Genomes of Stigonematalean Cyanobacteria (Subsection V) and the Evolution of Oxygenic Photosynthesis from Prokaryotes to Plastids

Tal Dagan[1,*,†], Mayo Roettger[2,†], Karina Stucken[1], Giddy Landan[1,2], Robin Koch[1], Peter Major[2], Sven B. Gould[2], Vadim V. Goremykin[3], Rosmarie Rippka[4], Nicole Tandeau de Marsac[4,10], Muriel Gugger[5], Peter J. Lockhart[6], John F. Allen[7,8], Iris Brune[9], Irena Maus[9], Alfred Pühler[9], and William F. Martin[2]

[1]Institute of Genomic Microbiology, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

[2]Institute of Molecular Evolution, Heinrich-Heine-University Düsseldorf, Düsseldorf, Germany

[3]IASMA Research and Innovation Center, Fondazione Edmund Mach, San Michele all'Adige (TN), Italy

[4]Institut Pasteur, Unité des Cyanobactéries, Paris, France

[5]Institut Pasteur, Laboratoire Collection des Cyanobactéries, Paris, France

[6]Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand

[7]School of Biological and Chemical Sciences, Queen Mary, University of London, London, United Kingdom

[8]Research Department of Genetics Evolution and Environment, University College London, London, United Kingdom

[9]Center for Biotechnology, University of Bielefeld, Bielefeld, Germany

[10]Present address: Aix-Marseille University, Laboratoire de Chimie Bactérienne (LCB), Marseille, France

[†]These authors contributed equally to this work.

***Corresponding author:** E-mail: tal.dagan@hhu.de; tal.dagan@uni-duesseldorf.de.

Accepted: December 4, 2012

Data deposition: Genomes have been deposited in NCBI under accessions PRJNA104961, PRJNA104963, PRJNA104969, PRJNA104967, PRJNA104965, and PRJNA157363.

Abstract

Cyanobacteria forged two major evolutionary transitions with the invention of oxygenic photosynthesis and the bestowal of photosynthetic lifestyle upon eukaryotes through endosymbiosis. Information germane to understanding those transitions is imprinted in cyanobacterial genomes, but deciphering it is complicated by lateral gene transfer (LGT). Here, we report genome sequences for the morphologically most complex true-branching cyanobacteria, and for *Scytonema hofmanni* PCC 7110, which with 12,356 proteins is the most gene-rich prokaryote currently known. We investigated components of cyanobacterial evolution that have been vertically inherited, horizontally transferred, and donated to eukaryotes at plastid origin. The vertical component indicates a freshwater origin for water-splitting photosynthesis. Networks of the horizontal component reveal that 60% of cyanobacterial gene families have been affected by LGT. Plant nuclear genes acquired from cyanobacteria define a lower bound frequency of 611 multigene families that, in turn, specify diazotrophic cyanobacterial lineages as having a gene collection most similar to that possessed by the plastid ancestor.

Key words: plastid evolution, endosymbiosis, phylogenomics, true-branching cyanobacteria, nitrogen fixation.

Introduction

Cyanobacteria are crucial players in Earth and life history because they generated the oxygen that has been present in the Earth's atmosphere for the last 2.4 billion years (Bekker et al. 2004) and because one uniquely fateful cyanobacterium became, via endosymbiosis, the ancestor of all plastids among photosynthetic eukaryotes (Gould et al. 2008). Though they continue to impact global geochemical cycles through N_2-fixation (Moisander et al. 2010), and the

sequestering of trace metals (Morel and Price 2003) as well as phosphorous (van Mooy et al. 2009), their main ecological significance is the oxygen-producing photosynthetic apparatus that fuels most contemporary food chains. Their main evolutionary significance is that they mediated two pivotal innovations in life's history—water-splitting photosynthesis and the origin of primary plastids. Clues to both of those major evolutionary transitions should, in principle, be imprinted in cyanobacterial genomes. But reconstructing those events is not straightforward, because lateral gene transfer (LGT) redistributes genes among prokaryote genomes (Ochman et al. 2000), and among cyanobacterial genomes in particular (Raymond et al. 2002; Mulkidjanian et al. 2006; Dufresne et al. 2008; Shi and Falkowski 2008), over geological time.

By necessity, and perhaps more so than for any other prokaryotic group, LGT has always been hard-wired into the bigger picture of cyanobacterial evolution. To explain the origin of cyanobacterial water-splitting photosynthesis, both of the main competing theories require LGT to account for the distribution of photosystems across prokaryotic groups (Xiong and Bauer 2002; Hohmann-Marriot and Blankenship 2011). This is because the reaction centers of photosystems I and II clearly share common ancestry (Baymann et al. 2001; Hohmann-Marriot and Blankenship 2011), but without specifying how they entered the cyanobacterial ancestor genome. One theory posits that the two photosystems evolved in independent lineages and became merged in the founder cyanobacterium via LGT (Baymann et al. 2001), while the alternative has it that the photosystems diverged within a photosynthetic (protocyanobacterial) ancestor and were subsequently exported via LGT to some anoxygenic photosynthetic lineages (Xiong and Bauer 2002; Mulkidjanian et al. 2006; Sharon et al. 2009). Compatible with a role for LGT in photosystem evolution is the finding that the genes for both photosystems I and II are mobile in marine phage metagenomes (Lindell et al. 2004; Sharon et al. 2009).

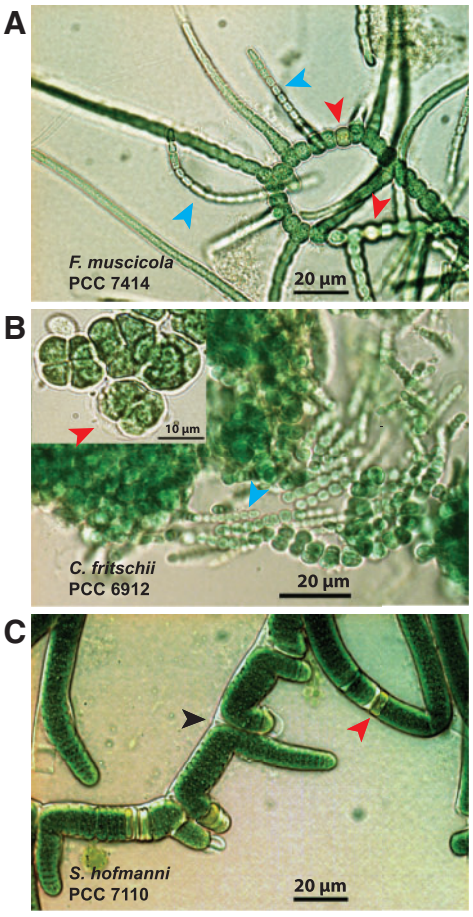
LGT also figures into the origin of plastids, because many genes were transferred from endosymbiont to host. Chloroplasts were once free-living cyanobacteria and contained approximately 2,000 proteins (Richly and Leister 2004), a number comparable with a cyanobacterium, yet the genomes of modern plastids contain only 5–10% as many genes as those of their free-living cousins. This suggests that hundreds or thousands of the plastid ancestor's genes were either lost or relocated to the host nucleus during the course of plant evolution via endosymbiotic gene transfer (EGT) (Gould et al. 2008). Furthermore, the phylogenetic identity of the plastid ancestor remains debated because of LGT. Different phylogenetic trees trace the plastid ancestor near the base of cyanobacterial diversification (Criscuolo and Gribaldo 2011), near coccoid cyanobacteria within the *Synechococcus–Prochlorococcus* (SynPro) clade (Reyes-Prieto et al. 2010), near the nitrogen-fixing *Cyanothece* clade (Deschamps et al. 2008), or near filamentous, heterocyst-forming cyanobacterial lineages (Deusch et al. 2008). The simplest explanation for such findings—in an evolutionary context that incorporates LGT—is that the plastid ancestor donated one (chimeric) genome's worth of genes to the host, and that LGT has been reassorting the homologs of these genes among free-living cyanobacterial and other prokaryote genomes ever since (Deusch et al. 2008). Because of LGT over time, the question of which "lineage" of cyanobacteria gave rise to the plastid loses meaning (Doolittle and Bapteste 2007), because the genomes and nature of the "lineages" have changed since the time of plastid origin over 1.2 billion years ago (Deusch et al. 2008; Gross et al. 2008). However, comparison of plant genes acquired from the plastid ancestor with cyanobacterial homologs can reveal which modern cyanobacteria harbor a collection of genes most similar to that of the plastid ancestor.

So far, missing in genomic studies of cyanobacterial evolution are sequences from the group designated as subsection V (Rippka et al. 1979). Subsection V cyanobacteria grow as filaments that differentiate heterocysts (specialized N_2-fixing cells), they produce cyst-like resting cells (akinetes) as well as differentiated motile trichomes (hormogonia), and most exhibit true branching. The developmental and morphological variety of subsection V cyanobacteria places them among the most complex of prokaryotes, for which reason they were even long thought to be the direct ancestors of all eukaryotes but only in the days before the endosymbiotic origin of plastids has been postulated (Mereschkowsky 1905) and eventually gained compelling support (Doolittle 1980). To better understand the role of subsection V species in cyanobacterial evolution and their possible relationship to the plastid ancestor, we have sequenced five genomes sampling a broad spectrum of filamentous, true-branching architecture (fig. 1*A* and *B*), and diverse geographical locations including rice fields in India (*Fischerella muscicola* PCC 73103 and *Chlorogloeopsis fritschii* PCC 6912), and hot springs in New Zealand (*F. muscicola* PCC 7414), Wyoming, USA (*F. thermalis* PCC 7521), and in Spain (*C. fritschii* PCC 9212) (Rippka et al. 1979). In addition *Scytonema hofmanni* PCC 7110, a Nostocales representative (subsection IV) isolated from a limestone cave (Crystal cave, Bermuda) (Rippka et al. 1979), whose filaments form false branches (fig. 1*C*) and exhibit aerial growth, was included for comparison.

