

Genomic Footprints of a Cryptic Plastid Endosymbiosis in Diatoms

Ahmed Moustafa,^{1*} Bank Beszteri,^{2*} Uwe G. Maier,³ Chris Bowler,^{4,5}
Klaus Valentin,² Debashish Bhattacharya^{1,6†}

Diatoms and other chromalveolates are among the dominant phytoplankters in the world's oceans. Endosymbiosis was essential to the success of chromalveolates, and it appears that the ancestral plastid in this group had a red algal origin via an ancient secondary endosymbiosis. However, recent analyses have turned up a handful of nuclear genes in chromalveolates that are of green algal derivation. Using a genome-wide approach to estimate the "green" contribution to diatoms, we identified >1700 green gene transfers, constituting 16% of the diatom nuclear coding potential. These genes were probably introduced into diatoms and other chromalveolates from a cryptic endosymbiont related to prasinophyte-like green algae. Chromalveolates appear to have recruited genes from the two major existing algal groups to forge a highly successful, species-rich protist lineage.

Diatoms are well-studied members of the putative supergroup Chromalveolata [fig. S1 and supporting online material (SOM) text] and comprise unicellular, photosynthetic, dominant taxa in the marine phytoplankton. Diatoms are central to understanding oceanic primary production and biogeochemistry (1). Much effort is currently being expended to develop some taxa as models for genetic and genomic research as well as sources for biofuel (2) and nanotechnology (3). We conducted a phylogenomic analysis of the diatom proteome using complete genome data from *Thalassiosira* and *Phaeodactylum*. This procedure identified 2423 and 2533 (2423/2533) *Phaeodactylum* and *Thalassiosira* genes, respectively (this order of results is used throughout the paper and SOM), that are derived from red or green algal sources. Contrary to the expectation of the chromalveolate hypothesis (4), however, >70% of these genes are of green (not red) lineage provenance (Fig. 1, table S1, and fig. S7). This green gene contribution constitutes ~16% of the diatom proteome. Two of the major topological classes that were uncovered are shown in fig. S2. The first class (fig. S2A) contains 442/442 trees in which both red and green algae are present, but there is robust bootstrap support for the green algae plus diatom (and other chromalveolates) clade. Of these trees, 144/133 show the green algal and diatom sequences to diverge within the Plantae

kingdom, which is composed of green algae and plants, glaucophytes, and red algae (5, 6). An example is phytoene desaturase (fig. S2A), which is an early enzyme in plastid carotenoid biosynthesis. It was previously reported that 5 of the 16 genes in this photoprotective pathway are of green algal origin (7), and their occurrence in chromalveolates probably ensures a high photosynthetic efficiency under fluctuating light (8). The remaining 298/309 trees indicate an independent origin of the gene in the donor green algae, with respect to other Plantae. A second major class of trees shows an independent gene origin in prasinophytes relative to other green algae and plants, before being transferred to diatoms and other chromalveolates (Fig. 2). The absence of red algal homologs in some trees in this class may be explained by gene loss in the reduced nuclear genome of the red algal representative in our database, *Cyanidioschyzon merolae* (SOM text). An example tree from this class is a member of the isoprenylcysteine carboxyl methyltransferase superfamily (fig. S2B). We found four genes that encode the following gene products: naphthoate synthase (GenBank GI number 219114006), heme oxygenase (GenBank GI number 219117865), pyruvate dehydrogenase (GenBank GI number 219119135), and GUN4-like protein (GenBank GI number 219127880), which are retained in red algal plastid genomes but absent from this red-derived organelle genome in diatoms. These sequences are present in the diatom nucleus but are of green algal derivation. This suggests that red plastid-encoded genes were lost if green homologs were already present in the host nucleus.

