Supplementary Data

Data S1. Table S1 and experimental procedures. Table S1. Location and number of thioredoxin system-related genes and genes in the Calvin–Benson cycle (CB) and oxidative pentose phosphate pathway (OPPP) in different photosynthetic organisms (*Synechocystis* sp. PCC 6803, *Cyanidioschyzon merolae*, *Arabidopsis thaliana*, *Paulinella micropora* KR01).

Data S2. (A) Phylogeny of thioredoxin (TRX) in KR01. **(B)** Phylogeny of glucose-6-phosphate dehydrogenase (G6PDH) in KR01 and sequence alignment of redox-active Cys region of chloroplastic G6PDH. Asterisk indicate redox-active Cys sites. **(C)** Phylogeny of malate dehydrogenase (MDH) in KR01. Bootstrap support values >90 are shown at the branches. Well-supported clades with interest are marked with gradient color boxes. Color code: blue, *Paulinella* species; green, *Arabidopsis thaliana*; gray, bacteria and archaea species.

Supplemental Data:

Independent evolution of the thioredoxin system with diverse phylogenetic origins in photosynthetic *Paulinella* species

Duckhyun Lhee, Debashish Bhattacharya, Hwan Su Yoon

Supplementary Table

Table S1. Location and number of thioredoxin system-related genes and genes in the Calvin–Benson cycle (CB) and oxidative pentose phosphate pathway (OPPP) in different photosynthetic organisms (*Synechocystis* sp. PCC 6803, *Cyanidioschyzon merolae*, *Arabidopsis thaliana*, *Paulinella micropora* KR01).

	Enzyme	Cyanobacteria (*Synechocystis* sp. PCC 6803)	Red algae (*Cyanidioschyzon merolae*)		Green plant (*Arabidopsis thaliana*)		*Paulinella micropora* KR01	
			Plastid genome	Nuclear genome	Chloroplast genome	Nuclear genome	Chromatophore genome	Nuclear genome
Thiol-oxidoreductase	TRX	4	1	7	-	20	2	9
	GRX	3	-	2	-	28	1	2
Thioredoxin reductase	FTR	2	1 (catalytic chain)	1 (variable chain)	-	3 (catalytic beta chain), 3 (variable subunit)	2 (catalytic chain, variable chain)	NA
	NTR	1	-	1	-	4	1	3
CB	Rubisco	1 (large subunit), 2 (small subunit)	1 (large subunit), 1 (small subunit)	-	1 (large subunit)	8 (small subunit)	1 (large subunit), 1 (small subunit)	NA
	Activase	-	-	-	-	3	NA	NA
	PRK	1	-	1	-	1	1	NA
	GAPDH	2	-	4	-	16	1	2
	CP12	1	-	1	-	3	NA	1 (partial domain)
	F/SBPase	1	-	-	-	-	1	NA
	FBPase	1	-	2	-	2	NA	1
	SBPase	-	-	2	-	1	NA	NA
	MDH	1	-	2	-	7	NA	4
OPPP	OpcA	1	-	-	-	-	NA	NA
	G6PDH	1	-	2	-	12	NA	3

Supplemental Experimental Procedures

Survey of genes in this paper.
In this study, we used data from *Paulinella micropora* KR01[S1], *Paulinella chromatophora* CCAC 0185[S2], and *Paulinella ovalis*[S3]. The chromatophore proteome data from *P. chromatophora* CCAC 0185[S4] was used to predict the location of proteins in KR01. The presence of a chromatophore transit peptide in KR01 proteins and

gene expression data under high light stress and the diurnal cycle provided corroborating data for reaching conclusions[S1]. To find thioredoxin system-related genes and genes in the Calvin–Benson cycle and oxidative pentose phosphate pathway in *Paulinella* species, the reference genes of *Arabidopsis thaliana* and *Synechocystis* sp. strain PCC 6803 were used as starting queries in BLASTp searches (e-value cutoff: 1e-5)[S5] against the *Paulinella* database. Domain structure was analyzed using the web-based CD-search[S6]. After manually checking the alignment, only genes that were well-aligned with conserved domains were selected for downstream analysis.

	Gene name	**Encoded location**	***P. chromatophora* ortholog**	**Predicted location**	**DEG (p-value < 0.05 and |log2 (fold change)| > 1)**	**Chromatophore transit peptides**	**Diurnal rhythmic expression**
TRX	g18255.t1	Nuclear genome	scaffold26240-size528|m.128845	Lysate			
TRX	g28294.t1	Nuclear genome	scaffold7220-size1567|m.58497	Lysate			
TRX	g35100.t1	Nuclear genome	scaffold8548-size1418|m.65412	Chromatophore		crTP	
TRX	g38191.t1	Nuclear genome	scaffold12955-size1047|m.85383	Lysate			
TRX	g4155.t1	Nuclear genome	scaffold11899-size1123|m.80931				
TRX	g41691.t1	Nuclear genome	scaffold15851-size878|m.96520				
TRX	g63005.t1	Nuclear genome	scaffold7336-size1552|m.59121				
TRX	g69762.t1	Nuclear genome	scaffold16773-size830|m.99823				
TRX	g76893.t1	Nuclear genome	scaffold14993-size922|m.93369				
TRX	PCKR_721	Chromatophore genome	PCC_0738	Chromatphore	NA	NA	NA
TRX	PCKR_755	Chromatophore genome	PCC_0772	Chromatphore	NA	NA	NA
GRX	g65652.t1	Nuclear genome	scaffold4665-size1929|m.43315	Lysate			
GRX	g77183.t1	Nuclear genome	scaffold27528-size502|m.132296, scaffold24812-size560|m.124887				Diurnal
GRX	PCKR_810	Chromatophore genome	PCC_0826	Chromatophore	NA	NA	NA
NTR	g2886.t1	Nuclear genome					
NTR	g17500.t1	Nuclear genome					
NTR	g3229.t1	Nuclear genome			TP06_UP		
NTR	PCKR_028	Chromatophore genome	PCC_0027	Chromatophore	NA	NA	NA
FTR	PCKR_644	Chromatophore genome	PCC_0658	Chromatophore	NA	NA	NA
FTR	PCKR_213	Chromatophore genome	PCC_0219	Chromatophore	NA	NA	NA
GAPDH	g28715.t1	Nuclear genome	scaffold4957-size1875|m.45177	Lysate			
GAPDH	g38229.t1	Nuclear	scaffold3834-	Lysate			

