Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis

Raphaël Méheust[a], Ehud Zelzion[b], Debashish Bhattacharya[b], Philippe Lopez[a], and Eric Bapteste[a,1]

[a]Unité Mixte de Recherche 7138 Evolution Paris Seine, Institut de Biologie Paris Seine, Université Pierre et Marie Curie, Centre National de la Recherche Scientifique, Sorbonne Universités, 75005 Paris, France; and [b]Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, NJ 08901

Edited by John M. Archibald, Dalhousie University, Halifax, Canada, and accepted by the Editorial Board February 14, 2016 (received for review September 8, 2015)

The integration of foreign genetic information is central to the evolution of eukaryotes, as has been demonstrated for the origin of the Calvin cycle and of the heme and carotenoid biosynthesis pathways in algae and plants. For photosynthetic lineages, this coordination involved three genomes of divergent phylogenetic origins (the nucleus, plastid, and mitochondrion). Major hurdles overcome by the ancestor of these lineages were harnessing the oxygen-evolving organelle, optimizing the use of light, and stabilizing the partnership between the plastid endosymbiont and host through retargeting of proteins to the nascent organelle. Here we used protein similarity networks that can disentangle reticulate gene histories to explore how these significant challenges were met. We discovered a previously hidden component of algal and plant nuclear genomes that originated from the plastid endosymbiont: symbiogenetic genes (S genes). These composite proteins, exclusive to photosynthetic eukaryotes, encode a cyanobacterium-derived domain fused to one of cyanobacterial or another prokaryotic origin and have emerged multiple, independent times during evolution. Transcriptome data demonstrate the existence and expression of S genes across a wide swath of algae and plants, and functional data indicate their involvement in tolerance to oxidative stress, phototropism, and adaptation to nitrogen limitation. Our research demonstrates the "recycling" of genetic information by photosynthetic eukaryotes to generate novel composite genes, many of which function in plastid maintenance.

gene fusion | endosymbiosis | photosynthesis | eukaryote evolution | novel gene origin

The genomes of the proteobacterium-derived mitochondrion and the cyanobacterium-derived plastid have undergone significant genome reduction due to outright gene loss or transfer to the nuclear genome (1, 2). Organelle gene loss by transfer to the nucleus is known as endosymbiotic gene transfer [EGT (a special form of horizontal gene transfer; HGT)] and has resulted in chimeric host nuclear genomes with, in the case of plastids, from ca. 200 to several thousand intact endosymbiont genes being relocated (3) (Fig. 1*A*). Plastid EGT has a long evolutionary history, extending back over a billion years in the case of primary plastid origin in the Archaeplastida (glaucophytes, red and green algae, and their sister group, plants) and several hundred million years for secondary plastids in groups such as diatoms, haptophytes, and dinoflagellates (4). A common fate for many nuclear-encoded organelle-derived proteins is to be targeted back to the compartment of origin via channels [i.e., translocons at the outer- and inner-envelope membrane of plastids and mitochondria (Toc/Tic and Tom/Tim, respectively)] to carry out organelle functions (5). Identification of EGT candidates generally relies on phylogenetic methods that use simultaneous alignment of colinear proteins sharing significant sequence similarity over all, or most, of their lengths to reconstruct the tree and its constituent branch lengths. An alternative approach is network methods that rely on reconstruction of both full and partial (i.e., protein domain; Fig. 1*B*) gene relationships using pairwise protein similarity values. These methods allow detection of reticulate sequence evolution, such as the fusion of domains derived from heterologous proteins (6–10). Here we used networks to ask the following two questions: (*i*) Did the Archaeplastida plastid endosymbiont contribute gene fragments to symbiogenetic genes (S genes) that are detectable in algal and plant nuclear genomes? (*ii*) If so, are these S genes expressed, and what putative functions did the novel domain combinations confer to the host lineage? These questions are motivated by the knowledge that although fundamental to the origin of complex life forms such as plants and animals, plastid endosymbiosis wrought significant challenges for the first algal lineages. These resulted from light harvesting, which can capture excess energy that must be dissipated, and oxygen evolution, which leads to the formation of reactive oxygen species (ROS) that need to be detoxified (11, 12).

Results and Discussion

We identified 67 families of expressed nuclear-encoded S genes (Fig. 2). These families are distributed in 349 algae and plants. Four S-gene families were likely present in the Archaeplastida ancestor, 11 S-gene families are shared by the red and the green lineages, and 28 S-gene families are found both in primary and secondary photosynthetic lineages, demonstrating their ancient origins and functional relevance (Fig. 3 and Fig. S1). The 55 S-gene candidates we focused on here are predicted to be plastid-targeted (Table S1), and at least 23 of these function in redox regulation and light and stress responses (Fig. 2).

Significance

Endosymbiotic gene transfer from the plastid genome to the nucleus comprises the most significant source of horizontal gene transfer in photosynthetic eukaryotes. We investigated genomic data at the infragenic level to determine whether the cyanobacterial endosymbiont also contributed gene fragments (i.e., domains) to create novel nuclear-encoded proteins. We found 67 such gene families that are expressed as RNA and widely distributed among plants and algae. At least 23 genes are putatively involved in redox regulation and light response, namely the maintenance of a photodynamic organelle. Our results add a new layer of complexity to plastid integration and point to the role of fused proteins as key players in this process.

Author contributions: R.M., P.L., and E.B. designed research; R.M. and E.Z. performed research; R.M. performed detection of S genes; E.Z. analyzed RNA-sequencing data; R.M., D.B., P.L., and E.B. analyzed data; and R.M., D.B., P.L., and E.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. J.M.A. is a guest editor invited by the Editorial Board.

Freely available online through the PNAS open access option.

Data deposition: The FASTA sequences of the S genes reported in this paper are available at www.evol-net.fr/downloads/S-genes.zip.

[1]To whom correspondence should be addressed. Email: epbapteste@gmail.com.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1517551113/-/DCSupplemental.

