The cyanobacterial genome core and the origin of photosynthesis

Armen Y. Mulkidjanian, Eugene V. Koonin, Kira S. Makarova, Sergey L. Mekhedov, Alexander Sorokin, Yuri I. Wolf, Alexis Dufresne, Frédéric Partensky, Henry Burd, Denis Kaznadzey, Robert Haselkorn, and Michael Y. Galperin

PNAS 2006;103;13126-13131; originally published online Aug 21, 2006;
doi:10.1073/pnas.0605709103

The cyanobacterial genome core and the origin of photosynthesis

Armen Y. Mulkidjanian*†‡, Eugene V. Koonin§, Kira S. Makarova§, Sergey L. Mekhedov§, Alexander Sorokin§, Yuri I. Wolf§, Alexis Dufresne¶, Frédéric Partensky¶, Henry Burd‖, Denis Kaznadzey‖, Robert Haselkorn†, and Michael Y. Galperin†§**

*School of Physics, University of Osnabrück, D-49069 Osnabrück, Germany; ‡A. N. Belozersky Institute of Physico–Chemical Biology, Moscow State University, Moscow 119899, Russia; §National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894; ¶Station Biologique, Unité Mixte de Recherche 7144, Centre National de la Recherche Scientifique et Université Paris 6, BP74, F-29682 Roscoff Cedex, France; ‖Integrated Genomics, Inc., Chicago, IL 60612; and **Department of Molecular Genetics and Cell Biology, University of Chicago, 920 East 58th Street, Chicago, IL 60637

Contributed by Robert Haselkorn, July 14, 2006

Comparative analysis of 15 complete cyanobacterial genome sequences, including ''near minimal'' genomes of five strains of *Prochlorococcus* spp., revealed 1,054 protein families [core cyanobacterial clusters of orthologous groups of proteins (core CyOGs)] encoded in at least 14 of them. The majority of the core CyOGs are involved in central cellular functions that are shared with other bacteria; 50 core CyOGs are specific for cyanobacteria, whereas 84 are exclusively shared by cyanobacteria and plants and/or other plastid-carrying eukaryotes, such as diatoms or apicomplexans. The latter group includes 35 families of uncharacterized proteins, which could also be involved in photosynthesis. Only a few components of cyanobacterial photosynthetic machinery are represented in the genomes of the anoxygenic phototrophic bacteria *Chlorobium tepidum*, *Rhodopseudomonas palustris*, *Chloroflexus aurantiacus*, or *Heliobacillus mobilis*. These observations, coupled with recent geological data on the properties of the ancient phototrophs, suggest that photosynthesis originated in the cyanobacterial lineage under the selective pressures of UV light and depletion of electron donors. We propose that the first phototrophs were anaerobic ancestors of cyanobacteria (''procyanobacteria'') that conducted anoxygenic photosynthesis using a photosystem I-like reaction center, somewhat similar to the heterocysts of modern filamentous cyanobacteria. From procyanobacteria, photosynthesis spread to other phyla by way of lateral gene transfer.

cyanobacteria | protein families | lateral gene transfer

Cyanobacteria are one of the earliest branching groups of organisms on this planet (1, 2). They are the only known prokaryotes to carry out oxygenic photosynthesis, and there is little doubt that they played a key role in the formation of atmospheric oxygen ≈2.3 Gyr ago (2). Despite its evolutionary, environmental, and geochemical importance, many aspects of cyanobacterial cell life remain obscure (3–5). Genome sequencing opened a new chapter in cyanobacterial research. In the last few years, complete genome sequences of several freshwater and marine cyanobacteria became available, providing ample data for systematic analysis. A comparison of the complete genomes from three different strains of *Prochlorococcus* spp. demonstrated a wide variety of gene complements within this genus due to massive genome reduction in some lineages (6, 7). Studies of the genes shared by cyanobacteria and other photosynthetic organisms allowed delineation of the ''photosynthetic gene set'' and demonstrated a significant extent of lateral gene transfer (LGT) among phototrophic bacteria (8–11). A somewhat surprising result of the latter work has been that genes for most proteins involved in photosynthesis (hereafter ''photosynthetic genes'') were not in the photosynthetic gene set.

We compared proteins encoded in 15 complete cyanobacterial genomes, including five genomes of *Prochlorococcus* spp., to define the minimal set of genes common to all cyanobacteria and to trace the conservation of these genes among other taxa. We analyzed the phylogenetic affinities of genes in this set and identified previously unrecognized candidate photosynthetic genes. We further used this gene set to address the identity of the first phototrophs, a subject of intense discussion in recent years (8, 9, 12–33). We show that cyanobacteria and plants share numerous photosynthesis-related genes that are missing in genomes of other phototrophs. This observation suggests, in agreement with geological evidence, that (now extinct) anoxygenic ancestors of cyanobacteria are the most plausible candidates for the ancestral photoautotrophs, which apparently disseminated parts of their photosynthetic apparatus to other bacteria by way of LGT.

Results

Common and Unique Protein Families in Cyanobacteria. Clustering of proteins encoded in the 15 complete cyanobacterial genomes yielded 3,188 protein families [cyanobacterial clusters of orthologous groups of proteins (CyOGs)] with members encoded in at least three genomes. Of these CyOGs, 892 were encoded in each cyanobacterial genome, and 162 more were encoded in 14 of 15 genomes (Table 2, which is published as supporting information on the PNAS web site). The combined set of 1,054 CyOGs that are missing in no more than one cyanobacterial genome is hereafter referred to as the core CyOGs. Predictably, cyanobacteria with small genomes are over-represented in the core CyOGs compared with species with larger genomes. Thus, core CyOGs include 52–66% of all proteins encoded in *Prochlorococcus* spp. but only 25% of *Anabaena* sp. PCC 7120 proteins (Table 3, which is published as supporting information on the PNAS web site).

