January 2011

Niche of harmful alga Aureococcus anophagefferens revealed through ecogenomics

Christopher Gobler
Stony Brook Universtiy

Dianna Berry
Stony Brook University

Sonya Dyhrman
Woods Hole Oceanographic Institution

Steven Wilhelm
University of Tennessee, Knoxville, wilhelm@utk.edu

Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics

Christopher J. Gobler[a,b,1], Dianna L. Berry[a,b,2], Sonya T. Dyhrman[c,2], Steven W. Wilhelm[d,2], Asaf Salamov[e], Alexei V. Lobanov[f], Yan Zhang[f], Jackie L. Collier[b], Louie L. Wurch[c], Adam B. Kustka[g], Brian D. Dill[h], Manesh Shah[i], Nathan C. VerBerkmoes[h], Alan Kuo[e], Astrid Terry[e], Jasmyn Pangilinan[e], Erika A. Lindquist[e], Susan Lucas[e], Ian T. Paulsen[j], Theresa K. Hattenrath-Lehmann[a,b], Stephanie C. Talmage[a,b], Elyse A. Walker[a,b], Florian Koch[a,b], Amanda M. Burson[a,b], Maria Alejandra Marcoval[a,b], Ying-Zhong Tang[a,b], Gary R. LeCleir[c], Kathryn J. Coyne[k], Gry M. Berg[l], Erin M. Bertrand[m], Mak A. Saito[m,n], Vadim N. Gladyshev[d], and Igor V. Grigoriev[e,1]

[a]School of Marine and Atmospheric Sciences, Stony Brook University, Southampton, NY 11968; [b]School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY 11794-5000; [c]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543; [d]Department of Microbiology, University of Tennessee, Knoxville, TN 37996; [e]US Department of Energy Joint Genome Institute, Walnut Creek, CA 94598; [f]Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115; [g]Department of Earth and Environmental Sciences, Rutgers University, Newark, NJ 07102; [h]Chemical Sciences and [i]Biosciences Divisions, Oak Ridge National Laboratory, Oak Ridge, TN 37830; [j]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney 2109, New South Wales, Australia; [k]College of Earth, Ocean, and Environment, University of Delaware, Lewes, DE 19958; [l]Department of Environmental Earth System Science, Stanford University, Stanford, CA 94305; [m]Massachusetts Institute of Technology and Woods Hole Oceanographic Institution Joint Program in Chemical Oceanography, Woods Hole, MA 02543; and [n]Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA 02543

Edited by David M. Karl, University of Hawaii, Honolulu, HI, and approved January 26, 2011 (received for review October 29, 2010)

Harmful algal blooms (HABs) cause significant economic and ecological damage worldwide. Despite considerable efforts, a comprehensive understanding of the factors that promote these blooms has been lacking, because the biochemical pathways that facilitate their dominance relative to other phytoplankton within specific environments have not been identified. Here, biogeochemical measurements showed that the harmful alga *Aureococcus anophagefferens* outcompeted co-occurring phytoplankton in estuaries with elevated levels of dissolved organic matter and turbidity and low levels of dissolved inorganic nitrogen. We subsequently sequenced the genome of *A. anophagefferens* and compared its gene complement with those of six competing phytoplankton species identified through metaproteomics. Using an ecogenomic approach, we specifically focused on gene sets that may facilitate dominance within the environmental conditions present during blooms. *A. anophagefferens* possesses a larger genome (56 Mbp) and has more genes involved in light harvesting, organic carbon and nitrogen use, and encoding selenium- and metal-requiring enzymes than competing phytoplankton. Genes for the synthesis of microbial deterrents likely permit the proliferation of this species, with reduced mortality losses during blooms. Collectively, these findings suggest that anthropogenic activities resulting in elevated levels of turbidity, organic matter, and metals have opened a niche within coastal ecosystems that ideally suits the unique genetic capacity of *A. anophagefferens* and thus, has facilitated the proliferation of this and potentially other HABs.

genomics | proteome | comparative genomics | eutrophication

Harmful algal blooms (HABs) are caused by phytoplankton that have a negative impact on ecosystems and coastal fisheries worldwide (1–4) and cost the US economy alone hundreds of millions of dollars annually (5). The frequency and impacts of HABs have intensified in recent decades, and anthropogenic processes, including eutrophication, have been implicated in this expansion (1–3). Although there is great interest in mitigating the occurrence of HABs, traditional approaches that have characterized biogeochemical conditions present during blooms do not identify the aspects of the environment that are favorable to an individual algal species. Predicting where, when, and under what environmental conditions HABs will occur has further been inhibited by a limited understanding of the cellular attributes that facilitate the proliferation of one phytoplankton species to the exclusion of others.

Aureococcus anophagefferens is a pelagophyte that causes harmful brown tide blooms with densities exceeding 10^6 cells mL^{-1} for extended periods in estuaries in the eastern United States and South Africa (6). Brown tides do not produce toxins that poison humans but have decimated multiple fisheries and seagrass beds because of toxicity to bivalves and extreme light attenuation, respectively (6). Brown tides are a prime example of the global expansion of HABs, because these blooms had never been documented before 1985 but have recurred in the United States and South Africa annually since that time (6). Like many other HABs, *A. anophagefferens* blooms in shallow, anthropogenically modified estuaries when levels of light and inorganic nutrients are low and organic carbon and nitrogen concentrations are elevated (1–3).

For this study, we used an ecogenomic approach to assess the extent to which the gene set of *A. anophagefferens* may permit its dominance under the environmental conditions present in estuaries during brown tides. We characterized the biogeochemical conditions present in estuaries before, during, and after *A. anophagefferens* blooms. Sequencing this HAB genome (*A. anophagefferens*), we compared its gene content to those of six phytoplankton species identified through metaproteomics to co-occur with this alga during blooms events. Using this ecogenomic approach, we investigated how the gene sets of *A. anophagefferens* differ from the six comparative phytoplankton species and how these differences may affect the ability of *A. anophagefferens* to compete in the physical (e.g., light harvesting), chemical (e.g., nutrients, organic matter, and trace metals), and ecological (e.g., defense against predators and allelopathy) environment present during brown tides.

Author contributions: C.J.G. and I.V.G. designed research; C.J.G., D.L.B., S.T,D., S.W.W., J.L.C., L.L.W., B.D.D., M.S., N.C.V., A.K., A.T., J.P., E.A.L., S.L., S.C.T., E.A.W., F.K., M.A.M., and I.V.G. performed research; I.T.P., G.M.B., and I.V.G. contributed new reagents/analytic tools; C.J.G., D.L.B., S.T.D., S.W.W., A.S., A.V.L., Y.Z., J.L.C., L.L.W., A.B.K., B.D.D., M.S., N.C.V., A.K., A.T., T.K.H.-L., A.M.B., Y.-Z.T., G.R.L., K.J.C., E.M.B., V.N.G., and I.V.G. analyzed data; and C.J.G., S.T.D., S.W.W., J.L.C., A.B.K., M.A.S., V.N.G., and I.V.G. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this paper has been deposited in the GenBank database (accession no. ACJI00000000).

[1]To whom correspondence may be addressed. E-mail: christopher.gobler@stonybrook.edu or IVGrigoriev@lbl.gov.

[2]D.B., S.D., and S.W. contributed equally to this work.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1016106108/-/DCSupplemental.

Results and Discussion

During an investigation of a US estuary, Quantuck Bay, NY, from 2007 to 2009, brown tides occurred annually from May to July, achieving abundances exceeding 10^6 cells mL^{-1} or 5×10^6 μm^3 mL^{-1} (Fig. 1). *A. anophagefferens* was observed to bloom after spring diatom blooms and outcompeted small (<2 μm) eukaryotic and prokaryotic phytoplankton (e.g., *Ostreococcus* and *Synechococcus*) during summer months (Fig. 1*D*), a pattern consistent with prior observations (7, 8). Concurrently, dissolved inorganic nitrogen levels were reduced to <1 μM during blooms, whereas dissolved organic nitrogen levels and light extinction were elevated, resulting in a system with decreased light availability and concentrations of dissolved organic nitrogen far exceeding those of dissolved inorganic nitrogen (Fig. 1*C*). Metaproteomic analyses of planktonic communities were performed to identify phytoplankton that *A. anophagefferens* may compete with during blooms by quantifying organism-specific peptides among the microbial community. Performing such analyses on the plankton present in this estuary highlights the dominance of *A. anophagefferens* and coexistence of the six phytoplankton species for which complete genome sequences have been generated (Fig. 1*E*): two coastal diatom species, *Phaeodactylum tricornutum* (clone CCMP632) (9) and *Thalassiosira pseudonana* [clone CCMP 1335 (10) isolated from an embayment that now hosts brown tides (6)], and coastal zone isolates of *Ostreococcus* (*O. lucimarinus* and *O. tauri*) (11) and *Synechococcus* [clones CC9311 (12) and CC9902] small eukaryotic and prokaryotic phytoplankton, respectively, (Fig. 1 and Table 1). To assess the extent to which the gene set of *A. anophagefferens* may permit its dominance within the geochemical environment found in this estuary (Fig. 1*C*), the gene complement of *A. anophagefferens* was determined by genome sequencing and was compared with those of the six competing phytoplankton species (Fig. 1*E* and Table 1).

Although phytoplankton genome size generally scales with cell size (15, 16), *A. anophagefferens* (2 μm) has a larger genome (56 Mbp) and more genes (~11,500) than the six competing phytoplankton species (2.2–32 Mbp and 2,301–11,242 genes) (Table 1 and *SI Appendix*, Tables S1, S2, S3, and S4). Its small cell size and thus larger surface area to volume ratio allows it to kinetically outcompete larger phytoplankton for low levels of light and nutrients (17), whereas its large gene content and more complex

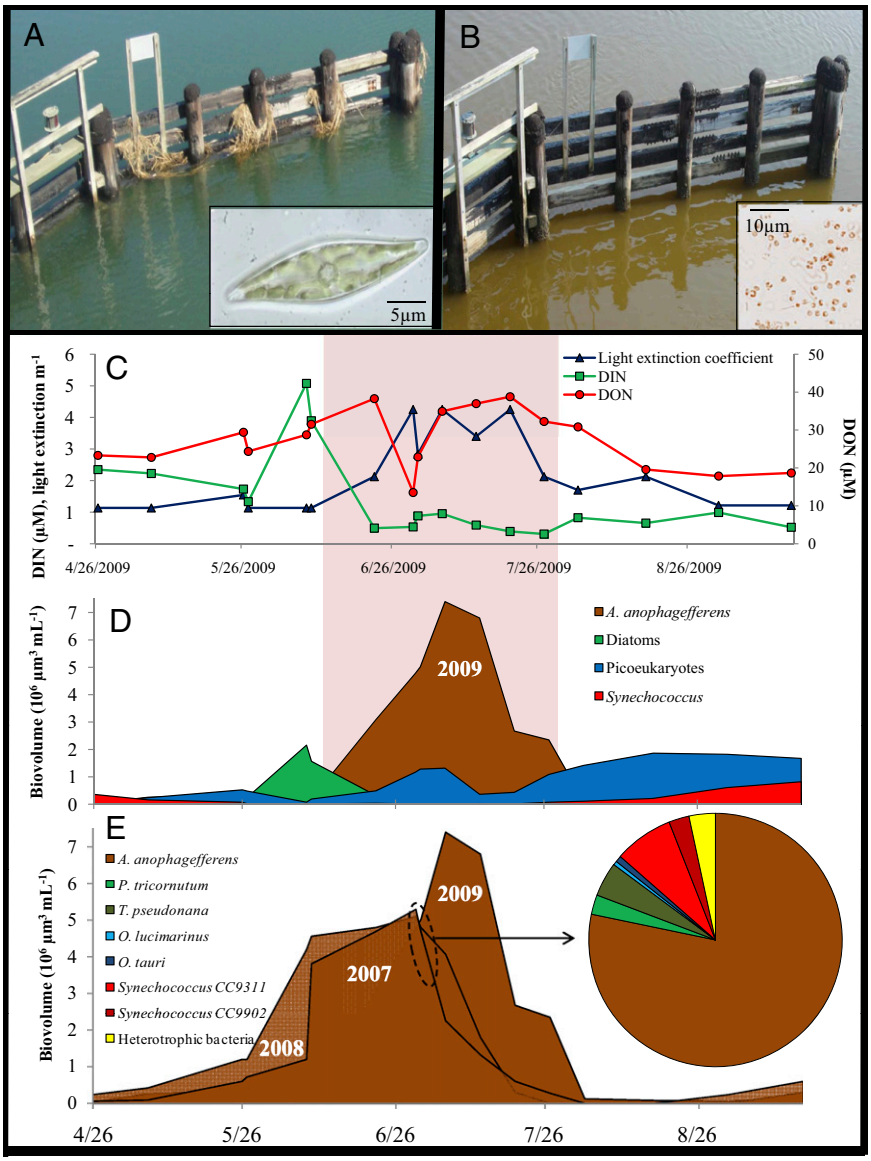


Fig. 1. Field observations from Quantuck Bay, NY. (*A*) Macro- and microscopic images (*Inset*) of an estuary (Quantuck Bay, NY) under normal conditions on June 9, 2009 before a brown tide (note the diatom in *Inset* micrograph image). (*B*) Similar macro- and microscopic images (*Inset*) taken July 6, 2009 during a harmful brown tide bloom caused by *A. anophagefferens* (note the dominance of *A. anophagefferens* in *Inset* micrograph). (*C*) The dynamics of dissolved inorganic nitrogen (DIN) and dissolved organic nitrogen (DON) and the extinction coefficient of light within seawater during the spring and summer of 2009 in Quantuck Bay. (*D*) The dynamics of phytoplankton during the spring and summer of 2009, a year when *A. anophagefferens* bloomed almost to the exclusion of other phytoplankton, including picoeukaryotes, which are often dominated by *Ostreococcus* sp. in estuaries that host brown tides (6–8), and *Thalassiosira* and *Phaeodactylum*, genera that are found in this system (6). The shaded regions in *C* and *D* indicate the period when *A. anophagefferens* blooms, highlighting that *A. anophagefferens* blooms when levels of DIN and light levels are low and DON levels are high and also highlighting that *A. anophagefferens* blooms can persist for more than 1 mo during the summer when this species dominates phytoplankton biomass inventories. (*E*) The dynamics of *A. anophagefferens* cell densities during 2007, 2008, and 2009, with the dates of samples collected for metaproteome analyses (June 26, 2007 and July 9, 2007) indicated within the dashed circled. *Inset* metaproteome pie chart specifically depicts the mean relative abundance of unique spectral counts of peptides matching proteins from *A. anophagefferens*, *P. tricornutum* (9), *T. pseudonana* (10), *O. tauri* (11), *O. lucimarinus* (11), *Synechococcus* (CC9311) (12), *Synechococcus* (CC9902), and heterotrophic bacteria.

genetic repertoire may provide a competitive advantage over other small phytoplankton with fewer genes. The *A. anophagefferens* genome contains the largest number of unique genes relative to the six competing phytoplankton examined here (209 vs. 12–79 unique genes) (Table 1). Many of these unique or enriched genes in *A. anophagefferens* are associated with light harvesting, organic matter use, and metalloenzymes as well as the synthesis of microbial predation and competition deterrents (*SI Appendix*, Tables S5, S6, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, and S17). These enriched and unique gene sets are involved in biochemical pathways related to the environmental conditions prevailing during brown tides (Fig. 1) and thus, are likely to facilitate the dominance of this alga during chronic blooms that plague estuarine waters.

Light Harvesting. Phytoplankton rely on light to photosynthetically fix carbon dioxide into organic carbon, but the turbid, low-light environment characteristic of estuaries and intense shading during dense algal blooms (Fig. 1 *B* and *C*) can strongly limit photosynthesis. *A. anophagefferens* is better adapted to low light than the comparative phytoplankton species, which requires at least threefold higher light levels to achieve maximal growth rates (Fig. 2*A*). Its genome contains the full suite of genes involved in photosynthesis, including 62 genes encoding light-harvesting complex (LHC) proteins (Fig. 2*A*). This is 1.5–3 times more than other eukaryotic phytoplankton sequenced thus far (Fig. 2*A* and *SI Appendix*, Table S7) and a feature that likely enhances adaptation to low and/or dynamic light conditions found in turbid estuaries. LHC proteins bind antenna chlorophyll and carotenoid pigments that augment the light-capturing capacity of the photosynthetic reaction centers (18, 19). Twenty-six *A. anophagefferens* LHC genes belong to a group that has only six representatives in *T. pseudonana* and one representative in *P. tricornutum* (branch PHYMKG in Fig. 3 and *SI Appendix*, Fig. S1) but are similar to the multicellular brown macroalgae, *Ectocarpus siliculosus* (20). Similar LHC genes in the microalgae *Emiliania huxleyi* have recently been shown to be up-regulated under low light (21). We hypothesize that these LHC genes encode the major light-harvesting proteins for *A. anophagefferens* and that the enrichment of these proteins imparts a competitive advantage in acquiring light under the low-irradiance conditions that prevail during blooms (Fig. 1*C*).

Organic Matter Use. In addition to being well-adapted to low light, *A. anophagefferens* also outcompetes other phytoplankton in estuaries with elevated organic matter concentrations (6) (Fig. 1*C*), and can survive extended periods with no light (22). Consistent with these observations, the genome of *A. anophagefferens* contains a large number of genes that may permit the degradation of organic compounds to support heterotrophic metabolism. For example, its genome encodes proteins involved in the transport of oligosaccharides and sugars that are not found in competing phytoplankton, including genes for glycerol, glucose, and D-xylose uptake (*SI Appendix*, Table S8). The *A. anophagefferens* genome also encodes more nucleoside sugar transporters and major facilitator family sugar transporters than other comparative phytoplankton species (*SI Appendix*, Table S8). It is highly enriched in genes associated with the degradation of mono-, di-, oligo-, and polysaccharides as well as sulfonated polysaccharides. *A. anophagefferens* possesses 47 sufatase genes, including those targeting sulfonated polysaccharides such as glucosamine-(*N*-acetyl)-6-sulfatases, whereas the diatoms contain a total of three to four sulfatases and the comparative picoplankton contain none (*SI Appendix*, Table S9). *A. anophagefferens* also possesses many more genes involved in carbohydrate degradation than competing phytoplankton (85 vs. 4–29 genes in comparative phytoplankton), including 29 such genes present only in *A. anophagefferens* (Fig. 4 and *SI Appendix*, Tables S10 and S11). Collectively, these genes (*SI Appendix*, Tables S9, S10, S11, and S12) provide this alga with unique metabolic capabilities regarding the degradation of an array of organic carbon compounds, many of which may not be accessible to other phytoplankton. In an ecosystem setting, such a supplement of organic carbon would be critical for population proliferation within the low-light environments present in estuaries, particularly during dense algal blooms (Fig. 1*C*).

