Molecular Biology and Evolution

Research article

Endosymbiotic Gene Transfer and Transcriptional Regulation of Transferred Genes in *Paulinella chromatophora*

Eva C.M. Nowack,*[,1,2] Heiko Vogel,[3] Marco Groth,[4] Arthur R. Grossman,[2] Michael Melkonian,†[,1] and Gernot Glöckner†[,4,5]

[1]Cologne Biocenter, Botanical Institute, University of Cologne, Köln, Germany

[2]Department of Plant Biology, Carnegie Institution for Science, Stanford, California

[3]Max Planck Institute for Chemical Ecology, Jena, Germany

[4]Leibniz Institute for Age Research–Fritz Lipmann Institute, Jena, Germany

[5]Leibniz Institute for Freshwater Ecology and Inland Fisheries, Berlin, Germany

†Both authors shared senior authorship.

***Corresponding author:** E-mail: enowack@stanford.edu.

Associate editor: Hervé Philippe

Abstract

Paulinella chromatophora is a cercozoan amoeba that contains "chromatophores," which are photosynthetic inclusions of cyanobacterial origin. The recent discovery that chromatophores evolved independently of plastids, underwent major genome reduction, and transferred at least two genes to the host nucleus has highlighted *P. chromatophora* as a model to infer early steps in the evolution of photosynthetic organelles. However, owing to the paucity of nuclear genome sequence data, the extent of endosymbiotic gene transfer (EGT) and host symbiont regulation are currently unknown. A combination of 454 and Illumina next generation sequencing enabled us to generate a comprehensive reference transcriptome data set for *P. chromatophora* on which we mapped short Illumina cDNA reads generated from cultures from the dark and light phases of a diel cycle. Combined with extensive phylogenetic analyses of the deduced protein sequences, these data revealed that 1) about 0.3–0.8% of the nuclear genes were obtained by EGT compared with 11–14% in the Plantae, 2) transferred genes show a distinct bias in that many encode small proteins involved in photosynthesis and photoacclimation, 3) host cells established control over expression of transferred genes, and 4) not only EGT, but to a minor extent also horizontal gene transfer from organisms that presumably served as food sources, helped to shape the nuclear genome of *P. chromatophora*. The identification of a significant number of transferred genes involved in photosynthesis and photoacclimation of thylakoid membranes as well as the observed transcriptional regulation of these genes strongly implies import of the encoded gene products into chromatophores, a feature previously thought to be restricted to canonical organelles. Thus, a possible mechanism by which *P. chromatophora* exerts control over the performance of its newly acquired photosynthetic organelle may involve controlling the expression of nuclear-encoded chromatophore-targeted regulatory components of the thylakoid membranes.

Key words: *Paulinella chromatophora*, endosymbiotic gene transfer, transcriptome, photosynthetic organelle, evolution.

Introduction

Paulinella chromatophora is a thecate amoeba of cercozoan affiliation that contains "chromatophores," photosynthetic inclusions of cyanobacterial origin. Like plastids in the Plantae, chromatophores confer phototrophy to their host cell and are consistently inherited by progeny cells as a consequence of tight synchronization of host and chromatophore cell cycles (Hoogenraad 1927). However, phylogenetic analyses of various molecular data sets have unambiguously demonstrated that chromatophores evolved by endosymbiosis of an α-cyanobacterium, independently from and more recently than the plastids that evolved by endosymbiosis of a β-cyanobacterium (Marin et al. 2005, 2007; Yoon et al. 2006; Reyes-Prieto et al. 2010). The α-cyanobacteria are a monophyletic group of cyanobacteria comprising the picoplanctonic genera *Prochlorococcus*, *Synechococcus*, and *Cyanobium*; their most distinctive feature is the possession of a proteobacterial form 1A RubisCO and α-carboxysomes obtained by horizontal gene transfer (HGT), in contrast to the remaining cyanobacteria, which contain a form 1B RubisCO and β-carboxysomes and are thus referred to as β-cyanobacteria (Badger et al. 2002; Marin et al. 2007). The complete sequence of the chromatophore genome of *P. chromatophora* strain CCAC 0185 revealed a genome of 1.02 Mb, which is substantially reduced compared with genomes of free-living cyanobacteria (Nowack et al. 2008). The chromatophore genome sequence of a second *Paulinella* isolate FK01—probably a second species (Yoon et al. 2009)—showed a similar degree of reduction and significant conservation of gene order compared with *P. chromatophora* strain CCAC 0185, but there were 66 differential gene losses between the two isolates (Reyes-Prieto et al. 2010). In both isolates, genome reduction was accompanied by the loss of coding capacity for essential biosynthetic functions, such as the synthesis of amino acids and cofactors, demonstrating

that chromatophores, like plastids, are dependent on their host for growth and survival. However, in contrast to plastids, chromatophore genomes are significantly less reduced, reflecting their younger evolutionary age. The status of the chromatophore as an endosymbiont or organelle has been controversial (Bhattacharya and Archibald 2006; Theissen and Martin 2006). The recent demonstration that two expressed genes, which encode low molecular weight subunits associated with photosystem I, were transferred from the chromatophore to the nuclear genome of *Paulinella* isolate FK01 (Nakayama and Ishida 2009; Reyes-Prieto et al. 2010), supports the notion that chromatophores should be regarded as organelles in an early stage of evolution. However, owing to the paucity of nuclear genome sequence data, the extent of endosymbiotic gene transfer (EGT) is not known, and questions concerning if and how the host regulates expression of transferred genes have not been addressed.

Despite recent advances in genome sequencing technology, the generation of a complete eukaryotic genome sequence is still challenging and requires significant time and resource commitments. Thus, to explore the status of EGT and host symbiont regulation, we relied on an in-depth characterization of the *P. chromatophora* transcriptome. For this purpose, we used a novel experimental approach based on a combination of 454 sequencing of a normalized *P. chromatophora* cDNA library and Illumina next generation sequencing of cDNA (not normalized) from cultures in the light and dark phases of the diel (i.e., over the course of the day) cycle. This approach resulted in the generation of a comprehensive reference transcriptome data set for *P. chromatophora* CCAC 0185 on which we then mapped the short Illumina reads. Together with extensive phylogenetic analyses, these data enabled us to assess the extent of EGT from the chromatophore to the host nucleus and provide insights into mechanisms by which host cells may regulate the performance of their newly acquired photosynthetic organelles.

Materials and Methods

Culture Conditions

Paulinella chromatophora strain CCAC 0185 (a subisolate of strain M0880/a) was grown as described before (Nowack et al. 2008) in a 14/10 h light/dark cycle at low light intensities (3–10 μE m^{-2} s^{-1}).

RNA isolation, library construction, and sequencing

For total RNA isolation, *P. chromatophora* cells from cultures in late logarithmic growth phase were harvested by centrifugation some hours after the onset of the light period and immediately plunged into liquid N_2. Total RNA from the frozen cells was isolated using Trizol reagent. The DNase-treated RNA was normalized (Zhulidov et al. 2004) to diminish levels of highly abundant transcripts and polyT primed to selectively enrich for mRNA. The resulting material served as template for the generation of a cDNA library that was sequenced using a single run of the Roche/454 FLX system according to the manufacturer's recommendations. To assess the effectiveness of the normalization process, 100 randomly chosen contigs consisting of more than one read were analyzed for their individual coverage. For analysis of dark/light-induced differential gene expression levels, cells were harvested under dim green light 1 h before the start of the light period (dark induced) and 20 min after start of light period (light induced). RNA from these light- and dark-treated cells was isolated using Trizol reagent, and polyT priming was used to selectively enrich for mRNA. The RNA was DNase treated and then sequenced using in an Illumina/Solexa system according to the manufacturer's protocols. The complete set of sequences generated by 454 and Illumina sequencing is available from the short read archive at the National Center for Biotechnology Information (NCBI) under the accession number SRX015452. The contigs of the transcriptome backbone are available as Supplemental file.

Backbone Construction and Analysis of Transcription Levels

The 454 sequences were assembled using Newbler (454/Roche) with default settings. The 36mers from the Illumina sequencing were considered to match 454 contig data if at least 30 of the 36 bases were identical to a sequence in the 454 contig data set. All other 36mers were assembled using MIRA (http://www.chevreux.org/projects_mira.html). The resulting contigs were checked for overlaps with the 454 contig data set and, if present, fused together to the "transcriptome backbone" data set. To analyze the expression level associated with individual contigs, all Illumina reads were used as input for the Illumina sequence mapping pipeline and mapped against the transcriptome backbone. Expression data were normalized by applying a correction factor derived from the division of the number of reads from the dark library with the number of reads from the light library.

Expression levels were grouped into four bins based on the sum of matching Illumina reads from both treatments. 1) No expression: transcriptome backbone sequences without or with up to 99 matches, 2) low expression genes: 100–999 matches, 3) moderate expression genes: 1,000–1,999 matches, and 4) high expression genes: 2,000 or more matches. Because genes were partly fragmented into several contigs and expression data have not been normalized to the size of the transcripts, these categories do not precisely reflect gene expression levels. However, because we used only four categories for definition, meaningful fragment lengths differ by a factor of approximately 5, and there is a slight bias of increased sequence density toward the 3′ end of a transcript so that we can expect that the length effect is ameliorated, we used these categories as an approximation of expression levels of nuclear-encoded genes.

