The Carboxy Terminus of YCF1 Contains a Motif Conserved throughout >500 Myr of Streptophyte Evolution

Jan de Vries[1,*], John M. Archibald[1,2], and Sven B. Gould[3]

[1]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada

[2]Canadian Institute for Advanced Research, Toronto, Ontario, Canada

[3]Molecular Evolution, Heinrich-Heine-University Düsseldorf, Germany

*Corresponding author: E-mail: jan.devries@dal.ca.

Accepted: January 30, 2017

Abstract

Plastids evolved from cyanobacteria by endosymbiosis. During the course of evolution, the coding capacity of plastid genomes shrinks due to gene loss or transfer to the nucleus. In the green lineage, however, there were apparent gene gains including that of *ycf1*. Although its function is still debated, YCF1 has proven to be a useful marker for plastid evolution. YCF1 sequence and predicted structural features unite the plastid genomes of land plants with those of their closest algal relatives, the higher streptophyte algae; YCF1 appears to have undergone pronounced changes during the course of streptophyte algal evolution. Using new data, we show that YCF1 underwent divergent evolution in the common ancestor of higher streptophyte algae and Klebsormidiophycae. This divergence resulted in the origin of an extreme, klebsormidiophycean-specific YCF1 and the higher streptophyte Ste-YCF1. Most importantly, our analysis uncovers a conserved carboxy-terminal sequence stretch within YCF1 that is unique to higher streptophytes and hints at an important, yet unexplored function.

Key words: plastid evolution, charophytes, YCF1, streptophyte evolution, plastid genomes.

Introduction

Plastid genomes are demonstrably homologous to those of cyanobacteria, but have experienced drastic reductive evolution (Martin et al. 1998; Green 2011; Archibald 2015). Of the more than 4000 genes found in sections IV and V cyanobacteria, which have been suggested to be the closest living relatives to the ancestor of the plastid (Dagan et al. 2013; for an alternative view, see Ponce-Toledo et al. 2017), plastid genomes have retained only about 100–200 due to gene loss and endosymbiotic gene transfer (Timmis et al. 2004). A particularly appealing hypothesis for why some genes have been retained is that they are important for the *in situ* regulation of the photosynthesis redox reactions (Allen 2015), which also applies to the retention of genes by mitochondria for the regulation of their key bioenergetic reactions as part of the electron transport chain. Although the genomes of primary green plastids (Chloroplastida, *cf.* Adl et al. 2012) tend to have fewer genes than those of red algae and glaucophytes (Allen et al. 2011), there are some exceptions. One is *Ycf1*.

A few things are unusual about *Ycf1*. It has no detectable homolog in cyanobacteria, red algae or glaucophytes (de Vries et al. 2015). The origin of *Ycf1* traces back to either a gain of the gene by the plastid genome, which appear to be extremely rare events, or a radical divergence upon the duplication of a plastid gene in the common ancestor of all Chloroplastida (Wicke et al. 2011; de Vries et al. 2015; Nakai 2015a). Regardless of its origin, *ycf1* knockouts were shown to be lethal in tobacco (Drescher et al. 2000) and *Chlamydomonas* (Boudreau et al. 1997), although the reason(s) for this lethality, or the function of the protein, were unknown at the time. Based on data gathered from *Arabidopsis*, it was later proposed that YCF1 is in fact TIC 214, a critical component of the protein translocation machinery at the inner chloroplast envelope (Kikuchi et al. 2013). If true, TIC214 would be the only TIC component encoded by the plastid genome. Moreover, *Ycf1* has been lost in Poaceae and possibly a few other land plants as well (Katayama and Ogihara 1996; de Vries et al. 2015). For these reasons, the

© The Author(s) 2017. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

possibility that YCF1 represents "a general TIC translocon" (Kikuchi et al. 2013) has been challenged (de Vries et al. 2015; Bölter and Soll 2016a, 2016b), with additional criticism revolving around the extreme variation in YCF1 sequence length, which remains to be explained.

In the context of streptophyte terrestrialization, we recently discussed the evolution of the algal Cte-YCF1 to the land plant version of the protein, Ste-YCF1 (de Vries et al. 2016), along with the transformation of FTSH into YCF2 (cf. Civáň et al. 2014). This analysis highlighted a clear distinction between the plastid genomes of the lower (Klebsormidiophyceae + Chlorokybophyceae + Mesostigmatophyceae = KCM) and higher branching streptophyte algae (Zygnematophyceae + Coleochaetophyceae + Charophyceae = ZCC). However, these inferences were hampered by the availability of only a single plastid genome from the lower branching Klebsormidiophyceae. Here we revisit the evolution of *ycf1* and its implications for the origin of Ste-YCF1 in land plants, taking advantage of newly sequenced plastid genomes from streptophyte algae (Lemieux et al. 2016). Specifically, we have discovered a conserved motif, currently of unknown function, in the YCF proteins of some ZCC algae and embryophytes.

