Published in final edited form as:

Nat Plants. ; 3: 17066. doi:10.1038/nplants.2017.66.

Expansion of the redox sensitive proteome coincides with the plastid endosymbiosis

Christian Woehle¹, Tal Dagan¹, Giddy Landan¹, Assaf Vardi², and Shilo Rosenwasser^{2,3}

¹Institute of Microbiology, Kiel University, 24118 Kiel, Germany

²Department of Plant and Environmental Sciences, Weizmann Institute of Science, Rehovot 7610001, Israel

Abstract

The redox sensitive proteome (RSP) consists of protein thiols, in which their biochemical characteristics changed upon oxidation, playing an important role in coordinating cellular processes. Here, we applied a large-scale phylogenomic reconstruction approach in the model diatom *Phaeodactylum tricornutum* to map the evolutionary origins of the eukaryotic RSP. The majority of *P. tricornutum* redox sensitive cysteines (76%) is specific to eukaryotes, yet these are encoded in genes that are mostly of a prokaryotic origin (57%). Furthermore, we find a three-fold enrichment in redox sensitive cysteines in genes that were gained by endosymbiotic gene transfer during the primary plastid acquisition. The secondary endosymbiosis event coincides with frequent introduction of reactive cysteines into existing proteins. While the plastid acquisition imposed an increase in the production of reactive oxygen species, our results suggest that it was accompanied by significant expansion of the RSP, providing redox regulatory networks to cope with fluctuated environmental conditions.

Introduction

The origin of eukaryotes represents a major evolutionary transition in life history. According to endosymbiotic theory, photosynthetic eukaryotes arose via two symbiotic associations of prokaryotic lineages [1]. The first endosymbiont evolved into the mitochondrion, a process that was accompanied by a massive lateral gene transfer (LGT) from the endosymbiont genome into the nucleus (termed endosymbiotic gene transfer; EGT) [2]. Photosynthetic

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: http://www.nature.com/authors/editorial_policies/license.html#terms

Corresponding authors: Correspondence to Christian Woehle or Shilo Rosenwasser.

³Present address: The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, The Hebrew University, Rehovot 7610001, Israel

Data availability

The authors declare that all information on accessing data analyzed in this study is included in the paper (and its supplementary information files). The datasets generated during the current study are available from the corresponding authors upon request.

Author contributions

C.W., G.L., S.R., A.V. and T.D. conceived the study. T.D., G.L. and C.W. designed the research strategy. C.W., G.L. and S.R. performed the analyses. All authors were involved in the interpretation of results and writing of the article.

Competing financial interests

The authors declare no competing financial interests.

eukaryotes evolved from a eukaryotic ancestor that already harbored the hallmarks of eukaryotic cells including the nucleus and mitochondrion [3]. The acquisition of primary plastids involved an endosymbiosis of a cyanobacterium symbiont within a eukaryotic host and led to the evolution of Archaeplastida [3]. Secondary plastids of algae ancestry are found in multiple eukaryotic lineages, yet the number of independent plastid acquisition events in these lineages is still highly debated [4,5].

The chimeric ancestry of eukaryotes is reflected in their genomes. A substantial portion of eukaryotic genes trace back to an archaeobacterial or eubacterial ancestry [6,7]. The ancestry of eukaryotic genes is commonly correlated with their cellular function [6,8]. Genes of archaeobacterial origin typically encode for proteins that function in information processing pathways (e.g., replication, transcription and translation). Genes of proteobacterial and cyanobacterial also encode for proteins that generally function in operational processes within the cell (e.g., energy metabolism, synthesis of biomolecules, cell envelope and regulatory functions). While organelle acquisition was fundamental for the evolution of eukaryotic complexity [9], it was most probably accompanied by increased ROS (Reactive Oxygen Species) production resulted from their oxygen-based metabolic processes [10,11]. Thus, it is reasonable to hypothesize that the organelles evolution was accompanied by evolution of the mechanisms required for ROS detoxification and their integration into signaling pathways.

Oxidative stress is a unique physiological state which is characterized by modulation of gene expression patterns [12–14] and metabolic activities where ROS play a role as secondary messengers within a complex signaling network [11,15]. Post-translational thiol oxidation of redox sensitive proteins, which results in the modification of their physical structure and biochemical activity, is considered the major mechanism for redox regulation of biological processes [16]. Thus, redox sensitive proteins constitute an important component in the regulation of organelle function in eukaryotes. Yet, the evolutionary history of the eukaryotic redox proteome has been so far largely understudied.

Diatoms are a heterogeneous clade of phytoplankton that are responsible for roughly 20% of global primary productivity [17]. They belong to the Stramenopiles within the SAR supergroup (Stramenopiles, Alveolata and Rhizaria [18]), which includes species harboring secondary plastids. Diatom genomes encode for a mosaic of bacterial, plant and animal traits [19,20]. Here we examined the phylogeny of redox sensitive proteins in the diatom *P. tricornutum*. We reconstruct the origin of redox sensitive Cys residues in *P. tricornutum* proteome (RSCys) and classify them according to their ancestry. Our analysis reveals two major expansions of the eukaryotic redox signaling network, which coincide with the primary and secondary plastid acquisition during eukaryote evolution.

