Genome evolution in cyanobacteria: The stable core and the variable shell

Tuo Shi[*][†] **and Paul G. Falkowski**[*][‡][§]

[*]Environmental Biophysics and Molecular Ecology Program, Institute of Marine and Coastal Sciences and [‡]Department of Earth and Planetary Sciences, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901

Contributed by Paul G. Falkowski, November 27, 2007 (sent for review September 18, 2007)

Cyanobacteria are the only known prokaryotes capable of oxygenic photosynthesis, the evolution of which transformed the biology and geochemistry of Earth. The rapid increase in published genomic sequences of cyanobacteria provides the first opportunity to reconstruct events in the evolution of oxygenic photosynthesis on the scale of entire genomes. Here, we demonstrate the overall phylogenetic incongruence among 682 orthologous protein families from 13 genomes of cyanobacteria. However, using principal coordinates analysis, we discovered a core set of 323 genes with similar evolutionary trajectories. The core set is highly conserved in amino acid sequence and contains genes encoding the major components in the photosynthetic and ribosomal apparatus. Many of the key proteins are encoded by genome-wide conserved small gene clusters, which often are indicative of protein–protein, protein–prosthetic group, and protein–lipid interactions. We propose that the macromolecular interactions in complex protein structures and metabolic pathways retard the tempo of evolution of the core genes and hence exert a selection pressure that restricts piecemeal horizontal gene transfer of components of the core. Identification of the core establishes a foundation for reconstructing robust organismal phylogeny in genome space. Our phylogenetic trees constructed from 16S rRNA gene sequences, concatenated orthologous proteins, and the core gene set all suggest that the ancestral cyanobacterium did not fix nitrogen and probably was a thermophilic organism.

horizontal (lateral) gene transfer | oxygenic photosynthesis | gene family | nitrogen fixation

Oxygenic photosynthesis is arguably the most important biological process on Earth. Approximately 2.3 billion years ago (Ga) (1–4), that energy transduction pathway transformed Earth's atmosphere and upper ocean, ultimately facilitating the development of complex life forms that depend on aerobic metabolism (5–7). Cyanobacteria are widely accepted as the progenitor of oxygenic photosynthesis, and the clade has evolved into one of the largest and most diverse groups of bacteria on this planet (8). Cyanobacteria contribute significantly to global primary production (9, 10), and diazotrophic taxa are central to global nitrogen cycle (11–13). Arguably, no other prokaryotic group has had a greater impact on the biogeochemistry and evolutionary trajectory of Earth, yet its own evolutionary history is poorly understood.

The availability of complete genomes of related organisms provides the first opportunity to reconstruct events of genomic evolution through the analysis of entire functional classes (14). Currently, cyanobacteria represent one of the densest clusters of fully sequenced genomes [supporting information (SI) Table 1]. Comparisons of genome sequences of closely related marine *Prochlorococcus* and *Synechococcus* species have demonstrated an intimate link between genome divergence in specific strains and their physiological adaptations to different oceanic niches (15, 16). This ecotypic flexibility appears to be driven by myriad selective pressures that govern genome size, GC content, gene gains and losses, and rate of evolution (17, 18). Moreover, phylogenetic analyses of genes shared by all of the five known phyla of photosynthetic bacteria, including cyanobacteria, purple bacteria (Proteobacteria), green sulfur bacteria (Chlorobi), green filamentous bacteria (Chloroflexi), and Gram-positive heliobacteria (Firmicutes), have provided important insights into the origin and evolution of photosynthesis, an intensively debated subject in the past decades (19–29). This information has been substantially extended by genome-wide comparative informatics (30–32). One of the major implications of the latter work is a significant extent of horizontal gene transfer (HGT) among these photosynthetic bacteria. The observation that cyanophages sometimes carry photosynthetic genes (33–35) provides one mechanism of rapid HGT among these phyla. However, HGTs almost certainly do not occur with equal probability for all genes. For example, informational genes (those involved in transcription, translation, and related processes), which are thought to have more macromolecular interactions than operational genes (those involved in housekeeping), are postulated to be seldom transferred (36, 37). The existence of a core of genes that remain closely associated and resistant to HGT has been reported in recent studies using relatively intensive taxon sampling (38, 39). Identification of such core genes potentially allows separation of true phylogenetic signals from "noise." It is, therefore, of considerable interest to transcribe all coherent genome data into pertinent phylogenetic information and to identify which genes are more susceptible to HGT.

Here, we report on identification and reconstruction of the phylogeny of 682 orthologs from 13 genomes of cyanobacteria. Our primary goals are twofold: (*i*) to examine the impact of HGT on the evolution of photosynthesis and the radiation of cyanobacterial lineages; and (*ii*) to identify a core set of genes that are resistant to HGT on which robust organismal phylogeny can be reconstructed. Our results reveal that >52% (359) of the orthologs are susceptible to HGT within the cyanobacterial phylum and hence are responsible for the inconsistent phylogenetic signal of this taxon in genome space. In contrast, the remaining 323 orthologs show broad phylogenetic agreement. This core set is comprised of key photosynthetic and ribosomal proteins. This observation suggests that the macromolecular interactions in complex protein structures (e.g., ribosomal proteins) and metabolic pathways (e.g., oxygenic photosynthesis) are strongly resistant to piecemeal HGT. Transfer was ultimately accomplished by wholesale incorporation of cyanobacteria into eukaryotic host cells, giving rise to primary photosynthetic endosymbionts that retained both photosynthetic genes and genes coding for their own ribosomes (40–44).

