in our experiments irrespective of the presence or absence of tags on the EFL protein and irrespective of the *Diplonema* or *Isochrysis* origin of the EFL gene.

Because horizontally transferred genes may be disadvantaged in codon usage, we have compared this parameter for both gene couples. Codon usage of MATX and both EFLs departs from the kinetoplastid consensus to about the same extent as the endogenous MAT and EF-1α (supplementary figs. S2–S8, Supplementary Material online). We therefore conclude that codon usage bias does not play significant role in establishment of horizontally acquired MATX and EFL.

The analysis of functional divergence in case of EFL (Keeling and Inagaki 2004) and the conservation of all functional residues in case of MATX (Sanchez-Perez et al. 2008) suggest that these distant paralogs can perform the function of their counterparts. Nevertheless, the overall amino acid identity between EF-1α and EFL (40–45%) is generally lower than the amino acid identity between MAT and MATX (55–64%, excluding insertions). The amino acid identities between the particular gene pairs that have been used in our experiments fell into the aforementioned ranges, *Euglena* MATX with *Trypanosoma* MAT, *Diplonema* EFL with *Trypanosoma* EF-1α, and *Isochrysis* EFL with *Trypanosoma* EF-1α share 55%, 44%, and 42% of amino acids, respectively. The lower identity between EFL and EF-1α might contribute to the incapability of EFL to functionally substitute EF-1α. The fact that EFL was not able to substitute its paralog while MATX was could also be explained by the complexity hypothesis (Cohen et al. 2010). This hypothesis posits that genes, whose products are involved in many interactions with other proteins or molecules (like elongation factors), are less prone to transfers than genes with less interactions (enzymes like MATX).

Judging by their phylogenies and distribution among eukaryotes, the cases of EF-1α/EFL and MAT/MATX look very similar, yet our laboratory experiments indicate that this similarity may only be superficial. In the case of MAT/MATX, both long-term coexistence and horizontal transfer

followed by instantaneous functional replacement are plausible. MAT/MATX pair therefore satisfies assumptions of both outermost scenarios how the patchy distribution might evolve, that is, ancestral presence of both paralogs followed by differential losses or origin in one lineage followed by spread via HGTs. Results of our experiments thus do not help judging which of these scenarios are more plausible in this particular case. In fact, it is reasonable to expect that both phenomena—differential loss and HGT—contributed to the evolution of MAT/MATX patchy distribution. In the case of EF-1α/EFL, we have demonstrated that the coexistence of both variants is possible. This result is in agreement with the hypotheses that this dual state could be maintained for millions of years in euglenids (Gile, Faktorová, et al. 2009), diatoms (Kamikawa et al. 2008), green algae (Noble et al. 2007; Cocquyt et al. 2009), or even for much longer time since the common ancestor of all extant eukaryotes (Kamikawa et al. 2010). On the contrary, the instantaneous functional replacement of endogenous EF-1α by exogenous EFL was not successful, so at least in our experimental setting, we were not able to show that the EF-1α/EFL pair fulfills the assumption of the multiple-HGT scenario, and this scenario therefore seems in this particular case less probable. This does not mean that after a sufficiently long period of coexpression, when the cells become adapted to the exogenous paralog, EFL could not be able to substitute EF-1α. This situation, depending on the length of the adaptation period, however, approaches to the long-term coexpression followed by differential losses scenario. It is theoretically possible that the EFL gene was horizontally transferred at some points of its evolutionary history yet there is better evidence that coexpression followed by losses played a major role in the shaping the distribution of EFL and EF1-α.

In summary, the process of HGT and functional replacement of paralogs was simulated in a step-by-step fashion, which allowed to directly demonstrate that two relatively divergent variants of essential proteins can be coexpressed in vivo. A trouble-free simultaneous expression represents a necessary assumption of scenarios invoking deep paralogy and differential losses to explain the complex distribution of paralogs in the eukaryotic tree. Our experiments thus increase plausibility of this scenario for both EF-1α/EFL and MAT/MATX paralog pairs. Unlike EFL, MATX exhibits also a natural capability to spread among eukaryotes by horizontal transfer and instantaneous functional replacement indicating that also this mechanism might play a role in the evolutionary history of this particular paralog pair.

Supplementary Materials

Supplementary figures S1–S8 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to acknowledge Gabino Sanchez-Perez and Dion G. Durnford for kindly providing the *E. gracilis* MATX clone and Bryony Williams and Patrick Keeling for sharing the *I. galbana* EFL clone and Zuzana Vavrova (Biology Centre) for help with some experiments. This work was supported by the Czech Science Foundation 204/09/1667 (to J.L.), the Czech Science Foundation P506/11/1320 (to V.H.), the Grant Agency of the Charles University 63409 (to J.S.), the Ministry of Education of the Czech Republic (2B06129, 6007665801 and 0021620828)(to J.L. and V.H.), and the Praemium Academiae award (to J.L).

References

Andersen GR, Nissen P, Nyborg J. 2003. Elongation factors in protein biosynthesis. *Trends Biochem Sci*. 28:434–441.

Andersson JO. 2005. Lateral gene transfer in eukaryotes. *Cell Mol Life Sci*. 62:1182–1197.

Andersson JO. 2006. Genome evolution of anaerobic protists: metabolic adaptation via gene acquisition. In: Katz LA, Bhattacharya D, editors. Genomics and evolution of microbial eukaryotes. Oxford: Oxford University press. p. 109–122.

Andersson JO. 2009. Horizontal gene transfer between microbial eukaryotes. In: Gogarten MB, Gogarten JP, Olendzenski L, editors. Horizontal gene transfer. New York: Humana Press. p. 473–487.

Andersson JO, Sjogren AM, Horner DS, Murphy CA, Dyal PL, Svard SG, Logsdon JM Jr., Ragan MA, Hirt RP, Roger AJ. 2007. A genomic survey of the fish parasite *Spironucleus salmonicida* indicates genomic plasticity among diplomonads and significant lateral gene transfer in eukaryote genome evolution. *BMC Genomics*. 8:51.

Armbrust EV, Berges JA, Bowler C, et al. (45 co-authors). 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306:79–86.

Bock R, Timmis JN. 2008. Reconstructing evolution: gene transfer from plastids to the nucleus. *Bioessays* 30:556–566.

Bowler C, Allen AE, Badger JH, et al. (77 co-authors). 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature* 456:239–244.

Carnes J, Trotter JR, Peltan A, Fleck M, Stuart K. 2008. RNA editing in *Trypanosoma brucei* requires three different editosomes. *Mol Cell Biol*. 28:122–130.

Chiang PK, Gordon RK, Tal J, Zeng GC, Doctor BP, Pardhasaradhi K, McCann PP. 1996. S-adenosylmethionine and methylation. *FASEB J*. 10:471–480.

Chuang SM, Chen L, Lambertson D, Nand M, Kinzy TG, Madura K. 2005. Proteasome-mediated degradation of cotranslationally damaged proteins involves translation elongation factor 1A. *Mol Cell Biol*. 25:403–413.

Cocquyt E, Verbruggen H, Leliaert F, Zechman FW, Sabbe K, De Clerck O. 2009. Gain and loss of elongation factor genes in green algae. *BMC Evol Biol*. 9:39.

Cohen O, Gophna U, Pupko T. 2010. The complexity hypothesis revisited: connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol Biol Evol*. 21:1643–1660.

Doolittle WF. 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet*. 14:307–311.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. *Proc Natl Acad Sci U S A*. 104:2043–2049.

Doolittle WF, Papke RT. 2006. Genomics and the bacterial species problem. *Genome Biol*. 7:116.

Dujon B, Sherman D, Fischer G, et al. (67 co-authors). 2004. Genome evolution in yeasts. *Nature* 430:35–44.

Esser C, Ahmadinejad N, Wiegand C, et al. (15 co-authors). 2004. A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol Biol Evol*. 21:1643–1660.

Gabaldón T, Huynen MA. 2004. Shaping the mitochondrial proteome. *Biochim Biophys Acta*. 1659:212–220.

Gile GH, Faktorová D, Castlejohn CA, Burger G, Lang BF, Farmer MA, Lukeš J, Keeling PJ. 2009. Distribution and phylogeny of EFL and EF-1α in Euglenozoa suggest ancestral co-occurrence followed by differential loss. *PLoS One*. 4:e5162.

Gile GH, Novis PM, Cragg DS, Zuccarello GC, Keeling PJ. 2009. The distribution of Elongation factor-1 alpha (EF-1alpha), Elongation factor-like (EFL), and a non-canonical genetic code in the ulvophyceae: discrete genetic characters support a consistent phylogenetic framework. *J Eukaryot Microbiol*. 56:367–372.

Gile GH, Patron NJ, Keeling PJ. 2006. EFL GTPase in cryptomonads and the distribution of EFL and EF-1alpha in chromalveolates. *Protist* 157:435–444.

Gogarten JP. 2003. Gene transfer: gene swapping craze reaches eukaryotes. *Curr Biol*. 13:R53–R54.

Ho P, Kong KF, Tang JSH, Wong JTY. 2007. An unusual S-adenosylmethionine synthetase gene from dinoflagellate is methylated. *BMC Mol Biol*. 8:87.

Kamikawa R, Inagaki Y, Sako Y. 2008. Direct phylogenetic evidence for lateral transfer of elongation factor-like gene. *Proc Natl Acad Sci U S A*. 105:6965–6969.

Kamikawa R, Sakaguchi M, Matsumoto T, Hashimoto T, Inagaki Y. 2010. Rooting for the root of elongation factor-like protein phylogeny. *Mol Phylogenet Evol*. 56:1082–1088.

Kamikawa R, Sanchez-Perez GF, Sako Y, Roger AJ, Inagaki Y. 2009. Expanded phylogenies of canonical and non-canonical types of methionine adenosyltransferase reveal a complex history of these gene families in eukaryotes. *Mol Phylogenet Evol*. 53: 565–570.

Kamikawa R, Yabuki A, Nakayama T, Ishida K, Hashimoto T, Inagaki Y. 2011. Cercozoa comprises both EF-1α-containing and EFL-containing members. *Eur J Protistol*. 47:24–28.

Keeling PJ. 2009. Functional and ecological impacts of horizontal gene transfer in eukaryotes. *Curr Opin Genet Dev*. 19: 613–619.

Keeling PJ, Inagaki Y. 2004. A class of eukaryotic GTPase with a punctuate distribution suggesting multiple functional replacements of translation elongation factor 1alpha. *Proc Natl Acad Sci U S A*. 101:15380–15385.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*. 9:605–618.

Kondrashov FA, Koonin EV, Morgunov IG, Finogenova TV, Kondrashova MN. 2006. Evolution of glyoxylate cycle enzymes in metazoa: evidence of multiple horizontal transfer events and pseudogene formation. *Biol Direct*. 1:31.

Kurland CG, Andersson SG. 2000. Origin and evolution of the mitochondrial proteome. *Microbiol Mol Biol Rev*. 64:786–820.

Liu G, Grant WM, Persky D, Latham VM Jr, Singer RH, Condeelis J. 2002. Interactions of elongation factor 1alpha with F-actin and beta-actin mRNA: implications for anchoring mRNA in cell protrusions. *Mol Biol Cell*. 13:579–592.

Long S, Jirků M, Mach J, Ginger ML, Sutak R, Richardson D, Tachezy J, Lukeš J. 2008. Ancestral roles of eukaryotic frataxin: mitochondrial frataxin function and heterologous expression of hydrogenosomal *Trichomonas* homologs in trypanosomes. *Mol Microbiol*. 69:94–109.

Martin W, Schnarrenberger C. 1997. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. *Curr Genet*. 32:1–18.

Noble GP, Rogers MB, Keeling PJ. 2007. Complex distribution of EFL and EF-1alpha proteins in the green algal lineage. *BMC Evol Biol*. 7:82.

Oborník M, Green BR. 2005. Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. *Mol Biol Evol*. 22: 2343–2353.

Papke RT. 2009. A critique of prokaryotic species concepts. In: Gogarten MB, Gogarten JP, Olendzenski L, editors. Horizontal gene transfer. New York: Humana Press. p. 379–395.

Ricard G, McEwan NR, Dutilh BE, et al. (17 co-authors). 2006. Horizontal gene transfer from bacteria to rumen ciliates indicates adaptation to their anaerobic, carbohydrate-rich environment. *BMC Genomics*. 7:22.

Richards TA, Dacks JB, Campbell SA, Blanchard JL, Foster PG, McLeod R, Roberts CW. 2006. Evolutionary origins of the eukaryotic shikimate pathway: gene fusions, horizontal gene transfer, and endosymbiotic replacements. *Eukaryot Cell*. 5:1517–1531.

Ruiz-Trillo I, Lane CE, Archibald JM, Roger AJ (17 co-authors). 2006. Insights into the evolutionary origin and genome architecture of the unicellular opisthokonts *Capsaspora owczarzaki* and *Sphaeroforma arctica*. *J Eukaryot Microbiol*. 53:379–384.

Sakaguchi M, Takishita K, Matsumoto T, Hashimoto T, Inagaki Y. 2009. Tracing back EFL gene evolution in the cryptomonads–haptophytes assemblage: separate origins of EFL genes in haptophytes, photosynthetic cryptomonads, and goniomonads. *Gene* 441:126–131.

Sanchez-Perez GF, Hampl V, Simpson AGB, Roger AJ. 2008. A new divergent type of eukaryotic methionine adenosyltransferase is present in multiple distantly related secondary algal lineages. *J Eukaryot Microbiol*. 55:374–381.

Stairs CW, Roger AJ, Hampl V. 2011. Eukaryotic pyruvate formate lyase and its activating enzyme were acquired laterally from a firmicute. *Mol Biol Evol*. 28:2087–2099.

Watkins RF, Gray MW. 2006. The frequency of eubacterium-to-eukaryote lateral gene transfers shows significant cross-taxa variation within amoebozoa. *J Mol Evol*. 63:801–814.

Welch RA, Burland V, Plunkett G 3rd, et al. (19 co-authors). 2002. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc Natl Acad Sci U S A*. 99:17020–17024.

Whitaker JW, McConkey GA, Westhead DR. 2009. Prediction of horizontal gene transfers in eukaryotes: approaches and challenges. *Biochem Soc Trans*. 37:792–795.

The evolution of paralogous enzymes MAT and MATX within the Euglenida and beyond

Szabová J., Yubuki N., Leander B. S., Triemer R. E., Hampl V. (2014). *BMC Evolutionary Biology* 14(25).

Szabová et al. BMC Evolutionary Biology 2014, 14:25
http://www.biomedcentral.com/1471-2148/14/25

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

The evolution of paralogous enzymes MAT and MATX within the Euglenida and beyond

Jana Szabová[1,2*], Naoji Yubuki[3], Brian S Leander[3], Richard E Triemer[4] and Vladimír Hampl[1,2*]

Abstract

Background: Methionine adenosyltransferase (MAT) is a ubiquitous essential enzyme that, in eukaryotes, occurs in two relatively divergent paralogues: MAT and MATX. MATX has a punctate distribution across the tree of eukaryotes and, except for a few cases, is mutually exclusive with MAT. This phylogenetic pattern could have arisen by either differential loss of old paralogues or the spread of one of these paralogues by horizontal gene transfer. Our aim was to map the distribution of MAT/MATX genes within the Euglenida in order to more comprehensively characterize the evolutionary history of MATX.

Results: We generated 26 new sequences from 23 different lineages of euglenids and one prasinophyte alga *Pyramimonas parkeae*. MATX was present only in photoautotrophic euglenids. The mixotroph *Rapaza viridis* and the prasinophyte alga *Pyramimonas parkeae*, which harbors chloroplasts that are most closely related to the chloroplasts in photoautotrophic euglenids, both possessed only the MAT paralogue. We found both the MAT and MATX paralogues in two photoautotrophic species (*Phacus orbicularis* and *Monomorphina pyrum*). The significant conflict between eukaryotic phylogenies inferred from MATX and SSU rDNA data represents strong evidence that MATX paralogues have undergone horizontal gene transfer across the tree of eukaryotes.

Conclusions: Our results suggest that MATX entered the euglenid lineage in a single horizontal gene transfer event that took place after the secondary endosymbiotic origin of the euglenid chloroplast. The origin of the MATX paralogue is unclear, and it cannot be excluded that it arose by a gene duplication event before the most recent common ancestor of eukaryotes.

Keywords: Methionine adenosyltransferase, Horizontal gene transfer, Deep paralogy, Gene evolution, Euglenozoa

Background

Methionine adenosyltransferase (MAT) is a cytosolic ubiquitous enzyme that synthesizes S-adenosyl-L-methionine (SAM), a molecule that is one of the most important metabolites in living cells. SAM serves as the major methyl donor to phospholipids, DNA, RNA and other small molecules and is the second most widely used enzyme substrate after ATP [1,2]. MAT is a well-conserved enzyme that is encoded in the genomes of most eukaryotes, eubacteria, and archaebacteria (which have a highly divergent version of the gene) and has been well studied at the primary, secondary, and tertiary structural levels [3-5]. Except for the mammalian MAT II, which is a hetero-oligomer [6], members of the MAT family are homo-oligomers that usually form tetramers consisting of four identical subunits; the two active sites are located between the subunits in each dimer [3]. Mammalian MAT III and archaeal MATs form dimers [7].

Multiple sequence alignments of MAT genes from a wide diversity of eukaryotes demonstrated a paralogue of MAT, named MATX, with distinctive features that are absent in all other eukaryotic MATs. These features include four specific insertions and a large number of unique substitutions [8]. The recombinant MATX from *Euglena gracilis* has been found to function as a homodimer with activities comparable to MATs from other eukaryotes [9]. Molecular phylogenetic analyses clearly showed that MATX is related to other eukaryotic MATs, but it forms a long branch in the eukaryotic subtree [8]. The majority of MATX paralogues occur in four distantly

* Correspondence: janca.sz@centrum.cz; vladimir.hampl@natur.cuni.cz
[1]Department of Parasitology, Charles University in Prague, Faculty of Science, Vinicna 7, Prague 2 128 44, Czech Republic
[2]Biotechnology and Biomedicine Center of the Academy of Sciences and Charles University in Vestec, Prague, Czech Republic
Full list of author information is available at the end of the article

related groups of photosynthetic eukaryotes: haptophytes, photosynthetic euglenids, diatoms, and dinoflagellates. MATX was also detected in a pelagophyte alga *Aureococcus anophagefferens* [10]. All organisms possess either the MAT or the MATX form of the gene, with the exception of five diatom species that have both paralogues and *A. anophagefferens* that harbors two different homologues of MAT in addition to MATX [8,10].

A similar punctate distribution of two paralogues with the same function was reported for "elongation factor 1-alpha" (EF-1α) and its paralogue "elongation factor like" (EFL), which are highly conserved members of a GTPase superfamily involved in translation. Like MAT/MATX, the EF-1α/EFL paralogues have a patchy distribution across the tree of eukaryotes and rarely occur together in the same organism. EFL has been localized so far in eight groups of unrelated organisms: dinoflagellates, haptophytes, cercozoans, green algae, choanoflagellates, fungi, diatoms, and radiolarians [11-17].

The punctate distributions of MAT/MATX and EF-1α/EFL across the tree of eukaryotes can be explained by two scenarios: (1) a deep paralogy, whereby both paralogues were present in an ancient common ancestor followed by differential loss of one or the other paralogue in descendant lineages; and (2) a horizontal (syn., lateral) gene transfer (HGT), whereby a more recent origin of one paralogue (most likely the less frequent one, such as MATX) in one lineage of eukaryotes is followed by the spread of this paralogue to other distantly related lineages via horizontal transfer.

These scenarios differ in their assumptions. The first scenario hypothesizes coexistence and probably co-expression of both paralogues in one cell for a long time without negative effects on the organism. This scenario explains the distribution purely by vertical transmission. In this case, MATX must have originated by gene duplication from the MAT already present in the common ancestor of all MATX containing taxa. This organism was very ancient and not very distantly related, maybe identical, to the most recent common ancestor of eukaryotes. Since that time, MAT and MATX must have been propagated side by side in the genomes of the descendants to much more recent nodes of eukaryotic evolution and in some cases (diatoms) even to extant organisms.

