Cyanophora paradoxa Genome Elucidates Origin of Photosynthesis in Algae and Plants

Dana C. Price, Cheong Xin Chan, Hwan Su Yoon, Eun Chan Yang, Huan Qiu, Andreas P. M. Weber, Rainer Schwacke, Jeferson Gross, Nicolas A. Blouin, Chris Lane, Adrián Reyes-Prieto, Dion G. Durnford, Jonathan A. D. Neilson, B. Franz Lang, Gertraud Burger, Jürgen M. Steiner, Wolfgang Löffelhardt, Jonathan E. Meusel, Matthew C. Posewitz, Steven Ball, Maria Cecilia Arias, Bernard Henrisatt, Pedro M. Coutinho, Stefan A. Rensing, Aikaterini Symeonidi, Harshvardhan Doddapaneni, Beverley R. Green, Veeran D. Rajah, Jeffrey Boore, Debashish Bhattacharya

The primary endosymbiotic origin of the plastid in eukaryotes more than 1 billion years ago led to the evolution of algae and plants. We analyzed draft genome and transcriptome data from the basally diverging alga Cyanophora paradoxa and provide evidence for a single origin of the primary plastid in the eukaryote supergroup Plantae. C. paradoxa retains ancestral features of starch biosynthesis, fermentation, and plastid protein translocation common to plants and algae but lacks typical eukaryotic light-harvesting complex proteins. Traces of an ancient link to parasites such as Chlamydiae were found in the genomes of C. paradoxa and other Plantae. Apparently, Chlamydia-like bacteria donated genes that allow export of photosynthetic from the plastid and its polymerization into storage polysaccharide in the cytosol.

Eukaryote evolution has largely been shaped by the process of primary endosymbiosis, whereby bacterial cells were taken up and over time evolved into double membrane–bound organelles, the plastid and the mitochondrion [e.g., (1, 2)]. The cyanobacterium-derived plastid is found in diverse photosynthetic organisms, including Glaucophyta, Rhodophyta, and green algae and their land plant descendants (the Viridiplantae). These three lineages are postulated to form a monophyletic group Plantae (or Archaeprotista) (3–6), a hypothesis that suggests the primary cyanobacterial endosymbiosis occurred exclusively in their single common ancestor. Plastid gene trees demonstrate a single origin of the Plantae (5, 7); however, many nuclear, multiprotein phylogenies provide little (8) or no support (9, 10) for their monophyly. These latter results may reflect a reticulate ancestry among genes that can mislead phylogenetic inference (11). Furthermore, glaucophytes retain ancestral cyanobacterial features not found in other Plantae (12)—such as the presence of peptidoglycan between the two bounding membranes of the plastid (13)—that cast doubt on their evolutionary history. It is therefore unclear whether the Plantae host and its plastid, with its associated complex machinery (e.g., for plastid protein import and solute transport) (14, 15), had a single origin or multiple origins. To elucidate the evolutionary history of key algal and land plant traits and to test Plantae monophyly, we have generated a draft assembly of the ~70 Mbp nuclear genome from the glaucophyte Cyanophora paradoxa CCMP329 (Pringsheim strain) (Fig. 1A).

A total of 27,921 C. paradoxa proteins were predicted from the genome data, and 4628 had significant BLASTp hits (e < 10^-10) to prokaryote and eukaryote genome data in our comprehensive local database (table S1). Using phylogenomics (16), we generated 4445 maximum likelihood trees from the C. paradoxa proteins and found that >60% support a sister-group relationship between glaucophytes and red and/or green algae with a bootstrap value ≥90% (Fig. 1B and fig. S1). The Plantae clade in many of these trees is, however, interrupted by chlorophyll a+ c containing “chroomateolates.” An example of this type of tree is fungus-1,6-bisphosphatase (Fig. 1C and fig. S2), which has cytosolic and plastidic isoforms. The gene for this enzyme, found in streptomycoses (e.g., diatoms) and haptophytes, originated from the red algal secondary endosymbiont that gave rise to the plastid in these taxa (2, 9). This sort of intracellular gene transfer associated with endosymbiosis (EGT) has greatly enriched algal and land plant genomes (17, 18).