Author contributions: T.S. and P.G.F. designed research; T.S. performed research; T.S. and P.G.F. analyzed data; and T.S. and P.G.F. wrote the paper.

The authors declare no conflict of interest.

[†]Present address: Department of Ocean Sciences, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064.

[§]To whom correspondence should be addressed. E-mail: falko@marine.rutgers.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0711165105/DC1.

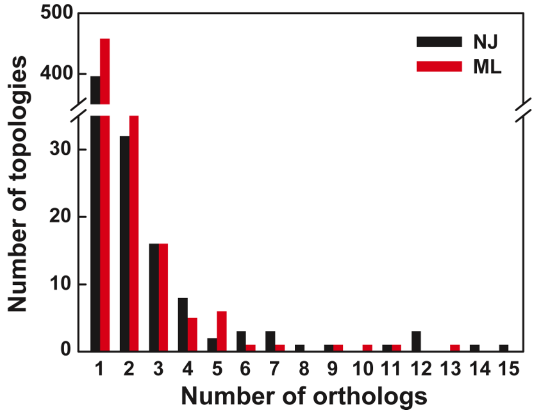

Fig. 1. Distribution of tree topologies among 682 sets of orthologs. Both NJ (black bars) and ML (red bars) tree topologies give similar distribution patterns. There is no unanimous support for a single topology; rather, most of the orthologs (58% and 67% for NJ and ML trees, respectively) appear as singletons that associate with unique topologies.

Results

Conserved Protein Families in Genomes of Cyanobacteria. Our pairwise genome comparison reveals a total of 682 orthologs common to all 13 genomes examined (SI Table 2). These orthologs constitute the core gene set and some define aspects of the genotype that are uniquely cyanobacterial. This core set represents only 8.9% (in the case of the largest genome, *Nostoc punctiforme*) to 39.7% (in the case of the smallest genome, *Prochlorococcus marinus* MED4) of the total number of protein-coding genes from each genome under study (see SI Table 1) but seems to account for all of the principal functions (SI Table 3). Our analysis leads to an estimate of the pool of orthologs similar to what has been identified from 10 cyanobacterial genomes (45), but nearly three times more than the number of cyanobacterial signature genes bioinformatically characterized by Martin *et al.* (46) and only 65% of the number of cyanobacterial clusters of orthologous groups (31). The discrepancy mostly results, in the case of the former, from a filtering procedure to remove homologs of chloroplasts and anoxygenic photoautotrophs and, in the case of the latter, from a less-stringent unidirectional BLAST hit scheme used. In addition, some of the incomplete genomes used in this study are still undergoing confirmation from the final assembly, hence equivalent genes may have been overlooked in some cases. It is highly possible that, because of the overly restrictive criterion (47), even without the use of any particular threshold (e.g., the default BLAST *e* value threshold is 10), the set of orthologs identified by the reciprocal top BLAST hit scheme would underestimate the actual number of orthologs (18).

Phylogenetic Incongruence Among Conserved Protein Families. Based on amino acid sequences, we built phylogenetic trees for each of the 682 orthologous protein families, using both neighbor joining (NJ) and maximum likelihood (ML) methods. Surprisingly, the frequency distribution of observed topologies fails to reveal a predominant, unanimous topology that represents a large number of orthologs (Fig. 1). In contrast, most of the orthologs (58% and 67% for NJ and ML, respectively) exhibit their own unique topologies. As a result, the maximum number of orthologs that share a particular topology accounts for only 1.9–2.1% of the orthologous datasets (Fig. 1).

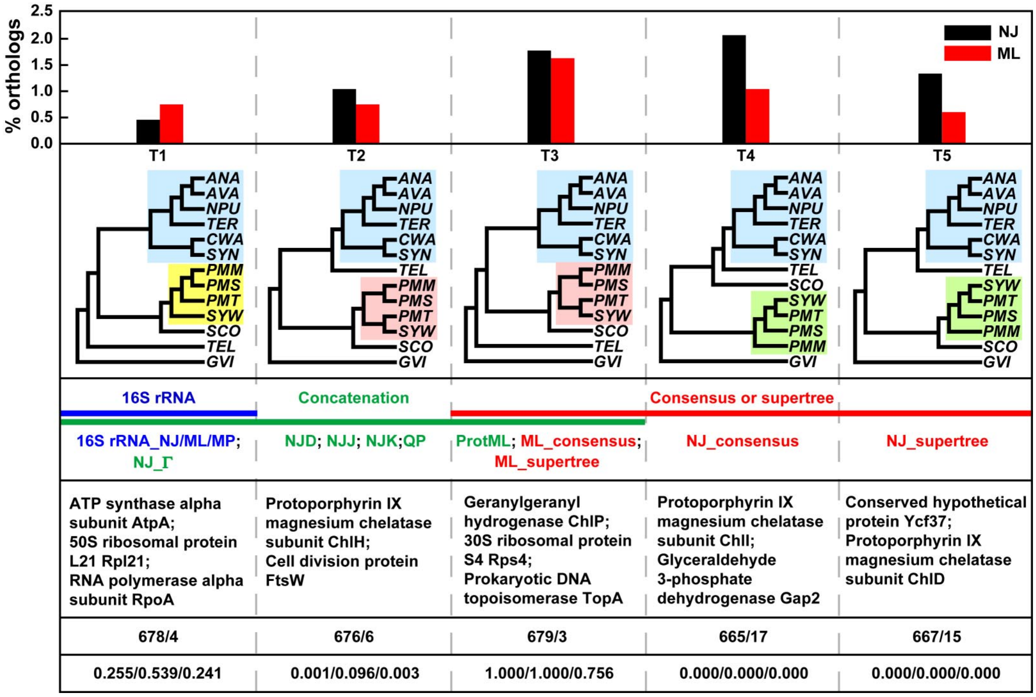
% orthologs
2.5
2.0
1.5
1.0
0.5
0.0
NJ
ML
T1
T2
T3
T4
T5
ANA
AVA
NPU
TER
CWA
SYN
PMM
PMS
PMT
SYW
SCO
TEL
GVI
16S rRNA
Concatenation
Consensus or supertree
16S rRNA_NJ/ML/MP; NJ_Γ
NJD; NJJ; NJK;QP
ProtML; ML_consensus; ML_supertree
NJ_consensus
NJ_supertree
ATP synthase alpha subunit AtpA; 50S ribosomal protein L21 Rpl21; RNA polymerase alpha subunit RpoA
Protoporphyrin IX magnesium chelatase subunit ChlH; Cell division protein FtsW
Geranylgeranyl hydrogenase ChlP; 30S ribosomal protein S4 Rps4; Prokaryotic DNA topoisomerase TopA
Protoporphyrin IX magnesium chelatase subunit ChlI; Glyceraldehyde 3-phosphate dehydrogenase Gap2
Conserved hypothetical protein Ycf37; Protoporphyrin IX magnesium chelatase subunit ChlD
678/4
676/6
679/3
665/17
667/15
0.255/0.539/0.241
0.001/0.096/0.003
1.000/1.000/0.756
0.000/0.000/0.000
0.000/0.000/0.000