A. anophagefferens, like many HABs, blooms when inorganic nitrogen levels are low but organic nitrogen levels are elevated (Fig. 1*C*) (1–3), and *A. anophagefferens* is known to efficiently metabolize organic compounds for nitrogenous nutrition (6, 23). Notably, this niche strategy is reflected within the *A. anophagefferens* genome, which encodes transporters specific for a diverse set of organic nitrogen compounds including urea, amino acids, purines, nucleotide sugars, nucleosides, peptides, and oligopeptides (*SI Appendix*, Table S8) (24). Relative to competing phytoplankton, *A. anophagefferens* is enriched in genes encoding enzymes that degrade organic nitrogen compounds, such as nitriles, asparagine, and urea (Fig. 2*B*). *A. anophagefferens* is also the only species among the phytoplankton genomes examined that possesses a membrane-bound dipeptidase, several histidine ammonia lyases, tripeptidyl peptidase, and several other enzymes (*SI Appendix*, Table S13) that could collectively play a role in metabolizing organic nitrogen compounds that are not bioavailable to other phytoplankton. Furthermore, the *A. anophagefferens* genome also contains enzymes that degrade amino acids, peptides, proteins, amides, amides, and nucleotides, often possessing more copies of these genes than competing phytoplankton (*SI Appendix*, Table S13). This characteristic, along with its unique gene set, may provide *A. anophagefferens* with a greater capacity to use organic compounds for nitrogenous nutrition compared with its com-

Table 1. Major features of the genomes of *A. anophagefferens* and six competing algal species: *P. tricornutum* (9), *T. pseudonana* (10), *O. tauri* (11), *O. lucimarinus* (11), *Synechococcus* (CC9311) (12), and *Synechococcus* (CC9902)

	A. anophagefferens	*P. tricornutum*	*T. pseudonana*	*O. tauri*	*O. lucimarinus*	*Synechococcus* (CC9311)	*Synechococcus* (CC9902)
Cell diameter (μm)	2.0	11.0	5.0	1.2	1.3	1.0	1.0
Cell volume (μm^3)	6	61	88	1.8	2.0	1.2	1.2
Genome size (Mbp)	57	27	32	13	13	2.6	2.2
Predicted gene number	11,501	10,402	11,242	7,892	7,651	2,892	2,301
Genes with known functions	8,560	6,239	6,797	5,090	5,322	1,607	1,469
Genes with Pfam domains	6,908	5,398	5,791	4,763	4,214	1,636	1,488
Genes with unique Pfam domains	209	79	75	23	51	55	12

Genes with known functions were identified using Swiss-Prot, a curated protein sequence database, with an *e*-value cutoff of $<10^{-5}$ (13). Pfam domains are sequences identified from a database of protein families represented by multiple sequence alignments and hidden Markov models (14). The compressed nature of *P. tricornutum* cells (11 × 2.5 μm) makes its biovolume smaller than *T. pseudonana*.

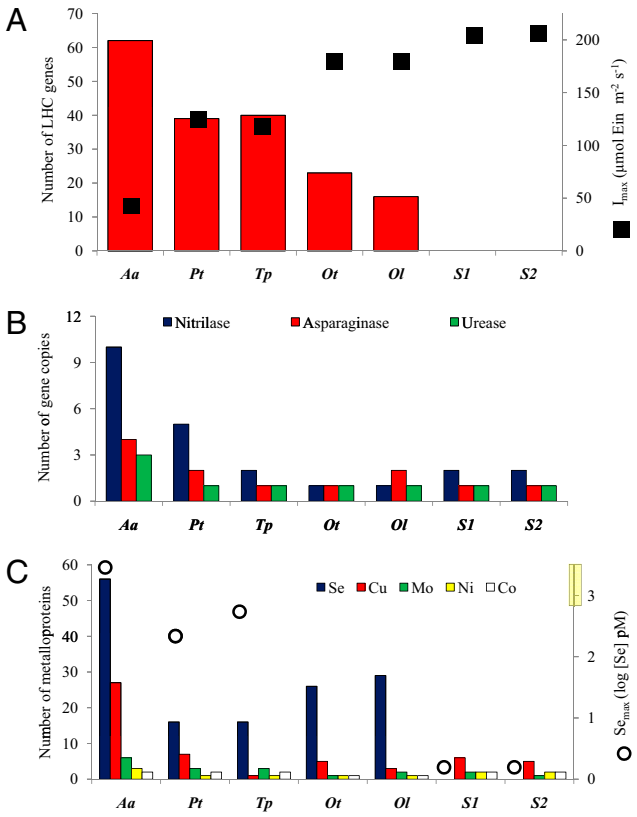


Fig. 2. Comparisons of gene compliment between *A. anophagefferens* and other co-occurring phytoplankton species. Aa, *A. anophagefferens*; Pt, *P. tricornutum*; Tp, *T. pseudonana*; Ot, *O. tauri*; Ol, *O. lucimarinus*; S1, *Synechococcus* clone CC9311; S2, *Synechococcus* clone CC9902. (*A*) The number of light-harvesting complex (LHC) genes present in each phytoplankton genome (red bars; left axis) and I_{max}, the irradiance level required to achieve maximal growth rates in each phytoplankton (black squares; right axis) are shown. Among these species, *A. anophagefferens* possesses the greatest number of LHC genes, achieves a maximal growth rate at the lowest level of light, and blooms when light levels are low. (*B*) The number of genes associated with the degradation of nitriles, asparagine, and urea in each phytoplankton genome. *A. anophagefferens* grows efficiently on organic nitrogen and possesses more nitrilase, asparaginase, and urease genes than other phytoplankton. (*C*) Interspecies comparison of the genes encoding proteins that contain the metals Se, Cu, Mo, Ni, and Co (left axis) and Se_{max}, the selenium level (added as selenite shown as log concentrations) required to achieve maximal growth rates in *A. anophagefferens*, *P. tricornutum*, *T. pseudonana*, and *Synechococcus* (white circles; right axis). The range of dissolved selenium concentrations found in estuaries is depicted as a yellow bar on the right *y* axis. *A. anophagefferens* has the largest number of proteins containing Se, Cu, Mo, and Ni and blooms exclusively in shallow estuaries where inventories of these metals are high. *SI Appendix* contains details of irradiance- and Se-dependent growth data and Se concentrations in estuaries.

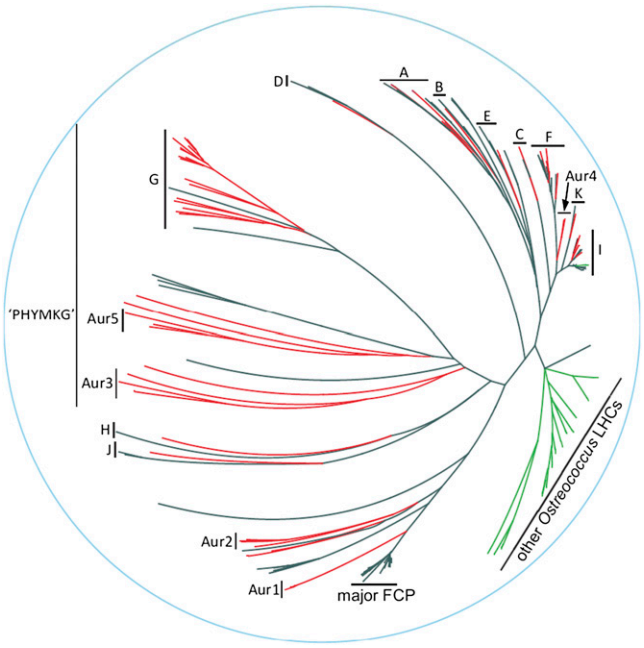


Fig. 3. Phylogenetic tree constructed from amino acid sequences of predicted LHC proteins from two diatoms (*P. tricornutum* and *T. pseudonana*; black branches), two *Ostreococcus* species (*O. tauri* and *O. lucimarinus*; green branches), and *A. anophagefferens* (red branches). The tree constructed in MEGA4 (*SI Appendix*, Fig. S1) is displayed here after manipulation of the original branch lengths in Hypertree (http://kinase.com/tools/HyperTree.html) to aid visualization of major features of the tree. None of the *Aureococcus* LHCs were closely related to green plastid lineage LHCs, although four belonged to a group found in both the green and red plastid lineages (group I). None of the *Aureococcus* LHCs clustered with the major fucoxanthin-chlorophyll binding proteins (FCP) of diatoms and other heterokonts (major FCP group). However, many *Aureococcus* LHCs did group with similar sequences from *P. tricornutum* and *T. pseudonana* (as well as LHCs from other red-lineage algae not included in this tree; groups A–K). There were also five groups of *A. anophagefferens* LHCs that were not closely related to any other LHCs (Aur1 to Aur5). Group G includes 16 LHCs from *A. anophagefferens* and 2 LHCs from *T. pseudonana*, and it shares a unique PHYMKG motif near the end of helix two, with 10 additional *A. anophagefferens* LHCs plus 5 more from the diatoms. Cyanobacteria such as *Synechococcus* do not possess LHC proteins.

petitors, a hypothesis supported by its dominance in systems with elevated ratios of dissolved organic nitrogen to dissolved inorganic nitrogen and the reduction in dissolved organic nitrogen concentrations often observed during the initiation of brown tides (6, 25).

Metalloenzymes. *A. anophagefferens* blooms in shallow, enclosed estuaries (6) where the concentrations of metals and elements like selenium are elevated (26–28), but it never dominates deep estuaries or continental shelf regions (6) that are characterized by lower metal and trace element inventories (26–28). *A. anophagefferens* has a large and absolute requirement for some trace elements, such as selenium (Fig. 2*C*). In comparison, phytoplankton, such as *Synechococcus*, do not require this element, whereas others, such as *T. pseudonana* and *P. tricornutum*, have lower selenium requirements for maximal growth (Fig. 2*C*). The *A. anophagefferens* genome is consistent with these observations, being enriched in numerous classes of proteins that require metals and elements like selenium as cofactors (Fig. 2*C*). It possesses at least 56 genes encoding selenocysteine-containing proteins, two times the number present in the *O. lucimarinus* genome, which previously had the largest known eukaryotic selenoproteome (11, 29), and fourfold more than the diatom genomes (Fig. 2*C*). The *A. anophagefferens* selenoproteome includes nearly all known eukaryotic selenoproteins as well as selenoproteins that were previously described only in bacteria (29) and several selenoproteins that have never been described in any other organism (*SI Appendix*, Table S14). In addition, several selenoprotein families are represented by multiple isozymes (*SI Appendix*, Table S14). One-half of the selenoproteins are methionine sulfoxide reductases, thioredoxin reductases, glutathione peroxidases, glutaredoxins, and peroxiredoxins (*SI Appendix*, Table S14). Together, these enzymes help protect cells against oxidative stress in the dynamic and ephemeral conditions present in estuaries through the removal of hydroperoxides and the repair of oxidatively damaged proteins. Moreover, selenocysteine residues are often superior catalytic groups compared with cysteine (30–32), and thus, they allow *A. anopha-*

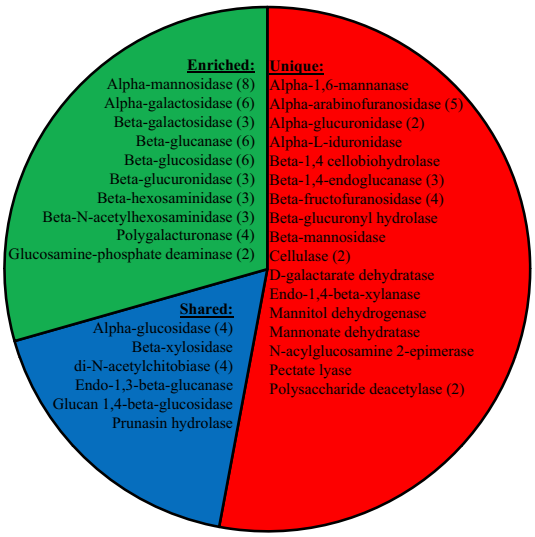


Fig. 4. Genes encoding for enzymes involved in degrading organic carbon compounds in *A. anophagefferens*. The graph displays the portion and names of the genes encoding for functions that are unique to *A. anophagefferens* (red; 53%), enriched in *A. anophagefferens* relative to the six comparative phytoplankton (34%; green), and present at equal or lower numbers in *A. anophagefferens* relative to the six comparative phytoplankton (13%; blue). The number of genes present in multiple copies in *A. anophagefferens* is shown in parentheses. Further details regarding these genes are presented in *SI Appendix*, Tables S10 and S11.

gefferens to more efficiently execute multiple metabolic processes and increase its competitiveness relative to other phytoplankton in the anthropogenically modified estuaries where it blooms.

The *A. anophagefferens* genome is also enriched in genes encoding for molybdenum-, copper-, and nickel-containing enzymes (Fig. 2*C*). For example, the *A. anophagefferens* genome includes two times the number of genes encoding molybdenum-containing oxidases found in competing species (6 vs. 1–3 genes) (Fig. 2*C* and *SI Appendix*, Tables S15 and S16) and has the largest number of molybdenum-specific transporters (*SI Appendix*, Table S8). Similarly, *A. anophagefferens* possesses four times more genes that encode copper-containing proteins than its competitors (27 vs. 1–6 genes) (Fig. 2*C*), including 5 multicopper oxidases and 20 tyrosinase-like proteins (*SI Appendix*, Tables S15 and S16). Several of the *A. anophagefferens* tyrosinase and multicopper oxidase family proteins are heavily glycosylated (more than four glycosylation sites) (*SI Appendix*, Table S16) and thus, are likely secretory proteins, whereas the few present in the other comparative algal species are not. These copper-containing enzymes degrade lignin, catalyze the oxidation of phenolics, and can have antimicrobial properties (33, 34) and thus, may provide nutrition or confer protection to *A. anophagefferens* cells. *A. anophagefferens* is also the only phytoplankton species with a homolog of the CutC copper homeostasis protein, which permits efficient cellular trafficking of this metal (*SI Appendix*, Table S8). With three nickel-requiring ureases, *A. anophagefferens* has more nickel-containing enzymes than other comparative phytoplankton (Fig. 2 *B* and *C*). Consistent with its ecogenomic profile, these ureases allow *A. anophagefferens* to meet its daily N demand from urea, whereas other phytoplankton do not (35). Perhaps to support the synthesis and use of ureases, *A. anophagefferens* is the only comparative phytoplankton species with a high-affinity nickel transporter (HoxN) (36). *A. anophagefferens* is not universally enriched in metalloenzymes, because other phytoplankton contain equal numbers of cobalt-containing enzymes (Fig. 2*C*). However, the formation of blooms exclusively in shallow estuaries ensures that *A. anophagefferens* has access to a rich supply of the selenium, copper, and nickel required to synthesize these ecologically important and catalytically superior enzymes (30, 31, 37).

Microbial Defense. Although genes associated with the adaptation to low light, the use of organic matter, and metals permit *A. anophagefferens* to dominate a specific geochemical niche found within estuaries, genes involved in the production of compounds that inhibit predators and competitors may further promote blooms (2). Although specific toxins have yet to be identified in *A. anophagefferens*, it is grazed at a low rate during blooms (2, 6), and its genome contains two to seven times more genes involved in the synthesis of secondary metabolites than the comparative phytoplankton genomes (*SI Appendix*, Fig. S2). *A. anophagefferens* also possesses a series of genes involved in the synthesis of putative antimicrobial compounds that are largely absent from the competing phytoplankton species (*SI Appendix*, Table S17). For example, *A. anophagefferens* has five berberine bridge enzymes involved in the synthesis of toxic isoquinoline alkaloids (38, 39) (*SI Appendix*, Table S17). *A. anophagefferens* uniquely possesses a membrane attack complex gene and multiple phenazine biosynthetase genes (*SI Appendix*, Table S17) that encode enzymes that may provide defense against microbes and/or protistan grazers (40, 41). There are two- to fourfold more ATP-binding cassette (ABC) transporters in *A. anophagefferens* compared with competing species (112 vs. 30–54 ABC transporters) (*SI Appendix*, Table S8), and it is specifically enriched in ABC multidrug efflux pumps that protect cells from toxic xenobiotics and endogenous metabolites (42, 43). Finally, the *A. anophagefferens* genome encodes 16-fold more Sel-1 genes (130 vs. 0–8 genes) (Table S6), 4-fold more ion channels (82 vs. 1–19 ion channels) (*SI Appendix*, Table S8), 4-fold more protein kinases, and 2-fold more WD40 domain genes than other phytoplankton (*SI Appendix*, Table S6). These genes may collectively mediate elaborate cell signaling and sensing by dense bloom populations (44–46), processes that would be important for detecting competitors, predators, other *A. anophagefferens* cells, and the environment. Together, genes involved in the synthesis of microbial deterrents, export of toxic compounds, and cell signaling may contribute to the proliferation of this species with reduced population losses and thus, assist in promoting these HABs (2).

