J Mol Evol (2004) 58:712–721
DOI: 10.1007/s00239-004-2593-z

JOURNAL OF MOLECULAR EVOLUTION

Presequence Acquisition During Secondary Endocytobiosis and the Possible Role of Introns

Oliver Kilian, Peter G. Kroth

Fachbereich Biologie, Universität Konstanz, 78457 Konstanz, Germany

Received: 23 October 2003 / Accepted: 9 December 2003

Abstract. Targeting of nucleus-encoded proteins into chloroplasts is mediated by N-terminal presequences. During evolution of plastids from formerly free-living cyanobacteria by endocytobiosis, genes for most plastid proteins have been transferred from the plastid genome to the nucleus and subsequently had to be equipped with such plastid targeting sequences. So far it is unclear how the gene domains coding for presequences and the respective mature proteins may have been assembled. While land plant plastids are supposed to originate from a primary endocytobiosis event (a prokaryotic cyanobacterium was taken up by a eukaryotic cell), organisms with secondary plastids like diatoms experienced a second endocytobiosis step involving a eukaryotic alga taken up by a eukaryotic host cell. In this group of algae, apparently most genes encoding chloroplast proteins have been transferred a second time (from the nucleus of the endosymbiont to the nucleus of the secondary host) and thus must have been equipped with additional targeting signals. We have analyzed cDNAs and the respective genomic DNA fragments of seven plastid preproteins from the diatom *Phaeodactylum tricornutum*. In all of these genes we found single spliceosomal introns, generally located within the region coding for the N-terminal plastid targeting sequences or shortly downstream of it. The positions of the introns can be related to the putative phylogenetic histories of the respective genes, indicating that the bipartite targeting sequences in these secondary algae might have evolved by recombination events via introns.

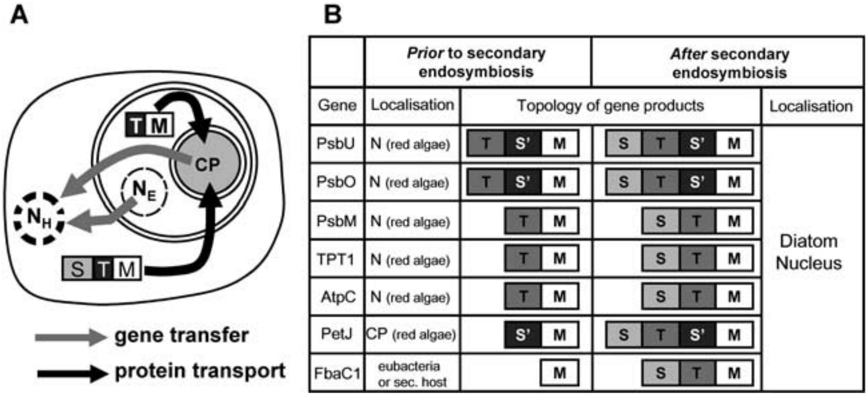
Key words: Chloroplast — Diatom — Endocytobiosis — Intron — Presequence — Targeting

Introduction

It is now widely accepted that mitochondria and chloroplasts arose by independent primary endocytobioses, that is by uptake of a free-living proteobacterium or a cyanobacterium, respectively, by a eukaryotic host cell (Delwiche and Palmer 1997; Martin et al. 1998; Martin and Müller 1998). Furthermore, phylogenetic analyses support the view that the plastids of red algae, glaucophytes, green algae, and land plants may be traced back to a single endocytobiotic event (Delwiche and Palmer 1997; Moreira et al. 2000; Palmer 2003). The subsequent transformation and reduction of the prokaryotic endosymbiont into an organelle led to the extinction of 90–95% of the genes from the endosymbiont compared to free-living cyanobacteria, most of them being transferred to the nucleus of the host cell (Martin et al. 2002). However, the protein content of the endosymbionts/organelles was not reduced that drastically: A large number of organelle specific reactions like photosynthesis and the production of typical secondary plant metabolites still take place

The nucleotide sequences have been deposited at GenBank under accession numbers AY191862, AY191863, AY191864, AY191865, AY191866, AY191867, and AY191868.

Correspondence to: Peter G. Kroth; *email:* Peter.Kroth@uni-konstanz.de



	Prior to secondary endosymbiosis		*After* secondary endosymbiosis	
Gene	Localisation	Topology of gene products		Localisation
PsbU	N (red algae)	T S' M	S T S' M	Diatom Nucleus
PsbO	N (red algae)	T S' M	S T S' M	
PsbM	N (red algae)	T M	S T M	
TPT1	N (red algae)	T M	S T M	
AtpC	N (red algae)	T M	S T M	
PetJ	CP (red algae)	S' M	S T S' M	
FbaC1	eubacteria or sec. host	M	S T M	

Fig. 1. Schematic drawing describing the putative situation during establishment of secondary endocytobiosis. **A** Genes for plastid proteins located either on the chloroplast (CP) genome or the nucleus of the endosymbiont (N_E) were transferred into the nucleus of the secondary host (N_H) (gray arrows). Gene products of these genes have to be transported into the four membrane-bound plastids depending on additional targeting signals (black arrows; see text). S, signal peptide; T, transit peptide; M, mature protein. **B** Suggested topology and location of gene products prior (in red algae or secondary host) and after secondary endocytobiosis (in diatoms). Abbreviations as in A. S′, thylakoidal signal peptide. The topology in red algae is based on known genes: PsbU (GenBank AB023805), PsbO (AV432421), PsbM (AV438629), Tpt1 (the presence of a transit peptide is based on the absence of this gene on red algal plastid genomes and on the analysis of incomplete *Porlhyra* EST fragments [av435094, av432927]), AtpC (AV433223), PetJ (NC_000925, AB040818, AB040818), and FBA (gene phylogenies indicate that the diatom plastid aldolase is related to cytosolic enzymes of the putative secondary host or to bacterial enzymes [Kroth, Schroers, and Kilian, unpublished]).

inside the chloroplasts. A recent study estimated that ~4500 genes in the *Arabidopsis thaliana* genome have been acquired from the former cyanobacterium, whereas the gene products of almost half of those is apparently targeted back into the plastids (Martin et al. 2002), mediated by translocators embedded in the plastid envelope membranes (Keegstra and Cline 1999; Jarvis and Soll 2001).

