MBE Advance Access published April 17, 2012

Intragenomic Spread of Plastid-Targeting Presequences in the Coccolithophore *Emiliania huxleyi*

Fabien Burki,†[,1] Yoshihisa Hirakawa,†[,1] and Patrick J. Keeling*[,1]

[1]Canadian Institute for Advanced Research, Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

†These authors contributed equally to this work.

***Corresponding author:** E-mail: pkeeling@mail.ubc.ca.

Associate editor: Charles Delwiche

Abstract

Nucleus-encoded plastid-targeted proteins of photosynthetic organisms are generally equipped with an N-terminal presequence required for crossing the plastid membranes. The acquisition of these presequences played a fundamental role in the establishment of plastids. Here, we report a unique case of two non-homologous proteins possessing completely identical presequences consisting of a bipartite plastid-targeting signal in the coccolithophore *Emiliania huxleyi*. We further show that this presequence is highly conserved in five additional proteins that did not originally function in plastids, representing de novo plastid acquisitions. These are among the most recent cases of presequence spreading from gene to gene and shed light on important evolutionary processes that have been usually erased by the ancient history of plastid evolution. We propose a mechanism of acquisition involving genomic duplications and gene replacement through non-homologous recombination that may have played a more general role for equipping proteins with targeting information.

Letter

Plastids, for example, chloroplasts, are the photosynthetic organelles of plants and algae, which are products of endosymbioses where once free-living organisms became fully integrated to the host eukaryotic cells (Archibald 2009). This integration resulted in thousands of endosymbiotic genes that transferred to the host nucleus, the bulk of which encode proteins that are targeted back across multiple membranes to service a fully functional plastid (Bock and Timmis 2008). The crossing of these plastid membranes requires that the nucleus-encoded plastid-targeted proteins are equipped with an N-terminal sequence preceding the mature protein and recognized by a dedicated import apparatus (Bolte et al. 2009). These presequences are composed of a transit peptide (TP) or, in the case of secondarily-derived plastids, more complex bipartite presequences consisting of an endoplasmic reticulum (ER)-targeting signal peptide (SP) followed by a TP (Patron and Waller 2007). It is thought that the acquisitions of these presequences played a crucial role in the course of integration that transformed endosymbionts to plastids and therefore constitutes an essential factor in the spread of photosynthesis across the eukaryotic tree of life.

How proteins acquire targeting presequences following endosymbiotic gene transfers (EGTs) remains poorly understood. Mitochondria represent a similar case since they derive from an alpha-proteobacteria that also lost or transferred most genes to the host nucleus following endosymbiosis (Gray 1999). These nucleus-encoded mitochondrion-targeted proteins have been extensively studied in flowering plants, and numerous examples exist of random integrations into preexisting genes already preceded by the targeting information, exon shuffling, de novo formation of presequences, or regions within the mature protein that act as transport signals (Adams et al. 2000; Adams et al. 2002; Choi et al. 2006; Liu et al. 2009b). The situation in algae with secondary plastids is more obscure and putative semi-exon or exon shuffling has been suggested only in a few species (Waller et al. 1998; van Dooren et al. 2002; Kilian and Kroth 2004; Vesteg et al. 2010). Generally, the low level of conservation among plastid-targeting presequences, with TP in particular lacking sequence similarity even among closely related species, has prevented the study of their origins. Only two examples of bipartite presequences sharing detectable sequence identity have been documented in euglenids, but the mechanisms of acquisition could not be deduced (Durnford and Gray 2006).