Results

Cysteine residue gain dynamics

Data of RSCys in the *P. tricornutum* proteome was obtained from Rosenwasser et al. 2014 [21]. All Cys residues that show no redox sensitivity or lack redox state information were cataloged as unclassified Cys (UNCys). To trace the origin of redox sensitive Cys in the *P.*

tricornutum redox sensitive proteome (RSP) we used an ancestral sequence reconstruction approach [22]. Homologs to *P. tricornutum* proteome were identified by comparing all *P. tricornutum* protein sequences to 132 proteomes from organisms representing different phyla across the tree of life, including archaeobacteria, eubacteria, protists, plants and animals (Supplementary Table S1). Each *P. tricornutum* protein sequence was aligned with its homologs and a phylogenetic tree was reconstructed using a maximum likelihood approach. The trees were rooted according to their taxonomic composition and the largest eukaryote specific clade that includes *P. tricornutum* was extracted for further analysis. This resulted in 7,118 eukaryotic gene trees. Ancestral sequences were reconstructed by PAML [23] and used to document Cys residue gains along the lineage leading to *P. tricornutum*. We distinguished between two possible scenarios for the evolution of Cys residues. Cys gains reconstructed into an existing gene were classified as *amino-acid (AA) replacement gains*, while Cys residues already present in the earliest occurrence of a gene were classified as *gene origin gains*.

An example of the Cys gain analysis is provided by 3-oxoacyl-[acyl-carrier-protein] synthase, which is involved in fatty acid metabolism and found to be redox sensitive in *P. tricornutum* as well as in tomato during infection response [24]. The *P. tricornutum* 3-oxoacyl-[acyl-carrier-protein] synthase (XP_002184832.1) includes one RSCys and ten UNCys residues and it has homologs in three other diatoms, *Symbiodinium* sp. clade B1, *Emiliana huxleyi* and *Guillardia theta*. An inference of Cys gain and loss dynamics reveals that five of the UNCys (Cys^{22,176,182,200,293}) were reconstructed to the common ancestor of all species represented in the tree. These Cys residues are classified as gains by gene origin. One RSCys (Cys⁷⁹) and two UNCys (Cys^{39,47}) residues were gained in the terminal branch leading to *P. tricornutum* (Fig. 1a). All three residues were gained in a gene already present in the genome of the diatom ancestor of *P. tricornutum*. Accordingly, they are classified as AA replacement in *P. tricornutum*. Because all homologs were found in organisms having a secondary plastid, the gene origin event of 3-oxoacyl-[acyl-carrier-protein] synthase is inferred to coincide with the secondary plastid acquisition. Another UNCys (Cys²¹⁶) was gained on an early branch that excludes *Emiliana huxleyi*. Two additional UNCys (Cys^{155,418}) are specific to diatoms, and are classified as AA replacements at the diatom ancestor. One of these UNCys (Cys¹⁵⁵) was lost again in *Thalassiosira pseudonana*.

Another example is the protein cluster of a protochlorophyllide reductase (XP_002179689.1), an enzyme involved in chlorophyll synthesis that its transcription level was shown to be sensitive to redox alterations [25]. Ancestral sequence reconstruction (Fig. 1b) shows that one RSCys (Cys³⁵⁶) and another UNCys (Cys²²⁴) are specifically gained in diatoms and *Emiliana huxleyi* by AA replacement. The remaining UNCys residues are not universal to all species in the tree, but were inferred to be present in the root and were thus classified as gene origin gains.

Similarly to the example above we analyzed 7,118 phylogenetic trees and inferred the evolution the Cys residues. In addition, we inferred the evolutionary origin of Cys residues in 841 protein sequences having a single eukaryotic homolog and 2,343 protein sequences with no homologs in other eukaryotic genomes. Our dataset comprises 54,873 Cys residues. All Cys gain events were mapped onto six main ancestral lineages along the path leading

from the last eukaryotic common ancestor (LECA) to *P. tricornutum* through the primary and secondary endosymbiosis and the evolution of Stramenopiles and diatoms (Fig. 2a). We note that eukaryotes harboring a secondary plastid may not be monophyletic [4,18]; here we grouped those taxa into a single ancestral lineage for practical reasons.

The mapping of Cys gains to ancestral lineages was determined according to the taxonomic depth of the earliest node on the protein phylogeny possessing the Cys residue. Thus, a *P. tricornutum* Cys residue that was reconstructed as present in a node ancestral to at least one member of the three non-photosynthetic eukaryotic groups (Opisthokonta, Amoebozoa and Excavata) was inferred to have been gained at the LECA (e.g., Supplementary Fig. S1a: malate synthase). Residues that were reconstructed as present in a node ancestral to members of Archaeplastida were inferred to have been gained at the ancestral lineage of primary plastid bearing eukaryotes, hence these gain events coincide with the primary plastid acquisition (e.g., Supplementary Fig. S1b: protoporphyrin IX magnesium chelatase subunit H). Nodes ancestral to any taxon whose evolution involves secondary plastid acquisition, including Haptophyta, Cryptophyceae, Rhizaria and Alveolata, were inferred as gains in the secondary plastid ancestral lineage (e.g., Supplementary Fig. S1c: 6-phosphogluconolactonase). Cys residues that were reconstructed as present in a node ancestral to a non-diatom Stramenopiles species were inferred as gains in the Stramenopiles lineage (e.g., Supplementary Fig. S1d: Nitrite reductase). Residues reconstructed as present in the ancestor of any diatom other than *P. tricornutum* were assigned to the diatoms lineage (e.g., Supplementary Fig. S1e: Cytochrome b6-f complex iron-sulfur subunit). Lastly, *P. tricornutum* specific Cys were reconstructed as gains at the *P. tricornutum* lineage (e.g., Fig. 1a: 3-oxoacyl-[acyl-carrier-protein] synthases). All RSCys and their ancestral reconstruction classification are detailed in Supplementary Table S2.