Klebsormidium spp. YCF1 Proteins Stand out

YCF1 sequences are diverse in sequence and size. We first used a phylogenetic approach to gain a preliminary glimpse at YCF1 evolution across streptophyte algae. Though a divergent protein, our YCF1 phylogeny (fig. 1*a*) shows remarkable resemblance to the most recent species phylogeny based on the phylogenetic analysis of 88 plastid-encoded proteins (Lemieux et al. 2016; fig. 1*b*). When comparing the YCF1 phylogeny to the species phylogeny (fig. 1*a* and *b*), two YCF1 proteins do not branch as expected within streptophyte algae, both of which are from the genus *Klebsormidium*. This likely reflects the pronounced structural changes characteristic of the plastid genomes of Klebsormidiophyceae (Lemieux et al. 2016). This prompted us to take a closer look at the protein sequences of streptophyte algal YCF1s.

Klebsormidiophyceaen YCF1 Has Undergone Drastic Sequence Expansion

YCF1 homologs exhibit a wide range of sizes (de Vries et al. 2015). There are nevertheless some features these proteins all share: they are predicted to contain six to eight transmembrane TM domains at their N-terminus and to encode charged repeats that cover much of the remaining sequence (cf. de Vries et al. 2015; Nakai 2015b). We surveyed all 20 YCF1 proteins with regard to their length, charge, and number of predicted TM domains.

Streptophyte algal YCF1 proteins range between 410 aa (*Mesostigma viride*) and 4186 aa (*K. flaccidum*) (fig. 1*c*). The largest portion of each YCF1 protein, except for the small YCF1s of the basal branching KCM streptophyte algae *M. viride* and *Chlorokybus atmophyticus*, is made up of the repetitively charged C-terminus. Noteworthy was a long negatively charged region from aa position 1777 to 1883 in the YCF1 of *Mesotaenium endlicherianum*, which is followed by similarly sized positively charged region spanning amino acid (aa) residues 1885–1943. In all cases, the YCF1 N-terminus was predicted to contain 6–7 TMs, usually located within the first 250 aa (fig. 1*c*). The only exceptions are the zygnematalean YCF1s of *M. endlicherianum* and *Cylindrocystis brebissonii*, which were sometimes predicted to have a seventh TM starting around position 1250 and a seventh and eighth TM at around positions 650 and 700, respectively. The longest streptophyte YCF1s in our dataset (and to our knowledge the longest of any streptophyte) are those of the Klebsormidiophyceae. This, together with the results of the phylogenetic analysis (fig. 1*a*), sets the klebsormidiophycean YCF1 clearly apart from both the Cte-YCF1 of the other KCM streptophyte and chlorophyte algae as well as the ZCC/embryophyte Ste-YCF1.

A Novel Motif in Ste-YCF1

Land plants and ZCC streptophyte algae are united by having a Ste-YCF1 type protein (de Vries et al. 2015). This separation was determined through reciprocal HMM detection, which uncovered Cte-YCF1 only when using Cte-YCF1 alignments as a seed and Ste-YCF1s only when using Ste-YCF1 alignments as a seed. Aligning all ZCC Ste-YCF1s with the KCM Cte-YCF1s reveals two regions with high sequence conservation: the first is the hydrophobic N-terminus with its TM domains (fig. 1*c*) and the second is at the very C-terminus (fig. 1*d* and *e*). Closer inspection of the C-terminal region uncovered a conserved stretch of sequence that includes a 12 aa long motif, RLEDLACMNRFW; we suspect that the single conserved cysteine residue is critical for the function of the motif. We henceforth consider these 12 aa as the 'core motif'. This core motif is present in Zygnematophyceae (the closest algal relatives of embryophytes [Wickett et al. 2014]), Coleochaetophyceae and embryophytes. Directly upstream of this core motif there are eight more conserved residues; the full consensus of the Ste-YCF1 motif is thus IKRFLWPTxRLEDLACMNRFW (fig. 1d and e). The only Charophycaeae sequence in this dataset, that of *Chara vulgaris*, bears a divergent, possibly rudimentary, form of the core motif including the cysteine and differing in four aa, RIEDL**V**CM**E**R**V**W (bold = different aa). The fact that *Chara* possesses the most rudimentary form of the core motif is consistent with the basal branching of the Charophyceae within ZCC algae (cf. Wickett et al. 2014; Delwiche and Cooper 2015).