Fructose bisphosphate aldolases (FBAs) are nucleus-encoded metabolic enzymes functioning in the cytosol and plastids where they are involved in glycolysis and carbon fixation, respectively (Flechner et al. 1999). They exist in two non-homologous but functionally equivalent forms referred to as FBA class I and class II (henceforth FBA I and FBA II), with no sequence similarity and different catalytic mechanisms (Marsh and Lebherz 1992) as well as a complex evolutionary history (Patron et al. 2004). Algae harboring secondary plastids, such as diatoms or coccolithophorids, generally possess two copies of both classes, with cytosolic and plastid-targeted copies of each (Allen et al. 2011). In the coccolithophore *Emiliania huxleyi*, one of the most abundant photosynthetic unicellular eukaryotes in the oceans (Liu et al. 2009a), the distribution of FBA proteins does not fit with the expected pattern. The genome of *E. huxleyi* encodes four FBA I, three of which were predicted

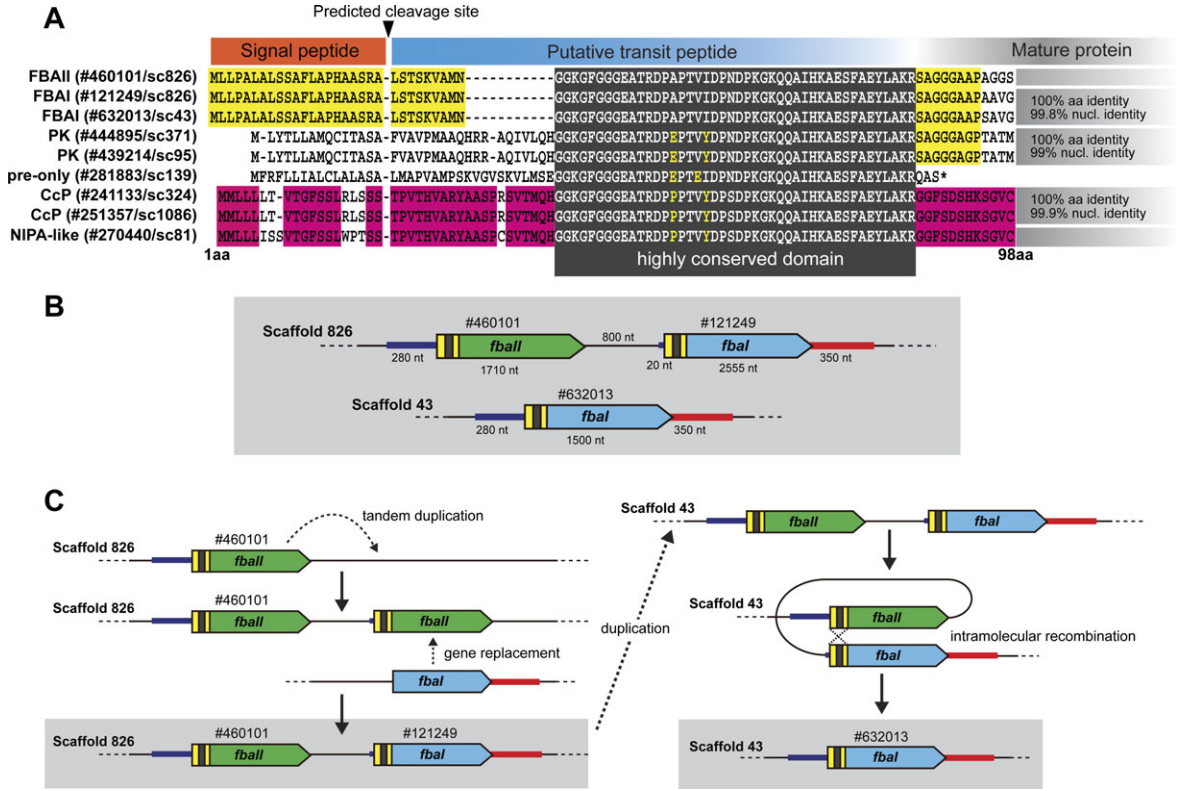

FIG. 1. (*A*) N-terminal sequences consisting of a bipartite plastid targeting signal (signal peptide and transit peptide) and a highly conserved domain among six non-homologous proteins and three recent paralogs (FBA I, PK, and CcP). The arrowhead indicates the predicted signal peptide cleavage sites, while cleavage sites between transit peptides and mature proteins are unclear. Identical residues are shown in colored boxes (yellow, pink, and black). Sequence identity between paralogs for the mature proteins is indicated (*B*) Genomic configuration of the *fba* genes on scaffolds 826 and 43. The yellow rectangles correspond to the presequence region and include the highly conserved domain (black); the blue and red bars represent conserved sequences directly upstream (280 or 20 nt) and downstream (350 nt) of the *fba* genes, respectively. (*C*) Possible process of presequence acquisition in two *fbaI* genes. A *fbaII* gene first tandemly duplicated and a *fbaI* gene flanked by 350 bp of downstream DNA replaced one of the mature protein coding region. The fragment encoding *fbaII* and *fbaI* on scaffold 826 probably duplicated to scaffold 43, followed by an intramolecular recombination event that led to the lost of the *fbaII* copy and the exchange of the presequence and adjacent DNA.