The frequency of RSCys gain events in all eukaryotic protein families, for each of the six ancestral lineages is summarized in Fig. 2b-d according to the two types of gains. The expected baseline evolutionary signal is provided by an identical analysis applied to UNCys and aspartic acid residues in the *P. tricornutum* proteome. Asp was chosen as an additional control because in the present dataset its conservation level is comparable to that of cysteine. Our analysis reveals that residue gain in the earlier lineages (e.g., LECA and primary plastid) is primarily via gene origin whereas in later nodes (e.g., Stramenopiles and diatoms) the proportion of residue gain via AA replacement increases. The relative frequency of the two types of gain is similar among the three tested residues in most ancestral lineages (Fig. 2b).

We next tested for enrichment of RSCys gain in each ancestral lineage by comparing the proportion of RSCys residues gained in the node to that of the two baseline residues ($\alpha=0.05$, using Fisher's exact test and false discovery rate (FDR)). A significant enrichment of RSCys gains via gene origin was observed in the LECA and the primary and secondary plastid endosymbiosis ancestral lineages (Fig. 2c). RSCys gains by gene origin at those three lineages are significantly more frequent than expected according to the baseline residues UNCys and Asp. Among these lineages, the highest enrichment (three-fold) was detected in the primary plastid endosymbiosis. A significant enrichment of RSCys gains via AA replacement is observed in the secondary plastid ancestral lineages (Fig. 2d). This points to

three major expansions of the RSP that coincided with the evolution of LECA (i.e., mitochondrion acquisition) and the primary and secondary plastid acquisitions. Earlier expansions at the LECA and primary plastid ancestral lineages were driven by gene origin, while the expansion at the secondary plastid acquisition was driven by gene origin as well as AA replacements in existing genes. The proportion of RSCys gains via gene origin in the *P. tricornutum* specific lineage is significantly depleted in comparison to the baseline amino acids (Fig. 2c). A similar trend is observed for RSCys gain via AA replacement at the diatoms ancestral lineage, but this observation has a weak statistical support in comparison to UNCys residues (Fig. 2d).

The observation that RSCys gain is correlated with the plastid acquisition is further supported by protein functional annotations. A test for enrichment of gene ontology (GO) terms [26] of protein sequences that contain RSCys revealed that RSCys gain events in the primary plastid ancestral lineage are enriched in protein sequences annotated with plastid- or ROS-related terms (e.g., plastid, peroxidase and chlorophyll biosynthesis). Significantly enriched GO terms associated with RSCys gains reconstructed to the secondary plastid acquisition are related to cofactor binding and pigment biosynthesis (Supplementary Table S3).

Origins of the redox sensitive proteome

The majority (76%) of *P. tricornutum* RSCys residues are observed only in eukaryotic homologs, hence they are eukaryotic specific. This indicates that the present-day RSP is mainly attributed to eukaryotic innovation. Nonetheless, the 24% RSCys with homologous residues in prokaryotes constitutes a significantly higher proportion than that observed for other Cys residues (8% p-value <0.001; Using Fisher test). Thus, a substantial proportion of RSCys was already present in the prokaryotic ancestors. We have already observed that during the early stages of eukaryote evolution the majority of RSP expansion occurred via gene origin and coincide with the organelle acquisitions. Given those two observations, we hypothesized that RSCys gained by gene origin at the three earliest ancestral lineages correspond to gene acquisition by endosymbiotic gene transfer.

We therefore classified *P. tricornutum* protein coding genes into several ancestry classes: cyanobacteria (plastid ancestor), proteobacteria (mitochondrion ancestor), bacteria, archaeobacteria and eukaryotic specific, based on the branching pattern in phylogenetic trees (see Methods). Most redox sensitive proteins (proteins with at least one RSCys residue) are observed to be of prokaryotic ancestry (57%) rather than eukaryote specific, and prokaryotic ancestry is significantly more frequent in genes encoding RSPs in comparison to other protein coding genes of *P. tricornutum* (32%; p-value <0.001 using Fisher test). Considering individual RSCys residues in comparison to the baseline residue frequencies in RSP genes, we find that genes of cyanobacterial ancestry harbor a significantly higher frequency of RSCys, nearly two-fold more frequent than the baseline residues (Table 1, $\alpha=0.05$, using Fisher test and FDR).

The distribution of protein ancestry within the ancestral lineages (Fig. 3) reveals that cyanobacterial ancestry account for 35% of the RSCys gains by gene origin in the primary plastid ancestral lineage and 47% of the gains in the secondary plastid ancestral lineage. Of

the RSCys gains via AA replacement at the secondary plastid ancestral lineage, 50% occur in genes of cyanobacterial ancestry (Supplementary Fig. S2). The high frequency of genes of cyanobacterial ancestry in the plastid ancestral lineages (primary and secondary) further supports the observation that RSCys reconstructed to those nodes were gained concomitantly with the plastid acquisition. This suggests a significant eukaryotic specific expansion of the RSP by AA replacement in genes of cyanobacterial ancestry (Fig. 3).

In proteins of cyanobacterial or proteobacterial ancestry, most RSCys gained by gene origin were also present in the prokaryotic lineages (23 out of 26 and 11 out of 15 respectively). This conservation is not restricted to the immediate ancestral group, as half of these RSCys are conserved also in other eubacterial or archaeobacterial species. Thus, several of the Cys residues that are redox sensitive in *P. tricornutum* can be traced back to the very root of the tree of life. One example of an RSCys residue with homologs in both eubacteria and archaeobacteria is the peroxiredoxin (EC45666.1; Supplementary Fig. S1f) with a catalytic Cys (Cys⁹⁸) that is conserved in all organisms, except *Aureococcus anophagefferens* that contains a Selenocysteine instead.