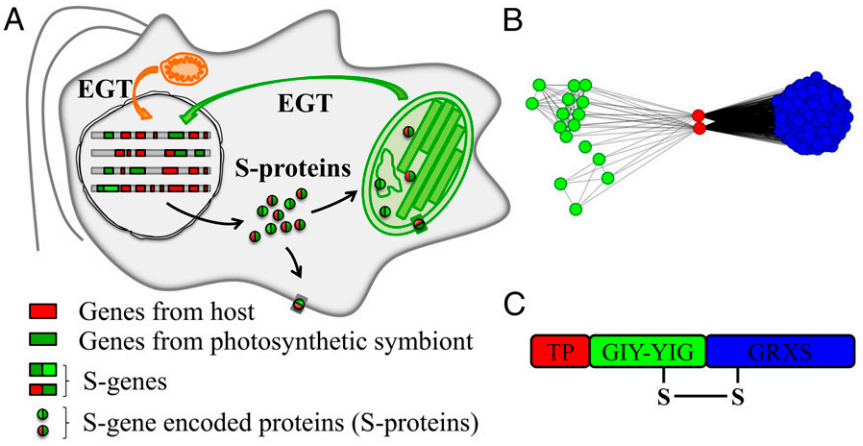

Fig. 1. Origin of composite genes in algae and plants. (*A*) The role of plastid endosymbiosis in providing the genetic toolkit for S-gene origin. (*B*) Network analysis of the *AtGRXS16* (family 14) S gene in *A. thaliana*. The red nodes identify the S genes; green and blue nodes are the components from GIY–YIG and GRXS domains, respectively, that gave rise to S genes through gene fusion. (*C*) Domain structure of AtGRXS16. An intramolecular disulfide bond can be formed between the two domains. TP, transit peptide.

Evidence That S Genes Are Not Assembly Artifacts. It is conceivable that the union of two unrelated protein domains that we report here as S genes could potentially be explained by misassembly of genomic or transcriptomic reads, an expected outcome of the analysis of large datasets. Given this concern, we used several approaches to validate the existence of S genes. The first was to collect RNA-seq (sequencing) data that could be used to map to coding sequences (CDSs) and genomic sequences of S genes. If the RNA reads mapped uniformly across the CDS or genomic DNA with no loss of coverage at the domain junctions, then we had evidence the coding region was authentic. We did this procedure for two taxa, a green alga *Picochlorum* and the model plant *Arabidopsis thaliana*. In the first case, we downloaded transcriptome reads from *Picochlorum* SE3 [National Center for Biotechnology Information (NCBI) BioProject accession no. PRJNA245752] and mapped these to the CDSs of S genes from its closely related sister species *Picochlorum oklahomensis* and *Picochlorum* RCC944 (Table S2). These results showed that for nine shared homologs, transcriptome coverage across the CDSs was nearly 100% and uniform across the domain junctions (Fig. S2). These results strongly support the existence of these S genes. Furthermore, we used PCR with genomic DNA from *Picochlorum* SE3 for five S genes to validate that they were intact fragments. These results are shown in Fig. S3, and sequencing of the nearly complete CDS fragments showed identity to the genomic region encoding the S gene. Mapping of RNA-seq reads to *A. thaliana* S gene-encoding genomic regions (i.e., exons and introns; Table S2) also showed robust and uniform mapping to the exonic regions (Fig. S2), again supporting the existence of intact S genes in this well-annotated genome.

We also checked whether S genes may result from gene misannotation (i.e., the annotation of two separate gene sequences as a single gene, or misincorporation of an exon from two overlapping genes into a gene annotation). We found evidence that 23 S-gene families have at least one gene with all domains being positioned in the same exon, thereby arguing against possible misincorporation of exon information (Table S3). Finally, although we did not validate every S gene cited in this study, we are buoyed by the fact that all families are found in at least one genome and one transcriptome, with many occurring in >10 taxa (Fig. 2 and Fig. S1), making it highly unlikely that these data are explained by artifacts due to misassembly. Although it is difficult to reconstruct robust and resolved domain phylogenies due to their small size, we assessed whether S genes may have been misannotated by reconstructing complete S-gene trees. For example, the phylogeny of an anciently derived S gene (family 31) limited to Viridiplantae is shown in Fig. S4 and supports the existence of this composite sequence in the green lineage ancestor. This tree is in agreement with the accepted relationship of green lineages, thereby showing no evidence of a complex history but rather persistence of the gene family across species. These results are summarized in Fig. 2, which also reports the number of transcriptomes and genomes of distinct organisms in which homologs of S genes were found. Because some transcriptomes are derived from phagotrophic protists (in particular, heterotrophic dinophytes such as *Oxyrrhis*), there is a risk of prey contamination (i.e., the S gene might derive from prey DNA). Therefore, identifying the S gene in multiple transcriptomes from a given taxonomic group provides stronger support for the presence of the S gene in that group.

S Genes Involved in Redox Regulation. Many S-gene families play a role in redox regulation, including family 14, which contains AtGRXS16, a plastid-localized protein in *A. thaliana* (Fig. 2 and Fig. S2). This gene family is widely distributed in Viridiplantae (green algae and plants), and may also be present in a small number of other species (Fig. S1). *AtGRXS16* is composed of two fused domains that do not exist together elsewhere in the tree of life. This S-gene family encodes an N-terminal GIY–YIG (GlyIleTyr–TyrIleGly) endonuclease fold of cyanobacterial origin and a C-terminal CGFS-type monothiol GRXS (glutaredoxin; disulfide oxidoreductase) of bacterial (yet noncyanobacterial) origin that are negatively regulated by the formation of an intramolecular disulfide bond (Fig. 1*C*). This association allows ROS scavenging via the GRXS domain coupled with the ability of the GIY–YIG endonuclease to repair oxidative stress-induced DNA double-strand breaks in plant plastid genomes (13). Consequently, this anciently derived S gene plays an important role in coordinating redox regulation and DNA repair in response to ROS (13). Consistent with these observations are RNA-seq data (14) that show a ca. threefold up-regulation of *AtGRXS16* ($P < 0.01$) in *A. thaliana* seedlings in light versus dark conditions (*S-Gene Expression Analysis*).