The second scenario assumes that one (MATX) can be horizontally transferred and is capable of functional replacement of the MAT form soon after the transfer. Our previous work on the model systems of *Euglena gracilis* and *Trypanosoma brucei* indicates that MATX fulfills the assumptions for both of these scenarios, because this paralogue can be co-expressed with MAT and can immediately take over its function [18]. By contrast, EFL was capable of long-term co-expression, but was not able to functionally replace EF1-α. Based on these results, neither of the two evolutionary scenarios can be refuted for MAT/MATX. However, in the case of EF1-α/EFL, HGT is apparently more difficult and likely played a less important role in the evolutionary history of this paralogue couple [18].

There are several questions associated with the putative HGT explanation for the origin and distribution of the MATX paralogue that remain unanswered. For instance, under what circumstances would the highly divergent MATX evolve within one recent group of eukaryotes and in which lineage could it happen? One hypothesis posits that MATX evolved during a secondary endosymbiotic origin of plastids from the endosymbiont copy of the MAT gene, which was released from purifying selection and underwent accelerated sequence evolution [8]. Therefore, an analysis of the distribution of MAT/MATX in euglenids provides an opportunity to evaluate this possibility.

The Euglenida is a large group of marine and freshwater eukaryotic flagellates with diverse modes of nutrition, including phagotrophy, osmotrophy, photoautotrophy, and a recently discovered example of mixotrophy (a euglenid capable of both phagotrophy and photosynthesis) [19,20]. Photosynthetic and secondarily osmotrophic euglenids (i.e., colorless euglenids that have lost photosynthesis) form a monophyletic group that is the sister lineage to the mixotrophic *Rapaza viridis* and is nested within a paraphyletic assemblage of phagotrophic euglenids. It is inferred that the secondary chloroplast was gained through secondary endosymbiosis in the most recent common ancestor of all photosynthetic euglenids, including *R. viridis* [19-22]. The marine flagellate *Pyramimonas* (Pyramimonadales, Prasinophyta) is inferred to be the closest known relative of the euglenid chloroplasts (Turmel et al. 2009). In this study, we investigated the distribution of MAT and MATX in euglenids and *Pyramimonas* in order to evaluate whether the origin of MATX occurred simultaneously with the secondary endosymbiotic origin of the euglenid chloroplast. These data were also expected to provide insights into whether euglenids were the first group of eukaryotes to evolve the MATX paralogue.

Results

MAT and MATX phylogeny and distribution of MATX in euglenids

We generated six new sequences of MAT and 20 new sequences of MATX. The MAT sequences were obtained from heterotrophic euglenids (*Petalomonas cantuscygni* and *Distigma* sp.), the mixotroph *Rapaza viridis*, two photoautotrophic euglenids (*Phacus orbicularis* and *Monomorphina pyrum*) and the prasinophyte alga *Pyramimonas parkeae*. The MATX sequences were obtained from all investigated photoautotrophic euglenids, except *Rapaza viridis* (Table 1). The sequences retrieved from transcriptome projects were complete; sequences amplified

Table 1 Sources of sequences applied in this study

Taxon	Protein MAT/MATX	SSU
Euglena clara	† supplement	AJ532423.1*
Euglena stellata	† supplement	AF150936.1*
Euglena gracilis	† supplement	AY029409.1*
Euglena hiemalis	† supplement	DQ140157.1*
Euglena proxima	† supplement	EU624027.1*
Euglena viridis	† supplement	AJ532415.1*
Euglenaria anabaena	† supplement	AF242548.1*
Eutreptiella braarudii	† supplement	AJ532397.1*
Eutreptiella gymnastica	▲ KF383289	▲ KF559331
Distigma sp.	▲ KF383287	
Eutreptia viridis	† supplement	AF157312.1*
Lepocinclis tripteris	† supplement	AF286210.1*
Lepocinclis playfairiana	† supplement	KF267871*
Monomorphina aenigmatica	▲ KF383291	AF283313.1*
Monomorphina parapyrum	† supplement	AF112874
Monomorphina pyrum	▲ KF383286 MAT ▲ KF383290 MATX	▲ KF559330
Phacus inflexus	† supplement	FJ719629.1*
Phacus orbicularis	† supplement	AF283315.1*
Pyramimonas parkeae	▲ KF383285	
Rapaza viridis	▲ KF383288	AB679269.1*
Trachelomonas ellipsoidalis	† supplement	DQ140135.1*
Trachelomonas sp.	▲ KF383292	AJ532447.1*
Trachelomonas volvocina	† supplement	AF096995.1*
Strombomonas accuminata	† supplement	EU624029.1*

The sequences downloaded from GenBank are marked by *; sequences obtained by Sanger sequencing method in this study are marked by ▲, sequences obtained from transcriptome projects sequenced by Roche 454 sequencing were marked by † and are available in supplement.

from cDNA (*Pyramimonas parkeae, Trachelomonas* sp., *Distigma* sp., *Monomorphina aenigmatica* and *Monomorphina pyrum*) were partial (approximately 430 amino acids). We found additional so far unnoticed partial MATX homologues in GenBank from the haptophyte *Prymnesium*, the plant *Lactuca serriola* and the beetle *Dendroctonus frontalis*. Further database searches revealed that *Lactuca* and *Dendroctonus* also contain the MAT paralogue. The presence of the MATX paralogue in the single species of plant and metazoa is highly suspicious, and we treat this data with caution because we cannot exclude the possibility of contamination by foreign RNA in the *Lactuca* and *Dendroctonus* transcriptome data sets. The MAT sequences of *Rhodomonas* sp., *Rhodomonas salina, Thalassionema sp*. and *Peranema trichophorum* and the MATX sequence of *Karenia brevis* retrieved from GenBank were also incomplete. Despite their incompleteness, all MAT and MATX sequences were suitable for determining the paralogue type and for phylogenetic analyses; therefore, all sequences were added to the alignment with published MAT/MATX sequences for phylogenetic analysis (Figure 1).

In the phylogenetic tree (Figure 1), MATX paralogues formed a well-supported clade that was separated from the MAT paralogues by a long stem. The tree was rooted by five bacterial outgroups within the MAT paralogues, with *Trichomonas vaginalis* MAT being the most basal branch. However, the backbone topology of the MAT tree was weakly supported, and the MATX branch was situated only one node apart from prokaryotes. We used Kishino Hasegawa (KH), weighted KH (WKH), Shimodaria Hasegawa (SH) and weighted SH (WSH) tests to evaluate whether the root position between MAT and MATX paralogues is significantly worse than the suggested root on the *T. vaginalis* branch. The tests showed that this root position cannot be excluded ($p = 0.076$ for KH and WKH, $p = 1.00$ for SH and $p = 0.945$ for WSH).

The MATX sequences from photoautotrophic euglenids formed a well-supported subclade (bootstrap 77%) within the more inclusive MATX clade and branched as the sister group to a clade consisting of *Lactuca*, dinoflagellates and *Dendroctonus*. The MAT sequences from the heterotrophic euglenids clustered together with kinetoplastids; the MAT sequence from *P. parkeae* branched together with other green algae; and the MAT sequences from *M. pyrum* and *P. orbicularis* clustered with ciliates and *Aureococcus*, respectively.

We also performed an independent analysis of MATX sequences that enabled us to use more alignment positions to reconstruct the phylogenetic relationships within the MATX clade (Figure 2). The tree was rooted with the branch of diatoms, haptophytes and *Aureococcus* according to Figure 1.

Comparison of MATX and SSU rRNA gene phylogeny

We investigated whether or not the phylogeny of the MATX paralogues differs significantly from the species phylogeny. Significant differences would indicate that MATX has not evolved vertically but instead experienced HGTs between the MATX containing taxa. As "species trees", we have used topologies inferred from small subunit (SSU) rRNA gene sequences and also manually constructed topologies reflecting current view of species relationships. The SSU rRNA gene tree and manual species topologies differed in minor details and they are reported in Additional file 1 and in Additional file 2: Figure S1 and Additional file 3: Figure S3. We used the KH and SH tests to compare the species topologies with the best MATX topology and the set of 500 bootstrap topologies calculated from MATX alignment (Table 2). The tests showed that the "species topologies" are strongly rejected (p value = $< 7*10^{-6}$). To be sure that the conflict with the

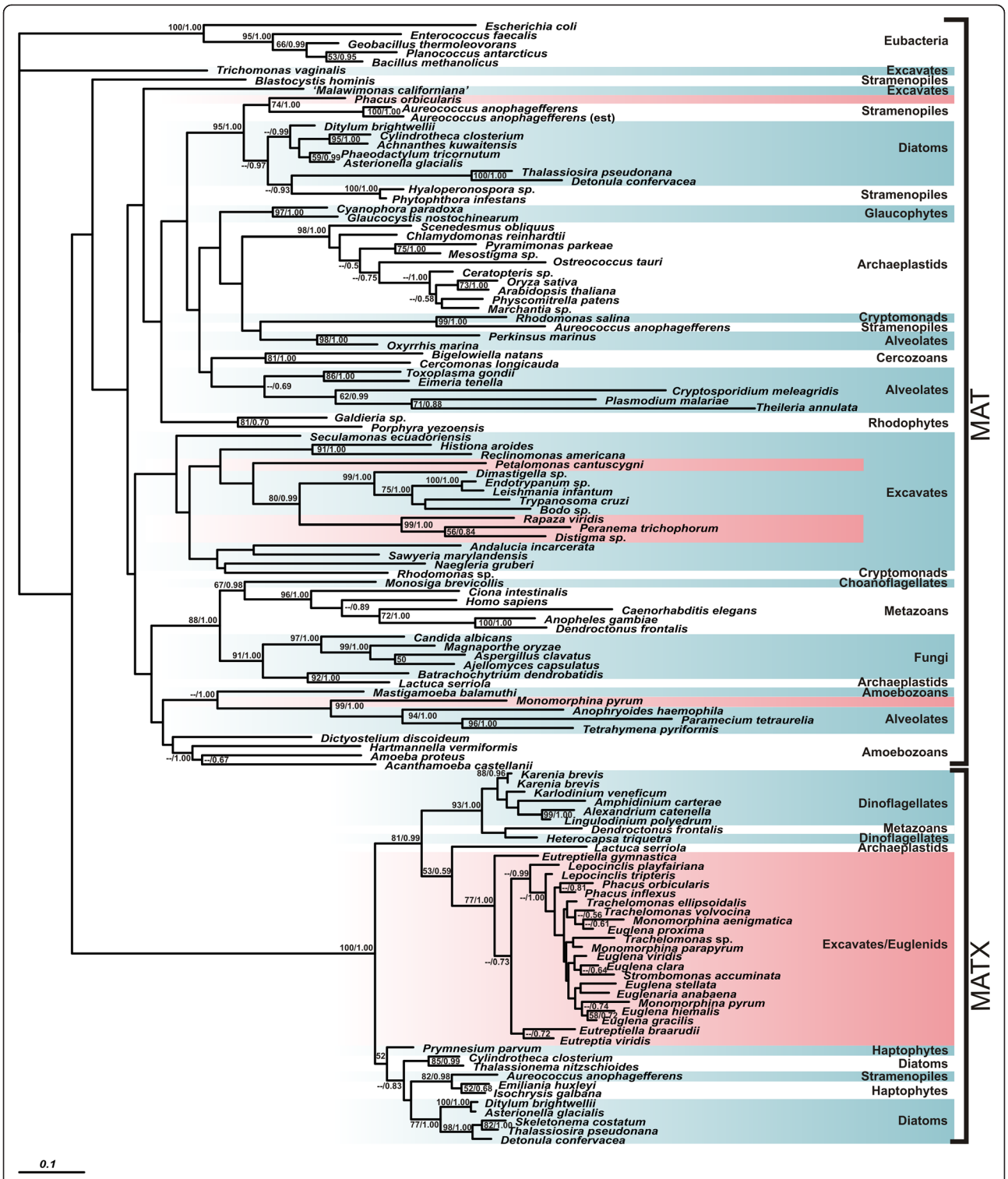

Figure 1 Maximum likelihood phylogeny of MAT and MATX. The tree was constructed by maximum likelihood method in RAxML from the 347 amino acid positions. The values at nodes represent maximum likelihood bootstraps/Bayesian posterior probabilities; only values above 50% and 0.5, respectively, are shown. Euglenid taxa are marked in red.

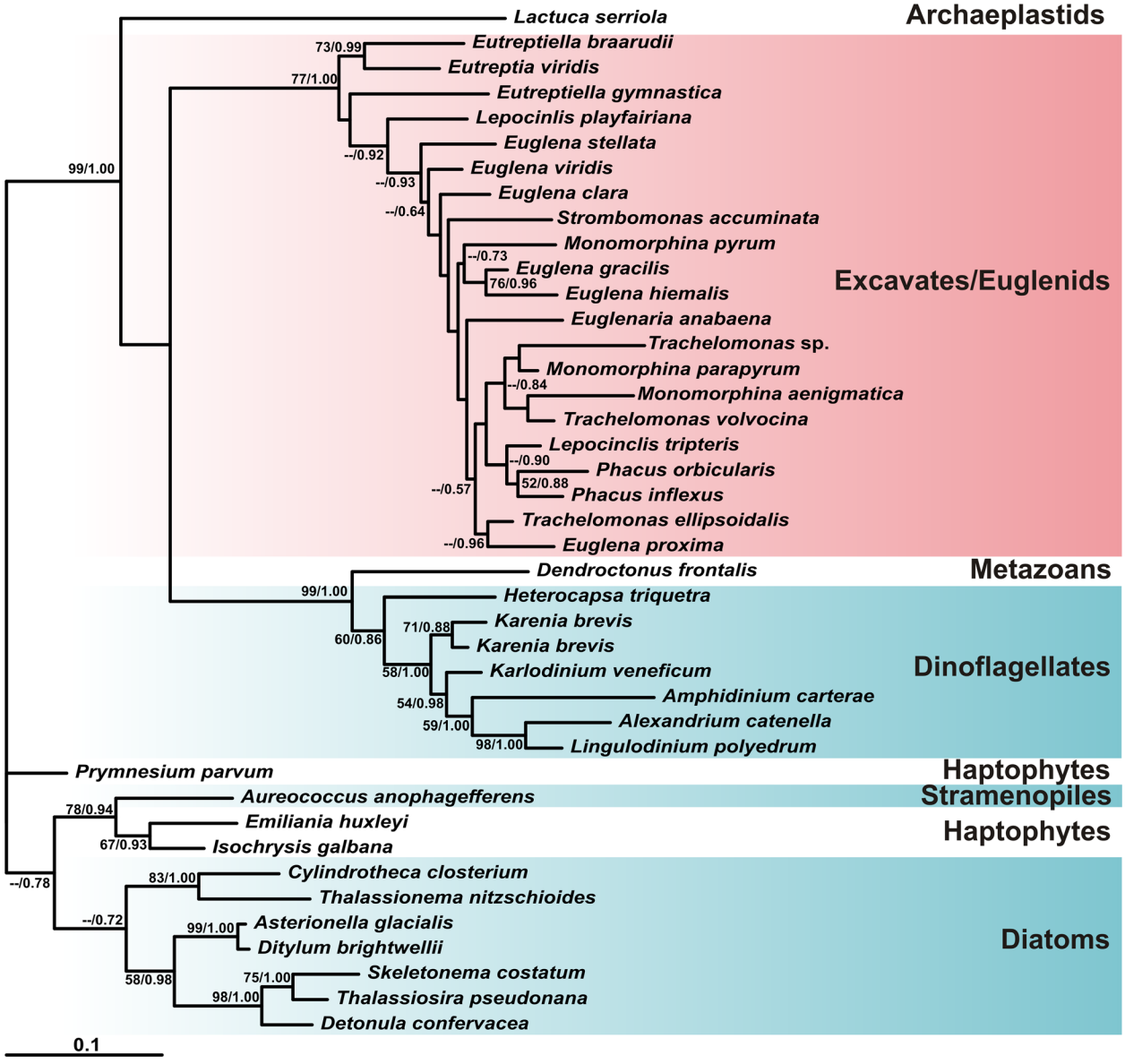

Figure 2 Maximum likelihood phylogeny of MATX clade. The tree was constructed by maximum likelihood method in RAxML from the 392 amino acid positions. The values at nodes represent maximum likelihood bootstraps/Bayesian posterior probabilities; only values above 50% and 0.5, respectively, are shown.

SSU rRNA gene tree topology is not caused only by the *Lactuca*, *Dendroctonus* and *Aureococcus* MATX sequences, whose origin is dubious, and *Prymnesium*, the sequence of which is very incomplete, we repeated the tests after exclusion of these four taxa. The "species topologies" were again rejected ($p = < 2*10^{-4}$). The "species topologies" were significantly excluded also if we compared topologies rooted by *Trichomonas* and *Escherichia*, although the significance was lower ($p = < 0.001$).

Similarly we compared the MATX topology (Additional file 4: Figure S2) with the SSU rRNA gene tree (Additional file 3: Figure S3) and manual species topologies (Additional file 1) of the subclade of photosynthetic euglenids. In this case, the tests showed that the euglenid "species topologies" cannot be rejected ($p > = 0.003$).

Discussion

Distribution of MAT and MATX paralogues in euglenids

Some genes are dispersed across the tree of eukaryotes in a punctate pattern, which means that they are present in unrelated taxa and absent in interspersed lineages. This observation suggests that the evolution of these genes was complicated and may involve events like gene duplications (the origin of paralogues), horizontal gene transfers, and

Table 2 Results of topology tests

	KH	WKH	SH	WSH
MATX (1)	**0/0**	**0/0**	**$7*10^{-6}$/0**	**0/0**
MATX excl. APLD (2)	**0/0**	**0/0**	**$1*10^{-4}$/$2*10^{-4}$**	**$5*10^{-5}$/$4*10^{-6}$**
MATX rooted (3)	**$8*10^{-6}$/0**	**$1*10^{-5}$/0**	**0.001/$2*10^{-5}$**	**$1*10^{-4}$/0**
MATX rooted exl. APDL (4)	**0/0**	**0/0**	**0.001/0.001**	**$2*10^{-4}$/$2*10^{-4}$**
MATX euglenids (5)	0.004/0.004	0.003/0.003	0.25/0.246	0.209/0.172

The p-values of significance for differences between likelihoods of MATX gene tree vs. likelihoods of species trees. In each cell are given p-values using species tree inferred from phylogeny of SSU rRNA/species tree based consensus from a literature. The tests were performed for five sets of taxa: (1) full MATX data set, (2) MATX excluding *Aureococcus*, *Prymnesium*, *Lactuca* and *Dendroctonus* (excl. APLD), (3) rooted full MATX data set, (4) rooted MATX excl. APLD and (5) MATX of euglenids. Four tests were used: Kishino Hasegawa (KH), weighted Kishino Hasegawa (WKH), Shimodaria Hasegawa (SH), weighted Shimodaria Hasegawa (WSH). P-values = < 0.001 are given in bold.

gene losses. Deciphering the history of such a gene is often difficult. Two of the most enigmatic examples are (1) elongation factor 1-alpha (EF-1α) and its paralogue elongation factor-like (EFL) and (2) methionine adenosyl transferase (MAT) and its paralogue MATX [8,11]. In both cases, these essential genes come in two paralogues that exhibit a patchy distribution among eukaryotes and are mutually, almost strictly, exclusive in their occurrence. We considered two scenarios to explain the possible evolution of the distribution of MAT and MATX: (A) a deep paralogy scenario and (B) a horizontal gene transfer scenario. MAT and MATX gene histories in euglenids according to these two scenarios are shown in Figure 3.

We detected MATX only in photoautotrophic euglenids. *Rapaza viridis*, which contains secondary chloroplasts and represents the earliest diverging lineage within the photoautotrophic clade, apparently possesses only the MAT form of the gene; the same holds for the heterotrophic euglenids (*Petalomonas*, *Distigma* and *Peranema*) and *Pyramimonas parkeae*, which contains the closest known relative of the euglenid chloroplast. Therefore, our results suggest that MATX is specific for the clade of photoautotrophic euglenids after the split of *Rapaza*. We also found two exceptions within the clade of photoautotrophic euglenids; *P. orbicularis* and *M. pyrum* both possess the MAT and MATX paralogues in their cDNAs, so both genes are transcribed in these species. The MATX form in these two species is located within the MATX clade with other photoautotrophic euglenids, while the MAT form is unrelated to euglenid MATs; the MAT of *P. orbicularis* branches together with the MAT sequences from *Aureococcus*, and the MAT in *M. pyrum* branches together with the MAT sequence from ciliates. These facts are most likely explained by two independent horizontal gene transfers of MATs from two different sources into two different lineages of euglenids.