Fig. 2. Representative backbone tree topologies. Phylogenetic trees were constructed by using both 16S rRNA gene and orthologous proteins through phylogenomic approaches (see *Materials and Methods* for details). Phylogenetic tree construction methods are highlighted with colored horizontal bars and text. Conserved monophyletic subgroups are shaded. Row one shows the proportion of orthologs giving a particular tree topology (NJ, black bar; ML, red bar). Also shown are examples of proteins corresponding to that topology. Row five indicates number of datasets accepting (*Left*) or rejecting (*Right*) a particular topology in a Shimodaira–Hasegawa (SH) (55) test (SI Fig. 7). Row six shows the evaluation of the five backbone topologies, using the concatenated 323-core-gene set (Fig. 3) through Kishino–Hasegawa (72) (*Left*), SH (55) (*Center*), and expected likelihood weight (73) (*Right*) tests, which infer a confidence tree set. *ANA*, *Anabaena* sp. PCC7120; *AVA*, *Anabaena variabilis* ATCC29413; *CWA*, *Crocosphaera watsonii* WH8501; *GVI*, *G. violaceus* PCC7421; *NPU*, *N. punctiforme* ATCC29133; *PMM*, *P. marinus* MED4; *PMT*, *P. marinus* MIT9313; *PMS*, *P. marinus* SS120; *SCO*, *S. elongatus* PCC7942; *SYW*, *Synechococcus* sp. WH8102; *SYN*, *Synechocystis* sp. PCC6803; *TEL*, *T. elongatus* BP-1; *TER*, *Trichodesmium erythraeum* IMS101.