We estimated the “footprint” of cyanobacterium-derived EGT in Plantae genomes. The proportion of cyanobacterium-derived nuclear genes varies from 18% in Arabidopsis thaliana (19) to ~7% in mesopholic red algae and 6% in Chlamydomonas reinhardtii (20, 21). Phylogenomic analysis of the predicted C. paradoxa proteins showed 274 to be of cyanobacterial provenance (22). This constitutes ~6% of proteins in the glaucophyte that have significant BLASTp hits (i.e., 274 out of 4628), as found in other algae (20, 21). BLASTp analysis identified 2029 proteins that are putatively destined for the plastid, of which 293 contain the transition sequence for plastid import [identified by the presence of phenylalanine (F) within the first four amino acids: MF, MAF, MNAF, and MAAF] (23, 24) (fig. S4B). Of these 293 proteins, 80% are derived from Cyanobacteria.

Another source of foreign genes in Plantae is horizontal gene transfer (HGT), which is not associated with endosymbiosis. Using 35,126 bacterial sequences as a query, we found 444 noncyanobacterial gene families with a common origin shared amongst Bacteria and Plantae. Among them, 15 genes are present in all three Plantae phyla. An example of a gene derived from Bacteria after an ancient HGT event that is shared by
Plantae is that encoding a thiamine pyrophosphate-dependent pyruvate decarboxylase family protein involved in alcohol fermentation (Fig. 1D). An- other 60 genes were present in only two of the tree phyla (i.e., 24, 10, and 26 genes in Glaucophyta-Viridiplantae, Glaucophyta-Rhodophyta, and Rhodophyta-Viridiplantae, respectively) (22). We sequenced the mitochondrial genome from C. paradoxa and from the distantly related glaucophyte Glacocystis nostochinearum and generated a near-complete plastid genome sequence from G. nostochinearum that was added to the existing plastid genome data from C. paradoxa (GenBank NC_001675). The mitochondrial DNAs (mtDNAs) of C. paradoxa and G. nostochinearum share similar characteristics, including a large gene content, but differ markedly in size [Fig. 3A, tables S3 and S4, and discussion in the supporting online material (SOM)]. A concatenated multigene (17,049 aligned amino acid positions) phylogeny of the plastid data shows the expected monophyly of Plantae plastids (2, 5, 6) and places glauco- phytes very close to the divergence point of red and green algae (fig. SSA).

Building on the support for Plantae monophyly provided by the phylogenetic analyses, we analyzed the evolution of landmark traits that are associated with plastid endosymbiosis. Previous work suggests that in red and green algae, the initial link for carbon metabolism between the