Analysis of CyOGs that apparently had no members in one or more cyanobacterial genomes revealed 31 (mostly short) proteins that are encoded in the respective genomes but were not called by gene-finding programs, such as subunits VI (PetL) and VII (PetM) of the cytochrome b_6f complex (34). We also found five full-length genes that were annotated as pseudogenes in the original genome submissions and whose products were not included in the protein database (Table 4, which is published as supporting information on the PNAS web site).

The stringent criteria used to define the core CyOGs led to

Conflict of interest statement: No conflicts declared.

Freely available online through the PNAS open access option.

Abbreviations: CyOGs, cyanobacterial clusters of orthologous groups of proteins; LGT, lateral gene transfer; PSI, photosystem I; PSII, photosystem II; RC, reaction center.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. DQ831217–DQ831236).

†To whom correspondence may be addressed: E-mail: amulkid@uos.de, r-haselkorn@uchicago.edu, or galperin@ncbi.nlm.nih.gov.

Table 1. Distribution of photosynthesis-related genes in genomes of phototrophic bacteria

Gene products	Cyanobacteria	Plants	Ctep	Rpal	Caur	Hmob
Photosystem I proteins						
Core RC1 subunit PsaA/PsaB	All	Y	Y	—	—	Y
RC1 subunits PsaC, PsaD, PsaE, PsaF, PsaL, PsaM	All	Y	—	—	—	—
RC1 subunits PsaI, PsaJ, PsaK	Missing in Gvio	Y	—	—	—	—
Photosystem II proteins						
Core RC subunit PsbA/PsbD	All	Y	—	Y	Y	—
RC2 subunits PsbB, PsbC, PsbE, PsbF, PsbH, PsbI, PsbJ, PsbK, PsbL, PsbM, PsbO, PsbP, PsbT, PsbW(Psb28), PsbX	All	Y	—	—	—	—
RC2 subunit PsbN	Missing in one Pmar	Y	—	—	—	—
RC2 subunits PsbZ (Ycf9) and Psb27	Missing in Gvio	Y	—	—	—	—
RC2 subunit PsbY (Ycf32)	Missing in Gvio and one Pmar	Y	—	—	—	—
Extrinsic protein PsbU, cytochrome c_{550} PsbV	Missing in four Pmar	Y*	—*	—	—	—
Extrinsic protein PsbO	Missing in Gvio and all Pmar	Y	—	—	—	—
Chlorophyll-binding proteins						
Light-harvesting complexes PcbABC/IsiA	Missing in four genomes	—	—	—	—	—
High-light-inducible proteins HLIP/ELIP/Scp	All	Y	—	—	—	—
Chlorophyll biosynthesis enzymes						
ChlB, ChlL, ChlN, ChlD, ChlH, ChlI, ChlM, ChlG, ChlP	All	Y	Y	Y	Y	Y
Aerobic protoporphyrin IX cyclase AcsF (Ycf59)	All	Y	—	Y	Y	—
Anaerobic protoporphyrin IX cyclase BchE	Only in *Synechocystis*	—	Y	Y	Y	Y
Cytochrome b_6f complex subunits						
PetB, PetC	All	Y	Y	Y	—	Y
PetA, PetD, PetG, PetM, PetN	All	Y	—	—	—	—†
PetL	Missing in Gvio	Y	—	—	—	—†
Water-soluble electron carriers						
Plastocyanin PetE	Missing in Telo	Y	—	—	Y	—
Cytochrome *c6* (PetJ, c_{553})	Missing in one Pmar	Y	—	—	—	—†
Calvin cycle enzymes						
RubisCO large subunit RbcL	All	Y	Y‡	Y	—	Y‡
RubisCO small subunit RbcS	All	Y	—	Y	—	—
Phosphoribulokinase, ribose-5-phosphate isomerase	All	Y	—	Y	Y	—
Regulatory and uncharacterized chloroplast proteins						
Gun4, Tab2, Cp12, PM23, Ycf3, Ycf4, Ycf23, Ycf33, Ycf34, Ycf35, Ycf37, Ycf51, Ycf58, Ycf60	All	Y	—	—	—	—
Cp097, Ycf20, Ycf21, Ycf55, Ycf66	Missing in Gvio	Y	—	—	—	—

The presence or absence of genes coding for proteins of each type in the complete genomes of 15 cyanobacteria, plants, *Chlorobium tepidum* (Ctep), *Rhodopseudomonas palustris* (Rpal), *Chloroflexus aurantiacus* (Caur), and *Heliobacillus mobilis* (Hmob) was assessed as described in *Materials and Methods*. RubisCO, ribulose-1, 5-bisphosphate carboxylase/oxygenase; Gvio, *Gleobacter violaceus;* Pmar, *Prochlorococcus marinus;* Telo, *Thermosynechococcus elongatus;* Y, gene present; —, gene absent. A detailed version of this table appears as Table 5, which is published as supporting information on the PNAS web site.

*PsbU and PsbV are found in red algae, haptophytes, diatoms, and dinoflagellates, but so far not in green plants (29). PsbU is encoded in the genome of *Prosthecochloris aestuarii,* a representative of Chlorobi.

†Proteins of *H. mobilis,* annotated in ref. 20 as PetA, PetD, PetJ, and PetL, are typical components of heterotrophic electron transfer chains, which are only distantly related to the corresponding cyanobacterial and plant proteins. *H. mobilis* genes analyzed in this work have been deposited in GenBank (accession nos. DQ831217–DQ831236).

