Genes of Cyanobacterial Origin in Plant Nuclear Genomes Point to a Heterocyst-Forming Plastid Ancestor

Oliver Deusch, Giddy Landan,† Mayo Roettger,* Nicole Gruenheit,* Klaus V. Kowallik,* John F. Allen,‡ William Martin,* and Tal Dagan**

*Institut für Botanik III, Heinrich-Heine Universität Düsseldorf, Universitätsstrasse 1, Düsseldorf, Germany; †Department of Biology and Biochemistry, University of Houston; and ‡School of Biological and Chemical Sciences, Queen Mary, University of London, London, United Kingdom

Plastids are descended from a cyanobacterial symbiosis which occurred over 1.2 billion years ago. During the course of endosymbiosis, most genes were lost from the cyanobacterium's genome and many were relocated to the host nucleus through endosymbiotic gene transfer (EGT). The issue of how many genes were acquired through EGT in different plant lineages is unresolved. Here, we report the genome-wide frequency of gene acquisitions from cyanobacteria in 4 photosynthetic eukaryotes—*Arabidopsis*, rice, *Chlamydomonas*, and the red alga *Cyanidioschyzon*—by comparision of the 83,138 proteins encoded in their genomes with 851,607 proteins encoded in 9 sequenced cyanobacterial genomes, 215 other reference prokaryotic genomes, and 13 reference eukaryotic genomes. The analyses entail 11,569 phylogenies inferred with both maximum likelihood and Neighbor-Joining approaches. Because each phylogenetic result is dependent not only upon the reconstruction method but also upon the site patterns in the underlying alignment, we investigated how the reliability of site pattern generation via alignment affects our results: if the site patterns in an alignment differ depending upon the order in which amino acids are introduced into multiple sequence alignment—N- to C-terminal versus C- to N-terminal—then the phylogenetic result is likely to be artifactual. Excluding unreliable alignments by this means, we obtain a conservative estimate, wherein about 14% of the proteins examined in each plant genome indicate a cyanobacterial origin for the corresponding nuclear gene, with higher proportions (17–25%) observed among the more reliable alignments. The identification of cyanobacterial genes in plant genomes affords access to an important question: From which type of cyanobacterium did the ancestor of plastids arise? Among the 9 cyanobacterial genomes sampled, *Nostoc* sp. PCC7120 and *Anabaena variabilis* ATCC29143 were found to harbor collections of genes which are—in terms of presence/absence and sequence similarity—more like those possessed by the plastid ancestor than those of the other 7 cyanobacterial genomes sampled here. This suggests that the ancestor of plastids might have been an organism more similar to filamentous, heterocyst-forming (nitrogen-fixing) representatives of section IV recognized in Stanier's cyanobacterial classification. Members of section IV are very common partners in contemporary symbiotic associations involving endosymbiotic cyanobacteria, which generally provide nitrogen to their host, consistent with suggestions that fixed nitrogen supplied by the endosymbiont might have played an important role during the origin of plastids.

Introduction

The idea that plastids arose from cyanobacteria through endosymbiosis is old and was scorned for many decades (Mereschkowsky 1905) but is no longer debated (Douglas 1998; Delwiche 1999; Matsuzaki et al. 2004; McFadden and van Dooren 2004; Archibald 2006). However, the issue of how, exactly, cyanobacteria contributed to plant genome evolution is still unresloved, as is the issue of what kind of cyanobacterium participated in plastid origin (Sato 2006). As an evolutionary mechanism, endosymbiosis differs substantially from point mutation because a genome's worth of new genetic material is the currency unit of genetic change, rather than a succession of fixed nucleotide polymorphisms or duplicated genes with mutated promoters. Accordingly, the evolutionary transition that transformed an oxygen-producing, prokaryotic endosymbiont into the diverse spectrum of photosynthetic organelles found among modern photosynthetic eukaryotes involved both inheritance from the prokaryote and invention by the eukaryote (Sato 2001).

From the standpoint of cell function and physiology at plastid origin, the most important inheritance was photosynthesis itself, which conferred the photolithoautotrophic lifestyle upon a heterotrophic host, whereas the most important invention was the protein import machinery (McFadden and van Dooren 2004; Soll and Schleiff 2004), which permitted the endosymbiont to import nuclear-encoded proteins. The invention of a protein import apparatus allowed the ancestral plastid to relinquish genes to the nucleus over evolutionary time without relinquishing those biochemical functions which are germane to the cyanobacterial lifestyle (Allen 2003). Protein import thus marked a crucial turning point in the evolutionary process that gave rise to plastids. Prior to the invention of protein import, the cyanobacterial endosymbiont was able to donate genes to its host, but unable to import the encoded products, such that, from the host's standpoint, the symbiont served as a virtually inexhaustible source of new and divergent genes for functions in various cell compartments (Martin and Schnarrenberger 1997; Martin and Herrmann 1998; Allen 2003; Bogorad 2008). Once the protein import apparatus had evolved, products of nuclear genes of cyanobacterial origin could be targeted to the plastid compartment, allowing the endosymbiont-encoded copies of such genes to escape purifying selection, and thus undergo pseudogenization and loss. The process just described, endosymbiotic gene transfer (EGT; Martin et al. 1993), resulted in the genetic integration of the endosymbiont with its host (Timmis et al. 2004; Sato 2006; Reyes-Prieto et al. 2007), accompanied by the transition of the former into a double membrane–bound organelle of the latter and by the origin of the eukaryotic lineage that

Key words: endosymbiosis, phylogenomics, multiple sequence alignment, plastid origin, nitrogen.

E-mail: tal.dagan@uni-duesseldorf.de.

Mol. Biol. Evol. 25(4):748–761. 2008
doi:10.1093/molbev/msn022
Advance Access publication January 24, 2008

possesses primary plastids, the archaeplastida (Adl et al. 2005). In the present work, we aim to address 2 questions.

First, we wish to address the quantitative scope of EGT from plastids using whole genome data. Modern plastids are estimated to contain anywhere from about 2,100 to 4,800 different proteins (Richly and Leister 2004), whereas plastid genomes encode only 60–200 proteins in various photosynthetic lineages (Timmis et al. 2004). Plastids thus contain roughly as many proteins as their free-living cyanobacterial cousins but have retained only a handful of the corresponding genes (Allen and Raven 1996). Various and differing estimates for the total number of genes that were acquired by plants from cyanobacteria have been reported. A phylogenomic investigation of *Arabidopsis* including 3 cyanobacterial genomes but only 1 reference eukaryotic genome suggested that about 18% of *Arabidopsis* proteins are cyanobacterial acquisitions (Martin et al. 2002), whereas an EST-based analysis of *Cyanophora paradoxa* indicated that about 11% of the proteins in that genome are cyanobacterial acquisitions (Archibald 2006; Reyes-Prieto et al. 2006).

Those differences are likely to have methodological causes, and it is well recognized that the reliability of phylogenetic inference with highly divergent proteins can affect such estimates (Martin et al. 2002; Reyes-Prieto et al. 2006). Comparisons of plant and cyanobacterial proteins are replete with alignments of highly divergent sequences because the origin of plastids dates back at least 1.2 billion years (Butterfield 2000; Yoon et al. 2004). In order to take the influence of highly divergent alignments into account in this phylogenomic study, we have made use of recent findings showing that phylogenetic results using real data differ markedly in comparisons of alignments generated by reading sequences from N- to C-terminus (the default in all current alignment programs, the "heads" orientation) to those generated using the same program but reading sequences in from C-terminus to N-terminus (the "tails" orientation) (Landan and Graur 2007). The underlying reasoning is straightforward: If a phylogenetic result is contingent upon the order in which amino acids are introduced into multiple sequence alignment—N- to C-terminal versus C- to N-terminal or heads versus tails—then it is likely to be an artifact of phylogeny inference and one that is rooted at the alignment step, where the site patterns for phylogenetic inference are generated. The reproducibility of site pattern generation as a function of heads versus tails alignment provides a criterion for separating phylogenetic wheat from chaff. Its utility to refine estimates for the number of genes that plants acquired from cyanobacteria is explored in a phylogenomic analysis of proteins encoded by *Arabidopsis*, rice, *Chlamydomonas*, and the red alga *Cyanidioschyzon* in comparison with 9 cyanobacterial and 228 other reference genomes.

Second, we wish to investigate the nature of the plastid ancestor, specifically which genes it contained and to which lineages of modern cyanobacteria it was most closely related as inferred from cyanobacterial genes present in plant nuclear genomes. Previous studies using alignments of single loci (Morden and Golden 1989; Turner et al. 1999; Marin et al. 2005) or concatenated genes encoded in plastid DNA in comparison to homologues from cyanobacteria (Rodriguez-Ezpeleta et al. 2005) have been inconclusive, although Sato (2006) suggested that the *Anabaena–Synechocystis* lineage might be closer to the ancestor of plastids than other cyanobacteria sampled in that study. However, the individual phylogenies of plastid-encoded genes can differ significantly even though they are related by the same evolutionary process (Martin et al. 1998; Lockhart et al. 1999), a prime example of the limitations involved in phylogenetic inference as one goes farther back in time (White et al. 2007). But among cyanobacteria the situation is worse because in addition to the problem of phylogenetic error, they can and do exchange genes both among themselves and with other bacterial groups via lateral gene transfer (LGT) (Raymond et al. 2002; Zhaxybayeva et al. 2006), such that concatenation will tend to mix phylogenetic signals, rather than amplify them (Bapteste et al. 2007), with bootstrap or Bayesian support values offering no guide to confidence, because large alignments tend to yield support values close to unity, regardless of whether or not the topology is correct (Phillips et al. 2004).