Conclusions. The global expansion of human populations along coastlines has led to a progressive enrichment in turbidity (47), organic matter, including organic nitrogen (1, 47, 48), and metals (26, 28) in estuaries. Matching the expansion of HAB events around the world in recent decades, *A. anophagefferens* blooms were an unknown phenomenon before 1985 but have since become chronic, annual events in US and South African estuaries (6), with the potential for further expansion. The unique gene complement of *A. anophagefferens* encodes a disproportionately greater number of proteins involved in light harvesting and organic matter use as well as metal and selenium-requiring enzymes relative to competing phytoplankton. Collectively, these genes reveal a niche characterized by conditions (low light, high organic matter, and elevated metal levels) that have become increasingly prevalent in anthropogenically modified estuaries, suggesting that human activities have enabled the proliferation of these HABs. In estuaries that host *A. anophagefferens* blooms, anthropogenic nutrient loading promotes algal growth and as a result, elevated levels of organic matter and turbidity (6), whereas high concentrations of metals have been attributed to maritime paints and some fertilizers (27, 49). Collectively, these findings establish a context within which to prevent and control HABs, specifically by ameliorating anthropogenically altered aspects of marine environments that harmful phytoplankton are genomically predisposed to exploit. Like *A. anophagefferens*, many HAB-forming dinoflagellates are known to exploit organic

forms of carbon and nitrogen for growth (1–4), grow well under low light (50), and have elevated requirements of copper, molybdenum, and selenium (51, 52). Continued ecogenomic analyses of HABs will reveal the extent to which these events can be attributed to human activities that have transformed coastal ecosystems to suit the genetic capacity of these algae.

Materials and Methods

The environmental conditions and plankton community composition within a brown tide-prone estuary (Quantuck Bay, NY) were monitored biweekly from spring to fall of 2007, 2008, and 2009. Nutrient levels were assessed by wet chemical and combustion techniques, whereas the composition of the plankton community was assessed by immunofluorescent assays, flow cytometry, and standard microscopy. Metaproteomes were generated using 2D nano-liquid chromatography tandem MS (LC-MS/MS), and spectra were analyzed using SEQUEST and DTASelect algorithms. The genome of *A. anophagefferens* was sequenced using the whole-genome shotgun approach using the Sanger platform assembled, with the JAZZ assembler, and annotated using JGI Annotation tools. Complete information regarding all methods used for all analyses reported here is available in *SI Appendix*.

ACKNOWLEDGMENTS. Assembly and annotations of *A. anophagefferens* are available from JGI Genome Portal at http://www.jgi.doe.gov/Aureococcus. Genome sequencing, annotation, and analysis were conducted by the US Department of Energy Joint Genome Institute supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. Efforts were also supported by National Oceanic and Atmospheric Administration Sea Grant Awards NA07OAR4170010 and NA10OAR4170064 to Stony Brook University via New York Sea Grant, National Oceanic and Atmospheric Administration Center for Sponsored Coastal Ocean Research Award NA09NOS4780206 to Woods Hole Oceanographic Institution, National Institutes of Health Grant GM061603 to Harvard University, and National Science Foundation Award IOS-0841918 to University of Tennessee.

1. Heisler J, et al. (2008) Eutrophication and harmful algal blooms: A scientific consensus. *Harmful Algae* 8:3–13.
2. Sunda WG, Graneli E, Gobler CJ (2006) Positive feedback and the development and persistence of ecosystem disruptive algal blooms. *J Phycol* 42:963–974.
3. Anderson DM, et al. (2008) Harmful algal blooms and eutrophication: Examining linkages from selected coastal regions of the United States. *Harmful Algae* 8:39–53.
4. Smayda TJ (1997) Harmful algal blooms: Their ecophysiology and general relevance to phytoplankton blooms in the sea. *Limnol Oceanogr* 42:1137–1153.
5. Hoagland P, Scatasta S (2006) *Ecology of Harmful Algae*, eds Graneli E, Turner J (Springer, Berlin), pp 391–402.
6. Gobler CJ, Lonsdale DJ, Boyer GL (2005) A synthesis and review of causes and impact of harmful brown tide blooms caused by the alga, *Aureococcus anophagefferens*. *Estuaries* 28:726–749.
7. O'Kelly CJ, Sieracki ME, Their EC, Hobson IC (2003) A transient bloom of *Ostreococcus* (Chlorophyta, Prasinophyceae) in West Neck Bay, Long Island, New York. *J Phycol* 39: 850–854.
8. Sieracki ME, Gobler CJ, Cucci T, Thier E, Hobson I (2004) Pico- and nanoplankton dynamics during bloom initiation of Aureococcus in a Long Island, NY bay. *Harmful Algae* 3:459–470.
9. Bowler C, et al. (2008) The *Phaeodactylum* genome reveals the evolutionary history of diatom *genomes*. *Nature* 456:239–244.
10. Armbrust EV, et al. (2004) The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. *Science* 306:79–86.
11. Palenik B, et al. (2007) The tiny eukaryote *Ostreococcus* provides genomic insights into the paradox of plankton speciation. *Proc Natl Acad Sci USA* 104:7705–7710.
12. Palenik B, et al. (2006) Genome sequence of Synechococcus CC9311: Insights into adaptation to a coastal environment. *Proc Natl Acad Sci USA* 103:13555–13559.
13. Boeckmann B, et al. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 31:365–370.
14. Finn RD, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38:D211–D222.
15. Connolly JA, et al. (2008) Correlated evolution of genome size and cell volume in diatoms (Bacillariophyceae). *J Phycol* 44:124–131.
16. Hessen DO, Jeyasingh PD, Neiman M, Weider LJ (2010) Genome streamlining and the elemental costs of growth. *Trends Ecol Evol* 25:75–80.
17. Raven JA, Kubler JE (2002) New light on the scaling of metabolic rate with the size of algae. *J Phycol* 38:11–16.
18. Green BR, Durnford DG (1996) The chlorophyll-carotenoid proteins of oxygenic photosynthesis. *Annu Rev Plant Physiol Plant Mol Biol* 47:685–714.
19. Durnford DG, et al. (1999) A phylogenetic assessment of the eukaryotic light-harvesting antenna proteins, with implications for plastid evolution. *J Mol Evol* 48:59–68.
20. Cock JM, et al. (2010) The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature* 465:617–621.
21. Lefebvre SC, et al. (2010) Characterization and expression analysis of the LHCf gene family in *Emiliania huxleyi* (Haptophyta) reveals differential responses to light and CO_2. *J Phycol* 46:123–134.
22. Popels LC, MacIntyre HL, Warner ME, Yaohong Z, Hutchins DA (2007) Physiological responses during dark survival and recovery in *Aureococcus anophagefferens* (Pelagophyceae). *J Phycol* 43:32–42.
23. Mulholland MR, Gobler CJ, Lee C (2002) Peptide hydrolysis, amino acid oxidation, and nitrogen uptake in communities seasonally dominated by *Aureococcus anophagefferens*. *Limnol Oceanogr* 47:1094–1108.
24. Wurch LL, Haley ST, Orchard ED, Gobler CJ, Dyhrman ST (2011) Nutrient-regulated transcriptional responses in the brown tide-forming alga *Aureococcus anophagefferens*. *Environ Microbiol* 13:468–481.
25. LaRoche J, et al. (1997) Brown tide blooms in Long Island's coastal waters linked to variability in groundwater flow. *Glob Change Biol Bioenergy* 3:397–410.
26. Sañudo-Wilhelmy SA, Flegal AR (1993) Comparable levels of trace-metal contamination in two semi-enclosed embayments: San Diego Bay and South San Fransiscio Bay. *Environ Sci Technol* 27:1934–1936.
27. Breuer E, Sañudo-Wilhelmy SA, Aller RC (1999) Distributions of trace metals and dissolved organic carbon in an estuary with restricted river flow and a brown tide. *Estuaries* 22:603–615.
28. Cutter GA, Cutter LS (2004) Selenium biogeochemistry in the San Francisco Bay estuary: Changes in water column behavior. *Estuar Coast Shelf Sci* 61:463–476.
29. Lobanov AV, et al. (2007) Evolutionary dynamics of eukaryotic selenoproteomes: Large selenoproteomes may associate with aquatic life and small with terrestrial life. *Genome Biol* 8:R198.
30. Stadtman TC (1996) Selenocysteine. *Annu Rev Biochem* 65:83–100.
31. Hatfield DL, Gladyshev VN (2002) How selenium has altered our understanding of the genetic code. *Mol Cell Biol* 22:3565–3576.
32. Kim HY, Gladyshev VN (2005) Different catalytic mechanisms in mammalian selenocysteine- and cysteine-containing methionine-R-sulfoxide reductases. *PLoS Biol* 3:e375.
33. Score AJ, Palfreyman JW, White NA (1997) Extracellular phenoloxidase and peroxidase enzyme production during interspecific fungal interactions. *Int Biodeterior Biodegradation* 39:225–233.
34. Mayer AM (2006) Polyphenol oxidases in plants and fungi: Going places? A review. *Phytochemistry* 67:2318–2331.
35. Fan C, Glibert PM, Alexander J, Lomas MW (2003) Characterization of urease activity in three marine phytoplankton species, *Aureococcus anophagefferens, Prorocentrum minimum*, and *Thalassiosira weissflogii*. *Mar Biol* 142:949–958.
36. Wolfram L, Friedrich B, Eitinger T (1995) The Alcaligenes eutrophus protein HoxN mediates nickel transport in Escherichia coli. *J Bacteriol* 177:1840–1843.
37. Messerschmidt A, Huber R, Wieghart K, Poulos T (2005) *Handbook of Metalloproteins* (Wiley, New York), Vols 1–3.
38. Facchini PJ (2001) Alkaloid biosynthesis in plants: Biochemistry, cell biology, molecular regulation, and metabolic engineering applications. *Annu Rev Plant Physiol Plant Mol Biol* 52:29–66.
39. Schmeller T, Latz-Brüning B, Wink M (1997) Biochemical activities of berberine, palmatine and sanguinarine mediating chemical defence against microorganisms and herbivores. *Phytochemistry* 44:257–266.
40. Rosado CJ, et al. (2007) A common fold mediates vertebrate defense and bacterial attack. *Science* 317:1548–1551.
41. Pierson LS, 3rd, Gaffney T, Lam S, Gong F (1995) Molecular analysis of genes encoding phenazine biosynthesis in the biological control bacterium. *Pseudomonas aureofaciens* 30–84. *FEMS Microbiol Lett* 134:299–307.
42. Sharom FJ (2008) ABC multidrug transporters: Structure, function and role in chemoresistance. *Pharmacogenomics* 9:105–127.
43. van Veen HW, Konings WN (1998) The ABC family of multidrug transporters in microorganisms. *Biochim Biophys Acta* 1365:31–36.
44. Mittl PRE, Schneider-Brachert W (2007) Sel1-like repeat proteins in signal transduction. *Cell Signal* 19:20–31.
45. Quarmby LM (1994) Signal transduction in the sexual life of *Chlamydomonas*. *Plant Mol Biol* 26:1271–1287.
46. Neer EJ, Schmidt CJ, Nambudripad R, Smith TF (1994) The ancient regulatory-protein family of WD-repeat proteins. *Nature* 371:297–300.
47. Lotze HK, et al. (2006) Depletion, degradation, and recovery potential of estuaries and coastal seas. *Science* 312:1806–1809.
48. Paerl HW, Pinckney JL, Fear JM, Peierls BL (1998) Ecosystem responses to internal and watershed organic matter loading: Consequences for hypoxia in the eutrophying Neuse river estuary, North Carolina, USA. *Mar Ecol Prog Ser* 166:17–25.
49. McBride MB, Spiers G (2001) Trace element content of selected fertilizers and dairy manures as determined by ICP-MS. *Commun Soil Sci Plant Anal* 32:139–156.
50. MacIntyre HL, et al. (2004) Mediation of benthic-pelagic coupling by microphytobenthos: An energy- and material-based model for initiation of blooms of *Aureococcus anophagefferens*. *Harmful Algae* 3:403–437.
51. Quigg A, et al. (2003) The evolutionary inheritance of elemental stoichiometry in marine phytoplankton. *Nature* 425:291–294.
52. Doblin MA, Blackburn SI, Hallegraeff GM (1999) Comparative study of selenium requirements of three phytoplankton species: *Gymnodinium catenatum, Alexandrium minutum* (Dinophyta) and *Chaetoceros cf. tenuissimus* (Bacillariophyta). *J Plankton Res* 21:1153–1169.

SUPPORTING INFORMATION

Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics;

Gobler, C.J. et al

MATERIALS AND METHODS

1. *A. anophagefferens* DNA isolation

Eight 2-liter flasks with one liter of *Aureococcus anophagefferens* CCMP1984 culture each were grown axenically to mid-exponential growth phase (10^7 cells ml^{-1}). The eight flasks were combined in a carboy and harvested with the Sharples continuous flow centrifuge. Cells were scraped from the mylar film, rinsed, and concentrated with a tabletop centrifuge to yield a wet weight of ~ 1 gram. The tube was temporarily stored in liquid N_2 vapors. After temporary storage, the bottom half of the tube was placed in a 68º C water bath and heated enough to free the pellet from the tube at which point it was placed in a mortar filled with liquid nitrogen. The pellet was ground using a pestle, and placed into 5 different 50 mL plastic centrifuge tubes, to which Cetyltrimethyl Ammonium Bromide (CTAB) at 68º C was added. Beta-mercaptoethanol (1% final volume) was added to each tube. Tubes were periodically mixed by partial inversion and were incubated for three hours at 68ºC. An equal volume of phenol:chloroform:isoamyl alcohol (25:24:1 ratio) was then added to the cell suspension/CTAB. The tubes were mixed gently by partial inversion and centrifuged in a tabletop centrifuge. The aqueous fraction was removed and an equal volume of chloroform:isoamyl alcohol (24:1) was added and mixed by partial inversion, and then centrifuged. After the second chloroform extraction, the aqueous phase was transferred to a new tube and 0.6 volume of cold (-20º C) isopropanol was added. The DNA spooled in the upper 1/3 or 1/2 of the tube, and a large amount of white flocculent

substance settled to the bottom half of the tube (likely carbohydrate). The DNA was drawn into a 5 mL pipette tip (cut back to increase the opening) and placed into a separate tube. Ultimately, the DNA was pooled all into one tube, and as much isopropanol as possible was removed with a 1000 µL pipette tip. The DNA was then dissolved in TE buffer, quantified with a Nanodrop-1000 UV-spectrophotometer, and run in an agarose gel with Lambda *Hin*dIII markers.

2. Genome sequencing and assembly

The genome of *Aureococcus anophagefferens* was sequenced using WGS strategy. Three libraries with an insert size of 2-3 KB, 6-8 KB, and 35-40 KB were used. The sequenced reads were screened for vector using cross_match, trimmed for vector and quality (1), and filtered to remove reads shorter than 100 bases, which resulted in the following dataset:

306,657 2-3 KB reads, containing 215 MB of sequence.

301,713 6-8 KB reads, containing 215 MB of sequence.

37,362 35-40 KB reads, containing 18 MB of sequence.

The data was assembled using release 2.10.6 of Jazz, a WGS assembler developed at the JGI (1). A word size of 13 was used for seeding alignments between reads. The unhashability threshold was set to 40, preventing words present in the data set in more than 40 copies from being used to seed alignments. A mismatch penalty of -30.0 was used, which will tend to assemble together sequences that are more than about 97% identical. After excluding redundant (<5Kb, with 80% of total length contained in scaffolds > 5 KB) and short (<1Kb) scaffolds from the initial assembly, there remained 59.6 MB of scaffold sequence, of which 6.0 MB (10.1%) was gaps. The filtered assembly contained 1,202 scaffolds, with a scaffold N/L50 of 13/1.3 MB,

and a contig N/L50 of 405/34 KB. The sequence depth derived from the assembly was 7.00 ± 0.13.

To estimate the completeness of the assembly, a set of 49,961 ESTs was BLAT-aligned to the unassembled trimmed data set, as well as the assembly itself. 48,963 ESTs (98.1%) were more than 80% covered by the unassembled data, 49,312 (98.7%) were more than 50% covered, and 49,500 (99.1%) were more than 20% covered. By way of comparison, 49,097 ESTs (98.3%) had mapped to the assembly.

3. EST sequencing:

Aureococcus anophagefferens CCMP 1984 was obtained from the Provasoli-Guillard Center for the Culture of Marine Phytoplankton (CCMP). The cultures were grown at 18°C on a 12 h:12 h light:dark cycle (45 µmol quanta m^{-2} s^{-1}) on f/2 medium with acetamide as the sole nitrogen source and harvested under nitrogen limitation according to Berg et al (2). *A. anophagefferens* poly A+ RNA was isolated from total RNA using the Absolutely mRNA Purification kit according to manufacturer's instructions (Stratagene, La Jolla, CA). cDNA synthesis and cloning was performed using a modified procedure based on the "SuperScript plasmid system with Gateway technology for cDNA synthesis and cloning" (Invitrogen). 1-2 µg of poly A+ RNA, reverse transcriptase SuperScript II (Invitrogen) and oligo dT-NotI primer (5'-GACTAGTTCTA GATCGCGAGCGGCCGCCCTTTTTTTTTTTTTTT -3') were used to synthesize first strand cDNA. Second strand synthesis was performed with *E. coli* DNA ligase, polymerase I, and RNaseH followed by end repair using T4 DNA polymerase. The SalI adaptor (5'- TCGACC CACGCGTCCG and 5'- CGGACGCGTGGG) was ligated to the cDNA, ligation products were digested with NotI (NEB), and subsequently size selected by gel electrophoresis

(1.1% agarose). Size ranges of cDNA were cut out of the gel (L: 600-1.2kb, M: 1.2kb-2kb, H: >2kb) and directionally ligated into the SalI and NotI digested vector pCMVsport6 (Invitrogen). The ligation was transformed into ElectroMAX T1 DH10B cells (Invitrogen).