‡These RbcL homologs appear to participate in methionine salvage, rather than in CO_2 fixation (50, 51).

the exclusion of many previously characterized cyanobacterial proteins. Because marine picocyanobacteria are unicellular, proteins that are involved in filament formation and heterocyst differentiation (3, 4) did not make it into the core set. Certain components of photosystems I (PSI) and II (PSII) are also missing from the core set. For example, the 12-kDa extrinsic subunit, PsbU, and a low-potential cytochrome c_{550}, PsbV, which both contribute to stabilization of the oxygen-evolving complex, are missing in four *Prochlorococcus* genomes (35). In contrast, PSI components PsaI, PsaJ, and PsaK and PSII component PsbZ, which are missing in the thylakoid-less cyanobacterium *Gloeobacter violaceus* (36, 37), are found in all other cyanobacterial genomes and hence were included in the core set, as was plastocyanin, the electron donor to PSI, which is missing in *Thermosynechococcus elongatus* (Table 1). Owing to the poor representation of genes involved in environmental sensing and signal transduction in the genomes of marine picocyanobacteria, most likely due to their adaptation to nutrient-poor and relatively constant oceanic environments (35, 38), there are few regulatory genes in the core set. In 85 of the 162 core CyOGs that lacked representatives from a single organism, that organism was *G. violaceus*. Proteins from one of the *Prochlorococcus* strains were missing in 31 core CyOGs, the thermophile *T. elongatus* was missing in 22 core CyOGs, and *Synechocystis* sp. was missing in 20 core CyOGs (Table 3).

Most core CyOGs comprise tight clusters, with various cyanobacterial proteins showing much higher similarity to each other than to any proteins from other organisms (Fig. 2, which is published as supporting information on the PNAS web site). However, certain proteins are only distantly related to other members of the CyOG and might represent examples of relatively recent LGT in the corresponding lineage. Examples include *G. violaceus* genes for arginyl-tRNA synthetase *glr4279*, chorismate synthase *glr3393*, and γ-glutamyl phosphate reductase *gll3923*, *Synechocystis* sp. genes for 6-pyruvoyl-tetrahydropterin synthase *slr0078*, α- (*slr1239*) and

β- (*slr1434*) subunits of NAD/NADP transhydrogenase, and many others.

Phylogenetic Affinities of the Core Cyanobacterial Genes. Of the 1,054 core CyOGs, 936 are shared with other bacteria. This set includes primarily housekeeping proteins that are involved in DNA replication and repair, transcription, translation, key metabolic pathways, and energy metabolism. Approximately 50 core CyOGs shared with other bacteria are formed by "conserved hypothetical" proteins whose functions are unknown and cannot be predicted from sequence similarity (39). Almost one-third of the families that are shared with other bacteria (291 CyOGs) are also encoded in plant genomes. In addition to the ribosomal proteins and other components of the chloroplast transcription and translation machinery, this list includes enzymes of heme biosynthesis, subunits of the respiratory complexes I (NADH dehydrogenase) and III (cytochrome *c* oxidase), and F_0F_1-ATP synthase, as well as subunits of the cytochrome b_6f complex (Table 2).

Eighty-four core CyOGs are shared exclusively with plants, such as *Arabidopsis thaliana* and *Oryza sativa*, the red alga *Cyanidioschyzon merolae*, and the diatom *Thalassiosira pseudonana*. Approximately half of these proteins have known functions and participate in photosynthesis as components of PSI, PSII, light-harvesting systems, or members of the high-light-inducible protein (HLIP)/early light-inducible protein (ELIP) superfamily (Table 1). Thirty-five CyOGs with the same phylogenetic profile (i.e., encoded in at least 14 cyanobacterial genomes and in at least some chloroplast-containing eukaryotes, but not in any other of the >350 prokaryotic and eukaryotic genomes) have no known function or have general function only (Table 6, which is published as supporting information on the PNAS web site). This profile suggests that the functions of these proteins are related to photosynthesis. Indeed, several recently characterized proteins with similar phylogenetic profiles turned out to participate in chlorophyll biosynthesis (40), photosynthesis (41), and light-driven NAD(P) reduction (42, 43).

Unusual Phylogenetic Profiles. Although the great majority of core CyOGs are shared with other bacteria and/or plants, some were also found in other eukaryotes that carry vestigial plastids, such as apicoplasts in *Plasmodium falciparum* and other apicomplexans. The proteins of likely cyanobacterial origin that are shared by apicomplexans and other plastid-carrying eukaryotes but that are missing in other eukaryotes include enzymes of the deoxyxylulose pathway of terpenoid biosynthesis, fatty acid biosynthesis, DNA gyrase, peptide deformylase, and several others. Most of these proteins are validated targets for antimalarial drugs (44, 45). Other proteins with similar phylogenetic profiles, such as translation initiation factor IF-1 (InfA), phosphoenolpyruvate carboxylase, and several poorly characterized proteins (e.g., seed maturation protein PM23) are also likely to function in apicoplasts and might merit exploration as additional drug targets. Another interesting group includes genes that are found exclusively in phototrophs, for example, *sll0608* and *sll0609* (ycf49), which are encoded in all cyanobacteria and are shared with plants and representatives of Chlorobi.

Cyanobacterial Synapomorphies. In addition to tight clustering of their core proteins (Fig. 2), cyanobacteria possess other features identifying them as members of a distinct phylogenetic lineage (clade). The apparent cyanobacterial synapomorphies (unique features shared by members of a clade) include 50 core CyOGs that do not have close homologs in other organisms (see *Materials and Methods*). These CyOGs are listed in Table 7, which is published as supporting information on the PNAS web site. Remarkably, the function of only one of these genes is known: SomA (Slr0042) is an outer membrane porin (46). Although functions of the other synapomorphic core CyOGs remain unknown, some of their members are expressed under stress conditions, e.g., Sll1507, Slr1160, and Slr1915 are induced by high salt (47). Conservation of these proteins in (almost) all cyanobacteria makes them attractive targets for future experimental studies (see ref. 39).