Classification of Contigs Based on Their Presumptive Origin and Inference of EGT

Contigs of the transcriptome backbone were classified into possible EGT candidates, nuclear genes, chromatophore-

encoded genes, and bacterial contamination by Blast analysis using default parameters. The complete refseq database from NCBI was used for finding general similarities; to detect bacterial contamination, the subset of the microbial proteins refseq database was used. Sequences identical to those present in the previously analyzed chromatophore genome were grouped as chromatophore transcripts. Sequences that yielded a cyanobacterium as best BlastX hit with a score >95 but that were not classified as chromatophore-encoded genes were regarded as possible EGT candidates and were extracted for more detailed analyses (see below). Contigs with a score of >150 for a prokaryote gene but without a significant hit to cyanobacterial genes (score <150) were considered as potential bacterial contaminants. For the rest of the contigs, we made a further subdivision into nuclear genes with (nuclear with DB [score >150]) and without database hit (nuclear without DB).

To ensure a cyanobacterial origin of the broad set of possible EGT candidates, only those contigs were considered further, which yielded more than 15 cyanobacteria (or plastid-containing organisms) among the best 30 BlastX hits when the sequences were compared with the non-redundant (nr) protein database from NCBI using default settings. This arbitrary criterion exploits the fact that there are currently more than 59 completely sequenced cyanobacterial genomes. This approach was considered more suitable than regarding the first three or five best hits because α-cyanobacteria are known to have obtained numerous genes by HGT from proteobacteria (Marin et al. 2007; Reyes-Prieto et al. 2010) and the chromatophore branches at the base of the α-cyanobacteria. Thus, matches to chromatophore-derived genes might include, within the top Blast hits, proteobacterial sequences. Contigs fulfilling our criterion were integrated into a global cyanobacterial amino acid sequence alignment (for taxon sampling, see supplementary table S1, Supplementary Material online). When similar noncyanobacterial sequences were available, eubacterial and eukaryotic sequences were also downloaded from NCBI and integrated into the alignments. Multiple sequence alignments were generated using ClustalW and refined manually. Unambiguously alignable sequence blocks that were covered by contigs were manually extracted from the alignments and used for phylogenetic analysis (alignments available on request from the authors). Protein phylogenies were inferred by maximum likelihood (ML) analyses using PhyML v2.4.4 (Guindon and Gascuel 2003) within the framework of a model of amino acid sequence evolution determined with ProtTest v1.3 software (Abascal et al. 2005). The robustness of branches was tested by bootstrap analysis using 500 replicates.

Computational Prediction of Protein Subcellular Localization

For the prediction of protein subcellular localizations, the following Web-based tools were used: SignalP (Emanuelsson et al. 2007), TargetP (Emanuelsson et al. 2007), Signal-CF (Chou and Shen 2007), and Protein Prowler (Bodén and Hawkins 2005).

Determination of Genomic Sequence of EGT Candidates

Paulinella chromatophora cultures were washed in sterile culture medium three times by differential sedimentation. Genomic DNA was extracted using the Qiagen DNeasy Plant Mini Kit. Polymerase chain reaction (PCR) primers were designed based on cDNA sequence information (see supplementary table S2, Supplementary Material online), and PCRs were performed testing various annealing temperatures and elongation times for each primer pair on a gradient thermocycler (T Gradient; Biometra, Göttingen, Germany). PCR products were sequenced using the ABI sequencing system (Big Dye Terminator Kit v1.1; Applied Biosystems, Foster City, CA).

Rapid Amplification of cDNA Ends–PCR

To determine full-length cDNA sequences of several EGT candidates, rapid amplification of cDNA ends (RACE) with PCR was performed on Trizol extracted *P. chromatophora* RNA using the GeneRacer Kit and Superscript III reverse transcriptase (both: Invitrogen, Carlsbad, CA). Primers used for PCR are listed in supplementary table S2, Supplementary Material. If no single bands could be obtained for RACE-PCRs followed by agarose gel electrophoresis, largest band was eluted from the gel, cloned into pCR4-TOPO (Invitrogen, Carlsbad, CA), and introduced into chemically competent *Escherichia coli* cells (One Shot TOP10; Invitrogen). Colony PCR was performed on 10 colonies that grew on selective lysogeny broth agar plates; for these reactions, the same primers as for RACE-PCR were used. If electrophoresis revealed fragments of different lengths, only clones containing the longest fragment were used for sequencing.

Results

Construction of a Reference Transcriptome Data Set

To assess the extent of chromatophore-derived EGT in *P. chromatophora*, we generated 125 Mb of raw data by 454 sequencing of a normalized *P. chromatophora* cDNA library. Assembly yielded 27,073 contigs comprising more than 10 Mb of unique sequence, making the mean coverage of all contigs approximately 10×. None of the randomly inspected contigs exceeded 20× coverage, indicating that the normalization procedure was effective. Illumina sequencing of two *P. chromatophora* cDNA libraries generated from light- and dark-induced cells, respectively (see Materials and Methods), yielded another 544 Mb of raw data. By fusing 454 and Illumina sequences (for details, see Materials and Methods), a large reference transcriptome data set containing a total of 32,012 contigs comprising more than 12 Mb of unique sequence with an average contig size of 386 bp was generated (table 1); this data set is referred to as the "transcriptome backbone." The observed number of contigs is too large to reflect an accurate number of unique transcripts. Fragmentation of the transcriptome data is caused by 1) too few reads and/or uneven read coverage of mRNA sequences, 2) alternatively spliced mRNAs from the same locus, and 3) different transcription

Table 1. The Backbone Sequences for the Mapping of Transcriptome Data.

	Number	Bases
Normalized library (reads from 454 run)	508,943	125,512,073
Normalized library (contigs from 454 run)	27,073	10,454,782
Differential expression libraries (reads from Illumina runs)	15,121,472	544,372,992
Additional contigs from MIRA assembly of Illumina read without 454 backbone match	6,123	2,233,482
Overlapping contigs	690	505,726
Illumina only contigs (after connecting contigs with overlaps of >20 bases)	5,177	1,727,756
All contigs of the transcriptome backbone	32,012	12,363,940

start and stop sites of individual transcripts from the same locus. The transcript variability generated by situations (2) and (3) poses problems for current assembly programs, which tend to assemble such reads into nonoverlapping contigs (see, e.g., Newbler user manual).

To estimate coverage of the *P. chromatophora* transcriptome by our data, we extracted ribosomal protein sequences from the red alga *Cyanidioschyzon merolae* for which a complete nuclear genome sequence is available and blasted them against the *P. chromatophora* transcriptome backbone. Of the 77 core proteins of the large and small ribosomal subunits, 92% could be detected in the transcriptome backbone.

EGT and HGT

Because we used a nonaxenic culture of *P. chromatophora*, which was only checked for the absence of cyanobacteria (and of eukaryotes), it is not possible to distinguish bacterial contamination from true HGT for those genes that do not have a clear cyanobacterial origin. Thus, we focused on two possible transfer events: genes transferred from the chromatophore to the nuclear genome (EGT) and genes transferred from the genome of cohabitating cyanobacteria to the *P. chromatophora* nuclear genome (HGT). To identify both classes of genes, we determined contigs with the highest similarities to cyanobacterial sequences by BlastX analysis of the contigs comprising the transcriptome backbone. A low Blast score cutoff of 95 was applied to enable detection of sequences encoding small polypeptides and short protein fragments generated from contigs truncated within the coding sequence. This cutoff was empirically determined to yield a reasonable balance between sensitivity and specificity. By increasing the cutoff, true EGT candidates were gradually lost, whereas decreasing the cutoff resulted in accumulation of false positives, that is, sequences similar to a cyanobacterial sequence by chance alone. The Blast score cutoff of 95 yielded a list of 229 contigs of tentative cyanobacterial origin. To extract those sequences for which the cyanobacterial hit indicates a biological relationship rather than a spurious similarity, only those contigs with more than 15 cyanobacteria (or plastid-containing organisms) among the best 30 BlastX hits were further considered. This criterion further reduced the number of possible EGT candidates to 83 contigs. Random tests with the other 146 contigs confirmed that these contigs were of mixed origin and had no clear cyanobacterial affiliation.

Phylogenetic analyses revealed an α-cyanobacterial origin for 43 of the remaining 83 contigs (table 2 and fig. 1 and supplementary fig. S1, Supplementary Material online; nucleotide sequences of these contigs are provided in supplementary table S3, Supplementary Material online). For another ten contigs of cyanobacterial origin, no clear α-cyanobacterial origin was observed (see below). For the remaining contigs, a cyanobacterial origin could either be excluded (4 contigs) or phylogenetic relationships could not be determined unambiguously (26 contigs) due to either a lack of sufficient phylogenetic signal, strongly divergent sequences leading to long branches, which are difficult to position in phylogenetic analyses, or frequent HGTs within bacteria, including the cyanobacteria.

The 43 contigs with α-cyanobacterial affiliation matched to at least 32 different genes, hereafter referred to as "EGT candidates" (table 2). Nonoverlapping contigs that aligned with the same gene (see EGT candidates no. 21, 22, and 27 in table 2) were regarded as derived from a single gene and subjected to concatenated analysis because the trees for each of the contigs did not conflict with each other (data not shown), although existence of two gene copies with similar sequences cannot be excluded. Other contigs clearly represented different diverged copies of the same gene: there are two genes encoding photosystem I subunit X (PsaK), and two genes encoding a hypothetical protein ortholog of WH5701_10105, with both represented in the data set by two nonoverlapping contigs. There are also several genes encoding members of the high light–inducible (Hli) protein family: 11 of the sequences appear to be full length (or almost full length), whereas 3 are partial. The latter were, owing to their short overlap with the global alignment, excluded from further analysis. One additional *hli* copy, which had not been identified by the screening procedure, was identified by TBlastN searches of cyanobacterial *hli* sequences against the transcriptome backbone. Manual inspection of the *hli* assemblies revealed that numerous contigs were composed of several distinct versions of *hli* genes that differed at only a few nucleotide positions. Thus, a copy number of 15 is an underestimation of the real number of *hli* genes present in the transcriptome backbone.