Evolution of the MAT and MATX paralogues

We will focus on how well the observed data fit within the context of the two alternative hypotheses for the evolution of MAT and MATX in euglenids in particular and eukaryotes in general: (A) the deep paralogy scenario and (B) the horizontal gene transfer scenario (Figure 3). Let us first suppose that the deep paralogy scenario (Figure 3A) is correct. This scenario requires at least four independent losses of the MATX gene to explain its distribution in euglenids and many more losses of MATX to explain its distribution within the tree of eukaryotes. Gene losses are frequent events and many losses are not in themselves unlikely. Slightly suspicious, however, is the discrepancy in the number of MAT losses versus the number of MATX losses in this scenario. MAT was lost in euglenids (and within the Euglenozoa) only once, while MATX was lost at least four times only within euglenids. A similar disproportion of losses is present in the tree of eukaryotes. If we compare the MAT/MATX history to the case of EF-1α/EFL, the discrepancy is not as significant in the EF-1α/EFL case; the occurrence of EFL is more fragmented not only in euglenids but also in other eukaryotic groups [15-17,23]. To our knowledge, it is impossible to evaluate the significance of the observed disproportion between the number of losses of one paralogue compared to the other, so we must conclude that in this respect our observations do not contradict the deep paralogy scenario.

Moreover, if the deep paralogy scenario is correct (Figure 3A), then we would expect both paralogues MAT and MATX to be present in the most recent common ancestor of all MATX-containing taxa, which is likely identical to the most recent common ancestor of eukaryotes. If so, then we would expect that the root of the tree in Figure 1 will be positioned between the MAT and MATX lineages. This is true for EF-1α/EFL tree [11]. In the case of MAT/MATX, the bacterial outgroups form the sister branch to MAT of *Trichomonas vaginalis*, and the MATX clade is positioned within the MAT lineages. However, the bootstrap values supporting the backbone of the MAT/MATX tree are very low (Figure 1), and the root position on the MATX branch was not rejected by the statistical tests. In this respect our data do not contradict the deep paralogy scenario.

The deep paralogy scenario also assumes that the two paralogues can be co-expressed together in one organism.

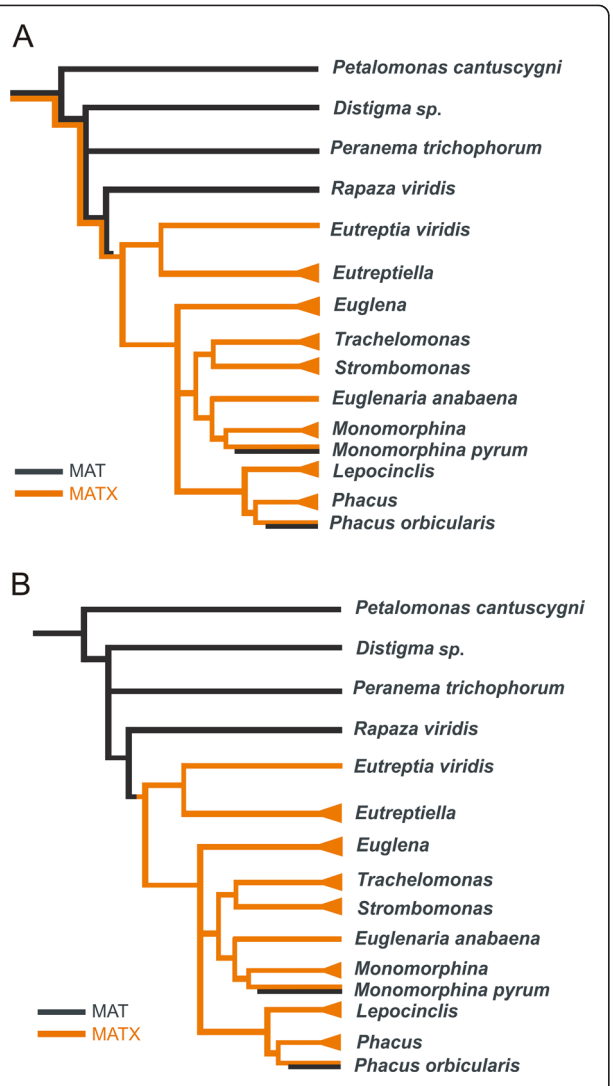

Figure 3 Schematic trees illustrating two possible scenarios of MAT/MATX evolution mapped on the currently accepted phylogenetic relationships of euglenids. The presence of MATX is marked with orange color and MAT is colored with black. **(A)** Scenario involving deep paralogy followed by differential losses. **(B)** Scenario involving horizontal gene transfer.

The observation that the two paralogues are simultaneously present in the transcriptomes of two different euglenids (*P. orbicularis* and *M. pyrum*), five diatoms, and *Aureococcus* [10] demonstrates that this is indeed possible. Moreover, we have confirmed this fact experimentally on the model system of *Euglena gracilis* and *Trypanosoma brucei* [18]. In this respect the data do not contradict the deep paralogy scenario.

Finally, the deep paralogy scenario expects that the relationships between the eukaryotic groups in the MATX part of the tree will correspond to the accepted eukaryotic phylogeny, because the gene, despite being lost in many lineages, has evolved vertically. This is apparently not true, because MATX sequences in dinoflagellates form a relatively robust sister branch to MATX sequences in euglenids (bootstrap = 81%), even though dinoflagellates are in fact more closely related to apicomplexans, ciliates, stramenopiles (including diatoms) and haptophytes. More importantly, the conflict between the global MATX phylogeny and the species phylogeny of the MATX containing taxa was significant in statistical tests. Within the clade of photoautotrophic euglenids, the MATX phylogeny also differed from species tree, but this difference was not significant. In this last respect, therefore, our data do contradict the scenario of deep paralogy followed by differential losses in its purest form. In order to explain this observation, we must invoke either horizontal gene transfers within the MATX clade or at least two more gene duplications and subsequent differential losses of putative paralogues within the MATX clade. The latter case would assume that some ancestral organisms would harbor at least four paralogues of this enzyme, which is inconsistent with the observation that most extant species contain only one paralogue (see Additional file 1); therefore, we conclude that MATX has not evolved vertically.

Let us now suppose that the horizontal gene transfer scenario is correct. The first assumption of this scenario is that the MATX paralogue is capable of horizontal transfer. The ability of the MATX paralogue to substitute the function of MAT has been proven experimentally in *E. gracilis* and *T. brucei* [18]. In this study, we have also revealed two relatively clear cases of MAT horizontal transfers from different sources into *P. orbicularis* and *M. pyrum*. In order to explain the distribution of MATX in euglenids through HGT, we only require a single horizontal gene transfer shortly after *Rapaza viridis* split from the other photoautotrophic euglenids (Figure 3B); only a few more horizontal gene transfers would be necessary to explain the distribution of MATX in all eukaryotes. Taken together, the data suggests that MATX is capable of HGT and the number of required events is low. In this respect, the data do not contradict the horizontal gene transfer scenario.

The second assumption of the HGT scenario is that there was a eukaryotic group in which the MATX first evolved and then subsequently spread into other lineages of eukaryotes. Such a group would ideally appear as a paraphyletic assemblage near the very base of MATX clade. At the same time, the root of the MAT/MATX tree would be situated inside the MAT paralogues. The data collected so far do not suggest any source group, because the taxa with MATX either form monophyletic groups (e.g., euglenids and dinoflagellates) or have unclear phylogenetic positions (e.g., diatoms, haptophytes and *Aureococcus*). Our working hypothesis that the MATX originated during the secondary endosymbiotic origin of the euglenid chloroplast ([8]) is not supported by the fact

that the MATX paralogue is absent in both *Rapaza viridis* and the closest relative of the euglenid chloroplast, *Pyramimonas*. Moreover, the MATX paralogues in euglenids do not form a paraphyletic group, but instead form a robust clade within the more inclusive MATX clade. The position of the root between MAT and MATX lineages cannot be rejected, and both paralogues might have been present in the common ancestor of all eukaryotes. The current data are in this respect not in direct conflict but, at the same time, they are also not supportive of the horizontal gene transfer scenario.

Conclusions

Our data are not entirely consistent with either of the two scenarios for MAT/MATX evolution in their purest forms. The hypothesis of deep paralogy followed by differential losses is rejected by the fact that MATX did not evolve purely by vertical transmission. The hypothesis of a more recent origin of MATX followed by spread via horizontal gene transfers is complicated by the absence of a source of the first MATX paralogue and the fact that both paralogues could be present in the most recent common ancestor of all eukaryotes. Therefore, we infer that the MATX paralogue spread among eukaryotes via HGT; however, the original source of MATX is not yet known and it could originate by gene duplication from MAT in the last eukaryotic common ancestor.

We also infer that euglenids were not the group in which the MATX paralogue evolved. Instead, a foreign MATX paralogue substituted the ancestral euglenid MAT paralogue in a single horizontal gene transfer event that occurred after the secondary endosymbiotic origin of the euglenid chloroplast (Figure 3B). Although the donor of the euglenid MATX paralogue is not known, the MATX paralogue, once established, may have evolved vertically within the clade of photoautotrophic euglenids. Two photoautotrophic euglenids (*P. orbicularis* and *M. pyrum*) regained a new version of the MAT paralogue by recent horizontal gene transfers from two different eukaryotic lineages and now contain both paralogues. Overall, the case study of MAT/MATX illustrates the complex evolutionary histories of some eukaryotic genes and highlights the prevalence of gene duplications, differential losses of paralogues, and horizontal gene transfer events during the course of eukaryotic evolution.

Methods

Euglenid strains and culture conditions

All cultures used in this study are listed in Table 1. Strains of *Eutreptiella gymnastica* (SCCAP K-0333), *Trachelomonas* sp. (SCCAP K-1380) and *Pyramimonas parkeae* (SCCAP K-0007) were obtained from the Scandinavian Culture Collection of Algae and Protozoa (SCCAP). Strains of *Monomorphina pyrum* (CCAP 1261/4B) and *Monomorphina aenigmatica* (CCAP 1261/9) were obtained from the Culture Collection of Algae and Protozoa (CCAP). *Distigma* sp. was isolated from samples collected from freshwater sediment from Czech Republic (50°27'N, 13°20'E). This culture was not monoeukaryotic and contained various other protists, therefore, we used a method of single cell cloning by serial dilution to obtain a monoclonal *Distigma* sp. culture. *Rapaza viridis* was isolated and cultured from marine sediment samples from Canada (48° 47.551' N, 125° 06.974' W) [20]. *Euglena clara* (SAG 25.98), *Euglena gracilis* (SAG 1224-5/25), *Euglena proxima* (SAG 1224-11a), *Eutreptia viridis* (SAG 1226-1c), were obtained from the Culture Collection of Algae at Goettingen, Germany. *Euglena stellata* (UTEX 372), *Trachelomonas volvocina* (UTEX 1327), *Monomorphina parapyrum* (UTEX 2354) and *Euglenaria anabaena* (UTEX 373) were obtained from the Culture Collection of Algae at the University of Texas, Austin Texas, USA. *Euglena viridis* (ATCC PRA110) was from the American Type Culture Collection, Manassas, Virginia, USA and *Eutreptiella braarudii* (CCMP 1594) was obtained from the National Center for Marine Algae and Protozoa, East Boothbay, Maine, USA. *Phacus inflexus* (ACOI 1336) and *Phacus orbicularis* (ACOI 996) were obtained from the Coimbra Collection of Algae, Coimbra, Portugal. Culture of *Petalomonas cantuscygni* (CCAP 1259/1) was provided by Dr. Mark Farmer at the University of Georgia, Athens, Georgia, USA and it was originally obtained from the Culture Collection of Algae and Protozoa. *Strombomonas accuminata* NJ, S 716 and *Trachelomonas ellipsoidalis* NJ, ST1 are cultures maintained in the Triemer lab which were originally isolated from pond samples from New Jersey, USA; *Lepocinclis tripteris* MI 101 and *Lepocinclis playfairiana* MI 102 are cultures isolated from ponds near Michigan State University, East Lansing, MI, USA.

DNA, RNA isolation and preparation of cDNA

Genomic DNA from *Eutreptiella gymnastica*, *Trachelomonas* sp., *Pyramimonas parkeae*, *Monomorphina pyrum, Monomorphina aenigmatica,* and *Distigma* sp. was extracted from strains using the Qiagen Blood and Tissue kit and total RNA was isolated from 150 ml of well-grown cultures (approx. $25*10^6$ cells) using TRIzol Reagent (Invitrogen). Total RNA from *Rapaza viridis* was isolated using Ambion® RNAqueous-Micro Kit (Life technologies). mRNA was purified from total RNA with the use of Dynabeads mRNA Purification Kit (Invitrogen). cDNA was then prepared using Smarter PCR cDNA Synthesis Kit (Clontech) according to the manufacturer's protocol with 15 to 27 cycles of cDNA amplification (depending on the amount of mRNA used in the first-strand synthesis).

In case of *E. gracilis, M. parapyrum, S. accuminata* and *L. playfairiana* the total RNA was extracted by grinding

wet biomass in liquid nitrogen followed by purification using RNA/DNA Maxi Kit (Qiagen); mRNA, whenever used for cDNA synthesis, was purified from total RNA using Qiagen Oligotex mRNA Maxi Kit. cDNA was prepared using Smart (later Smarter) cDNA synthesis Kit (Clontech) or by similar technology provided by MINT cDNA synthesis Kit (Evrogen). cDNA libraries were normalized using Trimmer cDNA normalization Kit (Evrogen). The resulting normalized cDNA was adapted for Roche 454 sequencing by performing a multiple last amplification step, pooling the PCR products in order to achieve the overall amount of cDNA acceptable for sequencing.

For the remaining euglenid strains, total RNA was isolated using RNAzol RT RNA Isolation Reagent (Molecular Research Center, Inc.). High level purification of total RNA was achieved using MEGAclear Kit (Ambion). Next, mRNA was isolated using MIcroPoly(A)Purist Kit (Ambion). Preparation of cDNA suitable for the next generation sequencing was according to cDNA Rapid Library Preparation Manual (Roche, GS FLX Titanium Series, later GS FLX + Series - XL+).

Amplification, sequencing and assembly

In case of *Pyramimonas parkeae, Eutreptiella gymnastica, Trachelomonas* sp., *Distigma* sp., *Monomorphina aenigmatica* and *Monomorphina pyrum* we have amplified the MAT or MATX genes from cDNA template using slightly modified primers of Kamikawa et al. [10]: Forward primer MATA3-F (5'-GAGYMMGTSAVYGARGGYCAYCCXGACAA-3') directed at the consensus amino acid (aa) sequence GHPDK and the reverse primer MATB3-R (5'-CCRTGNGCNCCCCADCCDCCRTAXGT-3') directed at the eukaryotic consensus aa sequence TYGGWGAH inside a conserved block. Amplification was carried out in 25-μl reactions with 1.5 μl of the diluted cDNA as a template using EmeraldAmp MAX PCR Master Mix (TaKaRa Bio Inc.) and the following program: a hot start at 95°C for 4 min, followed by 35 cycles of denaturation at 95°C for 30 s, annealing at 55°C for 60 s and extension at 72°C for 90 s, finishing with an extension at 72°C for 15 min. The PCR products were excised from the gel, cloned into pGEM-T Easy Vector System (Promega) and sequenced. The new sequences were deposited in GenBank under the accession numbers listed in Table 1.

Small subunit (SSU) ribosomal RNA gene from *E. gymnastica* was amplified from genomic DNA with "universal" eukaryote SSU primer pairs Medlin A (5'-CTGGTTGATCCTGCCAG-3'), Medlin B (5'-TGATCCTTCTGCAGGTTCACCTAC-3') described by Medlin et al. [24]. Amplification was carried out using the following program: a hot start at 95°C for 4 min, followed by 35 cycles of denaturation at 95°C for 30 s, annealing at 55°C for 60 s and extension at 72°C for 90 s, finishing with an extension at 72°C for 15 min. Medlin A, Medlin B, EPA-23 (5'- GTCATATGCTTYKTTCAAGGRCTAAGCC -3'), EPA-2286 (5'- TCACCTACARCWACCTTGTTACGAC -3') according to Müllner et al. [25] and our primers SSU 633-F (5'-GGCAGCAGGCRCGCAAATTGC -3') and SSU 2031-R (5'- TCAACCAGACAAATCACTYCACCAA -3') were used for sequencing of PCR products.

Small subunit (SSU) ribosomal RNA gene from *L. playfairiana* and *M. parapyrum* was amplified from genomic DNA with nuclear SSU primers 18S_1A (AAYCTGGTTGATCCTGCCAGT) and 18S_1520B (TGATCCTTCTGCAGGTTCACCTAC). Amplifications were carried out using 5 min of denaturation at 94°C and 30 cycles of the following: 94°C for 30 s, 45°C – 50°C for 1 min, 72°C for 2 min, a final extension at 72°C for 11 min. For sequencing of PCR products were used primers 18S_1A, 18S_1520B, 18S_300F (WGGGTTYGATTCCGGAG), 18S_528F (CGGTAATTCCAGCTCC), 18S_516R (ACCAGACTTGCYCTCC), 18S_960F (TTTGACTCAACRCGGG) and 18S_1055R (CGGCCATGCACCACC).

For the 454 sequences obtained from cDNAs, the raw reads (SFF File format) from 454 were filtered to remove reads shorter than 50 bp and all reads which had more than 30% of the bases with a Phred quality score less than 30 using NGS QC TK [26] were excluded. The resulting high quality reads were assembled using Roche's proprietary "Newbler" software version 2.6 with "cDNA" option. Assembled contigs shorter than 200 bp were excluded.

The full length of euglenid MATX genes were 1290 bp. Some of the sequences were incomplete: *P. orbicularis* (length 1257 bp), *M. pyrum* (length 906 bp), *M. aenigmatica* (length 843 bp), *Trachelomonas* sp. (length 909 bp) and *E. anabaena* (length 1266 bp). The length of the MAT genes were 1167 bp for *P. cantuscygni*, 1137 for *P. orbicularis*, 795 bp for *R. viridis*, 774 bp for *M. pyrum*, 765 bp for *Distigma* sp. and 720 bp for *P. parkeae*.

Phylogenetic analyses

The MAT and MATX protein sequences were aligned in ClustalX [27], the SSU rRNA gene sequences were aligned in MAFFT (http://www.genome.jp/tools/mafft/) using G-INS-I option [28]. The alignments were manually refined in BioEdit 7.0.5.3. [29]. The regions, which could not be unambiguously aligned, were excluded from the analyses.

A phylogeny of eukaryotic MAT and MATX was inferred from 123 sequences using 347 aligned amino acid positions; the phylogenetic relationships within the MATX clade were inferred from 41 sequences and 405 positions; the phylogenetic relationships within the euglenid subgroup of the MATX clade were inferred from 21 sequences and 399 alignment positions. Maximum likelihood trees were estimated by RAxML_HPC version 2.3.3 [30] using the best fitting models as determined by Prottest (http://darwin.uvigo.es/software/prottest2_server.html) [31] and

10 replicates of starting tree construction. The models were PROTGAMMALG for MAT + MATX and MATX of euglenids and PROTGAMMAWAG for analysis of eukaryotic MATX clade. Bootstrap supports (BS) were calculated from 500 replicates. Bayesian trees were estimated by MrBayes version 3.1.2 (Ronquist and Huelsenbeck 2003) using the WAG + GAMMA + Invariants + covarion model of substitution. In case of MAT + MATX analysis (Figure 1), two MCMC were run for 5 860 000 generations, trees from the first 1000 000 generations were discarded as burn-in. In case of MATX analysis (Figure 2), two MCMC were run for 17 775 000 generations, trees from the first 2 818 500 generations were discarded as burn-in.