Photosynthetic Genes in the Conserved Cyanobacterial/Plant Core. The vast majority of cyanobacterial photosynthetic genes had no detectable homologs in anoxygenic phototrophic bacteria (Table 1). Genomes of two such phototrophs, the purple α-proteobacterium *Rhodopseudomonas palustris* and the green sulfur bacterium *Chlorobium tepidum*, have been sequenced, allowing us to establish orthologous relationships between their genes and those of cyanobacteria (48, 49). In addition, the genome of the green nonsulfur bacterium *Chloroflexus aurantiacus* was released by the Department of Energy Joint Genome Institute in an unfinished form, and the genome of the Gram-positive phototrophic bacterium *Heliobacillus mobilis* has been sequenced by Integrated Genomics, Inc. (Chicago, IL) and was kindly made available to us for some BLAST searches. Although these phototrophic bacteria and cyanobacteria share numerous typically bacterial proteins (8, 9, 11), we found that anoxygenic photosynthetic bacteria possess very few photosynthetic genes shared by cyanobacteria and plants; furthermore, even these shared genes differ in different bacteria (Table 1). Of the seven groups of cyanobacterial genes that are directly related to photosynthesis, only some genes of (bacterio)chlorophyll biosynthesis are shared by all prokaryotic phototrophs.

For example, whereas all cyanobacteria encode the full set of enzymes of the Calvin–Benson–Bassham cycle, the *Chlorobium* and *Heliobacillus* genomes lack the genes for phosphoribulokinase and ribose-5-phosphate isomerase, as well as the gene encoding the small subunit of ribulose-1,5-bisphosphate carboxylase/oxygenase (RubisCO) (Table 1). These organisms encode proteins similar to the large subunit of RubisCO; however, these are likely to participate in methionine salvage, rather than in CO_2 fixation (50, 51). Autotrophic CO_2 fixation by *Chloroflexus* is known to occur by way of the 3-hydroxypropionate cycle (52). This finding leaves *R. palustris* as the only anoxygenic phototroph in Table 1 to use the Calvin cycle.

Analysis of the unfinished genomes of two other phototrophs, β-proteobacterium *Rubrivivax gelatinosus* and γ-proteobacterium *Thermochromatium tepidum*, revealed a pattern of presence and absence of key photosynthetic genes that was very similar to that of *R. palustris* (data not shown) and is likely to be common to all purple phototrophic bacteria.

Discussion

The "Core Set" Versus the "Genomic Signature." Owing to the high level of sequence conservation among orthologous proteins from different cyanobacteria (Fig. 2), delineation of cyanobacterial gene clusters is a relatively straightforward task. In several earlier studies, this delineation was accomplished for such purposes as improved genome annotation (53), delineation of the cyanobacterial genomic signature (54), calculation of the number of cyanobacterial genes in plants (10), and tracing the evolution of the oxygen-evolving center of PSII (29). However, all these studies relied on arbitrarily set, usually conservative, threshold similarity values to infer orthology. As described previously, the cluster of orthologous groups (COG) approach, which does not depend on such thresholds, is more flexible and allows delineation of protein families with low, as well as high, levels of similarity (49). This procedure, however, can be used reliably only for complete genomes, which is why unfinished cyanobacterial genomes were not included in this work. Besides, certain CyOGs could contain homologous genes whose functions have diverged during evolution. For instance, homologous genes

for phycoerythrin (*cpeB*) and phycocyanin (*cpcB*) β-chains included in CyOG00868 are most likely in paralogs (55) rather than true orthologs.

A comparison of eight genomes, two finished and six unfinished, has been used to delineate a genomic signature of 181 cyanobacteria-specific proteins (54). A comparison of the core CyOGs with this genome signature showed that 131 of the 181 signature protein families survived inclusion of genomes of three more strains of *Prochlorococcus* spp. and four more strains of *Synechococcus* spp. and were represented in the core CyOGs (Table 3). In contrast, our analysis identified 26 synapomorphic CyOGs that were not included in the cyanobacterial genome signature (54). The 50 protein families that did not make it into the core CyOGs were most often missing in *G. violaceus* and in one or two strains of *Prochlorococcus*. In addition, for at least 19 of the remaining 131 core CyOGs, close homologs have been found among recently sequenced bacterial or archaeal genomes. This finding is hardly surprising given the rapid growth of the protein database and the large scale of lateral gene transfer among various lineages. There is little doubt that the current list of 50 cyanobacteria-specific core CyOGs (Table 6) will soon shrink even further.

Evolution of Photosynthesis and Lateral Gene Transfer. The availability of the cyanobacterial genome core allowed us to reassess the origin of (bacterio)chlorophyll-based photosynthesis, which, in addition to cyanobacteria, is found in the Bacteroidetes/Chlorobi group (e.g., *C. tepidum*), Firmicutes (e.g., *H. mobilis*), α-Proteobacteria (e.g., *R. palustris*), β-Proteobacteria (e.g., *Rubrivivax gelatinosum*), γ-Proteobacteria (e.g., *Chromatium vinosum*), and Chloroflexi (e.g., *C. aurantiacus*). The first two phyla have photosynthetic reaction centers (RCs) that are similar to the cyanobacterial PSI and that use low-potential FeS clusters as electron acceptors (RC1 type). The RCs of proteobacteria and Chloroflexi (RC2-type) use bound quinones as ultimate electron acceptors and are similar to the cyanobacterial PSII (although lacking the oxygen-evolving complex). It is generally believed that the evolution of photosynthetic genes was accompanied by their dissemination by way of LGT between different groups of bacteria (12, 26, 30, 31, 56). This idea is supported by the apparent presence of nonphotosynthetic representatives in all of these phyla, except for Cyanobacteria; by the fact that the photosynthesis-related proteins are often encoded on a single contiguous chromosomal region (superoperon) (20, 57); by several phylogenetic analyses (8, 11, 58); and by the observation that photosynthetic genes can be transduced by cyanophages (59).