As it relates to the origin of genes that eukaryotes acquired from the ancestor of plastids, LGT among prokaryotes means that—even though all available evidence points to a single origin of plastids (McFadden and van Dooren 2004; Archibald 2006) and notwithstanding recent debate as to how to define the term plastid (Bodyl et al. 2007; Larkum et al. 2007)—plant genomes harbor a discrete sample of the collection of genes present in the genome of the plastid ancestor. However, neither that gene sample nor the collection of genes present in the free-living plastid ancestor is likely to have persisted in its original, contiguous state in any modern cyanobacterial chromosome (Martin 1999) because of LGT among cyanobacteria (Zhaxybayeva et al. 2006) and between cyanobacteria and other prokaryotes (Raymond et al. 2002) since the origin of plastids. The same problem exists with regard to finding the ancestor of mitochondria among α-proteobacteria (Esser et al. 2007). Despite this complication that LGT introduces with regard to identifying the "lineage" of cyanobacteria from which plastids arose (for a discussion of modern concepts regarding prokaryotic lineages, see Doolittle and Bapteste 2007), one can still address the question of which modern cyanobacterial genomes possess the highest frequencies of genes with strongest sequence similarity to their nuclear encoded homologues in photosynthetic eukaryotes, which we do, using 9 sequenced cyanobacterial genomes.

Materials and Methods

Data

The nuclear proteomes of *Arabidopsis thaliana* version January 2006 and *Oryza sativa* version May 2006 were downloaded from RefSeq database (Pruitt et al. 2005). The nuclear proteomes of *Cyanidioschyzon merolae* (Matsuzaki et al. 2004) version February 2005 and *Chlamydomonas reinhardtii* (Merchant et al. 2007) version 2.0 were downloaded from their genome projects. Multiple copy proteins were condensed into a single entry. Homologous proteins within prokaryotes and nonphotosynthetic eukaryotes were searched as follows. The proteins of the photosynthetic

eukaryotes were Blasted (Altschul et al. 1997) to a data set including 200 eubacteria, 24 archaebacteria, and 13 nonphotosynthetic eukaryotes (supplementary table S1, Supplementary Material online). The Blast hits were filtered for hits of *E* value $\leq 10^{-10}$ and $\geq 25\%$ amino acid identities and ranked by the percent of identities multiplied by the ratio of the query length and the Blast pairwise alignment length. In order to render phylogenetic analysis using maximum likelihood (ML) computationally tractable for over >11,000 phylogenies, the number of operational taxonomical units per alignment had to be restricted, hence only the first 3 hits from each phylum were selected for sequence alignment. Because we are addressing the question of which proteins are most closely related to the plant homologues, rather than the question of how proteins from all lineages are related to one another, the selection of the 3 best representatives from each phylum should not influence our estimates. Proteins from the cyanobacterial genomes were investigated (*Anabaena variabilis* ATCC29413, *Gloeobacter violaceus* PCC7421, *Nostoc* sp. PCC7120, *Prochlorococcus marinus* MIT9313, *Synechococcus* sp. CC9605, *Synechococcus elongatus* PCC7942, *Synechococcus* sp. WH8102, *Synechocystis* sp. PCC6803, and *Thermosynechococcus elongatus* BP-1), and other genomes were obtained from GenBank.

Sequence Alignment

Each protein was aligned with its homologues using Muscle (Edgar 2004) with a maximum of 16 iterations. For the tails alignment, the protein sequences were reversed using a PERL script and were aligned again with Muscle using the same parameters. For the comparison between the heads and tails alignments, the tails alignment is reversed and each amino acid is replaced by its serial number in the protein sequence. Gaps in the alignment are converted to "0." An identical heads and tails column includes exactly the same amino acids from each sequence in both alignments. The columns score (CS) is calculated as the proportion of columns in the heads alignment that had a matching column in the tails alignment. An identical heads and tails pair is a pair of amino acids in 2 different sequences that are aligned together in both alignments. The sum-of-pairs score (SPS) is calculated as the proportion of pairs in the heads alignment that had a matching pair in the tails alignment.

Phylogenetic Trees

The heads and tails alignments were used to reconstruct phylogenetic trees with Phyml (Guindon and Gascuel 2003) using the ungapped positions only and the Jones-Taylor-Thornton (JTT) model (Jones et al. 1992) assuming rate variation across sites according to a Gamma distribution with 8 rate categories with the alpha parameter estimated from the data and invariable sites taken into account and with Neighbor-Joining (NJ) (Saitou and Nei 1987) using JTT distances. The heads and tails trees were compared by counting the identical splits between the trees using Treedist of PHYLIP (Felsenstein 2005). The phylogenetic partitions score (PPS) is the proportion of tree splits (branches) that were reconstructed identically from the heads and tails alignments. The nearest neighbors were identified as taxa contained in the smallest clade that included the homologue of the photosynthetic eukaryote.

Pairwise Analysis

Reciprocal best Blast hits (BBHs) of *Arabidopsis* genes and rice or alternatively *Chlamydomonas* genes were defined as orthologous pairs. Each pair of genes was aligned using ClustalW (Thompson et al. 1994), and protein distances were calculated with Protdist (Felsenstein 2005) using the JTT substitution matrix.

Nuclear Plastid DNA Rescreening

To identify nuclear plastid DNAs (NUPTs) (Richly and Leister 2004) protein-coding sequences from the corresponding chloroplast genomes were Blasted to the RefSeq database version January 2008. Proteins with 100% identical amino acids over the total length were scored as probable NUPTs.

Functional Classification

Proteins were classified into functional categories according to their BBH above 25% amino acid identities found in Swiss-Prot database (Boeckmann et al. 2003). The function of each protein in Swiss-Prot is described by one or more keywords. We manually classified Swiss-Prot keywords into functional categories (the list is available upon request). Each Swiss-Prot protein was assigned to a functional category according to the most frequent category of its keywords.

Results and Discussion

In order to identify genes of cyanobacterial origin in the genomes of 4 photosynthetic eukaryotes containing a primary plastid, we compared the nonredundant set of 83,138 proteins encoded in their genomes to a data set of 851,607 proteins from 9 sequenced cyanobacterial genomes, 13 reference eukaryotic genomes, and 215 reference prokaryotic genomes using Blast. Alignments were constructed for those query sequences that detected at least one homologue in cyanobacteria and homologues in at least 2 other search phyla (supplementary table S1, Supplementary Material online) at an *E*-value threshold 10^{-10} and 25% amino acid identity in the pairwise Blast alignment. Within each phylum, hits were ranked by their similarity and the ratio of the hit length to the query length. The first 3 hits from each phylum were selected for alignment and phylogenetic analysis. A summary of the distribution of hits at this threshold for each phylum is shown in table 1. For each query, we thus obtained a set of sequences that included the best cyanobacterial matches and the best matches from at least 2 other phyla to address the cyanobacterial ancestry of the plant homologue via phylogenetic inference. For those

Table 1
The number of Blast hits for each phylum and the number of trees in which the phylum was included

		Arabidopsis		Rice		*Chlamydomonas*		*Cyanidioschyzon*	
	Phylum (number of genomes)	Hits	Trees	Hits	Trees	Hits	Trees	Hits	Trees
Eubacteria	*Cyanobacteria* (9)	5,199	4,670	3,524	3,186	2,767	2,500	1,363	1,213
	Actinobacteria (17)	5,265	3,760	3,934	2,617	3,437	2,143	1,298	1,056
	Aquificae (1)	1,550	1,387	911	811	857	792	603	564
	Bacteroidetes (4)	3,116	2,431	2,024	1,558	1,702	1,392	988	849
	Chlamydiae (6)	1,884	1,638	1,124	982	946	828	2,126	858
	Chlorobi (3)	2,264	2,105	1,404	1,283	1,325	1,204	813	758
	Chloroflexi (2)	1,232	1,157	882	691	874	742	485	464
	Deinococcus–Thermus (2)	2,598	2,158	1,912	1,455	2,061	1,478	890	770
	Firmicutes (45)	5,188	3,545	3,402	2,412	2,608	1,860	1,214	996
	Fusobacteria (1)	1,328	1,200	742	675	773	700	481	441
	Planctomycetes (1)	2,730	2,177	1,882	1,479	1,483	1,266	788	705
	Proteobacteria (105)	6,857	4,378	4,958	3,009	4,385	2,392	1,585	1,160
	Spirochaetes (5)	2,766	2,317	1,633	1,380	1,415	1,211	851	735
	Thermotogae (1)	1,615	1,368	952	817	854	757	582	530
Archaeabacteria	*Crenarchaeota* (5)	1,941	1,366	1,175	815	974	694	693	491
	Euryarchaeota (18)	4,100	2,902	2,716	1,911	1,982	1,450	1,122	801
	Nanoarchaeota (1)	429	208	248	122	217	131	202	106
Eukaryotes	*Ascomycota* (8)	10,734	3,850	7,853	2,624	4,881	1,819	2,303	922
	Basidiomycota (2)	10,047	3,545	7,685	2,497	4,151	1,619	2,126	858
	Microsporidia (1)	2,849	836	1,594	516	1,182	415	769	284
	Protists (2)	7,187	2,158	4,432	1,509	3,155	1,139	1,415	525

data sets, alignments were constructed and phylogenetic trees were inferred using maximum likelihood and NJ.