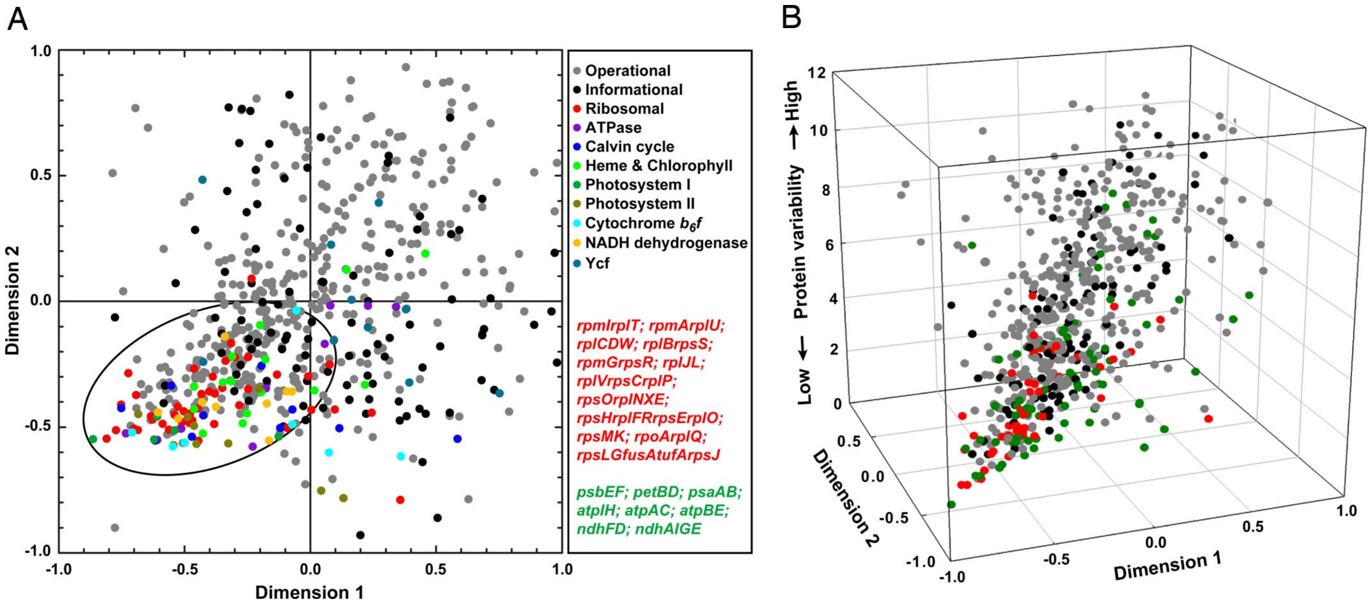

Fig. 3. PCoA of trees compared with topological distance. (*A*) Plot of the two first axes of the PCoA made from 628 ML trees. The other 54 genes are excluded as a result of axis demarcation. The same experiment with NJ trees gave very similar results. The ellipse depicts 323 orthologs in the densest region (the core) of the cloud that share a common phylogenetic signal, whereas trees present in the marginal area (the shell) are much more likely to be perturbed by horizontal transfers. Photosynthetic genes are color coded based on their respective pathways. Also shown are examples of conserved clusters of ribosomal (red text) and photosynthetic (green text) genes that are present in the core. (*B*) The PCoA plotted against the protein variability. Protein variability was measured by taking the total length of a corresponding tree as measured by total amino acid substitutions per site, divided by the number of sequences in the tree (48). The legend is the same except that photosynthetic genes are collectively designated as green dots.

Phylogenomic Reconcilement. To determine whether a common signal can be extracted from phylogenetic incongruence, we used the consensus, the supertree, and the reconstruction of phylogeny based on the concatenation of all of the 682 individual proteins. These approaches greatly resolve the topological incongruence, leading to five topologies as shown in Fig. 2. Specifically, the NJ and ML trees, using the concatenated sequences give three topologies in total (T1 and T2 for NJ; T2 and T3 for ML), one of which is in agreement with that of the 16S ribosomal RNA (rRNA) gene tree repeatedly obtained with NJ, ML, or maximum parsimony methods. The consensus and supertree built on the 682 individual NJ trees show two other topologies (T4 and T5), whereas those of ML trees reveal an identical topology to one of the concatenated ML trees. These five topologies are remarkably similar in that *Synechocystis* sp. PCC6803 and five diazotrophic species form a monophyletic clade, and that *Synechococcus* sp. WH8102 and three *Prochlorococcus* ecotypes form three different monophyletic clades. The notable conflicts concern the species *Synechococcus elongates* PCC7942 and the thermophilic *Thermosynechococcus elongates* BP-1, which tend to cluster at the base of the two major subgroups but form aberrant topologies.

Analyses of the fitness of a particular topology to the 682 sequence alignments (SI Figs. 6 and 7) indicate that almost all (97.5 to 99.6%) of the datasets support topologies T1-T5 at the 95% confidence level ($P = 0.95$), suggesting a lack of resolution of single gene phylogenies.

The Stable Core and the Variable Shell in Genome Space. To extrapolate evolutionary trajectories least affected by artificial paralogs, or genes potentially obtained by HGT, we calculated tree distances among all possible pairs of the orthologous sets. The pairwise distances were then used to conduct a principal coordinates analysis (PCoA). This results in a core set of 323 genes that share similar evolutionary histories (i.e., coevolving and rarely transferred) as opposed to the other 359 that exhibit divergent phylogenies (i.e., independently evolving and frequently transferred) (Fig. 3*A*). Ribosomal proteins are almost all grouped in the densest core, whereas the much sparser region of the cloud is formed largely by operational and nonribosomal informational genes. Additionally, the core is comprised of proteins constituting the scaffolds of the photosynthetic apparatus, and, at least partially, those that participate in ATP synthesis, chlorophyll biosynthesis, and the Calvin cycle. Based on an approach of "embedded quartets" that allows detection of HGT events with significantly improved resolution, Zhaxybayeva *et al.* (32) found that some of the major photosynthetic genes were subject to HGT and that the bias toward metabolic (operational) gene transfers was only detectable in transfers between cyanobacteria and other phyla. The apparent conflict between our analysis and that of Zhaxybayeva *et al.* is almost certainly due to methodological differences.

Using the sum of amino acid substitution per site in the tree as a "rough-and-ready" measure of protein variability (48), we compared the rates of evolution of genes in different functional categories. Both the PcoA (Fig. 3*B*) and the frequency distribution (Fig. 4) versus protein variability analyses reveal that ribosomal and photosynthetic genes are highly conserved, whereas the operational and other informational genes are strongly skewed toward high protein variability. A high degree of protein sequence conservation is significantly biased toward genes that are in the core and those encoded in genome-wide conserved gene clusters (Fig. 4), most likely because of the large number of ribosomal and photosynthetic genes present. This result suggests that the core gene set appears to have remained relatively stable throughout the evolutionary history of cyanobacteria, whereas genes in the shell are much more likely to be acquired via HGT.

We further reconstructed the phylogeny of the 13 genomes on the basis of a superalignment of 100,776 sites obtained via concatenating the 323 core proteins. We used three methods, all of which result in a tree having the same topology as that for the

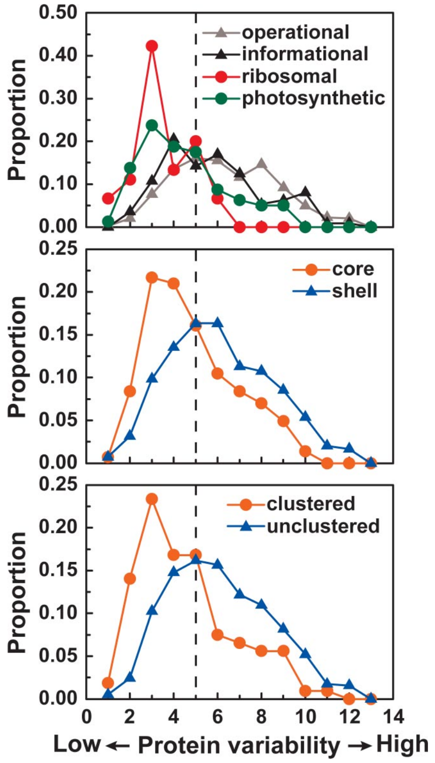

Fig. 4. Frequency distribution of genes belonging to designated categories within each 1.0 interval of protein variability. Protein variability was measured according to Rujan and Martin's method (48). Dashed lines denote the threshold that segregates the predominance of distribution of genes in different categories.