to be targeted to the plastid based on the presence of an N-terminal bipartite presequence (fig. 1*A*; supplementary fig. S1 and table S1, Supplementary Material online). This presequence consisted of a canonical SP followed by a hydrophilic region predicted to be a TP (see supplementary material, Supplementary Material online). Unexpectedly, however, phylogenetic analysis indicated that the single cytosolic copy (#308857) was strongly placed within an otherwise clearly plastid-targeted clade, whereas two plastid paralogs (#121249 and #632013) belonged to a clade inferred to be cytosolic (supplementary fig. S1, Supplementary Material online).

Most interestingly, the #121249 plastidic FBA I is encoded immediately downstream of a plastid-targeted FBA II (# 460101 on scaffold 826), and the presequences of both proteins share 100% identity (fig. 1*A*). Another plastid-targeted FBA I (#632013 on scaffold 43) was found to be nearly identical to the #121249 copy (99.8% nucleotide similarity in the coding region). Strikingly, the upstream untranslated region of the *fbaI* (#632013) gene shared 97% identity (over 280 bp) to that of *fbaII* (#460101), whereas the downstream untranslated region (350 bp) was conserved between the two *fbaI* genes (#632013 and #121249) (fig. 1*B*). One possible explanation for these observations is that an FBA I copy (#121249) functioning in the cytosol of *E. huxleyi* acquired a plastid-targeting presequence by replacing a tandemly duplicated *fbaII* gene (fig. 1*C*). This *fbaII* (#460101) – *fbaI* (#121249) arrangement on scaffold 826 may have further duplicated to scaffold 43, followed by an intramolecular recombination between their presequences that led to the loss of the *fbaII* gene on scaffold 43 (fig. 1*C*).

This model does not explain why recombination between two non-homologous *fba* genes would take place, and indeed the situation is more complex, because BLASTP searches against the *E. huxleyi* genome revealed homologous leader sequences preceding five additional unrelated proteins (fig. 1*A*). To confirm that these proteins (and the FBAs) were not assembly artifacts, we generated new cDNA sequences allowing us to refine the JGI gene models (the sequences were deposited in GenBank: JQ088180, 088184, 088186, and 088187). Two proteins corresponded to pyruvate kinases (PKs) and two were cytochrome c peroxidases (CcPs); we were unable to obtain cDNA for the last hit,