Fig. 1. (A) Schematic (left) and transmission electron micrograph (right) images of C. paradoxa. This freshwater microalga has two flagella (f) of unequal lengths, and the plastid (p) contains an electron-dense central body (cb). The nucleus (n), mitochondrion (m), starch granules in the cytoplasm (st), and peptidoglycan septa (s) formed during plastid division are also shown. (B) Percentage of single-protein Randomized Accelerated Maximum Likelihood (RaXML) trees (raw numbers shown in the bars) that support the monophyly of Glaucophyta (bootstrap ≥ 90%) solely with other Plantae members, or in combination with non-Plantae taxa that interrupt this clade. These latter groups of trees are primarily explained by red/green algal EGT into the nuclear genome of chromalveolates and euglenids. For each of these algal lineages, the set of trees with different numbers of taxa (N) ≥ 4, ≥ 10, ≥ 20, ≥ 30, and ≥ 40 and distinct phyla ≥ 3 in a tree are shown. The Plantae-only groups are Glaucophyta-Rhodophyta (GIR), Glaucophyta-Viridiplantae (GIVI), and Glaucophyta-Viridiplantae (GIRVI). Trees with evidence of EGT are shown as the single group, GIR/GIVI/GIRVI. An expanded analysis that shows the results when red or green algae are used as the query to search for support for Plantae monophyly is summarized in fig. S1. (C) Schematic ML phylogeny of a gene encoding a thiamine pyrophosphatase (TPP)—dependent pyruvate decarboxylase family protein involved in alcohol fermentation. RaXML bootstrap support values are shown at the nodes of the full tree is shown in fig. S2. (D) Schematic ML phylogeny of a gene encoding a thiamine pyrophosphatase (TPP)—dependent pyruvate decarboxylase family protein involved in alcohol fermentation. RaXML bootstrap support values are shown at the nodes of the trees in panels (C) and (D), in which glaucohytes, red algae, green algae, and chromalveolates are in purple, red, green, and brown, respectively.
host cell and plastid relied on sugar-phosphate transporters that evolved from existing host endomembrane nucleotide sugar transporters (NSTs) (I5, 25). Unexpectedly, we found that although six endomembrane-type NST genes exist in C. paradoxa (fig. S3), this alga lacks plastidial phosphate-translocator (PT) genes (Fig. 2A). We searched the genome for alternative candidate genes encoding homologs of bacterial carbon exporters and found two UhpC-type hexose-phosphate transporters (Fig. 2B) that are also present in red and green algae. One of these putative transporters (contig 37408) encodes a plastid-targeting signal containing a conserved phenylalanine (F) and proline (P) at the N terminus typical for C. paradoxa plastid-targeted proteins [e.g., (23)] (Fig. 2, B and C). The second UhpC homolog contains an F and a P at the N terminus and may also be plastid-targeted. These transporters are related to sequences in the parasites Chlamydiae and Legionella. We suggest that the UhpC gene originated in Plantae and Legionella species through independent HGT events from Chlamydiae. This direction of gene transfer for the prokaryotes is supported by the nested position of a single proteobacterial clade (i.e., Legionella) within Chlamydiae. Chlamydiae have contributed a substantial number of other genes to Plantae that are involved in plastid functions (26, 27).

A second landmark trait of photosynthetic eukaryotes is the presence of protein-conducting channels (translocons) in the outer and inner envelope membranes of plastids (Toc and Tic, respectively) [e.g., (14)], for which there is biochemical evidence in the C. paradoxa plastid (28). We identified nuclear-encoded homologs of the major Toc75 and Tic110 translocon proteins, two Toc34-like receptors, homologs of the plastid Hsp70 and Hsp93 chaperones, and a stromal processing peptide (table S2). This minimal set of components likely formed the ancestral plastid protein translocation system in Plantae (29). Candidates for additional translocon subunits Tic20 to Tic22, Tic32, Tic55, and Tic62 were also found in the glaucophyte. The presence of a conserved core of cyanobacterium-derived, homologous translocon subunits shared by C. paradoxa and other Plantae (i.e., Toc75, Tic20, and Tic22) or those apparently evolved de novo in the host (i.e., Tic34 and Tic110) (table S2 and fig. S4A) provides strong evidence that the primary plastid was established in a single ancestor of Plantae (I4, 29, 30).

A key component of plastids is light-harvesting complex (LHC) proteins that increase the capacity to capture incoming light. Cyanobacteria, red algae, and glaucophytes use phycobilisomes (protein complexes anchored to the thylakoid) for light harvesting, but red algae also have nuclear-encoded pigment-binding proteins homologous to the typical green algal LHClcs but specifically associated with photosystem I (31). Unexpectedly, no candidate genes with three membrane-spanning regions, characteristic of LHClcs in all other photosynthetic eukaryotes, were found in C. paradoxa. This alga does, however, encode several members of the LHC-like (LIL) protein family that contain a single chlorophyll-binding transmembrane helix, including two copies of the “one-helix proteins” (OHPs) (see fig. S8 and SOM).

The anaerobic capabilities of C. paradoxa provide parallels to the well-characterized anoxic metabolism of the green alga Chlamydomonas reinhardtii.