Materials and Methods

Cyanobacterial Cultures and DNA Isolation

Stock cultures were maintained at 37°C on slants (or plates) in BG11o medium (Rippka and Herdman 2002), supplemented with 5 mM $NaHCO_3$ and solidified with 0.9% (w/v) washed agar (Sigma, A 8678). For DNA isolation, cultures were grown at 37°C in BG11 medium (Rippka and Herdman 2002), with orbital shaking (100 rpm) in an Infors Incubator, at a PPFD of



D

Genus	*Chlorogloeopsis*		*Fischerella*			*Scytonema*
Species	*fritschii*	*fritschii*	*muscicola*	*muscicola*	*thermalis*	*hofmanni*
PCC strain	6912	9212	73103	7414	7521	7110
Assembled scaffolds	8	21	37	16	4	81
Bases in scaffolds	7,819,561	7,698,501	7,573,287	7,008,859	5,482,660	12,073,012
%GC	41.5	41.5	38.6	40.4	40.5	41.5
No. CDSs	7,439	7,571	7,231	7,167	5,340	12,356
CDS density (%)	82	82	78	79	79	83
Mean gene length (bp)	830	829	820	778	825	806
rRNA operons	4	3	3	3	4	3
No. tRNA	45	41	43	59	42	67
No. Plasmids	2	n. d.[(1)]	n. d.[(1)]	0	0	n. d.[(1)]

[(1)] Not determined.

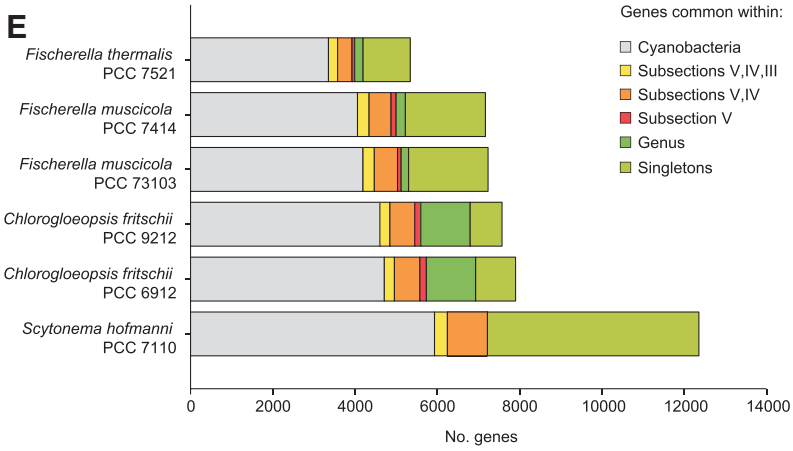


Fig. 1.—Genomes of Stigonematales and *Scytonema*. (*A*) *Fischerella muscicola* PCC 7414, forming true lateral branches. (*B*) *Chlorogloeopsis fritschii* PCC 6912, undergoing cell divisions in more than one plane but never producing lateral branches. Heterocysts and hormogonia, differentiated by members of both genera are marked by red and cyan arrows, respectively. (*C*) *Scytonema hofmanni* PCC 7110 showing false branching filaments (black arrow) and heterocysts (red arrow). (*D*) Genomic features of the six novel sequenced genomes. Genomes have been deposited in NCBI under accessions (PRJNA104961, PRJNA104963, PRJNA104969, PRJNA104967, PRJNA104965, and PRJNA157363). Fully annotated versions are available at www.molevol.de/resources. (*E*) Frequency distribution of protein coding genes in the new genomes, and comparison with other cyanobacterial genomes examined.

30 µmol quanta $m^{-2}s^{-1}$. Cultures were harvested after 3–6 weeks of incubation, depending on density of the inoculum and the growth rates of the strains. DNA isolation from strains of *Chlorogloeopsis* was performed as described (Franche and Damerval 1988), with the addition of 1% Sarkosyl during lysozyme treatment to remove polysaccharides and a final RNA digestion step. Polysaccharide-free high molecular weight genomic DNA (gDNA) from strains of *Fischerella* was obtained by following a protocol for polysaccharide-rich plants (Sharma et al. 2002).

Genome Sequencing and Annotation

Prior to genome sequencing the identity of the gDNA was verified by sequencing of the 16S rDNA with primers 101F (ACTGGCGGACGGGTGAGTAA) and 1047R (GACGACAGCC ATGCAGCACC), and comparison against cyanobacterial sequences available in NCBI. Genome sequencing was performed on the Genome Sequencer FLX using Titanium chemistry (Roche Applied Science, Penzberg, Germany) yielding a 10- to 32-fold coverage. Genome scaffolding was achieved by 3 kbp paired-end standard runs. The sequencing libraries were prepared from 4 µg of gDNA for whole genome shot gun sequencing and 5 µg of gDNA for paired-end sequencing, according to the supplier's instructions. Additionally, a fosmid library was constructed with the Copy Control Fosmid Library Production Kit (Biozym Scientific, Hess. Oldendorf, Germany). Terminal DNA sequences of cloned genomic inserts were determined with an ABI 3730xl DNA Analyzer (Life Technologies, Darmstadt, Germany). Furthermore, Sanger-reads were generated from fosmid clones to cover the gaps between contigs for each of the five genomes. Sequence data were assembled with the GS De Novo Assembler Software (ver. 2.0.01.14, 2.3, and 2.5.3). For each genome, large (>500 bp) and small contigs

(<500 bp) were obtained, including numerous repetitive elements and insertion segments. For finishing purposes, all DNA sequences were uploaded into the Consed program (Gordon et al. 1998). The final annotation including COGs (Tatusov et al. 2001) of the genome sequences was accomplished with the GenDB software (Meyer et al. 2003). Gene prediction was performed by means of combining results of the software tools GLIMMER (Delcher et al. 1999), CRITICA (Badger and Olson 1999), and GISMO (Krause et al. 2007).

Phylogenetic Analysis of Cyanobacterial Genomes

Fully sequenced cyanobacterial proteomes were downloaded from NCBI version March/2011. For the reconstruction of cyanobacterial gene families, we conducted an all-against-all BLAST search (Ver. 2.2.17) (Altschul et al. 1997) using the protein sequences. Reciprocal best BLAST hits (rBBH) were performed using a threshold of E value $\leq 10^{-10}$ and percent amino acid identity $\geq$30. For the clustering analysis, the overall protein sequence similarity between rBBH proteins, calculated as the percent of identical amino acids, was multiplied by the length ratio of the two proteins. Clusters of gene families were inferred from the rBBH similarity matrix using the MCL ver. 1.008 clustering procedure (Enright et al. 2002), with the inflation parameter (I) set to 2.0. For the reconstruction of a consensus tree phylogeny, 324 gene families present as single copies in all cyanobacterial genomes analyzed were aligned with MAFFT (Katoh et al. 2002) ver. 6.717b. Phylogenetic trees were reconstructed using the Neighbor-Joining (NJ) approach (Saitou and Nei 1987). Protein sequence distances were calculated with PROTDIST (Felsenstein 1993), and applying the JTT substitution model (Jones et al. 1992). Phylogenetic trees were reconstructed with NEIGHBOR (Felsenstein 1993). The consensus phylogeny was reconstructed with CONSENSE (Felsenstein 1993). A concatenated alignment was reconstructed from the aligned protein sequences, and all genes were weighted equally (supplementary fig. S1, Supplementary Material online). A phylogenetic tree was reconstructed from the concatenated alignment using the NJ approach and the software described as earlier. A phylogenetic network was reconstructed with SplitsTree Ver. 4.10 using the default parameters (Huson and Bryant 2006). A minimal lateral network (MLN) was reconstructed using the consensus phylogeny as the reference tree, and the gene families described earlier according to the approach described in Dagan et al. (2008). Maximum likelihood phylogeny was reconstructed using PhyML (Guindon et al. 2010) with LG model + I (estimation of invariant sites) + G (gamma distribution with 4 rate categories). Tree topology (SPR), branch length, and rate parameters were optimized.

