LBNL-58730
Published

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

The Genome of the Diatom *Thalassiosira Pseudonana*: Ecology, Evolution and Metabolism

G. Armbrust, J. Berges, C. Bowler, B. Green, D. Martinez, N. Putnam, S. Zhou, A. Allen, K. Apt, M. Bechner, M. Brzezinski, B. Chaal, A. Chiovitti, A. Davis, D. Goodstein, M. Hadi, U. Hellsten, M. Hildebrand, B. Jenkins, J. Jurka, V. Kapitonov, N. Kroger, W. Lau, T. Lane, F. Larimer, J. Lippmeier, S. Lucas, M. Medina, A. Montsant, M. Obornik, M. Schnitzler Parker, B. Palenik, G. Pazour, P. Richardson, T. Rynearson, M. Saito, D. Schwartz, K. Thamatrakoln, K. Valentin, A. Vardi, F. Wilkerson, D. Rokhsar

Genomics

October 2004

Submitted to
Science

DISCLAIMER

The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism

*E. Virginia Armbrust,[1] John A. Berges,[2] Chris Bowler,[3, 4] Beverley R. Green,[5] Diego Martinez,[6] Nicholas H Putnam,[6] Shiguo Zhou,[7] Andrew E. Allen,[8, 4] Kirk E. Apt,[9] Michael Bechner,[7] Mark A. Brzezinski,[10] Balbir K. Chaal,[5] Anthony Chiovitti,[11] Aubrey K. Davis,[12] Mark S. Demarest,[10] J. Chris Detter,[6] Tijana Glavina,[6] David Goodstein,[6] Masood Z. Hadi,[13] Uffe Hellsten,[6] Mark Hildebrand,[12] Bethany D. Jenkins,[14] Jerzy Jurka,[15] Vladimir V. Kapitonov,[15] Nils Kröger,[16] Winnie W.Y. Lau,[1] Todd W. Lane,[17] Frank W Larimer,[18,6] J. Casey Lippmeier,[9,19] Susan Lucas,[6] Mónica Medina,[6] Anton Montsant,[3, 4] Miroslav Obornik, [5,20] Micaela Schnitzler Parker, [1] Brian Palenik,[12] Gregory J. Pazour,[21] Paul M. Richardson,[6] Tatiana A. Rynearson,[1] Mak A. Saito,[22] David C. Schwartz,[7] Kimberlee Thamatrakoln,[12] Klaus Valentin,[23] Assaf Vardi,[4] Frances P. Wilkerson,[24] *D. S. Rokhsar,[6, 25]

[1]School of Oceanography, University of Washington, Seattle, WA 98195, USA. [2]Department of Biological Sciences, University of Wisconsin-Milwaukee, Milwaukee WI 53201, USA. [3]Laboratory of Molecular Plant Biology, Stazione Zoologica, Villa Comunale, I 80121 Naples, Italy. [4] CNRS/ENS FRE2433, Dept of Biology, Ecole Normale Supérieure, 75230 Paris, France. [5]Dept. of Botany, University of British Columbia,Vancouver, B.C., Canada, V6T 1Z4. [6]DoE Joint Genome Institute, Walnut Creek, California, 94598, USA. [7]Depts. of Genetics and Chemistry, University of

Wisconsin-Madison, Madison, WI 53706, USA. [8]Department of Geosciences, Princeton University, Princeton, NJ 08540, USA. [9]Martek Biosciences Corp, 6480 Dobbin Rd, MD, 21045. [10]The Dept. of Ecology, Evolution and Marine Biology and the Marine Science Institute, University of California, Santa Barbara, CA 93106, USA. [11]School of Botany, University of Melbourne, Victoria 3010, Australia. [12]Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92093, USA. [13]Lockheed Martin Corporation, Sandia National Laboratory, PO box 969, MS-9951, Livermore, CA 94551, USA. [14]Ocean Sciences Depart., University of California Santa Cruz, Santa Cruz, CA 95064, USA. [15]Genetic Information Research Institute, Mountain View, CA 94043, USA.[16]Lehrstuhl Biochemie I, Universität Regensburg, D-93053, Regensburg, Germany. [17]Biosystems Research Department, Sandia National Labs, Livermore California 94551-0969. [18]Genome Analysis Group, Oak Ridge National Laboratory, Oak Ridge, TN 37831. [19]Department of Biological Sciences, University of Hull, Hull HU6 7RX, UK. [20]current address: Institute of Parasitology ASCR, Branisovska 31, 370 05 Ceske Budejovice, Czech Republic. [21]Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA. [22]Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA. [23]Alfred Wegener Institute, 27570 Bremerhaven, Germany. [24]Romberg Tiburon Center, San Francisco State University, Tiburon, CA 94920, USA. [25]Center for Integrative Genomics, University of California at Berkeley, Berkeley, CA, USA

*To whom correspondence should be addressed. Email: armbrust@ocean.washington.edu (E.V.A.), dsrokhsar@lbl.gov (D.S.R.)