To further assess the degree of conservation of the RLEDLACMNRFW motif among embryophytes, we expanded

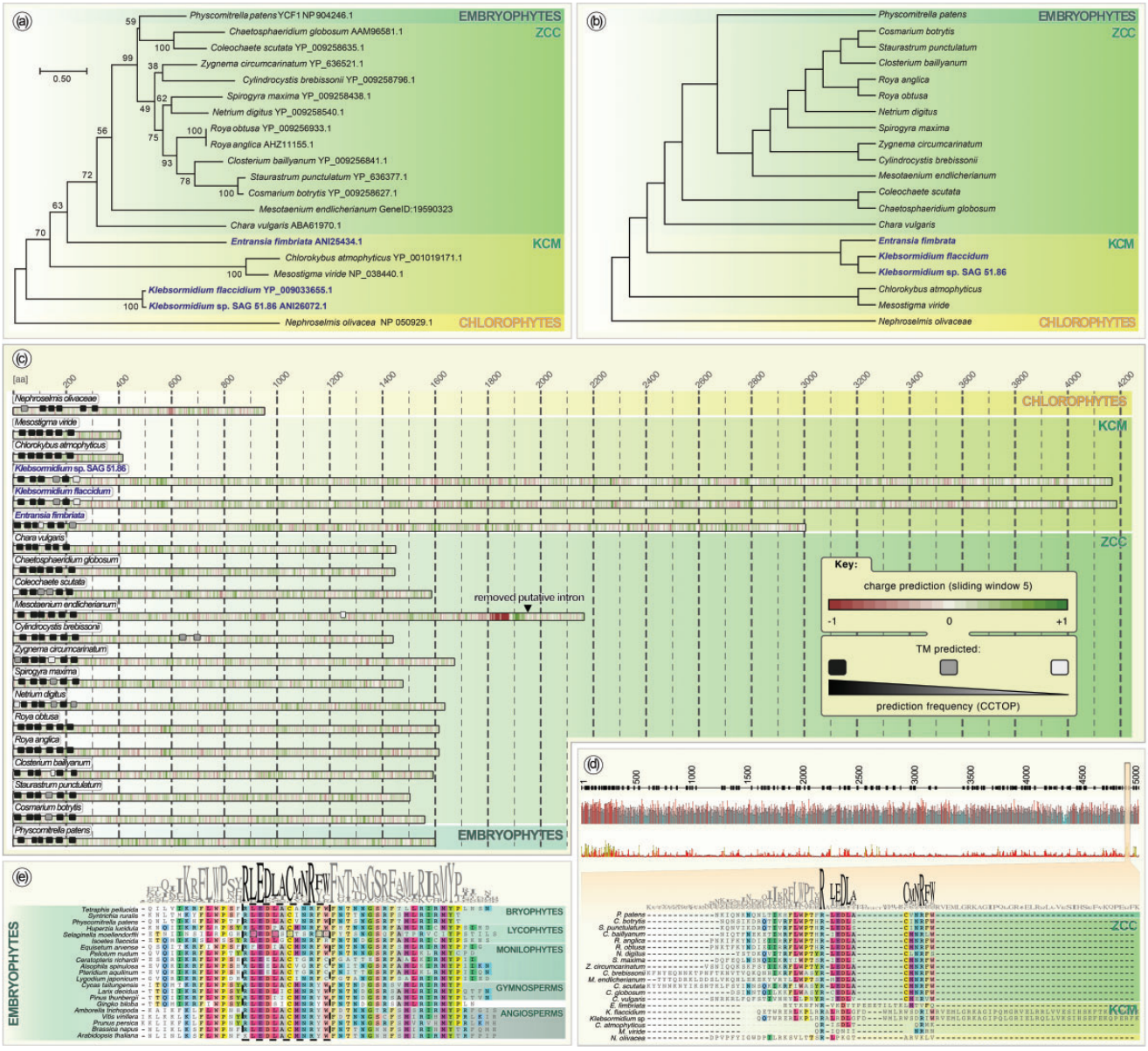

Fig. 1.—Phylogeny of klesbormidiophycean YCF1. *(a)* Maximum likelihood LG + G + I + F phylogeny (500 bootstraps, values shown at each node, partial deletion [95%], gamma category 5) of YCF1 protein sequences from the same species as in *(b)*. NCBI accession numbers of the YCF1 proteins are shown behind the species name. The tree is drawn to scale based on the substitution rate. Note the phylogenetic position of *Klebsormidium* spp. YCF1. (*b*) A reference cladogram of species phylogenetic relationships based on Lemieux et al. (2016). Relationships of the chlorophyte *N. olivaceae*, 5 lower branching KCM and 13 higher branching ZCC streptophyte algae, and the land plant *P. patens* are shown. Klebsormidiophyceae are highlighted in purple. *(c)* YCF1 proteins are drawn to scale based on their length in aa. Color gradient from red (negative) to white (none) to green (positive) indicate the predicted charge based on the EMBOSS explorer (sliding window size: 5 aa). Boxes on the proteins indicate predicted TM domains where, applying a majority rule, some consensus was found based on TMHMM, Phobius and SOSUI. The frequency with which the TMs were predicted by the different programs was evaluated using CCTOP and is depicted as black (predicted by all algorithms) to white (predicted by few algorithms) coloration of the boxes. Note the length of klebsormidiophycean YCF1s (purple font) and the position of the 200 bp intron removed prior to *in silico* translation of *M. endlicherianum*'s YCF1 (arrowhead). *(d)* Depiction of the YCF1 protein sequence alignment of the 20 YCF1s used in (*a*). The topmost line depicts the full 5015 positions of the alignment (of which 311 aligned residues were used for the phylogeny shown in [*a*]), black boxes indicate regions where based on a majority rule some consensus was found for at least half of proteins under consideration. The histograms below depict high (red) to low (blue) hydrophobicity (top) and 30–70% (yellow) to < 30% (red) sequence identity (bottom). A zoom-in into the very C-terminal region of the alignment shows that Ste-YCF1 of higher ZCC streptophyte algae shares the core aa-sequence "RLEDLACMNRFW" with land plants. Note the conserved hydrophobic block of N-terminal aa and the conserved C-terminal end. The removal of the intron in the *ycf1* ORF