consensus, supertree, and concatenation of all of the 682 protein families (T3 in Fig. 2 and tree presented in Fig. 5). It differs only slightly from other tested topologies that are not rejected by most individual alignments but exhibits a superior likelihood support (Fig. 2). Intriguingly, all of the diazotrophic cyanobacteria fall within a distinct group, and their divergence from other nondiazotrophic taxa appears to occur much later after the origin of the clade based on rooting with *Gloeobacter violaceus* PCC7421, most possibly the earliest lineage within the radiation of cyanobacteria (49), and *T. elongatus* BP-1, a unicellular thermophilic cyanobacterium that inhabits hot springs. The early diazotrophic cyanobacteria appear to have been nonheterocystous, with heterocyst-forming lineages emerging later, possibly as a result of elevated levels of atmospheric O_2 (50).

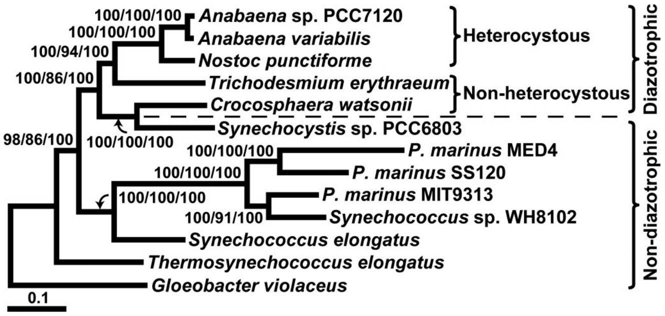

Fig. 5. Phylogenetic tree reconstructed based on the concatenation of the 323 core proteins. The topology shown agrees with the consensus topology of the 682 orthologs (T3 in Fig. 2) and is supported by almost all individual datasets (Fig. 2 and SI Fig. 7). Bootstrap probabilities estimated by NJ-Γ/QP/ProtML with 1,000 replications are shown for each internal branch. The scale bar refers to the number of amino acid substitutions per site. The dashed line designates the split between diazotrophic and nondiazotrophic taxa.

Discussion

Our analyses reveal an overwhelming phylogenetic discordance among the set of genes selected as likely orthologs (Fig. 1). Conflicting phylogenies can be a result of artifacts of phylogenetic reconstruction, HGT, or unrecognized paralogy. In our reciprocal best hit approach, we retained as orthologs those containing only one gene per species. Therefore, only orthologous replacement and hidden paralogy (i.e., differential loss of the two copies in two lineages) can occur in selected families. These two types of events are expected to be comparatively rare under application of the reciprocal hit criterion (51). Thus, phylogenetic incongruence is unlikely due to artifacts from a biased selection of orthologs. Furthermore, the overall phylogenetic disagreement does not seem to be caused by tree reconstruction or model selection artifacts because both NJ and ML individual trees unambiguously support plural partitions (Fig. 2). HGT is likely one of the most important driving forces that lead to the discrete evolutionary histories of the conserved protein families. Indeed, HGT has played an important role in the evolution of prokaryotic genomes (52–54). A hallmark of HGT is that the transferred genes often exhibit aberrant organismal distributions, which contrast with the relationships inferred from both the 16S rRNA gene tree and phylogenies of vertically inherited individual protein-coding genes. But how can this superficially random gene transfer event explain the conserved nature of many of the key genes that comprise the functional core across all cyanobacterial taxa?

Although phylogenomic approaches are capable of capturing the consensus or frequent partitions that silhouette the trend in genome evolution, they may not necessarily guarantee the paucity of a conflicting phylogenetic signal in genome space. The plural support for the consensus/supertree/concatenation topologies indicates that the five top topologies are not significantly different from each other; that is, >90% of the datasets do not discriminate among the topologies (Fig. 2). Do the consensus/supertree/concatenation trees accurately reflect organismal history, or, on the contrary, do they blur the vertical inheritance signal by incorporating potential HGTs? There is a large margin of uncertainty. Part of the uncertainty may be due to the strength of the Shimodaira–Hasegawa (SH) test (55), especially when examining the accuracy of similar topologies. Indeed, the SH test was based on the evaluation of only 15 of a total of 13,749,310,575 possible unrooted tree topologies for 13 species (SI Figs. 6 and 7). Although the majority of the possible topologies would not be supported by any dataset, the selection of a limited number of trees may have biased the analyses. But part of the uncertainty can also be attributed to the data, most notably genes that are subjected to HGT and homologous recombination between closely related species. This is even more pronounced in the PCoA, which demonstrates clearly that ≈53% of the orthologs are subject to HGT complicating/diluting the vertical inheritance signal within the cyanobacterial phylum (Fig. 3).