Abstract

Diatoms are unicellular algae with plastids acquired by secondary endosymbiosis. They are responsible for ~20% of global carbon fixation. We report the 34 Mbp draft nuclear genome of the marine diatom, *Thalassiosira pseudonana* and its 129 Kbp plastid and 44 Kbp mitochondrial genomes. Sequence and optical restriction mapping revealed 24 diploid nuclear chromosomes. We identified novel genes for silicic acid transport and formation of silica-based cell walls, high-affinity iron uptake, biosynthetic enzymes for several types of polyunsaturated fatty acids, utilization of a range of nitrogenous compounds and a complete urea cycle, all attributes that allow diatoms to prosper in the marine environment.

Diatoms are unicellular, photosynthetic, eukaryotic algae found throughout the world's oceans and freshwater systems. They form the base of short, energetically-efficient food webs that support large-scale coastal fisheries. Photosynthesis by marine diatoms generates as much as 40% of the 45-50 billion tonnes of organic carbon produced each year in the sea (*1*), and their role in global carbon cycling is predicted to be comparable to that of all terrestrial rainforests combined (*2*, *3*). Over geological time, diatoms may have influenced global climate by changing the flux of atmospheric carbon dioxide into the oceans (*4*).

A defining feature of diatoms is their ornately patterned silicified cell wall or frustule, which displays species-specific nano-structures of such fine detail that diatoms have long been used to test the resolution of optical microscopes. Recent attention has focused on biosynthesis of these nano-structures as a paradigm for future silica

nanotechnology (*5*). The long history (over 180 million years) and dominance of diatoms in the oceans is reflected by their contributions to vast deposits of diatomite, most cherts and a significant fraction of current petroleum reserves (*6*).

As photosynthetic heterokonts, diatoms reflect a fundamentally different evolutionary history from the higher plants that dominate photosynthesis on land. Higher plants and green, red and glaucophyte algae are derived from a primary endosymbiotic event in which a non-photosynthetic eukaryote acquired a chloroplast by engulfing (or being invaded by) a prokaryotic cyanobacterium. In contrast, dominant bloom-forming eukaryotic phytoplankton in the ocean, such as diatoms and haptophytes, were derived by secondary endosymbiosis whereby a non-photosynthetic eukaryote acquired a chloroplast by engulfing a photosynthetic eukaryote, probably a red algal endosymbiont (Fig. 1). Each endosymbiotic event led to new combinations of genes derived from the hosts and endosymbionts (*7*).

Prior to this project, relatively few diatom genes had been sequenced, few chromosome numbers were known, and genetic maps did not exist (*8*). The ecological and evolutionary importance of diatoms motivated our sequencing and analysis of the nuclear, plastid, and mitochondrial genomes of the marine centric diatom *Thalassiosira pseudonana*.

Whole-Genome Sequencing and Architecture

The *T. pseudonana* genome was sequenced using a whole genome shotgun approach, with over fourteen-fold sequence coverage (*9*) (Table S1). Genomic DNA was isolated from a clonal culture produced by repeated mitotic divisions of a single diploid

founder cell. Finished quality circular mitochondrial and plastid genomes were derived from the whole genome shotgun data and organelle-enriched DNA.

Sequence polymorphisms between nuclear chromosome copies were easily detected as discrepancies between aligning reads supported by two or more reads for each of two alternate sequences. This analysis showed that 0.75% of the nucleotides in the nuclear genome are polymorphic (Fig. S1), which is comparable to levels seen previously in marine animals (*10, 11*). Only two haplotypes were observed, consistent with descent from a single diploid founder.

Twenty-four pairs of nuclear chromosomes ranging in size from 0.34 to 3.3 Mb were characterized by optical restriction site (*Nhe I*) mapping, for a total of 34.5 Mb. Over 90% of the sequence can be assigned to individual chromosomes, including a 350 kb cluster of ~35 copies of ribosomal DNA repeats (Fig. 2). Nine chromosome pairs can be separated into their respective haplotypes based on restriction site polymorphisms and seven of these pairs can be further distinguished from one another by insertions or inverted duplications ranging from tens of kb to a megabase in size. Surprisingly, DNA sequence and optical maps indicate that chromosome 23 is nearly identical to a comparably sized segment of chromosome 21. Chromosome 23 was likely created by an extremely recent duplication whose divergence is less than the haplotypic variation in the genome, possibly since the isolation of this laboratory strain. The plastid genome appears in the optical map as concatemers of similar restriction patterns, suggesting that several variants of the plastid genome are present. The mitochondrial genome at ~44 Kbp is too small for accurate optical mapping.

A majority of interspersed repeats in the nuclear genome are relics of transposable elements that constitute as much as 2% of the genome (Table S2). Among the most common transposable elements are LTR retrotransposons including *Copia* (*Copia1-9_TP*) and *gypsy*; both categories of diatom retrotransposons appear to self-prime reverse transcription in a manner not previously observed (Fig. S2). A potentially novel superfamily of nonautonomous transposable elements was observed that is derived from *gypsy,* but encodes only integrase and is flanked by terminal inverted repeats (TE2_TP). The presence of multiple, autonomous transposons suggest that transposon-tagging may hold future promise for generating diatom mutants (*12*).