On average, about 14% of those archeaplastidan proteins with homologues in cyanobacteria and at least 2 other phyla branched as nearest neighbor to cyanobacterial homologues. However, among the gene trees investigated, we found that almost every phylum sampled appeared as a nearest neighbor of the query species (table 2), which can be due to phylogenetic error or LGT among prokaryotes. Contrasted to the proportion of the proteins from each phylum in our database, the cyanobacterial signal is substantial. For example, 13% of the trees deliver cyanobacterial nearest neighbors in *Arabidopsis*, yet only 3.8% of the genes in our database reside in cyanobacterial chromosomes: the proportions of cyanobacterial nearest neighbors in plant genomes are far above random similarities. If we divide the proportion of nearest neighbors for each phylum by its gene content in our data set, then we find that the next largest proportion of plant gene nearest neighbors stems from the eubacterial genomes in our sample, with a majority of proteobacterial genes (fig. 1 and table 2). These could be attributed either to genes that were acquired from the mitochondrion ancestor, biased phylogenetic signals, and/or the fluid nature of bacterial genomes over time (Esser et al. 2007).

EGT Inference Depends upon Sequence Conservation

The foregoing first approximations for the percentage of cyanobacterial acquisitions represent only the fraction of genes in each archaeplastidan genome for which we can construct trees with at least 4 taxa at the phylum level. They exclude genes with a more narrow distribution across prokaryotic phyla (many photosystem genes, for example) and those that do not fulfill our criteria for inclusion in multiple alignments because of low sequence conservation. Moreover, among those genes with sufficient phyletic distribution and sequence conservation to be included in the tree-building procedure, proteins with higher sequence conservation showed a tendency to display a cyanobacterial origin more often than more divergent proteins did. For example, 22% of the *Arabidopsis* gene trees with a mean branch length of $\leq$0.2 substitutions per site indicate cyanobacterial origin of the plant nuclear gene, whereas only 6% of the gene trees with mean branch length $\geq$1 do (fig. 2*a*). The same tendency is observed for rice (fig. 2*b*) and *Chlamydomonas* (fig. 2*c*). This could have either a biological or a methodological cause: 1) genes of cyanobacterial origin in plants might preferentially belong to the most slowly evolving proteins in the genome or 2) because phylogenetic inference is less accurate with more highly diverged

Table 2
The distribution of nearest neighbors of archaeplastidan genes within the different taxa. Mixed clades include members from various taxa

Nearest Neighbor	*Arabidopsis* (%)	*Oryza* (%)	*Chlamydomonas* (%)	*Cyanidioschyzon* (%)	No. proteins
Cyanobacteria	592 (13%)	432 (14%)	356 (14%)	207 (17%)	31,940
Proteobacteria	522 (11%)	308 (10%)	458 (18%)	104 (9%)	360,234
Other eubacteria	882 (19%)	588 (18%)	524 (21%)	257 (21%)	226,314
Archaeabacteria	45 (1%)	38 (1%)	25 (1%)	18 (1%)	56,513
Eukaryotes	2156 (46%)	1498 (47%)	895 (36%)	523 (43%)	176,606
Mixed	473 (10%)	323 (10%)	242 (10%)	106 (9%)	

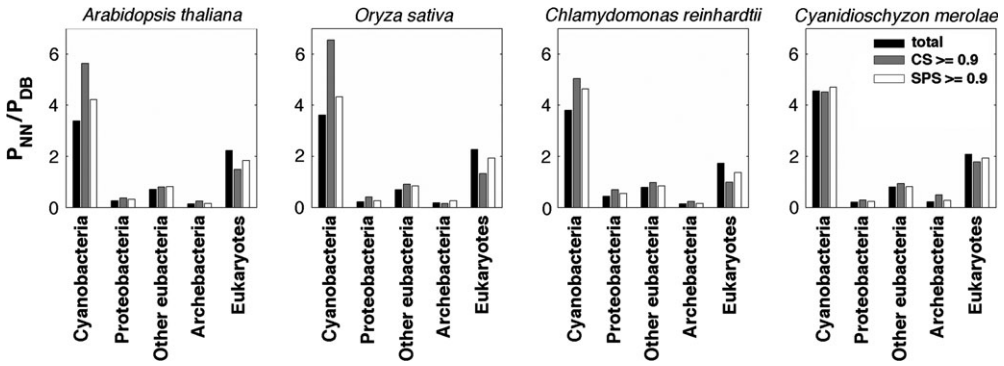

FIG. 1.—The ratio of the proportion of nearest neighbors in different taxa (P_{NN}; table 2) and the proportion of the proteins of the taxa in the database (P_{DB}).

proteins (Nei et al. 1995; Nei 1996), the rate of false negatives—in the present case, trees failing to recover a true cyanobacterial origin of some plant proteins—might increase with increasing sequence divergence.

To test the first possibility, we calculated pairwise protein distances of *Arabidopsis* proteins to their orthologs in rice and *Chlamydomonas* and plotted the proportion of proteins that were inferred as an EGT (or not) as a function of those pairwise distance distributions (fig. 2*e*). Comparison of the protein distances calculated for *Arabidopsis* and rice orthologs showed no significant difference between the protein distance distribution of EGTs and of non-EGT proteins ($P = 0.64$, using Wilcoxon test). The same procedure for *Arabidopsis* and *Chlamydomonas* orthologs showed that the protein distances of EGTs are somewhat smaller than the protein distances of non-EGTs ($P = 0.002$, using Wilcoxon test). Hence, the tendency to infer EGT among the less polymorphic alignments is not clearly attributable to a tendency for EGTs to preferentially occur among the more slowly evolving proteins in the genome, suggesting that phylogenetic inference itself might be preferentially generating an increased frequency of false negatives for EGT candidates among the more highly divergent sequences.

Divergent Proteins Produce Unreliable Alignments

If phylogenetic inference is producing false EGT negatives preferentially for the more highly divergent sequences in our sample, there are, in principle, 2 prime suspects as the possible source of error: Either the tree-building algorithms are failing or the alignments themselves are called into question. It is well known that phylogenetic inference is error prone (Phillips et al. 2004), particularly when sequence divergence exceeds 50% (Nei et al. 1995; Nei 1996), which is very commonly the case in genome-wide phylogenetic studies such as the present one (fig. 2). But the influence that the methodology of sequence alignment itself can have upon phylogeny is less well studied. A phylogenetic tree can hardly be more reliable than the alignment upon which it is based, but objective criteria to assess the quality of alignments in practice are, at best, extremely rare (Kumar and Filipski 2007). We therefore turned our attention to the alignments underlying the phylogenetic analyses in the present study, rather than to the minutiae of tree-building methods themselves, in order to better characterize their influence upon our ability to infer a particular biological process (gene acquisitions from plastids) from contemporary genome data.

For this purpose, we examined the utility of the Heads or Tails (HoT) method (Landan and Graur 2007) that quantifies the dependence of inferred residue homology in an alignment matrix upon the order in which characters are entered into the alignment. In the HoT method, the input sequences are aligned to create a heads (or forward) alignment. For the tails (or reverse) alignment, the input sequences are first reversed, so that the sequences read C- to N-terminus, and aligned independently using the same algorithm and settings, creating a second alignment of the same sequences. The 2 alignments are then compared using the CS (Thompson et al. 1999), which is the proportion of alignment columns that were reconstructed identically in the 2 alignments. When the forward and reverse alignments are nearly identical, the vast majority of columns are reproduced in both alignments and are thus consistent with respect to the character input order and the alignment can be considered reliable (or at least reproducible). But when the CS is low, the alignment is marked as being highly dependent upon an arbitrary variable, namely the order in which the amino acids are subjected to alignment. In this way, both an alignment and trees inferred from it can be identified as questionable or unreliable by virtue of their dependence upon a variable no less arbitrary than a coin toss (Landan and Graur 2007).