A recent publication of thiol oxidation data for the cyanobacterium *Synechocystis* sp. PCC 6803 during light/dark modulation and in response to photosynthesis inhibition [27] enables the comparison of *P. tricornutum* RSCys to that of a representative cyanobacterial species. A total of 962 protein-coding genes are homologous between *P. tricornutum* and *Synechocystis* sp. PCC 6803 (Supplementary Table S4). These homologs include 1,105 Cys residues, where oxidation information in both species is available for 100 residues. A comparison of the oxidation state of these homologous Cys revealed that Cys having the same redox sensitivity state in the two species are observed significantly more frequently than expected by chance (Supplementary Table S5a; p-value=0.0118, using Fisher test). The set of common RSCys consists of nine residues that are oxidized in both *P. tricornutum* and *Synechocystis* sp. PCC 6803 (Supplementary Table S5b). We note that these residues have been classified as RSCys under different physiological conditions [27]. Nonetheless the comparison shows that they are redox sensitive in both organisms. These observations, while based on a small sample size, bear witness to the contribution of plastid acquisition to the RSP, even while reflecting the divergence and independent evolution of the cyanobacterial and eukaryotic lineages following the origin of plastids.

The contribution of LGT to large scale eukaryote genome evolution outside the framework of EGT is harshly debated [28,29]. Yet, LGT is often invoked to explain sporadic occurrence of prokaryotic homologs in eukaryotic genomes (e.g., [30,31]). Our dataset includes 39 RSCys with unresolved eubacterial ancestry that cannot be clearly associated with EGT from the plastid or mitochondrion ancestors. Hypothetically, these may have been obtained via LGT from free-living prokaryotes (i.e., independently of EGT). If so, it would suggest a substantial role of LGT in the RSP expansion. However, the lack of significant enrichment for RSCys in genes of general bacterial ancestry does not indicate this (Table 1). It was recently proposed that the *P. tricornutum* genome [20] includes 587 protein coding genes acquired from prokaryotic donors at the diatoms ancestor or *P. tricornutum*. Of these, 28 proteins (comprising 34 RSCys) are included in the RSP of *P. tricornutum*. We note, however, that the original genome analysis did not make a distinction between LGT and

EGT. We reexamined the phylogeny of protein sequences included in our dataset and found that 15 of the 28 putative laterally transferred genes are of either proteobacterial or cyanobacterial ancestry. Hence, it is probable that these genes were acquired via EGT rather than LGT.

An example of a gene that was considered as an LGT-candidate in [20], but may have been classified as an EGT is a nitrite reductase protein (Nir, Supplementary Fig. S1d) that in *P. tricornutum* contains two RSCys and 19 UNCys residues, making it a potentially important redox sensitive protein. For the remaining 13 genes we could not find clear EGT signal, hence with the current data these may be considered as LGT events.

Discussion

Here we followed the evolution of the *P. tricornutum* redox sensitive protein network during major events in the diatom evolution by examining phylogenetic relations between homologous proteins and reconstruction of ancestral sequences. Our results demonstrate that the RSP is highly dynamic, with a constant flux of new genes and Cys residues expanding the repertoire of redox responsive network. The vast majority of RSP innovation is purely eukaryotic, and a substantial portion of it is recent and specific to the diatom lineage. Yet, underlying the massive eukaryotic innovations are clear footprints of two major expansions of the RSP, acquisition of redox sensitive proteins derived from the endosymbiotic origin of the chloroplast and introduction of RSCys residues into existing proteins during secondary endosymbiosis. Our results are supported by the observation of protein domains from plastid origin in proteins within eukaryotic redox regulation pathways [32] and the cyanobacterial origin of thioredoxins in plants [33]. These expansions correspond to major transitions in the eukaryotic lineage that were most probably accompanied by increased ROS pressures imposed by acquiring the mitochondrion and plastids that harbor oxygen-based metabolic processes. Our study suggests that organelle acquisitions not only generated a major ROS challenge, but also provided, via EGT, many of the proteins that are employed in the ROS signaling and response pathways.

The contribution of plastid ancestors to the evolution of redox pathways was gradual and different functions were incorporated in different stages (Supplementary Table S2). The RSP expansion during the primary plastid acquisition includes the gain of redox transmitters such as thioredoxin via EGT from cyanobacteria (Supplementary Table S2). Expanding the cellular capabilities to sense redox signals by increasing the number of redox signal transmitters such as thioredoxins and their protein targets can provide cells with highly modular sensing network of environmental cues. This plasticity can allow cells to integrate various signals in order to rapidly adjust cellular processes in a reversible manner, via reactive Cys and maintain homeostasis under fluctuated physiological conditions [34]. Our analysis uncovers a significant expansion of the redox signaling also during the secondary endosymbiosis event, indicating that a developed redox regulation could have been beneficial during the inhabitation of the new organelle. As the basic machinery for transmitting of redox signals was already acquired during the primary endosymbiosis, it is reasonable that this expansion was mainly characterized by Cys residue gains in pre-existing

proteins (Fig. 2). An example is the protochlorophyllide reductase, which is involved in chlorophyll metabolism, and gained a novel RSCys via AA replacement (Fig. 1b).

According to our analysis, the contribution of the mitochondrion acquisition to the RSP is not significantly larger than other contributors. On the other hand, enrichment of RSCys by gene origin at the LECA node and enrichment of prokaryotic inheritance was observed and may still suggest a contribution of the mitochondrial ancestor (Table 1; Fig. 2). It is possible that redox regulation was only one process among many other operational functions acquired from proteobacteria [6], and the RSP system was not over-represented among the genes acquired by EGT from the mitochondrion.