Overview of gene and protein features

A total of 11,242 protein coding genes are predicted in the diatom nuclear genome (*9*), along with 144 plastid- and 40 mitochondrial-encoded genes (Fig. S3, Table 1). Of the total predicted nuclear proteins, 5916 are supported by homology (score >200, E-value < 1 e-20) with public database proteins, 7007 have recognizable Interpro domains, and 6799 can be assigned to a eukaryotic cluster of orthologous groups (Fig. S4). The unusual assortment of Interpro protein domains (Table 2) provides insights into regulatory mechanisms underlying gene expression, intracellular signaling, and transport that can now be tested experimentally.

The major categories of transcription factor in *T. pseudonana* are of the heat shock family, which are relatively rare in other eukaryotes. Transcription factors containing homeobox and Rel homology regions (RHR) are also well represented, although the latter lack the MADS-box domains found in higher plants. The importance

of chromatin-level control of gene expression can be inferred from identification of many putative proteins with characteristic domains such as bromodomains, chromodomains, histone deacetylases, SET, HMG, and RCC1 (Table 2). Regulation by RNA processing is also suggested by the high proportion of gene models encoding RNA processing proteins with DEAD box helicase domains (Table 2, Fig. S4).

Unlike many other eukaryotes, the diatom does not appear to rely heavily upon receptor kinases or LRR-containing receptors, as neither of the large categories of proteins containing kinase domains or LRR repeats (Table 2) possess obvious transmembrane domains. Furthermore, no G-protein-coupled receptors (GPCR) were found, suggesting that the major class of diatom receptor awaits discovery.

From the high numbers of kinase-encoding domains in the genome (Table 2), phosphorelay-based signal transduction systems appear to be commonly used in *T. pseudonana*. Of these, bacterial two-component histidine kinase-based signaling represents an important subclass, as sensor and response regulator domains were detected in a range of configurations. The most notable examples are a phytochrome with sensor and kinase domains that apparently functions as a light-regulated histidine kinase (see Light harvesting, photoprotection, and photoperception) and members of a prokaryotic nitrate/nitrite sensing system (NifR3, NarL) (*13*) with a receiver domain and a DNA-binding output domain, which would be the first example of such a nitrogen sensing system in eukaryotes. The lack of obvious Hpt proteins suggests that diatoms employ novel relay systems.

The diatom contains a dramatically reduced number of major facilitator family (MFS) transporters relative to other eukaryotes (Table 2). A relatively high proportion of

the identified transporters are similar to ones known to confer multi-drug resistance or to transport excess metals out of cells or into vacuoles indicating that diatoms likely avoid toxicity by exporting potential toxins out of the cytoplasm.

Establishment of Secondary Endosymbiosis

The secondary endosymbiosis hypothesis (Fig. 1; reviewed in *7*) predicts different possible origins for nuclear-encoded diatom genes: nuclear or mitochondrial genomes of the secondary heterotrophic host; nuclear, plastid or mitochondrial genomes of the red algal endosymbiont. To infer gene origins in the modern diatom we compared its proteome with that of two extant photosynthetic eukaryotes (the green plant *A. thaliana* and the red alga *Cyanidioschyzon merolae*) and one heterotrophic eukaryote (*Mus musculus*).

Almost half the diatom proteins have similar alignment scores to their closest homologs in plant, red algal and animal genomes (Fig. S5), underscoring the evolutionarily ancient divergence of Plantae (red algae, green algae and plants), Opisthokonta (animals/fungi), and the unknown secondary host that gave rise to the heterokont (diatom) lineage. Interestingly, 806 diatom proteins align with mouse proteins but not green plant or red alga proteins (Score ≥ 100 with BLOSUM62 matrix, or E-value < 1e-5) (Fig. 3A). The most straightforward interpretation is that these "animal-like" genes were derived from the heterotrophic secondary host, although scenarios involving gene loss in the plant/red algal lineage cannot be ruled out. Many could encode components of the flagellar apparatus and basal bodies, as has recently been shown for proteins shared by animals and the green alga *Chlamydomonas,* but not *A. thaliana* (*14*).

Contrary to expectation, 182 diatom proteins have matches only to red algal proteins, while 980 align only with plant proteins. This probably reflects the fact that the *A. thaliana* proteome is more than four times larger than the *C. merolae* proteome (Table 2). A similar comparison that included the cyanobacterium *Nostoc sp*. PCC 7120 emphasizes the large number of proteins shared by all groups of photosynthetic eukaryotes (Fig. 3B, S5), many of which are likely involved in chloroplast functions.

The assemblage of genes identified in our modern diatom representative reflects large-scale transfers of genes between genomes during establishment of the endosymbiosis (cf. *7, 15*). In addition to *C. merolae*, we also compared proteins encoded by the nucleomorph genome (remnant red algal symbiont nucleus) of the cryptophyte *Guillardia theta* (*16*) with the *T. pseudonana* nuclear-encoded proteome. Six nuclear diatom genes are most closely related to nucleomorph genes (*dhm*, *cbbX*, *cpn60* (*groEL*), *ftsZ*, *hcf136*, and *hli*). Four of these genes are also found on red algal plastid genomes (*cbbX, cpn60, ftsZ, hli*), thus demonstrating successive stages in gene transfer from red algal plastid to red algal nucleus (nucleomorph) to heterokont host nucleus.