When all alignments underlying the trees in figure 2 are examined by HoT analysis, a strong correlation is found in all 4 genome data sets between sequence conservation as estimated by mean branch length in the trees and the proportion of identical columns recovered in the heads and tails alignments (fig. 3). CS among the most conserved sequences (mean branch length ≤ 0.2) ranges between a mean of 0.77 ± 0.17 in *Arabidopsis* and 0.76 ± 0.18 in *Chlamydomonas*, whereas among the most divergent sequences CS falls to a mean of 0.45 ± 0.37 in *Cyanidioschyzon* (mean branch length ≥ 0.8) or 0.11 ± 0.14 in *Arabidopsis* (mean branch length ≥ 1; fig. 3). In other words, for the most highly divergent sequences in the present data, about 80–90% of the columns generated by multiple alignment

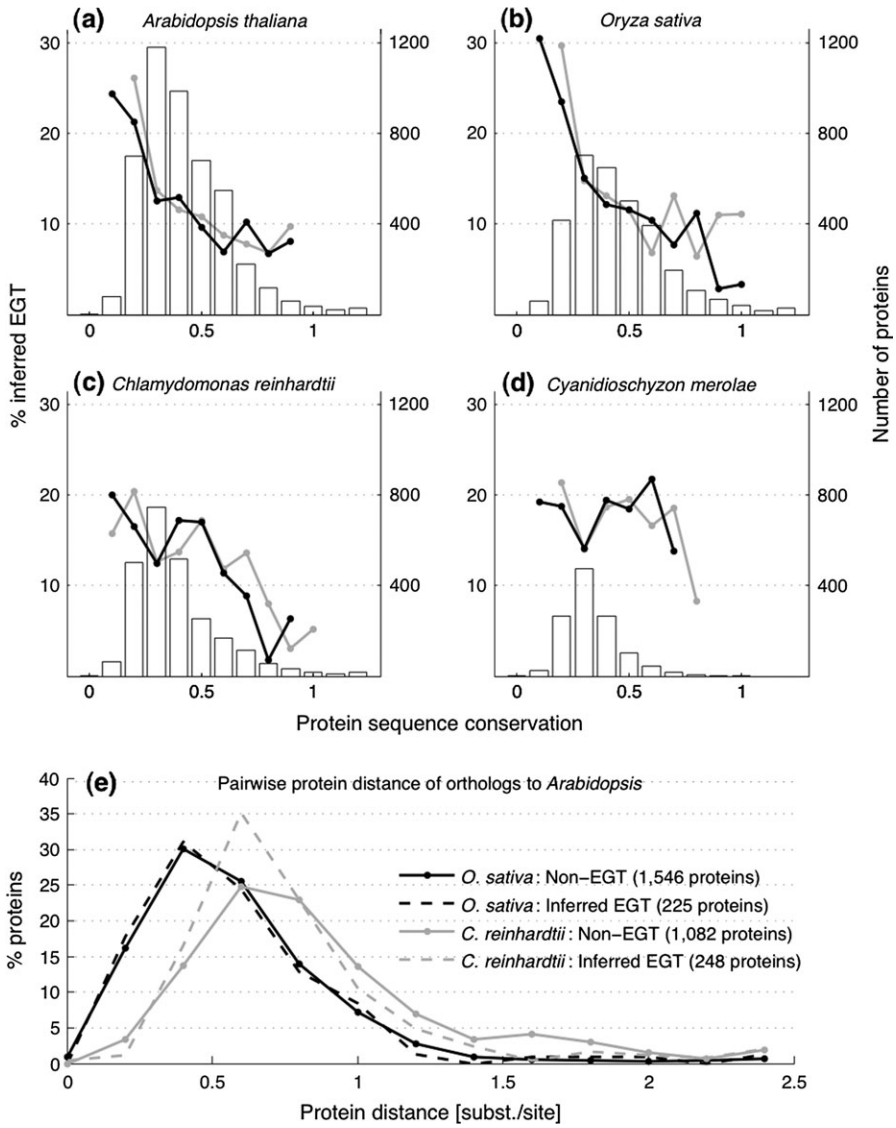

FIG. 2.—(*a–d*) The frequency of EGT as inferred from multiple sequence alignments (MSAs) of varying sequence conservation degrees. The distribution of sequence conservation as calculated by mean branch length of the phylogenetic trees (the sum of branch lengths divided by the number of branches) reconstructed using ML approach is presented in open bars (similar distribution is observed for NJ trees). The frequency of genes inferred as EGT is plotted for both ML tress (black) and NJ trees (gray). (*e*) Distribution of pairwise protein distances for rice and *Chlamydomonas* orthologs of *Arabidopsis* genes.

are dependent upon the order in which the amino acids are aligned. Such columns are irreproducible in the simplest test case and therefore contain unreliable, and possibly misleading, information.

Because EGT inference uses the alignments to reconstruct phylogenetic trees, we used the HoT analysis to construct a separate phylogeny for both the heads and tails alignments of each protein and compared the results. For each sequence set, the heads and the tails alignments were created using Muscle, both were purged of gapped sites, and both were used to infer trees with ML and NJ. The similarity of the heads and tails trees was quantified by the PPS, which is the fraction of internal edges that are common to both trees. The PPS is strongly dependent upon CS for all 4 genome data sets (table 3), and this observation holds for both ML and NJ tree topologies.

The correlation between PPS and CS is positive, except for the *Cyanidioschyzon* data, which probably relates to that genome's small sample size (only 4,762 proteins). But CS is not always a good proxy for PPS in the heads versus tails trees comparison because there are cases (~5% of the data) in which unreliable alignments still produce the same topology. Another measure for alignment reliability using HoT analysis is the comparison of identically reconstructed amino acids pairs (Landan and Graur 2007), in which all of the amino acid pairs between all pairs of sequences are tested for identical reconstruction, yielding a SPS (Thompson et al. 1999). SPS is a less strict measure of alignment uncertainty than CS because if a column differs with regard to only one or a few sequences in the heads or tails comparison, the corresponding HoT columns are scored as proportionately similar using SPS but as

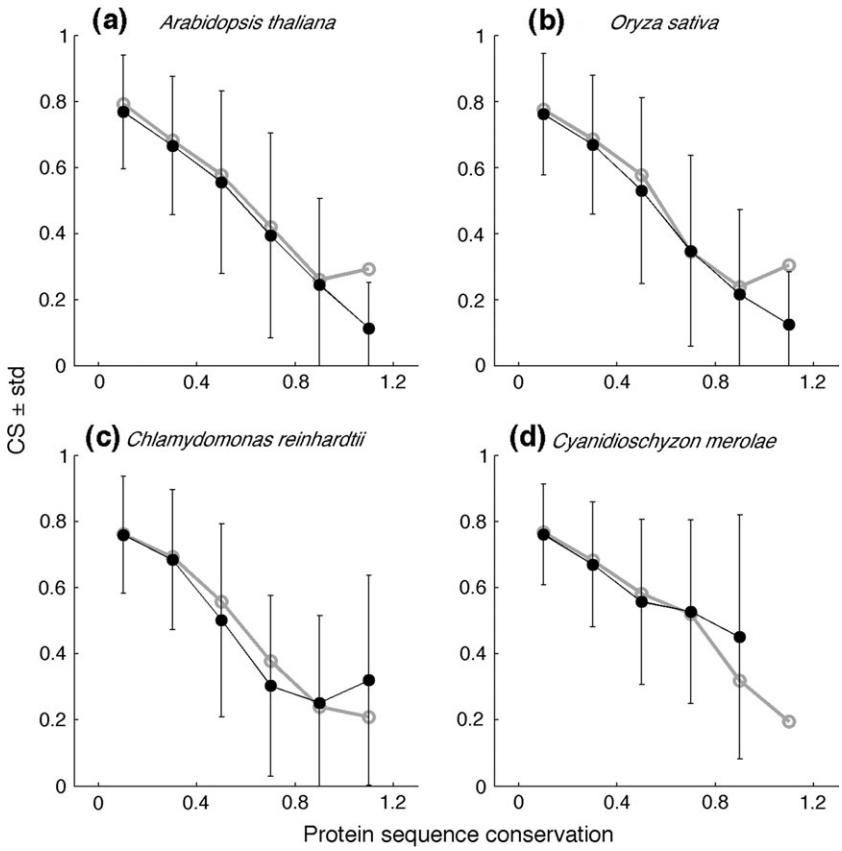

FIG. 3.—Conserved sequences produce more reliable alignments. The mean CS (dots) ± standard deviation is presented for ML trees (black) and NJ trees (gray).

nonidentical (0 of 0/1) using CS. SPS is therefore almost always higher than CS for a given alignment (supplementary fig. 1, Supplementary Material online), but when it is low, the alignment is extremely unreliable because almost all putatively homologous amino acid pairs that are presented to the tree reconstruction procedure are generated in a manner that is dependent upon the heads versus tails parameter of the alignment procedure. Accordingly, SPS correlates better with PPS (table 3), reaching a correlation coefficient of up to $r = 0.77$ in the rice data.

Better Alignments Give Higher Estimates for the Fraction of Acquired Genes

The correlation between sequence conservation and the reliability of the alignment (fig. 3) and its implications for the reliability of the inferred phylogentic tree (table 3), suggest that the inferred frequency of EGT depends primarily upon the alignment reliability and that the tree results are merely a secondary symptom thereof.

If we estimate the frequency of EGT-positive trees from the most conserved sequences only, we get higher estimates than if we also consider the poorly conserved sequences. The latter are, however, producing many false negatives because we observe that EGT-positive trees are not more prevalent among slowly evolving genes than EGT negatives by the measure of pairwise identity of the archaeplastidan homologues (fig. 2*e*). Because alignments determine phylogenetic results via the generation of homology patterns, the most accurate estimate should come from the best alignments.