Phylogenetic Analysis of the Plastid Ancestor

Sequences of nuclear-encoded proteins from the whole genomes of *Arabidopsis thaliana, Oryza sativa* subsp. *japonica*, *Physcomitrella patens, Chlamydomonas reinhardtii, Entamoeba histolytica, Dictyostelium discoideum, Filobasidiella neoformans, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster, Ciona intestinalis, Danio rerio, Gallus gallus, Canis lupus familiaris*, and *Homo sapiens* were obtained from RefSeq database release November 2009 (Pruitt et al. 2007). Nuclear proteomes of *Cyanidioschyzon merolae* version February 2005 (Matsuzaki et al. 2004), *Ostreococcus tauri* version 2.0 (Palenik et al. 2007), and *Xenopus tropicalis* release 4.1, August 2005 (Bowes et al. 2008), were downloaded from the respective genome project websites. Additionally, 650 fully sequenced genomes of prokaryotes, including those of 46 cyanobacterial representatives, were downloaded from NCBI RefSeq database release November 2009 (Pruitt et al. 2007). To avoid clustering artifacts of distantly related eukaryotic and prokaryotic sequences, the sequences of cyanobacteria and photosynthetic eukaryotes were first clustered into separate sets of protein families. Matrices of algal/plant and cyanobacterial sequences were constructed from reciprocal best BLAST hits using an all-against-all BLAST, and thresholds of E-value $\leq 10^{-10}$ and amino acid sequence identities $\geq$25%. Clusters of homologous protein sequences were reconstructed from each of the matrices using MCL (Enright et al. 2002) Ver. 08-312, 1.008, with scheme = 7 and $I = 2.0$. Protein sequences of noncyanobacterial prokaryotes and nonphotosynthetic eukaryotes were added to the plant/algal clusters of proteins, depending on their sequence homologies using the above threshold, and a limit of three sequences per phylum. Overlapping plant/algal and cyanobacterial clusters were joined. The sequences of protein families were aligned using MAFFT (Katoh et al. 2002) Ver. 6.717b (2009/12/03). Multiple sequence alignment quality was assessed using the HoT-method (Landan and Graur 2007). Plant/algal protein sequences with Sum of Pairs Score <80% were excluded from the cluster. Phylogenetic trees were reconstructed using maximum likelihood approach with PhyML (Guindon et al. 2010) and the best-fit model as inferred with ProtTest (Abascal et al. 2005). The search for a best-fit model using ProtTest was restricted for nuclear gene substitution models including JTT (Jones et al. 1992) and WAG (Whelan and Goldman 2001) matrices. These were tested with all combinations of +I (estimation of invariant sites), +G (gamma distribution with 4 rate categories), and +F (using amino acid frequencies from the alignment) parameters. Branch lengths, model, and topology were optimized. From among 35,862 trees in total, WAG model was found as the best fit in 89% of the trees, with WAG + I + G as the more prevalent choice (34%). Genes of endosymbiotic origin in algal and plant genomes were inferred from the phylogenetic trees by searching for sisterhood between cyanobacterial protein sequences and their counterparts encoded by the nuclear genes of the photosynthetic eukaryotes (Martin et al. 2002). Protein families in the latter phototrophs were counted as having resulted from EGT(s), if

at least one of them had a cyanobacterial sequence as the nearest neighbor. Concatenated alignments were analyzed and used for tree construction by the same methods as described earlier.

Results

Genomes of Subsection V (Stigonematales) and *Scytonema*

The genome size distribution of the five Stigonematales strains (5.9 ± 2 Mb; fig. 1*D*) is similar to that of subsection IV members (Nostocales) (Larsson et al. 2011). With only 5,340 CDSs, *F. thermalis* PCC 7521 has the smallest genome among the subsection V members, whereas the genome of *S. hofmanni* PCC 7110 (subsection IV) has 12,356 predicted ORFs, making it the most gene-rich prokaryote sequenced to date (fig. 1*D*). Clustering of all 223,941 CDSs encoded in 51 cyanobacterial genomes by protein sequence similarity resulted in 18,185 cyanobacterial protein families and 47,174 singletons. Protein families with metabolic or cellular functions have significantly more duplicates in strains of subsection V than in those of subsection IV ($P < 2.2 \times 10^{-16}$, paired *t* test). Subsection V and IV strains do not differ in gene copy number for information processing protein families ($P = 0.11$, paired *t* test). The genome of strain PCC 7521 contains fewer duplicates ($P < 2.2 \times 10^{-16}$, paired *t* test) than the other two representatives of *Fischerella*. The frequency of genes shared with other filamentous cyanobacteria and the distribution of gene function are similar (fig. 1*E* and supplementary fig. S2, Supplementary Material online) among the three phenotypically similar *Fischerella* strains (Rippka et al. 1979).

Patterns of gene presence and absence might identify genes related to cyanobacterial morphological diversity (Stucken et al. 2010; Larsson et al. 2011). A subset of 22 protein families is unique and common to all filamentous cyanobacteria in our sample (supplementary table S1, Supplementary Material online), only few of which have known function. Subsection V members share 7 ± 1% of their proteome with those of subsection IV, and 73 protein families are specific to heterocyst-forming strains (supplementary table S2, Supplementary Material online). Most of the remaining subsections IV- and V-specific genes fall into cell wall, membrane, and envelope biogenesis COGs, such as glycosyltransferases, exopolysaccharide synthesis, and secretion. Some of the subsection V-specific protein families might be involved in the multiseriate filament phenotype and formation of true branches. On average, only 2% of the proteins encoded in subsection V genomes are specific to true-branching forms. Only 46 gene families are uniquely shared among subsection V genomes (supplementary table S3, Supplementary Material online). Although their functions are yet unknown, their classifications entail mostly cell wall, membrane, envelope biogenesis, and signal transduction functions. The relative paucity of proteins comprising the core set of the true branching cyanobacteria suggests that this phenotype hinges upon very few expressed proteins, which may mainly affect regulation of cell division genes and/or localization of their products.

Vertical and Lateral Components of Cyanobacterial Genome Evolution

To reconstruct a cyanobacterial backbone phylogeny, we identified all 324 single-copy protein families common to all 51 cyanobacteria in our sample and reconstructed their phylogenetic trees. The consensus tree (fig. 2), rooted with *Gloeobacter violaceus*, indicates a single origin for the filamentous architecture, and the concatenated alignment (564,408 sites) yielded an identical topology with NJ (supplementary fig. S3, Supplementary Material online), where all branches are supported by 100% bootstrap replicates. Maximum likelihood reconstruction yielded a phylogeny in which filamentous cyanobacteria are polyphyletic (supplementary fig. S3, Supplementary Material online), the difference to NJ being the position of *Microcoleus chthonoplastes* PCC 7420, a filamentous strain isolated from salt marshes (Rippka et al. 1979). Current whole-genome cyanobacterial phylogenies group *Microcoleus* with subsection I (Criscuolo and Gribaldo 2011), yielding paraphyly for filamentous forms. Although 55 of our 324 single copy gene trees support that position for *Microcoleus*, 111 recover filamentous monophyly, discrepancies that might reflect the workings of LGT (Raymond et al. 2002; Mulkidjanian et al. 2006; Shi and Falkowski 2008; Dufresne et al. 2008). To test the consistency of the backbone (consensus) phylogeny, we reconstructed a phylogenetic network using SplitsTree (Huson and Bryant 2006). The resulting network reveals a paucity of conflicting splits in the data (supplementary fig. S4, Supplementary Material online). A total of 92 out of 212 splits are compatible with the NJ tree topology and their sum of split weight amounts to 96% of the total network; and thus, the NJ tree explains most of the split variability in the data.

To estimate the degree and distribution of LGT in cyanobacterial evolution, we reconstructed a MLN, which infers LGT frequencies by allowing increasing amounts of LGT per protein family across a given backbone phylogeny (here the consensus tree), and identifying for all gene families the LGT frequency at which the distributions of modern genome sizes and inferred ancestral genome sizes agree best (Dagan et al. 2008). The MLN analysis conservatively assumes that all gene trees for all protein families are compatible (Dagan et al. 2008) and entails no gene tree comparisons. It revealed that 6,068 (34%) of the cyanobacterial protein families require no LGT to account for their gene distributions, whereas 12,116 (66%) protein families have undergone at least one LGT event. Because the method does not tally conflicting gene trees for