Our results show that RSCys are mainly eukaryotic innovations (76%) and therefore indicate a massive expansion of redox regulation during eukaryote diversification. However, the reconstruction of protein ancestry shows that the contribution of the prokaryotic ancestors to the RSP expansion was significant. Different stages of eukaryotic evolution were accompanied by RSCys gain and those involved many eukaryotic specific genes, but also genes of different prokaryotic origins. This is exemplified by the 40s ribosomal protein S3 phylogeny (Supplementary Fig. S1g) that shows a clear archaeobacterial ancestry and an RSCys residue gain that coincides with the LECA. In a few cases we even observed RSCys residues that are conserved in both bacteria and archaeobacteria (e.g., peroxiredoxin, Supplementary Fig. S1f). Our ability to infer whether these Cys had a singling role at the very early origin of life is limited. Nonetheless, those Cys are as ancient as LUCA (Last Universal Common Ancestor). Indeed, the finding of antioxidant enzymes in many strict anaerobes [35] suggests the emergence of ROS production and regulation mechanisms prior to the rise of oxygen in the Earth atmosphere [36]. Thus, the very ancient Cys residues we uncovered here could, potentially, already have been part of redox signaling prior to the evolution of oxygenic photosynthesis. The finding of thioredoxins and their protein targets in a methane-producing archaeon [37] as well as the recent finding of redox regulation of sulfide-dependent anoxygenic photosynthetic electron transfer in *Rhodobacter capsulatus* [38] further support this hypothesis.

Oxidation and reduction of reactive thiol proteins play a key regulatory post-translational modification that enables integration of redox signals into cellular pathways [16]. However, over oxidation of redox sensitive Cys, under high ROS pressure, subvert this regulatory role and lead to loss of protein function and its subsequent degradation [39]. Therefore, accumulation of redox-sensitive Cys during evolution present a trade-off between extended ability to accurately sense redox signals on the one hand and its sensitivity to oxidative stress on the other. Following the significant expansion of the RSP during primary and secondary endosymbiosis, the subsequent evolution of *P. tricornutum* is characterized by a paucity of gains as compared to the baseline expectation provided by UNCys and Asp. (Fig. 2c, node 6). This reduction in gaining reactive Cys residues may indicate a cost associated with higher sensitivity to toxic ROS level, and that only Cys with key regulatory and metabolic value were fixed after the initial expansion.

To conclude, we provide evidence for the major contribution of endosymbiosis events to the redox-sensitive protein network evolution. Our results point for a major role of redox regulation in coordinating organelle endosymbiosis at the eukaryotic origins.

Material and Methods

Data

P. tricornutum data, including genes protein sequences, oxidation data, Gene Ontology (GO) and targeting were obtained from [21]. Eleven genes annotated as “predicted” or “hypothetical” (out of 186) were manually annotated according to the best BLASTP hits (E -value $<10^{-5}$) in closely related taxa (see also Table S2). GO term enrichment was assessed using Ontologizer [40] with the “Topology-Weighted” algorithm. Only the top ten GO-terms by p-value were considered. LGT-candidates in *P. tricornutum* were extracted from [20] by protein identifiers and their presence in the RSCys data was inferred from 100% identity BLASTP hits.

Taxonomic classification of 74 eukaryotic and 58 prokaryotic species was determined according to NCBI taxonomy [41] and Adl et al. 2012 [18] except for Hacrobia that represent an assemblage of Haptophyta and Cryptophyceae [42]. All 1,764,232 predicted protein sequences of the selected organisms were downloaded from RefSeq (Ver. Oct-Nov 2014) [43], GenBank [44] and additional online databases (Supplementary Table S1). Identical sequences of the same organism were clustered using CD-HIT [45].

Oxidation data for *Synechocystis* sp. PCC 6803 was obtained from [27]. Peptides were mapped to the *P. tricornutum* protein sequences using BLAT Ver. 34×12 [46]). Only the first Cys residue per mapped peptide was considered. In cases of multiple oxidation data for one Cys residue that point to different types redox sensitivity (RSCys or UNCys), only the redox sensitive one was considered. Otherwise the Cys residue was counted only once.

Protein families and phylogenetic trees

A search for eukaryotic homologs to *P. tricornutum* protein sequences was performed with BLAST Ver. 2.2.29+ [47] using an E -value $<10^{-5}$ threshold. Bidirectional best BLAST hits (BBH) [48] were extracted and globally aligned with needle (EMBOSS 6.6;[49]). Homologs having 30% amino acids similarity with the *P. tricornutum* sequence were clustered into 8,017 protein families, with 2,343 sequences remaining as singletons. All eukaryotic clusters were aligned using MAFFT Ver. 7 (“linsi” option; [50]), before the addition of prokaryotic homologs. The search for prokaryotic homologs was performed with PSIBLAST [51] Ver. 2.2.29+ using the eukaryotic alignments and the unclustered singleton sequences, applying an E -value $<10^{-5}$ threshold. The BBH in each prokaryotic species was pairwise globally aligned with all cluster members and retained if at least one eukaryotic cluster member had 30% identical amino acids. Prokaryotic homologs were merged into the clusters and aligned using MAFFT Ver. 7 (“linsi” option). Maximum likelihood trees were reconstructed using PhyML (Ver. 3.0 [52]) with the “BEST” search strategy, and the inferred substitution model (LG+G+I+F) chosen as the most frequently chosen PROTTEST (Ver. 3.2 [53]) result according to corrected Akaike information criterion.

Tree rooting, gene origin and ancestral sequence reconstruction

Trees with prokaryotic homologs were rooted on the branch that splits the largest eukaryotic-only group including *P. tricornutum*. Eukaryotic clades of 3 operational taxonomic units (OTUs) were extracted for further analysis. Remaining trees without prokaryotic homologs were rooted by the midpoint approach [54]. Gene ancestry was determined according to nearest neighbor of the root split (see Supplementary Fig. S3 for details).