The propensity of photosynthetic genes to be laterally transferred between distantly related organisms makes identification of the lineage that was the first to develop chlorophyll-based photosynthesis particularly challenging. The extremely small number of photosynthetic proteins that are shared between different photosynthetic bacteria (Table 1) forced the phylogenetic analyses to rely on surrogate protein sets, such as the enzymes involved in (bacterio)chlorophyll biosynthesis (14). These analyses contributed to the understanding of the evolution of (bacterio)chlorophyll (17) but did not purport to reflect evolution of photosynthesis in general. In contrast, the recent study of Xiong *et al.* (21) assumed that topology of the phylogenetic tree built for the (bacterio)chlorophyll biosynthesis enzymes is representative of the evolution of the photosynthetic machinery as a whole. Specifically, the authors' observations that proteobacteria branched first in the tree were interpreted as evidence that photosynthesis originally evolved in purple bacteria (21). Others, however, were either unable to reproduce this result (31) or have not observed this topology in phylogenetic trees for any other genes (11). In addition, Green and Gantt (22) noted that (*i*) branching of proteobacterial genes at the root of the tree meant that they were the most divergent in the set, not necessarily that they branched off earlier, and (*ii*) "ancient" genes in modern proteobacteria could have originated elsewhere, and purple bacteria could have acquired them by way of LGT. Thus, even if the observations of Xiong *et al.* (21) were valid, their conclusions on the origin of photosynthesis among purple bacteria do not follow from their results.

The data in Table 1, which show that only some enzymes of (bacterio)chlorophyll biosynthesis are found in all phototrophs, together with the observations of extensive LGT and recombination in cyanobacterial genomes, limit the contribution of the standard tree-based approach to the problem of the origin of photosynthesis. Analysis of phylogenetic patterns of key photosynthetic proteins might be more informative for this purpose.

Which Bacteria Were the First Phototrophs? In the past few years, photosynthesis has been proposed to have emerged in *Heliobacillus* (16, 27), *Chlorobium* (13), *Chloroflexus* (15), or proteobacterial (21) lineages (reviewed in refs. 24, 26, 30, and 32). Although the arguments in favor of proteobacteria do not appear valid (see above), there seems to be some support for each of the other candidates. Thus, apparently primitive homodimeric RCs of type I are found in *Chlorobium* (13) and *Heliobacillus* (60), whereas *Chloroflexus* is believed to be an early-branching lineage of phototrophs (15). Cyanobacteria are usually not considered explicitly as a lineage in which photosynthesis could have emerged because of the far greater complexity of their photosynthetic machinery. This fact, however, can be interpreted both ways. Indeed, the total number of genes involved in photosynthesis in cyanobacteria is much greater than that in any of the other prokaryotic phototrophs (Table 1). Only cyanobacteria possess photosynthetic reaction centers of both types, RC1 and RC2, and, in addition to chlorophyll- and phycobilin-containing light-harvesting systems, have chlorophyll-binding proteins whose function is believed to be dissipation of light energy to prevent photodamage (HLIPs; see Table 1). Thus, the majority of photosynthetic genes must have first appeared in the cyanobacterial lineage anyway (Fig. 1). This finding suggests that the same could be true for the core RC genes and that the ancestors of cyanobacteria ("procyanobacteria" or "pro-protocyanobacteria" in ref. 26) should also be considered as candidates for the role of the first phototrophs.

Sequence data alone do not allow one to establish the direction of ancient lateral transfer of photosynthetic genes; this requires additional information from independent sources. Important clues to the nature of the first phototrophs can be gained from geological data. Tice and Lowe (61, 62) have provided geological evidence that the Buck Reef Chert, a 250- to 400-m thick rock running along the South African coast, was produced by phototrophic microbial communities *ca.* 3.4 Gyr ago. They also noted the absence of traces of life in the deeper (>200 m) water environments (61). Tice and Lowe (61, 62) defined the inhabitants of the primordial microbial communities as partially filamentous phototrophs, which, according to the carbon isotopic composition, used the Calvin cycle to fix CO_2. The absence of oxidized iron and sulfur in the sediments indicated that neither iron (II) nor sulfide had been used as electron donors (62). By exclusion, this finding leaves atmospheric hydrogen as the most plausible electron donor (62, 63). Because the Calvin cycle is absent in Gram-positive (*Heliobacillus*) and green sulfur (*Chlorobium*) phototrophs that use other pathways for CO_2 fixation (56), it is unlikely that their ancestors were the phototrophic inhabitants of the Buck Reef Chert. *C. aurantiacus* does not have the Calvin cycle either, but this pathway has been reported in another representative of green nonsulfur bacteria, *Oscillochloris trichoides* (64). However, RC2, which is found in Chloroflexi and purple bacteria, would hardly be useful in a