Library quality was first assessed by randomly selecting 24 clones and PCR amplifying the cDNA inserts with the primers M13-F (GTAAAACGACGGCCAGT) and M13-R (AGGAAACAGCTATGACCAT). The number of clones without inserts was determined and 384 clones for each library were picked, inoculated into 384 well plates (Nunc) and grown for 18 hours at 37ºC. Each clone was amplified using RCA then the 5' and 3' ends of each insert was sequenced using vector-specific primers (FW: 5'- ATTTAGGTGACACTA TAGAA and RV 5' – TAATACGACTCACTATAGGG) and Big Dye chemistry (Applied Biosystems). The average read length and pass rate were 753 (Q20 bases) and 96% respectively.

Colonies from an *Aureococcus anophagefferens* cDNA library were plated onto agarose plates (254mm plates from Teknova, Hollister, CA) at a density of approximately 1000 colonies per plate. Plates were grown at 37ºC for 18 hours then individual colonies were picked and each used to inoculate a well containing LB medium with appropriate antibiotic in a 384 well plate (Nunc, Rochester, NY). 384 well plates were grown at 37ºC for 18 hours. Plasmid DNA for sequencing was produced by rolling circle amplification (Templiphi, GE Healthcare, Piscataway, NJ). Subclone inserts were sequenced from both ends using primers complimentary to the flanking vector sequence (Fwd: 5'- GTAAAACGACGGCCAGT, Rev: 5' - AGGAAACAGCTATGACCAT) and sequenced using Big Dye terminator chemistry on an ABI 3730 (ABI, Foster City, CA).

4. Genome Annotation

The genome assembly v1.0 of *Aureococcus anophagefferens* was annotated using the JGI annotation pipeline, which takes assembly scaffolds and ESTs as inputs to produce gene models and their annotations. It starts with masking assembly scaffolds using RepeatMasker (http://www.repeatmasker.org/) and a custom repeat library of 837 putative transposable element sequences. After masking, gene models were predicted using several methods: 1) putative full length genes derived from 16,280 cluster consensus sequences of over 50,000 clustered and assembled *A. anophagefferens* ESTs were mapped to genomic sequence, 2) homology-based gene models were predicted using FGENESH+ (3) and Genewise (4) seeded by Blastx alignments against sequences from NCBI non-redundant protein set, 3) *ab initio* gene predictor FGENESH (3) was trained on the set of putative full-length genes and reliable homology-based models. Genewise models were completed using scaffold data to find start and stop codons. ESTs and EST clusters were used to extend, verify, and complete the predicted gene models. Because multiple gene models were generated for each locus, a single representative model was chosen based on homology and EST support and used for further analysis.

All predicted gene models were annotated using InterProScan (5) and hardware-accelerated double-affine Smith-Waterman alignments (www.timelogic.com) against SwissProt (www.expasy.org/sprot) and other specialized databases like KEGG (6). Finally, KEGG hits were used to map EC numbers (http://www.expasy.org/enzyme/), and Interpro hits were used to map GO terms (7). In addition predicted proteins were annotated according to KOG (8) classification. In total 11,501 gene models were predicted with their characteristics summarized in tables S1 - S4. Predicted genes and annotations referred to in this work were then manually curated using web-based annotation tools at JGI Portal http://www.jgi.doe.gov/Aureococcus. We used Blastp at E-value threshold of 1e-05 to search for potential homologues of *Aureococcus*

gene family members in the complete genomes of the six comparative phytoplankton. Gene homologs were manually inspected to correct for possible annotation errors. Pfam domains were identified using TimeLogic implementation of HMMER package.

5. Phylogentic analysis of light harvesting complex genes and membrane transporters

Sixty-two light harvesting chlorophyll-binding protein (LHC) genes were identified in the *A. anophagefferens* genome by extensive BLAST of known LHC protein sequences from plants and other algae against predicted *A. anophagefferens* proteins. LHC genes were identified in the comparative genomes by the same process. The predicted protein sequences were initially aligned in BioEdit (9) with ClustalW after removing the N-terminal putative signal peptides as well as four sequences that disrupted the alignment because they were too short or too divergent (Thaps3:262313, Thaps3:38122, Thaps3:41655, and Ost99013:24978), and the alignments were improved manually following the protein-structure-based alignments of Eppard and Rhiel (10) and Eppard et al (11). The phylogenetic tree was constructed in MEGA4 (12) by the neighbor-joining method with pairwise deletion of gaps and distances calculated using the Poisson correction with rate variation among sites (gamma = 1). Comparisons of transporter proteins among groups were performed using TransportDB (http://www.membranetransport.org/) a comprehensive database resource of information on membrane transporters extensively described by Ren et al (13).

6. Searches for selenoprotein genes

To identify selenoprotein genes, the *Aureococcus* genome was analyzed with SECISearch (14), which searched for primary sequences and secondary structures and then calculated free

energy for various parts of the predicted. As in other organisms, selenocysteine (Sec) is inserted into *Aureococcus* selenoproteins with the help of Sec insertion sequence (SECIS) elements. The genome was analyzed with both default and loose patterns of SECISearch to accommodate identification of unusual SECIS structures, and the searches were extended to search for organism-specific structures by a modified SECISearch (14). After candidate SECIS elements were identified, ORFs were predicted in the regions upstream of the SECIS elements. An additional requirement was the presence of at least one homologous protein in the NCBI non-redundant database. The final step was a manual sequence and homology analysis of predicted selenoprotein ORFs located upstream of candidate SECIS elements.

Separately, the genome was analyzed with TBLAST against all known selenoprotein sequences to identify homologs of previously described selenoprotein genes. A third procedure was the use of an approach that searched for selenocysteine/cysteine pairs in homologous sequences (14). ORFs with in-frame UGA codons were extracted that satisfied these criteria: (i) conservation of selenocysteine-flanking regions; and (ii) occurrence of homologs containing cysteine in place of selenocysteine. We used TBLASTX to examine all potential ORFs with in-frame UGA codons against NCBI non-redundant protein database. All hits were then tested for the occurrence of SECIS elements with SECISearch. PSI-BLAST was used for the identification of more distant homologs. The datasets resulting from the three independent methods of selenoprotein identification were combined and the proteins classified as homologs of previously known selenoproteins, novel selenoproteins and candidate selenoprotein genes.

7. Identification of metalloproteins for copper, molybdenum, nickel and cobalt (in the form of vitamin B_{12})

For each metal, we used representative sequences of all known metal-dependent proteins (i.e., strictly metal-binding proteins) to search for homologous sequences in *Aureococcus anophagefferens* and other selected organisms via TBLASTN (15) with an e-value <0.1. We excluded proteins that bind alternative metals in different organisms. Considering that some proteins contain both metal-dependent and metal-independent subunits, we only used metal-binding-domain-containing proteins/subunits as metalloproteins. Distant homologs were further identified using iterative BLAST searches with default parameters. Orthologous proteins were defined using the COG database and bidirectional best hits (16, 17). Additional analyses, such as conservation of metal-binding ligands and phylogenetic analysis, were also used to help identify orthologs from numerous homologs. Because iron- and zinc-containing enzymes are significantly more diverse in function and metal binding modes (i.e. metal substitutions, variable binding) they were not targeted during our analyses.

8. Generation of metaproteomes from estuaries with *A. anophagefferens* blooms

Environmental samples were collected 6/26/2007 and 7/9/2007 from Quantuck Bay, NY, USA, a site of frequent brown tides (Fig 1 main body of text). Field samples were harvested via centrifugation at 2,000 *g* for 30 min. Cell pellets were lysed and proteins denatured in a solution of 6 M guanidine and 10 mM DTT in 50 mM Tris buffer (pH 7.6), with bead beating (0.l mm zirconia/silica beads, 2 min, 30s on/off, 20Hz) followed by 1h at 60^{o} C. The solution was then diluted 6-fold with 50 mM Tris buffer/10mM CaCl2 (pH 7.6), proteins were digested into peptides with 1:100 (wt/wt) sequencing grade trypsin (Promega, Madison, WI), and insoluble cellular material was removed by centrifugation (2,000 *g* for 10 min). Peptides were desalted off-line by C18 solid phase extraction (Waters, Milford, MA), concentrated, filtered and aliquoted as

previously described (18). Two-dimensional nano-LC MS/MS analysis of each sample was carried out on an LTQ-Orbitrap Velos mass spectrometer (Thermo Fisher, San Jose, CA) as described elsewhere (19,20). In brief, chromatographic separation of the tryptic peptides was conducted over a 22 h period of increasing (0-500 mM) pulses of ammonium acetate followed by a 2 h aqueous to organic solvent gradient. The LTQ was operated in a data-dependent manner as follows: MS/MS on top ten ion detected in full scan, two microscans for both, full and MS/MS scans, centroid data for all scans, and dynamic exclusion set at 1.

Resulting MS/MS spectra were searched using the SEQUEST algorithm (21) with a concatenated isolate database containing all predicted proteins (including chloroplast proteins and mitochondria proteins) from *Aureococcus anophagefferens* CCMP 1784 and *Alcanivorax borkumensis* DG881, Alcanivorax sp. DG881, *Ostreococcus lucimarinus, Phadodactylum tricornutum* CCAP1055,*Thalassiosira pseudonana* CCMP1335, *Silicibacter pomeroyi* DSS-3, *Roseobacter denitrificans* OCh-114, *Candidatus Pelagibacter ubique* HTCC1062, *Gramella forsetii* KT0803, *Alcanivorax borkumensis* SK2, *Synechococcus* sp. CC9311, *Synechococcus* sp. CC9902. In addition to these "targeted" species likely in the bloom sample, the following species were included as non-target/decoy: *Geobacter bemidjiensis*, *Rhodopseudomonas palustris* CGA009, *Shewanella oneidensis* MR-1, *Bacteroides fragilis* YCH46, *Bifidobacterium longum* NCC2705, *Campylobacter jejuni* RM1221, *Enterobacter sakazakii* ATCC BAA-894, *Escherichia coli* K12, *Ferroplasma acidarmanus*, *Helicobacter pylori* 26695, *Listeria monocytogenes* EGD-e along with common contaminants (i.e. keratins and trypsin). The output data files were then filtered and sorted with the DTASelect algorithm (22) using the following

parameters: DeltCN of at least 0.08 and cross-correlation scores (Xcorrs) of at least 1.8 (+1), 2.5 (+2) and 3.5 (+3), with at least two peptides identified within the same run.

9. Non-genome data

For data in Fig. 1D, counts of picoeukaryotes, which are often dominated by *Ostreococcus* sp. (23), *A. anophagefferens,* and *Synechococcus* sp. were made via a flow cytometer (24, 25) whereas counts of diatoms, such as *Phaeodactylum* and *Thalassiosira*, were made with a light microscope and were converted to biovolume based on the measured dimension of cells. Light extinction coefficients were estimated from secchi disk measurements and DIN and DON were measured using wet chemistry techniques (26, 27). Irradiance levels required to achieve maximal growth rates of each phytoplankton species displayed in Fig 2A. were obtained from MacIntyre et al (28) and Six et al (29, 30). Selenium concentrations required to achieve maximal growth rates were obtained from Harrison et al (31), Wang and Dei (32), and from culture experiments conducted with axenic *A. anophagefferens* clone CCMP1984 grown in G-medium made from artificial seawater (33) supplemented with differing concentrations of selenium added as selenite. Cultures were grown at 21° C in an incubator with a 12:12h light:dark cycle, illuminated by a bank of fluorescent lights that provided a light intensity of ~100 µmol quanta m^{-2} s^{-1} to cultures. Cultures were maintained for a minimum of four transfers at each concentration prior to the collection of final growth rate data to ensure that cells were fully acclimated to treatment conditions and that the carryover of selenium from the initial, full strength was eliminated. Cellular growth rates were calculated for cultures based on cell densities in exponential growth phases, using the formula $\mu = \ln (B_t/B_0)/t$, where B_o and B_t are the initial and final biomass, and t is the incubation duration in days. Growth rates were averaged over the

entire exponential phase, which typically persisted for 3 – 6 days, depending on the concentration of selenium in the media. Se-limitation of cultures was confirmed by the stimulation of growth of cultures in stationary phase at concentrations below 5 nM following the addition of 10 nM Se.

SUPPLEMENTAL REFERENCES

1. Aparecio, et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, 297, 1301-10 (2002).
2. Berg, G.M., Shrager, J., Glockner, G., et al. Understanding nitrogen limitation in *Aureococcus anophagefferens* (Pelagophyceae) through cDNA and qRT-PCR analysis. J. Phycol. 44, 1235-1249 (2008.)
3. Salamov, A.A. & Solovyev, V.V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* 10, 516-522 (2000).
4. Birney, E. & Durbin, R. Using GeneWise in the Drosophila annotation experiment. *Genome Res* 10, 547-548 (2000).
5. Zdobnov, E.M. & Apweiler, R. InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847-848 (2001).
6. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32, D277-280 (2004).
7. Ashburner, M. et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-29 (2000).
8. Koonin, E.V. et al. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biology* 5 (2004)
9. Hall, T. A. BioEdit: a user-friendly biological sequences alignment editor and analysis programfor Windows 95/98/NT. *Nucleic Acids Symp. Ser.*41, 95-98. (1999).
10. Eppard, M. & Rhiel, E. The genes encoding light-harvesting subunits of *Cyclotella cryptica* (Bacillariophyceae) constitute a complex and heterogeneous family. *Mol .General Genetics* 260, 335-345 (1998).
11. Eppard, M., Krumbein, W. E., von Haeseler, A. & Rhiel, E. Characterization of fcp4 and fcp12, two additional genes encoding light harvesting proteins of *Cyclotella cryptica* (Bacillariophyceae) and phylogenetic analysis of this complex gene family. *Plant Biol* 2, 283-289 (2000).
12. Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24, 1596-1599 (2007).
13. Ren, Q., Kaixi Chen, K. & Paulsen I.T. TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. Nucleic Acids Res. 35, D274–D279 (2007).
14. Lobanov, A.V., Fomenko, D.E., Zhang, Y., Sengupta, A., Hatfield, D.L. & Gladyshev, V.N. Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic and small with terrestrial life. *Genome Biol.* 8, R198 (2007).
15. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410 (1990).
16. Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* 28, 33-36 (2000).
17. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. A genomic perspective on protein families. *Science* 278, 631-637 (1997).
18. VerBerkmoes, N.C., Shah, M.B., Lankford, P.K., Pelletier, D.A., Strader, M.B., Tabb, D.L., McDonald, W.H., Barton, J.W., Hurst, G.B., Hauser L., et al (2006). Determination and comparison of the baseline proteomes of the versatile microbe *Rhodopseudomonas palustris* under its major metabolic states. Journal of Proteome Research *5*, 287-298.

19. Ram R.J., Verberkmoes, N.C., Thelen, M.P., Tyson, G.W., Baker, B.J., Blake, R.C. II, Shah, M., Hettich, R.L., Banfield, J.F. (2005). Community proteomics of a natural microbial biofilm. Science *308*, 1915-1920.
20. Brown, S.D., Thompson, M.R., Verberkmoes, N.C., Chourey, K., Shah, M., Zhou, J., Hettich, R.L., Thompson, D.K. (2006). Molecular dynamics of the *Shewanella oneidensis* response to chromate stress. Mol. Cell Proteomics *5*, 1054-1071.
21. Eng, J. K., McCormack, A. L., Yates III, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry *5*, 976-989.
22. Tabb, D.L., McDonald, W.H., Yates, J.R. III. (2002). DTASelect and Contrast: Tools for Assembling and Comparing Protein Identifications from Shotgun Proteomics. J. Proteome Res. *1*, 21-26.
23. O'Kelly, C.J., Sieracvki, M.E., Their, E.C. & Hobson, I.C. A transiet bloom of Osterococcus (Chlorophyt, prasinophyceae) in West Neck Bay, Long Island, New York., *J. Phycol.*39, 850–854 (2003).
24. Olson, R.J., Zettler, E.R., Chisholm, S.W. & Dusenberry, J.A. In: Particle Analysis in Oceanography. (ed. Demers, S.) p 351-399 (Springer-Verlag, 1991)
25. Stauffer, B.A., Schaffner R.A., Wazniak, C. & Caron, D.A. Immunofluorescence Flow Cytometry Technique for Enumeration of the Brown-Tide Alga, Aureococcus anophagefferens Appl. Environ. Microbiol. 74, 6931–6940 (2008).
26. Parsons, T.R., Maita, Y. & Lalli, C.M. A manual of chemical and biological methods for seawater analysis. (Pergamon Press, 1984)
27. Valderama, J.C. The simultaneous analysis of total nitrogen and phosphorous in natural waters. *Mar. Chem.* 10, 109-122 (1981).
28. MacIntyre, H. L., Lomas, M. W., Cornwell, J., Suggett, D. J., Gobler, C. J., Koch, E. W. & Kana, T. M. Mediation of benthic-pelagic coupling by microphytobenthos: and energy- and material-based model for initiation of blooms of *Aureococcus anophagefferens*. *Harmful Algae* 3, 403-37 (2004).
29. Six, C., Thomas, J.C., Brahamsha, B., Lemoine, Y. & Partensky, F. Photophysiology of the marine cyanobacterium Synechococcus sp. WH8102, a new model organism. *Aquat. Microbiol. Ecol.* 35, 17-29 (2004).
30. Six, C., Finkel, Z.V., Rodriguez, F., Marie, D., Partensky, F. & Campbell, D.A. Contrasting photoacclimation costs in ecotypes of the marine eukaryotic picoplankter Ostreococcus *Limnol. Oceanogr.*, 53,, 255–265 (2008).
31. Harrison, P.J., Yu, P.W., Thompson, P.A., Price, N.M. & Phillips, D.J. Survey of selenium requirements in marine phytoplankton. *Mar. Ecol. Prog. Ser.* 47, 89-96 (1988).
32. Wang, W.-X. & Dei, R.C.H. Effects of major nutrient additions on metal uptake in phytoplankton *Environ. Pollution* 111, 233-240 (2001).
33. Doblin, M.A., Blackburn, S.I. & Hallegraeff, G.M. Growth and biomass stimulation of the toxic dinoflagellate *Gymnodinium catenatum* (Graham) by dissolved organic substances. *J. Exp. Mar. Biol. Ecol.* 236, 33-47 (1999).