In figure 4, the fraction of EGT-positive trees is plotted across the distributions for bin intervals of heads-versus-tails alignment quality as estimated by the criteria CS,

Table 3
Correlation coefficient of the PPS and 2 alignment reliability measures: the CS and the SPS. Both *CS* and *SPS* correlate significantly ($P < 0.05$ using the Pearson's correlation) with the proportion of identical PPS as calculated for both NJ and ML trees in all tested genomes

	Arabidopsis thaliana		*Oryza sativa*		*Chlamydomonas reinhardtii*		*Cyanidioschyzon merolae*	
	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)	PPS (ML)	PPS (NJ)
CS	0.6	0.63	0.61	0.66	0.61	0.65	0.42	0.45
SPS	0.71	0.75	0.77	0.79	0.7	0.75	0.48	0.53

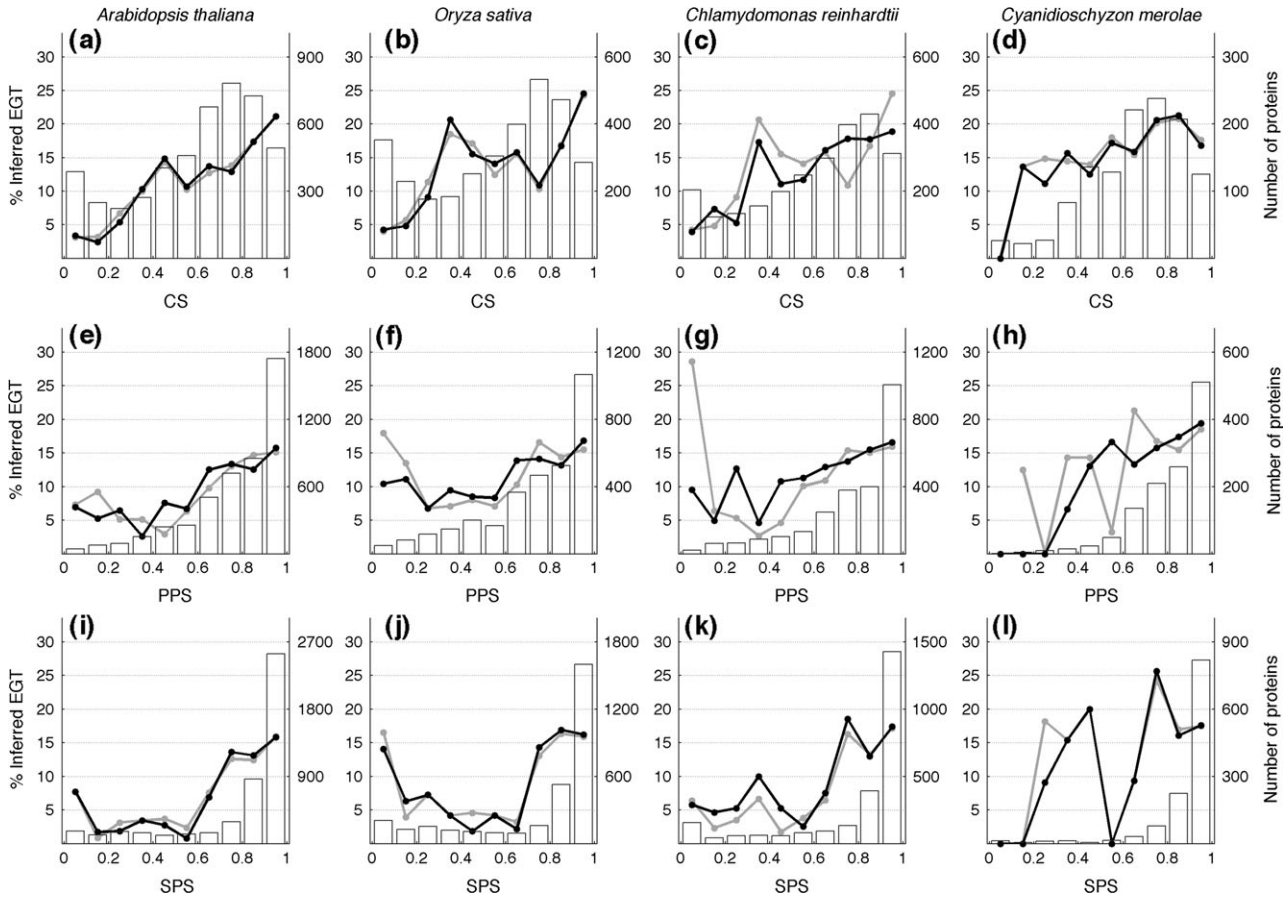

FIG. 4.—The frequency of EGT as inferred from MSAs of varying reliability degrees. The distribution of MSA reliability as estimated by the 3 measures is presented in open bars. The frequency of genes inferred as EGT is plotted above for ML trees (black) and NJ trees (gray).

PPS, and SPS for each archeaplastidan genome data set. For each criterion, the most reliable (reproducible) alignments give the highest estimate of EGT-positive trees, on the order of 17–25% of all trees examined, with a decreasing trend toward poorer alignments, notwithstanding variation among intervals containing small sample sizes. For alignments with CS $\geq$0.9 (fig. 4*a–d*), the highest estimates are obtained, but between 87% (*Chlamydomonas*) and 91% (*Oryza*) of the gene trees are excluded from consideration by this stringent criterion. For alignments with PPS $\geq$0.9 (fig. 4*e–h*), over 50% of all trees are excluded, but the estimates for the proportion of EGT-positive trees are very similar to the estimates for SPS $\geq$0.9 (fig. 4*i–l*), where between 50% (*Oryza*) and 67% (*Cyanidioschyzon*) of the gene trees are included in the estimate. Clearly, excluding alignments in which >20% of the site pairs differ in the toss-up comparison represents a very conservative threshold for excluding phylogenomic data; such data should not be subjected to phylogenetic inference because the phylogeny inference program will be optimizing parameters for site patterns of extremely uncertain homology. If we use the conservative value of SPS $\geq$0.9 as a cutoff for excluding data that is identified by the HoT method as irreproducibly alignable, hence unreliable, then 16%, 16%, 17%, and 18% of the alignments (gene trees) examined indicate a cyanobacterial origin of the plant nuclear gene for *Arabidopsis*, *Oryza*, *Chlamydomonas*, and *Cyanidioschyzon*, respectively.

How Do Genes Get to the Nucleus and Why Are Any Retained in the Plastid?

The quantitative contribution of EGT inferred here for plant genomes is substantial and raises the question of what gene transfer mechanism is involved. Various lines of evidence favor the view that the mechanism of EGT involves bulk recombination of organelle DNA that is released into the cytoplasm, perhaps by organelle lysis, with DNA of nuclear chromosomes (Timmis et al. 2004). One such line of evidence stems from fragments of organelle DNA that have been relocated to the nucleus and integrated into nuclear chromosomes (Bensasson et al. 2001; Richly and Leister 2004; Behura 2007; Hazkani-Covo and Graur 2007). Nuclear copies of organelle DNA do not preferentially comprise coding regions or particular segments of organelle DNA (Hazkani-Covo and Graur 2007). In *Arabidopsis*, a complete and almost intact 131-kb copy of the chloroplast genome is found near the centromere of nuclear chromosome 2, whereas in rice, a complete and nearly intact copy of the 367-kb mitochondrial genome is found near the centromere of chromosome 10. Both copies share >99%

sequence identity with their homologues in organelles (Huang et al. 2005), indicating that they were transferred intact and recently during evolution. In addition, laboratory studies employing transgenic mitochondria (Thorsness and Fox 1990) or transgenic chloroplasts (Huang et al. 2003) with suitable marker genes have directly demonstrated novel organelle-to-nucleus DNA transfer events involving bulk DNA recombination and at rates that compare to the rate of point mutation (Stegemann et al. 2003). Sequenced eukaryotic genomes are, in general, replete with copies and fragments of organelle DNA that have been relocated to the nucleus and integrated into nuclear chromosomes (Leister 2005). This, together with studies involving transgenic organelles, indicate that gene transfer from organelles to the nucleus via organelle lysis and bulk recombination of organelle chromosomes into nuclear chromosomes is the mechanism of organelle-to-nucleus gene transfer, with recombination, expression, mutation, and selection governing the fate of the transferred genes.

As in a previous study (Martin et al. 2002), the list of genes transferred to plant nuclei from the ancestor of plastids contains sequences encoding virtually all functional categories but mainly biosynthetic and metabolic functions (supplementary table S2, Supplementary Material online). The mechanism of transfer suggests that any gene can be transferred and fixed in the nucleus, so why should any genes be retained in the organelle? A number of suggestions have been offered in this regard (Allen 2003), but the one that most fully accounts for the observations is that individual plastids need to be able to regulate the expression of genes encoding proteins involved in maintaining redox balance in the photosynthetic membrane.

Seeking the Closest Relative of Plastids in Nuclear Genes and Allowing for LGT

In analyses of ribosomal RNA sequences, plastids tend to branch deeply within cyanobacterial phylogeny but to show no specific affinity to any particular cyanobacterial lineage (Turner et al. 1999; Marin et al. 2005). In analyses of alignments of concatenated plastid-encoded protein data, Rodriguez-Ezpeleta et al. (2005) obtained the same result, whereas Sato (2006) found weak phylogenetic evidence to suggest that plastids might stem from within the *Anabaena–Synechocystis* lineage. As outlined in the introduction, the concatenation approach usually assumes that the concatenated sequences in question all share the same history, and this assumption is generally problematic where prokaryotic genes are involved (Bapteste et al. 2007). Initial studies with concatenated sequences of plastid-encoded genes investigated the congruence of signals for individually analyzed plastid proteins (Goremykin et al. 1997; Martin et al. 1998; Vogl et al. 2003), but this step is usually omitted in newer analyses (Rodriguez-Ezpeleta et al. 2005), which would appear problematic, especially given the evidence that cyanobacteria exchange genes (Raymond et al. 2002; Zhaxybayeva et al. 2004, 2006) and that signal loss in plastid-encoded proteins is an issue even during algal phylogeny (White et al. 2007). Taken together, the available observations suggest that there is not enough information present in the ~45 proteins common to the genomes of photosynthetic plastids to confidently ascertain the closest relative of plastids among modern cyanobacteria. Furthermore, the observations suggest that LGT among cyanobacteria requires that the question be approached in such a manner as to avoid the concept of "sister group lineages" among prokaryotes at the whole genome level (Doolittle and Bapteste 2007), whereas for individual genes, the concept of the sister group seems unproblematic.