For the purposes of topology testing, pruned and rooted data sets of MATX clade were analyzed – 40 sequences (only one *Karenia brevis* sequence was used), 36 sequences (without *Aureococcus*, *Prymnesium*, *Dendroctonus* and *Lactuca*) and both previous data sets rooted by *Trichomonas* and *Escherichia* (i.e. 42 and 38 sequences). All alignments contained 405 amino acid positions and were analysed as described above. Phylogenetic trees of SSU rDNA were inferred by maximum likelihood method from the corresponding set of taxa – 40 and 36 sequences in unrooted, 42 and 38 sequences in rooted analyses of MATX clade and 21 sequences of MATX containing euglenids. Unrooted and rooted SSU alignments contained 1525 and 1282 positions respectively. A maximum likelihood trees were estimated by RAxML_HPC version 2.3.3 [30] using the GTRGAMMA model of nucleotide substitution, 10 replicates of starting tree construction and BS were calculated from 500 replicates.

All data sets and trees generated in this study have been deposited in TreeBASE (study accession number is 15062).

Topology testing

The Kishino Hasegawa (KH) [32] and Shimodaria Hasegawa tests [33] implemented in Consel 0.1j [34] were used for topology testing. We have decided not to report the results of approximately unbiased test [35] because we have realized that the test behaves very unstably for our data sets; re-testing of the same data sets produced very different p-values that sometimes differed in significance. Regarding the significance or non-significance at the $p = 0.001$ level, the results of the AU tests were in agreement with the results of KH and SH tests in most cases; however due to their instability, we have decided to report only the results of KH and SH tests.

A set of 503 topologies was created in order to test whether the relationships between MATX paralogues are in conflict with the relationship of MATX containing taxa as inferred from SSU rDNA sequences. This set of topologies contained the best topology inferred from an analysis of the MATX protein alignment by RAxML, 500 topologies from bootstrap permutations of the MATX alignment generated by RAxML, the best tree inferred by RAxML from the SSU rRNA alignment of the same set of taxa, and the manually constructed topology reflecting the current view of species relationships. The latter two topologies representing species trees are given in Additional file 1 and in Additional file 2: Figure S1 and Additional file 3: Figure S3. Site likelihoods for topologies 1–501 were inferred by Treepuzzle 5.2. [36] using MATX gene alignment, $WAG + I + \Gamma$ model of amino acid substitution and parameter values inferred from the topology nr. 1. Site likelihoods for topologies 502 and 503 were inferred by Treepuzzle using MATX gene alignment, $WAG + I + \Gamma$ model of amino acid substitution and parameter values inferred from these topologies. The sets of site likelihoods were then compared by the KH, weighted KH (WKH), SH and SH (WSH) test in Consel 0.1j [34]. The tests were performed for (1) the full set of MATX paralogues from 40 taxa, (2) a set of MATX paralogues, excluding MATX from *Aureococcus*, *Prymnesium*, *Dendroctonus* and *Lactuca*, (3) data set 1 rooted by *Trichomonas* and *Escherichia*, (4) data set 2 rooted by *Trichomonas* and *Escherichia*, and (5) a set of MATX paralogues from euglenids.

The same tests were used to evaluate whether or not the root position between MAT and MATX paralogues can be rejected. For these tests, we used topology shown in Figure 1, 500 bootstrap topologies calculated from the same alignment, and a topology that differed from Figure 1 only in the position of prokaryotic outgroups that were moved on the branch separating MAT and MATX paralogues. The tests were performed as described above.

Availability of supporting data

All the supporting data are included as additional files.

Additional files

Additional file 1: Reconciliation of MATX gene tree with species tree. We have used the software Jane (http://www.cs.hmc.edu/~hadas/jane/) to reconcile the MATX gene tree with the species tree. For this analysis we have excluded taxa with very incomplete sequence (*Prymnesium*) or taxa, whose MATX sequences could be result of contamination (*Lactuca* and *Dencroctonus*). If we set the cost of gene loss to 0, which could be a realistic value in case of loss of one of two paralogues, then the discrepancy between MATX gene tree and species tree can be explained by the same number of events if we consider duplications and differential losses (A) or horizontal gene transfers (B).

Additional file 2: Figure S1. Maximum likelihood phylogeny of MATX containing taxa based on SSU rRNA gene. The tree was constructed by maximum likelihood method in RAxML from the 1525 nucleotide positions. The values at nodes represent maximum likelihood bootstraps, only values above 50% are shown.

Additional file 3: Figure S3. Maximum likelihood phylogeny of MATX containing euglenid taxa based on SSU rRNA gene. The tree was constructed by maximum likelihood method in RAxML from the 1525

nucleotide positions. The values at nodes represent maximum likelihood bootstraps, only values above 50% are shown.

Additional file 4: Figure S2. Maximum likelihood phylogeny of euglenid MATX. The tree was constructed by maximum likelihood method in RAxML from the 399 amino acid positions. The values at nodes represent maximum likelihood bootstraps, only values above 50% are shown.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
JS participated on cDNA preparation (for *Distigma* sp., *P. parkeae*, *E. gymnastica*, *Trachelomonas* sp., *M. pyrum*, *M. aenigmatica* and *R. viridis*), data analysis, in the sequence alignments and drafted the manuscript. NY provided *Rapaza viridis* RNA and revised the manuscript. BSL revised the manuscript. RET provided the transcriptome data for the rest of euglenid species and revised the manuscript. VH supervised the study, performed the phylogenetic analyses and helped to draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements
The work on the project was supported by the project "BIOCEV – Biotechnology and Biomedicine Centre of the Academy of Sciences and Charles University" (CZ.1.05/1.1.00/02.0109), from the European Regional Development Fund and by the Czech Science Foundation (P506/11/1320) awarded to VH and by grants to BSL from the Tula Foundation (Centre for Microbial Diversity and Evolution at the University of British Columbia) and the Canadian Institute for Advanced Research, Program in Integrated Microbial Biodiversity. RET was supported by an Assembling the Tree of Life grant from the National Science Foundation (DEB-0830056). Many of the Euglenozoan MATX sequences were generated as part of this larger project. RET would like to thank his collaborators at Virginia Commonwealth University, Dr. Gregory A. Buck (PI on the grant), Dr. Vishal N. Kopardé and Dr. Andrey V. Matveyev for their roles in generating these sequences.

Author details
[1]Department of Parasitology, Charles University in Prague, Faculty of Science, Vinicna 7, Prague 2 128 44, Czech Republic. [2]Biotechnology and Biomedicine Center of the Academy of Sciences and Charles University in Vestec, Prague, Czech Republic. [3]Departments of Botany and Zoology, Canadian Institute for Advanced Research, Program in Integrated Microbial Biodiversity, University of British Columbia, Vancouver, British Columbia V6T 1Z4, Canada. [4]Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824, USA.

Received: 20 August 2013 Accepted: 30 December 2013
Published: 11 February 2014

References
1. Cantoni GL: **Biological methylation: selected aspects.** *Annu Rev Biochem* 1975, **44:**435–451.
2. Chiang PK, Gordon RK, Tal J, Zeng GC, Doctor BP, Pardhasaradhi K, McCann PP: **S-adenosylmethionine and methylation.** *FASEB J* 1996, **10:**471–480.
3. Takusagawa F, Kamitori S, Misaki S, Markham GD: **Crystal structure of S-adenosylmethionine synthetase.** *J Biol Chem* 1996, **271:**136–147.
4. Graham DE, Bock CL, Schalk-Hihi C, Lu ZJ, Markham GD: **Identification of a highly diverged class of S-adenosylmethionine synthetases in the archaea.** *J Biol Chem* 2000, **275:**4055–4059.
5. Gonzalez B, Pajares MA, Hermoso JA, Alvarez L, Garrido F, Sufrin JR, Sanz- Aparicio J: **The crystal structure of tetrameric methionine adenosyltransferase from rat liver reveals the methionine-binding site.** *J Mol Biol* 2000, **300:**363–375.
6. Kotb M, Kredich NM: **S-Adenosylmethionine synthetase from human lymphocytes purification and characterization.** *J Biol Chem* 1985, **260:**3923–3930.
7. Markham GD, Pajares MA: **Structure – function relationships in methionine Adenosyltransferases.** *Cell Mol Life Sci* 2009, **66:**636–648.
8. Sanchez-Perez GF, Hampl V, Simpson AGB, Roger AJ: **A new divergent type of eukaryotic methionine adenosyltransferase is present in multiple distantly related secondary algal lineages.** *J Eukaryot Microbiol* 2008, **55:**374–381.
9. Garrido F, Estrela S, Alves C, Sánchez-Pérez GF, Sillero A, Pajares MA: **Refolding and characterization of methionine adenosyltransferase from *Euglena gracilis*.** *Protein Expr Purif* 2011, **79:**128–136.
10. Kamikawa R, Sanchez-Perez GF, Sako Y, Roger AJ, Inagaki Y: **Expanded phylogenies of canonical and non-canonical types of methionine adenosyltransferase reveal a complex history of these gene families in eukaryotes.** *Mol Phylogenet Evol* 2009, **53:**565–570.
11. Keeling PJ, Inagaki Y: **A class of eukaryotic GTPase with a punctate distribution suggesting multiple functional replacements of translation elongation factor 1α.** *Proc Natl Acad Sci U S A* 2004, **101:**15380–15385.
12. Noble GP, Rogers MB, Keeling PJ: **Complex distribution of EFL and EF-1α proteins in the green algal lineage.** *BMC Evol Biol* 2007, **7:**82.
13. Kamikawa R, Inagaki Y, Sako Y: **Direct phylogenetic evidence for lateral transfer of elongation factor-like gene.** *Proc Natl Acad Sci U S A* 2008, **105:**6965–6969.
14. Keeling PJ, Palmer JD: **Horizontal gene transfer in eukaryotic evolution.** *Nat Rev Genet* 2008, **9:**605–618.
15. Gile GH, Faktorova D, Castlejohn CA, Burger G, Lang BF, Farmer MA, Lukes J, Keeling PJ: **Distribution and phylogeny of EFL and EF-1α in Euglenozoa suggest ancestral co-occurrence followed by differential loss.** *PLoS One* 2009, **4:**e5162.
16. Kamikawa R, Yabuki A, Nakayama T, Ishida K, Hashimota T, Inagaki Y: **Cercozoa comprises both EF-1α-containing and EFL-containing members.** *Eur J Protistol* 2011, **47:**24–28.
17. Ishitani Y, Kamikawa R, Yabuki A, Tsuchiya M, Inagaki Y, Takishita K: **Evolution of elongation factor-like (EFL) protein in Rhizaria is revised by radiolarian EFL gene sequences.** *J Eukaryot Microbiol* 2012, **59:**367–373.
18. Szabova J, Ruzicka P, Verner Z, Hampl V, Lukes J: **Experimental examination of EFL and MATX eukaryotic horizontal gene transfers: coexistence of mutually exclusive transcripts predates functional rescue.** *Mol Biol Evol* 2011, **28:**2371–2378.
19. Leander BS, Esson HJ, Breglia SA: **Macroevolution of complex cytoskeletal systems in euglenids.** *Bioessays* 2007, **29:**987–1000.
20. Yamaguchi A, Yubuki N, Leander BS: **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* 2012, **12:**29.
21. Leander BS: **Did trypanosomatid parasites have photosynthetic ancestors?** *Trends Microbiol* 2004, **12:**251–258.
22. Takahashi F, Okabe Y, Nakada T: **Origins of the secondary plastids of euglenophyta and chlorarachniophyta as revealed by an analysis of the plastid-targeting, nuclear-encoded gene psbO.** *J Phycol* 2007, **43:**1302–1309.
23. Gile GH, Novis PM, Cragg DS, Zuccarello GC, Keeling PJ: **The distribution of Elongation Factor-1 Alpha (EF-1alpha), Elongation Factor-Like (EFL), and a non-canonical genetic code in the ulvophyceae: discrete genetic characters support a consistent phylogenetic framework.** *J Eukaryot Microbiol* 2009, **56:**367–72.
24. Medlin L, Elwood HJ, Stickel S, Sogin ML: **The characterization of enzymatically amplified eukaryotes 16S like ribosomal RNA coding regions.** *Gene* 1988, **71:**491–499.
25. Müllner AN, Angeler DG, Samuel R, Linton EW, Triemer RE: **Phylogenetic analysis of phagotrophic, photomorphic and osmotrophic euglenoids by using the nuclear 18S rDNA sequence.** *Int J Syst Evol Microbiol* 2001, **51:**783–791.
26. Patel RK, Jain M: **NGS QC toolkit: a toolkit for quality control of next generation sequencing data.** *PLoS ONE* 2012, **7:**e30619.
27. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The clustalx windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucl Acids Res* 1997, **24:**4876–4882.
28. Katoh K, Asimenos G, Toh H: **Multiple alignment of DNA sequences with MAFFT.** *Methods Mol Biol* 2009, **537:**39–64.
29. Hall TA: **BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.** *Nucl Acids Symp Ser* 1999, **41:**95–98.
30. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22:**2688–2690.
31. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinformatics* 2005, **21:**2104–2105.

32. Kishino H, Hasegawa M: **Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea.** *J Mol Evol* 1989, **29:**170–179.
33. Shimodaira H, Hasegawa M: **Multiple comparisons of log-likelihoods with applications to phylogenetic inference.** *Mol Biol Evol* 1999, **16:**1114–1116.
34. Shimodaira H, Hasegawa M: **CONSEL: for assessing the confidence of phylogenetic tree selection.** *Bioinformatics* 2001, **17:**1246–1247.
35. Shimodaira H: **An approximately unbiased test of phylogenetic tree selection.** *Syst Biol* 2002, **51:**492–508.
36. Schmidt HA, Strimmer K, Vingron M, Von Haeseler A: **TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing.** *Bioinformatics* 2002, **18:**502–504.

doi:10.1186/1471-2148-14-25
Cite this article as: Szabová *et al.*: The evolution of paralogous enzymes MAT and MATX within the Euglenida and beyond. *BMC Evolutionary Biology* 2014 **14**:25.

The plastid genome of Eutreptiella provides a window into the process of secondary endosymbiosis of plastid in euglenids

Hrdá Š., Fousek J., **Szabová J.**, Hampl V., Vlček Č. (2012).

PLoS One 7(3):e33746.

OPEN ACCESS Freely available online

PLoS one

The Plastid Genome of *Eutreptiella* Provides a Window into the Process of Secondary Endosymbiosis of Plastid in Euglenids

Štěpánka Hrdá[1], Jan Fousek[2], Jana Szabová[1], Vladimír Hampl V[1]*, Čestmír Vlček[2]*

1 Charles University in Prague, Faculty of Science, Department of Parasitology, Prague, Czech Republic, **2** Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Prague, Czech Republic

Abstract

Euglenids are a group of protists that comprises species with diverse feeding modes. One distinct and diversified clade of euglenids is photoautotrophic, and its members bear green secondary plastids. In this paper we present the plastid genome of the euglenid *Eutreptiella*, which we assembled from 454 sequencing of *Eutreptiella* gDNA. Comparison of this genome and the only other available plastid genomes of photosynthetic euglenid, *Euglena gracilis*, revealed that they contain a virtually identical set of 57 protein coding genes, 24 genes fewer than the genome of *Pyramimonas parkeae*, the closest extant algal relative of the euglenid plastid. Searching within the transcriptomes of *Euglena* and *Eutreptiella* showed that 6 of the missing genes were transferred to the nucleus of the euglenid host while 18 have been probably lost completely. *Euglena* and *Eutreptiella* represent the deepest bifurcation in the photosynthetic clade, and therefore all these gene transfers and losses must have happened before the last common ancestor of all known photosynthetic euglenids. After the split of *Euglena* and *Eutreptiella* only one additional gene loss took place. The conservation of gene content in the two lineages of euglenids is in contrast to the variability of gene order and intron counts, which diversified dramatically. Our results show that the early secondary plastid of euglenids was much more susceptible to gene losses and endosymbiotic gene transfers than the established plastid, which is surprisingly resistant to changes in gene content.

Citation: Hrdá Š, Fousek J, Szabová J, Hampl V V, Vlček Č (2012) The Plastid Genome of *Eutreptiella* Provides a Window into the Process of Secondary Endosymbiosis of Plastid in Euglenids. PLoS ONE 7(3): e33746. doi:10.1371/journal.pone.0033746

Editor: Jonathan H. Badger, J. Craig Venter Institute, United States of America

Received November 23, 2011; **Accepted** February 16, 2012; **Published** March 20, 2012

Funding: This work was supported by Ministry of Education, Youth and Sport of the Czech Republic (project MSM0021620828), by Czech Science Foundation P506/11/1320 to VH, and by the Grant Agency of the Charles University 63409 to JS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: vlada@natur.cuni.cz (VH); vlcek@img.cas.cz(ČV)

Introduction

Euglenids are a relatively large group of protists that contains species with different types of feeding strategies: some euglenid species (e.g. *Rhabdomonas*) are osmotrophic and feed by pinocytosis; others developed phagotrophic apparatuses for catching bacteria (e.g. *Entosiphon*) or even eukaryotes (e.g. *Peranema*) [1,2]. One large clade of euglenids is photoautotrophic and its members bear green secondary plastids (e.g. *Euglena gracilis*). The plastid has been subsequently and independently lost in several branches within this clade (*Euglena longa*, prev. *Astasia longa*, *Euglena quartana*, prev. *Khawkinea quartana*, *Euglena hyalina*, *Euglena viridis hyalina* and *Phacus ocellatus*, prev. *Hylophacus ocellatus*) [3–5]. The phototrophic euglenids and their secondary heterotrophic descendents are classified as class Euglenophyceae [4]. Complete plastid genome sequences are known so far for only two closely related euglenid species, *Euglena gracilis* [6] and *Euglena longa* [7].

The fact that plastids are present in a single clade of euglenids favors a hypothesis that the ancestor of this clade acquired the plastid by engulfing a green alga [6,8]. Our current knowledge on the phylogeny of euglenids implies that this endosymbiotic event happened after the split of *Peranema* but before the split of *Eutreptiella* and *Eutreptia*, the basal lineages of the phototrophic clade [9]. This "plastid late" hypothesis is further indirectly supported by the fact that the autotrophic clade is derived from within the eukaryovorous euglenids; eukaryovory is regarded as the derived feeding mode in euglenids and at the same time it is a useful predisposition facilitating the engulfment of green algae [10]. The plastid of *Euglena gracilis* can be completely lost after bleaching with many enviromental and chemical agents without effect on cell viability, and this fact is also used as an argument for a relatively recent acquisition of the plastid, which has not yet been recruited for cellular functions other than photosynthesis [11]. Recent study of introns in the plastid targeting presequences also agrees with the plastid-late hypothesis [12]. An alternative but currently less-accepted plastid-early hypothesis postulates that the euglenid plastid was acquired early in the evolution of euglenids, or even in the common ancestor of euglenids and kinetoplastids (e.g. *Trypanosoma*), their nearest sister group [13,14]. The presence of genes of red algal origin in the photosynthetic *Euglena* as well as in the heterotrophic *Peranema* suggests that the lineage of euglenids might have experienced a cryptic red algal plastid endosymbiosis before the current green algal plastid was established [15].

Analyses of 70 plastidial genes and conservation of gene order on the plastid genome has pointed to *Pyramimonas* (Pyramimonadales, Prasinophyceae) as the closest extant relative of the euglenid plastid [7]. *Pyramimonas* comprises marine flagellates, suggesting that the endosymbiotic event happened in the marine environ-

ment. Although the majority of euglenids live in freshwater, the basal lineage of the autotrophic clade contains the marine species *Eutreptia* and *Eutreptiella*, corroborating the hypothesis of a marine origin of photosynthetic euglenids [4,5]. The comparative analysis of the gene content between the plastid genome of *Pyramimonas parkeae*, which encodes 110 conserved genes (81 protein and 29 RNA species) [16], and *Euglena gracilis*, which comprises 88 conserved genes (58 protein and 30 RNA species) [6], has revealed a substantial loss of genes (for example all genes of NADH-plastoquinone oxidoreductase of the plastidial respiratory chain) happening from the common ancestor of *P. parkeae* and *E. gracilis* to extant *E. gracilis*. This reduction of gene repertoire is explained as a consequence of secondary endosymbiosis, although comparable gene losses took place in the prasinophyte lineages leading to *Pycnococcus* and to the coccoid microalgae *Ostreococcus* and *Monomastix* [16]. Further gene loss in euglenids accompanying the loss of photosynthetic activity has been observed in the closely related but non-photosynthetic *Euglena longa*, which has maintained 56 conserved genes (26 protein and 30 RNA species) [7]. Despite the reduction of coding capacity of the *Euglena* plastid in comparison to that of *P. parkeae*, the size of the *E. gracilis* genome increased (143.2 vs. 101.6 kb in *P. parkeae*). The increase in the genome size should mainly be ascribed to the expansion of self-splicing introns. While *P. parkeae* features a single group II intron, the genome of the *E. gracilis* plastid contains 160 group II and group III introns (15 of which formed twintrons), which is by far the most of all known organellar genomes [17,18]. There are indications that the expansion of introns may be a feature specific to *E. gracilis* and its relatives [17,18]; however, no other plastid genome of euglenids has been completely sequenced, which would be necessary to enable comprehensive comparisons.