In contrast to the situation in higher plants (*17*) and dinoflagellates (*18*), no evidence was found for recent large-scale transfers of *T. pseudonana* plastid or mitochondrial DNA to the nuclear genome. However, closely related copies of the *psbW* gene (*psb28*), which encodes a photosystem II protein, were found on both plastid and nuclear genomes, suggesting at least one plastid to nucleus transfer "in progress". Only a plastid copy of the *psbW* gene is in *C. merolae*, supporting the recent nature of this transfer. No red algal mitochondrial genes could be detected in *T. pseudonana* (E-value < 1e-5) and phylogenetic analysis of multiple mitochondrial genes supported inclusion of

diatom and red algal sequences in separate clades (Fig. S6). Both results are consistent with the hypothesis that red algal mitochondria were lost by the ancestral heterokont with only the heterotrophic host mitochondria retained (Fig. 1).

Establishment of a stable secondary endosymbiosis required evolution of a protein import system to allow cytoplasmically synthesized proteins to traverse the two additional membranes that surround the plastid (Fig.1). An N-terminal signal peptide sequence directs proteins across the endoplasmic reticulum (ER) membrane, which is continuous with the outermost plastid membrane in most heterokonts (*19*), but the mechanism of transit across the next three membranes remains unclear. A majority of putative plastid transit sequences examined (57/67) have a Phe at the +1 position (after the predicted signal cleavage site), frequently followed by a Pro at position +3, +4, +5 or +7 (Fig. S7, Table S3). The first 17-18 positions are enriched in Ser, Thr and Arg and have few negatively charged amino acids, similar to transit sequences of higher plants (*20*), but unlike those of the alveolate parasite *Plasmodium falciparum* (*21*). Tic 110 (with limited similarity to *C. merolae* Tic 110) was the only identified member of the Tic and Toc proteins that make up the chloroplast protein translocation system in plants (*22*) suggesting that novel changes to the import system have occurred. Homologs of plant stromal- and thylakoid-processing peptidases were identified, as were components of the Sec, Tat, and SRP thylakoid translocation-insertion machinery suggesting that once proteins arrive in the plastid stroma, machinery derived from the cyanobacterial ancestor directs them to their final destination.

Metabolic Adaptations to Life in Aquatic Environments

The Frustule - Life in a Glass House

The most distinctive aspect of diatoms is the frustule, which consists of hydrated silicon dioxide (silica) and a small amount of organic material. The strength of the frustule is hypothesized to help protect diatoms from being crushed during predation (*23*). The frustule is composed of two unequally sized halves connected by a series of overlapping siliceous girdle bands (*24*). Replication of the frustule during mitotic divisions creates two differently sized daughter cells, one of which is smaller than the parent cell. Cell size is ultimately restored through sexual reproduction (*24*). In generating and maintaining their frustules, diatoms control biogenic cycling of silicic acid in the ocean to such an extent that every atom of silicon entering the oceans is incorporated into diatom frustules about 40 times before its burial in sea floor sediments (*25*).

Silicon biochemistry is largely uncharacterized in any organism. We identified three novel genes that encode transporters for active uptake of silicic acid (*26*). Silica precipitates within a silica deposition vesicle (SDV) in a process controlled by long-chain polyamines and a family of Lys-and Ser-rich phosphoproteins (silaffins) that are embedded within the forming silica frustule (*27*). Five novel silaffins were identified based on genome sequence and N-terminal sequences of biosilica-associated *T. pseudonana* proteins (*28*). No matches were found to the silaffin gene sil-1 of the diatom *Cylindrotheca fusiformis* or to silicateins that initiate silica precipitation in sponges (*29*). Long-chain polyamine biosynthesis presumably requires spermidine and spermine

synthase-like enzymes (Fig. 4), and *T. pseudonana* has at least four times more copies of genes encoding these enzymes than any other organism sequenced to date.

Dissolution of the frustule in water is minimized because it is surrounded by an organic casing that includes glycoproteins with high levels of rhamnose and xylose, two sugars rarely found in other eukaryotic glycans. Based on the presence of highly conserved calcium binding domains (*30*), four new frustulins (casing glycoproteins) were identified, but no pleuralins, which are associated with terminal girdle bands in *C. fusiformis* (*31*). Consistent with a high content of hydroxylated amino acids found in the organic casing (*32*), one of the more abundant gene families in *T. pseudonana* encodes prolyl-4 hydroxylases.

Staying afloat

Diatoms must compensate for their dense silica frustule to maintain position within the illuminated portion of the water column. Some diatoms, including members of *Thalassiosira*, apparently increase drag and thus decrease sinking rates by extruding chitin fibers from pores in the frustule (*24*). These fibers can represent as much as 40% of total cell biomass and 20% of total cellular nitrogen (*33*). Enzymes for chitin biosynthesis and at least 22 putative chitinases were identified. This suggests that chitin fiber length may be dynamic and perhaps regulated at different life cycle stages or in response to a balance between a need for flotation and a need for enhanced nutrient delivery; a reduction in chitin fiber length will decrease the thickness of the boundary layer around a cell thus enhancing nutrient flux to the cell surface. The multiple chitinases could also play roles in defense against fungal infection. Identification of a

chitooligosaccharide N-deacetylase indicates that *T. pseudonana* also generates chitin-based oligosaccharides.