of *Mesotaenium* restores a YCF1 that carries a RxEDLACMNRFW motif. *(e)* An alignment of 21 land plant YCF1s (3 bryophytes, 3 lycophytes, 6 monilophytes, 3 gymnosperms, *Gingko*, and 5 angiosperms) demonstrates that RLEDLACMNRFW is conserved throughout all embryophyte Ste-YCF1s. Note the putative RNA editing sites (marked with dotted boxes filled in the color corresponding to the aa after RNA editing) in the *S. moellendorffii* motif sequence (based on data from *S. unicata* [cf. Oldenkott et al. 2014]).

the alignment to include sequences from across the diversity of land plants. We found that the core motif RLEDLACMNRFW is conserved from moss to *Arabidopsis thaliana* (fig. 1e). Moreover, land plants have an extended conserved C-terminal region of 39 aa (fig. 1e). The only land plant sequence showing severe alterations from this core motif (including the cysteine residue) was that of *Selaginella moellendorffii*. *S. moellendorffii* is known for its unusual organellar biology characterized by massive RNA editing (Hecht et al. 2011). In its close relative *S. uncinata*, RNA editing of the *Ycf1* open reading frame (ORF) converts the codons from encoding R**P**ED**P**S**R**MNR**PR** into encoding R**L**ED**L**S**C**MNR**FW** (Oldenkott et al. 2014; fig. 1e). Ergo, RLEDLACMNRFW is a motif within YCF1 that *Arabidopsis* shares with the common ancestor of all land plants.

To determine whether this motif has been detected before, we used the 12, 21, and 39 aa motif versions as queries for InterPro scans. The former two did not return any hits. Using the 39 aa motif as a query, InterPro again predicted no family memberships, domains or functions, but detected an unintegrated (i.e. not yet curated) signature match against the unnamed PANTHER protein family PTHR33163 (cf. Mi et al. 2016), that consists of 74 land plant YCF1s and does not list the 12, 21, or 39 aa motifs as a conserved feature. None of the motif versions can be found in PFAM (cf. Finn et al. 2016) and motif searches of the PDB database (cf. Berman et al. 2000) returned no hits. Using the 12, 21, and 39 aa versions of the motif as queries for a BLASTp against the RefSeq database (applying a very low stringency *e* value cutoff of <1), we detected only streptophyte YCF1 sequences (see supplementary table S1, Supplementary Material online). Hence, the aa sequence RLEDLACMNRFW (and extended versions) indeed represents a novel motif restricted to higher streptophyte YCF1s.