Given the foregoing, we asked: Which of the 9 cyanobacterial genomes sampled here contains the highest frequency of genes that are scored as being of cyanobacterial origin in plant nuclear genomes? For each of the 4 plant genomes and for each of the 9 cyanobacterial genomes sampled, we tabulated the pairwise amino acid sequence identity for each cyanobacterial protein that was present in our alignments and that was scored as being of cyanobacterial origin in the respective plant genome. Color-coded versions of those tables are presented in figure 5*a*, where it is evident that the 9 cyanobacteria sampled differ with respect to their overall similarity to the collection of cyanobacterial genes that is present in plant nuclear genomes. The differences reflect gene presence or absence on the one hand, with the smallest genome (2,265 proteins in *P. marinus* MIT9313) harboring fewer proteins that share high sequence identity to the plant homologues. But the differences also reflect overall sequence similarity, with *G. violaceus* PCC 7421, a midsized genome with 4430 proteins, harboring many homologues of plant proteins of cyanobacterial origin, albeit with visibly lower sequence identity than the other cyanobacteria in the sample (fig. 5*a*). This finding consistent with the circumstance that *Gloeobacter* is generally regarded a primitive or early-branching cyanobacterium because of its lack of thylakoids and its position in some phylogenetic trees (Sato 2006; Tomitani et al. 2006).

The sequence similarity array in figure 5*a* summarizes many millions of individual sequence comparisons, but it does not deliver phylogenetic specifics. Figure 5*b* shows the frequency with which proteins from the given cyanobacterial genome occurred within the sister group to the plant nuclear gene among the 11,569 ML phylogenies inferred in the present study. Those frequencies are shown for all alignments, for the alignments with SPS $\geq$0.8, and for the alignments with SPS $\geq$0.9. In all cases, *Anabaena* and *Nostoc* had the highest frequency of harboring a sister of the plant protein in phylogenetic trees. This result is influenced to some extent by genome size because *Anabaena* and *Nostoc* also have the highest frequency of harboring a homologue of those plant proteins, which found hits in cyanobacteria only (fig. 5*c*), and gene presence/absence is directly affected by genome size. But the size of a genome within which a protein-coding gene resides does not affect sequence similarity in individual protein comparisons (fig. 5*a*) nor does it directly bias the phylogentic results, as seen in the comparison of genome size (fig. 5*d*) and sister group frequency in ML trees (fig. 5*b*) for *Gloeobacter*. In other words, among the 9 cyanobacterial genomes sampled, *Nostoc* sp. PCC7120 and *A. variabilis* ATCC29143 harbor collections of genes that are—in terms of presence/absence and sequence similarity—more like those possessed by

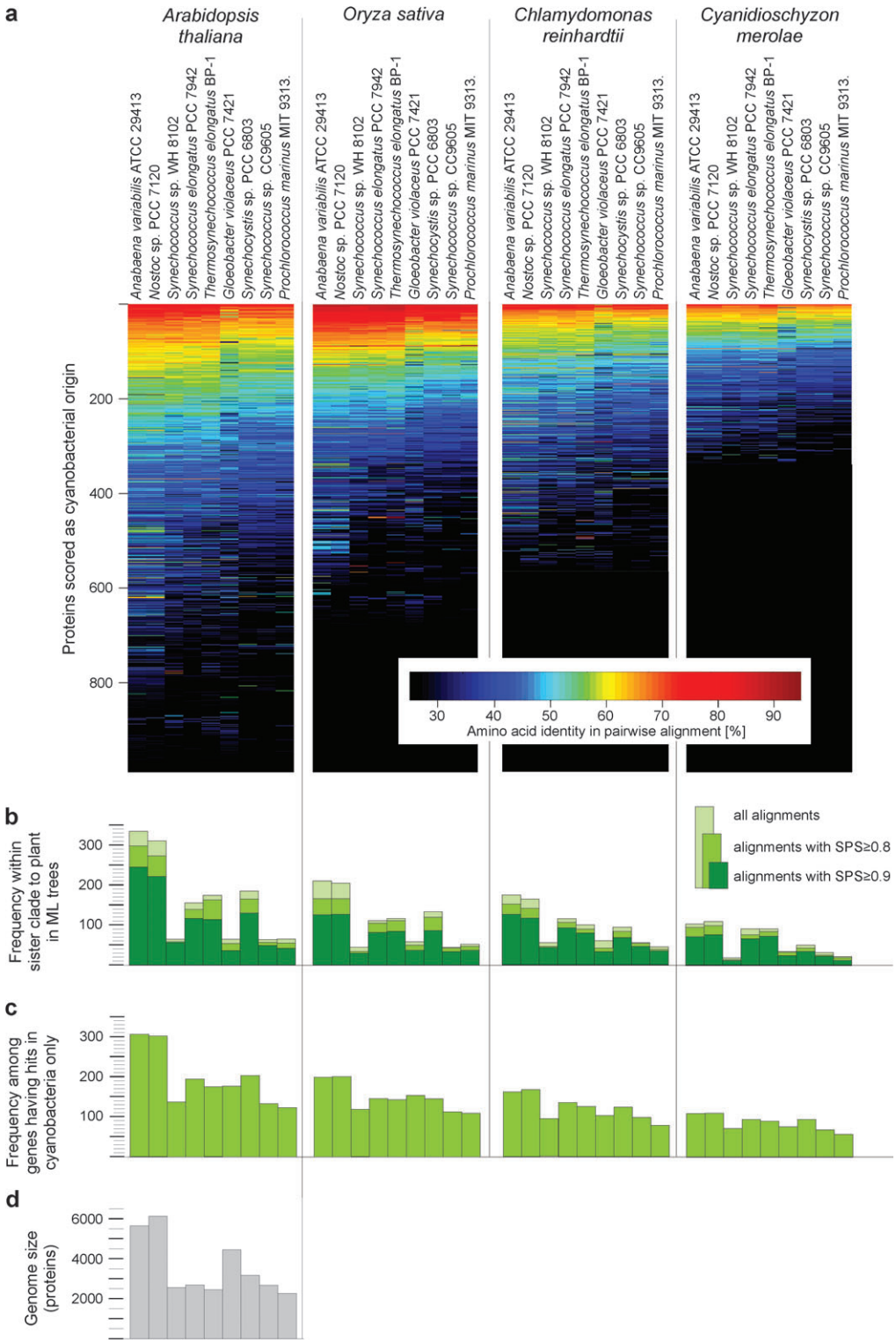

FIG. 5.—Overall similarity of proteins of cyanobacterial origin encoded in plant nuclear genomes to proteins encoded in cyanobacterial genomes. (*a*) For each of the 4 plant genomes, a color-coded table is shown. The rows correspond to the proteins scored as acquisitions from cyanobacteria; this includes cyanobacterial nearest neighbor inference and exclusive Blast hits within the cyanobacterial proteomes. The *Oryza sativa* RefSeq nuclear data includes a total of 51 genes that are annotated as nuclear proteins but probably are NUPTs (see Materials and Methods). Columns correspond to cyanobacterial genomes, the elements of the matrix contain the percent amino acid identity between the plant query, and the best hit in the respective cyanobacterial genome in the pairwise alignment generated by Blast. The scale of amino acid identity is given in the inset at lower right. (*b*) The frequency with which the homologue from the cyanobacterial genome indicated in (a) appeared in the cyanobacterial sister clade to the respective plant protein in ML trees. (*c*) The frequency with which the cyanobacterial genome indicated in (a) contained a homologue of the plant protein for plant proteins that found homologues in cyanobacteria only at the specified threshold. (*d*) Number of proteins encoded in each cyanobacterial genome indicated in (a).

the plastid ancestor than are those in the other 7 cyanobacterial genomes sampled here. With regard to *Nostoc* sp. PCC7120, our result is consistent with an earlier result based on a smaller cyanobacterial genome sample and involving *Nostoc punctiforme* (Martin et al. 2002). With regard to *Nostoc* and *Anabaena*, our result is also consistent with the conclusions of Sato (2006) as they apply to *Anabaena*.