---

**Fig. 2.** (A) Maximum likelihood (PhyML) phylogeny of endomembrane (NST) and plastidial phosphate-transporters (PTs). PhyML bootstrap values are shown above the branches, and RAxML bootstrap values are shown below the branches in italics. Only bootstrap values ≥50% are shown. Red algae, Viridiplantae, and chromalveolates are shown in red, green, and brown, respectively. The different plastid-targeted transporters are glucose 6-phosphate translocator (GPT), phosphono-ylpyruvate translocator (PTT), triose phosphate translocator (TPT), and xylulose 5-phosphate translocator (XPT). The numbers in the filled circles indicate transporters shown to the right that (1) resulted from primary endosymbiosis and (2) were transferred from the red algal secondary endosymbiont to the chromalveolate host. The full tree is shown in fig. S3. (B) RAxML phylogeny of UhpC-type hexose-phosphate transporters in algae, land plants, and Bacteria. RAxML bootstrap values are shown above the branches, and PhyML bootstrap values are shown below the branches in italics. Only bootstrap values ≥50% are shown. Red algae, Viridiplantae, and chromalveolates are shown in red, green, magenta, and pink, respectively. The numbers after the C. paradoxa contig names are bioinformatic predictions for plastid targeting of the encoded proteins with Predotar (P), WoLF PSORT (WP), TargetP (TP), and ChloroP (CP). These data suggest that contig 37408 (and perhaps also contig 54308) is destined for plastid targeting in C. paradoxa.

---

**C**: Introns (A) Genes encoding plastid-targeted transporters (B) encode a substantial number of other genes to Plantae that are involved in plastid functions (26, 27).

---

**Table 1**: Genome coverage depth
The predicted pathways in C. paradoxa suggest a complex heterofermentative state in the Plantae ancestor (Fig. 3B), with retention in Chlorophyta (e.g., Chlorophyceae and Trebouxiophyceae) and losses in Streptophyta. Orthologs of fermentation enzymes encoded by the C. paradoxa genome can also be found in Cyanobacteria (for lactate, formate, acetate, and ethanol) and hydrogenosomal euukaryotes (for H₂, lactate, formate, and ethanol). Therefore, the complex fermentative capabilities conserved between the distantly related C. paradoxa and green algae (33) likely represent an evolutionarily advantageous combination of anoxic enzymes from the eukaryote host and the cyanobacterial endosymbiont. Genome evidence for H₂ metabolism in C. paradoxa is demonstrated by the presence of a [FeFe]-hydrogenase and associated maturases (HydE, HydF, and HydG).

Putative carbohydrate metabolism enzymes in C. paradoxa were identified using the Carbohydrate-Active enZymes (CAZy) database (34) annotation.

**Fig. 3.** (A) Gene maps of glaucophyte mitochondrial genomes. Black blocks indicate genes that are arranged outside and inside the circle, indicating they are transcribed clockwise and counterclockwise, respectively. rRNA gene names are abbreviated in single–amino acid letter code; subscript numbers distinguish isoacceptor tRNAs. The color-coding identifies genes that are typically present in fungal and animal mtDNAs (black); additional genes found in land plants and protists are marked in blue and unidentified ORFs in green. The arc marks the genome region with long segmental duplications in C. paradoxa mtDNA. (B) Phylogeny of small subunit rRNA of organisms with sequenced genomes that share predicted fermentative orthologs encoded in the C. paradoxa genome. Predicted orthologs encoded by C. paradoxa are shown boxed. Enzyme abbreviations are as follows: Fe-ADH, iron-containing alcohol dehydrogenase; ADHE, iron-containing alcohol/alddehyde dehydrogenase; ACK, acetate kinase; HYDA, [FeFe]-hydrogenase; HYDE, radical S-adenosylmethionine HYDA maturase; HYDF, fused HYDE and HYDF maturase; HYDG, radical S-adenosylmethionine HYDA maturase; LDH, lactate dehydrogenase; NARF, nuclear prelamin A recognition factor; PTA, phosphotransacetylase; PDC, pyruvate decarboxylase; PFL, pyruvate:formate lyase; PFLA, pyruvate:formate lyase activase; PFR, pyruvate:ferredoxin oxidoreductase; PNO, pyruvate:NADP oxidoreductase. The darker filled circle represents PNO, a divided fill represents evidence for both enzymes, and white crosses indicate the heteromeric PFRs present in the Thermotogae. Shading is also used to discriminate the single-domain type II PTA isoforms. Single-letter abbreviations for the amino acids are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
pipeline. The genome of this alga encodes ~84 glycoside hydrolases (GHs) and 128 glycosyl transferases (GTs). This is far greater than in the marine green microalga *Ostreococcus lucimanus* (*CCE9901*) and the extremophilic red alga *Cyanidioschyzon merolae* but less than in *Arabidopsis thaliana* (Table 1). Consistent with these results, inspection of the number of CAZy families (22) shows that *C. paradoxa* contains twice the number of GH families and 20% more GT families than *O. lucimanus* and *C. merolae*. The smaller number of CAZy families in these unicellular algal lineages when compared to *A. thaliana* suggests less functional redundancy.