To confirm these BLAST-based results, we used all motif versions as queries in pHMMER searches (Finn et al. 2015) against the UniProt database (The UniProt Consortium 2015). The 12 aa motif did not return any significant hits. The 21 aa and 39 aa motifs returned 1788 and 2139 significant hits, respectively. All of these hits, except one, were streptophyte sequences that encompass land plant and 12 ZCC streptophyte algal Ste-YCF1s. The exception was the nuclear-encoded FAR-RED IMPAIRED RESPONSE 1 (FAR1) protein of *Medicago truncatula* (protein id: MTR_2g039370), containing a 39 aa long C-terminal stretch (F**K**L**FLWP**NY**RLEDLACINR**Y**WFNT**H**N**GS**H**F**S**I**L**RI**H**M**YP** [bold = identical residue]) that resembled the 39 aa version of the Ste-YCF1 motif. FAR1 proteins are transcription factors (Hudson et al. 1999; Lin et al. 2007) with the ability to bind DNA (Lin et al. 2007). This warrants further investigation, as nucleic acid binding capacities have been described for *Chlamydomonas* YCF1 (Boudreau et al. 1997). An involvement of the novel Ste-YCF1 motif in DNA/RNA-binding or even transcriptional regulation is hence conceivable (see below).

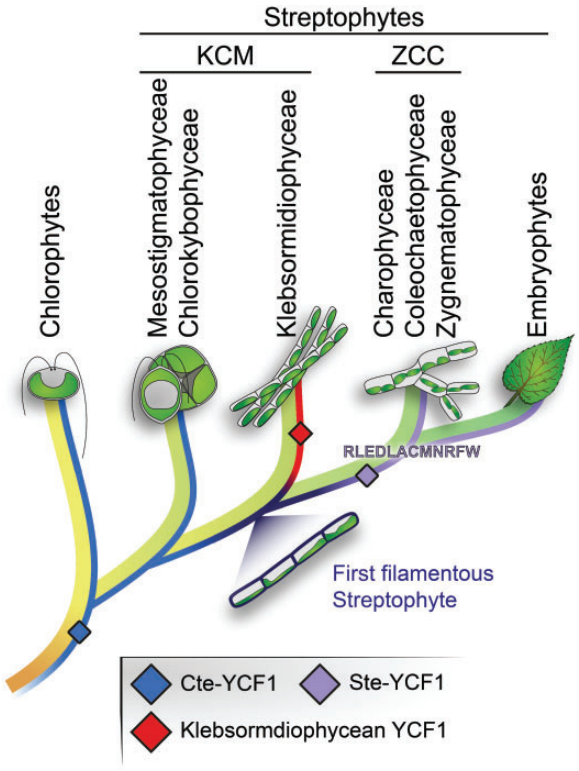

Fig. 2.—YCF1 sequence divergence during streptophyte algal evolution. A cladogram of the evolution of Chloroplastida (thick lines; chlorophytes = yellow, streptophytes = different shades of green) with the trajectory of YCF1 evolution projected on top. The ancestral Cte-YCF1 (blue) is shared by chlorophytes and the KCM streptophyte algae of the Mesostigmatophyceae and Chlorokybophceae. Along the trajectory of streptophyte evolution that includes the first filamentous streptophyte (one of the common ancestors of recent Klebsormidiophyceae and ZCC streptophyte algae + land plants) the Cte-YCF1 sequence started to change (dark purple). From there, YCF1 diverged into the klebsormidiophycean YCF1 (red) and the Ste-YCF1 (purple). The latter is found among ZCC streptophyte algae and land plants.

YCF1 Sequence Divergence and the First Filamentous Streptophyte Algae

What does this mean for the trajectory of YCF1 evolution and ultimately its function? Cte-YCF1 most likely evolved in the common ancestor of all Chloroplastida (Wicke et al. 2011; de Vries et al. 2015). After the split of streptophytes and chlorophytes (cf. Lewis and McCourt 2004; Becker and Marin 2009), the basal branching streptophyte algae Mesostigmatophyceae and Chlorokybophyceae retained the ancestral Cte-YCF1, sharing it hence with the chlorophytes (fig. 2). In the common ancestor of Klebsormidiophyceae and higher branching ZCC streptophyte algae, the Cte-YCF1 evolved into the highly expanded version found in Klebsormidiophyceae (>3000 aa length) and the Ste-YCF1 (bearing the RLEDLACMNRFW motif). Ste-YCF1 was vertically

inherited by land plants. Therefore, the changes in YCF1 sequence emerged along the trajectory that includes the first filamentous streptophyte (fig. 2).