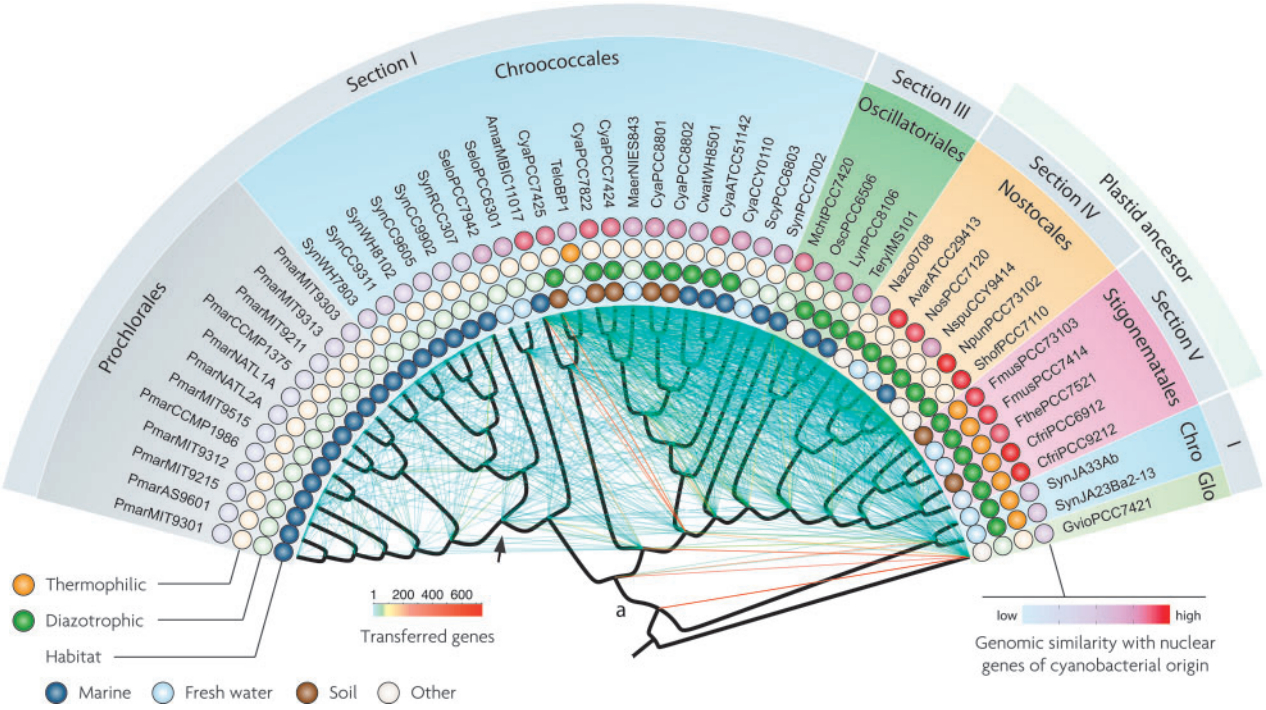


Fig. 2.—Vertical and lateral gene evolution in cyanobacterial genomes. NJ consensus (or backbone) tree, inferred from 324 single-copy protein families common to all 51 cyanobacteria in our sample, and rooted with *Gloeobacter violaceus* PCC 7421. Branches indicating vertical gene evolution are indicated in black. The MLN is indicated by edges that do not map onto the vertical component, with number of genes per edge indicated by a color gradient from cyan (1 gene) to orange (736 genes). The phylogenetic position of the eukaryotic clade reconstructed using 23 core genes is marked by "a." The SynPro clade is marked by an arrow.

homologous sequences, these are conservative lower bound estimates, in contrast to other recent studies (Raymond et al. 2002; Mulkidjanian et al. 2006; Shi and Falkowski 2008; Dufresne et al. 2008). Our estimate is found in agreement with earlier quantification of LGT frequency among cyanobacteria using an embedded quartets approach (Zhaxybayeva et al. 2006).

The MLN is presented in figure 2, and shows vertical components of cyanobacterial evolution and a network of 1,183 edges indicating laterally shared genes. Within the network, 358 edges (32%) represent a single laterally shared gene, whereas most edges (55%) carry $\leq$3 genes. Only 91 (7%) of edges carry >20 genes. Thus, bulk transfers of tens of genes or more are rare. The clade of marine *Prochlorococcus* and *Synechococcus* (SynPro) strains, which are recognized as being closely related environmental specialists of reduced genome size (Rocap et al. 2003; Dufresne et al. 2008), appear to have the lowest LGT frequency. The intertwined phylogenies within this clade (Zhaxybayeva et al. 2009) go undetected because the MLN is reconstructed from gene presence/absence data that are uninformative for the reconstruction of recombination events at the intra-species level (Dagan et al. 2008). The most highly connected nodes implicate the four contemporary strains *Acaryochloris marina* MBIC 11017, *Cyanothece* PCC 7425, *M. chthonoplastes* PCC 7420, and *S. hofmanni* PCC 7110 (fig. 2). Two of these strains, *A. marina*, an atypical marine unicellular cyanobacterium producing chlorophyll *d* as the primary photosynthetic pigment (Swingley et al. 2008), and *M. chthonoplastes*, a marine mat former, have the largest genomes (8.36 and 8.65 Mb, respectively) known for members of subsections I and III, and show an expansion of protein families (Larsson et al. 2011). The MLN pinpoints large genomes as harboring gene pools that are frequently transferred among cyanobacteria, and identifies subsection V strains as being more highly connected with strains of subsections IV and III (1.4 edges/node) than with unicellular strains (0.3 edges/node), also when strains of the SynPro clade are excluded (0.7 edges/node). This may suggest the existence of a LGT barrier between unicellular (mostly marine) and filamentous (mostly terrestrial) cyanobacteria.

The Nature of the Plastid Ancestor

To identify plant nuclear genes of cyanobacterial origin, we reconstructed 35,862 phylogenetic trees containing both eukaryotic and prokaryotic homologs and looked for trees in which plants and cyanobacteria branch together. In the present sample, considering all trees, between 8.7% and 11.5% of all nuclear genes in photosynthetic eukaryotes sampled branch with cyanobacterial homologs (table 1).

Table 1
Proportion of Plant Genes of Endosymbiotic Origin

	No. Proteins	Total Tree Set			CS ≥ 80%		≤ 3 homologues	
		No. Trees	No. Putative EGT	EGT Bootstrap Support	No. Trees	No. Putative EGT	No. Protein Families	No. Putative EGT
Arabidopsis thaliana	30,897	9,025	801 (8.9%)	87.89 ± 20.10	3,306	424 (12.8%)	2,091	136 (6.5%)
Oryza sativa	26,712	7,292	637 (8.7%)	84.82 ± 21.41	2,596	347 (13.4%)	1,623	95 (5.9%)
Physcomitrella patens	35,468	8,847	903 (10.2%)	84.74 ± 22.11	3,425	542 (15.8%)	1,402	78 (5.6%)
Ostreococcus tauri	7,715	3,495	403 (11.5%)	84.64 ± 21.20	1,232	247 (20.1%)	324	26 (8.0%)
Cyanidioschyzon merolae	4,761	2,688	307 (11.4%)	83.92 ± 20.05	844	167 (19.8%)	223	15 (6.7%)
Chlamydomonas reinhardtii	14,262	4,515	478 (10.6%)	83.81 ± 21.04	1,646	283 (17.2%)	599	41 (6.8%)
Total	119,815	35,862	3,529 (9.8%)	84.97 ± 20.98	13,049	2,010 (15.4%)	6,262	391 (6.2%)

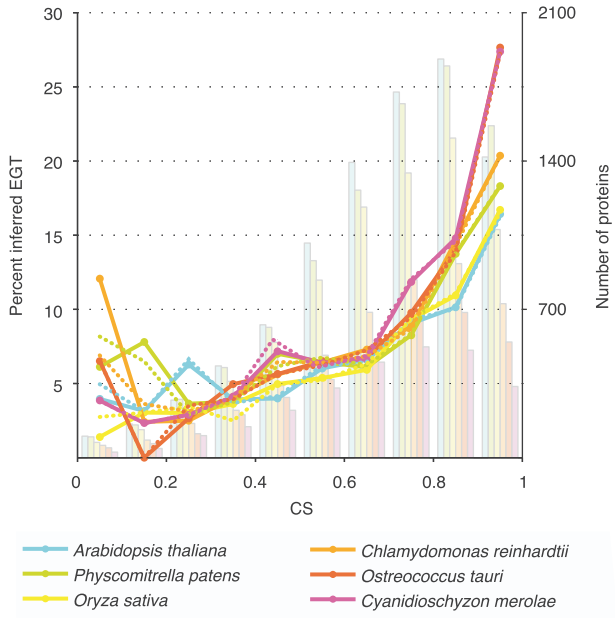


Fig. 3.—Phylogenetic characteristics of EGT inference. The frequency of EGT as inferred from alignments of varying reliability degrees. The distribution of alignment reliability as estimated by column score (CS) is presented in bars, colored according to the respective eukaryotes. The CS measure is calculated as the proportion of alignment sites whose reconstruction is independent upon the direction upon which the sequences are fed to the alignment algorithm (Landan and Graur 2007). The frequency of genes inferred as EGT is plotted above in the eukaryote-dependent strongly colored lines, with the proportions inferred from trees reconstructed by maximum likelihood and NJ approaches in solid and dashed lines, respectively.

For the most reliable alignments, where false negatives are less likely, the proportion of genes acquired from plastids ranges between 16% of the genes in *Arabidopsis* genome and >20% of the genes in the smaller genomes of *Ostreococcus* and *Cyanidioschyzon* (fig. 3), with energy metabolism and carbohydrate metabolism (99 genes) being the most frequent functional categories (supplementary fig. S5, Supplementary Material online). Clearly, the quantitative contribution of cyanobacteria to plant genomes was great, and the backbone of plant metabolism was acquired from them—plants are, biochemically, cyanobacteria wrapped in a bigger box.

To trace the nature of the plastid ancestor, we first assembled a dataset of 23 nuclear genes of plastid origin present in all plant and cyanobacterial genomes sampled. The tree of concatenated alignments, rooted by *G. violaceus* PCC 7421, shows a deep branch placing plastids basal among cyanobacteria (designated with an "a" in fig. 2). Expanding the data set to include 200 universal cyanobacterial gene families with a single, composite plant OTU (genes acquired from cyanobacteria and present in at least one plant) yielded the same long, deep branch. Long basal branches are characteristic of long-branch attraction (LBA), a well-known phylogenetic artifact. Compositional heterogeneity such as AT bias and heterotachy can cause LBA (Lockhart et al. 2006), and a basal position due to an LBA often involves the grouping of strains in which strain-specific character states are abundant (Stiller and Hall 1999). The sequences of the 23 universally distributed proteins in the six photosynthetic eukaryotes were found to contain significantly more unique substitutions than their cyanobacterial homologues ($P = 7 \times 10^{-66}$, one-tailed Kolmogorov–Smirnov test, supplementary fig. S6, Supplementary Material online), and an examination of the larger set of 200 phylogenetic trees reconstructed for genes of endosymbiotic origin shows that the eukaryotic clade branch length is on average 10-fold larger than that of the cyanobacterial branches. The basal position of plastids among cyanobacteria in the concatenated alignment tree (fig. 2 and supplementary fig. S6, Supplementary Material online) is attributable to LBA. Worse, given that LGT is frequent among cyanobacteria (Raymond et al. 2002; Mulkidjanian et al. 2006; Shi and Falkowski 2008; Dufresne et al. 2008), there is no reason to suspect that any "core" gene phylogeny will be a faithful proxy for the rest of the genome (Doolittle and Bapteste 2007).