Interestingly, the 32 EGT candidates found in the transcriptome backbone showed a strong bias toward short gene products (29 proteins are $\leq$250 amino acids based on the *Synechococcus* sp. WH5701 sequences and 24 have $\leq$106 amino acids). Because the majority of EGT candidates encode small polypeptides and some contigs do not cover the entire gene, many of the EGT candidates exhibited alignments of short lengths (15 of the 19 alignments were $<$80 amino acids), which resulted in low resolution in phylogenetic trees. Nevertheless, many EGT candidates were

Table 2. Description of Contigs Identified as EGT Candidates.

Gene No.	Tree in Figure[a]	Support for α-Cyanobacterial Affiliation[b]	Gene	Function	Deduced aa[c]	Length in WH5701 (aa)[d]	Copy in the Chromatophore[e]	Contig[f]	Length of Contig (nt) [g]	% GC (coding)[h]	Expression Level[i]	Up/ Downregulation[j]	PCR on Genomic DNA[k]
1	*1A*	—	PsaE	Photosynthesis	68c	69	—	C23634	333	56	High		+
2	S1A	74	PsaK	Photosynthesis	79c	85	—	C23556	231	63	High		+
3	S1A	74	PsaK, 2. copy	Photosynthesis	78c	85	—	C16172	318	65	High		+
4	*1B*	98	Ycf34, 15 aa C-terminal truncation	Unknown function in chloroplasts and cyanobacteria	68c	83	—	N52568	392	54	Low		+
5	*1C*	—	CsoS4A	Carbon concentration	76c	94	PCC_0912	C25650	408	52	Low		+
6	S1B	—	PsbN	Photosynthesis	43c	46	—	C02166	410	54	High		+
7	S1C	70	HP: WH5701_06721	Cytochrome b_6f complex	70c	77	—	C26707	315	52	High		+
8	S1D	85	CP12 domain-containing protein, 16 aa C-terminal truncation	Light regulation of Calvin cycle	58	87	—	C23188	413	48	Low		
9		—	Hli		68			C26523	498	55	High		
10		—	Hli, 2. copy		66c			C01252	771	52	Low	Up	
11		—	Hli, 3. copy		63			C26729	406	54	High		
12		—	Hli, 4. copy		48c			C26846	239	54	Low		
13		—	Hli, 5. copy		66			C26678	199	55	High	Up	
14		—	Hli, 6. copy		48c			C01765	347	50	High	Up	
15	S1E	—	Hli, 7. copy	Photoacclimation	61	40		C25348	476	52	High		
16		—	Hli, 8. copy		49	—	—	C02225	216	49	Low		
17		68	Hli, 9. copy		63c	106		C27015	233	53	High	Up	
18		—	Hli, 10. copy		67			C00119	231	50	Moderate		
19		—	Hli, 11. copy		56c			C22247	398	54	High		
20		—	Hli, 12. copy, found by TBlastN, see text		56c			C26233	338	53	High		
			Partial Hli, 13. copy		33			C23908	268	55	High	Up	+
			Partial Hli, 14. copy		27			C25576	214	56	High		
			Partial Hli, 15. copy		56			C18116	391	50	High		
21	S1F	100	RecG	DNA recombination and repair	320	862	—	C13970	254	54	No		—
								C08095	217	47	No		
								C04944	392	56	No		
								C06988	201	48	No		

Table 2. Continued

Gene No.	Tree in Figure[a]	Support for α-Cyanobacterial Affiliation[b]	Gene	Function	Deduced aa[c]	Length in WH5701 (aa)[d]	Copy in the Chromatophore[e]	Contig[f]	Length of Contig (nt) [g]	% GC (coding)[h]	Expression Level[i]	Up/ Downregulation[j]	PCR on Genomic DNA[k]
22	*1D*	99	Glutathione S-transferase	Protection against oxidative stress	367	417	—	N101035	284	51	Low		
								N30457	440	49	No		
								N65410	502	51	Low		—
								N99978	331	46	Low		
23	S1G	100	HP: WH5701_06946	Predicted amine oxidase/zeta-carotene desaturase	82	420	—	C02152	543	59	Low		
24	S1H	63	HP: WH5701_05850	Unknown	20	45	—	C26044	218	51	Low		
25	S1I	99	HP: WH5701_10105	Unknown	352	250	—	C16165	1057	46	Moderate		—
26			HP: WH5701_10105, 2. copy		111			C09104	334	46	Low		
			HP: WH5701_10105		37			C24377	113	46	No		
			HP: WH5701_10105, 2. copy		59			N39869	488	54	No		
27	S1J	90	HP: WH5701_09855	Unknown	151	204	—	N14067	332	58	No		—
								C15058	181	51	Low		
28	S1K	Occur only in α-cyanobacteria	HP: SYNW0763	Unknown	55c	80	—	C25354	184	51	Low		
29	S1L		HP: WH5701_13905	Unknown	59c	84	PCC_0022	C01215	715	51	High		
30	S1M		HP: WH5701_09129	Unknown	72	179	—	C04492	220	50	No		
31	S1N		HP: WH5701_13415	Unknown	69c	71	—	C04807	357	50	Moderate		—
32	S1O		HP: WH5701_08969	Unknown	59	161	—	C25548	226	48	Low		

NOTE.—aa, amino acid; nt, nucleotide; Hli, high light–inducible protein; HP, hypothetical protein.

[a] Indicates in which figure results of the ML analyses are presented. Main figures in bold; supplementary figures: in regular font.

[b] Bootstrap support (>50%) for α-cyanobacterial affiliation of the respective contig.

[c] Provides the lengths of the protein as deduced from the contig; a "c" following the aa number indicates that the deduced protein sequence is presumably complete.

[d] Length of the complete protein in the reference α-cyanobacterium *Synechococcus* sp. WH 5701.

[e] If a chromatophore-encoded copy of the respective gene exists, the locus tag is given.

[f] For each contig, identifier is indicated.

[g] For each contig, contig length is indicated.

[h] For each contig, average GC content in coding regions is indicated.

[i] Expression level of the contigs classified in high, moderate, low, and no expression (see experimental procedures).

[j] Up- or downregulation (>2-fold) on the onset of the light period.

[k] Contigs for which PCR was performed on genomic DNA are highlighted by a + (or a—when amplification failed).

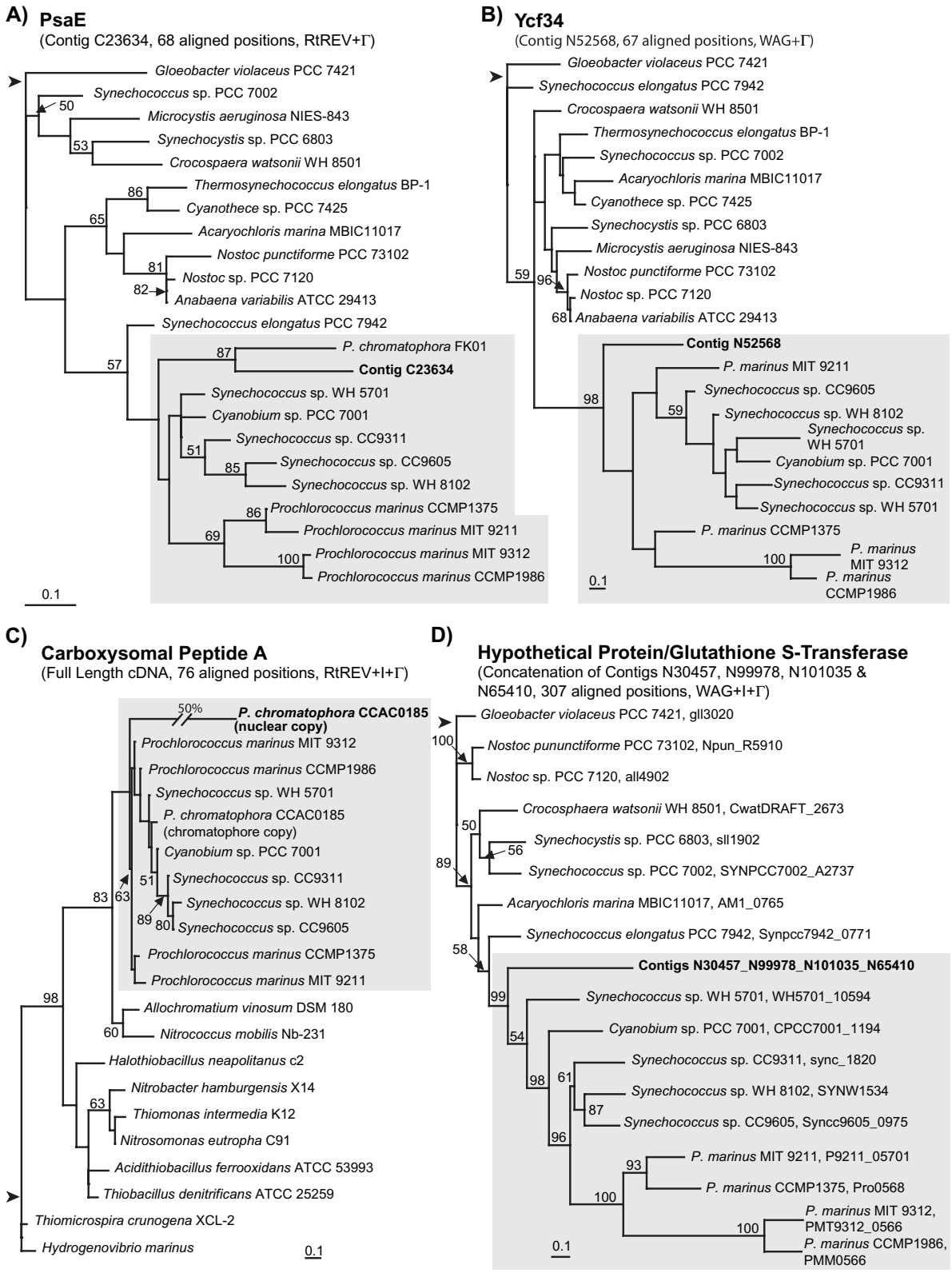

FIG. 1. Evidence for α-cyanobacterial origin of EGT candidates. (*A*) PsaE, (*B*) Ycf34, (*C*) CsoS4A, and (*D*) a glutathione S-transferase. The ML trees were inferred from amino acid sequences derived from the *Paulinella chromatophora* transcriptome backbone, cyanobacterial, and proteobacterial genomes. Number of aligned amino acid positions as well as evolutionary model of amino acid substitution used for the analyses is provided in brackets in the headings of each tree. Numbers at branches represent ML bootstrap values ≥50%. Strain designations (and for hypothetical proteins: gene identity tags) are indicated after the species name. α-cyanobacteria are shaded in gray. Arrowheads highlight the root of the trees. Nucleotide sequences of contigs of α-cyanobacterial origin are provided in supplementary table S3, Supplementary Material online.