Nitrogen Metabolism/Urea Cycle

The genome encodes multiple transporters for nitrate and ammonium, the primary inorganic nitrogen sources for diatoms, and a single copy of a plastid-localized nitrite transporter (Fig. 4). The presence of multiple transporters for essential inorganic nutrients (nitrate, ammonium, phosphate, sulfate, silicic acid) likely reflects differential regulation and/or substrate affinities. We also found evidence for uptake and utilization of organic forms of nitrogen and for catabolism of amino acids (e.g., transaminases, glutamate dehydrogenase) and purines (e.g., uricase, allantoicase) indicating that *T. pseudonana* uses multiple forms of nitrogen.

Identification of enzymes necessary for a complete urea cycle was a surprise because this pathway has not been previously described in a eukaryotic photoautotroph. Unlike other organisms with a urea cycle, diatoms are unlikely to excrete "waste" urea since they posses an active urease and can grow on urea as a sole nitrogen source. In *T. pseudonana*, the enzyme that catalyzes the first step of the urea cycle (carbamoyl phosphate synthase, CPS III) appears to be targeted to mitochondria (Table S4) as in other organisms with a urea cycle. Higher plants possess a plastid-localized version of CPS that uses glutamine rather than NH_4^+ and is required for the first step of pyrimidine biosynthesis. The diatom has lost this form of CPS, but has a second CPS III-type gene without any organellar targeting sequence, suggesting that pyrimidine biosynthesis occurs in the cytoplasm of diatoms as in heterotrophs, rather than in plastids as in higher

plants/green algae (Fig. 4). This observation raises the intriguing question of how pyrimidines are transported across the four plastid membranes.

The urea cycle appears to be fully integrated into diatom metabolism in ways not previously suspected. Two intermediates of the urea cycle, arginine and ornithine, feed into other pathways present in the diatom. Ornithine is used to make spermine and spermidine, the polyamines required for diverse functions in all eukaryotes (*34*), and probably also long-chain polyamines required for silica precipitation during frustule formation. Ornithine can also be converted directly to proline by ornithine cyclodeaminase. The genome encodes many proline-rich proteins and proteins containing proline-binding motifs. Arginine is used for the synthesis of the signaling molecule nitric oxide via nitric oxide synthase, which in higher plants plays a role in pathogen defense. We also find evidence for generation of the energy storage molecule, creatine phosphate, via a urea cycle branch pathway originating from arginine (Fig. 4) that is absent from *A. thaliana*, *C. reinhardtii*, and *C. merolae*. Urea can also serve as an osmolyte, although this seems less likely in *T. pseudonana* since enzymes involved in synthesis of the more common osmolytes, betaine and mannose, as well as two plant-like halotolerance proteins were identified.

Carbon fixation

Despite the global importance of diatom carbon fixation, considerable debate still surrounds how diatoms concentrate and deliver CO_2 to the active site of the carbon-fixation enzyme Rubisco. Reinfelder and colleagues have championed the hypothesis that diatoms rely upon a carbonic-anhydrase dependent C4-like carbon concentrating

mechanism (*35*). We identified complete cytoplasmically-localized glycolytic and gluconeogenic pathways, as well as additional enzymes necessary for interconversions and production of C4 compounds (Fig. 4). However, we found no evidence for plastid localization of decarboxylating enzymes required for delivery of CO_2 to Rubisco. The multiple carbonic anhydrases identified here catalyze interconversions between CO_2 and bicarbonate, but all appear to localize to the cytoplasm rather than the plastid as in C3-plants and green algae. This *in silico* analysis should help guide future experimentation.

Energy stores

Diatoms accumulate fixed carbon primarily as chrysolaminaran (1,3-β-D-glucan) during nutrient-replete conditions. This carbohydrate is metabolized rapidly during the dark and both an endo- and an exo-1,3-β-D-glucanase were identified. In addition, putative sugar transporters were identified, which would enable uptake of reduced carbon from the environment.

As in metazoans, longer-term storage of reduced carbon in diatoms involves lipids (*36*), which have the additional advantage of enhancing buoyancy. All plastid-containing organisms examined to date, including *P. falciparum* and now *T. pseudonana,* carry out *de novo* biosynthesis of fatty acids within the plastid via a type II fatty acyl synthase, followed by export of fatty acids to the cytoplasm (*37*). At least 25% of the fatty acids synthesized by *T. pseudonana* are polyunsaturated (*38*), including the nutritionally important eicosapentaenoic (EPA) and docosahexaenoic (DHA) acids. We identified a complete pathway for polyunsaturated fatty acid biosynthesis, including several microsomal elongases and desaturases (including a putative Δ4 desaturase) that

successively modify simpler fatty acids. Additional lipid-related components identified include a deduced sterol biosynthetic pathway that should produce cholesterol, cholestanol and epibrassicasterol (Fig. 4), and a C-24(28) sterol reductase, presumably involved in the synthesis of 24-methylene sterols.