Here we report the complete genome sequence of *Eutreptiella gymnastica*, a member of the basal lineage of the photosynthetic clade, and phylogenetically most distant from *Euglena gracilis* – the common ancestor of *E. gracilis* and *E. gymnastica* was the common ancestor of all currently known members of the photosynthetic lineage [1,19,20]. Comparative analysis of the gene content of euglenid plastids allows relatively precisely tracing the events of gene transfers and gene losses accompanying this particular case of secondary endosymbiosis. The vast differences in intron density suggest that the expansion of introns has happened specifically in the lineage leading to *E. gracilis*.

Results and Discussion

The complete size of the circular chloroplast DNA of *Eutreptiella gymnastica* is 67 622 bp. An overview of the general features of this genome and its closest relatives is given in Table 1. The genome sequence is numbered from the first nucleotide after the second 23S RNA gene (see a physical map of chloroplast DNA – Figure 1). The organization of the genome resembles those of higher plants and algae (including *Pyramimonas parkeae*) with a large single copy region (LSC), a small single copy region (SSC) and two inverted repeats (IR). Simplified maps of plastid genomes of *Eutreptiella gymnastica*, *Euglena gracilis*, *Euglena longa*, and *Pyramimonas parkeae* are illustrated in Figure 2 for comparison.

As is apparent from the genome map (Figure 1), the SSC region is reduced (to 1055 bp), containing only one ORF of unknown function (orf248). This is not surprising, because *E. gymnastica* (like *E. gracilis*) has lost most genes usually found in the SSC region (NADH dehydrogenase complex and a few others). Two of them (rpl32 and psaC) are relocated to other sites. The large single copy region (47 528 bp) contains most genes for proteins and tRNAs.

Two regions resembling inverted repeats (IR, 6304 bp) contain one 16S rRNA gene (1463 bp), one 23S rRNA gene (2999 bp), and a 1726-bp-long sequence with unknown function that contains 2 tandem repeats – VNTR (3×11 bp and 3,4×33 bp). Between the IR copies, the 23S rRNA genes differ in three bases, while all other sequences are identical. The IR copy on the plus strand further contains an insertion of a block of genes (tRNA-Ala, tRNA-Cys, rps2, atpI, atpH, atpF, atpA, and orf372), and the IR copy on the minus strand contains the insertion of tRNA-Ile. The gene cluster of rps2, atpI, atpH, atpF, and atpA found within the IR is one of the ancestral gene clusters conserved in streptophyte and prasinophyte plastid genomes, but it is usually located in the LSC region [16]. The tRNA-Ala and tRNA-Ile genes are present also in the IR of *P. parkeae*.

The inverted repeats do not contain 5S RNA, and in fact *Eutreptiella* lacks it completely. Absence of 5S RNA was also recorded in the plastid genome of *Pyramimonas parkeae* and *Pycnococcus provasolii*, but the possibility exists that its sequence was unrecognized [16]. Interestingly, transcriptional analysis of the *E. gracilis* plastid chromosome showed that, although the genes for 5S, 23S and 16S RNA make one operon [6], the abundance of 5S RNA is much lower than the abundance of 23S and 16S RNA [21]. If 5S RNA is present but remains unrecognized in the plastid genome of *Eutreptiella*, it probably is not localized within the RNA operon, as the 16S RNA gene is very closely followed by neighboring genes. The symmetrical arrangement of tandem repeats in the non-coding part of the IRs suggests that this region may function as the origin of replication. According to the classical model [22], which has recently been challenged [23], the replication of plant and some green algal plastid genomes starts simultaneously from both IRs, and expands unidirectionally towards the SC region, forming two D-loop structures. After it passes the initiation site of the opposing D-loop, the two D-loops fuse to form Cairn-type bidirectional forks that move away from each other and meet approximately 180 degrees from the starting point. *E. gracilis* and *E. longa* plastid genomes lack IRs (Figure 2) and, so far, no model of their replication has been proposed. The origin of replication in the plastid genome of *E. gracilis* has been localized into the region of tandem repeats approximately 6 kb upstream from the extra 16S rRNA gene (Figure 2) [6,24,25]. From this site, the replication probably proceeds in both directions [6]. Because most genes are coded on the leading strand of replication, these genomes have a strikingly non-random distribution of genes. Starting from the ORI site, in one half of the circle, most genes are coded by the plus strand, and in the other half on the minus strand [6,26]. A similar situation is in *Eutreptiella*, but the switch of the coding strands is situated approximately 2/3 of the way through the circle (Figure 1 and 2).

The size of the *E. gymnastica* plastid genome is less than half of that of *E. gracilis*, though the number of conserved genes in both species is not very different (Table 1). The difference in the genome size is caused by different numbers of self-splicing introns. The genome of *E. gracilis* plastid contains 160 group II and group III introns, which is by far the most of all known organellar genomes [6,17]. The plastid genome of *Eutreptiella* apparently contains only two putative introns, and in this respect it resembles the plastid genome of *Pyramimonas parkeae*, which contains only one [16]. We have not found any sequential, structural or positional homology either between the introns of *Eutreptiella* and *Pyramimonas* or between the introns of *Eutreptiella* and *Euglena gracilis*. The first putative intron of *Eutreptiella* (1480 bp) is located in the psaA gene. This intron apparently contains an orf386 (1158 bp) that shows very weak homology to reverse transcriptases. The homology is so weak that it was revealed only after iteration in PSI-BLAST. As

Table 1. General features of euglenid and *Pyramimonas* cpDNA.

Feature	*Eutreptiella gymnastica*	*Pyramimonas parkae*	*Euglena gracilis*	*Euglena longa*
Genome size:	67 622	101 605	143 171	73 345
GC percentage:	34,32	34,7	26,13	22,41
Gene-unique loci:	91	123	96	76
Unique rRNA (count/bases):	2/8 924	2/9 086	3/15 057	3/15617
Unique tRNA (count/bases):	26/1 959	27/2 393	27/2 764	27/2122
CDS (conserved genes/all):	59/63	81/94	58/66	26/46
non-spliced (count/bases):	61/38 145	93/69 072	26/15 873	29/15 822
spliced (count/bases):	2/5 511	1/1 467	40/34 449	17/16 299
Introns (count/bases):	2/1 630	1/2 757	160/55 702	61/NA
Density (genes per kb):	0,932	0,925	0,468	0,627
Average length (excl. introns):	692	750	751	698
Coding percentage (excl. introns):	64,5	69,4	35,1	43,7
Intergenic sequences (excl.RNA):	12 614	18 720	25 535	NA
Overlaping sequences:	1 161	1 890	6 209	NA

doi:10.1371/journal.pone.0033746.t001

group II introns often encode for reverse transcriptases, which probably help with their splicing and retroposition [27,28], the homology should be taken seriously. The second intron is much shorter (152 bp), without an ORF, and is located in the rpoB gene. The sizes of both introns (excluding ORFs) are smaller than typical group II and longer that group III introns, and we have not been able to find any noticeable similarities in the secondary structure with self-splicing introns in *E. gracilis* or elsewhere. Therefore, their ability to self-splice as well as their exact boundaries should be considered only putative. Besides the orf386 in intron 1, the *Eutreptiella* plastid genome encodes three other ORFs with homology to reverse transcriptases or intron maturases. Two of them (orf291 and orf372) have no close homologues, and their evolutionary origin cannot be traced. The third (mat1) is clearly homologous to mat1 (ycf13) of *E. gracilis* and other euglenids, and in the tree (Figure S1) it forms a well supported branch (98%) with these genes. Mat1 was apparently present in the last common ancestor of euglenid plastids but interestingly this reverse transcriptase is unrelated to the single reverse transcriptase found in the plastid genome of *Pyramimonas* (orf608) (Figure S1). Mat1 is also remarkably conservative regarding its position in the genome. In almost all investigated euglenids, including relatively closely related *Eutreptia*, it is situated in the internal group III intron of the 4th intron in the psbC gene [18]. In *Eutreptiella* it is located right next to the psbC gene, which in *Eutreptiella* does not contain any intron. Mat1 was found also in the chloroplast of *E. longa*. As this organism has no psbC gene, the mat1 gene is situated in different loci [7]. The RT and X domains of *E.gracilis* and *E.longa* mat1 deviate from the consensus sequence of 34 group II intron-encoded proteins [29]. Sequence alignment of mat1 in *Eutreptiella* and *E.gracilis* shows the loss of at least two conserved domains. Comparison between the genomes of *E. gracilis*, *Eutreptiella* and *Pyramimonas* suggests that the genome of the common ancestor of euglenid plastids was intron-poor but encoded at least one reverse transcriptase (mat1). Expansion of introns is apparently a feature specific to *E. gracilis* and its relatives, as already suggested by Thompson et al. [17]. On the other hand, the small number of introns, their unusual sizes and structures and the loss of the otherwise-conserved intron in psbC indicate the suppression of introns in *Eutreptiella*. The evidence for the recent horizontal transfer of a group II intron from a cyanobacterial donor was found in the chloroplast genome of *Euglena myxocylindracea* [30]. This intron (in the psbA gene) includes ORF575, named mat4, which resembles cyanobacterial reverse transcriptases. Mat4 is also homologous to the maturase of *Pycnococcus provasolii* and *Volvox carteri* (Figure S1), which is located in an intron of the atpB gene [16].

The content of the unique protein coding genes is surprisingly similar between *Euglena gracilis* and *Eutreptiella gymnastica* plastid genomes (Figure 3). The *Eutreptiella* plastid encodes for the same photosynthetic proteins (31), transcription/translation proteins (5), ribosomal proteins (21), and maturase mat1 as *Euglena gracilis*. There are only 5 extra ORFs in *Eutreptiella* as compared with *E. gracilis* – four ORFs without strong similarity to known proteins (orf291, orf386, orf248 and orf372) and one conserved protein with homology to *P. parkeae* ycf65 (putative ribosomal protein rpl3). Similarly, only eight genes (including intron maturases mat2, roaA and orf506) are specific to *E. gracilis*. Not surprisingly, many of the shared proteins have been lost in *Euglena longa*, whose plastid has lost photosynthetic activity. Given this almost exact match of protein coding capacity of two genomes, whose last common ancestor was at the same time the last common ancestor of all known euglenid plastid genomes, we can with reasonable confidence expect that the *Eutreptiella* plastid genome also matches the coding capacity of this last common ancestor. Using the *Pyramimonas parkeae* plastid genome to represent the closest relative to the plastid endosymbiont, we may trace quite precisely the changes in the protein coding capacity of the plastid genome that took place right before and during the process of the secondary endosymbiogenesis. This coding capacity was reduced compared to *Pyramimonas* by the set of genes coding for: 10 proteins of NADH dehydrogenase complex, 2 proteins of cytochrome B_6F (petA, petN), 3 proteins of chlorophyll metabolism (ChlL, ChlN, ChlB), heme binding protein ccsA, photosystem I subunit psaI, initiation factor infA, the protease subunit of clp protease clpP, chloroplast division protein FtsH, and several conserved and non-conserved ORFs with unknown function. A BLAST search of 23,372 transcriptome sequences of *E. gracilis* in GenBank and 268 530 transcriptome sequences of *Eutrepriella* produced by us (unpublished data) revealed that transcripts for some of these proteins,

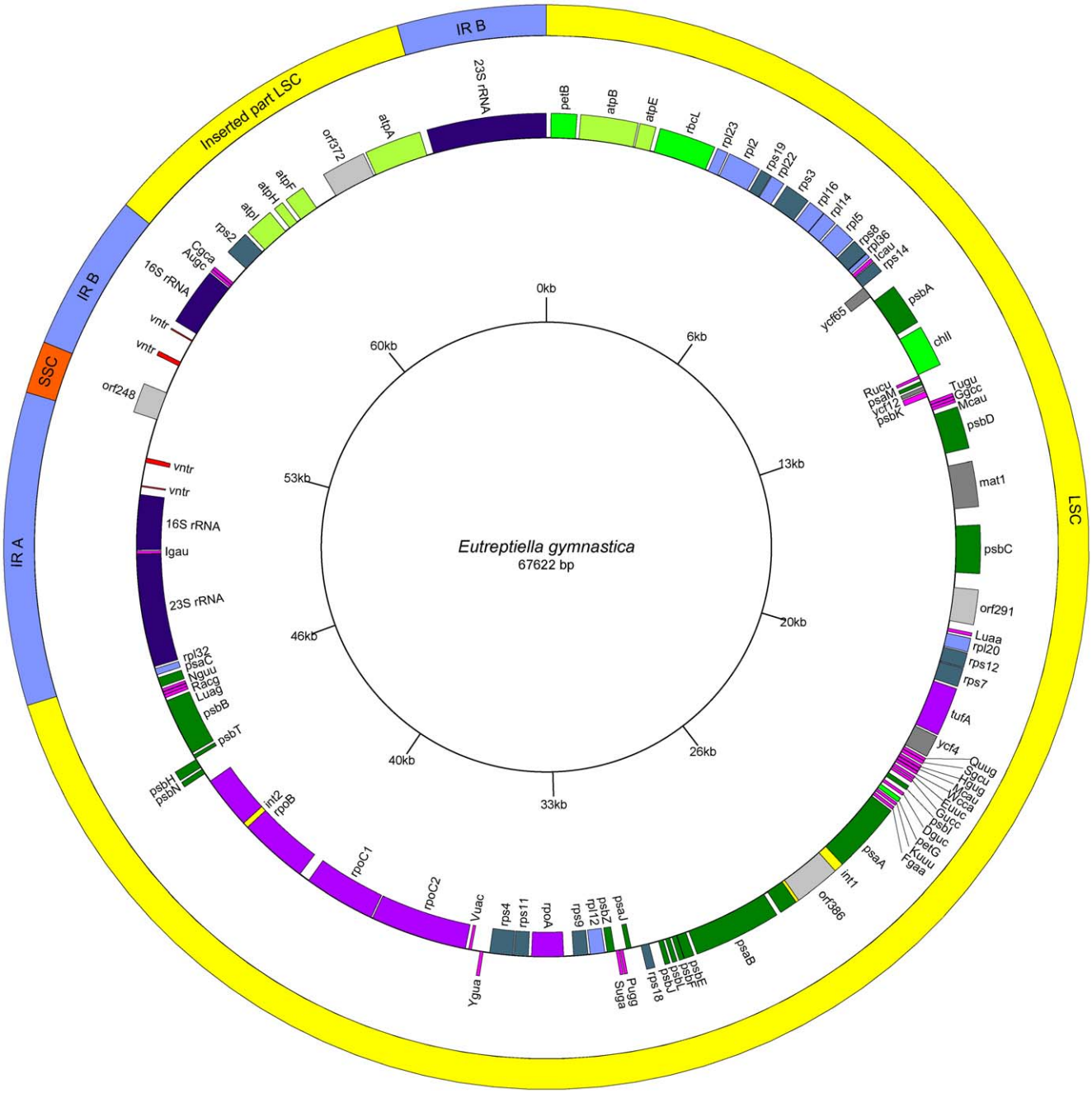

Figure 1. Map of the plastid genome of *Eutreptiella gymnastica*. Outer circle shows the large single copy region (LSC) (yellow), short single copy region (SSC) (red) and inverted repeats (IR) (blue). The inner circle shows genes and their division layout in respect to the DNA strands. The genes are color coded according to their function: photosynthesis (shades of green), translation (except maturases) (shades of blue), transcription (violet), tRNA (pink), maturases and unknown function (gray).
doi:10.1371/journal.pone.0033746.g001

namely petA, petN, ycf3, clpP, and ftsH, are present in both transcriptomes, indicating that these genes were probably transferred into the nucleus of the common ancestor of photosynthetic euglenids during the endosymbiogenesis. The gene ccsA is present only in the transcriptome of *Euglena*, suggesting that it was transferred into the nucleus of the common ancestor of photosynthetic euglenids, but retained in *Euglena* while probably lost in *Eutreptiella*. The rest of these genes were not found in any transcriptome. Although we cannot rule out the possibility that their transcripts were missed by transcriptome sequencing (e.g. due to the low abundance of transcripts), the observations here suggest that they might have been lost completely, either in the evolution of green algal ancestor of euglenid plastid after the split of the *Pyramimonas* branch, or later during endosymbiogenesis itself.

In contrast to the highly conserved gene content of *E. gracilis* and *Eutreptiella gymnastica* plastid genomes, the conservation of gene order is much lower between the two and also in comparison to *Pyramimonas*, indicating that many genome rearrangements have taken place. To get a rough estimate of the degree of gene conservation we counted the number of neighboring gene pairs

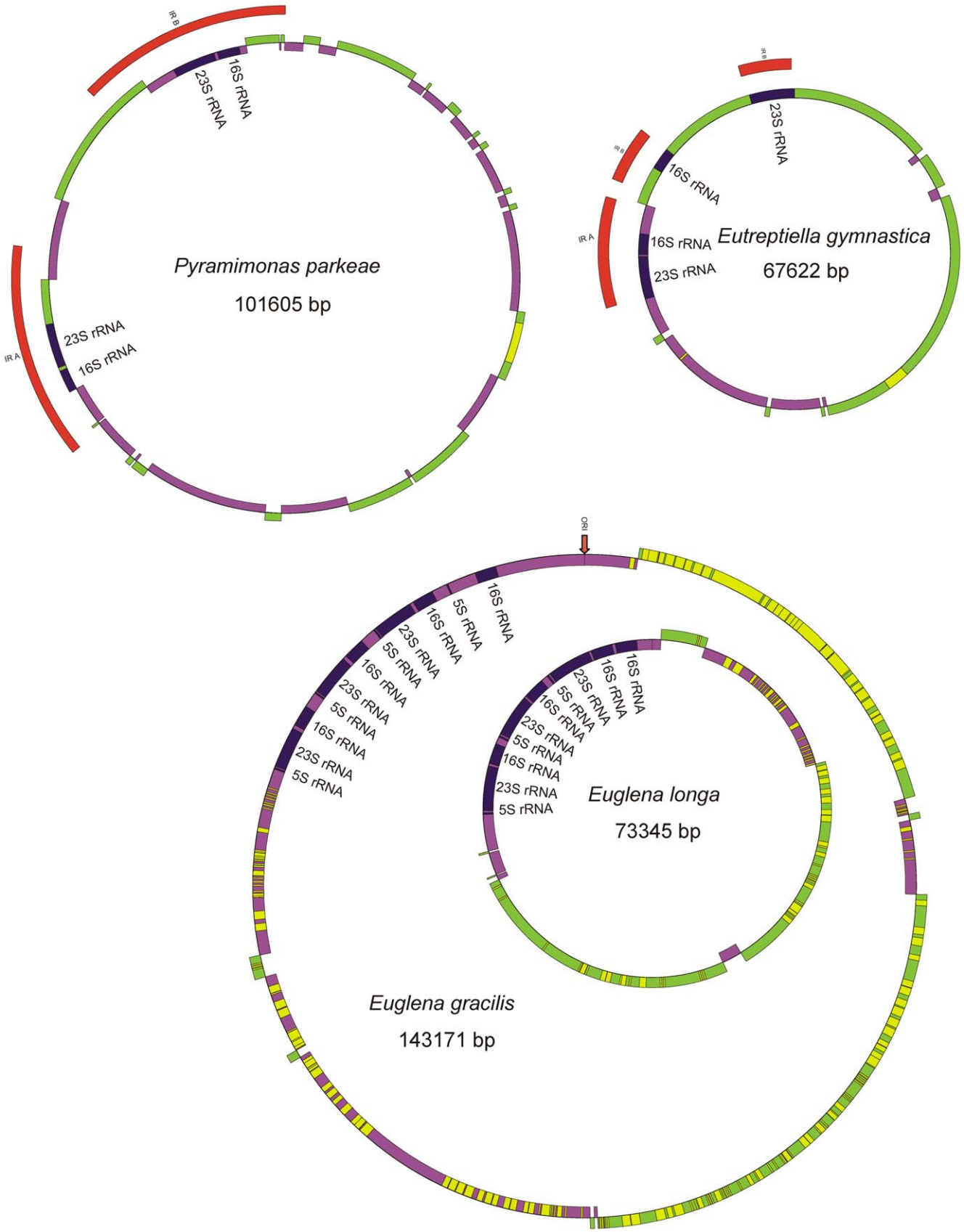

Figure 2. Simplified maps of the plastid genomes of *Eutreptiella gymnastica*, *Euglena gracilis*, *Euglena longa* and *Pyramimonas parkeae*. The maps are in scale to their sizes. The colors indicate the coding strands (plus-green and minus-violet), the ribosomal RNAs (blue) and introns (yellow). The inverted repeats IRA and IRB in *Pyramimonas* and *Eutreptiella* are marked in red. The ori site in *Euglena gracilis* is marked by an arrow.
doi:10.1371/journal.pone.0033746.g002