Our results reveal that both photosynthetic and ribosomal genes share similar evolutionary histories and belong to the cyanobacterial genome core (Fig. 3). This finding of limited HGT in proteins with extraordinarily conserved primary structure is consistent with the complexity hypothesis; that is, genes coding for large complex systems that have more macromolecular interactions are less subject to HGT than genes coding for small assemblies of a few gene products (37). Translation in prokaryotes requires coordinated assembly of at least 100 gene products, including ribosomal small and large subunits, which interact with 5S, 16S, and 23S rRNA; numerous tRNA and mRNA; initiation and termination factors; ions; etc. Similarly, the oxygenic photosynthetic apparatus needs an investment of a huge number of proteins, pigments, cofactors, and trace elements for effective functionality. All of the components required in both machines are presumed to be present in a potential host, and the complexity of gene product interactions is a significant factor that restricts their successful HGT rates relative to the high HGT rates observed for operational genes. It is noteworthy, however, that not all photosynthetic genes are significantly resistant to HGT. Photosynthetic genes outside the core, including genes encoding proteins whose functions are yet to be confirmed (*ycf*) and those that tend to form the periphery (i.e., supplemental "add ons") of the photosynthetic scaffold, may be less critical to biophysical interactions and hence more readily transferred between cyanobacteria compared with the large integral membrane proteins that belong to the functional core of the photosynthetic apparatus. The impact of membrane protein interactions appears to continue to limit the transfer of core photosynthetic genes to the nucleus in higher plants and algae, even after the endosymbiosis event. This is supported by the fact that proteins whose genes are most resistant to transfer to the nucleus constitute the functional physical core of the photosynthetic apparatus (56). In contrast, some genes that belong to the peripheral scaffold of photosynthesis, for example, the *petC* and *psbO*, have been transferred to the nucleus as did thousands of other easily transferred cyanobacterial genes (57). A striking feature of these HGT-resistant components is that they tend to cluster together in a putative operon, containing two to four genes, that is conserved among all cyanobacteria and plastids (58). The mechanism underpinning the conservation of gene order is unknown. It could be an advantage in gene expression for coordinated transcription of the genes and assembly of the subunits of a multimetric complex. However, it is more likely that protein–protein, protein–cofactor, and protein–membrane interactions exert a strong selection pressure to maintain synteny to reduce the chance of being perturbed by HGT via genetic recombination (58, 59). These interactions not only govern the conservation of gene order, but also the tempo of evolution of these genes (Figs. 3 and 4). There seems to be a link between the tempo of evolution and resistance to HGT; the probability of HGT increases with decreased conservation of amino acid sequence in a gene product (48). Moreover, the complexity of oxygenic photosynthetic machinery makes it difficult to transfer components piecemeal to nonphotosynthetic prokaryotes. Indeed, operon splitting of the photosynthetic apparatus requires many independent transfers of noncontiguous operons. Although large-scale HGT among photosynthetic prokaryotes (30) may suggest a complex nonlinear process of evolution that results in a mosaic structure of photosynthetic pathway (60), transfer of the key photosynthetic genes are very rare (33, 34). Transfer of this key pathway was only achieved by wholesale incorporation of cyanobacteria into eukaryotic host cells (40–44).

Phylogenetic analysis of the cyanobacterial genome core strongly suggests that the last common ancestor of extant cyanobacteria was incapable of N_2 fixation (Fig. 5). This metabolic pathway appears to have been acquired via HGT much later after the origin of this clade, possibly as a result of a "fixed nitrogen crisis" in the late Archean and early Proterozoic eons (61). In this scenario, the accumulation of a small concentration of oxygen (resulting from oxygenic photosynthesis) would have led to massive denitrification of the upper ocean with a concomitant loss of fixed inorganic nitrogen for growth of marine photoautotrophs. This process created an evolutionary bottleneck, which potentially selected for the stable transfer of the *nif* operon from a (presumably) heterotrophic prokaryote to cyanobacteria. It should be noted that selection of the specific nitrogenase was probably not related to metal availability, because Fe was abundant under these mildly oxidizing conditions (62). It was only after the "great oxidation event," ≈2.3 Ga (1–3) and later, that Fe would become limiting, leading to the sequential selection for V- and Mo-containing nitrogenases. Thus, under the mildly oxidizing conditions that prevailed in the late Archean to early Proterozoic, Fe-based nitrogenases would have been naturally selected within the archaea and subsequently transferred to a large group of bacteria via HGT (63). In the late Proterozoic and throughout the Phanerozoic, oxygenic photosynthesis ultimately led to precipitation of insoluble oxidized (ferric) Fe, thereby making this element a major factor limiting N_2 fixation in the ocean, a condition that appears to continue to limit the productivity in the contemporary ocean (64). Macrogenomic features of cyanobacteria potentially provide clues regarding the ability of these organisms to acquire nitrogenase and other genes. For example, all cyanobacterial diazotrophs have significantly larger genomes than their nonfixing counterparts (SI Table 1), suggesting that the genomes may have been exposed to frequent HGT and are more competent to incorporate genes.

Materials and Methods

Gene Family Selection. We performed all-against-all BLAST (65) comparisons of protein sequences for all possible pairs of the 13 genomes of cyanobacteria (SI Table 1), using an *e* value of 10^{-4} as a lower limit cutoff, and reciprocal genome-specific best hits were identified. A total number of 682 protein families consisting of one gene per genome were retrieved and assigned to functional categories according to those defined for the cluster of orthologous groups (66).

Alignments and Tree Construction. Protein sequences were aligned with ClustalW (67), followed by selecting unambiguous parts of the alignments excluding all gap sites. ML trees were computed with PHYML (68), using the JTT model of substitution and the Gamma (Γ)-based method for correcting the rate heterogeneity among sites. Neighbor joining (NJ) trees were constructed by using the distance matrix provided by TREE-PUZZLE (69) under a Γ-based model of substitution (alpha parameter estimated, eight Γ rate categories) and bootstrapped by using SEQBOOT and CONSENSE from PHYLIP (70). See *SI Methods* for concatenation, consensus, and supertree reconstruction.