Domains in S genes can be reused for redox regulation, as illustrated by family 4. This gene is found in the red, green, and secondary plastid-derived lineages and is composed of two fused domains. The N-terminal region again encodes a GIY–YIG endonuclease fold of cyanobacterial origin, whereas the C terminus encodes a NifU domain of cyanobacterial origin that is involved in iron–sulfur (Fe–S) cluster assembly (15). Bioinformatic evidence was found for plastid targeting of this protein (Fig. 2).

Another S gene involved in redox regulation that is widely distributed in Viridiplantae is family 19 (Fig. 2). This modular gene (*SufE3*) encodes quinolate synthetase and defines a novel combination of two biochemically interacting domains: a SufE domain of cyanobacterial origin and a NadA domain of (noncyanobacterial) prokaryotic origin (16). The quinolate activity of the NadA domain relies on a highly oxygen-sensitive (4Fe–4S) cluster, whose formation depends on a cysteine residue present in its novel genetic partner, the SufE domain, which is involved in the long-term competence of the enzyme (16). Because this nuclear-encoded quinolate synthetase is plastid-localized (17), it is likely to be exposed to high levels of oxidative stress. The SufE domain has been proposed to continuously repair/reconstitute

Family	Domains (gene name)	Assumed function	Taxonomic distribution										Evidence						
			Glaucophyta	*Rhodophyta*	*Chlorophyta*	*Streptophyta*	*Cryptophyta*	*Haptophyta*	*Cercozoa*	*Ochrophyta*	*Bacillariophyta*	*Dinophyta*	*Euglenozoa*	Dataset genomes	Other genomes	Transcriptomes	Published evicence	RNA-seq	RT-PCR	Read mapping	Same exon
1	**hydrolase†*** + LPLAT† (*ELT* or *PES*)	maintenance of membrane integrity	1	6	12	59		7		5	22	12	1	7	63	59					
2	**NTF2-like†*** + SOUL† (*SOUL3*)	eyespot related in *C.reinhardtii*		5	12	54	1	10	2	10	26	2		11	57	55					
3	**PAPS reductase** + thioredoxin	sulfur assimilation		1	5	44	1	10	2	5	32	9		5	46	62					
4	**GIY-YIG†*** + **NifU†**	DNA repair + redox	1	3	8	1		1		5	8	4		6	3	22					
5	**SufE*** + BolA (*SufE1*)	redox		5	6	35		2		8	14	2		5	36	34					
6	**photolyase** + **hydrolase†***	DNA repair + light response		2	5	45		1		2		1		2	45	9					
7	**DUF3593†** + **DUF2499†***	-		3	4			6		5	16	5		2	3	37					
8	**3-dehydroquinate synthase** + O-methyltransferase	shikimate biosynthesis		1					1	4	9	5	1	1	3	17					
9	**serine protease** + LIM zinc finger domain	-	1						1	2	1	13		3	0	15					
10	PPIase† + **rhodanese†** (*PIN3*)	auxin efflux in *A.thaliana*		6	7	21		3		3				4	25	11					
11	**psbD†** + **psbC†**	photosynthesis		1	2	1		2			1			1	1	5					
12	**DUF760†*** + **DUF760†***	-			1	4		4		3		1		1	4	8					
13	**tsf†*** + **tsf†***	translation			3	1			2			10	1	4	3	10					
14	**GIY-YIG†*** + GRXS (*AtGRXS16*)	DNA repair + redox		1	3	45						1		2	45	3					
15	**peroxiredoxine** + thioredoxin†	redox		1	1					5		1		1	1	6					
16	**atpB†** + **atpE†**	ATP synthase		2		4		1			2			1	4	4					
17	PPIase + **tRNA-i(6)A37 methylthiotransferase†**	tRNA transferase		4						4	9	1		1	1	17					
18	**ferredoxin†*** + tetratricopeptide repeat	redox			5	43				1	1			3	43	4					
19	**SufE†*** + NadA (*SufE3*)	quinolate synthetase			14	51				5			1	4	55	13					
20	DUF2256 + **unknown domain 3†**	-			2			3			11	4		3	2	16					
21	glycosyltransferase + **glycosyltransferase***	-						1		1		1	1	1	0	3					
22	**phytochrome†** + PAS domain	photosensory signaling protein	1			127	1							1	130	2					
23	TPR repeat + RING + **ATP-dependant protease**	-		1	3	43								1	43	3					
24	DnaJ + **ferredoxin†***	redox			4	5						1		3	5	2					
25	**RNA methyltransferase†** + 2OG-Fe(II) oxygenase	-			3			2				1		1	1	5					
26	CobU† + **DUF4346†**	cobalamin biosynthesis					1	1			15			2	1	14					
27	**acetyltransferase†** + methyltransferase	-					1	6				1		1	0	7					
28	**tic-20†** + calmodulin	translocation								7	29	14		3	3	47					
29	**allophycocyanin beta subunit†*** + **allophycocyanin beta subunit†***	photosynthesis		1			1							1	0	1					
30	RPS13 + **RPS11** + **RNA polymerase alpha subunit†**	transcription/translation		1							1			1	0	1					
31	**RIBR†** + **DUF1768** (*PyrR*)	riboflavin biosynthesis			11	55								6	57	3					
32	**DUF2499†*** + unknown domain 1	-			1						3			1	0	3					
33	**PDZ** + FKBP	periplasmic protease			7							1		2	3	3					
34	**ferredoxin nitrite reductase†** + oxido-reductase + rubredoxin	nitrogen assimilation			7							1		2	2	4					
35	NTF2-like + **unknown domain 2†**	-						5			1			1	1	4					
36	**DUF2930†*** + **DUF2930†***	-							1			4		1	0	4					
37	**NTF2-like†*** + **lipase†**	-								4		8		1	0	13					
38	**ftsH†*** + **ftsH†***	maintenance of photosynthesis									1	10		1	0	10					
39	**phycobilisome linker polypeptide duplication†***	photosynthesis	2											1	0	1					
40	**SufE†*** + tRNA 5-methylaminomethyl-2-thiouridylate methyltransferase	tRNA transferase		2										1	0	2					
41	glutamylcyclotransferase + **hydrolase†*** + LPLAT† (*ELT* or *PES*)	maintenance of membrane integrity		6										3	1	2					
42	DnaJ† + **phycocyanobilin lyase†**	light response + redox		9										3	1	5					
43	**O-methyltransferase*** + deoxyadenosine/deoxycytidine kinase	-		2										1	0	1					
44	acyltransferase + **glyoxalase†**	Lactoylglutathione lyase		2										1	0	1					
45	**DUF89** + Fructose-1,6-bisphosphatase†	Calvin cycle enzyme			4									1	2	1					
46	ABC transporter + **DUF2246**	ABC transporter			3									1	0	2					
47	S1† + S1† + **tsf*** + **tsf†*** (*PETs*)	translation			3									2	0	1					
48	carbohydrate Binding Module 48 + **sucrose phosphatase**	carbohydrate metabolism			2									1	0	1					
49	**9-cis-epoxycarotenoid dioxygenase** + glutathione S-transferase†	abscisic acid biosynthesis + redox			8									2	3	3					
50	**fasciclin** + chlorophile a-b binding protein	surface protein			7									2	2	3					
51	**peroxiredoxin-like†** + **PAP-fibrillin†**	-			3									1	0	2					
52	**excinuclease B†** + **excinuclease C**	DNA repair + light response			3									1	1	1					
53	ankyrin + **DEAD/DEAH box helicase** + **DSHCT†**	-			2									1	0	1					
54	**transcription activator TenA** + HMP-P kinase† + TMP-Ppase†	thiamin metabolism			2									1	0	1					
55	**glycosyltransferase*** + SNARE	-			4									1	1	2					
56	**DUF393** + polyphosphate glucokinase† + EF-hand + thioredoxin	-			5									1	2	2					
57	**ribokinase** + kinase + **CHAT domain**	pentose phosphate pathway						4						1	0	3					
58	**arsC transcriptional regulator†** + arsM	-						5						1	0	4					
59	**SNARE** + **2-polyprenyl-6-methoxyphenol hydroxylase**	FAD-dependent oxidoreductase						2						1	0	1					
60	**phosphatase** + EF-hand + **CBS**	-									9			1	0	8					
61	G6PD† + **6PGD**	pentose phosphate pathway									7			1	0	6					
62	**SufE†*** + cysteine--tRNA synthetase†	tRNA synthetase										10		1	0	9					
63	methyltransferase + **cytochrome b6/f complex, subunit V†**	-										2		1	0	1					
64	bacteriorhodopsin-like + PAS + **transduction signal**	photosensory signaling protein										16		1	0	15					
65	**CobW** + TNF receptor superfamily + EPS_sugar_tfrase	-										11		1	0	10					
66	**ferredoxin†*** + ferritin†	iron storage										5		1	0	4					
67	DUF2358 + **NTF2-like†*** + **hydrolase***	-										1		1	0	1					