recovered in the ML tree topology at the base of the α-cyanobacteria (fig. 1 and supplementary fig. S1*A*, *E*, and *J*, Supplementary Material online), which represents the phylogenetic position expected for genes of chromatophore origin (Marin et al. 2005, 2007). Furthermore, five contigs are homologs of genes that encode hypothetical proteins that exclusively affiliated with α-cyanobacteria, that is, BlastP searches against the nr database at NCBI using the full-length proteins of *Synechococcus* strain WH5701 (or, if not available, of *Synechococcus* strain WH8102), and default settings yielded no hits in β-cyanobacteria (table 2 and supplementary fig. S1*K*–*O*, Supplementary Material online).

Contigs with cyanobacterial affiliation but no clear α-cyanobacterial affiliation matched to ten genes (supplementary table S4, Supplementary Material online) that are hereafter referred to as "HGT candidates." Although the phylogenetic position within the cyanobacteria for most of these genes could not be resolved (supplementary fig. S2, Supplementary Material online), two of them—one annotated as a putative esterase/lipase/thioesterase and the other as a hypothetical protein (homolog to sll0364 of *Synechocystis* sp. PCC6803)—clearly clustered with a specific clade/species of cyanobacteria: the putative esterase/lipase/thioesterase formed a well-supported sister group (bootstrap: 100%) with the non-N_2-fixing unicellular β-cyanobacterium *Synechococcus* sp. PCC7002 (supplementary fig. S2*A*, Supplementary Material online), whereas the sll0364 homolog was a member (bootstrap: 98%) of a clade of filamentous, N_2-fixing cyanobacteria (supplementary fig. S2*F*, Supplementary Material online). In contrast to the EGT candidates, HGT candidates that we identified did not share the bias toward short gene products, but the deduced average length of the ten proteins is 305 amino acids (based on homologous proteins from *Synechococcus* WH5701; supplementary table S4, Supplementary Material online).

Whereas the large majority of transferred genes are not represented by a chromatophore-encoded copy, for two EGT and two HGT candidates, a second gene copy is present on the chromatophore genome. The EGT candidates are the carboxysomal peptide A (*csoS4A*, formerly *orfA* or *pepA*), a gene obtained by the ancestor of the chromatophore from a proteobacterium by HGT (see outgroup in fig. 1C and Marin et al. 2007), and EGT candidate no. 29 encoding a hypothetical protein of unknown function (table 2). The HGT candidates encode phosphoglycerate kinase (*pgk*) and inositol monophosphatase (supplementary table S4, Supplementary Material online). Three of the four transferred genes with counterparts in the chromatophore displayed highly divergent sequences compared with those of the chromatophore and cyanobacteria. This is evident for CsoS4A and Pgk, based on their long branches in phylogenetic trees (fig. 1C and supplementary fig. S2*E*, Supplementary Material online), and for EGT candidate no. 29 based on its alignment to the chromatophore and cyanobacterial sequences, which is presented in the supplementary figure S1*L*, Supplementary Material online (a phylogenetic tree was not informative owing to very high sequence conservation of this gene within α-cyanobacteria that resulted in very short or zero length branches).

Functions of Transferred Genes

Intriguingly, the EGT candidates display a conspicuous functional bias. The majority of proteins for which a function is known are involved in photosynthesis. The Hli proteins are involved in managing absorbed excitation energy or serve as chlorophyll carriers and are required for viability of cyanobacteria in high light (Montané and Kloppstech 2000; He et al. 2001; Bhaya et al. 2002; Xu et al. 2004). Other EGT candidates are involved in electron transport with regulatory rather than structural functions; these include PsaE (Barth et al. 1998), PsaK (Fujimori et al. 2005; Düring et al. 2007), and the homolog to WH5701_06721 (Volkmer et al. 2007). The exact function of PsbN, which is likely not a subunit of PSII but may bind transiently to it (Kashino et al. 2002; Plöscher et al. 2009), is still unclear; nonetheless, there is a dedicated nuclear-encoded sigma factor associated with the plastid-encoded RNA polymerase that is known to specifically regulate transcription of the plastid genome-localized *psbN* gene in *Arabidopsis thaliana* (Zghidi et al. 2007), suggesting an important regulatory function of the gene product. CsoS4A is crucial for carboxysome function as a CO_2 leakage barrier (Cai et al. 2009) and thus enhances performance of RubisCO. CP12 domain-containing proteins mediate light regulation of Calvin cycle activity in an nicotinamide adenine dinucleotide phosphate-dependent manner in β-cyanobacteria and the Plantae (Wedel and Soll 1998; Trost et al. 2006). However, only the C-termini of CP12 domain-containing proteins of α-cyanobacterial can be aligned with β-cyanobacterial proteins (see the long branch in supplementary fig. S1*D*, Supplementary Material online). Furthermore, the EGT candidate homolog to CP12 domain-containing proteins of α-cyanobacteria is C-terminally truncated and lacks the highly conserved CP12 domain, making its function questionable. Notably, all the aforementioned functions are localized inside of the chromatophore. Only DNA helicase RecG, the putative glutathione S transferase, and the putative carotene desaturase, which have 862, 417, and 420 amino acids, respectively, in *Synechococcus* WH5701, and are thus much longer than the other EGT candidates, do not directly conform to the same functional profile. However, it should be noted that RecG, glutathione S transferase, and carotenoid desaturase may all be involved in ameliorating the potential damaging effects of absorbing excess light energy. The cellular functions of the other eight EGT candidates are currently unknown.

The same functional bias is not obvious for the HGT candidates. Proteins encoded by the HGT candidates, where a function is known, are involved in various metabolic processes, but not in photosynthetic processes (supplementary table S4, Supplementary Material online).

Adaptation of EGT Candidates to the Nuclear Environment

Our data set displays clear differences in nucleotide composition of chromatophore-encoded and nuclear-encoded

genes (supplementary table S5, Supplementary Material online). The nuclear genes have an average GC content of 49.6% (55.3% GC in the third codon position). Within the nuclear genes, highly expressed genes—represented by ribosomal protein genes—have an average GC content of 52.8% (64.4% GC in the third codon position), whereas genes that are expressed at lower levels have an average GC content of 49.7% (55.0% GC in the third codon position). In contrast, chromatophore-encoded genes have an average GC content of 40.4% (with the very low GC content of 26.7% in the third codon position). This is typical for endosymbiotic prokaryotes and differs from most free-living cyanobacteria (e.g., GC content of protein-coding genes of *Synechococcus* sp. WH5701 is 66%). Contrary to chromatophore-encoded genes, coding regions of our EGT candidates display markedly higher GC contents, with a range of 65–46% (table 2; average: 52.5%, with 57.5% GC in the third codon position) and are thus much closer in nucleotide composition to host gene sequences. We found no clear-cut nucleotide ratio criterion that would allow us to distinguish contaminating bacterial sequences from HGT and other nuclear genes.

To examine whether transferred genes have adapted to the nuclear environment by acquisition of introns, we amplified a number of EGT candidates from genomic DNA and determined their nucleotide sequences. Despite repeated attempts with various primer combinations, only 8 of 13 EGT candidates tested (table 2) yielded PCR products. For the six products representing *psaE*, both copies of *psaK*, *ycf34*, *csoS4A*, and the hypothetical protein homolog to WH5701_06721, introns were detected within the amplified region. The introns range from 74–335 bases (fig. 2*B*). The donor and acceptor sites have the canonical GT/AG residues that are recognized by the spliceosomal complex. In addition, only G at the last exon position of the donor site is conserved (supplementary fig. S3, Supplementary Material online). A nuclear copy of *psaE* was previously described in *P. chromatophora* strain FK01 (Nakayama and Ishida 2009). Interestingly, the intron positions in the two strains differ (fig. 2*A*). Moreover, only the coding portions of the genes can be aligned; intron sequences as well as 5′ and 3′ untranslated regions (UTRs) are completely different. As exemplified by the intron/exon structure of *csoS4A*, numerous introns may occur within a transferred gene, fragmenting the coding sequence into short exons (fig. 2*C*).