To identify the putative sources of the diatom green genes, we examined their distribution among the green lineages (Viridiplantae). The Viridiplantae comprise two well-supported phyla, the Chlorophyta (most green algae, such as *Chlamydomonas*, in the core chlorophytes and the prasinophytes) and the Streptophyta (charophyte green algae and all land plants; Fig. 2A). The prasinophytes include the world's smallest

eukaryotes (the picoeukaryote *Ostreococcus*; cell diameter ~1 μm), which are part of a morphologically diverse group of paraphyletic lineages diverging at the base of the Chlorophyta (9). We found that 637/716 diatom green genes (36/41%) trace their origin to the prasinophytes in our data set (*Micromonas* and *Ostreococcus*; Mamiellales clade) of which 167/175 are shared with other Chlorophyta (71/67 genes; *Chlamydomonas* and *Volvox*) or Streptophyta (23/40 genes; *Arabidopsis*, *Oryza*, *Physcomitrella*, and *Zea*) or by both phyla (73/68 genes; Fig. 2B). These 167/175 genes have a putative ancient origin in Viridiplantae. Streptophyte- and core chlorophyte-specific donors account for 192/177 and 145/170 genes, respectively (Fig. 2C). Many of these genes may ancestrally have been present in the Viridiplantae and lost by prasinophytes and/or other green lineage members, whereas the remainder represent independent horizontal gene transfers (HGTs) into streptophytes and core chlorophytes. In spite of the reduced nuclear genome of the prasinophytes in our study (~9000 protein-encoding genes) as compared to the larger genomes of core chlorophytes and streptophytes (~15,000 and ~30,000 protein-encoding genes, respectively), 470/541 genes are shared exclusively between prasinophytes and diatoms (Fig. 2, B and C), of which 462/502 (98 and 93%) are present in expressed sequence tag (EST) libraries from *Phaeodactylum* and *Thalassiosira*. Because of their specific affiliation with picoprasinophytes, these genes are unlikely to represent missing sequences from *Cyanidioschyzon*. This diatom green gene set may therefore be gene recruitments via HGT in picoprasinophytes that were later trans-

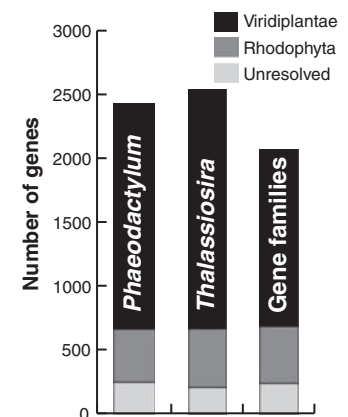


Fig. 1. Diatom genes of a red or green algal origin that were identified using phylogenomic analysis of complete genome data. Each bar represents the total number of algal genes in the corresponding diatom species. The "gene families" bar indicates the total number of transferred genes in both diatoms after clustering the data into gene families through single-linkage hierarchical clustering. The "unresolved" category indicates that red and green algae are sisters of each other in the tree and monophyletic with diatoms (and other chromalveolates).

¹Interdisciplinary Program in Genetics, University of Iowa, Iowa City, IA 52242, USA. ²Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, 27570 Bremerhaven, Germany. ³Zellbiologie, Philipps-Universität Marburg, Marburg, Germany. ⁴CNRS UMR8186, Department of Biology, Ecole Normale Supérieure, 46 rue d'Ulm, 75005 Paris, France. ⁵Stazione Zoologica Anton Dohrn, Villa Comunale, I-80121 Naples, Italy. ⁶Department of Biological Sciences and the Roy J. Carver Center for Comparative Genomics, University of Iowa, Iowa City, IA 52242, USA.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: debashi-bhattacharya@uiowa.edu

ferred to the diatom (chromalveolate) nucleus. These sequences could hold clues to the evolution of prasinophyte green algae and their great success in different aquatic environments (10).