Plastid-specific protein targeting in land plants occurs posttranslationally utilizing an N-terminal presequence (“transit peptide”), which is cleaved off after the successful transport of the protein into the organelle (Van der Vere et al. 1995; Bruce 2000). The coding sequences for transit peptides supposedly were not included in the respective genes before they were transferred from the genome of the endosymbiont to the nucleus, thus the respective sequence information must have been acquired during or after integration of the genes into the host genome. It is still unclear whether the genetic domains of transit peptides may have been fused to genes of plastidic proteins or whether they evolved by mutation. In a few cases the presence of introns in the respective regions indicates that processes like exon shuffling may be involved (Gantt et al. 1991; Waller et al. 1998). In mitochondria there is also some evidence that processes like simple DNA recombination or exon shuffling might have been the driving force toward the phylogenetic fusion of transit peptides and mature proteins (Wischmann and Schuster 1995; Long et al. 1996; Kubo et al. 1999; Adams et al. 2002). In metazoans the latter process, which is mediated by recombination between introns, is thought to have played an important role in fusion of gene modules encoding functional domains of different enzymes (Patthy 1999). Problems for identifying possible indicative introns (which originally might have been involved in exon shuffling) mainly arise from additional intron integration/excision processes that occurred after the primary endocytobiosis, resulting in a large number of introns per gene as observed in *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000).

A variety of algal genera like diatoms, brown algae, cryptophytes, and euglenoids has evolved by a process termed secondary endocytobiosis (Delwiche and Palmer 1997; Cavalier-Smith 2000; McFadden 2001). In this process, which is supposed to have happened successfully at least twice during evolution, eukaryotic algae (ancestors of red and green algae) were engulfed by eukaryotic host cells and transformed into plastids (Fig. 1A). Algae that evolved this way are typically characterized by possessing “complex” plastids that are surrounded by more than two—usually three or four—envelope membranes (van Dooren et al. 2001). In most of those algae, including diatoms, secondary endocytobiosis led to a complete extinction of the nucleus of the endosymbiont; only in cryptophytes and chlorarachniophytes is a drastically reduced genome still present between the second and the third membrane of the respective chloroplast envelope (Douglas et al. 2001; Gilson and McFadden 2002). As a consequence, most of the genes encoding plastid proteins formerly located in the nucleus of the endosymbiont had to be transferred to the nucleus of the new host to preserve functionality of the plastids (Fig. 1A). This second round of intracellular gene translocation resulted in a new challenge: New transport pathways and the respective targeting signals had to be developed to enable targeting of plastid proteins across the additional plastid envelope membranes (Cavalier-Smith 1999; van Dooren et al. 2001; Kroth 2002). Nucleus-encoded chloroplast proteins in these organisms have bipartite presequences, consisting of a signal peptide followed by a transit peptide (Fig. 1A). The functionality of these two domains has been demonstrated for various organisms with secondary plastids (Lang et al. 1998; Sulli et al. 1999; Waller et al. 2000). In diatoms apparently a first cotranslational transport step (presumably at the outer endoplasmic reticulum [ER] membrane surrounding the plastid, called chloroplast ER [CER] [Gibbs 1981]) and at least one further translocation step are involved (Bhaya and Grossman 1991; Lang et al. 1998); it is still unclear, however, whether additional targeting signals within the presequences are needed.

So far there are only a few indications how the bipartite presequences of nucleus-encoded proteins destined for secondary plastids may have been assembled, but similarly to the obvious acquisition of transit peptides in the course of primary endocytobiosis, the recombinatory addition of a gene fragment encoding either a signal peptide or a complete bipartite presequence (Waller et al. 1998; Schaap et al. 2001) may have contributed to the huge task of supplying a high number of transferred genes with adequate presequence domains.

We have analyzed seven genes from the diatom *Phaeodactylum tricornutum* and found single introns that might reflect the origin of the individual targeting domains and the way that they were fused to previously relocated genes in organisms with secondary plastids.

Materials and Methods

Culture Conditions and Cell Harvesting

Phaeodactylum tricornutum Bohlin (University of Texas Culture Collection, strain 646) was cultivated as described previously (Apt et al. 2002) under continuous light (35 μmol photons m^{-2} s^{-1}). Cells were grown to log phase and directly subjected to fluorescence microscopy. For nucleic acid isolation or for transformation the algae were harvested by centrifugation at room temperature in 500-ml beakers (10 min, 3.000*g*).

RNA Isolation and mRNA Purification

Cells were harvested as described above, frozen in liquid N_2, and crushed in a mortar. Four volumes of prewarmed (50°C) homogenization medium (0.33 *M* sorbitol, 0.2 *M* Tris–Cl, pH 8.5, 10 m*M* EDTA, 10 m*M* EGTA, 2% SDS, 0.1% diethylpyrocarbonate) and 2 vol warm (50°C) TE-saturated phenol (supplemented with 0.2% β-mercaptoethanol) were added and the mixture was rigorously shaken for 20 min. After centrifugation for 15 min at 3000*g*, the upper phase was extracted with 3 vol of chloroform and was centrifuged again. The water-soluble phase was isolated and precipitated overnight at −20°C by adding 1 vol of cold isopropanol and 1/20 vol of sodium acetate (pH 6.0). Precipitated nucleic acids were collected by centrifugation (30 min at 10,000*g*), gently dried, resuspended, and subjected to LiCl precipitation (2 *M* LiCl, 12 h of incubation on ice). The RNA was collected by centrifugation (4°C, 30 min at 10,000*g*), washed with 70% ethanol and resuspended in water containing 0.1% DEPC. mRNA enrichment was performed using the Oligotex mRNA Purification System (Qiagen, Hilden) according to the manual.