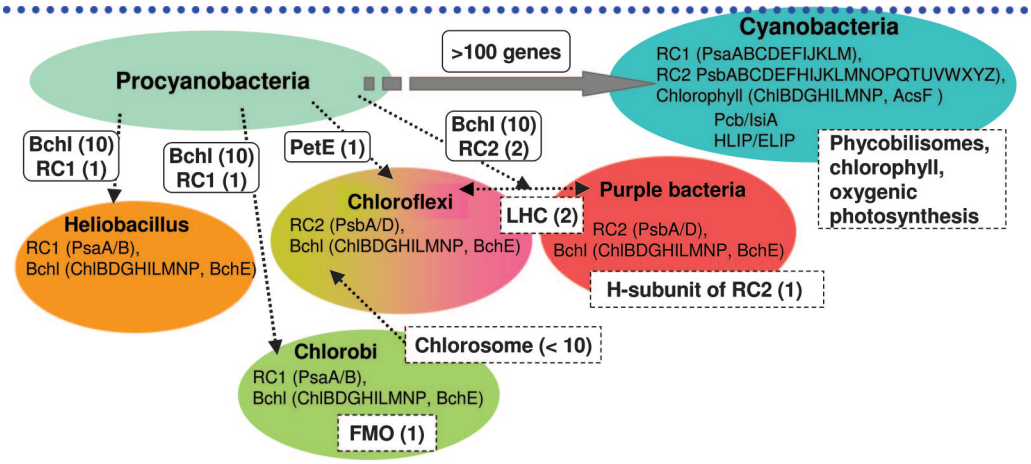

Fig. 1. Distribution of photosynthetic genes in different lineages of phototrophs and the directions of proposed lateral gene transfer. The phototrophic phyla are depicted in accordance with the depth of their location in modern (and perhaps primordial) microbial mats. Rounded boxes show the extent of photosynthetic gene transfer between the phyla, with the numbers of genes (CyOGs) transferred indicated in parentheses. Dashed boxes show major photosynthesis-relevant ''inventions'' that occurred outside the (pro-) cyanobacterial lineage. A detailed version of the figure, with appropriate references, appears as Fig. 3, which is published as supporting information on the PNAS web site.

hydrogen-driven metabolism. As noted by Olson (24), RC2 uses a quinone as the last electron acceptor, so it would therefore be over-reduced and kinetically incompetent under these conditions.

These observations leave the ancestors of cyanobacteria as the only phototrophs capable of inhabiting the Buck Reef Chert *ca.* 3.4 Gyr ago. Indeed, the mechanism of CO_2 fixation, the morphology of these organisms, and their location in the upper layer of the ancient microbial mat all unite them with modern cyanobacteria. Thus, analysis of the gene content (Table 1) and the geologic evidence both suggest that photosynthesis evolved in the cyanobacterial lineage. Because there is sufficient evidence that anoxygenic photosynthesis preceded oxygenic photosynthesis and was already taking place in the period between 3.5 and 2.5 Gyr ago, we propose that the first phototrophs were procyanobacteria (anoxygenic ancestors of the extant cyanobacteria) that could be responsible for the presence of cyanobacteria-specific biomarkers (2-methylhopanoides) in the 2.7-Gyr-old sediments (65). These anoxygenic procyanobacteria might have relied on RC1 to reduce $NAD(P)^+$ to NAD(P)H and resembled heterocysts, the specialized nitrogen-fixing cells that some modern filamentous cyanobacteria produce in response to starvation for fixed nitrogen. Heterocysts have PSI but no PSII and therefore do not conduct oxygenic photosynthesis (3, 4, 66). Instead, they maintain the anaerobic environment required for nitrogenase activity. Although modern heterocysts are a relatively recent invention (67), their formation can be viewed as a recapitulation of the ancestral cyanobacterial state and confirm the viability of cyanobacteria in a PSI-alone mode in the presence of suitable electron donors.

Driving Forces in the Origin and Evolution of Photosynthesis. The complexity of the photosynthetic machinery leaves no doubt that its origin and subsequent evolution must have occurred in multiple steps under constant selective pressure. This selective pressure could come from at least two key factors: the necessity for the cells to gain energy and to reduce the damaging effects of solar UV, which was orders-of-magnitude stronger in the absence of the ozone shield than it is now (68). As proposed by several authors, RC1 could evolve by way of multiple duplication events from simpler chlorophyll-binding membrane proteins, similar to the HLIPs of modern cyanobacteria (18, 19, 25). As argued previously, such proteins might serve to protect DNA from the damaging effects of UV light (18, 19). The emergence of RC1 could have been driven by the need for an alternative source of reducing power as the atmospheric hydrogen content gradually decreased. NAD(P)H, which is recycled by RC1, has a redox potential similar to that of hydrogen and can replace hydrogen in certain metabolic chains; the membrane hydrogenase and NADH-dehydrogenase are related enzymes that differ only in the substrate-binding module (69).

Upon gradual oxidation of the atmosphere, the need for further sources of redox equivalents could have driven the formation of a small, high-potential RC2, possibly through fission/reshuffling of RC1 (16, 18, 19, 28). Further depletion of electron donors upon oxidation of the available Fe(II), as discussed by several authors (e.g., refs. 24, 28, and 29), could have driven the evolution of RC2 into the water-oxidizing PSII.

In this framework, modern cyanobacteria inherited their photosynthetic apparatus from ancestral phototrophs, whereas other phototrophic bacterial lineages obtained theirs by way of LGT (Fig. 1). These transfer events must have happened at different stages of evolution: The ancestors of *Chlorobium* and *Heliobacterium* must have acquired their RC1 soon after its emergence, when it was still homodimeric, whereas Proteobacteria and *Chloroflexus* acquired RC2 before it ''learned'' to oxidize water. Anoxygenic phototrophs usually dwell in the depth of microbial mats. Perhaps, therefore, they were subject to a weaker selective pressure from light and oxygen than those (ancestors of modern cyanobacteria) that remained on the surface, resulting in preservation of ancestral features of the photosynthetic apparatus. Thus, photosynthetic enzymes of anaerobic bacteria can be considered snapshots of the ancient RCs: The homodimeric RC1 of *Heliobacillus mobilis* (PshA) and *Chlorobium tepidum* (PscA) are probably more similar to the ancient homodimeric RC1 than the highly evolved heterodimeric PSI (PsaA/PsaB) of modern cyanobacteria.