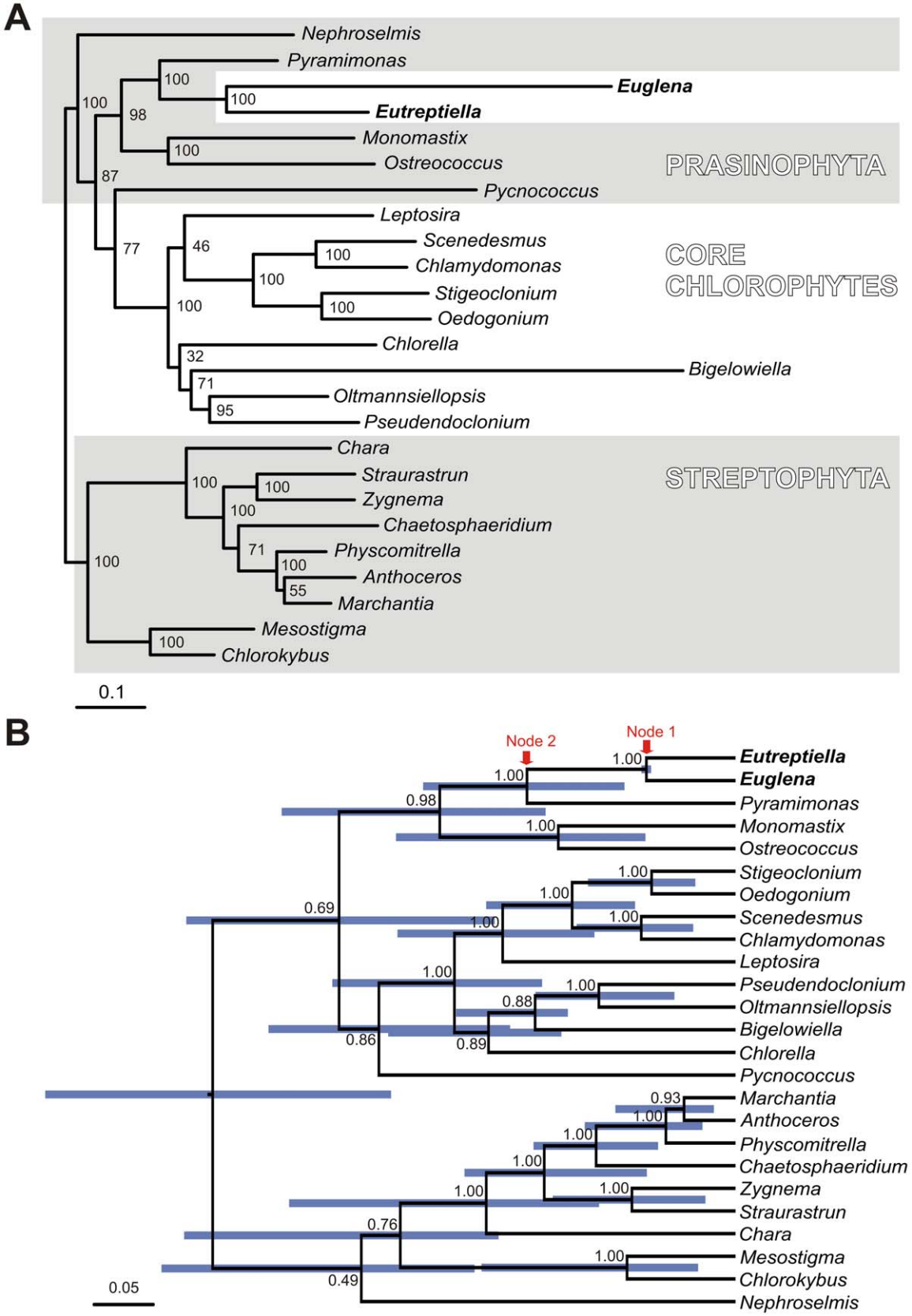

Figure 3. Venn diagrams showing overlaps in protein coding capacities between known euglenid plastid genomes and the plastid genome of *Pyramimonas parkeae*. The schematic representation of genome relationships is indicated in the left. Arrows indicate the probable fate of the genes absent from euglenid genomes. The genes are colour coded in respect to the functional group of their products: housekeeping proteins

(black), proteins involved in photosynthesis (green), maturases of introns (red) and genes with unknown function (gray). Maturases of introns included in the phylogenetic tree of maturases (Figure S1) are marked by asterisks.
doi:10.1371/journal.pone.0033746.g003

common for pairs of genomes. In this measure, the *E. gracilis* and *Eutreptiella* genomes are the closest as expected, sharing 61 adjacent gene couples; *Pyramimonas* shares with each of them 40 and 38 gene neighbors, respectively.

Phylogenomic analysis of 70 plastid protein coding genes confirmed with maximum bootstrap support the sister relationship of euglenid plastids and *Pyramimonas* (Figure 4) as reported by Turmel et al. [16]. The tip branch of *Euglena gracilis* is almost three times longer than the branch of *Eutreptiella*, probably a result of an accelerated substitution rate in the lineage leading to the genus *Euglena* (Figure 4A). The analyses with relaxed molecular clocks produced ultrametric trees (Figure 4B and Figure S2) that give estimates of relative ages of internal nodes. The branching order of these trees is virtually identical to the maximum likelihood tree. The relaxed clock analyses revealed that the common ancestor of *Euglena* and *Eutreptiella* (node 1 in Figure 4B and Figure S2) was not very recent, as it was approximately as old or older (depending on the clock model) as the common ancestor of vascular plants (common ancestor of *Marchantia*, *Anthoceros* and *Physcomitrella*). It also revealed that the age of the common ancestor of *Pyramimonas* and euglenid plastid (node 2 in Figure 4B and Figure S2), for the three clock models, was 1.2–2.3× older than the common ancestor of *E. gracilis* and *Eutreptiella* if considering the median of the age estimates and 1–5× older if considering the extreme values of the 95% confidence intervals of the age estimates (blue bars in Figures 4B and Figure S2). The time span from node 2 to node 1 was therefore similarly as long as or shorter than the time span from node 1 to the present time, but likely was not markedly longer. The period from node 2 to node 1 includes the green algal lineage that became the direct ancestor of the secondary euglenid plastid and then the stem branch of the secondary plastid before the split of genera *Euglena* and *Eutreptiella*. The exact point where the transition between alga and plastid happened is not known. During this period, 16 protein coding genes functioning in the plastid metabolism were possibly lost and six were transferred to the nucleus of the euglenid. This is in contrast to the at least comparable but very probably quite longer time of evolution that separates extant photosynthetic *E. gracilis* and *Eutreptiella* (twice the time from node 1 to present) during which only one gene (ycf65) was lost and none was transferred to the nucleus. The rapid slow-down of gene loss could be explained by the fact that the gene set was relatively quickly reduced to an essential core that must be preserved if the photosynthetic function is to be retained. The complete halt of endosymbiotic gene transfer from plastid to the host nucleus is, however, unexpected, as such transfers are also reported in plastids that have been established for a long time in their hosts [31,32]. Unlike the gene content the gene order evolved relatively uniformly – 61 gene couples remained in neighboring positions after the period separating *E. gracilis* and *E. gymnastica*, and correspondingly fewer (40 or 38) gene couples remained positionally fixed to each other after approximately double the period separating *P. parkeae* and *E. gracilis* or *P. parkeae* and *E. gymnastica*.

In conclusion, the plastid genome of *Eutreptiella* turned out to be almost identical to *Euglena gracilis* in protein coding gene content that is reduced when compared to *Pyramimonas*. This indicates that virtually all gene losses and endosymbiotic transfers of genes to the host nucleus took place in the period before the last common ancestor of the euglenid plastid. In contrast to the frozen protein content, the genome organization (gene order, inverted repeats) diversified significantly in the two sequenced lineages of euglenid plastids, and in the lineage leading to the genus *Euglena* it was furthermore accompanied by an accelerated substitutional rate in protein sequences and the expansion of self splicing introns. We have shown that the method of 454 sequencing could be widely applied to sequencing of organellar genomes.

Materials and Methods

Preparation of genomic DNA

A culture of *Eutreptiella gymnastica* strain SCCAP K-0333 was obtained from the Scandinavian Culture Collection of Algae and Protozoa and grown in TL30 medium in 12°C. 150 ml of well-grown culture (approx. $25*10^6$ cells) was used for DNA isolation. DNA was isolated using the Quiagen Blood and Tissue kit.

Sequencing and assembly of the plastid genome

1 µg of whole genomic DNA was subjected to 454 sequencing according to GS FLX Rapid Library Preparation Method protocol (Roche). In total 548 056 reads of average size 370 bases were produced. Automatic assembly of reads in Newbler 2.5.3 (Roche) resulted in 19 417 contigs (N50 contig size was 791 bases) and 9.2 Mb of unique sequence. Using a BLASTn homology search it was determined that two contigs, by far the longest (26 365 bp and 20 813 bp), represented parts of the plastid genome. It is expected that contigs derived from the plastid genome should have approximately the same coverage, and so those contigs that had coverage similar to contigs 1 and 2 (35× for contig 1 and 30× for contig 2) were selected from the assembly and all subjected to BLASTn homology search. Five of them were found to represent parts of the plastid genome. All plastid derived contigs were then manually assembled into a 67,274 bp long linear supercontig. Because we expected that the plastid genome would be a circular molecule, a PCR from the ends of the linear supercontig was used to amplify and sequence the missing part (primer F: 5′ - taacctgtgaacacgaag -3′ and primer R: 5′ - caaccagtaagttataggaa -3′). After adding 348 bases the genome was circularized.

Annotation

Annotation of ORFs was done using BLASTx homology search. tRNAs were found using tRNA Scan-SE [33], and rRNAs were annotated using a BLASTn homology search with their boundaries determined according to the alignment with rRNA from *Euglena gracilis* and *Pyramimonas parkeae*. The annotation was completed in Artemis 13.2.0 [34] and the annotated genome is deposited in the EMBL database under accession no. HE605038. The genome maps were plotted in GenomeV [35].

Intron secondary structures

The secondary structures of intron candidates were predicted by mFOLD version 2.3 [36] (http://mfold.rna.albany.edu/?q = mfold/RNA-Folding-Form2.3) using the default setting but with the temperature set to 12°C.

Phylogenetic analyses

The set of maturases was assembled from *Eutreptiella* mat1 and 121 homologues from GenBank representing both all available euglenid homologues and homologues from other taxa covering

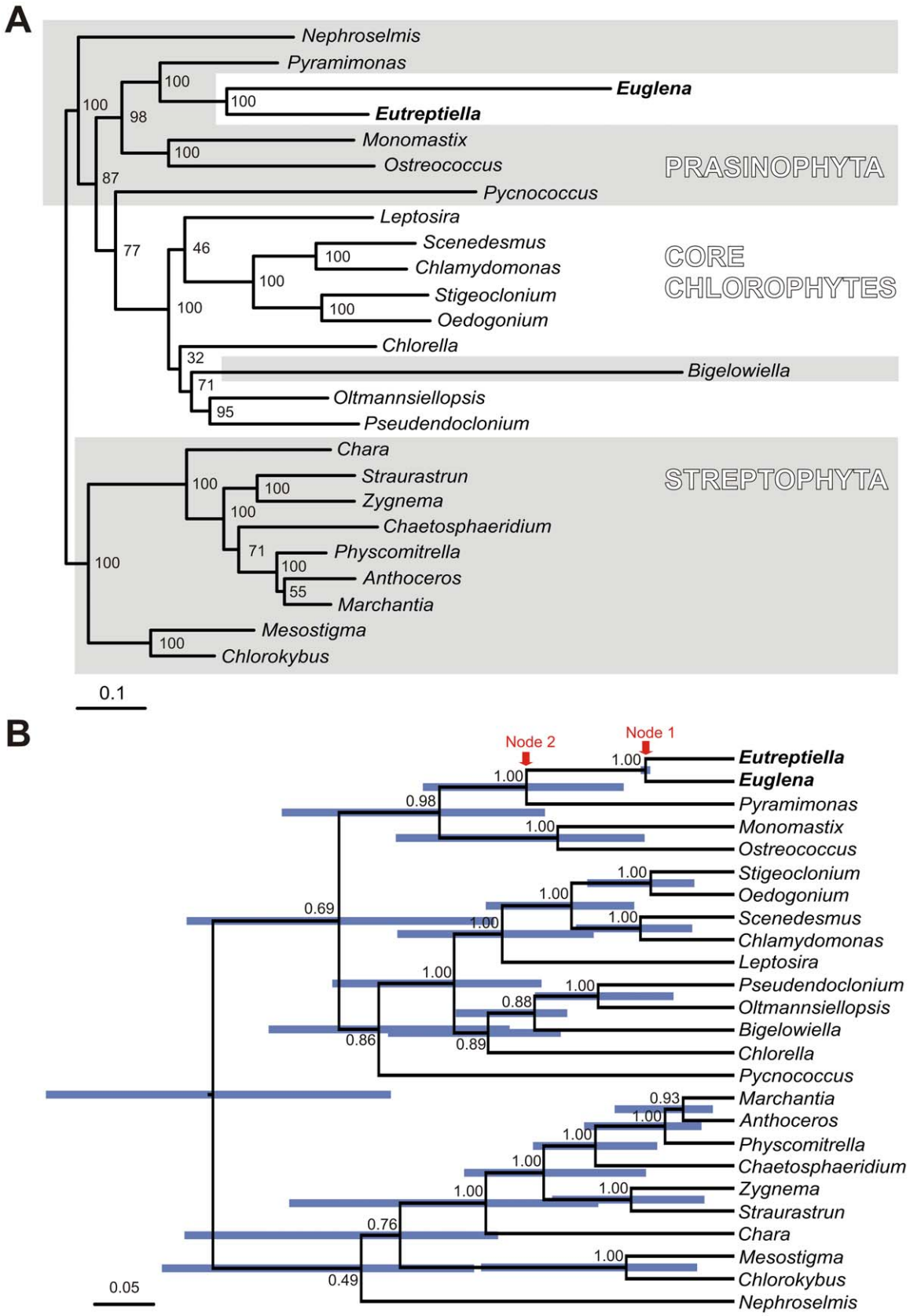

Figure 4. Phylogenies of plastid genomes of green algae, euglenids and *Bigelowiella* based on 70 genes. A. This phylogenetic tree was constructed using the maximum likelihood method implemented in RAxML, using the LG+I+G model selected by ProtTest. The bootstraps were estimated in 500 replicates. **B.** This tree was constructed in Beast v 1.6.1 using the WAG+I+Γ model of substitution and an uncorrelated exponential

model of relaxed molecular clock. MCMCs were run for $10*10^6$ generations; trees from the first $2*10^6$ generations were discarded as the burn-in. Node labels represent posterior probabilities, node bars represent the 95% confidence interval of relative node ages.
doi:10.1371/journal.pone.0033746.g004

the sequential diversity of this protein. The data set was aligned using ClustalX [37] and manually edited in Bioedit 7.0.5.3 [38]. The phylogenetic tree was constructed in RAxML v7.2.7 [39] using the PROTGAMMAILGF model. The bootstrap support was calculated using the same model and 500 permutations.

For the phylogenomic analysis we used the data set of 70 protein coding genes from 24 plastid genomes published by Turmel et al [17]. The *Eutreptiella* sequences were manually added to this set in Bioedit 7.0.5.3 [38], realigned using ClustalX [37], and the alignment was then manually edited in Bioedit 7.0.5.3 [38]. The phylogenetic tree was constructed in RAxML v7.2.7 [39] using a uniform PROTGAMMAILGF model for all gene partitions. The bootstrap support was calculated using the same model and 500 permutations. The analyses using relaxed molecular clocks were performed in Beast v 1.6.1 [40] using the WAG+I+Γ model of substitution and three models of relaxed molecular clock: an uncorrelated exponential model, an uncorrelated lognormal model and a random model. MCMC was run for $10*10^6$ generations; trees from first $2*10^6$, $7*10^6$ and $3*10^6$ generations were discarded as the burn-in, respectively.

Supporting Information

Figure S1 The phylogeny of intron maturases. The phylogenetic tree was constructed using the maximum likelihood method implemented in RAxML, using the LG+I+G model selected by ProtTest. The bootstraps were estimated in 500 replicates. The eukaryotic maturases are marked by red, the cyanobacterial are marked by cyan and other bacterial maturases are marked by black.
(DOCX)

Figure S2 Phylogenies of plastid genomes of green algae, euglenids and *Bigelowiella* based on 70 genes. These trees were constructed in Beast v 1.6.1 using the WAG+I+Γ model of substitution and an uncorrelated lognormal model of relaxed molecular clock (A) and random local model of relaxed molecular clock (B). MCMCs were run for $10*10^6$ generations; trees from the first $7*10^6$ and $3*10^6$ generations were discarded as the burn-in in A and B, respectively. Node labels represent posterior probabilities, node bars represent the 95% confidence interval of relative node ages.
(DOCX)

Acknowledgments

The authors would like to thank to Dr. Steve Zimmerly for e-mail consultations regarding group II and group III introns and Aaron Heiss for language corrections.

Author Contributions

Conceived and designed the experiments: VH CV. Performed the experiments: JF JS SH. Analyzed the data: SH VH CV. Contributed reagents/materials/analysis tools: CV VH. Wrote the paper: VH SH.