The fourfold higher abundance of green versus red genes in diatoms raises questions about the timing of the transfer of the green genes and whether these sequences were introduced via a single or multiple endosymbioses, or by unprecedented levels of HGT in chromalveolates. In order to address this issue, we determined the distribution of diatom green genes among

chromalveolates using complete genome data. Here, the distinction between gene origins via endosymbiotic gene transfer (EGT) (11) versus HGT reflects whether the genes can be traced back to a point source (prasinophyte-like algae) and are found in most if not all chromalveolates, versus sporadic gene origin in particular lineages and from multiple different sources, respectively. Neither of these outcomes is proof but rather argues for or against one hypothesis. Using this approach, we find that 85% of the green genes can be traced back to the ancestor of both dia-

toms and other Stramenopiles (Fig. 3). Diatoms share 46/55 green genes with the obligate parasites apicomplexans and 54/63 genes with the plastid-lacking ciliates. Analysis of genome data from the distantly related photosynthetic coccolithophorid *Emiliania huxleyi*, which is a haptophyte sister to cryptophytes (fig. S1), identified >400 green genes shared with diatoms. The inclusion of ESTs from dinoflagellates and cryptophyte algae shows that even when these partial data are used, 10 and 3% of the diatom green genes are shared with these groups, respectively (fig. S3). Given these results, we suggest that despite extensive gene losses among nonphotosynthetic lineages such as ciliates and apicomplexans, the most likely explanation is that a large proportion of the diatom green genes is of an ancient provenance and predate the split of cryptophytes and haptophytes from other chromalveolates.

Taken together, our results provide evidence of a prasinophyte-like endosymbiont in the common ancestor of chromalveolates. As discussed above, prasinophytes are an anciently diverged paraphyletic group of green algae (12) that was present early on in chromalveolate evolution. In the fossil record, prasinophytes are widely distributed by the Early Cambrian (13). These cells may well have been an abundant prey source for the chromalveolate ancestor. The alternative explanation of chromalveolate polyphyly would imply an unprecedented number of independent gains (~400) of the same green genes by diatoms and haptophytes. Therefore, our results provide strong support for a shared evolutionary history for these disparate chromalveolate lineages. In substantiated cases of serial endosymbiosis, the most recent endosymbiont provides the plastid, whereas the nuclear genome bears the footprints of past events. The dinoflagellates provide several independent examples of this phenomenon with the replacement of the broadly distributed red algal (peridinin-containing) plastid in different taxa with one of green, cryptophyte, or diatom origin (14, 15). Therefore, the presence of a red algal-derived plastid in most photosynthetic chromalveolates is most easily explained by the green algal endosymbiosis having predated the red algal capture (7, 16, 17).

A different interpretation of our green gene data is that these sequences did not derive from EGT and HGT but rather support a bona fide sister-group relationship between chromalveolates and green algae. Under this scenario, the chromalveolate ancestor contained a plastid of primary endosymbiotic origin [cyanobacterial (18)] that was shared with the green lineage and subsequently replaced by one of secondary (red algal) derivation. Although possible, this scenario is highly implausible because it not only argues against Plantae monophyly, which has been supported by recent phylogenomic studies (6, 19), but more importantly, demands that the vast majority of chromalveolate nuclear genes with nonplastid functions