DNA Isolation

Cells were pelleted as described above and 1 vol of 2× CTAB buffer (100 m*M*, Tris/HCl, pH 8, 1.5 *M* NaCl, 20 m*M* EDTA, 2% cetyltrimethylammonium bromide, supplemented with 0.2% β-mercaptoethanol prior to use) was added. The mixture was incubated at 65°C under gentle shaking, supplemented with 1 vol chloroform:isoamyl alcohol, and incubated for 1 additional h under gentle shaking at room temperature. After centrifugation (15 min, 3000*g*, room temperature) the liquid phase was mixed with the same volume of isopropanol and the DNA was precipitated by incubation at −20°C overnight. The DNA was pelleted by centrifugation (4°C, 30 min, 10,000*g*) and subsequently washed with ethanol (70%). After gentle air-drying the DNA was resuspended in TE buffer and stored at 4°C.

Library Construction and Sequencing

Purified mRNA has been reverse transcribed and cloned into Lambda Zap vector by the aid of the Lambda ZAP-CMV XR Library Construction Kit (Stratagene, La Jolla, CA) according to the instructions of the manufacturer. Isolated plasmids have been subjected to mass sequencing.

Construction of GFP-Fusion Proteins and Transformation into Diatoms

The gene for the enhanced green fluorescence protein (EGFP; Clontech, Palo Alto, CA) was fused with DNA sequences encoding complete presequences by using *Nco*I restriction sites, cloned into the pPha-T1-vector, and transformed into *Phaeodactylum tricornutum* as described previously (Zaslavskaia et al. 2000). Fluorescence microscopy of transformed diatoms was done with an Olympus BX51 fluorescence microscope equipped with a Nikon DMX-1200 camera using the filter sets HQ480/20 for GFP and U-MWSG2 for chlorophyll fluorescence (Olympus, Hamburg).

Isolation of Genomic Sequences

Genomic sequences corresponding to the previously isolated cDNA clones were isolated by standard PCR techniques using Taq polymerase (Eppendorf, Hamburg). The oligonucleotides (22 and 24 nucleotides length) matched the 5′ and 3′ untranslated regions of the corresponding cDNAs. PCR fragments were separated by agarose gel electrophoresis, isolated, cloned into the pGEM-T vector (Promega, Madison, WI), and sequenced.

Results and Discussion

Isolation of Diatom Genes Encoding Organellar Proteins and Prediction of the Subcellular Localization of their Products

The nucleotide sequences of seven cDNA fragments from the diatom *Phaeodactylum tricornutum* were determined: PsbU, extrinsic 12 kDa protein of photosystem II (PSBU); PsbO, oxygen-evolving enhancer protein 1 (OEE1); PsbM, subunit M of photosystem II (PSBM); Tpt1, triose phosphate/phosphate translocator (TPT1); AtpC, γ subunit of chloroplast ATPase (ATPC) (Apt et al. 2002); PetJ, cytochrome c_{553} (PETJ); and FbaC1, fructose-1,6-bisphosphate aldolase (FBAC1). Six of them were found to encode plastid proteins (being homologous to higher plant proteins), whereas the FbaC1 gene encodes a fructose-1,6-bisphosphate aldolase (FBA), which is homologous to bacterial and fungal class II FBA enzymes (see below). We further amplified and characterized the corresponding genomic DNA fragments by PCR techniques (the genomic sequence of PetJ has also been deposited in GenBank by other authors, under accession number AB078084). All of the deduced amino acid sequences contain similar N-terminal bipartite (FBAC1, TPT1, PSBM, γ subunit of chloroplast ATPase) or tripartite (PSBU, OEE1, PETJ) presequences (Fig. 3A and B). In diatoms the bipartite presequences are necessary for targeting of proteins from the cytosol to the plastid stroma, whereas tripartite leader peptides—consisting of a eukaryotic signal peptide, a plastidic transit peptide, and a prokaryotic signal peptide—are essential for correct targeting into the thylakoid lumen or the thylakoid membranes. Such complex tripartite presequences have also been found in other organisms with secondary plastids like *Euglena gracilis* (Vacula et al. 1999) or in dinoflagellates (Norris and Miller 1994). Computer analysis of the derived presequences by the program SignalP (http://www.cbs.dtu.dk/services/SignalP/[Nielsen et al. 1999]) in all cases predicted the presence of one N-terminal signal peptide which was supported by high probability values. The regions of the presequence cleavage sites in diatom plastid preproteins are very conserved and usually contain the motif ASAF or AFAP (Kroth 2002). The second presequence domain following the N-terminal signal peptide was analyzed by ChloroP (http://www.cbs.dtu.dk/services/ChloroP/[Emanuelsson et al. 1999]) and resulted in moderate scores for transit peptides. However, the ChloroP-derived prediction of the stromal peptidase cleavage site probably is not accurate, as the neural network used in ChloroP is trained on higher plant sequences. Earlier experiments, however, already had demonstrated that diatom transit peptides are functionally and structurally related to transit peptides of land plants (Lang et al. 1998). For thylakoidal proteins with tripartite presequences the additional third presequence domains were clearly identified as prokaryotic signal peptides by the SignalP program. In all cases the predicted transit peptide cleavage site is located in regions where the corresponding mature homologous proteins from other organisms have been processed, if preceded by a presequence at all.

In order to demonstrate the functionality of the plastid targeting sequences in vivo, we genetically fused all seven presequences (of AtpC, FbaC1, Tpt1, PsbO, PsbM, PsbU, PetJ) individually to the N-terminus of GFP as described in Materials and Methods and cloned these constructs into a diatom transformation vector. After transformation and selection of *Phaeodactylum* cells as described earlier (Apt et al. 2002) we analyzed those strains expressing GFP by microscopical analyses. We found GFP fluorescence in cells transformed with four of the constructs (AtpC, FbaC1, Tpt1, PsbO fused to GFP) to be generally located within the plastids (Fig. 2). For the OEE1:GFP construct we were also able to demonstrate the localization of GFP within the diatom thylakoids by purification and analysis of intact thylakoids (Ammon, Kilian, and Kroth, unpublished). Although we performed several transformation experiments, we never obtained transformants expressing PsbM:GFP, PsbU:GFP, and PetJ:GFP.