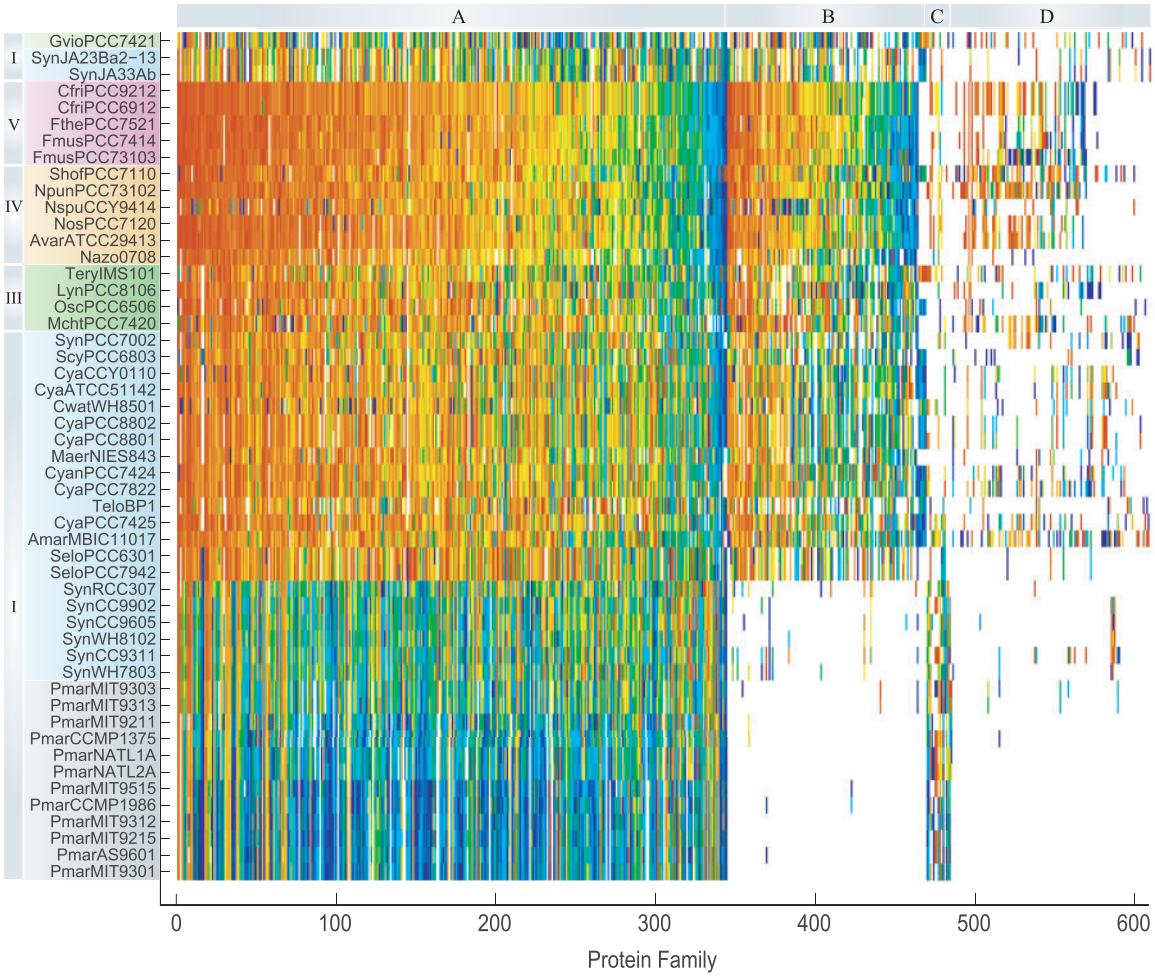


Fig. 4.—Presence/absence and sequence similarity patterns of cyanobacterial protein families by comparison with their homologs of endosymbiotic origin in six photosynthetic eukaryotes. Amino acid sequence similarity between the cyanobacterial proteins (*x* axis) and their counterparts in the eukaryotic plastid-derived set of protein families (*y* axis), as deduced for the genomes in the data set. Cell shades in the matrix correspond to the similarity ranking for each protein family (i.e., line) according to a color gradient from red (high similarity) to blue (low similarity). White cells correspond to genes lacking in the respective genomes. Protein families are ordered according to their distribution pattern into (*A*) nearly universal, (*B*) sparse representation or (*C*) highly frequent in the oceanic species, and (*D*) generally sparse representation. Cyanobacterial strains are ordered according to the MLN in fig. 2.

Therefore, we turned our attention to the larger set of nuclear genes of cyanobacterial origin whose homologs are not universally distributed among cyanobacteria. For 611 plant nuclear gene families identified as plastid acquisitions, we scored gene presence and absence, and protein sequence identity among cyanobacterial genomes (fig. 4). The SynPro clade lacks a substantial portion of these plastid ancestor gene families. A total of 245 (40%) protein families possessed by plants are absent in all *Prochlorococcus* strains, 137 (22%) are absent in all *Synechococcus* strains (fig. 4). The similarity map also shows that overall protein sequence similarity of plant nuclear genes is highest to homologs in members of subsection IV and V. For 225 (37%) protein families, the average amino acid identity between the cyanobacterial genes and their plant homologs is significantly higher for subsection V genomes ($\alpha = 0.05$, Kolmogorov–Smirnov test and FDR) than for subsection I genomes. When subsection IV and V genomes are combined and compared with those of subsection I, the value increases to 270 (44%) ($\alpha = 0.05$, Kolmogorov-Smirnov test and FDR). Thus, subsection IV and V genomes harbor more homologs of genes that plants acquired from cyanobacteria and those have higher sequence similarity to their plant homologs than genomes of subsection I. Similar amino acid usage in different organisms may sometimes lead to an overestimation of species relatedness (Rodríguez-Ezpeleta and Embley 2012). Here, we tested for such possible bias using a principle component analysis (PCA) for the amino acid frequencies encoded by the 611 genes of endosymbiotic origin. The transformation of amino acid usage into two principal components explains in total 89% of the variability observed

(supplementary fig. S7, Supplementary Material online). Furthermore, the PCA reveals that the eukaryotic species do not group with the filamentous cyanobacteria; hence, the protein sequence similarity observed between those two groups is not a result of biased amino acid usage. Consequently, we can conclude that in the present sample, the collection of genes possessed by the ancestor of plastids was most similar to that in filamentous, heterocyst-forming cyanobacteria (fig. 2).

Discussion

Possible Initial Benefits of Plastids

Today plastids supply fixed carbon to plant cells, but they also have a myriad of other functions in amino acid, lipid, and cofactor biosynthesis as well as nitrogen metabolism. What was the biochemical or physiological context of the symbiosis that gave rise to plastids—what initially associated the founder endosymbiont to its host in the first place? Traditional reasoning on the selective advantage that was crucial to the establishment of the plastid has it that the production of carbohydrates by the cyanobacterial endosymbiont was the key, a view that was clearly expressed by Mereschkowsky (1905, p. 605) in his initial formulation of endosymbiotic theory: "Plant cells receive with no effort whatsoever large amounts of preformed organic substrates (carbohydrates), which their chromatophores willingly supply."

An alternative suggestion is that the initial advantage of plastids may have simply been their uniquely useful metabolic end product, O_2, as a boost to respiration in early mitochondria (Martin and Müller 1998). The chemical benefit of O_2 could, of course, have only been of value if the initial endosymbiosis had taken place at a time in Earth's history, or in an environment, where O_2 was not freely available in sufficient amounts. Fossil evidence supports the notion that the primary plastid endosymbiosis occurred at least 1.2 billion years ago (Butterfield 2000) and molecular estimates suggest that plastids might have arisen by approximately 1.5 billion years ago (Parfrey et al. 2011). Geochemists have found over the last decade that an approximately 2 billion year span of protracted ocean anoxia ended only about 580 Ma (Anbar and Knoll 2002; Johnston et al. 2009; Lyons et al. 2009; Lyons and Reinhardt 2009; Sahoo et al. 2012). The six major eukaryotic assemblages or "supergroups" currently recognized, including plants, arose and diversified during that time (Parfrey et al. 2011), that is, while the oceans were still anoxic (Müller et al. 2012). Such geological context (ocean anoxia during most of the Proterozoic) would be compatible with a possible role for O_2 as an initial benefit in the plastid evolution. Indeed, for Stanier (1970), the production of O_2 was a reason to suggest that plastids arose before mitochondria did. Of course, Proterozoic ocean anoxia was likely less pronounced in the photic zone than below it (Johnston et al. 2009). A freshwater origin of plastids is also a possibility to consider, whereby the present data linking plastids phylogenomically more closely with freshwater cyanobacteria than with marine forms (fig. 2) would be compatible with that view.

Another suggestion is that the key to establishment of the plastid was the origin of carbon translocators in the plastid inner membrane and that the incorporation of a metabolite antiporter like the triose phosphate translocator in the ancestral plastid membrane was the essential step for establishing the primary endosymbiosis by allowing the plant ancestor to profit from cyanobacterial carbon fixation (Weber et al. 2006). In the same vein, it was furthermore argued that the key to establishment of the plastid entailed the insertion of additional host-controlled metabolite exchange proteins into plastid membranes fulfilling a similar export role (Gross and Bhattacharya 2009). A problem with theories that focus on carbon exporters as the key innovation at plastid origin is that cyanobacteria are well known to produce copious amounts of exopolysaccharides (De Philippis and Vincenzini 1998), such that there would be no need to evolve or insert transporters for provision of carbohydrates to be realized by the host.