Nostoc, *Anabaena*, Symbiosis, Plastids, and Nitrogen

With the caveat that the present cyanobacterial sample is quite limited, our finding is that *Nostoc* and *Anabaena* harbor more genes than other cyanobacteria of the type that plants acquired from cyanobacteria. *Nostoc* and *Anabaena* are filamentous cyanobacteria that produce heterocysts, differentiated and specialized cells that perform nitrogen fixation without producing oxygen (Rippka et al. 1979; Rajaniemi et al. 2005). Filamentous cyanobacteria that produce heterocysts are grouped in the section IV (without true branching) and section V (with true branching) as recognized in Roger Stanier's cyanobacterial classification (Rippka et al. 1979), which followed the same morphological principles as Geitler's (1932) traditional system. Tomitani et al. (2006) presented evidence to indicate that sections IV and V are the most derived among cyanobacteria, consistent with traditional views (Geitler 1932; Rippka et al. 1979) and furthermore argued that the origin of heterocysts antedates the rise of atmospheric oxygen about 2.3 billion years ago. Were the ancestor of plastids a member of section IV (a heterocyst-forming cyanobacterium), as the present data tend to suggest, that would put a weak bound on the maximum age of plastids at 2.3 Ga, with the minimum age well constrained by the fossil red alga *Bangiomorpha* at 1.2 Ga (Butterfield 2000; Yoon et al. 2004). This conclusion is consistent with the observation that akinetes, large spore-like reproductive cells produced only by some members of sections IV and V (Rippka et al. 1979; Rajaniemi et al. 2005), appear in the palaeontological record >0.4 Ga earlier than *Bangiomorpha* (Tomitani et al. 2006).

If the cyanobacterial ancestor of plastids was a member of section IV, it was able to fix nitrogen, raising the question of what, exactly, the cyanobacterial symbiont was doing for its host during the early stages of symbiosis. It was suggested that the retargeting of a host-derived carbon transporter was the initial step that allowed the cyanobacterium to export reduced carbon for its host (Weber et al. 2006), but normal cyanobacterial cell wall polysaccharides would also perform that function, and how such a transporter would be targeted across the cyanobacterial cell wall into the plasma membrane was not discussed. Other suggestions have also been put forth, namely local oxygen production, either before (Stanier 1970) or after (Martin and Müller 1998) the origin of mitochondria, and nitrogen has recently been considered (Kneip et al. 2007). Indeed, if we look around at the chemical function of cyanobacteria in modern symbiotic endosymbiotic associations where the physiology has been studied, nitrogen supply, in some cases without carbon export, stands in the foreground (Rai et al. 2000; Raven 2002). Well-studied examples of modern cyanobacterial endosymbioses include *Geosiphon pyriforme* (a fungus: Mollenhauer et al. 1996), *Rhopalodia gibba* (a diatom; Prechtl et al. 2004), *Azolla* species (an acquatic fern; Prasanna et al. 2006), coralloid roots in cycads (Costa et al. 2004), and *Gunnera* (a flowering plant; Chiu et al. 2005).

In those examples, the cyanobacterial partner is a member of section IV from the genus *Nostoc* or *Anabanea*, the exception is *Rhopalodia*, where the symbiont is related to *Cyanothece* (Prechtl et al. 2004), a coccoid member of section I. In the majority of modern cyanobacterial endosymbioses, the endosymbionts are diazotrophic (nitrogen fixing), in many cases they provide reduced nitrogen to their host (Rai et al. 2000; Raven 2002; Prechtl et al. 2004), and *Nostoc* is a very common partner. *Richellia* (Nostocales) endosymbionts living within *Hemiaulus* and *Rhizosolenia* provide nitrogen to their diatom hosts (Raven 2002) and are important contributors to marine N availability (Montoya et al. 2004). Also in many ectosymbiotic associations, for example lichens, where *Nostoc* species are typically the cyanobacterial partner (Rikkinen et al. 2002) or in open ocean epiphyitc associations of *Dichothrix* (Nostocales) with the brown alga *Sargassum* (Carpenter 1972), nitrogen plays a decisive role. The role of nitrogen in symbiosis has recently been reviewed by Kneip et al. (2007).

In the larger geological context, it was proposed that the current model of anoxic and sulfidic oceans during the Proterozoic would have limited nitrogen availability globally (Anbar and Knoll 2002) starting from about 2.3 Ga up until ~0.58 Ga as newer findings indicate (Fike et al. 2006; Canfield et al. 2007). That is the time during which plastids arose (Butterfield 2000; Yoon et al. 2004) and it corresponds to "*. . .a period of exceptional N stress for the biosphere*" (Anbar and Knoll 2002, p. 1140), owing to the limited marine availability of trace elements—the transition metals Mo, Fe, and V—the limitation arising from the insolubility of the corresponding transition metal sulfides in oceans sulfidic due to biological sulfate reduction (Anbar and Knoll 2002). Taken together, that environmental limitation during the time of plastid origin, the role of nitrogen in modern cyanobacterial endosymbioses (Rai et al. 2000), and our present data linking plastids to heterocyst-forming cyanobacteria among a yet limited sample would be compatible with the view that nitrogen could have played a role in the establishment of the symbiosis that led to plastids (Raven 2002; Kneip et al. 2007).

Conclusion

On average, 14% of the nuclear-encoded proteins in the genomes of 4 photosynthetic eukaryotes having homologues in cyanobacteria and at least 2 other phyla, regardless of their alignment quality, were inferred as gene acquisitions from cyanobacteria. Alignments with better sequence conservation—that is, alignments whose site patterns were independent of the order in which amino acids were aligned—consistently recover a higher proportion of inferred cyanobacterial origin for plant nuclear genes (16–18%; >20% if only very highly conserved sequences are considered) than the alignments of more poorly

conserved sequences. That suggests that the latter contain many false negatives and that the conservative average value of 14% is an underestimate, underscoring a quantitatively large contribution of cyanobacteria to the makeup of plant genomes.

It has long been known from simulation studies that sequences sharing $<50\%$ amino acid identity perform poorly in phylogenetic reconstruction (Nei 1996). But in the deeper regions of genome phylogenetics, such as plastid origins, sequence pairs sharing $\geq 50\%$ amino acid identity are comparatively rare. The heads-or-tails approach does not directly identify a demarcation line that may not be crossed in sequence analysis, but it does reveal when phylogenetic results are solely determined by a process no less arbitrary than a coin toss and hence provides a reality check of the data quality underlying phylogenomic analyses.

The present phylogenomic data point to filamentous, heterocyst-forming (nitrogen-fixing) cyanobacteria as the plastid ancestor, an inference that is compatible with the observation that the majority of contemporary, physiologically characterized, cyanobacterial endosymbioses entail nitrogen supply as a benefit from endosymbiont to host. The current model of Proterozoic ocean chemistry (Anbar and Knoll 2002; Canfield et al. 2007) would not only be compatible with the widespread occurrence of anaerobic forms of mitochondria among the major eukaryotic lineages (Theissen et al. 2003; Dietrich et al. 2006; Embley and Martin 2006), but also be highly compatible with the concept of a plastid that could fix and supply nitrogen, owing to the distinct possibility of nitrogen limitation in the Proterozoic ocean.

Supplementary Material

Supplementary figure S1 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Paul Falkowski, Peter Lockhart, and Uwe Maier for stimulating discussions. This work was funded by the Deutsche Forschungsgemeinschaft (SFB-Tr1; W.M.), the German-Israeli Foundation (T.D.) and the National Scientific Foundation (DBI-0543342; G.L.).

Literature Cited

Adl SM, Simpson AGB, Farmer MA, et al. (28 co-authors). 2005. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. J Euk Microbiol. 52:399–451.

Allen JF. 2003. The function of genomes in bioenergetic organelles. Philos Trans R Soc Lond B. 358:1i9–37.

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.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. 1997. Gapped Blast and PSI-Blast: a new generation of protein database search programs. Nucleic Acids Res. 25:338i9–3402.

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

Archibald JM. 2006. Algal genomes: exploring the imprint of endosymbiosis. Curr Biol. 16:R1033–R1035.

Bapteste E, Susko E, Leigh J, Ruiz-Trillo I, Bucknam L, Doolittle WF. 2007. Alternative methods for concatenation of core genes indicate a lack of resolution in deep nodes of the prokaryotic phylogeny. Mol Biol Evol. 25:83–91.

Behura SK. 2007. Analysis of nuclear copies of mitochondrial sequences in honey bee (Apis mellifera) genome. Mol Biol Evol. 24:1492–1505.

Bensasson D, Zhang D, Hartl DL, Hewitt GM. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol. 16:314–321.

Bodyl A, Mackiewicz P, Stiller JW. 2007. The intracellular cyanobacteria of *Paulinella chromatophora*: endosymbionts or organelles? Trends Microbiol. 15:295–296.

Boeckmann B, Bairoch A, Apweiler R, et al. (12 co-authors). 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 31:365–370.

Bogorad L. 2008. Evolution of early eukaryotic cells: genomes, proteomes, and compartments. Photosynth Res. 95:11–21.

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.

Canfield DE, Poulton SW, Narbonne GM. 2007. Late-Neoproterozoic deep-ocean oxygenation and the rise of animal life. Science. 315:92–95.

Carpenter EJ. 1972. Nitrogen fixation by a blue-green epiphyte on pelagic *Sargassum*. Science. 178:1207–1209.

Chiu WL, Peters GA, Levieille G, Still PC, Cousins S, Osborne B, Elhai J. 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.

Delwiche CW. 1999. Tracing the thread of plastid diversity through the tapestry of life. Am Nat. 154:S164–S177.

Dietrich LEP, Tice MM, Newmann DK. 2006. The co-evolution of life and earth. Curr Biol. 16:R395–R400.

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the Tree of Life hypothesis. Proc Natl Acad Sci USA. 104:2043–2049.