Fig. 2. Sixty-seven S-gene families identified in our study. Domains in bold originated from Cyanobacteria. Plastid-localized protein families (i.e., families with at least one protein predicted to be plastid-targeted according to ChloroP and ASAFind) are shaded in gray. *, domain of cyanobacterial origin occurring more than once per S gene. †, highly confident domain of cyanobacterial or prokaryotic (noncyanobacterial) origin (Table S4).

the Fe–S cluster in the NadA domain of the quinolate synthetase to maintain a functional protein (18). *SufE3* is ubiquitously expressed in all major plant organs and is embryo-lethal when knocked out in *A. thaliana* (18). In this model plant, there is ca. twofold gene (*At5g50210*) up-regulation ($P < 0.01$) in light versus dark conditions (14) (*S-Gene Expression Analysis*).

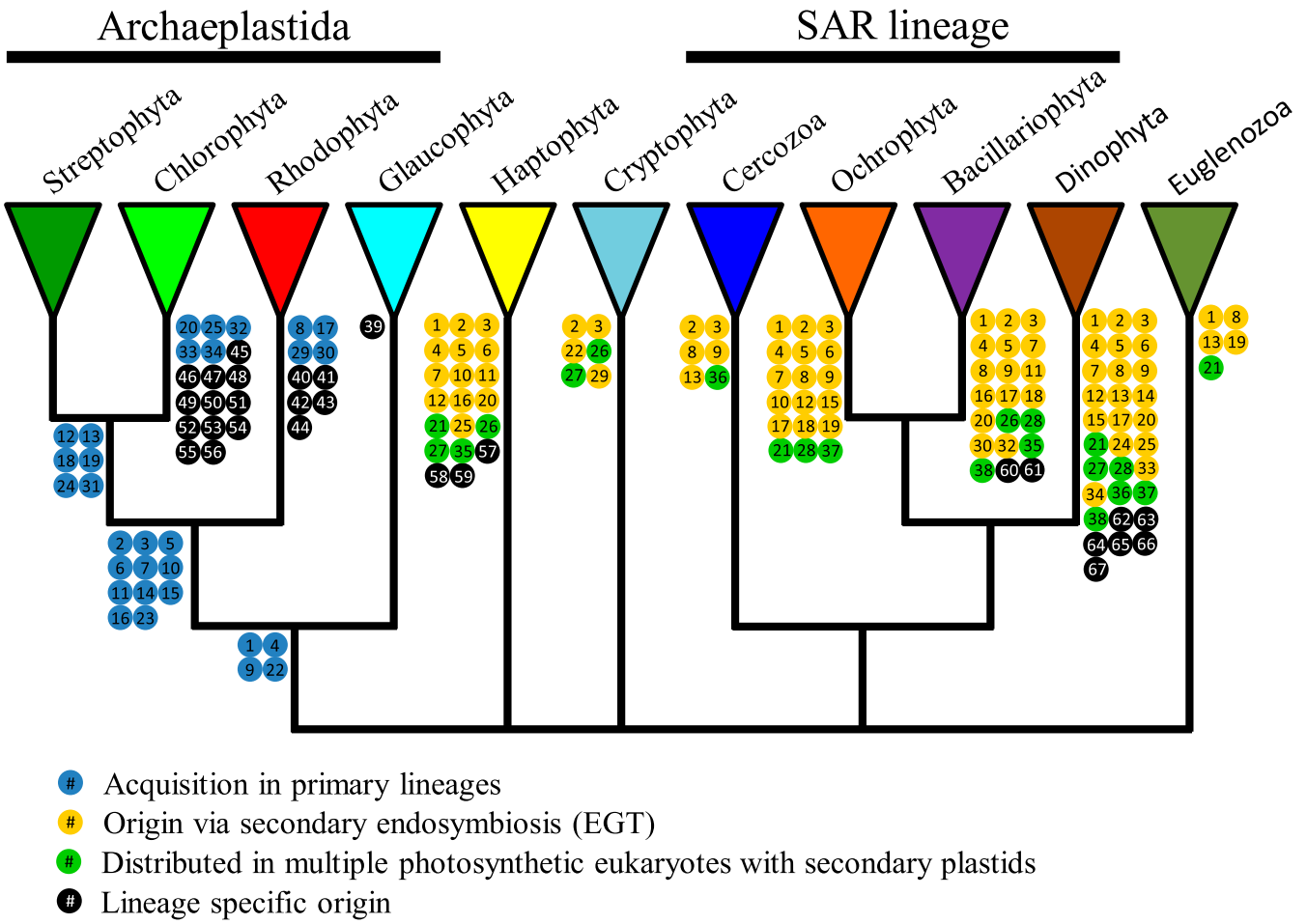

Fig. 3. Putative nuclear gene-based phylogeny of photosynthetic eukaryotes, showing the distribution of the 67 S-gene families we report. SAR, Stramenopiles-Alveolates-Rhizaria.

Another fascinating example is family 49 (Fig. 2), which is restricted to prasinophyte green algae and encodes two cyanobacterium-derived domains. The N-terminal region is a 9-*cis*-epoxycarotenoid dioxygenase (RPE65) domain involved in the production of abscisic acid from xanthophyll precursors (19), whereas the C terminus contains a glutathione S-transferase (GST) domain, which in plants plays a major role in reducing oxidative stress damage. Whereas responses to oxidative stress appear to be central to S-gene evolution, we also find examples of their roles in coordinating algal responses to light direction to optimize photosynthesis and growth.

S Genes Involved in Light Responses. S-gene family 2 (Fig. 2) defines the well-studied *AtHBP5* gene in *A. thaliana* and *SOUL3* in *Chlamydomonas reinhardtii*. This gene fusion is composed of an N-terminal region of cyanobacterial origin and a C-terminal region of prokaryotic derivation, and is present in the red and green lineages within Archaeplastida as well as in secondary plastid-containing algae. The heme-binding protein in *A. thaliana* (AtHBP5) is localized in plastoglobules, where it is likely involved in chlorophyll degradation (20). SOUL3 is localized to the plastid eyespot of *C. reinhardtii* (21) and, when knocked-out, the eyespot is reduced in size and its location is altered, negatively impacting phototaxis (21). AtHBP5 and SOUL3, which facilitate a coordinated response to light of the photosynthetic cell, produce an analogous phenotype to the communal phototropism of the well-known prokaryotic consortium *Chlorochromatium aggregatum* (22). In the latter case, cross-talk between photosynthetic epibiontic bacteria is transferred to a central motile, brown bacterium, thereby moving the collective to a location where epibionts can most efficiently perform photosynthesis (22).

Another family of S genes, family 10, is involved in phototropism and gravitropism (Fig. 2). This gene is composed of two domains, a peptidyl prolyl isomerase (PPIase) and a rhodanese superfamily domain, with the former of (noncyanobacterial) prokaryotic origin and the latter of cyanobacterial provenance. This S gene encodes a widely distributed PPIase in plants, red algae, haptophytes, and stramenopiles that is likely to be plastid-targeted (Fig. 2). In *A. thaliana*, this developmental protein (known as PIN3) is localized to the plasma membrane and reallocates auxin, affecting phototropism and gravitotropism of young sprouts (23).

S Genes Involved in Endosymbiont Stabilization. Achieving genetic integration also required innovations to stabilize the endosymbiont in the host cell. S genes were involved in this function as well, with some playing a role in scavenging organelle degradation products during abiotic stress. Family 1 (Fig. 2) encodes a plastid-localized composite protein in *A. thaliana* that contains two domains [e.g., an esterases/lipases/thioesterases (ELT) or phytyl ester synthase (PES) domain, and a hydrolase domain of cyanobacterial origin]. This protein is widely distributed in photosynthetic eukaryotes (Fig. S1), and in *A. thaliana* forms a gene family involved in fatty acid phytyl ester synthesis that is highly expressed during senescence and nitrogen deprivation (24); that is, these proteins scavenge toxic free phytol and fatty acids after thylakoid degradation. Family 41 (Fig. 2) is similar to family 1, albeit with an additional bacterium-derived gamma-glutamylcyclotransferase N-terminal domain involved in glutathione

metabolism. The taxonomic distribution of family 41 is restricted to red algae, suggesting that lineage-specific fusion events may have given rise to convergent functions to protect plastid membranes from abiotic stress.

S Genes with Potential Novel Functions in Photosynthetic Eukaryotes. Another important aspect of our network analysis was to provide the foundation for experimental analysis of novel genes, because S genes could also have introduced novel biochemical functions that are exclusive to photosynthetic eukaryotes. An example of this is family 42, which is restricted to red algae (Fig. S1). This S gene is composed of an N-terminal, bacterium-derived chaperone DnaJ domain fused to a phycocyanobilin (PCB) lyase domain of cyanobacterial origin. PCB lyases attach bilin chromophores to light-harvesting phycobiliproteins through thioether bonds to cysteine residues. This modular protein appears to be plastid-targeted in rhodophytes. Absent functional data, the biological relevance of family 42 remains unknown but suggests the possibility of stress-dependent regulation of PCB maturation via lyase-dependent chromophore attachment.