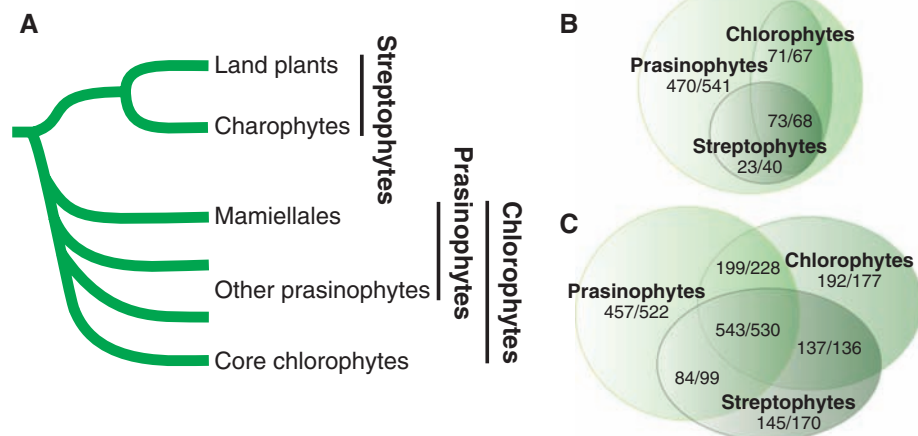
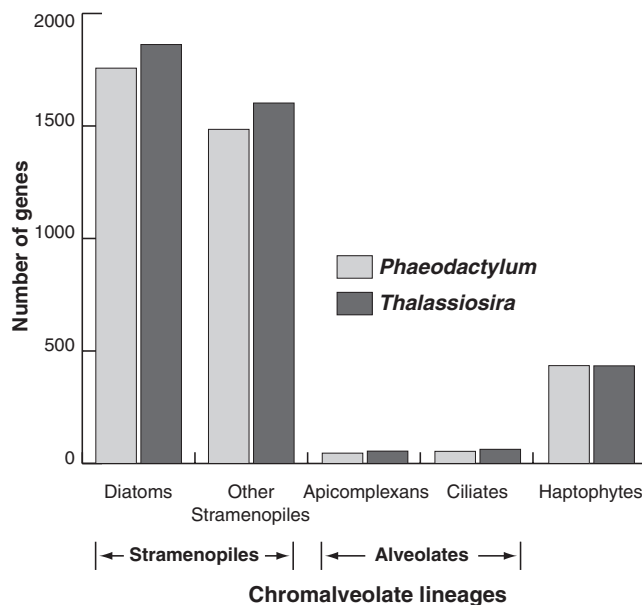


Fig. 2. Phylogenetic distribution of diatom genes of green algal origin among Viridiplantae. **(A)** Schematic tree that illustrates well-accepted phylogenetic relationships within the green lineage. **(B)** Venn diagram depicting the distribution of diatom green genes of prasinophyte origin. These genes support a specific sister-group relationship between prasinophytes and diatoms (and other chromalveolates). The two broad categories of gene sharing are as follows: (i) the gene is exclusive to prasinophytes (470/541), and (ii) the gene is shared with other Viridiplantae. **(C)** Venn diagram depicting the distribution of all diatom green genes among Viridiplantae. Here, 192/177 genes are of chlorophyte origin, whereas 145/170 genes are apparently derived from streptophytes. It should be noted that these are provisional values and will be affected by the strength of the phylogenetic signal in any given protein or the absence of data from particular groups; that is, some apparently streptophyte-specific diatom green genes may simply be explained by the loss of the genes in other Viridiplantae (such as prasinophytes).

Fig. 3. The distribution of diatom green genes among different chromalveolates. The value for each major chromalveolate lineage represents the number of proteins that satisfy two phylogenetic criteria: (i) monophyly of diatoms and the chromalveolate lineage in question, and (ii) monophyly of this clade with Viridiplantae. The category "other Stramenopiles" includes the pelagophyte *Aureococcus anophagefferens* and the oomycetes *Phytophthora capsici*, *P. ramorum*, and *P. sojae*, which have complete nuclear genome data available. Data from the remaining taxa in this category are organelle- or EST-derived.



(actin and tubulins) be directly related to Viridiplantae. Although most single- and multi-gene trees clearly demonstrate Viridiplantae monophyly (20), they do not, however, support a specific affiliation between greens and chromalveolates. There is no reason to expect that this phylogenetic signal would have been lost from chromalveolate genomes while being retained by Viridiplantae. Therefore, given the known proclivity of endosymbiosis to drive intracellular gene transfer (21, 22) and the absence of evidence for a specific phylogenetic relationship between Viridiplantae and chromalveolates, outside of the 16% reported here, we suggest that the green “footprint” in chromalveolates (although substantial) probably reflects a combination of EGT and HGT rather than a host affiliation.