Douglas SE. 1998. Plastid evolution: origins, diversity, trends. Curr Opin Genet Dev. 8:655–661.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Embley TM, Martin W. 2006. Eukaryotic evolution: changes and challenges. Nature. 440:623–630.

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

Felsenstein J. 2005. PHYLIP (phylogeny inference package). Version 3.6. Seattle (WA): Department of Genome Sciences, University of Washington.

Fike DA, Grotzinger JP, Pratt LM, Summons RE. 2006. Oxidation of the Ediacaran ocean. Nature. 444:744–747.

Geitler L. 1932. Rabenhorst's Kryptogamenflora von Deutschland, Österreich und der Schweiz. Vierzehnter Band: Cyanophyceae. Leipzig (Germany): Akademische Verlagsgesellschaft M.B.H., p. 1196.

Goremykin VV, Hansmann S, Martin WF. 1997. Evolutionary analysis of 58 proteins encoded in six completely sequenced

chloroplast genomes: revised molecular estimates of two seed plant divergence times. Plant Syst Evol. 206:337–351.
Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.
Jones DT, Taylor WR, Thornton JM. 1992. The rapid generation of mutation rate matrices from protein sequences. Comput Appl Biosci. 8:275–282.
Hazkani-Covo E, Graur D. 2007. A comparative analysis of numt evolution in human and chimpanzee. Mol Biol Evol. 24: 13–18.
Huang CY, Ayliffe MA, Timmis JN. 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature. 422:72–76.
Huang CY, Grunheit N, Ahmadinejad N, Timmis JN, Martin W. 2005. Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. Plant Physiol. 138:1723–1733.
Kneip C, Lockhart P, Voss C, Maier UG. 2007. Nitrogen fixation in eukaryotes—new models for symbiosis. BMC Evol Biol. 7:55.
Kumar S, Filipski A. 2007. Multiple sequence alignment: in pursuit of homologous DNA positions. Genome Res. 17: 127–135.
Landan G, Graur D. 2007. Heads or tails: a simple reliability check for multiple sequence alignments. Mol Biol Evol. 24:1380–1383.
Larkum AW, Lockhart PJ, Howe CJ. 2007. Shopping for plastids. Trends Plant Sci. 12:189–195.
Leister D. 2005. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet. 21:655–663.
Lockhart PJ, Howe CJ, Barbrook AC, Larkum AWD, Penny D. 1999. Spectral analysis, systematic bias, and the evolution of chloroplasts. Mol Biol Evol. 16:573–576.
Marin B, Nowack ECM, Melkonian M. 2005. A plastid in the making: evidence for a second primary endosymbiosis. Protist. 156:425–432.
Martin W. 1999. Mosaic bacterial chromosomes—a challenge en route to a tree of genomes. Bioessays. 21:99–104.
Martin W, Brinkmann H, Savona C, Cerff R. 1993. Evidence for a chimeric nature of nuclear genomes: eubacterial origin of eukaryotic glyceraldehyde-3-phosphate dehydrogenase genes. Proc Natl Acad Sci USA. 90:8692–8696.
Martin W, Herrmann RG. 1998. Gene transfer from organelles to the nucleus: how much, what happens, and why? Plant Physiol. 118:9–17.
Martin W, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. Nature. 392:37–41.
Martin W, Rujan T, Richly E, Hansen A, Cornelsen S, Lins T, Leister D, Stoebe B, Hasegawa M, Penny D. 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.
Martin W, Schnarrenberger C. 1997. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. Curr Genet. 32:1–18.
Martin W, Stoebe B, Goremykin V, Hansmann S, Hasegawa M, Kowallik KV. 1998. Gene transfer to the nucleus and the evolution of chloroplasts. Nature. 393:162–165.
Matsuzaki M, Misumi O, Shin-IT, et al. (42 co-authors). 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. Nature. 428:653–657.
McFadden GI, van Dooren GG. 2004. Evolution: red algal genome affirms a common origin of all plastids. Curr Biol. 14:R514–R516.
Merchant SS, Prochnik SE, Vallon O, et al. (117 co-authors). 2007. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science. 318:245–250.
Mereschkowsky C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biol Centralbl. 25:593–604. [English translation in Eur J Phycol. 34:287–295 (1999)].
Mollenhauer D, Mollenhauer R, Kluge M. 1996. Studies on initiation and development of the partner association in Geosiphon pyriforme (Kutz) v Wettstein, a unique endocytobiotic system of a fungus (Glomales) and the cyanobacterium *Nostoc punctiforme* (Kutz) Hariot. Protoplasma. 193:3–9.
Montoya JP, Holl CM, Zehr JP, Hansen A, Villareal TA, Capone DG. 2004. High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. Nature. 430:1027–1031.
Morden CW, Golden SS. 1989. psbA genes indicate common ancestry of prochlorophytes and chloroplasts. Nature. 337:382–385.
Nei M. 1996. Phylogenetic analysis in molecular evolutionary genetics. Annu Rev Genet. 30:371–403.
Nei M, Takezaki N, Sitnikova T. 1995. Assessing molecular phylogenies. Science. 267:253–254.
Phillips MJ, Delsuc F, Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol. 21:1455–1458.
Prasanna R, Kumar R, Sood A, Prasanna BM, Singh PK. 2006. Morphological, physiochemical and molecular characterization of *Anabaena* strains. Microbiol Res. 161:187–202.
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.
Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 33:D501–D504.
Rai AN, Söderbäck E, Bergman B. 2000. Cyanobacterium-plant symbioses. New Phytol. 147:449–481.
Rajaniemi P, Hrouzek P, Kastovská K, Willame R, Rantala A, Hoffmann L, Komárek J, Sivonen K. 2005. Phylogenetic and morphological evaluation of the genera *Anabaena, Aphanizomenon*, *Trichormus* and *Nostoc* (Nostocales, Cyanobacteria). Int J Syst Evol Microbiol. 55:11–26.
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. 326–1246.
Raymond J, Zhaxybayeva O, Gogarten JP, Gerdes SY, Blankenship RE. 2002. Whole-genome analysis of photosynthetic prokaryotes. Science. 298:1616–1620.
Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D. 2006. Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. Curr Biol. 16:2320–2325.
Reyes-Prieto A, Weber APM, Bhattacharya D. 2007. The origin and establishment of the plastid in algae and plants. Annu Rev Genet. 41:147–680.
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. 111:1–6.

Rodriguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Loffelhardt W, Bohnert HJ, Philippe H, Lang BF. 2005. Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Curr Biol. 15:1325–1330.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.

Sato N. 2001. Was the evolution of plastid genetic machinery discontinuous? Trends Plant Sci. 6:151–155.

Sato N. 2006. Origin and evolution of plastids: genomic view on the unification and diversity of plastids. In: Wise RR, Hoober JK, editors. The structure and function of plastids. Dordrecht (The Netherlands): Springer. p. 75–102.

Soll J, Schleiff E. 2004. Protein import into chloroplasts. Nat Rev Mol Cell Biol. 5:198–208.

Stanier RY. 1970. Some aspects of the biology of cells and their possible evolutionary significance. Symp Soc Gen Microbiol. 20:1–38.

Stegemann S, Hartmann S, Ruf S, Bock R. 2003. High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci USA. 100:8828–8833.

Stoebe B, Hansmann S, Goremykin V, Kowallik KV, Martin W. 1999. Proteins encoded in sequenced chloroplast genomes: an overview of gene content, phylogenetic information, and endosymbiotic gene transfer to the nucleus. In: Hollingsworth C, Bateman R, Gornall M, editors. Advances in plant molecular systematics. Andover (UK): Francis and Taylor. p. 327–352.

Theissen U, Hoffmeister M, Grieshaber M, Martin W. 2003. Single eubacterial origin of eukaryotic sulfide: quinone oxidoreductase, a mitochondrial enzyme conserved from the early evolution of eukaryotes during anoxic and sulfidic times. Mol Biol Evol. 20:1564–1574.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Thompson JD, Plewniak F, Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res. 27:2682–2690.

Thorsness PE, Fox TD. 1990. Escape of DNA from mitochondria to the nucleus in *Saccharomyces cerevisiae*. Nature. 346:376–379.

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

Tomitani A, Knoll AH, Cavanaugh CM, Ohno T. 2006. The evolutionary diversification of cyanobacteria: molecular-phylogenetic and paleontological perspectives. Proc Natl Acad Sci USA. 103:5442–5447.

Turner S, Pryer KM, Miao VPW, Palmer JD. 1999. Investigating deep phylogenetic relationships among cyanobacteria and plastids by small submit rRNA sequence analysis. J Euk Microbiol. 46:327–338.

Tyra HM, Linka M, Weber AP, Bhattacharya D. 2007. Host origin of plastid solute transporters in the first photosynthetic eukaryotes. Genome Biol. 8:R212.

Vogl C, Badger J, Kearney P, Li M, Clegg M, Jiang T. 2003. Probabilistic analysis indicates discordant gene trees in chloroplast evolution. J Mol Evol. 56:330–40.

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.

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.

Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol. 21:809–818.

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.

Zhaxybayeva O, Hamel L, Raymond J, Gogarten JP. 2004. Visualization of the phylogenetic content of five genomes using dekapentagonal maps. Genome Biol. 5:R20.

Takashi Gojobori, Associate Editor

Accepted January 20, 2008