Two pathways for ß-oxidation of fatty acids are present in *T. pseudonana*. One pathway is localized to mitochondria because a full set of the required enzymes possess predicted mitochondrial transit peptides (Fig. 4, Table S4). The second pathway appears to be localized to peroxisomes because it includes an acyl-CoA oxidase which is restricted to peroxisomes in other organisms (*39, 40*), and potential peroxisomal targeting motifs (*40*) were found for enzymes known to be specific for ß-oxidation of polyunsaturated fatty acids, including a 2, 4-dienoyl-CoA reductase and a Δ3, 5-Δ2, 4-dienoyl-CoA isomerase. The peroxisomal pathway is expected to generate significant quantities of H_2O_2 and a gene for catalase/peroxidase was found, although its protein localization could not be predicted. As with higher plants, peroxisomal pathway products in *T. pseudonana* presumably feed into the glyoxylate cycle and ultimately into gluconeogenesis for carbohydrate production (Fig. 4). Thus diatoms appear to use stored lipids for both metabolic intermediates and generation of ATP, which likely explains how diatoms can withstand long periods of darkness and begin growing rapidly upon a return to the light.

Light harvesting, photoprotection, and photoperception

Diatoms commonly dominate in well-mixed water columns where they must cope with dramatic changes in intensity and spectral quality of light over relatively short time

frames. Our *in silico* analyses indicate that diatoms likely perceive blue and red, but not green light. We identified putative homologs of cryptochromes, which function as blue light photoreceptors in other eukaryotes (*41*), and phytochrome, which is consistent with an earlier report hypothesizing that diatoms can perceive red/far red light (*42*). No obvious matches to phototropins or rhodopsins (putative blue and green receptors) were identified. The absence of a detectable green light receptor was a surprise since green light persists to the greatest depth in coastal waters, while red light and blue light are both absorbed at relatively shallow depths. This combination of photoreceptors may help diatoms perceive their proximity to the surface and/or detect red chlorophyll fluorescence from neighboring cells (*43*).

The light-harvesting complex (LHC) family in *T. pseudonana* includes at least 30 fucoxanthin-chlorophyll *a/c* proteins that absorb light and transfer it to photosynthetic reaction centers. No relicts of red algal phycobiliprotein genes were detected. No evidence was found for the PsbS protein essential for operation of the photoprotective xanthophyll cycle in higher plants (*44*). This is surprising since the major mechanism for dissipation of excess light energy in diatoms is an augmented xanthophyll cycle that involves the interconversion of diadinoxanthin and diatoxanthin in addition to the well-known violaxanthin-zeaxanthin interconversion found in higher plants (Fig. 4). No genes were found for other photoprotective members of the LHC superfamily (e.g. Elips, Seps) except for two small Hli proteins hypothesized to protect against damage from reactive oxygen species in cyanobacteria.

Damage from reactive oxygen species generated during photosynthesis could also be minimized by the two Fe type- and two Mn type- superoxide dismutases (SODs).

Similar to *P. falciparum*, no obvious match to a Cu/Zn-type SOD was found, nor was a match found for a Ni-containing SOD recently discovered in marine cyanobacteria (*45*). Components of several pathways associated with utilization of the antioxidants glutathione, ascorbate and alpha-tocopherol were also identified.

Iron Uptake

Productivity of major regions of the modern surface ocean is limited by low iron levels (*46*). Diatoms frequently dominate phytoplankton blooms created during large-scale iron fertilization experiments, emphasizing their important role in the marine carbon cycle (*47, 48*). We identified components of a high-affinity iron uptake system (*49*) composed of at least two putative ferric reductases that contain the required heme and co-factor binding sites necessary for activity. In addition, a multicopper oxidase and two iron permeases were identified that together could deliver Fe^{3+} to cells via reduction to ferrous iron. Diatoms may also use iron transport proteins found in cyanobacteria and they possess genes that appear to encode key enzymes necessary for the synthesis of enterobactin, an iron scavenging siderophore. Genes for metallothioneins and for phytochelatin synthases, which play important roles in metal homeostasis and detoxification, were also identified.

Conclusions

Sequence and optical mapping of the *T. pseudonana* genome showed that it is diploid with 24 chromosome pairs, data that could not be obtained by conventional cytological techniques. Analysis of predicted coding sequences demonstrated that it

possesses a full complement of transporters for the acquisition of inorganic nutrients and a wide range of metabolic pathways, as expected for a highly successful photoautotroph. Its origin by secondary endosymbiosis is supported by evidence for gene transfer from the nucleus of the red algal endosymbiont, and by the presence of ER signal sequences on chloroplast-targeted proteins.

About half the genes in the diatom cannot be assigned functions based on similarity to genes in other organisms, in part because diatoms have distinctive features that cannot be understood by appeal to model systems. Diatoms are unique in how they metabolize silicon to form their characteristically ornate silica frustule; protein transport into plastids is a more complicated system than is currently understood; the way by which CO_2 is delivered to RubisCo remains unclear; the high proportion of polyunsaturated fatty acids produced and their oxidation to feed intermediate metabolism is unusual among eukaryotes; even the receptors required to integrate environmental signals remain unknown. The presence of the enzymatic complement of the urea cycle is surprising; since there was no reason to suspect its presence, there is no current information about metabolic fluxes through the pathway. The unusual assortment of protein domains may reflect novel mechanisms of gene regulation.