Materials and Methods

Protein sets for *Anabaena* (*Nostoc*) sp. PCC 7120, *Synechocystis* sp. PCC 6803, *T. elongatus* BP-1, and *Prochlorococcus marinus* SS120 were extracted from GenBank (www.ncbi.nlm.nih.gov) and clustered by using the cluster of orthologous group (COG) method (48, 49). Proteins from an additional 11 cyanobacterial genomes (Table 2) were assigned to the resulting protein clusters (CyOGs) by using a modification of the COGNITOR

procedure (49), followed by manual verification and analysis of multidomain proteins. CyOGs that were missing representatives of one, two, or three species, as well as CyOGd that contained proteins shorter than 100-aa residues, were compared with the translation of the corresponding genomic DNA sequences by using TBLASTn (70). Detection of homologs of cyanobacterial proteins in organisms from other taxa was performed using Blastp searches against the National Center for Biotechnology Information (NCBI) nonredundant protein database. Phylogenetic distributions of homologs for each CyOG were analyzed by comparing them to prokaryotic and eukaryotic protein families and by checking for bidirectional best hits and domain architecture, as described in refs. 48 and 49. Cyanobacteria-specific CyOGs were defined as those consisting of proteins that did not retrieve noncyanobacterial hits after three iterations of PSI-BLAST run with the default inclusion parameter of $E = 0.001$.

We thank Beverly Green and Brian Palenik for helpful comments and Integrated Genomics, Inc., for allowing us access to the genomic sequence data for *H. mobilis* and *T. tepidum*. This study was supported by the Intramural Research Program of the National Institutes of Health, the National Library of Medicine (E.V.K., K.S.M., S.L.M., A.S., Y.I.W., and M.Y.G.), the European Community Program SynChips, the Network of Excellence Marine Genomics Europe, the Region Bretagne Program IMPALA (A.D. and F.P.), the Deutsche Forschungsgemeinschaft (A.Y.M.), and the Volkswagen Foundation (A.Y.M.).