Comparisons Among Trees. Trees were compared with Treedist program in PHYLIP using the branch score distance of Kuhner and Felsenstein (71) to generate an $n \times n$ distance matrix where n is the number of trees. Principal coordinates analysis (PCoA) was then performed with the multidimensional scaling procedure in SAS software, Version 8.2 (SAS Institute). PCoA allowed us to embed the n trees in a space of up to $n - 1$ dimensions. By plotting the objects (the trees) along the most significant two first dimensions, the major trends and groupings in the data can be visualized graphically.

For each of the 682 alignments, a comparison of the likelihood of the best topology with that of the candidate topologies (SI Figs. 6 and 7) was performed with the SH test (55) implemented in TREE-PUZZLE. Similarly, a comparison of the five backbone topologies (Fig. 2) was conducted with SH, Kishino–Hasegawa (72), and expected likelihood weight (73) tests, using the concatenated 323-core-gene set. All tests were conducted by using a 5% significance level and were performed by using the resampling of estimated log-likelihood method with 1,000 replications.

ACKNOWLEDGMENT. We thank Robert Blankenship and J. Peter Gogarten for critical reviews of the manuscript; William Martin, Colomban de Vargas, and Kay Bidle for stimulating discussions and insightful comments; and Lin Jiang for assistance with the principal coordinates analysis. This work was supported by the grants from the Agouron Foundation and National Aeronautics and Space Administration Exobiology Program Grant NX7AK14G (to P.G.F.).

1. Bekker A, *et al.* (2004) Dating the rise of atmospheric oxygen. *Nature* 427:117–120.
2. Anbar AD, *et al.* (2007) A whiff of oxygen before the great oxidation event? *Science* 317:1903–1906.
3. Kaufman AJ, *et al.* (2007) Late Archean biospheric oxygenation and atmospheric evolution. *Science* 317:1900–1903.
4. Knoll AH, Summons RE, Waldbauer JR, Zumberge JE (2007) in *Evolution of Primary Producers in the Sea*, eds Falkowski PG, Knoll AH (Academic, New York), pp 133–163.
5. Blankenship RE, Hartman H (1998) The origin and evolution of oxygenic photosynthesis. *Trends Biochem Sci* 23:94–97.