We previously hypothesized that changes in YCF1 occurred concomitantly with changes in FtsH/YCF2 (cf. de Vries et al. 2016). Although FtsH is a plastid-encoded component of the photosystem II maintenance machinery (de Vries et al. 2013; cf. Janska et al. 2013), YCF2's function remains unknown (Drescher et al. 2000). So, what about klebsormidiophycean FtsH/YCF2? *Entransia* is the only member of the Klebsormidiophyceae with a sequenced plastid genome that encodes the *ftsH* gene (cf. Lemieux et al. 2016). *Klebsormidium* spp. has probably lost *ycf2/ftsH*. Civáň and colleagues used a BLASTp approach to distinguish between higher streptophyte YCF2 and FtsH (among Chloroplastida, the latter can be found in *Mesostigma, Chlorokybus* and chlorophytes). Querying *Entransia* "FtsH" via BLASTp against the non-redundant protein database of NCBI (NCBI Resource Coordinators 2016) returned the YCF2 sequence of the conifer *Pseudolarix amabilis* as the best hit (YP_009268451.1; *e* value of 4×10^{-18}, but with a query coverage of only 22%), followed by many other land plant YCF2s with equally low query coverage. In accordance with the approach taken by Civan and colleagues, *Entransia* FtsH can hence be classified as YCF2. It is noteworthy that the closest streptophyte algal hit, when using the *Entransia* FtsH/YCF2 as query was found to be *Closterium*'s YCF2 (YP_009256827.1, *e* value of 2×10^{-14}, 27% query coverage and the 61st hit). We interpret this as klebsormidiophycean YCF2 being distinct from its orthologs in other streptophyte algae and land plants; this is reminiscent of the pattern we observed for klebsormidiophycean YCF1. Therefore, our new data further supports the notion that YCF1 and YCF2 sequence divergence occurred contemporaneously. The underlying reason(s) for this parallel divergence is as elusive as the functions of the proteins themselves.

In 1997, Boudreau and colleagues found that *Chlamydomonas* Cte-YCF1 was able to bind nucleic acids and suggested that it could bind plastid DNA. If so, YCF1 might be important, for example, during plastid division and in the context of a filamentous and, in the case of many ZCC algae and land plants, polyplastidic species. Alternatively, YCF1 might bind specific RNAs instead of DNA, which would point to a regulatory function at a certain stage of the plastid, perhaps regulating a stage transition associated with protein import. This regulatory role might be linked to the recent finding that exogenous treatment of the ZCC alga *Spirogyra pratensis* with the phytohormone ethylene resulted in a downregulation of the gene expression of Ste-YCF1, concomitant with elevated stress response signaling as well as decreased gene expression of photosynthesis-associated and chlorophyll biosynthesis-associated proteins (Van de Poel et al. 2016). In summary, the massive divergence in size, the conserved motif in Ste-YCF1, and the loss of the protein in grasses speaks in favor of a function that reaches beyond that of just a TIC component.

Conclusion

YCF1 remains an enigmatic protein. However, the broader the sampling of sequence data across the diversity of streptophytes becomes, the less puzzling the sequence diversity of YCF1 appears. In this study, we have shown that 1) klebsormidiophycean YCF1 is unique regarding its size expansion and 2) that Ste-YCF1s are united by a novel motif that is RLEDLACMNRFW. The function(s) of this motif remains to be explored, but it is likely significant considering its strict conservation across more than 500 Myr of evolution, which separates *Coleochaete* and *Arabidopsis* (cf. Parfrey et al. 2011).

Materials and Methods

To generate the YCF1 dataset, we sampled annotated ORFs of *ycf1* genes from 18 available streptophyte plastid genomes (Lemieux et al. 2000, 2007, 2016; Turmel et al. 2002, 2005, 2006; Civáň et al. 2014) as well as from the moss *Physcomitrella patens* (Sugiura et al. 2003) and the chlorophyte alga *Nephroselmis olivacea* (Turmel et al. 1999), and translated them into protein sequences using Geneious R8.1.8 (Kearse et al. 2012). For *M. endlicherianum*, we removed the intron suggested by Civáň et al. (2014) prior to translation.

Using MAFFT v.7 (Katoh and Standley 2013; G –INS –I), we generated an alignment of all 20 YCF1 proteins and used MEGA7 (Kumar et al. 2016) to compute a maximum likelihood phylogeny applying the LG + G + I + F model (Le and Gascuel 2008; 500 bootstraps, partial deletion [95%] leaving 311 positions, gamma category 5). The 21 embryophyte Ste-YCFs were aligned using the settings L–INS–I in MAFFT v.7.