Similarly, a central innovation in plastid evolution was the evolution of the plastid translocons (Toc/Tic) to allow the controlled entry of proteins translated in the cytosol into the organelle. We find here that domains present in translocon proteins can be recruited into S genes. This appears to be the case for family 28 (Fig. 2), which is absent from Archaeplastida but present in the red alga-derived plastid-containing stramenopiles and dinoflagellates. This modular protein is composed of an N-terminal calmodulin domain of prokaryotic (noncyanobacterial) origin fused with a cyanobacterium-derived Tic20-like domain. The Tic20 domain is widely distributed among photosynthetic eukaryotes (25, 26), where it plays an essential role in the creation of a preprotein-sensitive channel or contributes to retargeting proteins to the apicoplast in secondary plastid-containing organisms such as *Toxoplasma gondii* (27). The function of this novel S gene defies easy explanation; nonetheless, the combination of a calcium-sensing EF hand (two canonical domains exist in diatoms) with a plastid membrane channel protein suggests a role in calcium-dependent protein translocation in secondary photosynthetic eukaryotes. In pea, association between a calmodulin domain and the inner-envelope translocon component Tic32 protein has been reported, because a calmodulin binds to the C-terminal region of Tic32 in the inner chloroplast membrane, affecting channel activity (28). Interestingly, analysis of the N terminus of the S gene, uniting a calmodulin with Tic20, from the diatom *Phaeodactylum tricornutum* 219117465, provides evidence for a signal sequence cleavage site between residues 21 and 22 (SignalP 4.1) and a conserved ASAFAP motif typical for plastid-destined proteins in this species (29). RNA-seq analysis of *P. tricornutum* cultures under replete and nitrogen (N)-depleted conditions shows that the expression of this S gene is significantly down-regulated (ca. fivefold; $P = 2.57e\text{-}23$) under N stress (30) (*S-Gene Expression Analysis*).

Finally, gene family 64 (Fig. 2) might correspond to a new putative symbiogenetic bacteriorhodopsin (31–34). This protein unites a bacteriorhodopsin domain with a seven-transmembrane helical region in the N terminus, a PAS domain, and a transduction signal region of cyanobacterial origin in the C terminus. Interestingly, the transduction signal region is composed of two domains that are similar to the transduction signal region of ETR1 in *A. thaliana*: a signal transduction histidine kinase domain and a signal receiver domain (35, 36). Moreover, the N-terminal bacteriorhodopsin domain is preceded by 100 amino acids that may be involved in targeting. This S-gene family is present only in dinoflagellates.

Conclusions

In this study, we analyzed protein domain origins and identified at least 67 S genes (encompassing 2,153 coding regions) that had previously escaped detection using phylogenetic methods. S-gene functions include redox regulation, response to light, Fe–S cluster assembly, and, putatively, formation of protein channels. A total of 42% are present both in a primary photosynthetic lineage and in secondary plastid-bearing algae, suggesting their ancient emergence and their potential importance in the process of plastid establishment (Fig. 3). In contrast to these ancient S genes, 29 are lineage-specific families (43%) and were likely more recently formed, showing that cyanobacterial domain recycling is an ongoing process with a potential role in niche adaptation (Fig. 3). In addition, 55 of the S-gene products are demonstrated or predicted to be plastid-targeted (Fig. 2 and Table S1), suggesting their evolution offered an effective way to address the protein colocalization challenge in photosynthetic eukaryotes; that is, when fused with an N-terminal cyanobacterial domain that was already plastid-targeted, the novel protein did not need to "reinvent" or recruit the organelle-targeting sequence. Our results further underline the extent to which algae reuse genetic information to create not only complex structures such as the dinoflagellate "eye" (37) and metabolic pathways with chimeric gene origins (38–40) but now endosymbiont-derived composite genes with important roles in plastid maintenance. We suspect that because the number of proposed phylogenetically composite lineages continues to increase with the availability of novel genome data (41) [e.g., the photosynthetic sisters to Apicomplexa, *Chromera velia* and *Vitrella brassicaformis* (42)], our analysis provides a lower bound on S-gene numbers. Moreover, because our protocol excluded S-gene candidates present in nonphotosynthetic eukaryotes, composite genes retained in formerly photosynthetic lineages (e.g., relatives of apicomplexans) were not considered in our analysis. It is also likely that modular proteins with components derived from the mitochondrial endosymbiont will soon be discovered.

Materials and Methods

Dataset Construction. We assembled a protein sequence database by downloading every archaeal, viral, and plasmid genome that was annotated as "complete" according to the NCBI Genome database in November 2013 (152, 3,769, and 4,294 genomes, respectively). We also retrieved 230 eubacterial genomes, with 1 representative randomly chosen per eubacterial family, with the exception of cyanobacterial genomes, from which we selected 16 genomes. Finally, we sampled 38 unicellular eukaryotic genomes across the eukaryotic tree of life: 19 for photosynthetic organisms and 19 that are nonphotosynthetic, with a comparable total gene number and phylogenetic diversity in their ribosomal proteins. The resulting 2,192,940 protein sequences were compared pairwise using BLASTP (43) (version 2.2.26) (*E*-value cutoff 1e-5) (see Dataset S1 for the list of genomes used).

Detection of S-Gene Families. Composite genes and their associated component genes were detected with FusedTriplets (8) (*E* value <1e-5) by scanning the BLASTP output. Composite genes that were present in photosynthetic eukaryotes were compared with the entire nonredundant NCBI database (BLASTP; *E* value <10e-5 and ≥80% mutual sequence overlap) to confirm that these sequences had no full-length homologs outside photosynthetic eukaryotes. These composite genes were identified as candidate S genes. All sequences were also clustered into gene families according to a previous method (44, 45). Briefly, an undirected graph was constructed in which each node corresponds to a sequence and two nodes are linked if the corresponding sequences show a BLAST hit with an *E* value <1e-5, ≥30% sequence identity, and a mutual sequence overlap ≥80%. Connected components in this graph were considered to be gene families. For each candidate S gene, we retrieved the corresponding component sequences, as identified by FusedTriplets. Component sequences were clustered into component families according to the following rule: If two component sequences overlapped by more than 80% of their lengths on the protein composite, they belonged to the same component family. Component families were assigned a phylogenetic origin corresponding to their taxonomic composition. Component families were considered to be of eukaryotic origin if all their sequences belonged to eukaryotes. When one or more sequences from a component family contained prokaryotic sequences, we considered the component family to be of prokaryotic origin. If the three best prokaryotic component genes, according to their BLASTP bitscore against the composite gene, matched with the same prokaryotic phylum (e.g., Cyanobacteria), we considered the component to have more specifically originated from that prokaryotic phylum. All S-gene component origins were confirmed by BLAST analysis against an extensive prokaryotic dataset (2,982 prokaryotic genomes, 8,422,211 sequences). Only candidate S-gene families with at least one of their associated components assigned to a cyanobacterial origin (i.e., putative endosymbiotic origin) were retained.