2) SUPPLEMENTAL FIGURES

Figure S1. Phylogenetic tree constructed from amino acid sequences of predicted light harvesting chlorophyll-binding (LHC) proteins from two diatoms (*Phaeodactylum tricornutum*, Phatr, and *Thalassiosira pseudonana*, Thaps, both in black), two *Ostreococcus* species (*O. tauri*, Ossta, and *O. lucimarinus*, Ost9901, both in green), and *Aureococcus anophagefferens* (Auran, in red). Number after the colon is the protein identification number. Sixty-two LHC proteins were identified in the *A. anophagefferens* genome. This is approximately three times as many as have been found in the *Ostreococcus* strains (*O. tauri* has 16, *O. RCC809* has 19, and *O. lucimarinus* has 22), and 1.5 times as many as have been found in the diatoms (43 in *P. tricornutum* and 42 in *T. pseudonana*) (these numbers determined from extensive BLAST comparisons of LHCs among these genomes). None of the *A. anophagefferens* LHCs were most closely related to the major LHC types in the green plastid lineage, although four (LHC27, 36, 53, and 54) belonged to a group found in both the green and red plastid lineages (group I). None of the *A. anophagefferens* LHCs clustered with the 'major' fucoxanthin-chlorophyll binding proteins (FCP) of diatoms and other heterokonts (major FCP group). However, many *Aureococcus* LHCs grouped with other LHC sequences from *P. tricornutum* and *T. pseudonana* (as well as LHCs from other red-lineage algae not included in this tree). These groups (A to K) may share related (but as yet unknown) functions in the light-harvesting apparatus of *A. anophagefferens* and other red plastid lineage algae. One of these groups (group G) includes 16 LHCs from *A. anophagefferens* and 2 from *T. pseudonana*, and shares a unique PHYMKG motif near the end of helix two with 10 additional *A. anophagefferens* LHCs plus 5 more from the diatoms. EST and SAGE data show that at least 25 of the 26 in *A. anophagefferens* are expressed. There were also five groups of *A. anophagefferens* LHCs that were not closely related to diatom LHCs (Aur1 to Aur5). Aur1 was most similar to the major FCP cluster, Aur4 was similar to sequences from *Karlodinium micrum*, and Aur2 was loosely grouped with LHCs from several other dinoflagellates (dinoflagellate sequences not included in this tree). The phylogenetic tree was constructed in MEGA4 (*S12*) by the neighbor-joining method with pairwise deletion of gaps and distances calculated using the Poisson correction with rate variation among sites (gamma = 1). The optimal tree (sum of branch length = 85) is shown, the percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) are shown next to the branches, and the tree is drawn to scale with branch lengths in units of calculated amino acid substitutions per site.

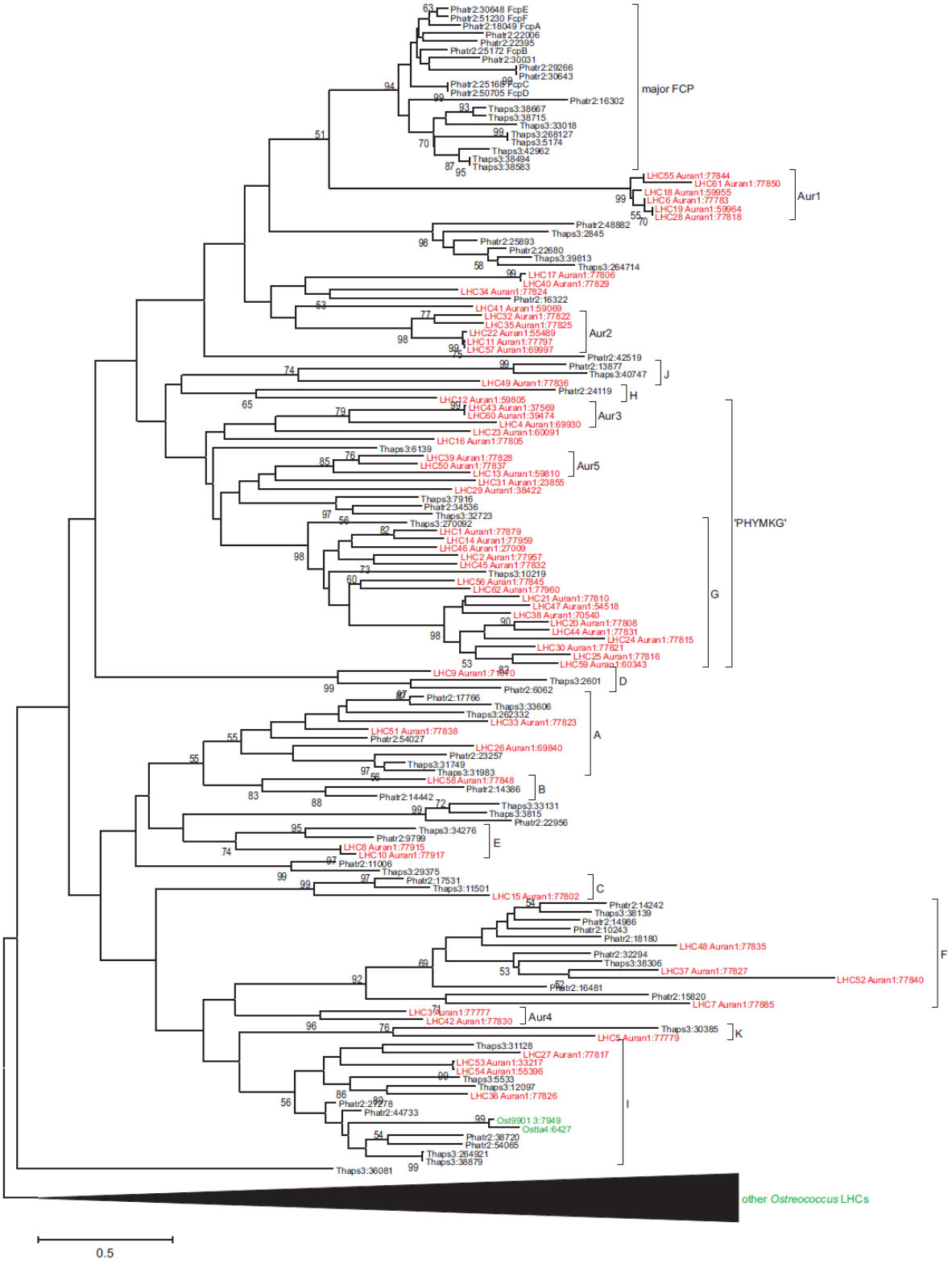
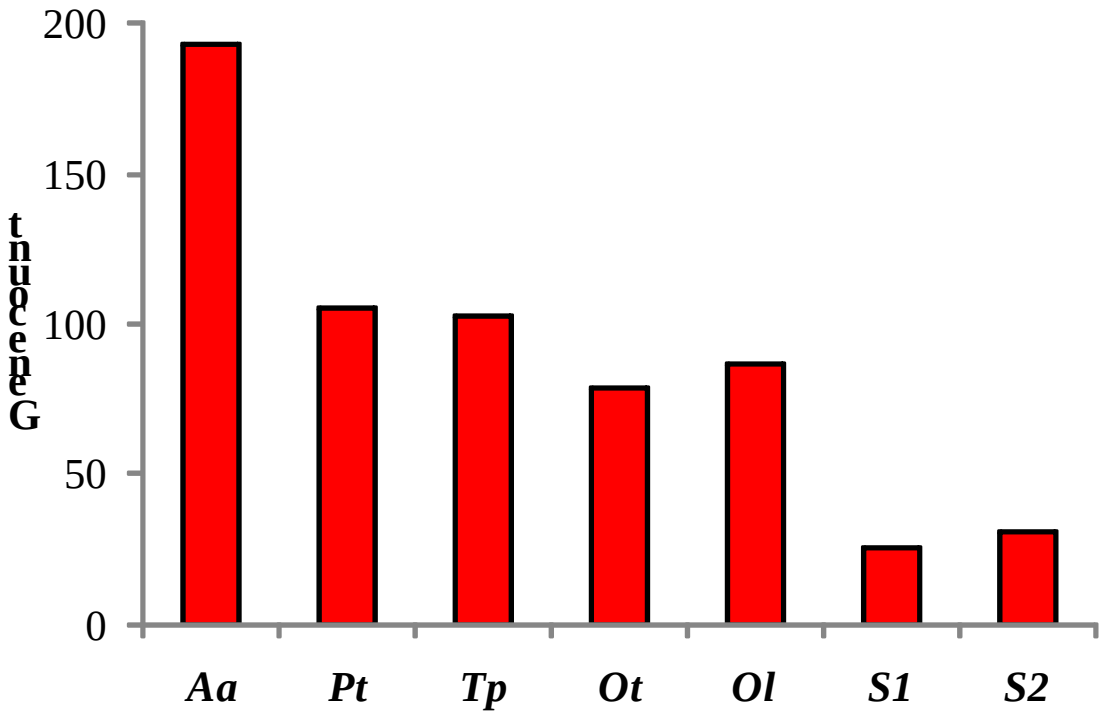

Phatr2:30648 FcpE
Phatr2:51230 FcpF
Phatr2:18049 FcpA
Phatr2:22006
Phatr2:22395
Phatr2:25172 FcpB
Phatr2:30031
Phatr2:29266
Phatr2:30643
Phatr2:25169 FcpC
Phatr2:50705 FcpD
Phatr2:16302
Thaps3:38667
Thaps3:38715
Thaps3:33018
Thaps3:268127
Thaps3:5174
Thaps3:42962
Thaps3:38494
Thaps3:38583
major FCP
LHC55 Auran1:77844
LHC61 Auran1:77850
LHC18 Auran1:59955
LHC6 Auran1:77783
LHC19 Auran1:59964
LHC28 Auran1:77818
Aur1
Phatr2:48882
Thaps3:2845
Phatr2:25893
Phatr2:22680
Thaps3:39813
Thaps3:264714
LHC17 Auran1:77806
LHC40 Auran1:77829
LHC34 Auran1:77824
Phatr2:16322
LHC41 Auran1:59069
LHC32 Auran1:77822
LHC35 Auran1:77825
LHC22 Auran1:55489
LHC11 Auran1:77797
LHC57 Auran1:69997
Aur2
Phatr2:42519
Phatr2:13877
Thaps3:40747
J
LHC49 Auran1:77836
Phatr2:24119
H
LHC12 Auran1:59805
LHC43 Auran1:37569
LHC60 Auran1:39474
LHC4 Auran1:69930
Aur3
LHC23 Auran1:60091
LHC16 Auran1:77805
Thaps3:6139
LHC39 Auran1:77828
LHC50 Auran1:77837
Aur5
LHC13 Auran1:59810
LHC31 Auran1:23855
LHC29 Auran1:38422
Thaps3:7916
Phatr2:34536
Thaps3:32723
Thaps3:270092
LHC1 Auran1:77879
LHC14 Auran1:77959
LHC46 Auran1:27009
LHC2 Auran1:77957
LHC45 Auran1:77832
Thaps3:10219
LHC56 Auran1:77845
LHC62 Auran1:77960
LHC21 Auran1:77810
LHC47 Auran1:54518
LHC38 Auran1:70540
LHC20 Auran1:77808
LHC44 Auran1:77831
LHC24 Auran1:77815
LHC30 Auran1:77821
LHC25 Auran1:77816
LHC59 Auran1:60343
'PHYMKG'
G
LHC9 Auran1:77970
Thaps3:2601
D
Phatr2:6062
Phatr2:17766
Thaps3:33606
Thaps3:262332
LHC33 Auran1:77823
LHC51 Auran1:77838
Phatr2:54027
LHC26 Auran1:69840
Phatr2:23257
Thaps3:31749
Thaps3:31983
LHC58 Auran1:77848
Phatr2:14386
B
Phatr2:14442
Thaps3:33131
Thaps3:3815
Phatr2:22956
A
Thaps3:34276
Phatr2:9799
LHC8 Auran1:77915
LHC10 Auran1:77917
E
Phatr2:11006
Thaps3:29375
Phatr2:17531
Thaps3:11501
LHC15 Auran1:77802
C
Phatr2:14242
Thaps3:38139
Phatr2:14986
Phatr2:10243
Phatr2:18180
LHC48 Auran1:77835
Phatr2:32294
Thaps3:38306
LHC37 Auran1:77827
LHC52 Auran1:77840
Phatr2:16481
Phatr2:15820
LHC7 Auran1:77885
F
LHC3 Auran1:77777
LHC42 Auran1:77830
Aur4
Thaps3:30385
LHC5 Auran1:77779
K
Thaps3:31128
LHC27 Auran1:77817
LHC53 Auran1:33217
LHC54 Auran1:55396
Thaps3:5533
Thaps3:12097
LHC36 Auran1:77826
Phatr2:2[illegible]78
Phatr2:44733
Ost9901 3:7949
Ostta4:6427
Phatr2:38720
Phatr2:54065
Thaps3:264921
Thaps3:38879
I
Thaps3:36081
other Ostreococcus LHCs
0.5

Figure S2. Genes associated with secondary metabolite biosynthesis, transport, and catabolism in *A. anophagefferens* and six comparative phytoplankton as identified via KOG analysis (EuKaryotic Orthologous Groups) for identifying ortholog and paralog proteins. Aa, Pt, Tp, Ot, Ol, S1, and S2 are *Phaeodactylum tricornutum*, *Thalassiosira pseudonana*, *Ostreococcus tauri*, *Ostreococcus lucimarinus*, *Synechococcus* clone CC9311 and *Synechococcus* clone CC9902, respectively.



3) SUPPLEMENTAL TABLES

Table S1. Sequence reads statistics of the *A. anophagefferens* genome

Insert Size	Untrimmed Sequence	Trimmed Sequence
2-3 KB	295.5 MB	214.6 MB
6-8 KB	289 MB	215 MB
35-40 KB	70 MB	18 MB
Total Untrimmed	654 MB	447 MB

Table S2. *A. anophagefferens* genome assembly statistics.

Scaffold Total:	1,202
Scaffold Sequence Total:	59.6 MB
Scaffold N50:	13
Scaffold L50:	1.3 MB
Contig Total:	5,776
Contig Sequence Total:	53.5 MB (10.1% gap)
Contig N50:	405
Contig L50:	34.0 KB
Estimated Depth:	7.0 +/- 0.13

Table S3. Mean characteristics of predicted gene models.

Gene length (bp)	2,138
Transcript length (bp)	1,601
Protein length (aa)	523
Exons per gene	2
Exon length (bp)	694
Intron length (bp)	412
Gene density /Mb	205

Table S4. Support for predicted gene models.

Gene count:	11,501
Supported by ESTs	2,435 (21%)
Supported by homology (Swissprot)	8,867 (77%)
Contain Pfam domain	7,834 (68%)

Table S5. Pfam domains unique to *A. anophagefferens* relative to six other phytoplankton**.**

Name	Number	Pfam domain number
Total	209	-
REJ domain	12	PF02010
Alpha-L-arabinofuranosidase B	5	PF09206
Lectin C-type domain	5	PF00059
Melibiase	5	PF02065
Myosin N-terminal SH3-like domain	5	PF02736
Poly(ADP-ribose) polymerase catalytic domain	5	PF00644
Scavenger receptor cysteine-rich domain	5	PF00530
Glycosyl hydrolases family 32 N terminal	4	PF00251
Laminin EGF-like (Domains III and V)	4	PF00053
Proprotein convertase P-domain	4	PF01483
Pro-kumamolisin, activation domain	4	PF09286
Berberine and berberine like	3	PF08031
Plastocyanin-like domain	3	PF02298
Glycosyl hydrolase family 20, domain 2	3	PF02838
Glycosyl hydrolase family 45	3	PF02015
Heparan sulfate 2-O-sulfotransferase (HS2ST)	3	PF05040
Neurotransmitter-gated ion-channel transmembrane region	3	PF02932
Polycystin cation channel	3	PF08016
PQQ enzyme repeat	3	PF01011
Plasmid pRiA4b ORF-3-like protein	3	PF07929
RhoGEF domain	3	PF00621
N-acetyltransferase	2	PF00797
Acyl-CoA thioesterase	2	PF02551
Cellulase	2	PF00150
CutC family	2	PF03932
Domain of Unknown Function (DUF1530)	2	PF07060
Domain of Unknown Function (DUF583)	2	PF04519
Domain of Unknown Function (DUF849)	2	PF05853
Endoplasmic reticulum protein ERp29, C-terminal domain	2	PF07749
F-actin capping protein alpha subunit	2	PF01267
Glycosyl hydrolase family 67 C-terminus	2	PF07477
Glycosyl hydrolase family 67 middle domain	2	PF07488
Glycosyl hydrolase family 92	2	PF07971
GMP-PDE, delta subunit	2	PF05351
Phospholipase B	2	PF04916
Neurotransmitter-gated ion-channel ligand binding domain	2	PF02931
Not1 N-terminal domain, CCR4-Not complex component	2	PF04065
NUMOD4 motif	2	PF07463
Pre-SET motif	2	PF05033
Radial spokehead-like protein	2	PF04712
Radial spoke protein 3	2	PF06098
UDP-glucoronosyl and UDP-glucosyl transferase	2	PF00201
WWE domain	2	PF02825
3-hydroxyanthranilic acid dioxygenase	1	PF06052
ab-hydrolase associated lipase region	1	PF04083
Allophanate hydrolase subunit 1	1	PF02682
Asparaginase	1	PF01112
BTB And C-terminal Kelch	1	PF07707
BCS1 N terminal	1	PF08740
Bestrophin	1	PF01062
Cysteine dioxygenase type I	1	PF05995
CUB domain	1	PF00431
Dioxygenase	1	PF00775
Down-regulated in metastasis	1	PF07539
Domain of Unknown Function (DUF108)	1	PF01958
Domain of Unknown Function (DUF1130)	1	PF06571
Domain of Unknown Function (DUF1446)	1	PF07287
Domain of Unknown Function (DUF1448)	1	PF07289
CEO family (DUF1632)	1	PF07857

Table S6. Pfam domains enriched in *A. anophagefferens* relative to six other phytoplankton.