YCF1 charge was predicted using the EMBOSS explorer (Rice et al. 2000). Predictions of TM domains were generated using TMHMM (Krogh et al. 2001), SOSUI (Hirokawa et al. 1998), and Phobius (Käll et al. 2004). To build a consensus between all TM predictions we used CCTOP (Dobson et al. 2015).

Supplementary Material

Supplementary data is available at *Genome Biology and Evolution* online.

Acknowledgments

We gratefully acknowledge financial support provided by the German Research Foundation (DFG) to S.B.G. (GO1825/4-1) and J.d.V. (VR132/1-1), and the Natural Sciences and Engineering Research Council of Canada awarded to J.M.A. J.M.A. is a Senior Fellow of the Canadian Institute for Advanced Research.

Literature Cited

Adl SM, et al. 2012. The revised classification of eukaryotes. J Eukaryot Microbiol. 59:429–493.

Allen JF, de Paula WBM, Puthiyaveetil S, Nield J. 2011. A structural phylogenetic map for chloroplast photosynthesis. Trends Plant Sci. 16:645–655.

Allen JF. 2015. Why chloroplasts and mitochondria retain their own genomes and genetic systems: colocation for redox regulation of gene expression. Proc Natl Acad Sci U S A. 112:10231–10238.

Archibald JM. 2015. Genomic perspectives on the birth and spread of plastids. Proc Natl Acad Sci U S A. 112:10147–10153.

Becker B, Marin B. 2009. Streptophyte algae and the origin of embryophytes. Ann Bot 103:999–1004.

Berman HM, et al. 2000. The Protein Data Bank. Nucleic Acids Res. 28:235–242.

Bölter B, Soll J. 2016a. Once upon a time – chloroplast protein import research from infancy to future challenges. Mol Plant 9:798–812.

Bölter B, Soll J. 2016b. Ycf1/Tic214 is not essential for the accumulation of plastid proteins. Mol Plant 10:219–221.

Boudreau E, et al. 1997. A large open reading frame (*orf1995*) in the chloroplast DNA of *Chlamydomonas reinhardtii* encodes an essential protein. Mol Gen Genet. 253:649–653.

Civáň P, Foster PG, Embley MT, Senéca A, Cox CJ. 2014. Analyses of charophyte chloroplast genomes help characterize the ancestral chloroplast genome of land plants. Genome Biol Evol. 6:897–911.

Dagan T, et al. 2013. Genomes of stigonematalean cyanobacteria (subsection V) and the evolution of oxygenic photosynthesis from prokaryotes to plastids. Genome. Biol Evol. 5:31–44.

de Vries J, et al. 2013. Is *ftsH* the key to plastid longevity in sacoglossan slugs?. Genome Biol Evol. 5:2540–2548.

de Vries J, Sousa FL, Bölter B, Soll J, Gould SB. 2015. YCF1: a green TIC?. Plant Cell 27:1827–1833.

de Vries J, Stanton A, Archibald JM, Gould SB. 2016. Streptophyte terrestrialization in light of plastid evolution. Trends Plant Sci. 21:467–476.

Delwiche CF, Cooper ED. 2015. The evolutionary origin of a terrestrial flora. Curr Biol. 25:R899–R910.

Dobson L, Reményi I, Tusnády GE. 2015. CCTOP: a Consensus Constrained TOPology prediction web server. Nucleic Acids Res. 43:W408–W412.

Drescher A, Ruf S, Calsa T, Carrer H, Bock R. 2000. The two largest chloroplast genome-encoded open reading frames of higher plants are essential genes. Plant J. 22:97–104.

Finn RD, et al. 2015. HMMER web server: 2015 update. Nucleic Acids Res. 43:W30–W38.

Finn RD, et al. 2016. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44:D279–D285.

Green BR. 2011. Chloroplast genomes of photosynthetic eukaryotes. Plant J. 66:34–44.

Hecht J, Grewe F, Knoop V. 2011. Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. Genome Biol Evol. 3:344–358.

Hirokawa T, Boon-Chieng S, Mitaku S. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 14:378–379.

Hudson M, Ringli C, Boyland MT, Quail PH. 1999. The *FAR1* locus encodes a novel nuclear protein specific to phytochrome A signaling. Genes Dev. 13:2017–2027.

Janska H, Kwasniak M, Szczepanowska J. 2013. Protein quality control in organelles — AAA/FtsH story. Biochim Biophys Acta. 1833:381–387.

Käll L, Krogh A, Sonnhammer ELL. 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 338:1027–1036.