Because most proteins encoded by EGT candidates in *P. chromatophora* have a function located in the chromatophore, the mechanism of potential targeting of these proteins to the chromatophore is of great interest. To identify possible presequences of EGT candidates that could mediate targeting of the proteins into the chromatophore, we analyzed nine genes for which sequence information 5′ of the presumed translation start site was either obtained by RACE-PCR or was covered by a contig of the transcriptome backbone (fig. 3). None of the transcripts analyzed had a classic translation initiation site (TIS), that is, an AUG codon, 5′ of the presumable start codon of the mature protein (determined from alignments with homologous cyanobacterial sequences). Possible alternative TISs (CUG, GUG, UUG, AUA, or ACG) were identified in only five of nine EGT candidates between the ATG of the mature protein and either the 5′ end of the sequence or the first in-frame stop codon (fig. 3). Potential presequences identified based on these alternative TISs were relatively short (15-1 amino acid). All nine possible presequences identified in this manner were analyzed using bioinformatic tools to determine if the putative presequences could predict the subcellular location of the mature protein (supplementary table S6, Supplementary Material online). Based on the longest possible presequences associated with PsbN and the *Synechococcus* WH5701_13415 homolog, signatures of endoplasmatic reticulum (ER)-targeting signal peptides (SPs) were found using TargetP (Emanuelsson et al. 2007). For the *Synechococcus* WH5701_13415 homolog, these predictions were confirmed by SignalP (Emanuelsson et al. 2007), Signal-CF (Chou and Shen 2007), and Protein Prowler (Bodén and Hawkins 2005). For PsbN only, Protein Prowler confirmed the TargetP prediction, whereas SignalP suggests that PsbN contains an uncleaved internal ER localization signal (signal anchor) and SignalCF predicts that PsbN is not a secreted protein. For both the PsbN and the *Synechococcus* WH5701_13415 homolog, the predicted SPs are 19 amino acids long, that is, cleavage sites are located well within the conserved core protein sequences (supplementary table S6, Supplementary Material online). For other proteins, either no predictions were obtained or, in the case of PsaE, a weakly supported mitochondria-targeting transit peptide with a cleavage site close to the C-terminal end of the protein was identified. Together, these results suggest that there is no common presequence feature associated with the EGT candidate sequences, making it unclear if and how these presumed chromatophore-targeted proteins enter the organelle and find their sites of function.

Analysis of Differential Expression

To determine expression levels of *P. chromatophora* nuclear genes, we sequenced cDNA libraries from light- and dark-induced cells on an Illumina machine. Mapping the resulting 36mer sequences back to the transcriptome backbone enabled us to characterize the transcriptional state of *P. chromatophora* under defined conditions without prior knowledge of the nuclear genome sequence (table 3). We found at least one matching Illumina sequence for more than 25,000 contigs (84%). Interestingly, all EGT candidates that encode proteins involved in photosynthesis (*psaE*, both copies of *psaK*, *psbN*, and the hypothetical protein homolog to WH5701_06721, with electron transport-regulating function in the cytochrome b_6f complex in *Synechocystis* PCC 6803; Volkmer et al. 2007), as well as 11 of the 15 Hli proteins, showed high levels of transcript abundance. The remaining EGT candidates displayed moderate to low levels of transcript abundance, with the exception of EGT candidate no. 29, which encodes a hypothetical protein of unknown function (table 2).

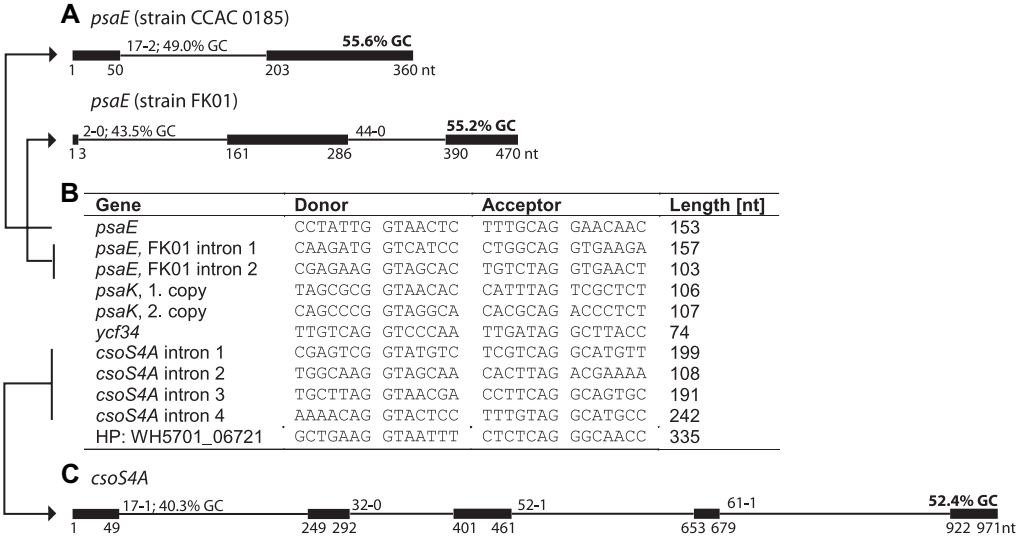

Fig. 2. Exon/intron structure of EGT candidates. Exon/intron structures of (*A*) nuclear *psaE* of different *Paulinella chromatophora* strains and (*C*) the nuclear *csoS4A* are represented schematically. Insertion positions of introns are indicated above the graphical representation of introns (number of amino acid—frame). Average GC contents are given for exons (bold, at the end of each gene) and for introns (regular font, above the first intron of each gene). Numbers below the graphical representation of the genes denote nucleotide positions of exon borders counting from the start codon of each gene. In (*B*), sequences for donor and acceptor sites as well as intron lengths are listed for all introns characterized in this study.

Although light is essential for growth of photosynthetic organisms, excess light energy can lead to severe damage of the cell. We therefore reasoned that major changes in gene expression linked to chromatophore function and light acclimation would occur after the switch from dark to light conditions. However, a global comparison of expression levels revealed that most light/dark expression data pairs cluster along a straight line with higher variability toward the genes expressed at lower levels (supplementary fig. S4, Supplementary Material online). However, major differences in expression levels between the two treatments were not observed. Only 88 contigs were found to have been upregulated (i.e., >2fold) on the onset of light (table 3). Notably, among these are five of the transferred *hli* genes. All remaining EGT candidates showed similar levels of transcripts both before and after the onset of light (table 2).

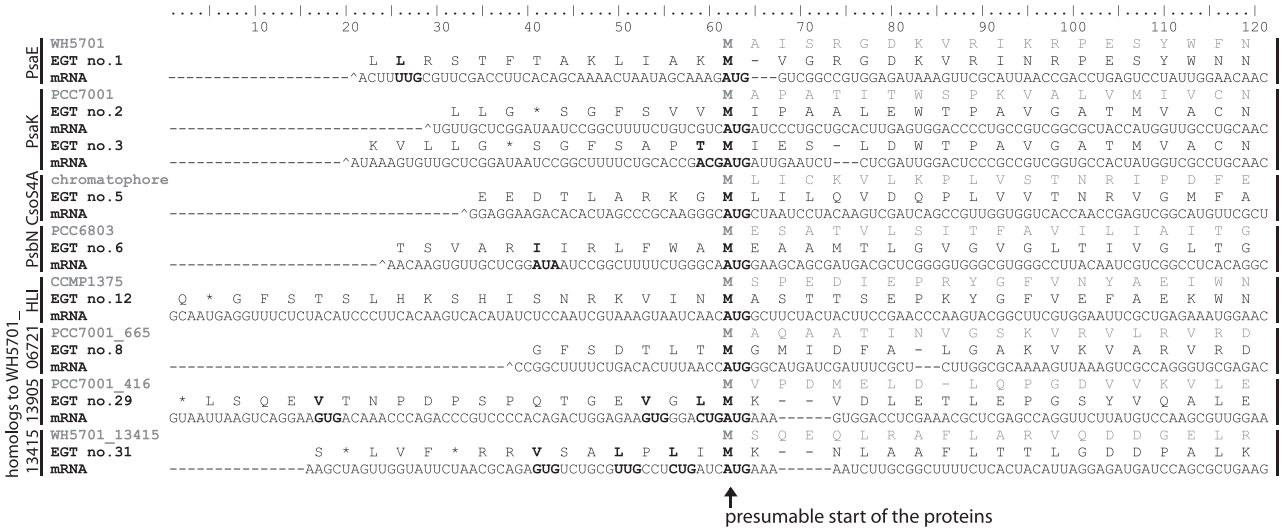

Fig. 3. Potential presequences of EGT candidates. Genes for which sequence information 5′ of the presumable start codon of the mature protein was either obtained by RACE PCR or covered by a contig of the transcriptome backbone were translated up to the end of the sequence available. Presumable starts of the respective proteins (in bold) are displayed aligned with homolog cyanobacterial or chromatophore sequences (in gray). Because N-termini of the proteins vary in length between cyanobacteria, cyanobacterial strains were chosen that encode a homolog protein more or less conserved in length (strain numbers: CCMP1375, *Prochlorococcus marinus*; PCC6803, *Synechocystis* sp.; PCC7001, *Cyanobium* sp.; and WH5701, *Synechococcus* sp.). mRNA sequences are given underneath the translation. "^" indicates the start of a sequence obtained by RACE-PCR, stop codons in the RNA translate into * in the amino acid sequence. Alignment gaps are represented by "-." Start codons as well as possible alternative TISs are highlighted in bold.

Table 3. Differential Gene Expression of the Transcriptome Backbone.