Domain name	*A. anophagefferens*	*P. tricornutum*	*T. pseudonana*	*O. tauri*	*O. lucimarinus*	*Synechococcus (CC9311)*	*Synechococcus (CC9902)*	Pfam domain number
Protein kinase domain	265	122	144	102	94	1	1	PF00069
Ankyrin repeat	176	60	68	55	50	0	0	PF00023
WD 40 domain	142	79	90	86	98	0	0	PF00400
ABC transporter	117	52	56	47	41	33	32	PF00005
DnaJ domain	103	54	59	57	50	9	8	PF00226
EF hand	147	39	49	26	24	1	0	PF00036
FKBP-type peptidyl-prolyl cis-trans isomerase	53	27	25	26	21	2	2	PF00254
Sel1 repeat	130	6	8	2	3	0	0	PF08238
Cyclic nucleotide-binding domain	84	11	21	11	10	3	3	PF00027
Cyclophilin type peptidyl-prolyl cis-trans isomerase/CLD	43	19	24	22	22	2	2	PF00160
Kinesin motor domain	60	15	26	15	15	0	0	PF00225
ABC1 family	40	22	19	21	20	4	3	PF03109
Aldo/keto reductase family	48	22	28	10	10	4	3	PF00248
Ion transport protein	65	12	17	13	12	2	1	PF00520
Core histone H2A/H2B/H3/H4	48	16	26	14	15	0	0	PF00125
WW domain	44	14	23	12	14	0	0	PF00397
Protein phosphatase 2C	30	20	18	14	13	0	0	PF00481
KR domain	33	17	12	10	8	8	5	PF08659
AMP-binding enzyme	34	14	16	11	10	3	2	PF00501
C2 domain	34	21	11	11	9	0	0	PF00168
Beta-ketoacyl synthase, N-terminal domain	46	3	4	8	18	2	2	PF00109
Ion channel	33	12	7	9	9	5	4	PF07885
MYND finger	54	8	7	4	3	0	0	PF01753
Regulator of chromosome condensation (RCC1)	24	6	16	11	10	0	0	PF00415
Papain family cysteine protease	29	8	11	9	9	0	0	PF00112
Glutathione S-transferase, N-terminal domain	23	8	8	12	7	4	3	PF02798
Phytanoyl-CoA dioxygenase (PhyH)	32	9	8	7	6	0	1	PF05721
ABC-2 type transporter	19	12	8	9	10	1	2	PF01061
Armadillo/beta-catenin-like repeat	38	8	8	2	3	0	0	PF00514
Zinc-binding dehydrogenase	17	11	9	10	9	2	0	PF00107
PH domain	29	10	11	5	2	0	0	PF00169
Sulfatase	46	4	3	0	0	0	0	PF00884
Aminotransferase class-V	14	8	6	7	6	6	6	PF00266
Myosin head (motor domain)	31	9	11	1	1	0	0	PF00063
U-box domain	25	6	9	7	5	0	0	PF04564
CobW/HypB/UreG, nucleotide-binding domain	17	5	8	9	6	4	3	PF02492
Aldehyde dehydrogenase family	19	8	8	6	6	3	1	PF00171
Ubiquitin family	25	10	5	5	4	0	0	PF00240
Sodium/calcium exchanger protein	20	6	9	4	4	3	2	PF01699
Phosphatidylinositol 3- and 4-kinase	16	6	10	9	7	0	0	PF00454
Tubulin/FtsZ family, GTPase domain	15	6	8	8	7	1	1	PF00091
Tubulin/FtsZ family, C-terminal domain	14	6	8	8	7	1	1	PF03953
Dual specificity phosphatase, catalytic domain	16	5	8	8	5	1	0	PF00782
HECT-domain (ubiquitin-transferase)	15	9	7	8	4	0	0	PF00632
Tetratricopeptide repeat	25	4	1	6	4	1	1	PF07721
EGF-like domain	18	1	9	7	5	0	0	PF07974
Actin	15	7	4	6	6	0	0	PF00022
Cobalamin synthesis protein cobW C-terminal domain	11	4	7	6	4	3	2	PF07683
Acyl-CoA dehydrogenase, middle domain	12	6	6	6	5	0	0	PF02770
FG-GAP repeat	29	0	2	2	1	0	0	PF01839
Cation transporter/ATPase, N-terminus	12	7	6	5	3	0	0	PF00690
Phosphopantetheine attachment site	11	3	3	6	3	2	2	PF00550
Acyl-CoA dehydrogenase, C-terminal domain	11	6	6	4	3	0	0	PF00441
Clathrin adaptor complex small chain	9	5	5	6	5	0	0	PF01217
Cathepsin propeptide inhibitor domain (I29)	14	5	7	1	3	0	0	PF08246
Phospholipase/Carboxylesterase	10	3	6	3	4	1	1	PF02230
SelR domain	11	5	2	3	3	2	2	PF01641
Sulfotransferase domain	13	2	8	2	1	1	1	PF00685
'Cold-shock' DNA-binding domain	11	5	5	3	4	0	0	PF00313
Dynamin family	12	5	5	3	3	0	0	PF00350
Formin Homology 2 Domain	13	5	6	2	2	0	0	PF02181
Serine carboxypeptidase	13	6	4	2	2	0	0	PF00450
Cupin superfamily protein	11	5	5	3	3	0	0	PF08007
Synaptobrevin	9	6	6	3	3	0	0	PF00957
Amidohydrolase family	8	4	3	4	3	2	2	PF01979
Dynein heavy chain	18	1	3	2	2	0	0	PF03028
Dynein heavy chain, N-terminal region 2	16	1	4	2	2	0	0	PF08393
tRNA synthetases class I (E and Q), catalytic domain	6	4	4	3	3	2	2	PF00749
PPPDE putative peptidase domain	11	2	4	3	3	0	0	PF05903
GDSL-like Lipase/Acylhydrolase	11	5	2	1	2	0	1	PF00657
Domain of unknown function DUF21	6	4	4	2	2	2	2	PF01595
Zinc finger, C2H2 type	8	1	4	5	4	0	0	PF00096
Diacylglycerol acyltransferase	7	4	3	4	4	0	0	PF03982
Patched family	7	3	2	4	4	0	0	PF02460
Inorganic pyrophosphatase	8	3	3	1	1	2	2	PF00719
RIO1 family	6	3	3	4	4	0	0	PF01163
Aspartyl/Asparaginyl beta-hydroxylase	11	5	2	0	1	0	1	PF05118
Cytosol aminopeptidase family, catalytic domain	8	4	2	2	2	1	1	PF00883
BNR/Asp-box repeat	10	1	1	2	3	1	1	PF02012
Ribulose-phosphate 3 epimerase family	5	3	2	3	3	2	1	PF00834
Peptidase family M20/M25/M40	6	3	3	2	2	1	2	PF01546
Peptidase dimerisation domain	6	3	3	2	2	1	2	PF07687
SH3 domain	16	0	2	0	0	0	0	PF00018
Trehalose-phosphatase	7	2	3	3	3	0	0	PF02358
Calponin homology (CH) domain	10	2	2	2	2	0	0	PF00307
Proteasome/cyclosome repeat	7	2	3	3	3	0	0	PF01851
BT1 family	6	3	3	3	3	0	0	PF03092
Exostosin family	7	3	2	3	2	0	0	PF03016
Transport protein particle (TRAPP) component, Bet3	7	3	3	2	2	0	0	PF04051
Ribosomal protein L7/L12 C-terminal domain	6	3	2	2	2	1	1	PF00542
IBR domain	7	2	2	3	2	0	0	PF01485
Paired amphipathic helix repeat	7	2	3	2	2	0	0	PF02671
Glycosyltransferase family 20	5	2	3	2	2	1	1	PF00982
Glutamine synthetase, catalytic domain	5	2	3	1	1	3	1	PF00120
3-Oxoacyl-[acyl-carrier-protein (ACP)] synthase III C terminal	4	1	2	2	2	2	2	PF08541
FYVE zinc finger	8	4	3	0	0	0	0	PF01363
Acyl-CoA dehydrogenase, N-terminal domain	7	4	4	0	0	0	0	PF02771

Table S6. Continued.

Beta-lactamase	6	1	0	2	2	2	2	PF00144
Semialdehyde dehydrogenase, NAD binding domain	3	2	2	2	2	2	2	PF01118
Tubulin-tyrosine ligase family	10	1	2	1	1	0	0	PF03133
Aminomethyltransferase folate-binding domain	5	2	2	2	1	1	1	PF01571
Amidohydrolase	5	2	2	2	1	1	1	PF04909
Variant SH3 domain	12	0	2	0	0	0	0	PF07653
Acyltransferase family	7	2	1	1	1	2	0	PF01757
Pirin C-terminal cupin domain	5	3	2	2	2	0	0	PF05726
S-adenosylmethionine synthetase, central domain	5	1	2	2	1	1	1	PF02772
Chlamydia polymorphic membrane protein (Chlamydia_PMP)	10	1	0	1	1	0	0	PF02415
DJ-1/PfpI family	3	2	2	2	2	2	0	PF01965
Ribosomal RNA adenine dimethylase	3	2	2	2	2	1	1	PF00398
S-adenosylmethionine synthetase, N-terminal domain	5	1	2	2	1	1	1	PF00438
Glycine cleavage T-protein C-terminal barrel domain	5	1	2	2	1	1	1	PF08669
CAP-Gly domain	7	2	3	1	0	0	0	PF01302
Glycosyl hydrolase family 3 N terminal domain	8	3	0	0	0	1	1	PF00933
Uracil DNA glycosylase superfamily	5	2	2	1	1	1	1	PF03167
TruB family pseudouridylate synthase (N terminal domain)	3	2	2	2	2	1	1	PF01509
S-adenosylmethionine synthetase, C-terminal domain	4	1	2	2	1	1	1	PF02773
SAM domain (Sterile alpha motif)	6	1	2	1	2	0	0	PF00536
Domain of unknown function (DUF323)	4	2	2	1	1	1	1	PF03781
Dynein heavy chain, N-terminal region 1	5	1	3	2	1	0	0	PF08385
Flavin-binding monooxygenase-like	5	1	2	2	2	0	0	PF00743
Ribonucleotide reductase, barrel domain	6	2	2	1	1	0	0	PF02867
SFT2-like protein	5	3	1	2	1	0	0	PF07770
FMN-dependent dehydrogenase	6	2	2	1	1	0	0	PF01070
Dihydrodipicolinate synthetase family	5	1	2	1	1	1	1	PF00701
ATP cone domain	5	2	1	1	1	1	1	PF03477
SAM domain (Sterile alpha motif)	5	2	1	2	2	0	0	PF07647
RNA polymerase Rpb2, domain 4	4	2	2	2	2	0	0	PF04566
Coatomer WD associated region	4	2	2	2	2	0	0	PF04053
Family of unknown function (DUF500)	5	3	3	0	0	0	0	PF04366
GCC2 and GCC3	8	0	0	2	1	0	0	PF07699
TIP49 C-terminus	3	2	2	2	2	0	0	PF06068
Glycosyltransferase family 25 (LPS biosynthesis protein)	7	0	0	0	3	0	1	PF01755
MoeZ/MoeB domain	3	2	2	1	1	1	1	PF05237
LeuA allosteric (dimerisation) domain	3	2	2	1	1	1	1	PF08502
AFG1-like ATPase	3	2	2	2	2	0	0	PF03969
Adenylate kinase, active site lid	3	2	2	2	2	0	0	PF05191
PQ loop repeat	4	1	2	2	2	0	0	PF04193
Adenylylsulphate kinase	4	1	2	1	1	1	1	PF01583
NHL repeat	4	0	1	2	2	1	1	PF01436
Conserved region in glutamate synthase	3	2	2	1	1	1	1	PF01645
Mo-co oxidoreductase dimerisation domain	4	2	2	2	1	0	0	PF03404
Ribonucleotide reductase, all-alpha domain	5	2	2	1	1	0	0	PF00317
Putative esterase	3	2	2	2	1	1	0	PF00756
Protein of unknown function, DUF393	3	0	2	2	2	1	1	PF04134
DNA mismatch repair protein, C-terminal domain	3	2	2	2	2	0	0	PF01119
Domain of unknown function (DUF227)	5	1	2	1	1	0	0	PF02958
Glycosyl hydrolases family 16	6	2	1	1	0	0	0	PF00722
O-methyltransferase	5	1	0	2	2	0	0	PF01596
Uncharacterised protein family UPF0066	3	1	2	2	2	0	0	PF01980
MraW methylase family	3	1	1	2	1	1	1	PF01795
Sterol methyltransferase C-terminal	3	1	2	2	2	0	0	PF08498
Glucose-6-phosphate dehydrogenase, C-terminal domain	3	2	1	1	1	1	1	PF02781
Dynein light chain type 1	5	1	2	1	1	0	0	PF01221
Annexin	3	2	2	1	1	0	0	PF00191
Glycosyl hydrolases family 2, TIM barrel domain	4	1	1	2	1	0	0	PF02836
Vacuolar protein sorting-associated protein 26	3	1	1	2	2	0	0	PF03643
Hydroxymethylglutaryl-coenzyme A synthase N terminal	7	1	1	0	0	0	0	PF01154
Glycosyl hydrolase family 20, catalytic domain	7	2	0	0	0	0	0	PF00728
NADPH-dependent FMN reductase	4	2	1	0	0	1	1	PF03358
Leucine Rich Repeat	6	1	0	0	2	0	0	PF07723
6-phosphofructo-2-kinase	3	2	2	1	1	0	0	PF01591
Domain of Unknown Function (DUF1000)	3	2	1	2	1	0	0	PF06201
Protein-L-isoaspartate(D-aspartate) O-methyltransferase (PCMT)	3	2	2	1	1	0	0	PF01135
Cas1p-like protein	3	2	2	1	1	0	0	PF07779
Glycosyl hydrolases family 2, sugar binding domain	4	1	1	2	1	0	0	PF02837
Protein of unknown function (DUF938)	2	1	1	1	1	1	1	PF06080
Ketopantoate reductase PanE/ApbA	4	1	0	1	1	1	0	PF02558
Urease alpha-subunit, N-terminal domain	2	1	1	1	1	1	1	PF00449
Thymidylate kinase	2	1	1	1	1	1	1	PF02223
Dehydratase family	2	1	1	1	1	1	1	PF00920
MutL C terminal dimerisation domain	3	0	2	1	2	0	0	PF08676
Urease, gamma subunit	2	1	1	1	1	1	1	PF00547
Histone methylation protein DOT1	5	2	0	1	0	0	0	PF08123
Acyl transferase domain	2	1	1	1	1	1	1	PF00698
Glucokinase	3	1	1	1	0	1	1	PF02685
Urease beta subunit	2	1	1	1	1	1	1	PF00699
R3H domain	3	0	0	1	2	1	1	PF01424
Ferrochelatase	2	1	1	1	1	1	1	PF00762
Glycosyl hydrolases family 2, immunoglobulin-like beta-sandwich domain	3	1	1	2	1	0	0	PF00703
Protein kinase C terminal domain	4	1	0	2	1	0	0	PF00433
Cyclin	4	2	2	0	0	0	0	PF08613
Ribosomal protein L5	2	1	1	1	1	1	1	PF00281
Repeat of unknown function (DUF1126)	4	0	1	2	1	0	0	PF06565
Myotubularin-related	6	1	1	0	0	0	0	PF06602
Dynamin central region	4	1	1	1	1	0	0	PF01031
Porphobilinogen deaminase, dipyromethane cofactor binding domain	2	1	1	1	1	1	1	PF01379
Tctex-1 family	7	0	1	0	0	0	0	PF03645
Formate/nitrite transporter	2	1	1	1	1	1	1	PF01226
GUN4-like	2	1	1	1	1	1	1	PF05419
Aminopeptidase P, N-terminal domain	3	1	0	1	1	1	1	PF05195
Glycosyl hydrolase family 3 C terminal domain	5	3	0	0	0	0	0	PF01915
C-terminal regulatory domain of Threonine dehydratase	2	1	1	1	1	1	1	PF00585
ribosomal L5P family C-terminus	2	1	1	1	1	1	1	PF00673
Protein of unknown function, DUF590	4	2	2	0	0	0	0	PF04547
Mandelate racemase / muconate lactonizing enzyme, N-terminal domain	4	0	1	0	0	2	0	PF02746
Multicopper oxidase	4	0	1	0	1	1	0	PF07732
G-protein alpha subunit	4	1	2	0	0	0	0	PF00503
Prp19/Pso4-like	3	1	1	1	1	0	0	PF08606

Table S6. Continued.