The genomic information provided by this project suggests starting points for a number of new experimental investigations of the biology of these globally important organisms, and their interaction with the marine environment in which they thrive. Using genome sequence to infer ocean ecology provides a powerful new approach to explore ecosystem structure.

References

1. D. M. Nelson, P. Tréguer, M. A. Brzezinski, A. Leynaert, B. Quéguiner, *Glob. Biogeochem. Cycle* **9**, 359-372 (1995).
2. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. G. Falkowski, *Science* **281**, 237-240 (1998).
3. D. G. Mann, *Phycologia* **38** (1999).
4. M. A. Brzezinski *et al.*, *Geophys. Res. Lett.* **29**, 10.1029/2001GL014349 (2002).
5. J. Parkinson, R. Gordon, *Trends Biotechnol.* **17**, 190-6 (1999).
6. J. S. S. Damste´ *et al.*, *Science* **304**, 584-587 (2004).
7. P. G. Falkowski *et al.*, *Science* **305**, 354-360 (2004).
8. A. Falciatore, C. Bowler, *Ann. REv. Plant Biol.* **53**, 109-130 (2002).
9. Materials and methods are available as supporting material on *Science* Online.
10. P. Dehal *et al.*, *Science* **298**, 2157 (2002).
11. S. Aparicio *et al.*, *Science* **297**, 1301 (2002).
12. V. Walbot, *Curr. Opin. Plant Biol.* **3**, 103-107 (2000).
13. V. Stewart, P. J. Bledsoe, *J. Bacteriol.* **185**, 2104-2111 (2003).
14. J. B. Li *et al.*, *Cell* **117**, 541-552 (2004).
15. P. J. Keeling, J. M. Archibald, N. M. Fast, J. D. Palmer, (submitted).
16. S. Douglas *et al.*, *Nature* **410**, 1091-1096 (2001).
17. W. Martin, *PNAS* **100**, 8612-8614 (2003).
18. J. D. Hackett *et al.*, *Curr. Biol.* **14**, 213-218 (2004).
19. P. G. Kroth, *Int Rev Cytol.* **221**, 191-255 (2002).
20. X.-P. Zhang, E. Glaser, *Trends in Plant Science* **7**, 14-21 (2002).
21. B. J. Foth *et al.*, *Science* **299**, 705-708 (2003).
22. K. Cline, in *Light- Harvesting Antennas in Photosynthesis. Advances in Photosynthesis and Respiration* B. Green, W. Parson, Eds. (Kluwer Academic Publishers, 2003) pp. 353-372.
23. C. E. Hamm *et al.*, *Nature* **421**, 841-843 (2003).
24. F. E. Round, R. M. Crawford, D. G. Mann, *The diatoms: Biology and morphology of the genera* (Cambridge University Press, 1990).
25. P. Tréguer *et al.*, *Science* **268**, 375-379 (1995).
26. M. Hildebrand, B. E. Volcani, W. Gassmann, J. L. Schroeder, *Nature* **385**, 688–689 (1997).
27. M. Sumper, N. Kröger, *J. Mat. Chem.* **14**, 2059-2065 (2004).
28. N. Poulsen, N. Kröger, *J. Biol. Chem.* **in press** (2004).
29. K. Shimizu, J. Cha, G. D. Stucky, D. E. Morse, *Proc. Natl. Acad. Sci. USA* **96**, 361 (1998).
30. N. Kröger, C. Bergsdorf, M. Sumper, *EMBO.* **13**, 4676-4683 (1994).
31. N. Kröger, R. Wetherbee, *Protist* **151**, 263 (2000).
32. B. E. Volcani, in *Silicon and siliceous structures in biological systems* R. Simpson, B. E. Volcani, Eds. (Springer, New York, 1981) pp. 157-2000.
33. J. McLachlan, A. G. McInnes, M. Falk, *Can. J. Bot.* **43**, 707 (1965).
34. P. Coffino, *Proc. Natl. Acad. Sci. USA* **97**, 4421-4423 (2000).
35. J. R. Reinfelder, A. M. L. Kraepiel, F. M. M. Morel, *Science* **407**, 996-999 (2000).