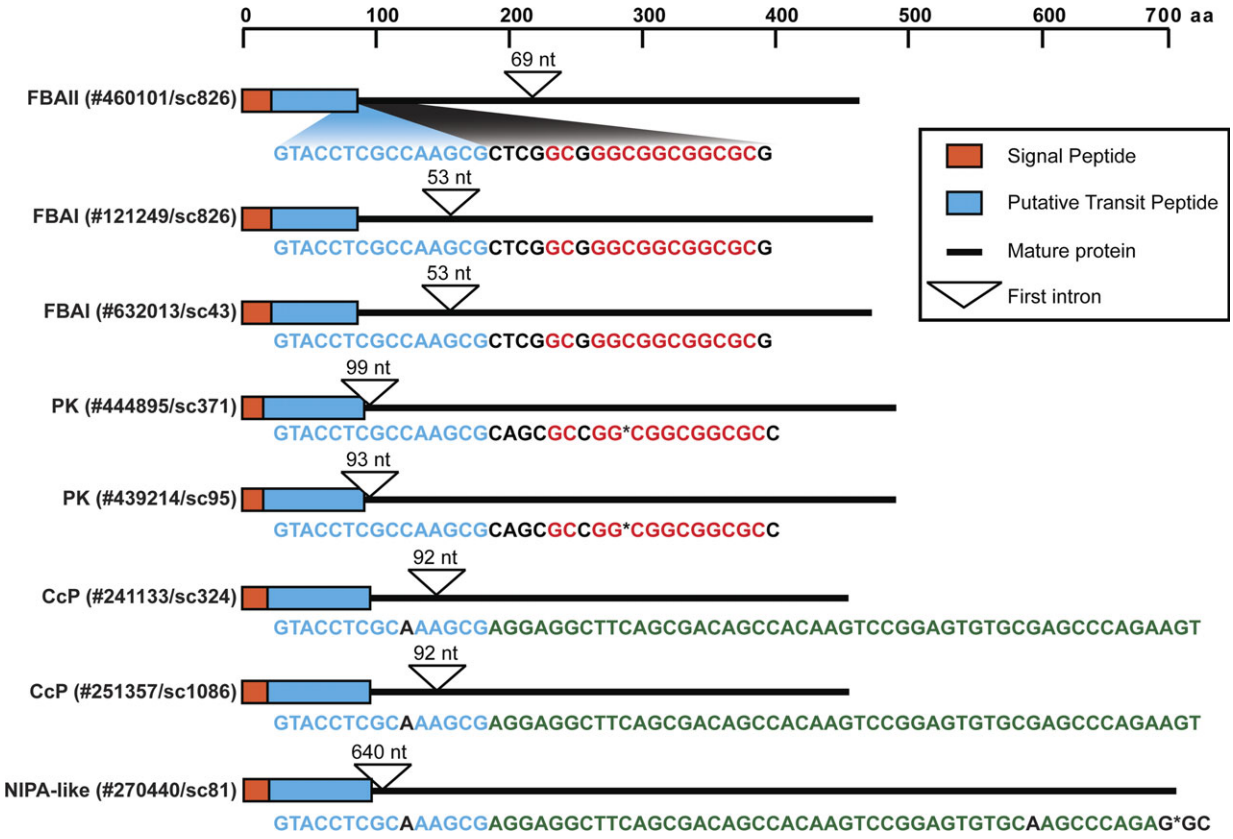

Fig. 2. Position and size of the first intron indicated on five non-homologous proteins and three paralogs (FBA I, PK, and CcP). The nucleotide sequence corresponding to the region directly downstream of the highly conserved domain of the predicted transit peptide is shown. Nucleotides in blue show the highly conserved region among all these genes, and red nucleotides represent the conserved region among the *fba* genes and *pk* genes. Nucleotides in green represent the conserved region between the *ccp* genes and *nipa*-like. *Indicates the first splice site in *pk* and *nipa*-like. The scale bar indicates the length of the proteins in amino acid.