Gene Expression and Gene Distribution Investigation. To gain insights into gene expression and distribution of S genes, composite sequences were compared with the predicted proteins of the combined assemblies of the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (46) and additional rhodophyte samples from the MMETSP (data.imicrobe.us/project/view/104) (BLASTP, *E* value <1e-5, ≥80% mutual sequence overlap) (see Dataset S1 for the list of combined assemblies used).

Prediction of Plastid Localization. ChloroP (47) (version 1.1) and ASAFind (29) (version 1.1.7) were used to predict the putative cellular localization of the 67 S proteins listed in Fig. 2. Proteomic data were also used for four species: *A. thaliana* (48), *C. reinhardtii* (49), *Cyanophora paradoxa* (50), and *Ostreococcus tauri* (51).

Exon Analysis. A total of 13 genomes had GenBank files available. For these taxa, we retrieved each exon sequence for each S gene. Exon sequences were blasted against S genes; if one exon contained all domains from the S gene according to the Conserved Domain Database (52), the corresponding S-gene family was considered as not to be subject to exon misincorporation.

ACKNOWLEDGMENTS. We thank Nicole Wagner (Rutgers) for doing the PCR and sequence analysis of the *Picochlorum* species. E.Z. and D.B. are grateful to the Rutgers University School of Environmental and Biological Sciences and members of the Genome Cooperative at School of Environmental and Biological Sciences for supporting this research. E.B. is funded by the European Research Council (FP7/2007-2013 Grant Agreement 615274).