1. Altermann, W. & Kazmierczak, J. (2003) *Res. Microbiol.* **154,** 611–617.
2. Bekker, A., Holland, H. D., Wang, P. L., Rumble, D., III, Stein, H. J., Hannah, J. L., Coetzee, L. L. & Beukes, N. J. (2004) *Nature* **427,** 117–120.
3. Haselkorn, R. (1998) *Science* **282,** 891–892.
4. Meeks, J. C. & Elhai, J. (2002) *Microbiol. Mol. Biol. Rev.* **66,** 94–121.
5. Bhaya, D. (2004) *Mol. Microbiol.* **53,** 745–754.
6. Hess, W. R. (2004) *Curr. Opin. Biotechnol.* **15,** 191–198.
7. Dufresne, A., Garczarek, L. & Partensky, F. (2005) *Genome Biol.* **6,** R14.
8. Raymond, J., Zhaxybayeva, O., Gogarten, J. P., Gerdes, S. Y. & Blankenship, R. E. (2002) *Science* **298,** 1616–1620.
9. Raymond, J., Zhaxybayeva, O., Gogarten, J. P. & Blankenship, R. E. (2003) *Philos. Trans. R. Soc. London B* **358,** 223–230.
10. Sato, N. (2002) *Genome Inform.* **13,** 173–182.
11. Zhaxybayeva, O., Hamel, L., Raymond, J. & Gogarten, J. P. (2004) *Genome Biol.* **5,** R20.
12. Blankenship, R. E. (1992) *Photosynth. Res.* **33,** 91–111.
13. Buttner, M., Xie, D. L., Nelson, H., Pinther, W., Hauska, G. & Nelson, N. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 8135–8139.
14. Burke, D. H., Hearst, J. E. & Sidow, A. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7134–7138.
15. Pierson, B. K. (1994) in *Early Life on Earth: Nobel Symposium No. 84*, ed. Bengtson, S. (Columbia Univ. Press, New York), pp. 161–180.
16. Vermaas, W. F. (1994) *Photosynth. Res.* **41,** 285–294.
17. Lockhart, P. J., Larkum, A. W., Steel, M., Waddell, P. J. & Penny, D. (1996) *Proc. Natl. Acad. Sci. USA* **93,** 1930–1934.
18. Mulkidjanian, A. Y. & Junge, W. (1997) *Photosynth. Res.* **51,** 27–42.
19. Mulkidjanian, A. Y. & Junge, W. (1999) in *The Phototrophic Prokaryotes*, eds. Peschek, G. A., Löffelhardt, W. & Schmetterer, G. (Kluwer Academic/Plenum, New York), Vol. 51, pp. 805–812.
20. Xiong, J., Inoue, K. & Bauer, C. E. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 14851–14856.
21. Xiong, J., Fischer, W. M., Inoue, K., Nakahara, M. & Bauer, C. E. (2000) *Science* **289,** 1724–1730.
22. Green, B. R. & Gantt, E. (2000) *J. Phycology* **36,** 983–985.
23. Baymann, F., Brugna, M., Muhlenhoff, U. & Nitschke, W. (2001) *Biochim. Biophys. Acta* **1507,** 291–310.
24. Olson, J. M. (2001) *Photosynth. Res.* **68,** 95–112.
25. Garczarek, L., Poupon, A. & Partensky, F. (2003) *FEMS Microbiol. Lett.* **222,** 59–68.
26. Green, B. R. (2003) in *Light-Harvesting Antennas in Photosynthesis*, eds. Green, B. R. & Parson, W. W. (Kluwer, Dordrecht, The Netherlands), pp. 129–168.
27. Gupta, R. S. (2003) *Photosynth. Res.* **76,** 173–183.
28. Rutherford, A. W. & Faller, P. (2003) *Philos. Trans. R. Soc. London B* **358,** 245–253.
29. De Las Rivas, J., Balsera, M. & Barber, J. (2004) *Trends Plant Sci.* **9,** 18–25.
30. Olson, J. M. & Blankenship, R. E. (2004) *Photosynth. Res.* **80,** 373–386.
31. Mix, L. J., Haig, D. & Cavanaugh, C. M. (2005) *J. Mol. Evol.* **60,** 153–163.
32. Nelson, N. & Ben-Shem, A. (2005) *BioEssays* **27,** 914–922.
33. Olson, J. M. (2006) *Photosynth. Res.* **88,** 109–117.
34. Zhang, H. & Cramer, W. A. (2004) *Methods Mol. Biol.* **274,** 67–78.
35. Dufresne, A., Salanoubat, M., Partensky, F., Artiguenave, F., Axmann, I. M., Barbe, V., Duprat, S., Galperin, M. Y., Koonin, E. V., Le Gall, F., *et al.* (2003) *Proc. Natl. Acad. Sci. USA* **100,** 10020–10025.
36. Nakamura, Y., Kaneko, T., Sato, S., Ikeuchi, M., Katoh, H., Sasamoto, S., Watanabe, A., Iriguchi, M., Kawashima, K., Kimura, T., *et al.* (2002) *DNA Res.* **9,** 123–130.
37. Inoue, H., Tsuchiya, T., Satoh, S., Miyashita, H., Kaneko, T., Tabata, S., Tanaka, A. & Mimuro, M. (2004) *FEBS Lett.* **578,** 275–279.
38. Palenik, B., Brahamsha, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E. E., McCarren, J., *et al.* (2003) *Nature* **424,** 1037–1042.
39. Galperin, M. Y. & Koonin, E. V. (2004) *Nucleic Acids Res.* **32,** 5452–5463.
40. Larkin, R. M., Alonso, J. M., Ecker, J. R. & Chory, J. (2003) *Science* **299,** 902–906.
41. Dauvillee, D., Stampacchia, O., Girard-Bascou, J. & Rochaix, J. D. (2003) *EMBO J.* **22,** 6378–6388.
42. Prommeenate, P., Lennon, A. M., Markert, C., Hippler, M. & Nixon, P. J. (2004) *J. Biol. Chem.* **279,** 28165–28173.
43. Battchikova, N., Zhang, P., Rudd, S., Ogawa, T. & Aro, E. M. (2005) *J. Biol. Chem.* **280,** 2587–2595.
44. Gornicki, P. (2003) *Int. J. Parasitol.* **33,** 885–896.
45. Tripathi, R. P., Mishra, R. C., Dwivedi, N., Tewari, N. & Verma, S. S. (2005) *Curr. Med. Chem.* **12,** 2643–2659.
46. Hansel, A., Pattus, F., Jurgens, U. J. & Tadros, M. H. (1998) *Biochim. Biophys. Acta.* **1399,** 31–39.
47. Huang, F., Fulda, S., Hagemann, M. & Norling, B. (2006) *Proteomics* **6,** 910–920.
48. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. (1997) *Science* **278,** 631–637.
49. Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. (2000) *Nucleic. Acids Res.* **28,** 33–36.
50. Hanson, T. E. & Tabita, F. R. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 4397–4402.
51. Ashida, H., Danchin, A. & Yokota, A. (2005) *Res. Microbiol.* **156,** 611–618.
52. Herter, S., Farfsing, J., Gad'On, N., Rieder, C., Eisenreich, W., Bacher, A. & Fuchs, G. (2001) *J. Bacteriol.* **183,** 4305–4316.
53. Nakamura, Y., Kaneko, T. & Tabata, S. (2000) *Nucleic. Acids Res.* **28,** 72.
54. Martin, K. A., Siefert, J. L., Yerrapragada, S., Lu, Y., McNeill, T. Z., Moreno, P. A., Weinstock, G. M., Widger, W. R. & Fox, G. E. (2003) *Photosynth. Res.* **75,** 211–221.
55. Sonnhammer, E. L. & Koonin, E. V. (2002) *Trends Genet.* **18,** 619–620.
56. Overmann, J. & Garcia-Pichel, F. (2000) in *The Prokaryotes: An Evolving Electronic Resource for the Microbiological Community*, Release 3.2, ed. Dworkin, M. (Springer, New York), available at http://link.springer-ny.com/link/service/books/10125. Accessed March 3, 2006.
57. Choudhary, M. & Kaplan, S. (2000) *Nucleic Acids Res.* **28,** 862–867.
58. Igarashi, N., Harada, J., Nagashima, S., Matsuura, K., Shimada, K. & Nagashima, K. V. (2001) *J. Mol. Evol.* **52,** 333–341.
59. Lindell, D., Sullivan, M. B., Johnson, Z. I., Tolonen, A. C., Rohwer, F. & Chisholm, S. W. (2004) *Proc. Natl. Acad. Sci. USA* **101,** 11013–11018.
60. Liebl, U., Mockensturm-Wilson, M., Trost, J. T., Brune, D. C., Blankenship, R. E. & Vermaas, W. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7124–7128.
61. Tice, M. M. & Lowe, D. R. (2004) *Nature* **431,** 549–552.
62. Tice, M. M. & Lowe, D. R. (2006) *Geology* **34,** 37–40.
63. Nisbet, E. G. & Sleep, N. H. (2001) *Nature* **409,** 1083–1091.
64. Berg, I. A., Keppen, O. I., Krasil'nikova, E. N., Ugol'kova, N. V. & Ivanovskii, R. N. (2005) *Mikrobiologiia* **74,** 258–264.
65. Summons, R. E., Jahnke, L. L., Hope, J. M. & Logan, G. A. (1999) *Nature* **400,** 554–557.
66. Golden, J. W. & Yoon, H. S. (2003) *Curr. Opin. Microbiol.* **6,** 557–563.
67. Tomitani, A., Knoll, A. H., Cavanaugh, C. M. & Ohno, T. (2006) *Proc. Natl. Acad. Sci. USA* **103,** 5442–5447.
68. Garcia-Pichel, F. (1998) *Origins Life Evol. Biosphere* **28,** 321–347.
69. Friedrich, T. & Scheide, D. (2000) *FEBS Lett.* **479,** 1–5.
70. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zheng, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.