6. Falkowski PG, *et al.* (2005) The rise of oxygen over the past 205 million years and the evolution of large placental mammals. *Science* 309:2202–2204.
7. Raymond J, Segre D (2006) The effect of oxygen on biochemical networks and the evolution of complex life. *Science* 311:1764–1767.
8. Whitton BA, Potts M (2000) in *The Ecology of Cyanobacteria*, eds Whitton BA, Potts M (Kluwer Academic, Dordrecht, The Netherlands), pp 1–11.
9. Waterbury JB, Watson SW, Guillard RRL, Brand LE (1979) Widespread occurrence of a unicellular, marine, planktonic, cyanobacterium. *Nature* 277:293–294.
10. Chisholm SW, *et al.* (1988) A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* 334:340–343.
11. Capone DG, Zehr JP, Paerl H, Bergman B, Carpenter EJ (1997) *Trichodesmium*, a globally significant marine cyanobacterium. *Science* 276:1221–1229.
12. Zehr JP, *et al.* (2001) Unicellular cyanobacteria fix N_2 in the subtropical North Pacific Ocean. *Nature* 412:635–638.
13. Karl D, *et al.* (2002) Nitrogen fixation in the world's oceans. *Biogeochemistry* 57/58: 47–98.
14. Koonin EV, Aravind L, Kondrashov AS (2000) The impact of comparative genomics on our understanding of evolution. *Cell* 101:573–576.
15. Palenik B, *et al.* (2003) The genome of a motile marine *Synechococcus*. *Nature* 424:1037–1042.
16. Rocap G, *et al.* (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424:1042–1047.
17. Hess WR (2004) Genome analysis of marine photosynthetic microbes and their global role. *Current Opin Biotech* 15:191–198.
18. Dufresne A, Garczarek L, Partensky F (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol* 6:R14.
19. Olson JM, Pierson BK (1987) Origin and evolution of photosynthetic reaction centers. *Orig Life* 17:419–430.
20. Blankenship RE (1992) Origin and early evolution of photosynthesis. *Photosynth Res* 33:91–111.
21. Vermaas WFJ (1994) Evolution of heliobacteria: Implications for photosynthetic reaction center complexes. *Photosynth Res* 41:285–294.
22. Xiong J, Inoue K, Bauer CE (1998) Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from *Heliobacillus mobilis*. *Proc Natl Acad Sci USA* 95:14851–14856.
23. Xiong J, Fischer WM, Inoue K, Nakahara M, Bauer CE (2000) Molecular evidence for the early evolution of photosynthesis. *Science* 289:1724–1730.
24. Baymann F, Brugna M, Muhlenhoff U, Nitschke W (2001) Daddy, where did (PS)I come from? *Biochim Biophys Acta Bioenerget* 1507:291–310.
25. Gupta RS (2003) Evolutionary relationships among photosynthetic bacteria. *Photosynth Res* 76:173–183.
26. Rutherford AW, Faller P (2003) Photosystem II: Evolutionary perspectives. *Philos Trans R Soc Lond B* 358:245–253.
27. Olson JM, Blankenship RE (2004) Thinking about the evolution of photosynthesis. *Photosynth Res* 80:373–386.
28. Sadekar S, Raymond J, Blankenship RE (2006) Conservation of distantly related membrane proteins: Photosynthetic reaction centers share a common structural core. *Mol Biol Evol* 23:2001–2007.
29. Xiong J (2006) Photosynthesis: What color was its origin? *Genome Biol* 7:245.
30. Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE (2002) Whole-genome analysis of photosynthetic prokaryotes. *Science* 298:1616–1620.
31. Mulkidjanian AY, *et al.* (2006) The cyanobacterial genome core and the origin of photosynthesis. *Proc Natl Acad Sci USA* 103:13126–13131.
32. 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.
33. Lindell D, *et al.* (2004) Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. *Proc Natl Acad Sci USA* 101:11013–11018.
34. Millard A, Clokie MRJ, Shub DA, Mann NH (2004) Genetic organization of the *psbAD* region in phages infecting marine *Synechococcus* strains. *Proc Natl Acad Sci USA* 101:11007–11012.
35. Coleman ML, *et al.* (2006) Genomic islands and the ecology and evolution of *Prochlorococcus*. *Science* 311:1768–1770.
36. Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci USA* 95:6239–6244.
37. Jain R, Rivera MC, Lake JA (1999) Horizontal gene transfer among genomes: The complexity hypothesis. *Proc Natl Acad Sci USA* 96:3801–3806.
38. Daubin V, Gouy M, Perriere G (2002) A phylogenomic approach to bacterial phylogeny: Evidence of a core of genes sharing a common history. *Genome Res* 12:1080–1090.
39. Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokaryotes: The case of the gamma-proteobacteria. *PLoS Biol* 1:101–109.
40. Martin W, *et al.* (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393:162–165.
41. Bhattacharya D, Medlin L (1998) Algal phylogeny and the origin of land plants. *Plant Physiol* 116:9–15.
42. Delwiche CF (1999) Tracing the thread of plastid diversity through the tapestry of life. *Am Nat* 154:S164–S177.
43. Grzebyk D, Schofield O, Vetriani C, Falkowski PG (2003) The mesozoic radiation of eukaryotic algae: The portable plastid hypothesis. *J Phycol* 39:259–267.
44. Falkowski PG, Knoll AH (2007) *Evolution of Primary Producers in the Sea* (Academic, New York).
45. Zhaxybayeva O, Lapierre P, Gogarten JP (2004) Genome mosaicism and organismal lineages. *Trends Genet* 20:254–260.
46. Martin K, *et al.* (2003) Cyanobacterial signature genes. *Photosynth Res* 75:211–221.
47. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. *Science* 278:631–637.
48. Rujan T, Martin W (2001) How many genes in *Arabidopsis* come from cyanobacteria? An estimate from 386 protein phylogenies. *Trends Genet* 17:113–120.
49. Nakamura Y, *et al.* (2003) Complete genome structure of *Gloeobacter violaceus* PCC 7421, a cyanobacterium that lacks thylakoids. *DNA Res* 10:137–145.
50. Berman-Frank I, Lundgren P, Falkowski P (2003) Nitrogen fixation and photosynthetic oxygen evolution in cyanobacteria. *Res Microbiol* 154:157–164.
51. Zhaxybayeva O, Gogarten JP (2003) An improved probability mapping approach to assess genome mosaicism. *BMC Genomics* 4:37.
52. Doolittle WF (1999) Lateral genomics. *Trends Cell Biol* 9:M5–M8.
53. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405:299–304.
54. Gogarten JP, Doolittle WF, Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Mol Biol Evol* 19:2226–2238.
55. Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol Biol Evol* 16:1114–1116.
56. Race HL, Herrmann RG, Martin W (1999) Why have organelles retained genomes? *Trends Genet* 15:364–370.
57. 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 USA* 99:12246–12251.
58. Shi T, Bibby TS, Jiang L, Irwin AJ, Falkowski PG (2005) Protein interactions limit the rate of evolution of photosynthetic genes in cyanobacteria. *Mol Biol Evol* 22:2179–2189.
59. Dandekar T, Snel B, Huynen M, Bork P (1998) Conservation of gene order: A fingerprint of proteins that physically interact. *Trends Biochem Sci* 23:324–328.
60. Blankenship RE (2001) Molecular evidence for the evolution of photosynthesis. *Trends Plants Sci* 6:4–6.
61. Fennel K, Follows M, Falkowski PG (2005) The co-evolution of the nitrogen, carbon and oxygen cycles in the Proterozoic ocean. *Am J Sci* 305:526–545.
62. Anbar AD, Knoll AH (2002) Proterozoic ocean chemistry and evolution: A bioinorganic bridge? *Science* 297:1137–1142.
63. Raymond J, Siefert JL, Staples CR, Blankenship RE (2004) The natural history of nitrogen fixation. *Mol Biol Evol* 21:541–554.
64. Falkowski PG (1997) Evolution of the nitrogen cycle and its influence on the biological sequestration of CO_2 in the ocean. *Nature* 387:272–275.
65. Altschul SF, *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucl Acids Res* 25:3389–3402.
66. Tatusov RL, *et al.* (2001) The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucl Acids Res* 29:22–28.
67. Thompson J, Higgins D, Gibson T (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl Acids Res* 22:4673–4680.
68. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
69. Strimmer K, von Haeseler A (1996) Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Mol Biol Evol* 13:964–969.
70. Felsenstein J (2002) PHYLIP (Phylogeny Inference Package), Version 3.6. (Department of Genetics, University of Washington, Seattle, WA).
71. Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol* 11:459–468.
72. Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J Mol Evol* 29:170–179.
73. Strimmer K, Rambaut A (2002) Inferring confidence sets of possibly misspecified gene trees. *Proc R Soc Lond B* 269:137–142.