	Contigs of Transcriptome Backbone	Number of Contigs with Illumina Matches			Number of Contigs without Any Illumina Match (%)	Number of Contigs in Classes of Expression Levels Based on Overall Number of Illumina Matches				Number of Contigs Showing Differential Regulation	
		Exclusively under Light Conditions	Exclusively under Dark Conditions	Under Both Conditions		Low (%)	Moderate (%)	High (%)	Very low or no (%)	>2× up (% of l + m + h)	>2× down (% of l + m + h)
Nuclear with database hit (score >150)	4,851	6	17	4,025	803 (16.5)	2,056 (42.4)	319 (6.6)	470 (9.7)	2,006 (41.1)	7 (0.2)	46 (1.6)
Nuclear without database hit	25,187	468	875	19,730	4,114 (16.3)	6,111 (24.3)	596 (2.4)	713 (2.8)	17,767 (70.5)	67 (0.9)	179 (2.4)
Tentative cyanobacterial candidates	229	3	4	186	36 (15.7)	84 (36.7)	11 (4.8)	28 (12.2)	106 (46.3)	5 (4.1)	1 (0.8)
Chromatophore genome	809	10	17	737	45 (5.6)	266 (32.9)	23 (2.8)	30 (3.7)	490 (60.6)	5 (1.6)	23 (7.2)
Potential bacterial contamination	936	6	9	700	221 (23.6)	365 (39)	58 (6.2)	59 (6.3)	454 (50.5)	4 (0.8)	7 (1.5)
Sum	32,012	493	922	25,378	5,219 (16.3)	8,882 (27.7)	1,007 (3.1)	1,300 (4.1)	20,823 (65.0)	88 (0.8)	256 (2.3)

Note.—l + m + h, low + moderate + high.

Discussion

Generation of a Comprehensive *P. chromatophora* Reference Transcriptome Data Set

Analysis of transcriptomes from organisms for which there is no nuclear genome sequence is challenging. We therefore employed a novel strategy to cover a large fraction of the transcriptome of *P. chromatophora*. Sequencing a normalized library constructed from *P. chromatophora* RNA using 454 technology captured even rare transcripts as demonstrated by the high percentage of contigs yielding no or very low numbers of matches with the (nonnormalized) Illumina reads. The merged assembly of 454 and Illumina reads yielded contigs with approximately 12 Mb of unique sequence. The mean coding capacity of a eukaryote genome ranges from 5,000 to 20,000 coding sequences (CDS), with a mean CDS length of 1,500 bases (see, e.g., Eichinger et al. 2005). Thus, a data set covering a complete eukaryotic transcriptome should be in the range of 7.5–30 Mb. Because not all CDS are transcribed and polyT-primed transcript data are incomplete, with a bias of increased sequence density toward the 3′ end of the transcript, we can assume that we have assembled at least parts of nearly all transcribed genes in our data set. This assumption is further supported by a recent study in which 4.1 Mb of transcript data of the mosquito *Aedes aegyptii* mapped to 27% of the reference transcriptome (Gibbons et al. 2009). If there are a comparable number of CDS for the *P. chromatophora* nuclear genome as for the *A. aegyptii* genome (15,419 CDS), our data set would encompass about 80% of the *P. chromatophora* transcriptome. The finding that the transcriptome backbone covers >90% of the expected eukaryotic ribosomal proteins, also supports a wide-ranging coverage of the transcriptome by our sequencing approach.

Within the transcriptome data, we identified more than 900 contigs with matches to genes in bacteria other than cyanobacteria. Only one of the genes represented by these contigs is highly expressed and appears to be induced by light by more than 2-fold. This gene codes for a carbonic anhydrase, which might be involved in concentrating CO_2 for optimization of photosynthetic reactions (data not shown). Thus, although the majority of these contigs is likely derived from bacterial contaminants in the culture, some might originate from HGT and be integrated into the nuclear genome of *P. chromatophora*. Because the bacterial kingdom is sufficiently covered to enable detection of the majority of genes of bacterial origin in the data set, the remaining more than 25,000 transcript units (if not classified as chromatophore-encoded or tentative cyanobacterial candidate) were regarded as derived from the nuclear genome. However, we cannot exclude the possibility that some of the contigs are derived from bacteria that have not yet been characterized. Because no Rhizaria genome is currently available for comparison, many protein-coding gene sequences specific to this clade will not be detected by our approach. Some transcribed units may also represent 5′ or 3′ UTRs, which are not well conserved even in closely related species.

Extent of EGT in *P. chromatophora*

Our screening for EGTs in *P. chromatophora* significantly extended the record of EGTs from the chromatophore to the nuclear genome from two genes reported previously (Nakayama and Ishida 2009; Reyes-Prieto et al. 2010) to a minimum of 32 genes. This finding demonstrates that the transfer of genes from the chromatophore to the nucleus is a general trend in the *P. chromatophora* system. The high GC content in coding regions of EGT candidate contigs and the presence of introns in six of the genomic sequences corroborates the nuclear localization of the EGT candidates identified. The high intron density observed in *csoS4A* offers an explanation for why no PCR amplicons could be obtained for other EGT candidates despite repeated trials (introns may disrupt potential binding sites of primers, which were designed based on cDNA sequence information).

Because our EGT inference procedure 1) relied on contigs that yielded cyanobacterial sequences as best Blast hits, 2) involved the analysis of relatively short gene fragments, and 3) excluded from the list of EGT candidates those genes for which phylogenetic affiliation remained ambiguous, thereby restricting the number of false positives, the data set presented represents a conservative estimate of the extent of EGT in *P. chromatophora*, and the 32 EGT candidates identified should be regarded as a minimal number. Conversely, it is difficult to estimate the total number of genes transferred from the chromatophore to the host nucleus for the following reasons: Although the chromatophore branches distantly from the plastids in the cyanobacterial radiation which, at the time of this study, was represented by 59 sequenced genomes (in NCBI), it is still possible that convergent evolution of cyanobacterial endosymbionts under similar selective pressures might have resulted in sequence similarities between chromatophore- and plastid-derived genes, causing photosynthetic eukaryotes to appear as a best BlastX hit for chromatophore-derived genes in some cases. These genes would fail to be identified by our inference strategy. Furthermore, the EGTs identified included highly similar copies of *hli* genes that were not distinguished by the assembly pipeline and counted as single genes. Additionally, for 26 contigs with a cyanobacterium as best Blast hit, phylogenetic relationships could not be determined unambiguously due to a lack of phylogenetic signal. Finally, a few of the HGT candidates might in fact have been derived from the chromatophore (discussed below). These considerations corroborate the conclusion that the 32 EGTs reported here is an underestimation of the total number of genes that were transferred from the chromatophore to the host nucleus. However, based on our data, it also seems unlikely that the total number of EGTs in the transcriptome backbone of *P. chromatophora* exceeds 100 genes.

Various estimates of the total number of nuclear genes that were acquired from cyanobacteria during the course of plastid evolution have been published; these range from 14% of the nuclear genome in *Arabidopsis*, rice, *Chlamydomonas*, and *Cyanidioschyzon* (Deusch et al. 2008), to 11% in *Cyanophora paradoxa* (Reyes-Prieto et al. 2006). Without a complete nuclear genome sequence of *P. chromatophora*, a comparison of the above numbers with those obtained from the *P. chromatophora* transcriptome is difficult. However, assuming that the nuclear genome of *P. chromatophora* comprises some 15,000 CDSs and that roughly 80% of these were sequenced (see above), the total number of EGTs should range between 40 and 125 (i.e., 32 to 100 $\times$ 1.25), representing 0.3–0.8% of the presumptive number of nuclear CDSs. Although the difference in the extent of EGT observed between members of the Plantae and *P. chromatophora* may in part be explained by the different methodological approaches used, it certainly also reflects the estimated 20-fold younger evolutionary age (Nowack et al. 2008) of the chromatophores compared with plastids and is in line with the 5–10 fold larger chromatophore genome size relative to that of typical plastid genomes (Nowack et al. 2008).

That EGT in *P. chromatophora* is an ongoing process is revealed by comparison with the recently completed chromatophore genome sequence of a different *Paulinella* isolate (Reyes-Prieto et al. 2010). In strain FK01, 27 chromatophore genes were identified that have been lost from the chromatophore genome of strain CCAC 0185 (conversely, 39 genes are still present in CCAC 0185 but absent from FK01). Of 19 annotated genes present on the chromatophore genome of FK01 but absent from CCAC 0185, at least two genes (a *hli* gene and *psaK*) were apparently transferred to the nucleus in CCAC 0185 and are expressed, demonstrating that these EGTs (in case of *psaK* also a gene duplication and the acquisition of spliceosomal introns) occurred after the two *Paulinella* isolates diverged from their common ancestor.

Also HGT from β-cyanobacteria Forged the Nuclear Genome of *P. chromatophora*

Apparently, HGT from cyanobacteria distinct from the ancestor of the chromatophore also contributed genes to the nuclear genome of *P. chromatophora*, although this has occurred to a minor extent. Some of the HGT candidates might in fact originate from the chromatophore but were misplaced in phylogenetic trees due to long-branch attraction artifacts (e.g., supplementary fig. S2*G* and *H*, Supplementary Material online); for others, however, a non-α-cyanobacterial origin is well supported and hence persuasive (e.g., supplementary fig. S2*A* and *F*, Supplementary Material online). In phagotrophic protists, HGT from food organisms is well documented (Doolittle 1998; Andersson 2005; Keeling and Palmer 2008). Because *P. chromatophora* had a phagotrophic existence before acquisition of the chromatophore (as its marine nonphotosynthetic relatives still do today, feeding on cyanobacteria), the HGTs identified in *P. chromatophora* likely reflect previous feeding habits of the amoeba in which the food sources may have included a diverse range of cyanobacteria. Notably, none of the proteins encoded by HGT candidates has a function that is essentially

localized in the chromatophore, which is reasonable if a presymbiotic origin of the HGTs is assumed.