References

1. 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). Int J Syst Evol Microbiol 51: 751–8.
2. Leander BS, Esson HJ, Breglia SA (2007) Macroevolution of complex cytoskeletal systems in euglenids. Bioessays 29: 987–1000.
3. Linton E, Hittner D, Levandowski CF, Auld T, Triemer RE (1999) A molecular study of euglenoid phylogeny using small subunit rDNA. J Eukaryot Microbiol 46: 217–223.
4. Marin B, Palm A, Klingberg M, Melkonian M (2003) Phylogeny and taxonomic revision of plastid-containing euglenophytes based on SSU rDNA sequence comparison and synapomorphic signatures in the SSU rRNA secondary structure. Protist 154: 99–145.
5. Marin B (2004) Origin and fate of chloroplasts in the euglenoida. Protist 155: 13–14.
6. Hallick RB, Hong L, Drager RG, Favreau MR, Monfort A, et al. (1993) Complete sequence of Euglena gracilis chloroplast DNA. Nucleic Acids Res 21: 3537–44.
7. Gockel G, Hachtel W (2000) Complete gene map of the plastid genome of the nonphotosynthetic euglenoid flagellate Astasia longa. Protist 151: 347–51.
8. Gibbs SP (1978) The chloroplasts of Euglena may have evolved from symbiotic green algae. Can J Bot 56: 2883–2889.
9. Leander BS, Triemer RE, Farmer MA (2001) Character evolution in heterotrophic euglenids. Eur J Protistol 37: 337–356.
10. Leander BS (2004) Did trypanosomatid parasites have photosynthetic ancestors? Trends Microbiol 12: 251–8.
11. Krajčovič J, Ebringer L, Schwartzbach SD (2002) Reversion of endosymbiosis? In Seckbach J, ed. Symbiosis: Mechanisms and Models. Cellular Origin in Extreme Habitats, Vol. 4. Dordrecht: Kluwer Academic Publisher. pp 185–206.
12. Vesteg M, Vacula R, Steiner JM, Mateášiková B, Löffelhardt W, et al. (2010) A possible role for short introns in the acquisition of stroma-targeting peptides in the flagellate Euglena gracilis. DNA Res 17: 223–231.
13. Hannaert V, Saavedra E, Duffieux F, Szikora JP, Rigden DJ, et al. (2003) Plant-like traits associated with metabolism of Trypanosoma parasites. Proc Natl Acad Sci USA 100: 1067–1071.
14. Bodył A, Mackiewicz P, Milanowski R (2010) Did trypanosomatid parasites contain an eukaryotic alga-derived plastid in their evolutionary past? J Parasitol 96: 465–75.
15. Maruyama S, Suzaki T, Weber APM, Archibald JM, Nozaki H (2011) Eukaryote-to-eukaryote gene transfer gives rise to genome mosaicism in euglenids. BMC Evolutionary Biology 11: 105.
16. Turmel M, Gagnon MC, O'Kelly CJ, Otis C, Lemieux C (2009) The chloroplast genomes of the green algae Pyramimonas, Monomastix, and Pycnococcus shed new light on the evolutionary history of prasinophytes and the origin of the secondary chloroplasts of euglenids. Mol Biol Evol 26: 631–48.
17. Thompson MD, Copertino DW, Thompson E, Favreau MR, Hallick RB (1996) Evidence for the late origin of introns in chloroplast genes from an evolutionary analysis of the genus Euglena. Nucleic Acids Res 23(23): 4745–52. Errata: Nucleic Acids Res 24: 542, 24: 1792, 1996.
18. 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? Mol Biol Evol 15(1): 76–86.
19. Müllner AN, Angeler DG, Samuel R, Linton EW, Triemer RE (2001) Phylogenetic analysis of phagotrophic, photomorphic and osmotrophic euglenoids by using the nuclear 18S rDNA sequence. Int J Syst Evol Microbiol 51: 783–91.
20. Busse I, Preisfeld A (2003) Systematics of primary osmotrophic euglenids: a molecular approach to the phylogeny of Distigma and Astasia (Euglenozoa). Int J Syst Evol Microbiol 53: 617–24.
21. Geimer S, Belicova A, Legen J, Slavikova S, Herrmann RG, et al. (2009) Transcriptome analysis of the Euglena gracilis plastid chromosome. Current Genetics 55: 425–238.
22. Heinhorst S, Canon GC (1993) DNA replication in chloroplasts. J Cell Sci 104: 1–9.
23. Bendich AJ (2004) Circular chloroplast chromosomes: the grand illusion. Plant Cell 16: 1661–6.
24. Ravel-Chapuis P, Heizmann P, Nigon V (1982) Electron microscopic localization of the replication origin of Euglena gracilis chloroplast DNA. Nature 300: 78–81.
25. Koller B, Delius H (1982) Origin of replication in chloroplast DNA of Euglena gracilis located close to the region of variable size. EMBO J 1: 995–8.
26. Morton BR (1999) Strand asymmetry and codon usage bias in the chloroplast genome of Euglena gracilis. Proc Natl Acad Sci U S A 96: 5123–8.
27. Michel F, Umesono K, Ozeki H (1989) Comparative and functional anatomy of group II catalytic introns-a review. Gene 82: 5–30.
28. Michel F, Ferat JL (1995) Structure and activities of group II introns. Annu Rev Biochem 64: 435–61.
29. 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 Res 21: 4991–7.
30. Sheveleva EV, Hallick RB (2004) Recent horizontal transfer to a chloroplast genome. Nucleic Acids Res 32: 803–810.

31. Huang CY, Ayliffe MA, Timmis JN (2003) Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature 422: 72–6.
32. Bock R, Timmis JN (2008) Reconstructing evolution: gene transfer from plastids to the nucleus. Bioessays 30: 556–66.
33. Lowe TM, Eddy SR (1997) tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucl Acids Res 25: 955–964.
34. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944–5.
35. Conant GC, Wolfe KH (2008) GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics 24: 861–2.
36. Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res 31: 3406–3415.
37. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.
38. Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41: 95–98.
39. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688–90.
40. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214.

Conclusions

The patchy distribution of MAT/MATX gene couple in the eukaryotic tree points to the complicated evolutionary history of this enzyme, which could be the result of: i) either HGT or ii) ancient paralogy with subsequent differential losses of one its counterpart paralogs in different eukaryotic lineages (Fig. 4).

These two scenarios differ in their assumptions. Deep paralogy scenario assumes: i) the long-term coexistence of both paralogs and their co-expression in the cell, ii) the congruence of phylogenetic relationships within individual paralogs with the organismal phylogeny and iii) the position of the root of the gene tree between the two paralogs. On the other hand, HGT scenario expects that: i) one paralog (probably the less frequent and divergent – in our case MATX) should be able to substitute the function of the other (ancestral) paralog in a short time, ii) the tree topology of at least one paralog is not consistent with the organismal topology, iii) there was a donor lineage, in which the divergent paralog MATX evolved and afterwards spread to other lineages. Then the root of MAT/MATX tree would be situated inside the MAT paralog and the donor lineage would be at the base of the MATX clade.

Because we had no information clarifying how MAT and MATX fulfill the assumptions of one or the other scenario, we decided to investigate it partly experimentally by means of an *Euglena gracilis*/*Trypanosoma brucei in-vivo*-model system (paper no. 1), and partly by phylogenetic reconstruction of gene trees using the most complete and updated set of homologues available at present (paper no. 2). We found out that transfection of *T. brucei* cells (containing the ancestral MAT form of the gene) with MATX from *E. gracilis*, and the co-expression of both genes had no impact on cells viability and growth. Thus the coexistence of both paralogs in one cell is possible indeed. This result indicates that MAT/MATX fulfills one assumption for the deep paralogy scenario. After initiation of RNA interference, the original MAT was knocked-down but the cells were able to survive because MATX substituted its function. This successful replacement is the assumption for the HGT scenario. Our *in vivo* experiments have shown that MATX is capable of evolving and spreading among different organisms by HGT as well as by ancestral coexistence with MAT and differential loss (paper no. 1).

We performed the same experiment with the EF-1α and its paralog EFL. While *T. brucei* cells were viable with both paralogs expressed in the same cell, after RNA interference induction and knocking down of the ancestral EF-1α, the foreign EFL was not able to substitute this essential function and cells died. This showed the differences between MAT/MATX and EF-1α/EFL gene couples. EFL was able to coexist with EF-1α but it was not capable of HGT in any of experiments carried out (paper no. 1).

In the next study (paper no. 2), we focused on the possible origin of MATX in the group of euglenids. We collected MATX sequences from various species of euglenids, both heterotrophic and photoautotrophic. We were also able to collect data from a recently discovered euglenid, mixotroph *Rapaza viridis*. We found MATX in all investigated photoautotrophic euglenids, while the mixotroph *Rapaza viridis*, which contains the secondary plastid and is the most basal lineage in photoautotrophic euglenid clade, possessed only the MAT form of the gene. The MAT form was also found in heterotrophic euglenids (*Petalomonas*, *Distigma* and *Peranema*) and the prasinophyte alga *Pyramimonas parkeae*, which represents the closest known relative of the euglenid plastid. From these results, we infer that MATX is specific for the clade of photoautotrophic euglenids, but it appeared in this lineage after the split of the mixotroph *R. viridis*, e.g. not simultaneously with the endosymbiosis of the euglenid plastid, which took place before *R. viridis* split. Based on these findings, we reject the former hypothesis of Sánchez-Pérez et al. (2008) claiming that MATX evolved in the nucleomorph of the emerging euglenid secondary plastid.

In order to study the evolutionary history of these enzymes in eukaryotes we performed phylogenetic analyses from available eukaryotic-MATX sequences (paper no. 2). The eukaryotic MATX phylogeny did not match the organismal phylogeny because MATX sequences from euglenids branched together with the dinoflagellates. The significance of these differences was confirmed in statistical tests. This is in contrast with one assumption of the deep paralogy scenario. The root in MAT/MATX phylogenetic tree was not localized between MAT and MATX clade, but the statistical tests did not exclude the possibility of this position, therefore MATX could be presented in the eukaryotic common ancestor, what is in agreement with one assumption of the deep paralogy scenario.

Regarding to euglenids, we also found that within this group the MATX phylogeny did not correspond to the organismal phylogeny, but the differences were not statistically significant and, consequently, the congruence was not rejected. Therefore, in the phototrophic clade of euglenids, MATX could have evolved by vertical inheritance. In order to explain the distribution of MATX paralog in euglenids by deep paralogy, there would be needed at least four independent losses of MATX. These gene losses would be increased, much more, when trying to explain the distribution of MATX within the whole eukaryotic tree. To explain the distribution of the MAT paralogue by deep paralogy we would need just one loss in euglenids and only few in the whole eukaryotic tree. Although, gene losses are frequent events among organisms, the fact that losses of one paralog are much more frequent than for the other one is suspicious. When we look on the MAT/MATX distribution from the HGT point of view, there will be needed only one HGT of MATX gene into photoautotrophic euglenids clade after the split of *R. viridis* and only few others to explain the situation in whole eukaryotes. Nevertheless, simple counts of these evolutionary events, do not allow making any serious conclusion, as we have no information about their relative probabilities.

During analyses of the distribution of these two paralogs in euglenids, we found also two dual MAT/MATX-containing euglenids (*Monomorphina pyrum* and *Phacus orbicularis*). MAT genes of these species branch on phylogenetic tree with *A. anophagefferens* and with ciliates, respectively, but not with the rest of MAT from Euglenozoa. Although the presence of MAT from ciliates looks slightly suspicious and it could be a contamination, we never saw ciliates in the culture. We also tried to amplify the 5' terminus of this MAT sequence to see whether it contained the typical splice leader, but unfortunately we did not succeed (unpublished data). At this moment, we favor the explanation of this pattern by recent HGTs into two lineages of phototrophic euglenids but the possibility of contamination was not excluded yet.

The biggest problem of the HGT scenario is that there is no clear donor group of MATX. From the MATX phylogeny we cannot even speculate about the directions of the possible transfers between eukaryotic groups. There is a theoretical option that the donor lineage of MATX could become extinct and therefore cannot be determined. This

hypothesis was firstly introduced by Fournier et al. (2009) in order to explain the distribution of the amino-acid pyrrolysine.

Based on both *in-vivo* experiments and phylogenetic reconstructions we infer that MAT/MATX fulfill the complete set of assumptions of neither of these two scenarios. MATX has not evolved purely vertically as supposes the deep paralogy scenario, whilst at the same time, the absence of donor lineage argues against the HGT hypothesis.

To reconcile the evidence that does not support any of the two scenarios in their pure form, we propose following hypothesis of evolution of MAT and MATX genes. MAT and MATX were both present in the eukaryotic common ancestor. Subsequently, one or the other paralog was lost in different eukaryotic lineages. Except for differential more or less random losses, few HGTs of MAT and MATX took place between lineages. In this way, MAT from ciliates and *A. anophagefferens* was transferred to two photoautotrophic euglenids. One transfer of MATX introduced this gene into the lineage of phototrophic euglenids from a still unknown source. This HGT happened after the acquisition of the secondary plastid, as no MATX was found in *R. viridis* or in *P. parkeae*. Then, after MATX was established in this lineage, it substituted the original MAT and subsequently evolved by vertical descent.

In the last part of our work (paper no. 3) we sequenced the plastid genome of *Eutreptiella gymnastica*, which was, before discovering *R. viridis*, the most basal and deepest lineage of photoautotrophic euglenids. The purpose of this study was to look closer into the changes involving the gene content as well as rearrangements in the plastid genome, which occurred after the ingestion of a green alga. These genome changes are common processes that could theoretically lead to the formation of MATX in euglenids or other secondary algae, although we currently do not support the hypothesis that MATX evolved this way.

We found out that the plastid genome of *Eutreptiella* is less than half of the size of *E. gracilis* plastid genome, as it does not have such a large amount of introns like *E. gracilis* (over 150 group II. and III. introns) (Hallick et al. 1993). In contrast with *Euglena*, *Eutreptiella* has only eight introns; two of them identified by us (paper no. 3), five by Pombert et al. (2012) and the last one by us after recent intensive screening of the *E. gymnastica* plastid genome (unpublished results). This corroborates the

hypothesis that the euglenid plastid ancestor was intron-poor (the *P. parkeae* plastid genome contains only one intron) (Turmel et al. 2009; Pombert et al. 2012). The number of conserved genes in *Euglena gracilis* and *Eutreptiella gymnastica* is almost identical, but both of them are reduced in comparison to the plastid genome of *Pyramimonas*. This suggest that during formation of plastid from alga, up to 18 protein coding genes were lost completely or were transferred to the nucleus; while during the period following that event, only one gene was lost and none was transferred to the nucleus. The coding capacity was therefore reduced, probably very soon, after the plastid acquisition in the period before last euglenid-plastid common ancestor (Hrdá et al. 2012). This points to the fact that major changes in the genome content took place in the early period of its evolution. Once the plastid was established, it became resistant to changes in gene content. But when we compare the plastid genomes from various euglenids sequenced so far, there are still changes in the gene order.

In summary, we have found out that the plastid genome of euglenids experienced a reduction in the number of genes soon after endosymbiosis, followed by the expansion of introns within the line leading to *E. gracilis*. However, the introduction of MATX into photosynthetic euglenids had no connection to the plastid endosymbiosis and it took place after the euglenid-plastid acquisition. We propose that the evolutionary history of the MAT/MATX in eukaryotes represents a mixture of both HGT and deep paralogy scenarios, as our data are not entirely consistent with either of these scenarios in their purest form.

References

Almeida, F. C., Leszczyniecka, M., Fisher, P. B., and Desalle, R. 2008. “Examining Ancient Inter-Domain Horizontal Gene Transfer.” *Evolutionary Bioinformatics* 4: 109–19.

Andersson, J. O. 2009. “Horizontal Gene Transfer between Microbial Eukaryotes.” In *Horizontal Gene Transfer*, edited by MariaBoekels Gogarten, JohannPeter Gogarten, and LorraineC. Olendzenski, 532:473–87. Methods in Molecular Biology. Humana Press.

Andersson, J. O. 2005. “Lateral Gene Transfer in Eukaryotes.” *Cellular and Molecular Life Sciences* 62 (11): 1182–97.

Andersson, J. O, Doolittle, W. F., and Nesbø, C. L. 2001. “Are There Bugs in Our Genome?” *Science* 292 (5523): 1848–50.

Andersson, J. O., Hirt, R. P., Foster, P. G., and Roger, A. J. 2006. “Evolution of Four Gene Families with Patchy Phylogenetic Distributions: Influx of Genes into Protist Genomes.” *BMC Evolutionary Biology* 6: 27.

Andersson, J. O., and Roger, A. J. 2002. “A Cyanobacterial Gene in Nonphotosynthetic Protists--an Early Chloroplast Acquisition in Eukaryotes?” *Current Biology* 12 (2): 115–19.

Andersson, J. O., and Roger, A. J. 2003. “Evolution of Glutamate Dehydrogenase Genes: Evidence for Lateral Gene Transfer within and between Prokaryotes and Eukaryotes.” *BMC Evolutionary Biology* 3 (1): 14.

Atkinson, G. C., Kuzmenko, A., Chicherin, I., Soosaar, A., Tenson, T., Carr, M., Kamenski, P., and Hauryliuk, V. 2014. “An Evolutionary Ratchet Leading to Loss of Elongation Factors in Eukaryotes.” *BMC Evolutionary Biology* 14 (1). BMC Evolutionary Biology: 35.

Behe, M. J., and Snoke, D. W. 2004. “Simulating Evolution by Gene Duplication of Protein Features That Require Multiple Amino Acid Residues.” *Protein Science* 13: 2651–64.

Blanch, A., Robinson, F., Watson, I. R., Cheng, L. S., and Irwin, M. S. 2013. “Eukaryotic Translation Elongation Factor 1-Alpha 1 Inhibits p53 and p73 Dependent Apoptosis and Chemotherapy Sensitivity.” *PloS One* 8 (6): e66436.

Cantoni, G. L. 1975. “Biological Methylation: Selected Aspects.” *Annual Review of Biochemistry* 44: 435–51.

Cantoni, G. L. 1951. “Activation of Methionine for Transmethylation.” *J. Biol. Chem* 189: 745–54.

Dixon, M. M., Huang, S., Matthews, R. G., and Ludwig, M. 1996. “The Structure of the C-Terminal Domain of Methionine Synthase: Presenting S-Adenosylmethionine for Reductive Methylation of B12.” *Structure (London, England : 1993)* 4 (11): 1263–75.

Doolittle, W. F. 1998. “You Are What You Eat: A Gene Transfer Ratchet Could Account for Bacterial Genes in Eukaryotic Nuclear Genomes.” *Trends in Genetics* 14 (8): 307–11.

Dreher, T. W., Uhlenbeck, O. C., and Browning, K. S. 1999. “Quantitative Assessment of EF-1 α·GTP Binding to Aminoacyl-tRNAs, Aminoacyl-Viral RNA, and tRNA Shows Close Correspondence to the RNA Binding Properties of EF-Tu.” *J. Biol. Chem* 274 (2): 666–72.

Duttaroy, A., Bourbeau, D., Wang, X. L., and Wang, E. 1998. “Apoptosis Rate Can Be Accelerated or Decelerated by Overexpression or Reduction of the Level of Elongation Factor-1 Alpha.” *Experimental Cell Research* 238 (1): 168–76.

Dzidic, S., and Bedeković, V. 2003. “Horizontal Gene Transfer-Emerging Multidrug Resistance in Hospital Bacteria.” *Acta Pharmacologica Sinica* 24 (6): 519–26.

Espartero, J., Pintor-Toro, J. A., and Pardo, J. M. 1994. “Differential Accumulation of S-Adenosylmethionine Synthetase Transcripts in Response to Salt Stress.” *Plant Molecular Biology* 25 (2). Kluwer Academic Publishers: 217–27.

Fontecave, M., Atta, M., and Mulliez, E. 2004. “S-Adenosylmethionine: Nothing Goes to Waste.” *Trends in Biochemical Sciences* 29 (5): 243–49.

Fournier, G. P., Huang, J., and Gogarten, J. P. 2009. “Horizontal Gene Transfer from Extinct and Extant Lineages: Biological Innovation and the Coral of Life.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 364 (1527): 2229–39.

Garrido, F., Alfonso, C., Taylor, J. C., Markham, G. D., and Pajares, M. A. 2009. “Subunit Association as the Stabilizing Determinant for Archaeal Methionine Adenosyltransferases.” *Biochimica et Biophysica Acta* 1794 (7):1082–90.

Garrido, F., Estrela, S., Alves, C., Sánchez-Pérez, G. F., Sillero, A., and Pajares, M. A. 2011. “Refolding and Characterization of Methionine Adenosyltransferase from Euglena Gracilis.” *Protein Expression and Purification* 79 (1):128–36.

Gile, G. H., Faktorová, D., Castlejohn, C. A., Burger, G., Lang, B. F., Farmer, M. A., Lukeš, J., and Keeling, P. J. 2009. “Distribution and Phylogeny of EFL and EF-1alpha in Euglenozoa Suggest Ancestral Co-Occurrence Followed by Differential Loss.” *PloS One* 4 (4): e5162.

Glansdorff, N. 2000. “About the Last Common Ancestor, the Universal Life-Tree and Lateral Gene Transfer: A Reappraisal.” *Molecular Microbiology* 38 (2): 177–85.

Gogarten, J. P., and Townsend, J. P. 2005. “Horizontal Gene Transfer, Genome Innovation and Evolution.” *Nature Reviews. Microbiology* 3 (9): 679–87.

Gómez-Gómez, L., and Carrasco, P. 1998. “Differential Expression of the S-Adenosyl-L-Methionine Synthase Genes during Pea Development.” *Plant Physiology* 117 (2): 397–405.

Graham, D. E., Bock, C. L., Schalk-Hihi, C., Lu, Z. J., and Markham, G. D. 2000. “Identification of a Highly Diverged Class of S-Adenosylmethionine Synthetases in the Archaea.” *The Journal of Biological Chemistry* 275 (6): 4055–59.

Gross, S. R., and Kinzy, T. G. 2005. “Translation Elongation Factor 1A Is Essential for Regulation of the Actin Cytoskeleton and Cell Morphology.” *Nature Structural & Molecular Biology* 12 (9):772–78.