Single Spliceosomal Introns Located Within the Regions of Genes Encoding the Presequence of Plastid Preproteins

A comparison of the cDNA sequences of the seven genes with their respective genomic DNA fragments obtained by PCR revealed the presence of one single spliceosomal intron in each gene. Interestingly, all introns were found in regions coding for presequences or the very N-terminus of the mature proteins as indicated in Fig. 3A and B. All of them contain 5′-GT and 3′-AG borders which are typical for spliceosomal introns (Tolstrup et al. 1997). Additionally, polypyrimidine stretches are present in the regions adjacent to the 3′-acceptor sites. The intron/exon borders are poorly conserved; the derived consensus sequence for the sample set is ---(G/C)*$\underline{\textit{GT}}$XX(A/G)(T/C) --------- (T/C/G)(G/C/A)C$\underline{\textit{AG}}$* --- (* marks the splice positions). Intron size ranges from 182 to 632 bp. Putative branch point sequences (bps) are present at a distance of 27–36 bp upstream of the acceptor site. The consensus sequence of the bps is NNCT(G/C)A(T/C) and is not as strictly conserved as observed in yeast (TACTAAC) (Tolstrup et al. 1997). Four introns (of seven introns analyzed) are in phase 0. Nuclear genes of *P. tricornutum* generally have a relatively high GC content, about 55% (Scala et al. 2002). The GC content of the analyzed introns is about 10% lower than that of the adjacent downstream exons and most of the exons positioned upstream of the respective introns have a GC content comparable to that of the downstream exon. Comparisons in pairs and total alignments of the introns did not reveal significantly homologous regions.

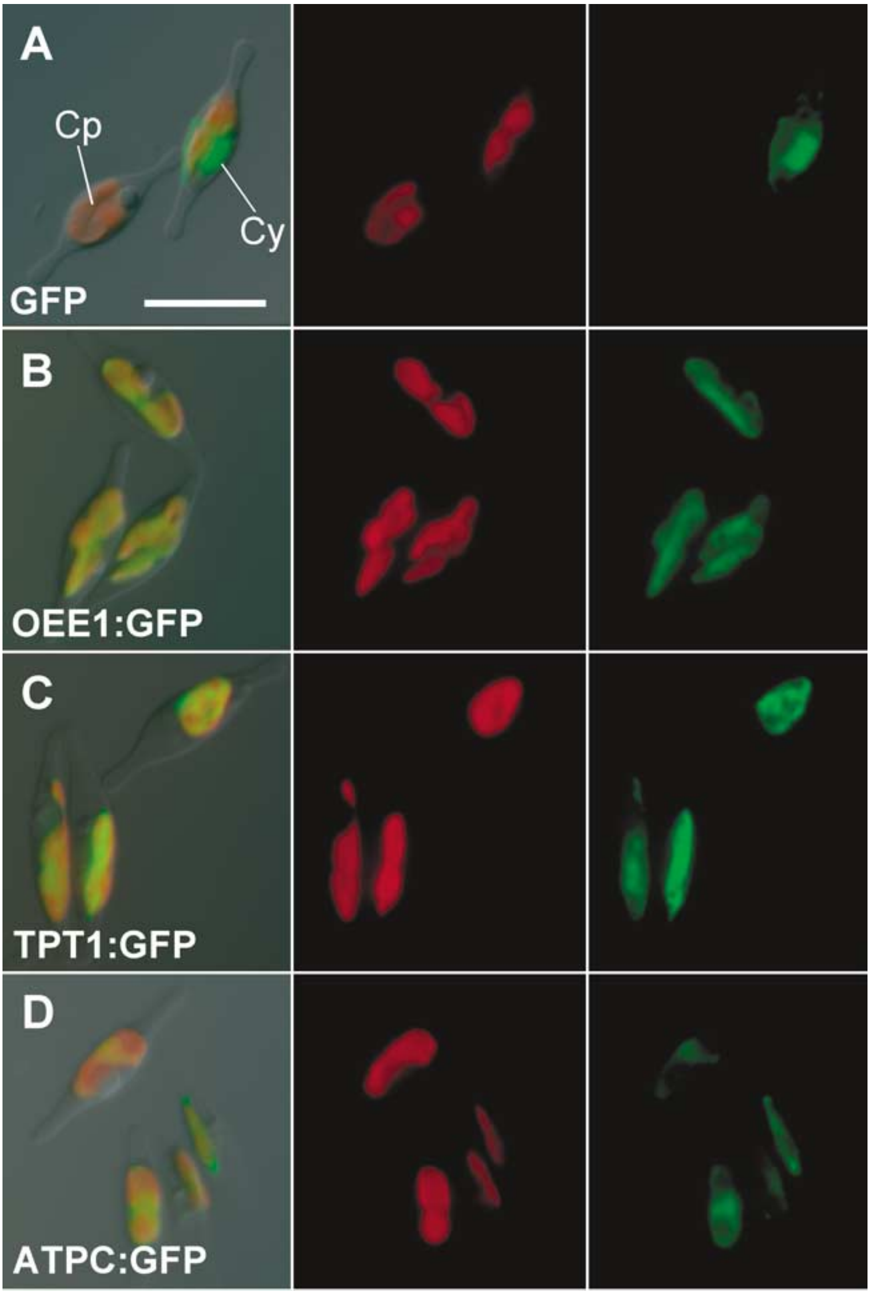
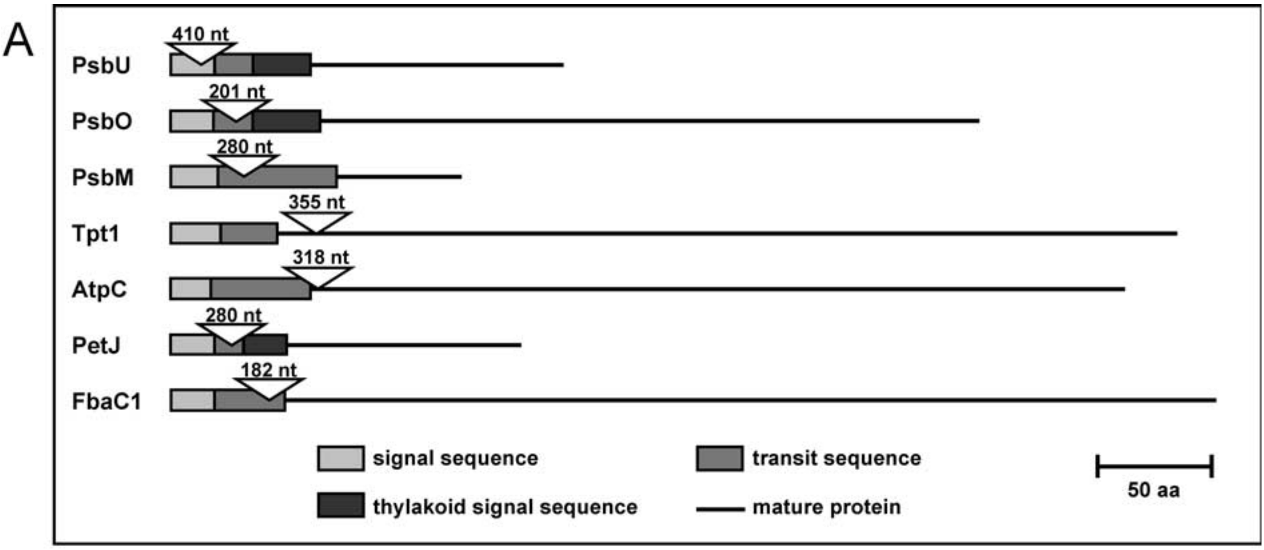