Host Cells Established Control Over Expression of Transferred Genes

EGT from plastids to nuclei is an ongoing process in Plantae that is observed at surprisingly high rates (Huang et al. 2003; Stegemann et al. 2003). Furthermore, large tracts of DNA of the widespread bacterial endosymbiont *Wolbachia* have been transferred to the nuclear genome of their nematode and insect hosts (Dunning Hotopp et al. 2007; Nikoh et al. 2008). For this reason, it was not surprising to discover EGT in *P. chromatophora*. Conversely, functional establishment of the transferred genes in the nuclear genome appears much rarer. Therefore, an important question is whether proteins encoded by EGT candidates are functionally integrated into the host cell. Slow erosion of gene sequence and loss of function are conceivable for EGT candidates 1) for which a second copy is encoded in the chromatophore and that form long branches in phylogenetic analyses (*csoS4A* and the hypothetical protein EGT candidate no. 29) or 2) for genes with pronounced truncations (*ycf34* and the CP12 domain-containing protein). Still, the data set presented indicates that the large majority of EGT candidates is functional: 1) the bias in their functional categories (see below) is highly unlikely to have arisen by chance alone, as would be required for DNA sequences without functional gene products, 2) their GC signature that clearly differs from chromatophore genes but is similar to nuclear genes and the presence of spliceosomal introns provides evidence that the EGT candidates are well integrated into the nuclear genome, and 3) expression levels of EGT candidates appear to be nonrandom: Highest expression levels were found for genes encoding proteins involved in construction of the photosynthetic membranes and their protection from the absorption of excess excitation energy. This expression pattern appears to be reasonable, considering that a main function of the chromatophore in the symbiotic association would be to provide photosynthate to the host. Increases in transcript abundances from some of the *hli* genes following the onset of illumination resemble regulation patterns observed for this class of genes in free-living cyanobacteria (Salem and van Waasbergen 2004).

In plastids, performance is controlled not only by regulation of transcription of nuclear-encoded plastid-targeted genes but also by direct modulation of transcription of plastid-encoded genes. This is achieved, at least in part, by a nuclear-encoded plastid-targeted RNA polymerase and nuclear-encoded sigma factors that modulate promoter selectivity of the plastid-encoded RNA polymerase (Shiina et al. 2009). The chromatophore genome encodes only two sigma factors, much fewer than that are found in free-living cyanobacteria (Nowack et al. 2008). However, in our data set, no additional sigma factor–like genes were discovered among the EGT candidates, nor is there evidence of a nuclear-encoded chromatophore-targeted RNA polymerase. Consequently, at this early stage of symbiosis, control over chromatophore performance by the host cell seems to be mainly mediated by regulation of expression of nuclear-encoded chromatophore-targeted proteins that localize to the thylakoid membranes and have regulatory functions in photosynthetic processes.

Functional Categories of EGT Candidates Strongly Suggest Import of Gene Products into the Chromatophore

Proteins encoded by genes transferred from the plastid to the nuclear genome in the Plantae not only fulfill various functions, mostly inside plastids, but also acquired new functions in other cell compartments (Martin et al. 2002; Reyes-Prieto et al. 2006). For proteins with a function directly linked to the thylakoid membrane, acquisition of new functions outside the photosynthetic compartment appears to be highly unlikely. Hence, the presence and transcriptional regulation of a considerable number of EGT candidates in *P. chromatophora* encoding proteins with predicted localization to the thylakoid membranes strongly implies that their gene products are indeed imported into the chromatophore; a feature thought to be restricted to canonical organelles (Theissen and Martin 2006).

In the Plantae, the large majority of nuclear-encoded plastid-targeted proteins contains N-terminal chloroplast-targeting transit peptides (cTPs) (Bruce 2001; Patron and Waller 2007), which mediate the import of the protein into the plastid via the Tic–Toc complex, although a few proteins with N-terminal ER-targeting SPs may enter the plastid through an alternative mechanism that involves the secretory pathway (Villarejo et al. 2005; Nanjo et al. 2006). The chromatophore is surrounded by two membranes, separated by a peptidoglycan wall. Thus, if EGT candidates are translated in the host cytosol, their passage into the chromatophore would require the traversal of two membranes. Although the inner membrane is clearly derived from the former cyanobacterial cytoplasmic membrane, the outer membrane may be derived from the host phagocytotic membrane rather than from the cyanobacterial outer membrane (Kies 1974; Nowack et al. 2008), which differs from the origin assumed for the outer membrane of primary plastids (Duy et al. 2007). On the basis of known protein import mechanisms into plastids, two mechanisms are conceivable: proteins may be imported via an independently evolved transport system similar to the Tic–Toc complex or access of the protein to the inter membrane space may depend on the secretory pathway, with subsequent transport across the inner membrane requiring a pore similar to the Tic complex. Indeed a few components of the Tic–Toc complex with cyanobacterial origin seem to be conserved on the chromatophore genome, although other components of the complex considered essential for protein import into primary plastids appear to have been lost—and were not found in the transcriptome backbone (data not shown)—making the use of the remaining components in a homologous protein import system in chromatophores questionable (Bodył et al. 2010). Either import pathway would require the association of the EGT candidates with presequences that

function in protein translocation. For several reasons, the acquisition of targeting capability by proteins encoded by those genes transferred from an organellar to the nuclear genome is regarded as more straightforward than, for example, the attainment of regulatory information to enable suitable expression levels (Martin and Herrmann 1998): the N-terminal sequences of many bacterial genes already appear to have protein-targeting potential (Lucattini et al. 2004). Furthermore, randomly cloned fragments of *E. coli* DNA as well as eukaryotic gene fragments were experimentally shown to successfully induce mitochondrial targeting in yeast in 2.7% or 5% of the cases, respectively (Baker and Schatz 1987). Finally, known transit peptides vary considerably in length and generally show no strong consensus sequences (Emanuelsson et al. 2007). However, the amino acid sequences deduced from translation of the nucleotide sequence upstream of the conserved translation start site of the EGT candidates did not yield clear evidence for targeting capability: 1) Compared with typical cTPs that are 20–100 amino acids in length, and typical eukaryotic SPs that are 15–30 amino acids in length (Emanuelsson et al. 2007), the sequences obtained based on RACE-PCR analyses are very short. However, although RACE experiments are meant to identify full-length mRNA sequences, there is no proof that the products obtained in our analyses do represent the full-length sequences. But the lengths of the possible presequences are also limited in several of the EGT candidates by in-frame stop codons. 2) Furthermore, classical TISs, that is, in-frame AUG codons, are not present in EGT sequences upstream of the AUG that defines the start position of the mature protein. Although for bacteria, the use of alternative start codons is common (the chromatophore itself is predicted to use 11.9% GUG, 7.2% UUG, and 0.8% CUG), for eukaryotes translation initiation at non-AUG codons is less frequent. However, the use of various non-AUG codons (CUG, GUG, UUG, AUA, or ACG) as TISs, in particular for presequences, has been increasingly reported for a variety of eukaryotes including *Arabidopsis* (Christensen et al. 2005), yeast (Chang and Wang 2004; Tang et al. 2004), and mammals (Touriol et al. 2003). However, if we consider translation initiation from non-AUG TISs, only five of nine EGT candidates would have potential presequences. 3) Computational examination of these presequences revealed no sequences similar to cTPs, which is not really surprising as even if a protein import system similar to the Tic–Toc complex evolved, it might require completely different signatures for targeting. Also, SP-like signatures were found in only two of the potential presequences. However, these SP predictions are not robust because both proteins with the putative SP-like sequence were predicted to have cleavage sites within the conserved part of the protein, and SP prediction for one of the proteins (PsbN) was not consistent among the different prediction algorithms.

From the above analyses, we would have to hypothesize that the EGT candidates have unusual alternative TISs that generate nonconventional N-terminal presequences that are involved in protein targeting. Alternatively, some internal sequence information or unidentified partner proteins may be involved in protein targeting. To elucidate the mechanism critical for the passage of proteins into the chromatophore would require biochemically oriented experimentation.

It is noteworthy that the EGT candidates are dominated by genes that encode small polypeptides (including *psaE* and *psaI* in *P. chromatophora* strain FK01; Nakayama and Ishida 2009; Reyes-Prieto et al. 2010). The use of Blast searches and ML analyses in our screening procedure should not lead to the preferential detection of short genes; indeed, the opposite would be expected. Therefore, the predominance of small proteins must have biological significance. Because the bias toward small proteins is largely confined to proteins with predicted localization to the chromatophore, the currently unknown mechanism of protein import into the chromatophores might favor such proteins. Although protein import of EGT candidates into the chromatophore still remains to be directly demonstrated, all our data concerning the chromatophore of *P. chromatophora* strongly suggests that it can be regarded as a photosynthetic organelle in an early stage of evolution.

Supplementary Material

Supplementary figures S1–S4 and supplementary tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank three anonymous reviewers for their constructive comments on the manuscript. E.C.M.N. was partially supported by Deutsche Forschungsgemeinschaft grant NO 888/1-1.

References

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

Andersson JO. 2005. Lateral gene transfer in eukaryotes. *Cell Mol Life Sci.* 62:1182–1197.

Badger MR, Hanson D, Price GD. 2002. Evolution and diversity of CO_2 concentrating mechanisms in cyanobacteria. *Funct Plant Biol.* 29:161–173.

Baker A, Schatz G. 1987. Sequences from a prokaryotic genome or the mouse dihydrofolate reductase gene can restore the import of a truncated precursor protein into yeast mitochondria. *Proc Natl Acad Sci U S A.* 84:3117–3121.

Barth P, Lagoutte B, Sétif P. 1998. Ferredoxin reduction by photosystem I from *Synechocystis* sp. PCC 6803: Toward an understanding of the respective roles of subunits PsaD and PsaE in ferredoxin binding. *Biochemistry* 37:16233–16241.

Bhattacharya D, Archibald JM. 2006. The difference between organelles and endosymbionts—response to Theissen and Martin. *Curr Biol.* 16:R1017–R1018.

Bhaya D, Dufresne A, Vaulot D, Grossman A. 2002. Analysis of the *hli* gene family in marine and freshwater cyanobacteria. *FEMS Microbiol Lett.* 215:209–219.

Bodén M, Hawkins J. 2005. Prediction of subcellular localization using sequence-biased recurrent networks. *Bioinformatics* 21:2279–2286.