Guigó, R., Muchnik, I., and Smith, T. F. 1996. “Reconstruction of Ancient Molecular Phylogeny.” *Molecular Phylogenetics and Evolution* 6 (2): 189–213.

Halim, A. B., LeGros, L., Chamberlin, M. E., Geller, A., and Kotb, M. 2001. “Regulation of the Human MAT2A Gene Encoding the Catalytic Alpha 2 Subunit of Methionine Adenosyltransferase, MAT II: Gene Organization, Promoter Characterization, and Identification of a Site in the Proximal Promoter That Is Essential for Its Activity.” *The Journal of Biological Chemistry* 276 (13): 9784–91.

Hallick, R. B., Hong, L., Drager, R. G., Favreau, M. R., Monfort, A., Orsat, B., Spielmann, A., and Stutz, E. 1993. “Complete Sequence of Euglena Gracilis Chloroplast DNA.” *Nucleic Acids Research* 21 (15): 3537–44.

Hannaert, V., Saavedra, E., Duffieux, F., Szikora, J. P., Rigden, D. J., Michels, P. A. M., and Opperdoes, F. R. 2003. “Plant-like Traits Associated with Metabolism of *Trypanosoma* Parasites.” *Pnas* 100 (3): 1067–71.

Henk, D. A., and Fisher, M. C. 2012. “The Gut Fungus Basidiobolus Ranarum Has a Large Genome and Different Copy Numbers of Putatively Functionally Redundant Elongation Factor Genes.” *PloS One* 7 (2): e31268.

Hrdá, Š., Fousek, J., Szabová, J., Hampl, V., and Vlček, Č. 2012. “The Plastid Genome of Eutreptiella Provides a Window into the Process of Secondary Endosymbiosis of Plastid in Euglenids.” *PLoS ONE*.

Huang, J. 2013. “Horizontal Gene Transfer in Eukaryotes: The Weak-Link Model.” *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* 35 (10): 868–75.

Huang, J., and Gogarten, J. P. 2009. “Ancient Gene Transfer as a Tool in Phylogenetic Reconstruction.” In *Horizontal Gene Transfer SE - 7*, edited by MariaBoekels Gogarten, JohannPeter Gogarten, and LorraineC. Olendzenski, 532:127–39. Methods in Molecular Biology. Humana Press.

Huh, J. H., Kim, D. J., Zhao, X. Q., Li, M., Jo, Y. Y., Yoon, T. M., Shin, S. K., Yong, J. H., Ryu, Y. W., Yang, Y. Y., and Suh, J. W. 2004. “Widespread Activation of Antibiotic Biosynthesis by S-Adenosylmethionine in Streptomycetes.” *FEMS Microbiology Letters* 238 (2): 439–47.

Cherest, H., Surdin-Kerjan, Y., Exinger, F., and Lacroute, F. 1978. “S-Adenosyl Methionine Requiring Mutants in Saccharomyces Cerevisiae: Evidences for the Existence of Two Methionine Adenosyl Transferases.” *Molecular and General Genetics MGG* 163 (2). Springer-Verlag: 153–67.

Chiang, P. K., and Cantoni, G. L. 1977. “Activation of Methionine for Transmethylation. Purification of the S-Adenosylmethionine Synthetase of Bakers’ Yeast and Its Separation into Two Forms.” *Journal of Biological Chemistry* 252 (13): 4506–13.

Chiang, P. K., Gordon, R. K., Tal, J., Zeng, G. C., Doctor, B. P., Pardhasaradhi, K., and McCann, P. P. 1996. “S-Adenosylmethionine and Methylation.” *FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology* 10: 471–80.

Ishitani, Y., Kamikawa, R., Yabuki, A., Tsuchiya, M., Inagaki, Y., and Takishita, K. 2012. “Evolution of Elongation Factor-Like (EFL) Protein in Rhizaria Is Revised by Radiolarian EFL Gene Sequences.” *Journal of Eukaryotic Microbiology* 59 (4): 367–73.

Juhas, M., van der Meer, J. R., Gaillard, M., Harding, R. M., Hood, D. W., and Crook, D. W. 2009. “Genomic Islands: Tools of Bacterial Horizontal Gene Transfer and Evolution.” *FEMS Microbiology Reviews* 33 (2): 376–93.

Kamikawa, R., Brown, M. W., Nishimura, Y., Sako, Y., Heiss, A. A., Yubuki, N., Gawryluk, R., Simpson, A. G. B., Roger, A. J., Hashimoto, T., and Inagaki, Y. 2013. “Parallel Re-Modeling of EF-1α Function: Divergent EF-1α Genes Co-Occur with EFL Genes in Diverse Distantly Related Eukaryotes.” *BMC Evolutionary Biology* 13 (131).

Kamikawa, R., Inagaki, Y., and Sako, Y. 2008. “Direct Phylogenetic Evidence for Lateral Transfer of Elongation Factor-like Gene.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (19): 6965–69.

Kamikawa, R., Sánchez-Pérez, G. F., Sako, Y., Roger, A. J., and Inagaki, Y. 2009. “Expanded Phylogenies of Canonical and Non-Canonical Types of Methionine Adenosyltransferase Reveal a Complex History of These Gene Families in Eukaryotes.” *Molecular Phylogenetics and Evolution* 53 (2):565–70.

Kamikawa, R., Yabuki, A., Nakayama, T., Ishida, K., Hashimoto, T., and Inagaki, Y. 2011. “Cercozoa Comprises Both EF-1α-Containing and EFL-Containing Members.” *European Journal of Protistology* 47 (1):24–28.

Karlberg, O., Canbäck, B., Kurland, C. G., and Andersson, S. G. 2000. “The Dual Origin of the Yeast Mitochondrial Proteome.” *Yeast* 17 (3): 170–87.

Kawalleck, P., Plesch, G., Hahlbrock, K., and Somssich, I. E. 1992. “Induction by Fungal Elicitor of S-Adenosyl-L-Methionine Synthetase and S-Adenosyl-L-Homocysteine Hydrolase mRNAs in Cultured Cells and Leaves of Petroselinum.” *Proceedings of the National Academy of Sciences of the United States of America* 89: 4713–17.

Keeling, P. J., and Inagaki, Y. 2004. “A Class of Eukaryotic GTPase with a Punctate Distribution Suggesting Multiple Functional Replacements of Translation Elongation Factor 1alpha.” *Proceedings of the National Academy of Sciences of the United States of America* 101 (43): 15380–85.

Keeling, P. J., and Palmer, J. D. 2008. “Horizontal Gene Transfer in Eukaryotic Evolution.” *Nature Reviews. Genetics* 9 (8): 605–18.

Khacho, M., Mekhail, K., Pilon-Larose, K., Pause, A., Côté, J., and Lee, S. 2008. “eEF1A Is a Novel Component of the Mammalian Nuclear Protein Export Machinery.” *Molecular Biology of the Cell* 19 (12): 5296–5308.

Kim, D. J., Huh, J. H., Yang, Y. Y., Kang, C. M., Lee, I. H., Hyun, C. G., Hong, S. K., and Suh, J. W. 2003. “Accumulation of S -Adenosyl-L-Methionine Enhances Production of Actinorhodin but Inhibits Sporulation in Streptomyces Accumulation of S -Adenosyl- L -Methionine Enhances Production of Actinorhodin but Inhibits Sporulation in Streptomyces Lividans TK23.” *Journal of Bacteriology* 185 (2): 592–600.

Koonin, E. V., Makarova, K. S., and Aravind, L. 2001. “Horizontal Gene Transfer in Prokaryotes: Quantification and Classification.” *Annual Review Microbiology* 55: 709–42.

Koonin, E. V., and Wolf, Y. I. 2008. “Genomics of Bacteria and Archaea: The Emerging Dynamic View of the Prokaryotic World.” *Nucleic Acids Research* 36 (21): 6688–6719.

Kotb, M., and Geller, A. M. 1993. “Methionine Adenosyltransferase: Structure and Function.” *Pharmacology & Therapeutics* 59 (2): 125–43.

Kotb, M., and Kredich, N. M. 1985. “S-Adenosylmethionine Synthetase from Human Lymphocytes. Purification and Characterization.” *The Journal of Biological Chemistry* 260 (7): 3923–30.

Lawrence, J. G., and Ochman, H. 1997. “Amelioration of Bacterial Genomes: Rates of Change and Exchange.” *Journal of Molecular Evolution* 44 (4): 383–97.

Leander, B. S. 2004. “Did Trypanosomatid Parasites Have Photosynthetic Ancestors?” *Trends in Microbiology* 12 (6): 251–58.

Lee, J. H., Chae, H. S., Hwang, B., Hahn, K. W., Kang, B. G., and Kim, W. T. 1997. "Structure and Expression of Two cDNAs Encoding S-Adenosyl-L-Methionine Synthetase of Rice (Oryza Sativa L.)." *Biochimica et Biophysica Acta* 1354 (1): 13–18.

LeGros, H. L., Halim, A. B., Geller, A. M., and Kotb, M. 2000. "Cloning, Expression, and Functional Characterization of the Beta Regulatory Subunit of Human Methionine Adenosyltransferase (MAT II)." *The Journal of Biological Chemistry* 275 (4): 2359–66.

LeGros, H. L., Geller, A. M., and Kotb, M. 1997. "Differential Regulation of Methionine Adenosyltransferase in Superantigen and Mitogen Stimulated Human T Lymphocytes." *Journal of Biological Chemistry* 272 (25): 16040–47.

Lieber, C. S., and Packer, L. 2002. "S -Adenosylmethionine : Molecular, Biological, and Clinical Aspects-an Introduction." *The American Journal of Clinical Nutrition* 76: 1148–50.

Lu, S. C., Gukovsky, I., Lugea, A., and Reyes, C. N. 2003. "Role of S-Adenosylmethionine in Two Experimental Models of Pancreatitis." *The FASEB Journal*, 56–58.

Lu, Z. J., and Markham, G. D. 2002. "Enzymatic Properties of S-Adenosylmethionine Synthetase from the Archaeon Methanococcus Jannaschii." *The Journal of Biological Chemistry* 277 (19): 16624–31.

Makarova, K. S., Wolf, Y. I., Mekhedov, S. L., Mirkin, B. G., and Koonin, E. V. 2005. "Ancestral Paralogs and Pseudoparalogs and Their Role in the Emergence of the Eukaryotic Cell." *Nucleic Acids Research* 33 (14): 4626–38.

Markham, G. D., DeParasis, J., and Gatmaitan, J. 1984. "The Sequence of metK, the Structural Gene for S-Adenosylmethionine Synthetase in Escherichia Coli." *The Journal of Biological Chemistry* 259 (23): 14505–7.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M., and Penny, D. 2002. "Evolutionary Analysis of Arabidopsis, Cyanobacterial, and Chloroplast Genomes Reveals Plastid Phylogeny and Thousands of Cyanobacterial Genes in the Nucleus." *Proceedings of the National Academy of Sciences of the United States of America* 99 (19): 12246–51.

Mato, J. M., Alvarez, L., Ortiz, P., and Pajares, M. A. 1997. "S-Adenosylmethionine Synthesis: Molecular Mechanisms and Clinical Implications." *Pharmacology & Therapeutics* 73 (3): 265–80.

Mikhailov, K. V., Janouškovec, J., Tikhonenkov, D. V., Mirzaeva, G. S., Diakin, A. Y., Simdyanov, T. G., Mylnikov, A. P., Keeling, P. J., and Aleoshin, V. V. 2014. "A Complex Distribution of Elongation Family GTPases EF1A and EFL in Basal Alveolate Lineages." *Genome Biology and Evolution* 6(9): 2361-7.

Mudd, S. H., and Cantoni, G. L. 1958. “ACTIVATION OF METHIONINE FOR TRANSMETHYLATION : III . THE METHIONINE-ACTIVATING ENZYME OF BAKERS ’ YEAST.” *The Journal of Biological Chemistry* 231: 481–92.

Newman, E. B., Budman, L. I., Chan, E. C., Greene, R. C., Lin, R. T., Woldringh, C. L., and D’Ari, R. 1998. “Lack of S-Adenosylmethionine Results in a Cell Division Defect in Escherichia Coli.” *Journal of Bacteriology* 180 (14): 3614–19.

Noble, G. P., Rogers, M. B., and Keeling, P. J. 2007. “Complex Distribution of EFL and EF-1alpha Proteins in the Green Algal Lineage.” *BMC Evolutionary Biology* 7: 82.

Nordgren K. K. S., Peng Y., Pelleymounter L. L., Moon I., Abo R., Feng Q., Eckloff B., Yee V. C., Wieben E., and Weinshilboum R. M. 2011. “Methionine Adenosyltransferase 2A/2B and Methylation: Gene Sequence Variation and Functional Genomics.” *Drug Metabolism and Disposition* 39 (11): 2135–47.

Nozaki, H., Matsuzaki, M., Takahara, M., Misumi, O., Kuroiwa, H., Hasegawa, M., Shin-i, T., Kohara, Y., Ogasawara, N., and Kuroiwa, T. 2003. “The Phylogenetic Position of Red Algae Revealed by Multiple Nuclear Genes from Mitochondria-Containing Eukaryotes and an Alternative Hypothesis on the Origin of Plastids.” *Journal of Molecular Evolution* 56 (4):485–97.

Ochi, K., and Freese, E. 1982. “A Decrease in S-Adenosylmethionine Synthetase Activity Increases the Probability of Spontaneous Sporulation.” *Journal of Bacteriology* 152 (1): 400–410.

Okamoto, S., Lezhava, A., Hosaka, T., Okamoto-Hosoya, Y., and Ochi, K. 2003. “Enhanced Expression of S -Adenosylmethionine Synthetase Causes Overproduction of Actinorhodin in Enhanced Expression of S -Adenosylmethionine Synthetase Causes Overproduction of Actinorhodin in Streptomyces Coelicolor A3 (2)†.” *Journal of Bacteriology* 185 (2): 601–9.

Page, R. D., and Charleston, M. A. 1997. “From Gene to Organismal Phylogeny: Reconciled Trees and the Gene Tree/species Tree Problem.” *Molecular Phylogenetics and Evolution* 7 (2): 231–40.

Pajares, M. A., and Markham, G. D. 2011. “Methionine Adenosyltransferase (S-Adenosylmethionine Synthetase).” In *Advances in Enzymology*, 449–521.

Palenik, B. 2002. “The Genomics of Symbiosis: Hosts Keep the Baby and the Bath Water.” *Proceedings of the National Academy of Sciences of the United States of America* 99 (19): 11996–97.

Park, H. S., Shin, S. K., Yang, Y. Y., Kwon, H. J., and Suh, J. W. 2005. “Accumulation of S-Adenosylmethionine Induced Oligopeptide Transporters Including BldK to Regulate Differentiation Events in Streptomyces Coelicolor M145.” *FEMS Microbiology Letters* 249 (2): 199–206.

Patron, N. J., Rogers, M. B., and Keeling, P. J. 2006. “Comparative Rates of Evolution in Endosymbiotic Nuclear Genomes.” *BMC Evolutionary Biology* 6: 46.

Peleman, J., Boerjan, W., Engler, G., Seurinck, J., Botterman, J., Alliotte, T., Van Montagu, M., and Inzé, D. 1989. “Strong Cellular Preference in the Expression of a Housekeeping Gene of Arabidopsis Thaliana Encoding S-Adenosylmethionine Synthetase.” *The Plant Cell* 1: 81–93.

Pombert, J. F., James, E. R., Janouškovec, J., and Keeling, P. J. 2012. “Evidence for Transitional Stages in the Evolution of Euglenid Group II Introns and Twintrons in the Monomorphina Aenigmatica Plastid Genome.” *PloS One* 7 (12): e53433.

Rouillon, A., Surdin-kerjan, Y., and Thomas, D. 1999. “Transport of Sulfonium Compounds: CHARACTERIZATION OF THE S-ADENOSYLMETHIONINE PERMEASES FROM THE YEAST SACCHAROMYCES CEREVISIAE*.” *Journal of Biological Chemistry* 274 (40): 28096–105.

Sakaguchi, M., Takishita, K., Matsumoto, T., Hashimoto, T., and Inagaki, Y. 2009. “Tracing Back EFL Gene Evolution in the Cryptomonads-Haptophytes Assemblage: Separate Origins of EFL Genes in Haptophytes, Photosynthetic Cryptomonads, and Goniomonads.” *Gene* 441: 126–31.

Sánchez-Pérez, G. F., Bautista, J. M., and Pajares, M. A. 2004. “Methionine Adenosyltransferase as a Useful Molecular Systematics Tool Revealed by Phylogenetic and Structural Analyses.” *Journal of Molecular Biology* 335 (3): 693–706.

Sánchez-Pérez, G. F., Hampl, V., Simpson, A. G. B., and Roger, A. J. 2008. “A New Divergent Type of Eukaryotic Methionine Adenosyltransferase Is Present in Multiple Distantly Related Secondary Algal Lineages.” *The Journal of Eukaryotic Microbiology* 55 (5): 374–81.

Schlesier, J., Siegrist, J., Gerhardt, S., Erb, A., Blaesi, S., Richter, M., Einsle, O., and Andexer, J. N. 2013. “Structural and Functional Characterisation of the Methionine Adenosyltransferase from Thermococcus Kodakarensis.” *BMC Structural Biology* 13 (1). BMC Structural Biology: 22.

Schröder, G., Eichel, J., Breinig, S., and Schröder, J. 1997. “Three Differentially Expressed S-Adenosylmethionine Synthetases from Catharanthus Roseus: Molecular and Functional Characterization.” *Plant Molecular Biology* 33 (2): 211–22.

Soukal, P. 2013. “Search for the Remnant of Plastid in the Cell of Rhabdomonas Sp.”

Syvanen, M. 1994. “Horizontal Gene Transfer: Evidence and Possible Consequences.” *Annual Review of Genetics* 28: 237–61.

Szabová, J., Yubuki, N., Leander, B. S., Triemer, R. E., and Hampl, V. 2014. “The Evolution of Paralogous Enzymes MAT and MATX within the Euglenida and beyond.” *BMC Evolutionary Biology* 14 (25).

Tabor, C. W., and Tabor, H. 1984. “Polyamines.” *Annual Review of Biochemistry* 53 (1):749–90.

Takusagawa, F., Kamitori, S., Misaki, S., and Markham, G. D. 1996. “Crystal Structure of S-Adenosylmethionine Synthetase.” *The Journal of Biological Chemistry* 271 (1): 136–47.

Taylor, J. C., and Markham, G. D. 2000. “The Bifunctional Active Site of S-Adenosylmethionine Synthetase. ROLES OF THE BASIC RESIDUES.” *Journal of Biological Chemistry* 275 (6): 4060–65.

Taylor, J. S., and Raes, J. 2004. “Duplication and Divergence: The Evolution of New Genes and Old Ideas.” *Annual Review of Genetics* 38: 615–43.

Turmel, M., Gagnon, M. C., O’Kelly, C. J., Otis, C., and Lemieux, C. 2009. “The Chloroplast Genomes of the Green Algae Pyramimonas, Monomastix, and Pycnococcus Shed New Light on the Evolutionary History of Prasinophytes and the Origin of the Secondary Chloroplasts of Euglenids.” *Molecular Biology and Evolution* 26 (3): 631–48.

Wagner, A. 2010. “On the Energy and Material Cost of Gene Duplication.” In *Evolution after Gene Duplication*, 207–14.

Whittaker, D. J., Smith, C. S., and Cardner, R. C. 1995. “Three cDNAs Encoding S-Adenosyl-I-Methionine Synthetase from Actinidia Chinensis.” *Plant Physiology* 108: 1307–8.

Williams, D., Andam, C. P., and Gogarten, J. P. 2010. “Horizontal Gene Transfer and the Formation of Groups of Microorganisms.” In *Molecular Phylogeny of Microorganisms*.

Yamaguchi, A., Yubuki, N., and Leander, B. S. 2012. “Morphostasis in a Novel Eukaryote Illuminates the Evolutionary Transition from Phagotrophy to Phototrophy: Description of Rapaza Viridis N. Gen. et Sp. (Euglenozoa, Euglenida).” *BMC Evolutionary Biology* 12: 29.

Yocum, R. R., Perkins, J. B., Howitt, C. L., and Pero, J. 1996. “Cloning and Characterization of the metE Gene Encoding S-Adenosylmethionine Synthetase from Bacillus Subtilis.” *Journal of Bacteriology* 178 (15): 4604–10.