which was a hypothetical protein with similarity to NIPA-like proteins. In addition, we identified one cDNA that encoded only this presequence (pre-only in fig. 1*A*), which was followed by a stop codon, 291 nucleotides of untranslated region, and a poly-A tail. Surprisingly, none of these additional proteins are classically targeted to the plastid or were inferred to belong to a plastid-targeted clade (phylogenies not shown). An intron was present close to the junction between the presequence and inferred mature protein (ranging from 8 nucleotides in PK to 182 nucleotides in CcP) in all cases except the gene encoding only the presequence (fig. 2). Moreover, six amino acids directly downstream of the highly conserved putative TP region were identical between FBA I and II and the two copies of PK (fig. 1*A*), corresponding to a GC-rich stretch that forms the splice sites in *pk* (AGCGCCGG*CGGCGGCGCC). Similarly, nearly complete identity (51/53 nucleotides) was observed between *nipa*-like and *ccp* directly downstream of the conserved presequence, and a splice site is present in *nipa*-like at the end of this sequence (fig. 2). Even though the mechanism of presequence acquisition in these additional cases cannot be completely assessed, the presence of introns suggests that exon or semi-exon shuffling might be involved. To determine if this process is genome wide or instead associated with this presequence specifically, we identified 105 additional putative plastid-targeted proteins and searched for homologs of their presequences in other regions of the genome. In no other case could homologous presequences be identified in non-homologous proteins, suggesting that this particular presequence is the only one that has engaged in such a spread, at least so recently that it can be detected.

These remarkable cases of presequence conservation and acquisition in several unrelated proteins have a number of implications for plastid evolution in *E. huxleyi*. Most obviously, the spread of this presequence has introduced a number of new proteins to the plastid, which makes these events quite different from previously described examples of presequence acquisition where proteins are back-targeting to their original compartment following EGTs (Adams et al. 2000; Liu et al. 2009b). Here, presequence acquisitions represent de novo acquisitions of new functions by the plastid. We postulate that the presequence originated with FBA II (#460101), since this is the only protein that could be traced back to an endosymbiotic origin (supplementary fig. S2 and table S2, Supplementary Material online). In contrast, the function of NIPA-like proteins is unknown and CcP is usually imported into mitochondria where it catalyzes the reduction of hydrogen peroxide (Daum et al. 1982). Two canonical mitochondrial

CcP copies are also present in *E. huxleyi* (JGI accessions 68109 and 438123) in addition to the two plastid-targeted copies, but the function of the plastid copies is unclear. PK is involved in the glycolytic pathway in the cytosol (Mertens 1993; Liapounova et al. 2006), but in *E. huxleyi*, PK has previously been noted not to localize to cytoplasm, but rather to the plastid (Tsuji et al. 2009). Interestingly, three additional enzymes functioning in the same pathway, namely phosphoglycerate mutase, enolase, and pyruvate carboxylase, were also predicted to have a bipartite plastid targeting signal in *E. huxleyi* (Tsuji et al. 2009).

The degree of sequence conservation between the presequences described here has no precedent, and the lack of comparable cases in other eukaryotes, or even within *E. huxleyi*, might suggest that this particular sequence possesses properties unlike other presequences, for example, mobility (a possibility also hinted at by the presence of an expressed mRNA encoding only the presequence). Alternatively, the presequence might be under unusual selective pressure, in which case it might have moved by transposition and illegitimate recombination, but in either case, it gives a unique view into how protein targeting can spread and how plastids can acquire novel functions rather rapidly.

Supplementary Material

Supplementary material, tables S1–S3, and figures S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (227301) and by a grant from the Tula Foundation to the Centre for Microbial Diversity and Evolution. The work conducted by the US Department of Energy Joint Genome Institute is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. In particular, we thank J. M. Archibald, G. I. McFadden, M. W. Gray, and C. Lane for project contributions to the sequencing of the nuclear genome of *G. theta* and *B. natans*, and Betsy Read for *E. huxleyi*. We thank the French ANR program POSEIDON (Colomban de Vargas) for giving access to unpublished genetic data (*Coccolithus braarudii*). P.J.K. is a Fellow of the Canadian Institute for Advanced Research.

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.

Allen AE, Moustafa A, Montsant A, Eckert A, Kroth PG, Bowler C. 2011. Evolution and functional diversification of fructose bisphosphate aldolase genes in photosynthetic marine diatoms. *Mol Biol Evol.* 29:367–379.

Archibald JM. 2009. The puzzle of plastid evolution. *Curr Biol.* 19:R81–R88.