Fig. 2. Localization of GFP fusion proteins in *Phaeodactylum tricornutum*. Cells expressing GFP without any presequence (**A:** control), and presequence:GFP fusion proteins (**B**–**E**) have been analyzed by bright-field and fluorescence microscopy. In the second and third rows, red chlorophyll autofluorescence and green GFP fluorescence are shown, respectively. In the first row, images obtained by bright-field and fluorescence microscopy of the same cells are merged, clearly showing that in B–E chlorophyll and GFP fluorescence are colocalized and that the GFP fusion proteins accumulate within the diatom chloroplasts (Cp), whereas unaltered GFP (A) accumulates within the cytosol (Cy). Scale bar: 10 μm.

Different Routes of Gene Traffic During the Establishment of Secondary Endocytobiosis

From Fig. 1B it is obvious that genes transferred during secondary endocytobiosis must have been equipped with additional targeting signals. As described below a complete bipartite presequence had to be fused to the FbaC1 and PetJ proteins to allow targeting into the secondary plastids; in all other cases only a signal peptide had to be added (either by



B

Gene	AA sequence	Phase
PsbU	**MKLAVFAVLIS** ▽ **TVAS***FVAPNGVQRAATTELNA▼*ERREFLSAAAVAAGLAFPLTAN*-AIRDYENVGYLGG·	2
PsbO	**MKFTAACSLALVASASA***FAPIPSVSVS ▽ RTTDLSM▼*SLQKDLANVGKVAAAGALAFGLATAPANAL*-TKS·	0
PsbM	**MKSFQLLTLFALIAASLA***FAPNQAPQQ ▽ VAKAAFTKAAGALPVAALAAPAFALDSIVEALPTNTVA▼LEV·	1
Tpt1	**MKVATTLTLAFICCASA***FGLNGQTTSVMKKVGFGAGSKPM▼VQAIDVQGNRLGSNM ▽ QPLKSAVANEDAP·	0
AtpC	**MRSFCIAALFAVASA***FTTQPTSFTVKTANVGERASGVFPEQSSAHRTRKATIVM▼DGKAN ▽ AIRDRITSV·	1
PetJ	**MKLAVIATLLATASA***FSIQAEFSK ▽ VAKG▼*AAAVGVGAVIAAAPALAG*-DVGAGEQIFNANCAACHAGGQ·	0
FbaC1	**MKLSTAALFFIPAVVA***FAPPQAAFRSNRALFATETAAEKT ▽ TFSKM▼PASVKPGVVTGKALVDLLQHAKK·	0

Fig. 3. Position of introns found in this work. **A** Intron localization in genomic DNA sequences of *Phaeodactylum tricornutum* analyzed in this work. The domains within the deduced amino acid sequences are shown; the length of the scale bar represents 50 amino acids. Shaded boxes stand for the respective targeting domains as indicated below. Triangles represent the positions of introns in the respective genes; the intron sizes in nucleotides (nt) are shown above. **B** Deduced amino acid sequences of the N-terminal parts of diatom plastid preproteins refer to the gene names. Signal peptides, transit peptides, and thylakoid targeting signal peptides are printed in boldface, underlined, and italic letters, respectively. Symbols indicate putative protease processing sites for the signal peptide (∗), the transit peptide (▼), and—if present—the thylakoid signal peptide (–). Open triangles indicate the intron positions within the respective DNA sequences. The individual phases of the introns are indicated.

direct addition of a signal peptide or by exchange of the existing transit peptide by a bipartite presequence). Figure 1B is based on the assumption that the ancestor of diatom plastids essentially had the same distribution of nuclear and plastid genes as modern red algae. According to this we can assign three types of gene transfers/substitutions.

Transfer of Nuclear Genes Within the Same Cell

Genes related to photosynthesis (PetJ, PsbU, PsbM, PsbO, AtpC) most likely originate from the cyanobacterium taken up during primary endocytobiosis. All these genes, with the exception of PetJ, are nucleus encoded in red algae as evidenced by available public sequence databases (see legend to Fig. 1. for details). As these genes are also nucleus encoded in diatoms, we conclude that PsbO, PsbM, PsbU, and AtpC were transferred twice in evolution: in the first step, from the cyanobacterial genome to the nucleus of the first host, and then in the second step, to the nucleus of the second host cell. The triose phosphate/phosphate translocator TPT1 apparently derived from the host of the primary endocytobiosis, as there is no known prokaryotic counterpart; therefore the respective gene probably should already have been present in the nuclear genome of the first host.

Transfer of a Chloroplast Gene into the Host Nucleus

PetJ is still located in the chloroplast genome of red algae (Reith and Munholland 1995), therefore it probably was also there in the ancestor of diatom plastids. This would mean that the gene was transferred directly from the plastid genome into the nuclear genome of the secondary host.