Ancestral sequence reconstruction was performed using PAML Ver. 4.7 [23] with the LG substitution model applying the marginal reconstruction approach. The input constituted the rooted trees with the corresponding OTU sequences from the multiple sequence alignments. Gapped positions and ambiguous characters were not considered. Resulting support values for ancestral states exceeded 0.9 for 79% of the RSCys residues and 61% of the UNCys residues. AA replacements were inferred as amino acid substitutions resulting in the residue found in the genome of *P. tricornutum*. If no AA replacements were determined for a specific amino acid residue, it was defined to be acquired at the root of the tree and classified as a gain by gene origin, as were amino acids in singleton proteins (i.e. those without eukaryotic homologs). In protein clusters with two members only, amino acid gains were classified as gene origin in their common ancestor if they are conserved or as AA replacement if the amino acid residue is present only in *P. tricornutum*.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Anne Kupzok, Judith Ilhan, Julia Weissenbach, Claudia Walda, Andrea Mrnjavac and Tanita Wein for critical comments on the manuscript. This project was supported by the European Research Council (Grant No. 281357 awarded to T.D. and 280991 awarded to A.V.), the Israeli Science Foundation (Grant 25 No. 712233 awarded to A.V.) and the cluster of excellence The Future Ocean (funded within the framework of the Excellence Initiative by the Deutsche Forschungsgemeinschaft (DFG) on behalf of the German federal and state governments).

References

1. Embley TM, Martin W. Eukaryotic evolution, changes and challenges. *Nature*. 2006 Mar 30; 440(7084):623–30. [PubMed: 16572163]
2. Timmis, JN., Ayliffe, MA., Huang, CY., Martin, W. *Nat Rev Genet*. Vol. 5. Nature Publishing Group; 2004 Feb. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes; p. 123-35.
3. Archibald JM. The puzzle of plastid evolution. *Curr Biol*. 2009 Jan 27; 19(2):R81–8. [PubMed: 19174147]
4. Baurain D, Brinkmann H, Petersen J, Rodriguez-Ezpeleta N, Stechmann A, Demoulin V, et al. Phylogenomic Evidence for Separate Acquisition of Plastids in Cryptophytes, Haptophytes, and Stramenopiles. *Mol Biol Evol*. 2010 Jun 13; 27(7):1698–709. [PubMed: 20194427]
5. Gould SB, Maier U-G, Martin WF. Protein import and the origin of red complex plastids. *Curr Biol*. 2015 Jun 15; 25(12):R515–21. [PubMed: 26079086]
6. Esser, C., Ahmadinejad, N., Wiegand, C., Rotte, C., Sebastiani, F., Gelius-Dietrich, G., et al. *Mol Biol Evol*. Vol. 21. Oxford University Press; 2004 Sep. A genome phylogeny for mitochondria

- among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes; p. 1643-60.
7. Deusch, O., Landan, G., Roettger, M., Gruenheit, N., Kowallik, KV., Allen, JF., et al. *Mol Biol Evol.* Vol. 25. Oxford University Press; 2008 Apr. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor; p. 748-61.
8. Alvarez-Ponce, D., McInerney, JO. *Genome Biol Evol.* Vol. 3. Oxford University Press; 2011. The human genome retains relics of its prokaryotic ancestry: human genes of archaeobacterial and eubacterial origin exhibit remarkable differences; p. 782-90.
9. Lane, N., Martin, W. *Nature.* Vol. 467. Nature Research; 2010 Oct 21. The energetics of genome complexity; p. 929-34.
10. Mittler R, Vanderauwera S, Gollery M, Van Breusegem F. Reactive oxygen gene network of plants. *Trends Plant Sci.* 2004 Oct; 9(10):490–8. [PubMed: 15465684]
11. Halliwell, B. *Plant Physiol.* Vol. 141. American Society of Plant Biologists; 2006 Jun. Reactive species and antioxidants. Redox biology is a fundamental theme of aerobic life; p. 312-22.
12. Rosenwasser, S., Fluhr, R., Joshi, JR., Leviatan, N., Sela, N., Hetzroni, A., et al. *Plant Physiol.* Vol. 163. American Society of Plant Biologists; 2013 Oct. ROSMETER: a bioinformatic tool for the identification of transcriptomic imprints related to reactive oxygen species type and origin provides new insights into stress responses; p. 1071-83.
13. Gadjev, I., Vanderauwera, S., Gechev, TS., Laloi, C., Minkov, IN., Shulaev, V., et al. *Plant Physiol.* Vol. 141. American Society of Plant Biologists; 2006 Jun. Transcriptomic footprints disclose specificity of reactive oxygen species signaling in Arabidopsis; p. 436-45.
14. Willems, P., Mhamdi, A., Stael, S., Storme, V., Kerchev, P., Noctor, G., et al. *Plant Physiol.* Vol. 171. American Society of Plant Biologists; 2016 Jul. The ROS Wheel: Refining ROS Transcriptional Footprints; p. 1720-33.
15. Laloi C, Apel K, Danon A. Reactive oxygen signalling: the latest news. *Curr Opin Plant Biol.* 2004 Jun; 7(3):323–8. [PubMed: 15134754]
16. Winterbourn CC, Hampton MB. Thiol chemistry and specificity in redox signaling. *Free Radic Biol Med.* 2008 Sep 1; 45(5):549–61. [PubMed: 18544350]
17. Nelson DM, Tréguer P, Brzezinski MA, Leynaert A, Quéguiner B. Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochem Cycles.* 1995 Sep 1; 9(3):359–72.
18. Adl SM, Simpson AGB, Lane CE, Lukes J, Bass D, Bowser SS, et al. The revised classification of eukaryotes. *J Eukaryotic Microbiology.* 2012 Sep; 59(5):429–93. [PubMed: 23020233]
19. Allen, AE., Dupont, CL., Oborník, M., Horák, A., Nunes-Nesi, A., McCrow, JP., et al. *Nature.* Vol. 473. Nature Publishing Group; 2011 May 12. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms; p. 203-7.
20. Bowler, C., Allen, AE., Badger, JH., Grimwood, J., Jabbari, K., Kuo, A., et al. *Nature.* Vol. 456. Nature Publishing Group; 2008 Nov 13. The Phaeodactylum genome reveals the evolutionary history of diatom genomes; p. 239-44.
21. Rosenwasser, S., Graff van Creveld, S., Schatz, D., Malitsky, S., Tzfadia, O., Aharoni, A., et al. *Proc Natl Acad Sci USA.* Vol. 111. National Acad Sciences; 2014 Feb 18. Mapping the diatom redox-sensitive proteome provides insight into response to nitrogen stress in the marine environment; p. 2740-5.
22. Yang, Z., Kumar, S., Nei, M. *Genetics.* Vol. 141. Genetics Society of America; 1995 Dec. A new method of inference of ancestral nucleotide and amino acid sequences; p. 1641-50.
23. Yang, Z. *Mol Biol Evol.* Vol. 24. Oxford University Press; 2007 Aug. PAML 4: phylogenetic analysis by maximum likelihood; p. 1586-91.
24. Balmant, KM., Parker, J., Yoo, M-J., Zhu, N., Dufresne, C., Chen, S. *Hortic Res.* Vol. 2. Nature Publishing Group; 2015. Redox proteomics of tomato in response to *Pseudomonas syringae* infection; p. 15043
25. Queval, G., Foyer, CH. *Philos Trans R Soc Lond B Biol Sci.* Vol. 367. The Royal Society; 2012 Dec 19. Redox regulation of photosynthetic gene expression; p. 3475-85.

26. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000 May; 25(1):25–9. [PubMed: 10802651]
27. Guo, J., Nguyen, AY., Dai, Z., Su, D., Gaffrey, MJ., Moore, RJ., et al. *Mol Cell Proteomics*. Vol. 13. American Society for Biochemistry and Molecular Biology; 2014 Dec. Proteome-wide light/dark modulation of thiol oxidation in cyanobacteria revealed by quantitative site-specific redox proteomics; p. 3270-85.
28. Ku C, Nelson-Sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, et al. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature.* 2015 Aug 27; 524(7566):427–32. [PubMed: 26287458]
29. Koutsovoulos, G., Kumar, S., Laetsch, DR., Stevens, L., Daub, J., Conlon, C., et al. *Proc Natl Acad Sci USA*. Vol. 113. National Acad Sciences; 2016 May 3. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*; p. 5053-8.
30. Schönknecht, G., Chen, W-H., Ternes, CM., Barbier, GG., Shrestha, RP., Stanke, M., et al. *Science*. Vol. 339. American Association for the Advancement of Science; 2013 Mar 8. Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote; p. 1207-10.
31. Boothby, TC., Tenlen, JR., Smith, FW., Wang, JR., Patanella, KA., Osborne Nishimura, E., et al. *Proc Natl Acad Sci USA*. National Acad Sciences; 2015 Nov 23. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. 201510461
32. Méheust, R., Zelzion, E., Bhattacharya, D., Lopez, P., Baptiste, E. *Proc Natl Acad Sci USA*. Vol. 113. National Acad Sciences; 2016 Mar 29. Protein networks identify novel symbiogenetic genes resulting from plastid endosymbiosis; p. 3579-84.
33. Balsera M, Uberegui E, Schürmann P, Buchanan BB. Evolutionary development of redox regulation in chloroplasts. *Antioxid Redox Signal.* 2014 Sep 20; 21(9):1327–55. [PubMed: 24483204]
34. Scheibe, R., Dietz, K-J. *Plant Cell Environ.* Vol. 35. Blackwell Publishing Ltd; 2012 Feb. Reduction-oxidation network for flexible adjustment of cellular metabolism in photoautotrophic cells; p. 202-16.
35. Imlay, JA. *J Biol Inorg Chem*. Vol. 7. Springer-Verlag; 2002. What biological purpose is served by superoxide reductase?; p. 659-63.
36. Slesak, I., Slesak, H., Kruk, J. *Astrobiology*. Vol. 12. Mary Ann Liebert, Inc; 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA; 2012 Aug. Oxygen and hydrogen peroxide in the early evolution of life on earth: in silico comparative analysis of biochemical pathways; p. 775-84.
37. Susanti, D., Wong, JH., Vensel, WH., Loganathan, U., DeSantis, R., Schmitz, RA., et al. *Proc Natl Acad Sci USA*. Vol. 111. National Acad Sciences; 2014 Feb 18. Thioredoxin targets fundamental processes in a methane-producing archaeon, *Methanocaldococcus jannaschii*; p. 2608-13.
38. Shimizu T, Shen J, Fang M, Zhang Y, Hori K, Trinidad JC, et al. Sulfide-responsive transcriptional repressor SqrR functions as a master regulator of sulfide-dependent photosynthesis. *Proc Natl Acad Sci USA*. 2017 Feb 28; 114(9):2355–60. [PubMed: 28196888]
39. Poole LB. The basics of thiols and cysteines in redox biology and chemistry. *Free Radic Biol Med*. 2015 Mar;80:148–57. [PubMed: 25433365]
40. Bauer, S., Grossmann, S., Vingron, M., Robinson, PN. *Bioinformatics*. Vol. 24. Oxford University Press; 2008 Jul 15. Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration; p. 1650-1.
41. Federhen, S. *Nucleic Acids Res*. Vol. 40. Oxford University Press; 2012 Jan. The NCBI Taxonomy database; p. D136-43.
42. Okamoto, N., Chantangsri, C., Horák, A., Leander, BS., Keeling, PJ. Molecular phylogeny and description of the novel katablepharid *Roombia truncata* gen. et sp. nov., and establishment of the *Hacrobia* taxon nov. In: Stajich, JE., editor. *PLoS One*. Vol. 4. Public Library of Science; 2009. p. e7080
43. O'Leary, NA., Wright, MW., Brister, JR., Ciufo, S., Haddad, D., McVeigh, R., et al. *Nucleic Acids Res*. Oxford University Press; 2015 Nov 8. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. gkv1189