Katayama H, Ogihara Y. 1996. Phylogenetic affinities of the grasses to other monocots as revealed by molecular analysis of chloroplast DNA. Curr Genet. 29:572–581.

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.

Kearse M, et al. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

Kikuchi S, et al. 2013. Translocon at the chloroplast inner envelope membrane. Science 339:571–574.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol. 305:567–580.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. Mol Biol Evol. 33:1870–1874.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. Mol Biol Evol. 25:1307–1320.

Lemieux C, Otis C, Turmel M. 2000. Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature 403:649–652.

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

Lemieux C, Otis C, Turmel M. 2016. Comparative chloroplast genome analyses of streptophyte green algae uncover major structural alterations in the Klebsormidiophyceae, Coleochaetophyceae and Zygnematophyceae. Front Plant Sci. 7:697.

Lewis LA, McCourt RM. 2004. Green algae and the origin of land plants. Am J Bot 91:1535–1556.

Lin R, et al. 2007. Transposase-derived transcription factors regulate light signaling in *Arabidopsis*. Science 318:1302–1305.

Martin W, et al. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature 393:162–165.

Mi H, et al. 2016. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Res. 45:D183–D189.

Nakai M. 2015a. The TIC complex uncovered: the alternative view on the molecular mechanism of protein translocation across the inner envelope membrane of chloroplasts. Biochim Biophys Acta. 1847:957–967.

Nakai M. 2015b. YCF1: a Green TIC: response to the de Vries *et al.* Commentary. Plant Cell 27:1834–1838.

NCBI Resource Coordinators. 2016. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 44:D7–D19.

Oldenkott B, Yamaguchi K, Tsuji-Tsukinoki S, Knie N, Knoop V. 2014. Chloroplast RNA editing going extreme: more than 3400 events of C-to-U editing in the chloroplast transcriptome of the lycophyte *Selaginella uncinata*. RNA 20:1499–1506.

Parfrey LW, Lahr DJG, Knoll AH, Katz LA. 2011. Estimating the timing of early eukaryotic diversification with multigene molecular clocks. Proc Natl Acad Sci U S A. 108:13624–13629.

Ponce-Toledo RI, Deschamps P, López-García P, Zivanovic Y, Benzerara K, Moreira D. 2017. An early-branching freshwater cyanobacterium at the origin of plastids. Curr Biol. 27:386–391.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 16:276–277.

Sugiura C, Kobayashi Y, Aoki S, Sugita C, Sugita M. 2003. Complete chloroplast DNA sequence of the moss *Physcomitrella patens*: evidence for the loss and relocation of *rpoA* from the chloroplast to the nucleus. Nucleic Acids Res. 31:5324–5331.

The UniProt Consortium. 2015. UniProt: a hub for protein information. Nucleic Acids Res. 43:D204–D212.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nat Rev Genet. 5:123–135.

Turmel M, Otis C, Lemieux C. 1999. The complete chloroplast DNA sequence of the green alga *Nephroselmis olivaceae*: insights into the architecture of ancestral chloroplast genomes. Proc Natl Acad Sci U S A. 96:10248–10253.

Turmel M, Otis C, Lemieux C. 2002. The chloroplast and mitochondrial genome sequences of the charophyte *Chaetosphaeridium globosum*: insights into the timing of the events that restructured organelle DNAs within the green algal lineage that led to land plants. Proc Natl Acad Sci U S A. 99:11275–11280.

Turmel M, Otis C, Lemieux C. 2005. The complete chloroplast DNA sequences of the charophyceaen green algae *Staurastrum* and *Zygnema* reveal that the chloroplast genome underwent extensive changes during the evolution of the Zygnematales. BMC Biol. 3:22.

Turmel M, Otis C, Lemieux C. 2006. The chloroplast genome sequence of *Chara vulgaris* sheds light into the closest green algal relatives of land plants. Mol Biol Evol. 23:1324–1338.

Van de Poel B, Cooper ED, Van Der Straeten D, Chang C, Delwiche CF. 2016. Transcriptome profiling of the green alga Spirogyra pratensis (Charophyta) suggests an ancestral role for ethylene in cell wall metabolism, photosynthesis, and abiotic stress responses. Plant Physiol. 172:533–545.

Wicke S, Schneeweiss GM, dePamphilis CW, Müller KF, Quandt D. 2011. The evolution of the plastid chromosome in land plants: gene content, gene order, gene function. Plant Mol Biol. 76:273–297.

Wickett NJ, et al. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc Natl Acad Sci U S A. 111:E4859–E4868.

Associate editor: Geoff McFadden