Bodył A, Mackiewicz P, Stiller JW. 2010. Comparative genomic studies suggest that the cyanobacterial endosymbionts of the

amoeba *Paulinella chromatophora* possess an import apparatus for nuclear-encoded proteins. *Plant Biol.* 12:639–649.

Bruce BD. 2001. The paradox of plastid transit peptides: conservation of function despite divergence in primary structure. *Biochim Biophys Acta—Mol Cell Res.* 1541:2–21.

Cai F, Menon BB, Cannon GC, Curry KJ, Shively JM, Heinhorst S. 2009. The pentameric vertex proteins are necessary for the icosahedral carboxysome shell to function as a CO_2 leakage barrier. *PLoS One*. 4:e7521.

Chang K-J, Wang C-C. 2004. Translation initiation from a naturally occurring non-AUG codon in *Saccharomyces cerevisiae*. *J Biol Chem.* 279:13778–13785.

Chou KC, Shen HB. 2007. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. *Biochem Biophys Res Commun.* 357:633–640.

Christensen AC, Lyznik A, Mohammed S, Elowsky CG, Elo A, Yule R, Mackenzie SA. 2005. Dual-domain, dual-targeting organellar protein presequences in *Arabidopsis* can use non-AUG start codons. *Plant Cell.* 17:2805–2816.

Deusch O, Landan G, Roettger M, Gruenheit N, Kowallik KV, Allen JF, Martin W, Dagan T. 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Mol Biol Evol.* 25:748–761.

Doolittle WE. 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends Genet.* 14:307–311.

Dunning Hotopp JC, Clark ME, Oliveira D, et al. (20 co-authors). 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 317: 1753–1756.

Düring U, Ossenbühl F, Wilde A. 2007. Late assembly steps and dynamics of the cyanobacterial photosystem I. *J Biol Chem.* 282:10915–10921.

Duy D, Soll J, Philippar K. 2007. Solute channels of the outer membrane: from bacteria to chloroplasts. *Biol Chem.* 388:879–889.

Eichinger L, Pachebat JA, Glöckner G, et al. (97 co-authors). 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* 435:43–57.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2:953–971.

Fujimori T, Hihara Y, Sonoike K. 2005. PsaK2 subunit in photosystem I is involved in state transition under high light condition in the cyanobacterium *Synechocystis* sp. PCC 6803. *J Biol Chem.* 280:22191–22197.

Gibbons JG, Janson EM, Hittinger CT, Johnston M, Abbot P, Rokas A. 2009. Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. *Mol Biol Evol.* 26:2731–2744.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.

He QF, Dolganov N, Björkman O, Grossman AR. 2001. The high light-inducible polypeptides in *Synechocystis* PCC6803: expression and function in high light. *J Biol Chem.* 276:306–314.

Hoogenraad HR. 1927. Zur Kenntnis der Fortpflanzung von *Paulinella chromatophora* Lauterb. *Zool Anz.* 72:140–150.

Huang CY, Ayliffe MA, Timmis JN. 2003. Direct measurement of the transfer rate of chloroplast DNA into the nucleus. *Nature* 422:72–76.

Kashino Y, Koike H, Yoshio M, Egashira H, Ikeuchi M, Pakrasi HB, Satoh K. 2002. Low-molecular-mass polypeptide components of a photosystem II preparation from the thermophilic cyanobacterium *Thermosynechococcus vulcanus*. *Plant Cell Physiol.* 43:1366–1373.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet.* 9:605–618.

Kies L. 1974. Electron microscopical investigations on *Paulinella chromatophora* Lauterborn, a thecamoeba containing blue-green endosymbionts (cyanelles). *Protoplasma* 80:69–89.

Lucattini R, Likić VA, Lithgow T. 2004. Bacterial proteins predisposed for targeting to mitochondria. *Mol Biol Evol.* 21:652–658.

Marin B, Nowack ECM, Glöckner G, Melkonian M. 2007. The ancestor of the *Paulinella* chromatophore obtained a carboxysomal operon by horizontal gene transfer from a *Nitrococcus*-like gamma-proteobacterium. *BMC Evol Biol.* 7:85.

Marin B, Nowack ECM, Melkonian M. 2005. A plastid in the making: evidence for a second primary endosymbiosis. *Protist.* 156:425–432.

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, 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 U S A.* 99:12246–12251.

Montané MH, Kloppstech K. 2000. The family of light-harvesting-related proteins (LHCs, ELIPs, HLIPs): was the harvesting of light their primary function? *Gene* 258:1–8.

Nakayama T, Ishida K. 2009. Another acquisition of a primary photosynthetic organelle is underway in *Paulinella chromatophora*. *Curr Biol.* 19:R284–R285.

Nanjo Y, Oka H, Ikarashi N, Kaneko K, Kitajima A, Mitsui T, Munoz FJ, Rodriguez-Lopez M, Baroja-Fernandez E, Pozueta-Romero J. 2006. Rice plastidial N-glycosylated nucleotide pyrophosphatase/phosphodiesterase is transported from the ER-Golgi to the chloroplast through the secretory pathway. *Plant Cell.* 18:2582–2592.

Nikoh N, Tanaka K, Shibata F, Kondo N, Hizume M, Shimada M, Fukatsu T. 2008. *Wolbachia* genome integrated in an insect chromosome: evolution and fate of laterally transferred endosymbiont genes. *Genome Res.* 18:272–280.

Nowack ECM, Melkonian M, Glöckner G. 2008. Chromatophore genome sequence of *Paulinella* sheds light on acquisition of photosynthesis by eukaryotes. *Curr Biol.* 18:410–418.

Patron NJ, Waller RF. 2007. Transit peptide diversity and divergence: a global analysis of plastid targeting signals. *Bioessays* 29:1048–1058.

Plöscher M, Granvogl B, Zoryan M, Reisinger V, Eichacker LA. 2009. Mass spectrometric characterization of membrane integral low molecular weight proteins from photosystem II in barley etioplasts. *Proteomics* 9:625–635.

Reyes-Prieto A, Hackett JD, Soares MB, Bonaldo MF, Bhattacharya D. 2006. Cyanobacterial contribution to algal nuclear genomes a primarily limited to plastid functions. *Curr Biol.* 16:2320–2325.

Reyes-Prieto A, Yoon HS, Moustafa A, Yang EC, Andersen RA, Boo SM, Nakayama T, Ishida K, Bhattacharya D. 2010. Differential gene retention in plastids of common recent origin. *Mol Biol Evol.* 27:1530–1537.

Salem K, van Waasbergen LG. 2004. Light control of *hliA* transcription and transcript stability in the cyanobacterium *Synechococcus elongatus* strain PCC 7942. *J Bacteriol.* 186:1729–1736.

Shiina T, Ishizaki Y, Yagi Y, Nakahira Y. 2009. Function and evolution of plastid sigma factors. *Plant Biotechnol.* 26:57–66.

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

Tang H-L, Yeh L-S, Chen N-K, Ripmaster T, Schimmel P, Wang C-C. 2004. Translation of a yeast mitochondrial tRNA synthetase

initiated at redundant non-AUG codons. *J Biol Chem.* 279:49656–49663.

Theissen U, Martin W. 2006. The difference between organelles and endosymbionts. *Curr Biol.* 16:R1016–R1017.

Touriol C, Bornes S, Bonnal S, Audigier S, Prats H, Prats A-C, Vagner S. 2003. Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. *Biol Cell.* 95:169–178.

Trost P, Fermani S, Marri L, Zaffagnini M, Falini G, Scagliarini S, Pupillo P, Sparla F. 2006. Thioredoxin-dependent regulation of photosynthetic glyceraldehyde-3-phosphate dehydrogenase: autonomous vs. CP12-dependent mechanisms. *Photosynth Res.* 89:263–275.

Villarejo A, Burén S, Larsson S, et al. (14 co-authors). 2005. Evidence for a protein transported through the secretory pathway *en route* to the higher plant chloroplast. *Nat Cell Biol.* 7:1224–1231.

Volkmer T, Schneider D, Bernát G, Kirchhoff H, Wenk SO, Rögner M. 2007. Ssr2998 of *Synechocystis* sp. PCC 6803 is involved in regulation of cyanobacterial electron transport and associated with the cytochrome b_6f complex. *J Biol Chem.* 282:3730–3737.

Wedel N, Soll J. 1998. Evolutionary conserved light regulation of Calvin cycle activity by NADPH-mediated reversible phosphoribulokinase/CP12/glyceraldehyde-3-phosphate dehydrogenase complex dissociation. *Proc Natl Acad Sci U S A.* 95:9699–9704.

Xu H, Vavilin D, Funk C, Vermaas W. 2004. Multiple deletions of small Cab-like proteins in the cyanobacterium *Synechocystis* sp. PCC 6803: consequences for pigment biosynthesis and accumulation. *J Biol Chem.* 279:27971–27979.

Yoon HS, Nakayama T, Reyes-Prieto A, Andersen RA, Boo SM, Ishida K, Bhattacharya D. 2009. A single origin of the photosynthetic organelle in different *Paulinella* lineages. *BMC Evol Biol.* 9:11.

Yoon HS, Reyes-Prieto A, Melkonian M, Bhattacharya D. 2006. Minimal plastid genome evolution in the *Paulinella* endosymbiont. *Curr Biol.* 16:R670–R672.

Zghidi W, Merendino L, Cottet A, Mache R, Lerbs-Mache S. 2007. Nucleus-encoded plastid sigma factor SIG3 transcribes specifically the *psbN* gene in plastids. *Nucleic Acids Res.* 35:455–464.

Zhulidov PA, Bogdanova EA, Shcheglov AS, et al. (11 co-authors). 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. *Nucleic Acids Res.* 32:15.