The rise to prominence in the oceans by diatoms and other chromalveolates such as dinoflagellates and haptophytes after the end-Permian mass extinction (250 million years ago) has been interpreted as the victory of red plastid lineages over the predominant green plastid taxa such as prasinophytes. Changing nearshore ocean chemistry is thought to underlie this globally important phenomenon (13, 23). In contrast to current thinking, our findings show that chromalveolates were already green before they acquired the red plastid. Although ancient, these two endosymbioses that were

supplemented by subsequent HGTs supplied chromalveolates such as diatoms [≈100,000 extant species (24)] with the genetic potential to become some of the most ecologically successful and dominant marine primary producers on our planet.

References and Notes

- C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, *Science* **281**, 237 (1998).
- G. C. Dismukes, D. Carrieri, N. Bennette, G. M. Ananyev, M. C. Posewitz, *Curr. Opin. Biotechnol.* **19**, 235 (2008).
- N. Kroger, N. Poulsen, *Annu. Rev. Genet.* **42**, 83 (2008).
- T. Cavalier-Smith, *J. Eukaryot. Microbiol.* **46**, 347 (1999).
- A. Reyes-Prieto, D. Bhattacharya, *Mol. Phylogenet. Evol.* **45**, 384 (2007).
- N. Rodriguez-Ezpeleta *et al.*, *Curr. Biol.* **15**, 1325 (2005).
- R. Frommolt *et al.*, *Mol. Biol. Evol.* **25**, 2653 (2008).
- S. Coesel, M. Obornik, J. Varela, A. Falciatore, C. Bowler, *PLoS One* **3**, e2896 (2008).
- M. Turmel, M. C. Gagnon, C. J. O’Kelly, C. Otis, C. Lemieux, *Mol. Biol. Evol.* **26**, 631 (2008).
- E. Derelle *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11647 (2006).
- W. Martin, H. Brinkmann, C. Savonna, R. Cerff, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 8692 (1993).
- J. Steinkötter, D. Bhattacharya, I. Semmelroth, C. Bibeau, M. Melkonian, *J. Phycol.* **30**, 340 (1994).
- P. G. Falkowski, A. H. Knoll, *Evolution of Primary Producers in the Sea* (Elsevier Academic, Amsterdam, 2007).
- K. Ishida, B. R. Green, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 9294 (2002).
- T. Tengs *et al.*, *Mol. Biol. Evol.* **17**, 718 (2000).
- J. Petersen, R. Teich, H. Brinkmann, R. Cerff, *J. Mol. Evol.* **62**, 143 (2006).

- A. Reyes-Prieto, A. Moustafa, D. Bhattacharya, *Curr. Biol.* **18**, 956 (2008).
- M. M. Hauber, S. B. Muller, V. Speth, U. G. Maier, *Bot. Acta* **107**, 383 (1994).
- F. Burki, K. Shalchian-Tabrizi, J. Pawlowski, *Biol. Lett.* **4**, 366 (2008).
- H. S. Yoon *et al.*, *BMC Evol. Biol.* **8**, 14 (2008).
- W. Martin *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12246 (2002).
- A. Reyes-Prieto, J. D. Hackett, M. B. Soares, M. F. Bonaldo, D. Bhattacharya, *Curr. Biol.* **16**, 2320 (2006).
- H. R. Thierstein, J. R. Young, *Coccolithophores: From Molecular Processes to Global Impact* (Springer, Berlin, 2004).
- F. E. Round, R. M. Crawford, D. G. Mann, *The Diatoms: Biology and Morphology of the Genera* (Cambridge Univ. Press, Cambridge, 1990).
- D.B. was supported by grants from NSF and NIH (EF 04-31117 and R01E5013679, respectively). A.M. was supported by an Institutional National Research Service Award (T 32 GM98629) from NIH. U.G.M. thanks the Deutsche Forschungsgemeinschaft (grant SFB-TR1) for support. We thank T. Mock for providing tiling array-generated diatom transcripts and J. E. DeReus at the High Performance Computing Facility at the University of Iowa for technical support.