Many *C. paradoxa* CAZymes are involved in starch metabolism (22). Synthesis of the polysaccharide within Viridiplantae plastids relies on enzymes of the GT5 CAZY (34) family associated with glycogen synthesis in Bacteria such as adenosine diphosphate (ADP)-glucose pyrophosphorylases and ADP-glucose using starch synthases (SSs). Eukaryotes synthesize glycogen from uridine diphosphate (UDP)-glucose using either a GT3 (all fungi, some Amoebozoa, and animals) or a GT5 (alveolates, parabasalids, and Amoebozoa) type of transferase. The major *C. paradoxa* enzyme is phylogenetically related to the GT5 UDP-glucose-specific enzyme of heterotrophic eukaryotes (35) and has been partially purified from this alga (36). This suggests the absence of ADP-glucose pyrophosphorylase in *C. paradoxa*. Therefore, it was unexpected to find a second gene in the glucophyte genome whose gene product is related to the SSIII-SSIV (GT5) type of starch synthases in Viridiplantae. This gene is phylogenetically related to glucan synthase in Chlamydiae, Cyanobacteria, and some Proteobacteria (fig. S5B) and likely played a key role in linking the biochemistry of the host and the endosymbiont. The SSIII-SSIV enzyme uses ADP-glucose in Bacteria and land plants (35), suggesting that *C. paradoxa* or, alternatively, the common ancestor of Viridiplantae and glucophytes may have used both types of nucleotide sugars for starch synthesis.

Analysis of the gene-rich *C. paradoxa* genome unambiguously supports Plantae monophyly (2, 5, 6, 14, 15) (see discussion of Plantae branching order in the SOM), laying to rest a long-standing issue in eukaryote evolution. Plantae share many genes with an EGT or HGT origin that have essential functions such as photosynthesis, starch biosynthesis, plastid protein import, plastid soluble transport, and alcohol fermentation. The alternative explanation of a polyphyletic Plantae would require the unlikely combination of a large number of independent HGT events in its major phyla followed by gene loss in all (or many) other eukaryotes. Consolidation of the Plantae allows insights into the gene inventory of their common ancestor. It is now clear that the Plantae ancestor contained many of the key innovations that characterize land plant and algal genomes, including extensive EGT from the cyanobacterial endosymbiont and retargeting of plastid-destined proteins, the minimal machinery required for plastid protein translation, and complex pathways for fermentation and starch biosynthesis. In spite of the fact that glucophytes retain peptidoglycan, an ancestral trait lost by all other algae (except the primary plastid in *Paulinella* (37) and land plants, glucophytes are not a lineage of “living fossils.” Rather, the *C. paradoxa* genome contains a unique combination of ancestral, novel, and “borrowed” (e.g., via HGT) genes, similar to the genomes of other Plantae (38).

References and Notes