B9 protein	5	0	2	0	0	0	0	PF07162
CHORD	3	1	1	1	1	0	0	PF04968
Thioesterase domain	3	0	0	2	2	0	0	PF00975
Cyanate lyase C-terminal domain	3	1	1	0	0	1	1	PF02560
Rapamycin binding domain	3	1	1	1	1	0	0	PF08771
Formamidopyrimidine-DNA glycosylase N-terminal domain	2	1	1	1	0	1	1	PF01149
Transporter associated domain	2	1	1	0	1	1	1	PF03471
Chalcone and stilbene synthases, C-terminal domain	3	1	1	0	0	1	1	PF02797
Natural resistance-associated macrophage protein	3	0	1	2	1	0	0	PF01566
Nicotinate phosphoribosyltransferase (NAPRTase) family	3	1	1	1	1	0	0	PF04095
Filamin/ABP280 repeat	4	0	1	0	2	0	0	PF00630
UPF0126 domain	3	0	0	1	2	1	0	PF03458
Guanylate-binding protein, N-terminal domain	5	1	1	0	0	0	0	PF02263
Scramblase	4	1	1	0	0	0	0	PF03803
Ribosomal protein L21e	2	1	1	1	1	0	0	PF01157
Vps52 / Sac2 family	2	1	1	1	1	0	0	PF04129
Coatamer beta C-terminal region	2	1	1	1	1	0	0	PF07718
Domain of unknown function (DUF1900)	4	1	1	0	0	0	0	PF08954
LEM3 (ligand-effect modulator 3) family / CDC50 family	2	1	1	1	1	0	0	PF03381
Histone deacetylase (HDAC) interacting	2	1	1	1	1	0	0	PF08295
SGS domain	2	1	1	1	1	0	0	PF05002
Eukaryotic DNA topoisomerase I, catalytic core	2	1	1	1	1	0	0	PF01028
SEP domain	2	1	1	1	1	0	0	PF08059
Glycosyl hydrolases family 28	4	2	0	0	0	0	0	PF00295
Receptor family ligand binding region	4	0	0	0	0	2	0	PF01094
Ribosomal protein L35Ae	2	1	1	1	1	0	0	PF01247
Eukaryotic protein of unknown function (DUF895)	4	2	0	0	0	0	0	PF05978
Malate/L-lactate dehydrogenase	2	1	1	1	1	0	0	PF02615
DSBA-like thioredoxin domain	2	1	1	1	1	0	0	PF01323
Phosphomannose isomerase type I	2	1	1	1	1	0	0	PF01238
Plus-3 domain	2	1	1	1	1	0	0	PF03126
Nep1 ribosome biogenesis protein	2	1	1	1	1	0	0	PF03587
Ribosomal protein S26e	2	1	1	1	1	0	0	PF01283
CRM1 C terminal	2	1	1	1	1	0	0	PF08767
RNA polymerase Rpc34 subunit	2	1	1	1	1	0	0	PF05158
Anti-silencing protein, ASF1-like	2	1	1	1	1	0	0	PF04729
Poly-adenylate binding protein, unique domain	2	1	1	1	1	0	0	PF00658
Ribosomal S13/S15 N-terminal domain	2	1	1	1	1	0	0	PF08069
Protein phosphatase 2A regulatory B subunit (B56 family)	2	1	1	1	1	0	0	PF01603
Macrocin-O-methyltransferase (TylF)	3	2	1	0	0	0	0	PF05711
Coatomer (COPI) alpha subunit C-terminus	2	1	1	1	1	0	0	PF06957
Calreticulin family	2	1	1	1	1	0	0	PF00262
Eukaryotic DNA topoisomerase I, DNA binding fragment	2	1	1	1	1	0	0	PF02919
Uncharacterized protein family UPF0027	2	0	1	1	1	0	0	PF01139
Protein of unknown function (DUF890)	2	0	1	1	1	0	0	PF05971
SRP19 protein	2	1	0	1	1	0	0	PF01922
Electron transfer flavoprotein-ubiquinone oxidoreductase	3	1	1	0	0	0	0	PF05187
Carboxylesterase	4	1	0	0	0	0	0	PF00135
Palmitoyl protein thioesterase	3	1	0	0	1	0	0	PF02089
Peptidyl-tRNA hydrolase PTH2	2	1	0	1	1	0	0	PF01981
Domain of unknown function (DUF298)	2	0	1	1	1	0	0	PF03556
Glutaredoxin 2, C terminal domain	3	1	1	0	0	0	0	PF04399
Frataxin-like domain	2	1	0	1	1	0	0	PF01491
Pectinacetylesterase	3	0	0	1	1	0	0	PF03283
Phenazine biosynthesis-like protein	3	0	1	1	0	0	0	PF02567
Tricarboxylate carrier	3	1	1	0	0	0	0	PF03820
Ribosomal L29e protein family	2	1	1	1	0	0	0	PF01779
Protein of unknown function (DUF339)	2	1	1	0	1	0	0	PF03937
p25-alpha	3	0	0	1	1	0	0	PF05517
Platelet-activating factor acetylhydrolase, plasma/intracellular isoform II	2	1	1	1	0	0	0	PF03403
Protein of unknown function (DUF1222)	3	1	1	0	0	0	0	PF06762
Polysaccharide deacetylase	2	1	1	0	0	0	0	PF01522
Domain of unknown function (DUF1899)	2	1	1	0	0	0	0	PF08953
Aldehyde oxidase and xanthine dehydrogenase, molybdopterin binding domain	2	1	1	0	0	0	0	PF02738
F5/8 type C domain	2	0	1	0	1	0	0	PF00754
CO dehydrogenase flavoprotein C-terminal domain	2	1	1	0	0	0	0	PF03450
Na$_{قي}$ antiporter 1	2	0	0	1	1	0	0	PF06965
Glucosamine-6-phosphate isomerases/6-phosphogluconolactonase	2	0	0	0	0	1	1	PF01182
HRDC domain	2	1	1	0	0	0	0	PF00570
ELMO/CED-12 family	2	0	0	1	1	0	0	PF04727
OmpA family	3	0	0	0	1	0	0	PF00691
MyTH4 domain	2	0	0	1	1	0	0	PF00784
FAD binding domain in molybdopterin dehydrogenase	2	1	1	0	0	0	0	PF00941
Peptidase family C69	3	1	0	0	0	0	0	PF03577
Ribonuclease B OB domain	2	0	0	0	0	1	1	PF08206
RhoGAP domain	2	1	1	0	0	0	0	PF00620
FAE1/Type III polyketide synthase-like protein	2	1	1	0	0	0	0	PF08392
[2Fe-2S] binding domain	2	1	1	0	0	0	0	PF01799
HpcH/HpaI aldolase/citrate lyase family	2	1	1	0	0	0	0	PF03328
Double-stranded DNA-binding domain	2	1	1	0	0	0	0	PF01984
Protein of unknown function DUF84	2	0	1	0	0	0	0	PF01931
START domain	2	1	0	0	0	0	0	PF01852
PPP5	2	1	0	0	0	0	0	PF08321
Glycosyl hydrolases family 35	2	1	0	0	0	0	0	PF01301
Methyl-CpG binding domain	2	0	1	0	0	0	0	PF01429
Common central domain of tyrosinase	2	1	0	0	0	0	0	PF00264
Elongation factor 1 gamma, conserved domain	2	1	0	0	0	0	0	PF00647
Calx-beta domain	2	0	0	0	0	1	0	PF03160
Arp2/3 complex, 34 kD subunit p34-Arc	2	0	0	0	1	0	0	PF04045
HELP motif	2	0	0	1	0	0	0	PF03451
NADH pyrophosphatase zinc ribbon domain	2	1	0	0	0	0	0	PF09297
VHS domain	2	0	0	1	0	0	0	PF00790

Table S7. Sixty two light harvesting complex genes in the *A. anophagefferens* genome. The number of genes in competing phytoplankton genomes with BlastP match e-value of < 10^{-5} is also depicted.

Pid	*P. tricornutum*	*T. pseudonana*	*O. lucimarinus*	*O. tauri*	*Synechococcus (CC9902)*	*Synechococcus (CC9311)*
77879	25	21	0	1	0	0
77957	16	16	1	1	0	0
77783	37	36	10	11	0	0
77885	29	24	0	0	0	0
59805	29	30	0	1	0	0
77959	30	23	0	0	0	0
77802	15	18	0	1	0	0
77805	37	36	10	11	0	0
77806	24	21	0	0	0	0
77808	29	26	0	0	0	0
55489	30	32	2	3	0	0
77816	29	29	0	0	0	0
77817	22	17	0	0	0	0
38422	21	17	0	0	0	0
77821	37	34	1	2	0	0
23855	5	6	0	0	0	0
77822	22	23	4	4	0	0
77825	23	19	0	0	0	0
77826	24	22	0	0	0	0
77829	23	20	0	1	0	0
59069	35	35	5	7	0	0
77830	24	22	0	0	0	0
77831	26	23	0	0	0	0
54518	26	24	0	1	0	0
77836	21	22	0	0	0	0
77838	28	28	1	1	0	0
55396	37	33	3	3	0	0
77845	15	15	0	0	0	0
69997	21	19	0	0	0	0
39474	18	18	3	2	0	0
77850	14	17	0	0	0	0
77960	33	24	4	7	0	0
69930	24	22	0	0	0	0
77882	15	18	0	0	0	0
77915	27	26	0	0	0	0
71070	28	30	10	9	0	0
77917	33	34	2	2	0	0
77921	28	26	0	0	0	0
59810	34	28	8	8	0	0
59955	21	22	5	5	0	0
59964	35	35	1	1	0	0
77810	37	33	3	4	0	0
60091	40	38	10	7	0	0
77815	18	19	0	0	0	0
69840	24	23	1	2	0	0
77818	19	22	8	8	0	0
77823	24	25	1	1	0	0
77824	36	34	2	3	0	0
77827	37	36	8	7	0	0
70540	10	9	3	2	0	0
77828	23	20	0	0	0	0
37569	23	20	0	0	0	0
77832	33	33	4	5	0	0
77835	21	19	0	0	0	0
77840	24	24	1	1	0	0
77844	38	35	5	4	0	0
77848	32	24	5	4	0	0
60343	20	18	5	3	0	0
77880	33	35	5	6	0	0
33217	29	32	5	6	0	0
27009	25	20	1	1	0	0
77837	19	21	1	0	0	0

Table S8. Number of genes encoding transporters proteins in *A. anophagefferens* and six other phytoplankton identified according to Ren et al (2007).

	A. anophagefferens	*P. tricornutum*	*T. pseudonana*	*O. tauri*	*O. lucimarinus*	*Synechococcus (CC9311)*	*Synechococcus (CC9902)*
Total Transporter Proteins:	647	518	476	321	311	127	78
ATP-binding Cassette (ABC)	112	54	45	38	33	39	30
Neurotransmitter receptor, ligand-gated	13	0	0	0	0	0	0
Voltage-gated Ion Channels	69	18	19	16	16	5	1
Total ion channel proteins	82	18	19	16	16	5	1
Oligopeptide - ion symporter	1	0	0	0	0	0	0
Peptide - ion symporter	1	0	0	0	0	0	0
Purine - ion symporters	2	0	0	0	0	0	0
D-xylose - ion symporter	4	0	0	0	0	0	0
Myo-inositol - ion symporter	1	0	0	0	0	0	0
Iron-phytosiderophore - ion symporter	1	0	0	0	0	0	0
Major facilitator superfamily sugar	7	3	4	2	2	0	0
Nitrate transporters	1	6	3	1	1	1	2
Amino acid transporters	16	23	23	5	6	7	5
Nucleoside transporters	1	0	0	1	1	0	0
Nucleotide-sugar transporters	15	9	10	6	7	0	0
Urea transporters	2	3	3	1	1	6	0
Glucose transporter	4	2	2	0	0	2	0
UDP-sugar transporter	2	0	0	0	0	0	0
Large, neutral amino acid transporter	1	0	0	0	0	0	0
Glycerol transporter	1	0	0	0	0	0	0
Molybdenum transporter	3	2	0	0	0	0	0
CutC family, Copper transporter	1	0	0	0	0	0	0
High-affinity Ni transporter (HoxN)	1	0	0	0	0	0	0

Table S9. Sulfatases in *A. anophagefferens* and six competing phytoplankton genomes as determined via BlastP matches with e-values of $< 10^{-5}$.

Protein	PID	*P. tricornutum*	*T. pseudonana*	*O. lucimarinus*	*O. tauri*	*Synechococcus (CC9902)*	*Synechococcus (CC9311)*	E-value
Total sulfatases (47 in A.a.)	-	4	3	0	0	0	0	
Arylsulfatase	23475	1	1	0	0	0	0	1.30E-54
Arylsulfatase	32660	1	1	0	0	0	0	2.40E-48
Arylsulfatase	23174	1	1	0	0	0	0	5.20E-47
Arylsulfatase	1628	2	2	0	0	0	0	2.60E-56
Arylsulfatase	27667	1	1	0	0	0	0	3.70E-20
Arylsulfatase	2359	1	1	0	0	0	0	1.40E-50
Arylsulfatase	2707	1	1	0	0	0	0	3.10E-43
Arylsulfatase	18760	1	1	0	0	0	0	4.20E-42
Arylsulfatase	13369	1	1	0	0	0	0	1.70E-26
Arylsulfatase	22180	2	1	0	0	0	0	1.10E-45
Arylsulfatase	22237	0	2	0	0	0	0	1.30E-08
Arylsulfatase	22610	1	1	0	0	0	0	1.70E-36
Arylsulfatase	63018	1	1	0	0	0	0	1.10E-43
Arylsulfatase	64040	1	1	0	0	0	0	3.80E-25
Arylsulfatase	64145	2	0	0	0	0	0	5.80E-41
Arylsulfatase	64446	2	1	0	0	0	0	2.00E-49
Arylsulfatase	30634	1	1	0	0	0	0	2.50E-52
Arylsulfatase	33025	1	1	0	0	0	0	6.50E-49
Arylsulfatase	60668	2	1	0	0	0	0	9.40E-56
Arylsulfatase	62515	1	1	0	0	0	0	5.80E-44
Arylsulfatase	62683	0	1	0	0	0	0	1.30E-07
Arylsulfatase	65375	1	1	0	0	0	0	6.50E-63
Arylsulfatase	66030	1	1	0	0	0	0	2.80E-25
Arylsulfatase	66827	1	1	0	0	0	0	3.10E-21
Arylsulfatase	68604	1	1	0	0	0	0	3.90E-47
Arylsulfatase	68689	2	3	0	0	0	0	4.10E-18
Arylsulfatase	70521	1	2	0	0	0	0	1.90E-09
Arylsulfatase	70842	1	1	0	0	0	0	2.40E-35
Arylsulfatase	71025	1	1	0	0	0	0	1.50E-32
Arylsulfatase	71524	1	1	0	0	0	0	8.00E-41
Arylsulfatase	72312	1	1	0	0	0	0	1.70E-36
Glucosamine (N-acetyl)-6-sulfatase	12528	2	2	0	0	0	0	3.20E-13
Glucosamine (N-acetyl)-6-sulfatase	29931	1	1	0	0	0	0	1.50E-37
Glucosamine (N-acetyl)-6-sulfatase	61173	1	1	0	0	0	0	9.00E-34
Glucosamine (N-acetyl)-6-sulfatase	65638	1	3	0	0	0	0	1.20E-15
Heparanase-like protein	3240	1	0	0	0	0	0	9.20E-15
Iduronate 2 sulfatase	67199	1	1	0	0	0	0	3.10E-12
Sulfatase	26323	1	1	0	0	0	0	3.00E-10
Sulfatase	5411	1	2	0	0	0	0	1.70E-10
Sulfatase	27261	1	1	0	0	0	0	2.20E-60
Sulfatase	37517	1	0	0	0	0	0	1.10E-39
Sulfatase	62077	1	1	0	0	0	0	9.30E-15
Sulfatase	64538	1	1	0	0	0	0	1.50E-31
Sulfatase	64729	1	1	0	0	0	0	5.00E-31
Tripeptidyl peptidase, sulfatase	62802	0	1	0	0	0	0	1.50E-35

Table S10. Genes encoding associated with sugar and oligosaccharide catabolism in *A. anophagefferens*. The number of genes in six competing phytoplankton genomes with a BlastP match with an e-value of $< 10^{-5}$ is also depicted.