Gene Substitution

FBAs occur in two forms, called class I and class II. Class I FBAs are usually found in the cytosol as well as in the plastids of higher plants. The more ancient class II FBAs are common in bacteria, fungi, and oomycetes; the last group of organisms is supposed to be related to the host of the secondary endocytobiosis leading to diatoms and related algae (Plaumann et al. 1997). Phylogenetic and enzymatic analyses clearly indicate that the nucleus-encoded FbaC1 gene presented here encodes a class II enzyme that is targeted to the plastid (Kroth, Schroers, and Kilian, unpublished). Therefore, either a gene encoding a class II aldolase originating from the secondary host may have been duplicated or a eubacterial gene may have been recruited by lateral gene transfer, followed by a redirection of the gene product into the plastid, replacing the original aldolase of the endosymbiont. However, in both cases a complete bipartite presequence had to be added.

Introns Dissecting Chloroplast Genes May Be Remnants of Recombination Events

One of the key questions is whether the introns are related to recombination events which, during secondary endocytobiosis, led to the addition of presequences to the gene products after relocation of the respective genes to the nucleus of the host. If so, they should have been introduced after secondary endocytobiosis. We compared the location of introns in respective genes from *Arabidopsis* to see whether they might share a common history. In some of the respective genes from *Arabidopsis* (AtpC, Tpt1) there are no introns. Other genes do not share a common history: PsbM is plastid encoded in land plants, and PetJ is nucleus encoded in *Arabidopsis* but still plastid encoded in red algae; there is no homologous gene for plastidic class II aldolases in *Arabidopsis*, whereas PsbU was not found on the *Arabidopsis* genome. Thus the only possibly relevant introns are present in PsbO of *Arabidopsis*, however, the positions do not match to the intron position in the respective diatom gene.

There are at least two scenarios to explain the fusion of coding DNA sequences: direct recombination of coding sequences within exons and recombination within introns via exon shuffling. While simple recombination merely depends on random in-frame combination of coding DNA fragments, exon shuffling is supposed to occur locally within introns which are present in both genes involved in this process (Patthy 1999). Recombination within these intron regions may result in fusion of protein domains provided that the intron/exon borders are in the same phase. Is it likely that presequence acquisition during secondary endocytobiosis in the predecessors of diatoms occurred with the assistance of introns? If not, the respective transferred genes may have integrated randomly into preexisting genes already preceded by sequence information encoding the needed targeting signal. Indeed, such a phenomenon has been observed by Adams and coworkers (2000), who identified several recent independent losses and transfers of mitochondrial genes in a variety of angiosperm genera. There are numerous examples in different genera demonstrating that gene products of individually transferred mitochondrial ribosomal protein genes (rps) were redirected into mitochondria (Kadowaki et al. 1996; Kubo et al. 1999, 2000; Adams et al. 2002). For example, mitochondrial import signals were generated by activation of internal transport signals (Kubo et al. 2000) or by random in-frame recombination with other nuclear genes encoding mitochondrial proteins already equipped with transit peptides. An indicator for such "parasitism" of preexisting nuclear genes can be the inclusion of parts of the transit-peptide donating gene within the resulting gene as demonstrated by Adams et al. (2000). In diatom plastid preproteins, however, we were not able to identify any remnants of such recombination events, indicating that the signal peptides may not have been generated this way. Also, regions within the mature protein that by chance might act as transport signals can be ruled out, because all nucleus-encoded chloroplast preproteins in diatoms identified so far are preceded by a signal sequence. Finally, individual transformation of transit peptide domains into bipartite presequences simply by mutation is also unlikely regarding the large number of the genes (approx. 1000 to 3000) that had to acquire a signal peptide after secondary endocytobiosis.

Therefore the singular occurrence of introns and their position within the presequence-encoding domains in all diatom genes analyzed in this work indicate a participation of these introns in the process of targeting sequence addition. Single introns in the region coding for presequences of plastid proteins have also been demonstrated in other organisms with secondary plastids. In the brown alga *Laminaria* one intron was found in the gene for an FCP protein (Caron et al. 1996) as also in a GAPDH gene of *Phaeodactylum* (Liaud et al. 2000). In the malaria parasite *Plasmodium* introns also were shown in apicoplast precursor proteins and have been discussed to be involved in exon shuffling (Waller et al. 1998; Schaap et al. 2002), however, in these genes several introns are present.

A classical exon shuffling process, however, is rather unlikely, as it would imply that not only the donor genes but also all relocated acceptor genes needed suitable introns. So are there any other possible ex-