The theory for the initial benefit of plastids that is currently best founded in direct observation, we would argue, is that nitrogen fixation was a key to the establishment of the symbiosis (Kneip et al. 2007). This view is supported by the circumstance that in modern symbioses involving cyanobacteria, nitrogen (not reduced carbon) is usually the key nutrient underlying the success of the partnership (Rai et al. 2000; Raven 2002). Accordingly, the cyanobacterial endosymbionts are nitrogen fixing forms and combined nitrogen (ammonium) is the nutrient provided by the cyanobacterium. This is true for diatoms with N_2-fixing cyanobacterial endosymbionts (Prechtl et al. 2004; Kneip et al. 2008), prymnesiophytes with associated N-fixing cyanobacteria that might be ectosymbionts (Thompson et al. 2012), cyanobionts in lichens (Rikkinen et al. 2002), coralloid roots of cycads (Costa et al. 2004), the angiosperm *Gunnera* (Chiu et al. 2005), and the water-fern *Azolla* (Ran et al. 2010). In the case of *Azolla* and *Rhopalodia*, the N_2-fixing cyanobacteria live as intracellular endosymbionts (Kneip et al. 2008; Ran et al. 2010).

Recent studies have suggested that a filamentous phenotype and heterocyst differentiation may have been hallmark phenotypic characteristics of the plastid ancestor (Deusch et al. 2008; Ran et al. 2010; Larsson et al. 2011). Indeed, in modern cyanobacterial symbioses, fixed nitrogen is the main currency of benefit that the cyanobacterial symbiont provides to its host (Kayley et al. 2007). The early physiological association of the plastid ancestors with their host might thus have been similar to that of the unicellular nitrogen-fixing endosymbiont and its diatom host *Rhopalodia* (Kneip et al. 2008), or the highly reduced *Nostoc azollae*, an obligate cyanobiont of water-ferns, whose genome has drastically been reduced,

with a large portion of the remaining genes specifically dedicated to heterocyst differentiation and nitrogen fixation (Ran et al. 2010). A potential problem with this view is that nitrogen fixation has not been retained by any modern plant (Allen and Raven 1996). Why not? One possible reason concerns the circumstance that cyanobacterial O_2 production led to an oxidation state of the environment in which nitrate became abundant (Falkowski et al. 2008)—in a world of abundant nitrate, nitrogenase is less necessary, hence less likely to be retained, although one should recall that modern cyanobacterial symbionts do fix nitrogen for their hosts. Perhaps, more importantly, in oxic environments cyanobacteria that express nitrogenase must exhibit either temporal separation of photosynthesis and nitrogen fixation (N_2-fixation occurring mainly in the dark; Mitsui et al. 1986), or other means of protecting the notoriously O_2-sensitive enzyme from inactivation such as diazocyte differentiation in *Trichodesmium* (Sandh et al. 2012), or heterocyst formation in subsections IV and V (Kumar et al. 2010). It is possible that such nitrogenase-protecting strategies, whereas readily accessible to genetically autonomous prokaryotes, are not among the realm of possibilities that plastids, which relinquished most of their genetic autonomy, can developmentally attain.

Many Endosymbionts, or Only One with Many Genes?

Gene transfer following plastid origin readily explains plant nuclear genes that branch with cyanobacteria. However, many plant-specific genes branch with other prokaryotes (fig. 5*A*). Plant genes that branch with chlamydial homologs have led to the inferences that a chlamydial endosymbiont accompanied the origin of plastids (Brinkman et al. 2002; Huang and Gogarten 2007; Price et al. 2012). This theory postulates that the plant ancestor consumed cyanobacteria as food and was parasitized by environmental chlamydias (Huang and Gogarten 2007; Moustafa et al. 2008), whereby the chlamydias were key to establishing the plastid because chlamydia-like bacteria donated genes that allowed export of photosynthate from the cyanobacterial plastid ancestor and its polymerization into storage polysaccharide in the cytosol (Price et al. 2012). The flaw with this theory is that it is based on the uncritical interpretation of computational results of genome comparisons that, as has long been known (Rujan and Martin 2001; Martin et al. 2002; Esser et al. 2007; Dagan et al. 2008), would implicate many other groups of prokaryotes far more strongly than they would implicate chlamydias as active bystanders at the origin of plastids. The focus on chlamydia as opposed to, say spirochaetes or proteobacteria, is arbitrary and to some extent ad hoc. If one were to take the chlamydia theory seriously, or think it through in full, the transiently symbiotic and gene-dealing "chlamydioplast" would have to take a number and wait in line next to the actinobacterioplast, the clostridioplast, the bacilloplast, the bacteriodetoplast, and the spirochaetoplast, and so forth. (fig. 5*A*). Beyond the

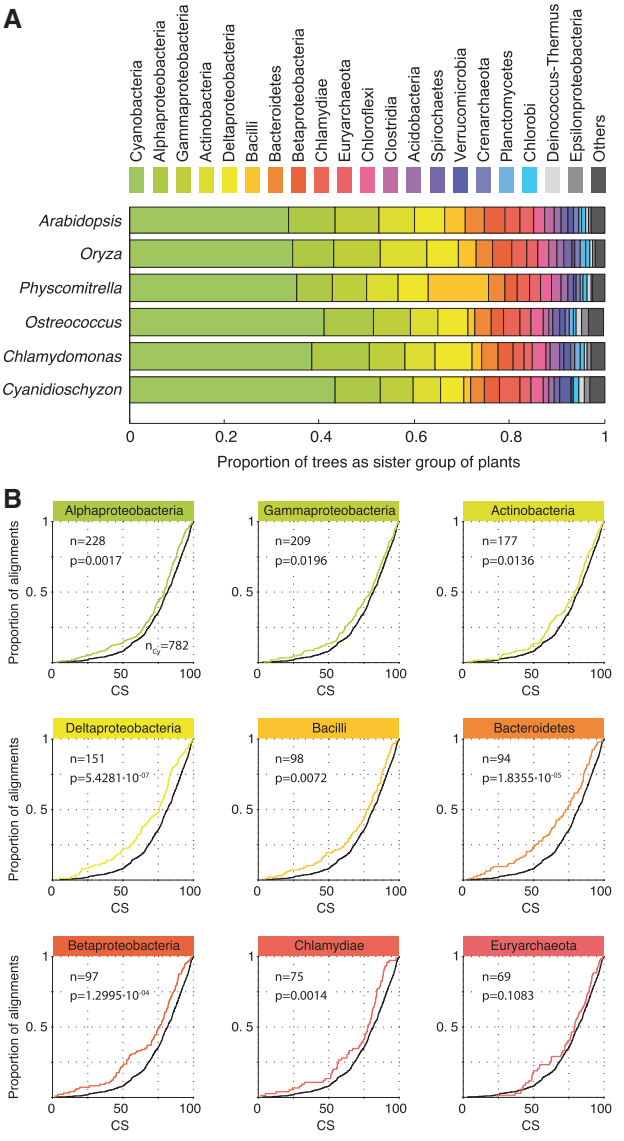


Fig. 5.—Taxon distribution of nearest neighbors to plant genes. (*A*) Tree samples distribute as following: *Arabidopsis*: 2,324; *Oryza*: 1,792; *Physcomitrella*: 2,511; *Ostreococcus*: 968; *Chlamydomonas*: 1,218; and *Cyanidioschyzon*: 693. Microbial taxonomic groups having a low frequency of nearest neighbors were grouped into the "Others" bar. Those include Aquificae, Dictyoglomi, Elusimicrobia, Fibrobacteres, Fusobacteria, Gemmatimonadetes, Korarchaeota, Nanoarchaeota, Nitrospirae, Tenericutes, Thaumarchaeota, and Thermotogae. (*B*) A comparison of alignment quality (CS) between trees of *Arabidopsis* genes having a cyanobacterial nearest neighbor (black) and trees where a nearest neighbor from a different prokaryotic group was inferred (colored according to the taxa). In all groups but the Euryarchaeota, the alignment quality of trees where a noncyanobacterial nearest neighbor was inferred is significantly lower in comparison with tree topologies having cyanobacteria as their nearest neighbor (using Wilcoxon test, $\alpha = 0.05$). These results suggest that the inference of noncyanobacterial nearest neighbors to plant genes is less reliable than the inference of cyanobacterial nearest neighbors.

cyanobacterial signal, which corresponds to a tangible double membrane-bounded and DNA-containing organelle, the other putative phylogenetic signals in the data, especially that involving chlamydia, are better explained in terms of known phenomena, such as LGT among free-living prokaryotes (Dagan et al. 2008) and by phylogeny reconstruction errors (White et al. 2007; Stiller 2011) (fig. 5*B*), both of which we know to really exist, than in terms of gene dealing endosymbionts whose existence is inferred from a few gene trees. The null hypothesis for endosymbiotic theory in the age of genomes should be: The ancestors of plastids underwent LGT, just like modern cyanobacteria, whose genomes are chimeras of genes from many sources (Mulkidjanian et al. 2006), and the plastid ancestor genome was probably no different (Richards and Archibald 2011). LGT among prokaryotes accounts for the diverse sequence affinities of genes acquired from the single ancestor of plastids with far fewer corollaries than a one-symbiont-per-gene theory. We merely need to incorporate the effect that LGT among prokaryotes will have over geological time on the endosymbiotic origins of organelles.