1. Martin W, et al. (1998) Gene transfer to the nucleus and the evolution of chloroplasts. *Nature* 393(6681):162–165.
2. 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(23):2320–2325.
3. Martin W, et al. (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proc Natl Acad Sci USA* 99(19):12246–12251.
4. Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D (2004) A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol* 21(5):809–818.
5. McFadden GI (2014) Origin and evolution of plastids and photosynthesis in eukaryotes. *Cold Spring Harb Perspect Biol* 6(4):a016105.
6. Bapteste E, et al. (2012) Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proc Natl Acad Sci USA* 109(45):18266–18272.
7. Haggerty LS, et al. (2014) A pluralistic account of homology: Adapting the models to the data. *Mol Biol Evol* 31(3):501–516.
8. Jachiet P-A, Pogorelcnik R, Berry A, Lopez P, Bapteste E (2013) MosaicFinder: Identification of fused gene families in sequence similarity networks. *Bioinformatics* 29(7):837–844.
9. Jachiet P-A, Colson P, Lopez P, Bapteste E (2014) Extensive gene remodeling in the viral world: New evidence for nongradual evolution in the mobilome network. *Genome Biol Evol* 6(9):2195–2205.
10. Leonard G, Richards TA (2012) Genome-scale comparative analysis of gene fusions, gene fissions, and the fungal tree of life. *Proc Natl Acad Sci USA* 109(52):21402–21407.
11. Rockwell NC, Lagarias JC, Bhattacharya D (2014) Primary endosymbiosis and the evolution of light and oxygen sensing in photosynthetic eukaryotes. *Front Ecol Evol* 2(66).
12. Halliwell B (2006) Reactive species and antioxidants. Redox biology is a fundamental theme of aerobic life. *Plant Physiol* 141(2):312–322.
13. Liu X, et al. (2013) Structural insights into the N-terminal GIY–YIG endonuclease activity of *Arabidopsis* glutaredoxin AtGRXS16 in chloroplasts. *Proc Natl Acad Sci USA* 110(23):9565–9570.
14. Jiao Y, Ma L, Strickland E, Deng XW (2005) Conservation and divergence of light-regulated genome expression patterns during seedling development in rice and *Arabidopsis*. *Plant Cell* 17(12):3239–3256.
15. Gao H, et al. (2013) *Arabidopsis thaliana* Nfu2 accommodates [2Fe-2S] or [4Fe-4S] clusters and is competent for in vitro maturation of chloroplast [2Fe-2S] and [4Fe-4S] cluster-containing proteins. *Biochemistry* 52(38):6633–6645.
16. Schippers JHM, et al. (2008) The *Arabidopsis* onset of leaf death5 mutation of quinolinate synthase affects nicotinamide adenine dinucleotide biosynthesis and causes early ageing. *Plant Cell* 20(10):2909–2925.
17. Katoh A, Uenohara K, Akita M, Hashimoto T (2006) Early steps in the biosynthesis of NAD in *Arabidopsis* start with aspartate and occur in the plastid. *Plant Physiol* 141(3):851–857.
18. Narayana Murthy UM, et al. (2007) Characterization of *Arabidopsis thaliana* SufE2 and SufE3: Functions in chloroplast iron-sulfur cluster assembly and NAD synthesis. *J Biol Chem* 282(25):18254–18264.
19. Tan B-C, et al. (2003) Molecular characterization of the *Arabidopsis* 9-*cis* epoxycarotenoid dioxygenase gene family. *Plant J* 35(1):44–56.
20. Lundquist PK, et al. (2012) The functional network of the *Arabidopsis* plastoglobule proteome based on quantitative proteomics and genome-wide coexpression analysis. *Plant Physiol* 158(3):1172–1192.
21. Schulze T, et al. (2013) The heme-binding protein SOUL3 of *Chlamydomonas reinhardtii* influences size and position of the eyespot. *Mol Plant* 6(3):931–944.
22. Overmann J (2010) The phototrophic consortium "*Chlorochromatium aggregatum*"—A model for bacterial heterologous multicellularity. *Adv Exp Med Biol* 675:15–29.
23. Friml J, Wiśniewska J, Benková E, Mendgen K, Palme K (2002) Lateral relocation of auxin efflux regulator PIN3 mediates tropism in *Arabidopsis*. *Nature* 415(6873):806–809.
24. Lippold F, et al. (2012) Fatty acid phytyl ester synthesis in chloroplasts of *Arabidopsis*. *Plant Cell* 24(5):2001–2014.
25. Töpel M, Jarvis P (2011) The Tic20 gene family: Phylogenetic analysis and evolutionary considerations. *Plant Signal Behav* 6(7):1046–1048.
26. Kasmati AR, Töpel M, Patel R, Murtaza G, Jarvis P (2011) Molecular and genetic analyses of Tic20 homologues in *Arabidopsis thaliana* chloroplasts. *Plant J* 66(5):877–889.
27. van Dooren GG, Tomova C, Agrawal S, Humbel BM, Striepen B (2008) *Toxoplasma gondii* Tic20 is essential for apicoplast protein import. *Proc Natl Acad Sci USA* 105(36):13574–13579.