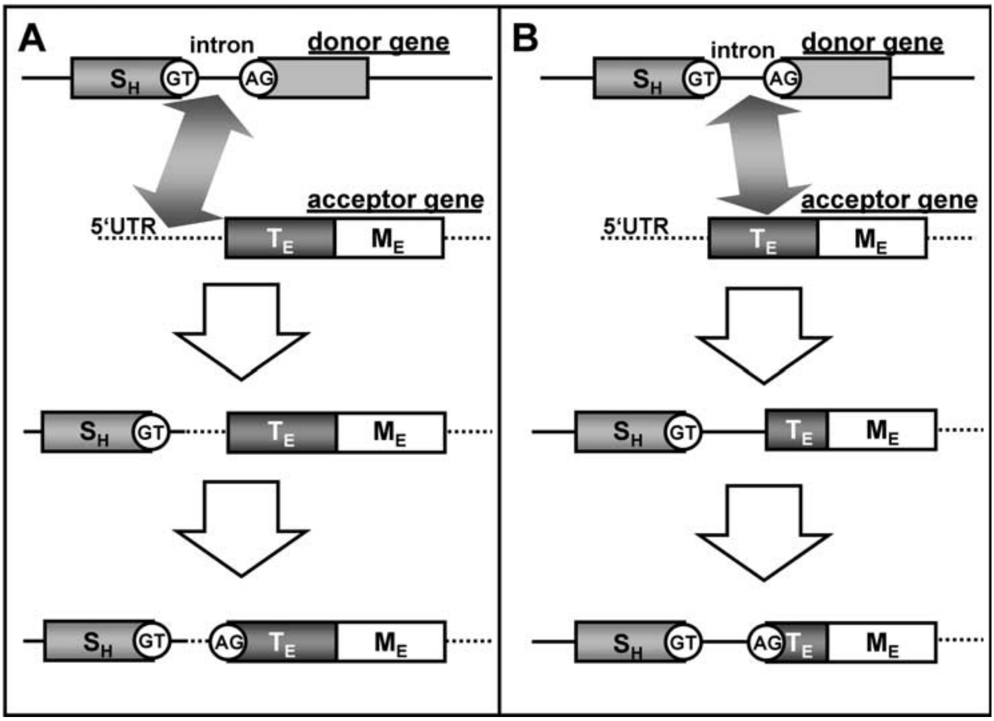


Fig. 4. Model for recombination events ("semi-exon shuffling") which could explain the results discussed in this work. The addition of a gene encoding a signal peptide to an acceptor gene already preceded by a transit peptide-encoding region is shown: recombination between a donor gene (consisting of a region encoding a signal peptide [S_H], an intron, and the region for the mature protein) and an acceptor gene (without intron) may result in the addition of a signal peptide. The origins of the introns and the 5′-untranslated regions (5′-UTR) are indicated by solid lines (donor gene) and by broken lines (acceptor gene), respectively. Intron borders are indicated by encircled GT (5′) and AG (3′). Recombination might occur between the intron and (**A**) the 5′-UTR or (**B**) the sequence encoding the transit peptide domain. Next, new 3′-intron borders may be generated by utilizing random AG nucleotides. Boxes labeled S, T, and M represent the DNA sequences encoding protein domains: S, signal peptide; T, transit peptide; M, mature protein; subscript H, originating from the host; subscript E, originating from the endosymbiont.

planations for this recombination processes? The introns within the diatom sequences are located either shortly upstream from (PsbU), within (PetJ, PsbM, PsbO, FbaC1), or shortly downstream from (AtpC1, Tpt1) the region encoding the putative transit peptide. Assuming that the respective genes of the endosymbiont had the same presequence topology as modern red algae (as shown in Fig. 1B), then the intron positions correspond either to the 5′-untranslated region (UTR) or the transit sequence encoding region of the respective genes before the transfer events. One possibility is that genes donating, e.g., signal peptides contained introns that recombined with acceptor genes lacking introns, a process which might be called "semi-exon shuffling" (Fig. 4). We propose that recombination might have occurred between the introns of the donating genes and the 5′-UTR or the transit peptide-encoding domains of the acceptor genes (Fig. 4A and B, respectively). Random AG nucleotides could have served as functional 3′-intron boundaries (if situated in the same phase as the respective 5′-GT border) to enable correct splicing of the respective transcripts. In cases where the original branching point sequences of the intron of the donating gene may have been lost, a new site had to be generated, which is not unlikely considering the poor conservation of putative branching point sequences we observed.

The recombination of an intron with the acceptor gene either within the 5′-UTR region or within the sequence encoding the transit peptide domain (see Fig. 4A and B) might explain the different locations of the introns we found, but also the variable lengths of the transit peptides in diatoms. Instead of signal peptides, also complete bipartite presequences may have been added in this way—partially replacing transit peptides if existing. This might have been the case for the FbaC1, AtpC, and Tpt1 genes.

It is still an open question why primary endocytobiosis of plastids apparently occurred only once successfully (which means to prevail in evolution), whereas we know of at least two and maybe more cases of successful secondary endocytobioses. Part of the reason may be that during secondary endocytobiosis essential genes had to be redirected from one eukaryotic to another eukaryotic compartment within the same cell. It is possible that the reutilization of regulatory elements, e.g., promoters and polyadenylation sites, may have sped up the process of secondary gene transfer. The proposed modified way of exon shuffling with an intron-free accepting gene might also have been helpful for enhancing the establishment of secondary plastids.

Acknowledgments. The authors are grateful to Andreas Renz, BASF Plant Bioscience, for constructing the *Phaeodactylum tricornutum* cDNA library and sequencing. This study was supported by the University of Konstanz and grants from the Deutsche Forschungsgemeinschaft (DFG; Project Kr 1661-2) and the European community (MARGENES; Project QLRT-2001-01226) to P.G.K.

References

Adams KL, Daley DO, Qiu YL, Whelan J, Palmer JD (2000) Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. Nature 408:354–357

Adams KL, Daley DO, Whelan J, Palmer JD (2002) Genes for two mitochondrial ribosomal proteins in flowering plants are derived from their chloroplast or cytosolic counterparts. Plant Cell 14:931–943

Apt KE, Zaslavkaia L, Lippmeier JC, Lang M, Kilian O, Wetherbee R, Grossman AR, Kroth PG (2002) *In vivo* characterization of diatom multipartite plastid targeting signals. J Cell Sci 115:4061–4069

Bhaya D, Grossman AR (1991) Targeting proteins to diatom plastids involves transport through an endoplasmic reticulum. Mol Gen Genet 229:400–404

Bruce BD (2000) Chloroplast transit peptides: Structure, function and evolution. Trends Cell Biol 10:440–447

Caron L, Douady D, Quinet-Szely M, deGoër S, Berkaloff C (1996) Gene structure of a chlorophyll a/c-binding protein from a brown alga: Presence of an intron and phylogenetic implications. J Mol Evol 43:270–280

Cavalier-Smith T (1999) Principles of protein and lipid targeting in secondary symbiogenesis: Euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. J Eukaryot Microbiol 46:347–366

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

Delwiche CF, Palmer JD (1997) The origin of plastids and their spread via secondary symbiosis. Plant Syst Evol 11:53–86

Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu XN, Reith M, Cavalier-Smith T, Maier UG (2001) The highly reduced genome of an enslaved algal nucleus. Nature 410:1091–1096

Emanuelsson O, Nielsen H, von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. Protein Sci 8:978–984