Clues to the Origin of Two Photosystems

One notable aspect of cyanobacterial phylogenomics presented in this study is that the marine cyanobacteria are not basal in the trees (fig. 2 and supplementary fig. S3, Supplementary Material online). These small unicellular cyanobacteria (diameter 1 μm or less) share reduced genome sizes (<3 Mb) as a common trait, and seem to have arisen from ancestors with larger genomes (Larsson et al. 2011) that, inferred from the phylogeny, lived in terrestrial, brackish, or perhaps freshwater environments (Sánchez-Baracaldo et al. 2005). This led Blank and Sánchez-Baracaldo (2010) to suggest that oxygenic photosynthesis arose in a freshwater environment. Our results support that view, and this conclusion has implications for the origin of water-splitting photosynthesis. Among many possibilities (Xiong and Bauer 2002; Hohmann-Marriot and Blankenship 2011; Williamson et al. 2011), it has been suggested that the progenitor of the cyanobacteria had genes for both type I (RCI) and type II (RCII) photosynthetic reaction centers (via gene duplication) but expressed either set of genes depending on the reducing conditions in the environment (Allen 2005): type RCI in the presence of H_2S for noncyclic electron flow, as in *Chlorobium* (or the facultative anaerobic cyanobacterium *Oscillatoria limnetica*); and type RCII in the absence of H_2S, for cyclic electron flow, as in *Rhodobacter* (Allen 2005). Were regulation to fail such that both type I and type II reaction centers became expressed in the absence of H_2S, the protocyanobacterium would oxidatively perish, unless it could extract electrons from an environmentally available donor.

Such an electron donor could have been aqueous $Mn^{II/III}$, which has the utilitous property of being photo-oxidized by ultraviolet light (Allen and Martin 2007), an abundant component of solar radiation incident on the Earth's surface prior to accumulation of atmospheric oxygen. Attaining suitably high concentrations of $Mn^{II/III}$ as an environmentally available electron donor in the ocean would be problematic, but not in a freshwater setting. Allen et al. (2012) have recently shown that an engineered, Mn-binding type II reaction center of *Rhodobacter sphaeroides* will produce O_2 from O_2^- in the presence of Mn in a light-dependent reaction in which photo-damage is impeded in comparison with that in a wild-type, Mn-free reaction center. Their observation (Allen et al. 2012) is likely an important clue to the origin of oxygenic photosynthesis, at which time a protocyanobacterial type II reaction center acquired, via natural selection, the ability to (photo-)oxidize $Mn^{II/III}$—itself ultimately rereduced by water—and then to reduce a newly constitutive type I reaction center. Transition from environmental (substrate) $Mn^{II/III}$ ions to the catalytic Mn_4Ca center of cyanobacterial RCII would then have permitted light-dependent CO_2 and/or nitrogen fixation, in the absence of electron donors other than water.

What Makes a Branching Cyanobacterium?

The morphological diversity of cyanobacteria poses an intriguing question in the biology and evolution of cell differentiation. Transposon mutagenesis of *Synechococcus elongatus* PCC 7942 (subsection I) revealed that the loss of several genes involved in cell division leads to filament formation (Miyagishima et al. 2005). However, our analysis revealed that all recognized cyanobacterial cell division genes are present in the genomes of filamentous cyanobacteria, including those of subsection V. This suggests that the filamentous phenotype in cyanobacteria of subsections III, IV, and V is not due to loss of genes for cell division, though it is currently unknown whether those that are present are all expressed. Genes common to both unicellular and filamentous cyanobacteria may also be important for determining trichome structure in members of subsections III–V. This is suggested by a recent study on the filamentous heterocystous strain *N. punctiforme* ATCC 29133 (Lehner et al. 2011), which showed that mutations of the *amiC2* gene, encoding an amidase involved in septa formation, will lead to a morphology similar to that of colonial unicellular cyanobacteria, and prevent heterocyst differentiation. Furthermore, filament formation in *S. elongatus* PCC 7942 can be induced by over-expression of the gene encoding FtsZ, which is known as a cell division protein (Mori and Johnson 2001). Thus, the lack of clear candidate genes whose distribution across cyanobacterial genomes correlate with cellular morphology and the experimental evidence that links between the expression level (rather than presence/absence) of cell division proteins and filament formation suggest that a filamentous

phenotype may result from modifications of the gene regulatory network and cell division program.

Supplementary Material

Supplementary figures S1–S7 and tables S1–S3 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The work in the authors' laboratories is supported by SFB-TR1 to T.D. and W.F.M., the European Research Council (grant no. 232975 to W.F.M.; grant no. 281357 to T.D.), and a Leverhulme Trust Research Grant (no. F07 476AQ to J.F.A.). The support by the Institut Pasteur and the Centre National de la Recherche Scientifique (URA 2172) is acknowledged by M.G., R.R., and N.T.M. The authors are grateful to T. Coursin and T. Laurent for technical assistance in maintaining the Pasteur Culture Collection of Cyanobacteria at the Institut Pasteur. Additional computational support and infrastructure was provided by "Zentrum fuer Informations- und Medientechnologie" (ZIM) at Heinrich-Heine-University, Duesseldorf, Germany.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics 21:2104–2105.

Allen JF. 2005. A redox switch hypothesis for the origin of two light reactions in photosynthesis. FEBS Lett. 579:963–938.

Allen JF, Martin W. 2007. Evolutionary biology: out of thin air. Nature 445: 610–612.

Allen JF, Raven JA. 1996. Free-radical-induced mutation vs redox regulation: costs and benefits of genes in organelles. J Mol Evol. 42:482–492.

Allen JP, et al. 2012. Light-driven oxygen production from superoxide by Mn-binding bacterial reaction centers. Proc Natl Acad Sci U S A. 109: 2314–2318.

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 35: 3389–3342.

Anbar AD, Knoll AH. 2002. Proterozoic ocean chemistry and evolution: a bioinorganic bridge. Science 297:1137–1142.

Badger JH, Olsen GJ. 1999. CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol. 16:512–524.

Baymann F, Brugna M, Mühlenhoff U, Nitschke W. 2001. Daddy, where did (PS)I come from? Biochim Biophys Acta. 1507:291–310.

Bekker A, et al. 2004. Dating the rise of atmospheric oxygen. Nature 427: 117–120.

Blank CE, Sánchez-Baracaldo P. 2010. Timing of morphological and ecological innovations in the cyanobacteria—a key to understanding the rise in atmospheric oxygen. Geobiology 8:1–23.

Bowes JB, et al. 2008. Xenbase: a *Xenopus* biology and genomics resource. Nucleic Acids Res. 36:D761–D767.

Brinkman FS, et al. 2002. Evidence that plant-like genes in Chlamydia species reflect an ancestral relationship between Chlamydiaceae, cyanobacteria, and the chloroplast. Genome Res. 12:1159–1167.

Butterfield NJ. 2000. *Bangiomorpha pubescens* n. gen., n. sp.: implications for the evolution of sex, multicellularity, and the mesoproterozoic/neoproterozoic radiation of eukaryotes. Paleobiology 263:386–404.

Chiu WL, et al. 2005. Nitrogen deprivation stimulates symbiotic gland development in *Gunnera manicata*. Plant Physiol. 139:224–230.

Costa JL, Romero EM, Lindblad P. 2004. Sequence based data supports a single *Nostoc* strain in individual coralloid roots of cycads. FEMS Microbiol Ecol. 49:481–487.

Criscuolo A, Gribaldo S. 2011. Large-scale phylogenomic analyses indicate a deep origin of primary plastids within cyanobacteria. Mol Biol Evol. 28:3019–3032.

Dagan T, Artzy-Randrup Y, Martin W. 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc Natl Acad Sci U S A. 105:10039–10044.

De Philippis R, Vincenzini M. 1998. Exocellular polysaccharides from cyanobacteria and their possible applications. FEMS Microbiol Rev. 22:151–175.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27: 4636–4641.

Deschamps P, et al. 2008. Metabolic symbiosis and the birth of the plant kingdom. Mol Biol Evol. 25:536–548.

Deusch O, et al. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. Mol Biol Evol. 25:748–761.

Doolittle WF. 1980. Revolutionary concepts in evolutionary biology. Trends Biochem Sci. 5:146–149.

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

Dufresne A, et al. 2008. Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. Genome Biol. 9:R90.

Enright AJ, van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 30: 1575–1584.

Esser C, Martin W, Dagan T. 2007. The origin of mitochondria in light of a fluid prokaryotic chromosome model. Biol Lett. 3:180–184.

Falkowski PG, Fenchel T, Delong EF. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science 320:1034–1039.

Felsenstein J. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Seattle (WA): University of Washington.

Franche C, Damerval T. 1988. Test on nif probes and DNA hybridizations. Methods Enzymol. 167:803–808.

Gordon D, Abajian C, Green P. 1998. Consed: a graphical tool for sequence finishing. Genome Res. 8:195–202.

Gould SB, Waller RF, McFadden GI. 2008. Plastid evolution. Annu Rev Plant Biol. 59:491–517.

Gross J, Bhattacharya D. 2009. Opinion: Mitochondrial and plastid evolution in eukaryotes: an outsiders' perspective. Nat Rev Genet. 10:495–505.