28. Chigri F, et al. (2006) Calcium regulation of chloroplast protein translocation is mediated by calmodulin binding to Tic32. *Proc Natl Acad Sci USA* 103(43):16051–16056.
29. Gruber A, Rocap G, Kroth PG, Armbrust EV, Mock T (2015) Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *Plant J* 81(3):519–528.
30. Levitan O, et al. (2015) Remodeling of intermediate metabolism in the diatom *Phaeodactylum tricornutum* under nitrogen stress. *Proc Natl Acad Sci USA* 112(2):412–417.
31. Béjà O, et al. (2000) Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science* 289(5486):1902–1906.
32. Slamovits CH, Okamoto N, Burri L, James ER, Keeling PJ (2011) A bacterial proteorhodopsin proton pump in marine eukaryotes. *Nat Commun* 2:183.
33. Avelar GM, et al. (2014) A rhodopsin-guanylyl cyclase gene fusion functions in visual perception in a fungus. *Curr Biol* 24(11):1234–1240.
34. Scheib U, et al. (2015) The rhodopsin-guanylyl cyclase of the aquatic fungus *Blastocladiella emersonii* enables fast optical control of cGMP signaling. *Sci Signal* 8(389):rs8.
35. Müller-Dieckmann HJ, Grantz AA, Kim SH (1999) The structure of the signal receiver domain of the *Arabidopsis thaliana* ethylene receptor ETR1. *Structure* 7(12):1547–1556.
36. Chang C, Kwok SF, Bleecker AB, Meyerowitz EM (1993) *Arabidopsis* ethylene-response gene ETR1: Similarity of product to two-component regulators. *Science* 262(5133):539–544.
37. Gavelis GS, et al. (2015) Eye-like ocelloids are built from different endosymbiotically acquired components. *Nature* 523(7559):204–207.
38. Oborník M, Green BR (2005) Mosaic origin of the heme biosynthesis pathway in photosynthetic eukaryotes. *Mol Biol Evol* 22(12):2343–2353.
39. Frommolt R, et al. (2008) Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Mol Biol Evol* 25(12):2653–2667.
40. Reyes-Prieto A, Bhattacharya D (2007) Phylogeny of Calvin cycle enzymes supports Plantae monophyly. *Mol Phylogenet Evol* 45(1):384–391.
41. Nelson-Sathi S, et al. (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517(7532):77–80.
42. Woo YH, et al. (2015) Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *eLife* 4:e06974.
43. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25(17):3389–3402.
44. Alvarez-Ponce D, Lopez P, Bapteste E, McInerney JO (2013) Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc Natl Acad Sci USA* 110(17):E1594–E1603.
45. Harel A, Karkar S, Cheng S, Falkowski PG, Bhattacharya D (2015) Deciphering primordial cyanobacterial genome functions from protein network analysis. *Curr Biol* 25(5):628–634.
46. Keeling PJ, 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(6):e1001889.
47. Emanuelsson O, Nielsen H, von Heijne G (1999) ChloroP, a neural network-based method for predicting chloroplast transit peptides and their cleavage sites. *Protein Sci* 8(5):978–984.
48. Sun Q, et al. (2009) PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res* 37(Database issue):D969–D974.
49. Terashima M, Specht M, Naumann B, Hippler M (2010) Characterizing the anaerobic response of *Chlamydomonas reinhardtii* by quantitative proteomics. *Mol Cell Proteomics* 9(7):1514–1532.
50. Facchinelli F, et al. (2013) Proteomic analysis of the *Cyanophora paradoxa* muroplast provides clues on early events in plastid endosymbiosis. *Planta* 237(2):637–651.
51. Le Bihan T, et al. (2011) Shotgun proteomic analysis of the unicellular alga *Ostreococcus tauri*. *J Proteomics* 74(10):2060–2070.
52. Marchler-Bauer A, et al. (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res* 43(Database issue):D222–D226.
53. Perrineau M-M, et al. (2014) Evolution of salt tolerance in a laboratory reared population of *Chlamydomonas reinhardtii*. *Environ Microbiol* 16(6):1755–1766.
54. Gorman DS, Levine RP (1965) Cytochrome *f* and plastocyanin: Their sequence in the photosynthetic electron transport chain of *Chlamydomonas reinhardti*. *Proc Natl Acad Sci USA* 54(6):1665–1669.
55. Foflonker F, et al. (2015) Genome of the halotolerant green alga *Picochlorum* sp. reveals strategies for thriving under fluctuating environmental conditions. *Environ Microbiol* 17(2):412–226.
56. Leliaert F, et al. (2012) Phylogeny and molecular evolution of the green algae. *CRC Crit Rev Plant Sci* 31(1):1–46.
57. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9):1312–1313.