Gantt JS, Baldauf SL, Calie PJ, Weeden NF, Palmer JD (1991) Transfer of rpl22 to the nucleus greatly preceded its loss from the chloroplast and involved the gain of an intron. EMBO J 10:3073–3078

Gibbs SP (1981) The chloroplast endoplasmic reticulum: Structure, function, and evolutionary significance. Int Rev Cytol 72:49–99

Gilson PR, McFadden GI (2002) Jam packed genomes—A preliminary, comparative analysis of nucleomorphs. Genetica 115:13–28

Jarvis P, Soil J (2001) Toc, Tic, and chloroplast protein import. Biochim Biophys Acta 1541:64–79

Kadowaki K, Kubo N, Ozawa K, Hirai A (1996) Targeting presequence acquisition after mitochondrial gene transfer to the nucleus occurs by duplication of existing targeting signals. EMBO J 15:6652–6661

Keegstra K, Cline K (1999) Protein import and routing systems of chloroplasts. Plant Cell 11:557–570

Kroth PG (2002) Protein transport into secondary plastids and the evolution of primary and secondary plastids. Int Rev Cytol 221:191–255

Kubo N, Harada K, Hirai A, Kadowaki K (1999) A single nuclear transcript encoding mitochondrial RPS14 and SDHB of rice is processed by alternative splicing: Common use of the same mitochondrial targeting signal for different proteins. Proc Natl Acad Sci USA 96:9207–9211

Kubo N, Jordana X, Ozawa K, Zanlungo S, Harada K, Sasaki T, Kadowaki K (2000) Transfer of the mitochondrial rps10 gene to the nucleus in rice: Acquisition of the 5′ untranslated region followed by gene duplication. Mol Gen Genet 263:733–739

Lang M, Apt KE, Kroth PG (1998) Protein transport into "complex" diatom plastids utilizes two different targeting signals. J Biol Chem 273:30973–30978

Liaud M-F, Lichtle C, Apt K, Martin W, Cerff R (2000) Compartment-specific isoforms of TPI and GAPDH are imported into diatom mitochondria as a fusion protein: Evidence in favor of a mitochondrial origin of the eukaryotic glycolytic pathway. Mol Biol Evol 17:213–223

Long M, de Souza SJ, Rosenberg C, Gilbert W (1996) Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c_1 precursor. Proc Natl Acad Sci USA 93:7727–7731

Martin W, Müller M (1998) The hydrogen hypothesis for the first eukaryote. Nature 392:37–41

Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa S, Kowallik KV (1998) Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393:162–165

Martin W, Rujan T, Richly E, Hansen A, Ansorge W, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci USA 99:12246–12251

McFadden GI (2001) Primary and secondary endosymbiosis and the origin of plastids. J Phycol 37:951–959

Moreira D, Le Guyader H, Philippe H (2000) The origin of red algae and the evolution of chloroplasts. Nature 405:69–72

Nielsen H, Brunak S, von Heijne G (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng 12:3–9

Norris BJ, Miller DJ (1994) Nucleotide sequence of a cDNA clone encoding the precursor of the peridinin-chlorophyll a-binding protein from the dinoflagellate *Symbiodinium* sp. Plant Mol Biol 24:673–677

Palmer JD (2003) The symbiotic birth and spread of plastids: How many times and whodunit? J Phycol 39:1–9

Patthy L (1999) Genome evolution and the evolution of exon-shuffling—A review. Gene 238:103–114

Plaumann M, Pelzer-Reith B, Martin WF, Schnarrenberger C (1997) Multiple recruitment of class-I aldolase to chloroplasts and eubacterial origin of eubacterial eukaryotic class-II aldolases revealed by cDNAs from *Euglena gracilis*. Curr Genet 31:430–438

Reith ME, Munholland J (1995) Complete nucleotide sequence of the *Porphyra purpurea* chloroplast genome. Plant Mol Biol Rep 13:333–335

Scala S, Carels N, Falciatore A, Chiusano ML, Bowler C (2002) Genome properties of the diatom *Phaeodactylum tricornutum*. Plant Physiol 129:993–1002

Schaap D, van Poppel NF, Vermeulen AN (2001) Intron invasion in protozoal nuclear encoded plastid genes. Mol Biochem Parasitol 115:119–211

Sulli C, Fang ZW, Muchhal US, Schwartzbach SD (1999) Topology of *Euglena* chloroplast protein precursors within endoplasmic reticulum to golgi to chloroplast transport vesicles. J Biol Chem 274:457–463

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Tolstrup N, Rouze P, Brunak S (1997) A branch point consensus from Arabidopsis found by non-circular analysis allows for better prediction of acceptor sites. Nucleic Acids Res 25:3159–3163

Vacula R, Steiner, Krajcovic J, Ebringer L, Löffelhardt W (1999) Nucleus-encoded precursors to thylakoid lumen proteins of *Euglena gracilis* possess tripartite presequences. DNA Res 6:45–49

Van der Vere PS, Bennett TM, Oblong JE, Lamppa GK (1995) A chloroplast processing enzyme involved in precursor maturation shares a zinc-binding motif with a recently recognized family of metalloendopeptidases. Proc Natl Acad Sci USA 92:7177–7181

van Dooren G, Schwartzbach SD, Osafune T, McFadden GI (2001) Translocation of proteins across the multiple membranes of complex plastids. Biochim Biophys Acta 1541:34–53

Waller RF, Keeling PJ, Donald RGK, Striepen B, Handman E, Lang-Unnasch N, Cowman AF, Besra GS, Roos DS, McFadden GI (1998) Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. Proc Natl Acad Sci USA 95:12352–12357

Waller RF, Reed MB, Cowman AF, McFadden GI (2000) Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. EMBO J 19:1794–1802

Wischmann C, Schuster W (1995) Transfer of rps10 from the mitochondrion to the nucleus in *Arabidopsis thaliana:* Evidence for RNA-mediated transfer and exon shuffling at the integration site. FEBS Lett 374:152–156

Zaslavskaia L, Lippmeier JC, Kroth PG, Grossman AR, Apt KE (2000) Transformation of the diatom *Phaeodactylum tricornutum* (Bacillariophyceae) with a variety of selectable marker and reporter genes. J Phycol 36:379–386