		genome	size2088|m.37882				
GAPDH	PCKR_414	Chromatophore genome	PCC_0418	Chromatophore	NA	NA	NA
MDH	g8927.t1	Nuclear genome	scaffold6154-size1705|m.52479	Chromatophore	TP06_UP, TP30_DOWN	crTP	Diurnal
MDH	g43748.t1	Nuclear genome	scaffold9817-size1295|m.71516	Lysate	TP06_DOWN		
MDH	g62328.t1	Nuclear genome	scaffold12090-size1110|m.81767	Lysate			
MDH	g55981.t1	Nuclear genome	scaffold9817-size1295|m.71516	Lysate	TP06_UP		Diurnal
CP12	g65620.t1	Nuclear genome					
FBPase	g72605.t1	Nuclear genome	scaffold7966-size1483|m.62430	Lysate	TP06_DOWN, TP18_UP		
FBPase	PCKR_261	Chromatophore genome	PCC0268	Chromatophore	NA	NA	NA
G6PDH	g9229.t1	Nuclear genome	scaffold2155-size2616|m.25388	Chromatophore		crTP	
G6PDH	g40784.t1	Nuclear genome	scaffold2155-size2616|m.25388	Chromatophore		crTP	Diurnal
G6PDH	g58255.t1	Nuclear genome	scaffold3969-size2060|m.38789	Lysate			

Phylogenetic analysis

Putative homologs for each target protein were identified using a BlastP search (*e*-value cutoff: 1e-5)[S5] against a local database, which contained taxa selected from NCBI RefSeq[S7] and the MMETSP database[S8] to provide a broad taxon sampling. CD-HIT[S9] was used to remove the redundant isoforms. Multiple amino acid alignments were done using Clustal Omega[S10] with the default options. These alignments were refined manually based on conserved domains. Maximum likelihood-based phylogenetic analysis and bootstrap resampling of the data was done using IQ-TREE[S11] with 1,000 ultrafast bootstrap replications[S12]. The evolutionary model was automatically selected using the model test option incorporated in IQ-TREE. Highly diverged or contaminant sequences, which exhibited a long-branch or taxonomic misplacement in the tree were removed and the analysis done anew. ETE3[S13] was used to visualize the tree and alignments.

Data Availability

Sequence alignments of phylogenetic trees are available from Dryad Digital Repository (https://doi.org/10.5061/dryad.8kprr4xmt).

Supplemental References

S1. Lhee, D., Lee, J., Ettahi, K., Cho, C.H., Ha, J.-S., Chan, Y.-F., Zelzion, U., Stephens, T.G., Price, D.C., Gabr, A., et al. (2020). Amoeba genome reveals dominant host contribution to plastid endosymbiosis. Mol. Biol. Evol.

S2. Nowack, E.C.M., Price, D.C., Bhattacharya, D., Singer, A., Melkonian, M., and Grossman, A.R. (2016). Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*. Proc. Natl. Acad. Sci. U.S.A. *113*, 12214-12219.

S3. Bhattacharya, D., Price, D.C., Yoon, H.S., Yang, E.C., Poulton, N.J., Andersen, R.A., and Das, S.P. (2012). Single cell genome analysis supports a link between phagotrophy and primary plastid endosymbiosis. Sci. Rep. *2*, 356-356.

S4. Singer, A., Poschmann, G., Muhlich, C., Valadez-Cano, C., Hansch, S., Huren, V., Rensing, S.A., Stuhler, K., and Nowack, E.C.M. (2017). Massive protein import into the early-evolutionary-stage photosynthetic organelle of the amoeba *Paulinella chromatophora*. Curr. Biol. *27*, 2763-2773 e2765.

S5. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T.L. (2009). BLAST+: architecture and applications. BMC Bioinformatics *10*, 421-421.

S6. Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C.J., Lu, S., Chitsaz, F., Derbyshire, M.K., Geer, R.C., Gonzales, N.R., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. *45*, D200-D203.

S7. O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. *44*, D733-745.

S8. Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., et al. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. PLoS Biol. *12*, e1001889-e1001889.

S9. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics (Oxford, England) *28*, 3150-3152.

S10. Sievers, F., and Higgins, D.G. (2018). Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. *27*, 135-145.

S11. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., and Minh, B.Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. *32*, 268-274.

S12. Minh, B.Q., Nguyen, M.A.T., and von Haeseler, A. (2013). Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol. *30*, 1188-1195.

S13. Huerta-Cepas, J., Serra, F., and Bork, P. (2016). ETE 3: Reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. *33*, 1635-1638.


A
95
95
B
C