Supporting Online Material

www.sciencemag.org/cgi/content/full/324/5935/1724/DC1
Materials and Methods
SOM Text
Figs. S1 to S7
Tables S1 and S2
References

3 March 2009; accepted 22 May 2009
10.1126/science.1172983

Solution Nuclear Magnetic Resonance Structure of Membrane-Integral Diacylglycerol Kinase

Wade D. Van Horn,^{1*} Hak-Jun Kim,^{1,2*} Charles D. Ellis,¹ Arina Hadziselimovic,¹ Endah S. Sulistijo,¹ Murthy D. Karra,¹ Changlin Tian,^{1,3} Frank D. Sönnichsen,⁴ Charles R. Sanders^{1†}

Escherichia coli diacylglycerol kinase (DAGK) represents a family of integral membrane enzymes that is unrelated to all other phosphotransferases. We have determined the three-dimensional structure of the DAGK homotrimer with the use of solution nuclear magnetic resonance. The third transmembrane helix from each subunit is domain-swapped with the first and second transmembrane segments from an adjacent subunit. Each of DAGK’s three active sites resembles a portico. The cornice of the portico appears to be the determinant of DAGK’s lipid substrate specificity and overhangs the site of phosphoryl transfer near the water-membrane interface. Mutations to cysteine that caused severe misfolding were located in or near the active site, indicating a high degree of overlap between sites responsible for folding and for catalysis.

Escherichia coli diacylglycerol kinase (DAGK) is encoded by the *dgkA* gene and catalyzes the direct phosphorylation of diacylglycerol (DAG) by Mg(II)-adenosine triphosphate (MgATP) to form phosphatidic acid as part of the membrane-derived oligosaccharide (MDO) cycle (1–3). In Gram-positive organisms, the *dgkA* homolog encodes an undecaprenol kinase, indicating a role in oligosaccharide assembly or in related signaling pathways (4). The DAGK homolog in *Streptococcus mutans* is known to be a virulence factor for smooth-

surface dental caries (5). DAGK was among the first integral membrane enzymes to be solubilized, purified, and mechanistically characterized (6). The wild-type protein is very stable (7, 8) and can spontaneously insert into lipid bilayers to adopt its functional fold (9, 10). Paradoxically, DAGK resembles many disease-linked human membrane proteins because it is highly susceptible to mutation-induced misfolding (10–12). DAGK functions as a 40-kD homotrimer, with a total of nine transmembrane (TM) helices and three active sites (13).

The structure of DAGK was determined using solution nuclear magnetic resonance (NMR) methods (14) under conditions in which the enzyme is a functional homotrimer solubilized in ~100-kD dodecylphosphocholine micelles; this work extends other solution NMR studies of >20-kD multispan membrane proteins (15–22). The backbone structure of the helical TM domain of DAGK (residues 26 to 121) was precisely determined by the data (Fig. 1, fig. S1, and table S4); however, motions associated with the N terminus (residues 1 to 25) have hindered determination of its conformation beyond confirming the presence of two stable amphipathic helices.

DAGK’s structure bears no resemblance to that of the water-soluble DAGK (23) (Fig. 1). The three-fold symmetry axis lies at the center of a parallel left-handed bundle formed by the second transmembrane (TM2) helices of the three subunits. TM2 has previously been proposed to play a central role in DAGK’s folding and stability (24, 25) and contains several highly conserved residues (Fig. 1), particularly near the

¹Department of Biochemistry and Center for Structural Biology, Vanderbilt University, Nashville, TN 37232, USA. ²Korea Polar Research Institute, Incheon 406-840, Korea. ³School of Life Science, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China. ⁴Otto Diels Institute for Organic Chemistry, Christian Albrechts University of Kiel, D-24098 Kiel, Germany.

*These authors contributed equally to this work.

†To whom correspondence should be addressed. E-mail: chuck.sanders@vanderbilt.edu