Gross J, Meurer J, Bhattacharya D. 2008. Evidence of a chimeric genome in the cyanobacterial ancestor of plastids. BMC Evol Biol. 8:117.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 59:307–321.

Hohmann-Marriott MF, Blankenship RE. 2011. Evolution of photosynthesis. Annu Rev Plant Biol. 62:515–548.

Huang J, Gogarten JP. 2007. Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? Genome Biol. 8:R99.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Johnston DT, Wolfe-Simon F, Pearson A, Knoll AH. 2009. Anoxygenic photosynthesis modulated Proterozoic oxygen and sustained Earth's middle age. Proc Natl Acad Sci U S A. 106:16925–16929.

Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation rate matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.

Kayley MU, Bergman B, Raven JA. 2007. Exploring cyanobacterial mutualism. Annu Rev Ecol Evol Syst. 38:255–273.

Kneip C, Lockhart P, Voss C, Maier UG. 2007. Nitrogen fixation in eukaryotes—new models for symbiosis. BMC Evol Biol. 7:55.
Kneip C, Voss C, Lockhart PJ, Maier UG. 2008. The cyanobacterial endosymbiont of the unicellular algae *Rhopalodia gibba* shows reductive genome evolution. BMC Evol Biol. 8:30.
Krause L, et al. 2007. GISMO-gene identification using a support vector machine for ORF identification. Nucleic Acids Res. 35:540–549.
Kumar K, Mella-Herrera RA, Golden JW. 2010. Cyanobacterial heterocysts. Cold Spring Harb Perspect Biol. 2:a000315.
Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.
Larsson J, Nylander JA, Bergman B. 2011. Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. BMC Evol Biol. 11:187.
Lehner J, et al. 2011. The morphogene AmiC2 is pivotal for multicellular development in the cyanobacterium *Nostoc punctiforme*. Mol Microbiol. 79:1655–1669.
Lindell D, et al. 2004. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Natl Acad Sci U S A. 101: 11013–11018.
Lockhart P, et al. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol Biol Evol. 23:40–45.
Lyons TW, Anbar AD, Severmann S, Scott C, Gill BC. 2009. Tracking euxinia in the ancient ocean: a multiproxy perspective and Proterozoic case study. Annu Rev Earth Planet Sci. 37:507–534.
Lyons TW, Reinhard CT. 2009. An early productive ocean unfit for aerobics. Proc Natl Acad Sci U S A. 106:18045–18046.
Martin W, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature 392:37–41.
Martin W, et al. 2002. Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proc Natl Acad Sci U S A. 99: 12246–12251.
Matsuzaki M, et al. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature 428:653–657.
Mereschkowsky C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol Centralbl. 25:593–604. [English translation in Eur J Phycol. 1999;34:287–295.]
Meyer F, et al. 2003. GenDB—an open source genome annotation system for prokaryote genomes. Nucleic Acids Res. 31:2187–2195.
Mitsui AS, et al. 1986. Strategy by which nitrogen-fixing unicellular cyanobacteria grow photoautotrophically. Nature 323:720–722.
Miyagishima SY, Wolk CP, Osteryoung KW. 2005. Identification of cyanobacterial cell division genes by comparative and mutational analyses. Mol Microbiol. 56:126–143.
Moisander PH, et al. 2010. Unicellular cyanobacterial distributions broaden the oceanic N2 fixation domain. Science 327:1512–1524.
Morel FM, Price NM. 2003. The biogeochemical cycles of trace metals in the oceans. Science 300:944–947.
Mori T, Johnson CH. 2001. Independence of circadian timing from cell division in cyanobacteria. J Bacteriol 183:2439–2444.
Moustafa A, Reyes-Prieto A, Bhattacharya D. 2008. Chlamydiae has contributed at least 55 genes to plantae with predominantly plastid functions. PLoS One 3:e2205.
Mulkidjanian AY, et al. 2006. The cyanobacterial genome core and the origin of photosynthesis. Proc Natl Acad Sci U S A. 103: 13126–13131.
Müller M, et al. 2012. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. Microbiol Mol Biol Rev. 76:444–495.
Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.
Palenik B, et al. 2007. The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. Proc Natl Acad Sci U S A. 104:7705–7710.
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.
Prechtl J, Kneip C, Lockhart P, Wenderoth K, Maier UG. 2004. Intracellular spheroid bodies of *Rhopalodia gibba* have nitrogen-fixing apparatus of cyanobacterial origin. Mol Biol Evol. 21:1477–1481.
Price DC, et al. 2012. *Cyanophora paradoxa* genome elucidates origin of photosynthesis in algae and plants. Science 335:843–847.
Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 35:D61–D65.
Rai AN, Söderbäck E, Bergman B. 2000. Cyanobacterium-plant symbioses. New Phytol. 147:449–481.
Ran L, et al. 2010. Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. PLoS One 5:e11486.
Raven JA. 2002. Evolution of cyanobacterial symbioses. In: Rai AN, Bergman B, Rasmussen U, editors. Cyanobacteria in symbiosis. Dordrecht (The Netherlands): Kluwer Academic Publishers. p. 326–246.
Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE. 2002. Whole-genome analysis of photosynthetic prokaryotes. Science 298:1616–1620.
Reyes-Prieto A, et al. 2010. Differential gene retention in plastids of common recent origin. Mol Biol Evol. 27:1530–1537.
Richards TA, Archibald JM. 2011. Cell evolution: gene transfer agents and the origin of mitochondria. Curr Biol. 21:R112.
Richly E, Leister D. 2004. An improved prediction of chloroplast proteins reveals diversities and commonalities in the chloroplast proteomes of *Arabidopsis* and rice. Gene 329:11–16.
Rikkinen J, Oksanen I, Lohtander K. 2002. Lichen guilds share related cyanobacterial endosymbionts. Science 297:357.
Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. 1979. Generic assignments, strain histories and properties of pure cultures of cyanobacteria. J Gen Microbiol. 111:1–61.
Rippka R, Herdman H. 2002. Pasteur culture collection of Cyanobacteria: catalogue and taxonomic handbook. I. Catalogue of strains. Paris: Institut Pasteur.
Rocap G, et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424:1042–1047.
Rodríguez-Ezpeleta N, Embley TM. 2012. The SAR11 group of alphaproteobacteria is not related to the origin of mitochondria. PLoS One 7:e30520.
Rujan T, Martin W. 2001. How many genes in *Arabidopsis* come from cyanobacteria? An estimate from 386 protein phylogenies. Trends Genet. 17:113–120.
Sahoo SK, et al. 2012. Ocean oxygenation in the wake of the Marinoan glaciation. Nature 489:546–549.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.
Sánchez-Baracaldo P, Hayes PK, Blank CE. 2005. Morphological and habitat evolution in the Cyanobacteria using a compartmentalization approach. Geobiology 3:145–165.
Sandh G, Xu Linghua, Bergman B. 2012. Diazocyte development in the marine diazotrophic cyanobacterium *Trichodesmium*. Microbiology 158:345–352.
Sharma AD, Gill PK, Singh P. 2002. DNA isolation from dry and fresh samples of polysaccharide-rich plants. Plant Mol Biol Rep. 20: 415a–415f.
Sharon I, et al. 2009. Photosystem I gene cassettes are present in marine virus genomes. Nature 461:258–262.
Shi T, Falkowski PG. 2008. Genome evolution in cyanobacteria: the stable core and the variable shell. Proc Natl Acad Sci U S A. 105:2510–2515.
Stanier RY. 1970. Some aspects of the biology of cells and their possible evolutionary significance. Symp Soc Gen Microbiol. 20:1–38.

Stiller JW. 2011. Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer. BMC Evol Biol. 11:259.
Stiller JW, Hall BD. 1999. Long-branch attraction and the rDNA model of early eukaryotic evolution. Mol Biol Evol. 16:1270–1279.
Stucken K, et al. 2010. The smallest known genomes of multicellular and toxic cyanobacteria: comparison, minimal gene sets for linked traits and the evolutionary implications. PLoS One 5:e9235.
Swingley WD, et al. 2008. Niche adaptation and genome expansion in the chlorophyll *d*-producing cyanobacterium *Acaryochloris marina*. Proc Natl Acad Sci U S A. 105:2005–2010.
Tatusov RL, et al. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. 29:22–28.
Thompson AW, et al. 2012. Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. Science 337:1546–1550.
van Mooy BA, et al. 2009. Phytoplankton in the ocean use non-phosphorus lipids in response to phosphorus scarcity. Nature 458: 69–72.
Weber AP, Linka M, Bhattacharya D. 2006. Single, ancient origin of a plastid metabolite translocator family in Plantae from an endomembrane-derived ancestor. Eukaryot Cell. 5:609–612.
Whelan S, Goldman N. 2001. A general model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.
White WT, Hills SF, Gaddam R, Holland BR, Penny D. 2007. Treeness triangles: visualizing the loss of phylogenetic signal. Mol Biol Evol. 24:2029–2039.
Williamson A, Conlan B, Hillier W, Wydrzynski T. 2011. The evolution of photosystem II: insights into the past and future. Photosynth Res. 107: 71–86.
Xiong J, Bauer CE. 2002. Complex evolution of photosynthesis. Annu Rev Plant Biol. 53:503–521.
Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP. 2009. Intertwined evolutionary histories of marine *Synechococcus* and *Prochlorococcus marinus*. Genome Biol Evol. 1:325–339.
Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. 2006. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 16:1099–1108.

Associate editor: John Archibald