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

Bolte K, Bullmann L, Hempel F, Bozarth A, Zauner S, Maier UG. 2009. Protein targeting into secondary plastids. *J Eukaryot Microbiol.* 56:9–15.

Choi C, Liu Z, Adams KL. 2006. Evolutionary transfers of mitochondrial genes to the nucleus in the Populus lineage and coexpression of nuclear and mitochondrial Sdh4 genes. *New Phytol.* 172:429–439.

Daum G, Bohni PC, Schatz G. 1982. Import of proteins into mitochondria. Cytochrome b2 and cytochrome c peroxidase are located in the intermembrane space of yeast mitochondria. *J Biol Chem.* 257:13028–13033.

Durnford DG, Gray MW. 2006. Analysis of *Euglena gracilis* plastid-targeted proteins reveals different classes of transit sequences. *Eukaryot Cell* 5:2079–2091.

Flechner A, Gross W, Martin WF, Schnarrenberger C. 1999. Chloroplast class I and class II aldolases are bifunctional for fructose-1,6-biphosphate and sedoheptulose-1,7-biphosphate cleavage in the Calvin cycle. *FEBS Lett.* 447:200–202.

Gray MW. 1999. Evolution of organellar genomes. *Curr Opin Genet Dev.* 9:678–687.

Kilian O, Kroth PG. 2004. Presequence acquisition during secondary endocytobiosis and the possible role of introns. *J Mol Evol.* 58:712–721.

Liapounova NA, Hampl V, Gordon PM, Sensen CW, Gedamu L, Dacks JB. 2006. Reconstructing the mosaic glycolytic pathway of the anaerobic eukaryote Monocercomonoides. *Eukaryot Cell* 5:2138–2146.

Liu H, Probert I, Uitz J, Claustre H, Aris-Brosou S, Frada M, Not F, de Vargas C. 2009a. Extreme diversity in noncalcifying haptophytes explains a major pigment paradox in open oceans. *Proc Natl Acad Sci U S A.* 106:12803–12808.

Liu SL, Zhuang Y, Zhang P, Adams KL. 2009b. Comparative analysis of structural diversity and sequence evolution in plant mitochondrial genes transferred to the nucleus. *Mol Biol Evol.* 26:875–891.

Marsh JJ, Lebherz HG. 1992. Fructose-bisphosphate aldolases: an evolutionary history. *Trends Biochem Sci.* 17:110–113.

Mertens E. 1993. ATP versus pyrophosphate: glycolysis revisited in parasitic protists. *Parasitol Today.* 9:122–126.

Patron NJ, Rogers MB, Keeling PJ. 2004. Gene replacement of fructose-1,6-bisphosphate aldolase supports the hypothesis of a single photosynthetic ancestor of chromalveolates. *Eukaryot Cell* 3:1169–1175.

Patron NJ, Waller R. 2007. Transit peptide diversity and divergence: A global analysis of plastid targeting signals. *Bioessays* 29:1048–1058.

Tsuji Y, Suzuki I, Shiraiwa Y. 2009. Photosynthetic carbon assimilation in the coccolithophorid *Emiliania huxleyi* (Haptophyta): Evidence for the predominant operation of the c3 cycle and the contribution of {beta}-carboxylases to the active anaplerotic reaction. *Plant Cell Physiol.* 50:318–329.

van Dooren GG, Su V, D'Ombrain MC, McFadden GI. 2002. Processing of an apicoplast leader sequence in *Plasmodium falciparum* and the identification of a putative leader cleavage enzyme. *J Biol Chem.* 277:23612–23619.

Vesteg M, Vacula R, Steiner JM, Mateasikova B, Loffelhardt W, Brejova B, Krajcovic J. 2010. A possible role for short introns in the acquisition of stroma-targeting peptides in the flagellate *Euglena gracilis*. *DNA Res.* 17:223–231.

Waller RF, Keeling PJ, Donald RG, 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 U S A.* 95:12352–12357.