Name	PID	*P. tricornutum*	*T. pseudonana*	*O. tauri*	*O. lucimarinus*	*Synechococcus (CC9311)*	*Synechococcus (CC9902)*	E_value
Total carbohydrate metabolism genes (85 in *A.a.*)	-	29	17	13	8	4	4	
Alpha-1,2-mannosidase	54067	0	0	0	0	0	0	
Alpha-1,2-mannosidase	68089	0	0	0	0	0	0	
Alpha-1,6-mannanase	3349	0	0	0	0	0	0	
Alpha-arabinofuranosidase	1957	0	0	0	0	0	0	
Alpha-arabinofuranosidase	2105	0	0	0	0	0	0	
Alpha-arabinofuranosidase	5150	0	0	0	0	0	0	
Alpha-arabinofuranosidase	21400	0	0	0	0	0	0	
Alpha-arabinofuranosidase	21424	0	0	0	0	0	0	
Alpha-galactosidase	2766	0	0	0	0	0	0	
Alpha-galactosidase	35927	0	0	0	0	0	0	
Alpha-galactosidase	27878	0	0	0	0	0	0	
Alpha-galactosidase	29988	0	0	0	0	0	0	
Alpha-galactosidase	20459	0	0	0	0	0	0	
Alpha-glucuronidase	10378	0	0	0	0	0	0	
Alpha-glucuronidase	66483	0	0	0	0	0	0	
Alpha-L-iduronidase	23523	0	0	0	0	0	0	
Beta-1,4- cellobiohydrolase	62046	0	0	0	0	0	0	
Beta-1,4-endoglucanase	5984	0	0	0	0	0	0	
Beta-1,4-endoglucanase	6432	0	0	0	0	0	0	
Beta-1,4-endoglucanase	6555	0	0	0	0	0	0	
Beta-fructofuranosidase	7475	0	0	0	0	0	0	
Beta-fructofuranosidase	61300	0	0	0	0	0	0	
Beta-fructofuranosidase	64125	0	0	0	0	0	0	
Beta-fructofuranosidase	65284	0	0	0	0	0	0	
Beta-glucuronyl hydrolase	70832	0	0	0	0	0	0	
Beta-mannosidase	67274	0	0	0	0	0	0	
Cellulase	12783	0	0	0	0	0	0	
Cellulase	23504	0	0	0	0	0	0	
D-galactarate dehydratase / Altronate hydrolase	20418	0	0	0	0	0	0	
di-N-acetylchitobiase	19716	0	0	0	0	0	0	
di-N-acetylchitobiase	26629	0	0	0	0	0	0	
di-N-acetylchitobiase	34613	0	0	0	0	0	0	
Endo-1,4-beta-xylanase	60931	0	0	0	0	0	0	
Mannitol dehydrogenase	70600	0	0	0	0	0	0	
Mannonate dehydratase	60041	0	0	0	0	0	0	
N-acylglucosamine 2-epimerase	33301	0	0	0	0	0	0	
Pectate lyase	70968	0	0	0	0	0	0	
Polysaccharide deacetylase	5548	0	0	0	0	0	0	
Polysaccharide deacetylase	71217	0	0	0	0	0	0	
Alpha-1,2-mannosidase	1504	2	4	2	2	0	0	6.70E-75
Alpha-1,2-mannosidase	10610	2	4	2	2	0	0	4.20E-67
Alpha-1,2-mannosidase	30199	2	4	2	2	0	0	3.00E-107
Alpha-1,2-mannosidase	53176	3	1	0	0	0	0	2.00E-56
Alpha-1,2-mannosidase	63227	1	3	1	2	0	0	1.00E-20
Alpha-1,2-mannosidase	71340	2	4	2	2	0	0	5.00E-142
Alpha-galactosidase	72107	1	1	0	0	0	0	2.50E-95
Alpha-glucosidase	10333	0	1	3	4	0	0	8.70E-83
Alpha-glucosidase	26418	1	1	0	1	0	0	2.10E-18
Alpha-glucosidase	67160	0	1	3	4	0	0	1.10E-60
alpha-glucosidase	70514	1	2	4	4	0	0	1.00E-168
Beta-galactosidase	31836	2	0	0	0	0	0	2.00E-13
Beta-galactosidase	61114	0	1	1	2	0	0	2.70E-08
Beta-galactosidase	66064	1	0	0	0	0	0	5.50E-23
Beta-glucanase	27453	2	1	0	1	0	0	4.60E-28
Beta-glucanase	24968	2	1	0	1	0	0	1.40E-24
Beta-glucanase	27689	2	1	0	1	0	0	9.40E-34
Beta-glucanase	19832	2	1	0	1	0	0	3.30E-26
Beta-glucanase	29716	2	1	0	1	0	0	1.50E-31
Beta-glucanase	70745	2	1	0	1	0	0	5.40E-11
Beta-glucosidase, Beta-xylosidase	539	6	0	0	0	1	1	5.20E-28
Beta-glucosidase, Beta-xylosidase	4612	3	0	0	0	0	0	1.70E-34
Beta-glucosidase, Beta-xylosidase	10207	5	0	0	0	0	1	2.30E-60
Beta-glucosidase, Beta-xylosidase	28037	2	0	0	0	0	0	1.50E-24
Beta-glucosidase, Beta-xylosidase	71699	5	0	0	0	0	0	1.50E-34
Beta-glucosidase, Beta-xylosidase	72703	4	0	0	0	0	0	4.40E-40
Beta-glucuronidase	3538	1	1	1	2	0	0	2.30E-42
Beta-glucuronidase	58877	1	1	1	2	0	0	4.00E-14
Beta-glucuronidase	71199	1	1	1	2	0	0	6.40E-09
Beta-hexosaminidase	22021	2	0	0	0	0	0	3.00E-27
Beta-hexosaminidase	60779	2	0	0	0	0	0	4.60E-22
Beta-hexosaminidase	61037	2	0	0	0	0	0	1.40E-16
Beta-N-acetylhexosaminidase	24518	2	0	0	0	0	0	5.30E-30
Beta-N-acetylhexosaminidase	28648	3	0	0	0	0	0	2.70E-52
Beta-N-acetylhexosaminidase	65096	2	0	0	0	0	0	1.90E-17
Beta-xylosidase	28884	3	0	0	0	0	0	1.00E-52
di-N-acetylchitobiase	61008	0	4	0	0	0	0	1.70E-07
Endo-1,3-beta-glucanase	63972	1	1	0	0	0	0	9.20E-46
Glucan 1,4-beta-glucosidase	64764	6	0	0	0	0	0	1.60E-22
Glucosamine-6-phosphate deaminase	9030	0	0	0	0	1	1	1.10E-09
Glucosamine-6-phosphate deaminase	29572	0	0	0	0	1	1	5.90E-12
Polygalacturonase	26417	2	0	0	0	0	0	2.80E-22
Polygalacturonase	36645	2	0	0	0	0	0	9.90E-15
Polygalacturonase	65850	2	0	0	0	0	0	1.90E-24
Polygalacturonase	71712	2	0	0	0	0	0	3.70E-13
Prunasin hydrolase	1771	1	1	2	1	0	0	1.00E-103

Table S11. The number of genes encoding associated with sugar and oligosaccharide catabolism in *A. anophagefferens* and six competing phytoplankton genomes as determined via a BlastP matches with an e-value of $< 10^{-5}$.

Genes	*A. anophagefferens*	*P. tricornutum*	*T. pseudonana*	*O. tauri*	*O. lucimarinus*	*Synechococcus (CC9311)*	*Synechococcus (CC9902)*
Genes unique to *A. anophagefferens*							
Alpha-1,6-mannanase	1	0	0	0	0	0	0
Alpha-arabinofuranosidase	5	0	0	0	0	0	0
Alpha-glucuronidase	2	0	0	0	0	0	0
Alpha-L-iduronidase	1	0	0	0	0	0	0
Beta-1,4 cellobiohydrolase	1	0	0	0	0	0	0
Beta-fructofuranosidase	4	0	0	0	0	0	0
Beta-glucuronyl hydrolase	1	0	0	0	0	0	0
Beta-mannosidase	1	0	0	0	0	0	0
Cellulase	2	0	0	0	0	0	0
D-galactarate dehydratase / Altronate hydrolase	1	0	0	0	0	0	0
endo-1,4-beta-xylanase	1	0	0	0	0	0	0
Endoglucanase	3	0	0	0	0	0	0
Mannitol dehydrogenase	1	0	0	0	0	0	0
Mannonate dehydratase	1	0	0	0	0	0	0
N-acylglucosamine 2-epimerase	1	0	0	0	0	0	0
Pectate lyase	1	0	0	0	0	0	0
Polysaccharide deacetylase	2	0	0	0	0	0	0
Total count	29	0	0	0	0	0	0
Genes enriched in *A. anophagefferens*							
Alpha-1,2-mannosidase	8	4	5	2	2	0	0
Alpha-galactosidase	6	2	1	0	1	2	2
Beta-galactosidase	3	1	1	1	2	0	0
Beta-glucanase	6	2	1	0	1	0	0
Beta-glucosidase, Beta-xylosidase	6	3	0	0	0	1	1
Beta-glucuronidase	3	1	1	1	2	0	0
Beta-hexosaminidase	3	2	0	0	0	0	0
Beta-N-acetylhexosaminidase	3	2	0	0	0	0	0
Polygalacturonase	4	3	0	0	0	0	0
Endo-1,3-beta-glucanase	1	1	1	0	0	0	0
Glucosamine-6-phosphate deaminase	2	0	0	0	0	1	1
Total count	45	21	10	4	8	4	4
Genes shared among the comparative phytoplankton							
Alpha-glucosidase	4	1	2	3	4	0	0
Beta-xylosidase	1	3	0	0	0	0	0
di-N-acetylchitobiase	4	0	4	0	0	0	0
Glucan 1,4-beta-glucosidase	1	3	0	0	0	0	0
Prunasin hydrolase	1	1	1	1	1	0	0
Total count	11	8	7	4	5	0	0

Table S12. Enzymes involved in degrading non-carbohydrate organic compounds in *A. anophagefferens* for which none of the competing phytoplankton genomes had a BlastP match with an e-value of $< 10^{-5.}$

Protein	Pid
Phospholipase B	11045
Phospholipase B	24092
Phospholipase D	59940
Lipase, ab-hydrolase	19589
Esterase	66525
Esterase	65650
Erythromycin esterase	33989
Hydrolase	25564
Hydrolase	69191
Hydrolase	72064
Hydrolase	65063
Hydrolase	64633
Hydrolase	63033

Table S13. Genes encoding enzymes involved in degrading organic nitrogen compounds in the *A. anophagefferens* genome. The number of genes in competing phytoplankton genomes with an e-value of < 10^{-5} during BlastP searches is also depicted.

Protein	*P. tricornutum*	*T. pseudonana*	*O. lucimarinus*	*O. tauri*	*Synechococcus (CC9902)*	*Synechococcus (CC9311)*	PID	E value
3-hydroxyanthranilic acid dioxygenase	0	0	0	0	0	0	25591	
4-hydroxyphenylacetate 3-hydroxylase family	0	0	0	0	0	0	34356	
Allophanate hydrolase	0	0	0	0	0	0	19876	
Asparaginase	0	0	0	0	0	0	24129	
Asparaginase/glutaminase	0	0	0	0	0	0	66990	
Aspartate dehydrogenase	0	0	0	0	0	0	13917	
Cysteine dioxygenase type I	0	0	0	0	0	0	33800	
Dioxygenase	0	0	0	0	0	0	64861	
Histidine ammonia-lyase	0	0	0	0	0	0	22572	
Lactam degradation enzyme, LamB/YcsF family	0	0	0	0	0	0	23531	
Membrane bound dipeptidase	0	0	0	0	0	0	59458	
Nitrile hydratase	0	0	0	0	0	0	25066	
Peptidase family C1 propeptide	0	0	0	0	0	0	72715	
Phenylalanine and histidine ammonia-lyase	0	0	0	0	0	0	22572	
Pro-kumamolisin	0	0	0	0	0	0	63942	
Proline racemase	0	0	0	0	0	0	23808	
Tripeptidyl peptidase	0	0	0	0	0	0	70813	
Urocanase	0	0	0	0	0	0	71046	
Acetamidase/Formidase	4	1	0	1	1	1	60068	2.80E-06
Acetamidase/Formidase	3	0	0	0	0	0	37987	1.20E-13
Aliphatic amidase	2	1	0	1	0	0	28333	1.00E-124
Asparaginase	2	1	2	1	1	1	25391	5.60E-80
Asparaginase	2	1	0	1	1	1	33691	7.70E-41
Beta-ureidopropionase-like protein	3	1	1	1	1	1	19125	7.40E-08
Cathepsin	6	4	2	2	0	0	52916	6.60E-26
Cyanase	1	1	0	0	1	1	21408	1.40E-15
Cyanase	1	1	0	0	1	1	58996	5.30E-47
Cyanase	1	1	0	0	1	1	60366	7.40E-51
Cystathionine beta lyase	4	3	4	4	3	3	58595	7.90E-16
Dihyrdolipoamide dehydrogenase	9	10	4	4	2	2	25795	2.80E-37
Dihyrdolipoamide dehydrogenase	11	11	4	4	2	2	26536	2.60E-09
Dipeptidase	7	4	0	0	2	2	68680	1.50E-33
Glycine cleavage protein	2	2	1	1	1	1	20469	2.40E-74
Glycine cleavage protein	1	1	1	1	0	0	59391	3.00E-59
Nitrilase	4	1	1	1	1	1	25106	2.30E-10
Nitrilase, hydratase	3	1	1	1	1	1	59241	1.10E-28
Nitrilase, hydratase	4	1	1	1	1	1	58715	3.70E-36
Nitrilase, hydratase	4	1	1	1	1	1	59719	3.60E-07
Nitrilase, hydratase	4	1	1	1	1	1	59666	2.80E-07
Nitriliase	4	1	1	1	1	1	27699	5.10E-26
Oxidoreductase	2	2	1	1	0	0	29204	3.30E-83
Peptidase	2	2	2	2	2	2	20552	0.00E+00
Proline dehydrogenase	2	1	1	1	0	0	25925	2.60E-18
Proline dehydrogenase	2	1	1	1	0	0	51906	2.00E-10
Pyroglutamyl peptidase	1	1	0	0	0	0	28248	1.30E-17
Urease	1	1	1	1	1	1	77851	1.00E-158
Urease	1	1	1	1	1	1	77852	0
Urease	1	1	1	1	1	1	77854	0
Urease accessory protein	1	1	1	1	1	1	30203	2.10E-74
Urease accessory protein	1	1	1	1	1	1	28624	1.50E-73
Urease accessory protein	1	1	3	1	0	0	65637	4.10E-11
5'-nucleotidase	2	2	1	1	0	0	29238	8.00E-106
5'-nucleotidase	2	2	1	1	0	0	20516	5.80E-37
5'-nucleotidase	1	1	0	1	0	0	21301	1.00E-129
5'-nucleotidase	0	0	0	0	0	0	28588	
5'-nucleotidase	3	2	1	1	0	0	60839	4.10E-16
Zinc carboxypeptidase	1	2	0	0	0	0	1129	1.80E-11
Zinc carboxypeptidase	1	2	0	0	0	0	10716	1.40E-09
Zinc carboxypeptidase	1	2	0	0	0	0	18992	1.40E-69
Zinc carboxypeptidase	1	2	0	0	0	0	23587	8.70E-09
Zinc carboxypeptidase	1	2	0	0	0	0	30345	4.40E-10
Zinc carboxypeptidase	0	1	0	0	0	0	60734	8.80E-07

Table S14. Selenocysteine-containing proteins in the *A. anophagefferens* genome.

Selenocysteine-containing proteins	PID
Thioredoxin domain containing protein	77962
Thioredoxin domain containing protein	78111
Thioredoxin domain containing protein	78110
Thioredoxin domain containing protein	78109
Thioredoxin domain containing protein	78108
Glutathione peroxidase	77994
Glutathione peroxidase	77995
Glutathione peroxidase	77996
Glutathione peroxidase	77997
Glutathione peroxidase	78003
Methionine sulfoxide reductase A	77998
Methionine sulfoxide reductase A	77999
Methionine sulfoxide reductase A	78000
Methionine sulfoxide reductase A	78001
Methionine sulfoxide reductase B	78002
Methionine sulfoxide reductase B	77940
Methionine sulfoxide reductase B	78004
Methionine sulfoxide reductase B	77942
Glutaredoxin	77970
Glutaredoxin	77969
Glutaredoxin	77966
Peroxiredoxin	78116
Peroxiredoxin	78117
Peroxiredoxin	77974
Selenoprotein T	77992
Selenoprotein W	77920
Selenoprotein W	77991
Selenoprotein W	78120
Selenoprotein W	78121
Selenoprotein W	78119
Selenoprotein O	77993
Selenoprotein U	77989
Selenoprotein U	77988
Selenoprotein M	77987
Selenoprotein M	78118
Fe-S oxidoreductase	77968
Selenoprotein Sep15	77986
Selenoprotein H	77985
Iodothyronine deiodinase	77984
Thioredoxin reductase	77983
Selenoprotein K	77979
Methyltransferase	78115
Methyltransferase	77963
Thiol:disulfide interchange protein	77965
Fe-S reductase	77972
UGSC-containing protein	77971
Membrane SelenoProtein	77964
GILT superfamily protein	77961
Rhodanase	78106
Protein disulfide isomerase	78005
Protein disulfide isomerase	77982
Protein disulfide isomerase	77977
Hypothetical protein	78107
Hypothetical protein	78112
Hypothetical protein	78113
Hypothetical protein	77980
Hypothetical protein	77978
Hypothetical protein	77975
Hypothetical protein	77967

Table S15. Distribution of metal-dependent proteins (Cu, Mo, Ni and Co) in *A. anophagefferens*

Metal	Metal-dependent protein family	Occurrence
Copper	Cu-Zn SOD	2
	Multi-copper oxidases	5
	Tyrosinase-like	20
Molybdenum	Sulfite oxidase	4
	Xanthine oxidase	2
Nickel	Urease	3
Cobalt (B12)	Methylmalonyl-CoA mutase	1
	Methionine synthase MetH	1

Table S16. Copper and molybdenum-containing proteins in the *A. anophagefferens* genome. The number of glycosylation sites per gene is also depicted.

Protein	Pid	Glycosylation sites
Tyrosinase-like	65038	5
Tyrosinase-like	65957	1
Tyrosinase-like	66243	
Tyrosinase-like	66567	
Tyrosinase-like	66629	
Tyrosinase-like	68995	
Tyrosinase-like	62628	
Tyrosinase-like	63818	1
Tyrosinase-like	63931	
Tyrosinase-like	64110	1
Tyrosinase-like	64134	
Tyrosinase-like	64228	
Tyrosinase-like	60940	
Tyrosinase-like	61043	8
Tyrosinase-like	62829	
Tyrosinase-like	63281	
Tyrosinase-like	64088	
Tyrosinase-like	64968	
Tyrosinase-like	72806	
Multicopper oxidase family	7686	
Multicopper oxidase family	27976	
Multicopper oxidase family	67552	
Multicopper oxidase family	71033	6
Multicopper oxidase family	72875	5
CuZn superoxide dismutase	7942	
CuZn superoxide dismutase	59136	
Sulfite oxidase family	64855	
Sulfite oxidase family	26887	
Sulfite oxidase family	55689	
Sulfite oxidase family	53391	
Xanthine oxidase family	36810	
Xanthine oxidase family	71657	

Table S17. Number of genes encoding enzymes associated with the deterrence of competitors and predators in the *A. anophagefferens* genome. The number of genes in competing phytoplankton genomes with an e-value of < 10^{-5} during BlastP searches is also depicted.

	P. tricornutum	*T. pseudonana*	*O. lucimarinus*	*O. tauri*	*Synechococcus (CC9902)*	*Synechococcus (CC9311)*	**Pid**
Berberine bridge enzyme	0	0	0	0	0	0	71018
Berberine bridge enzyme	0	0	0	0	0	0	62139
Berberine bridge enzyme	0	0	0	0	0	0	60769
Berberine bridge enzyme	0	0	1	1	1	0	67201
Berberine bridge enzyme	0	0	0	0	0	0	60770
Chloroquine transporter	0	0	0	0	0	0	65191
Membrane attack complex, Perforin domain	0	0	0	0	0	0	60790
Phenazine biosynthesis protein	0	1	0	0	0	0	19587
Phenazine biosynthesis protein	0	0	0	0	0	0	13446
Phenazine biosynthesis protein	0	0	0	0	0	0	35436
Erythromycin esterase	0	0	0	0	0	0	33989
ABC transporters, n=112	54	45	38	33	39	0	Muliple
Multi-drug ABC transporters, n=40	20	17	8	9	10	11	Muliple