44. Benson, DA., Clark, K., Karsch-Mizrachi, I., Lipman, DJ., Ostell, J., Sayers, EW. Nucleic Acids Res. Vol. 43. Oxford University Press; 2015 Jan. GenBank; p. D30-5.
45. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Oxford University Press. 2012 Dec 1; 28(23):3150–2.
46. Kent, WJ. Genome Res. Vol. 12. Cold Spring Harbor Lab; 2002 Apr. BLAT--the BLAST-like alignment tool; p. 656-64.
47. Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., et al. BMC Bioinformatics. Vol. 10. BioMed Central Ltd; 2009. BLAST+: architecture and applications; p. 421
48. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. Science. 1997 Oct 24; 278(5338):631–7. [PubMed: 9381173]
49. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000 Jun; 16(6):276–7. [PubMed: 10827456]
50. Katoh, K., Standley, DM. Mol Biol Evol. Vol. 30. Oxford University Press; 2013 Apr. MAFFT multiple sequence alignment software version 7: improvements in performance and usability; p. 772-80.
51. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997 Aug 31; 25(17):3389–402. [PubMed: 9254694]
52. Guindon, S., Delsuc, F., Dufayard, J-F., Gascuel, O. Methods Mol Biol. Vol. 537. Totowa, NJ: Humana Press; 2009. Estimating maximum likelihood phylogenies with PhyML; p. 113-37.
53. Darriba, D., Taboada, GL., Doallo, R., Posada, D. Bioinformatics. Vol. 27. Oxford University Press; 2011 Apr 15. ProtTest 3: fast selection of best-fit models of protein evolution; p. 1164-5.
54. Farris JS. Estimating phylogenetic trees from distance matrices. American Naturalist. 1972

Figure 1. Ancestral sequence reconstruction example:

The phylogenetic tree and an outline of the underlying multiple sequence alignment describe the Cys dynamics inferred from ancestral sequence reconstruction **(a)** for an 3-oxoacyl-[acyl-carrier-protein] synthase (XP_002184832.1) and **(b)** for a protochlorophyllide reductase (XP_002179689.1). RSCys (red) and UNCys (blue) labels indicate gains (+), losses (-), and residue position in *P. tricornutum*. Arcs highlight *P. tricornutum* (black) and other taxonomic groups. Bootstrap values >70 are shown. The outline of the multiple sequence alignment shows the neighborhood of Cys residues and their positions *P. tricornutum*. The order of sequences matches the OTUs in the phylogenetic tree.

Figure 2. Cys gains in the eukaryotic ancestry of *P. tricornutum*.

(a) Schematic tree for the classification of Cys gains into one of six ancestral lineages. Dashed lines indicate the unresolved relationship of plastids originated from secondary endosymbiosis and the grouping of Hacrobia with SAR. (b-d) Comparison gain frequencies of RSCys to UNCys and Asp, in each of the six ancestral lineages of (a). (b) Relative frequency of gains via gene origin or AA replacement. (c) Frequency of gains via gene origin. (d) Frequency of gains via AA replacement. Asterisks (*) mark significant differences between RSCys and UNCys and Asp, numbers are adjusted p-values (Fisher test)

with FDR correction for multiple testing with $\alpha=0.05$). Arrows mark enrichment ($\uparrow$) and depletion ($\downarrow$) in the RSCys column.

Figure 3. Prokaryotic gene ancestry of RSCys gains via gene origin in the six ancestral lineages. Layout as described in Fig. 2. Gene ancestry (colors) were inferred from the nearest neighbor of the eukaryotic clade (see Table 1). Prim. plastid, Primary plastid endosymbiosis; Sec. plastid, Secondary plastid endosymbiosis.

Table 1
Amino acid gains in redox sensitive proteins classified by gene ancestry.

Gene ancestry ^I	RSCys	UNCys	Asp
Cyanobacteria	42 (18%) [↑]	101 (10%) [*]	502 (11%) [*]
Proteobacteria	30 (13%)	153 (14%)	521 (12%)
Bacteria	39 (17%)	145 (14%)	650 (14%)
Archaea	11 (5%)	36 (3%)	126 (3%)
Eukaryotic	102 (44%) [↓]	611 (57%) [*]	2,654 (59%) [*]
Unclassified	7 (3%)	20 (2%)	81 (2%)
Total	231	1,066	4,533

^I Gene ancestry inferred from sister clade relationship in phylogenetic trees of RSP genes.

^{*} Significant enrichment ([↑]) or depletion ([↓]) in RSCys (Fisher test with FDR correction for multiple testing with $\alpha=0.05$)