36. G. A. Dunstan, J. K. Volkman, S. M. Barrett, C. D. Garland, *J. Appl. Phycol.* **5**, 71-83 (1993).
37. R. F. Waller *et al.*, *Proc Natl Acad Sci U S A* **95**, 12352-7 (1998).
38. G. A. Dunstan, J. K. Volkman, S. M. Barrett, J. M. Leroi, S. W. Jeffrey, *Phytochemistry* **35**, 155-161 (1994).
39. M. Fulda, J. Shockey, M. Werber, F. P. Wolter, E. Heinz., *The Plant Journal* **32**, 93 (2002).
40. A. T. J. Klein, M. van den Berg, G. Bottger, H. F. Tabak, B. Distel., *J. Biol. Chem.* **277**, 25011-25019 (2002).
41. M. Yanovsky, S. Kay, *Nature Rev. Mol. Cell Biol.* **4**, 265-275 (2003).
42. C. Leblanc, A. Falciatore, C. Bowler, (1999), *Plant Mol. Biol.* **40**, 1031–1044 (1999).
43. M. Ragni, M. Ribera, *J Plankton Res.* **26**, 433-443 (2004).
44. X.-P. Li *et al.*, *Nature* **403**, 391-395 (2000).
45. B. Palenik *et al.*, *Nature* **424**, 1037-1042 (2003).
46. J. K. Moore, S. C. Doney, D. M. Glover, I. Y. Fung, *Deep-Sea Res. II* **49**, 463-507 (2002).
47. P. W. Boyd *et al.*, *Nature* **407**, 695-702 (2000).
48. K. H. Coale *et al.*, *Science* **304**, 408-414 (2004).
49. N. J. Robinson, C. M. Procter, E. L. Connolly, M. L. M. L. Guerinot, *Nature* **397**, 696-697 (1999).
50. This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and the by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098 and Los Alamos National Laboratory under contract No. W-7405-ENG-36.; and DOE (DE-FG03-02ER63471 to E.V.A.), European Union Margens (QLRT-2001-01226 to C.B.), the CNRS Atip programme (2JE144 to C.B), and U.S. EPA (R827107-01-0, Basic to B.P., M.H.).

Supporting Online Material

www.sciencemag.org

Materials and Methods

Figs. S1 to S7

Tables S1 to S4

References

Table 1. General features of *Thalassiosira pseudonana* genomes

Feature	Value
Nuclear Genome	
Size (bp)	34,266,941
Chromosome number	24
Chromosome size range (bp)	360,000 – 3,300,000
G + C content (overall %)	47
G + C content (coding %)	48
Transposable elements (overall %)	2
Protein-coding genes	11, 242
Average gene size (bp)	992
Average number introns per gene	1.4
Gene density (bp per gene)	3,500
tRNAs	131 (includes at least 1 per codon)
Plastid Genome	
Size (bp)	128,813
G + C content (overall %)	31
Protein-coding genes	144
Gene density (bp per gene)	775
tRNAs	33
Mitochondrial Genome	
Size (bp)	43,827
G + C content (overall %)	30.5
Protein-coding genes	40
Gene density (bp per gene)	1137.5
tRNAs	22

Table 2. Protein domains (based on Interpro matches) in *T. pseudonana* and comparison with 4 other eukaryotes. Estimated proteome size is given in parentheses under the name of each organism.

	T. pseudonana (11,242)		*C. merolae* (6,229)		*A. thaliana* (26,187)		*N. crassa* (10,082)		*M. musculus* (27,098)		
Interpro ID	Proteins	Rank	Proteins	Rank	Proteins	Rank	Proteins	Rank	Proteins	Rank	Interpro Name
Ten most abundant domains in *T. pseudonana*											
IPR000719	202	1	71	4	1055	1	122	4	594	4	Protein kinase 1
IPR002290	136	2	61	5	228	2	101	7	351	6	Serine/threonine protein kinase 1
IPR003593	135	3	86	2	314	10	93	8	136	31	AAA ATPase
IPR001680	116	4	81	3	267	12	121	5	325	8	G-protein beta WD-40 repeats
IPR001245	111	5	52	8	43	163	90	10	250	7	Tyrosine protein kinase
IPR001440	105	6	27	21	137	33	37	25	165	29	TPR repeat
IPR002110	101	7	10	32	156	38	49	19	292	17	Ankyrin
IPR001410	97	8	57	6	153	28	74	13	119	44	DEAD/DEAH box helicase
IPR001611	96	9	10	32	544	4	14	44	257	19	Leucine-rich repeat
IPR001650	95	10	56	7	147	40	71	14	107	62	Helicase, C-terminal
Additional abundant domains in *T. pseudonana*											
IPR003590	32	41	2	>50	0	--	2	>50	0	--	Leucine-rich repeat ribonuclease
IPR002341	75	16	4	>50	23	>200	2	>50	36	>200	HSF/ETS DNA binding
IPR000232	65	20	3	>50	24	>200	3	>50	5	>200	Heat shock factor
IPR006671	51	27	7	>50	51	117	9	49	41	182	Cyclin, N-terminus
IPR000595	39	35	1	>50	29	>200	4	>50	0	--	Cyclic nucleotide binding
IPR006663	62	21	14	>50	78	72	14	44	52	111	Thioredoxin domain 2
IPR006662	50	29	9	>50	78	82	10	48	45	163	Thioredoxin
IPR000408	39	36	12	>50	52	134	13	45	46	157	Regulator of chromosome condensation (RCC)
IPR005123	43	31	3	>50	114	46	10	48	0	--	2OG-Fe (II) oxygenase
IPR007090	58	26	1	>50	332	11	0	--	0	--	Leucine-rich repeat, plant
Underrepresented domains in *T. pseudonana*											
IPR001810	2	>50	6	>50	648	4	36	26	82	81	Cyclin-like F-box
IPR007087	13	>50	17	36	182	28	90	10	730	4	Zn-finger, C2H2 type
IPR004827	10	>50	5	>50	74	93	24	34	62	116	Basic-leucine zipper (bZIP)
IPR007114	32	39	12	>50	87	71	117	6	95	67	MFS (transporter)

Figure legends

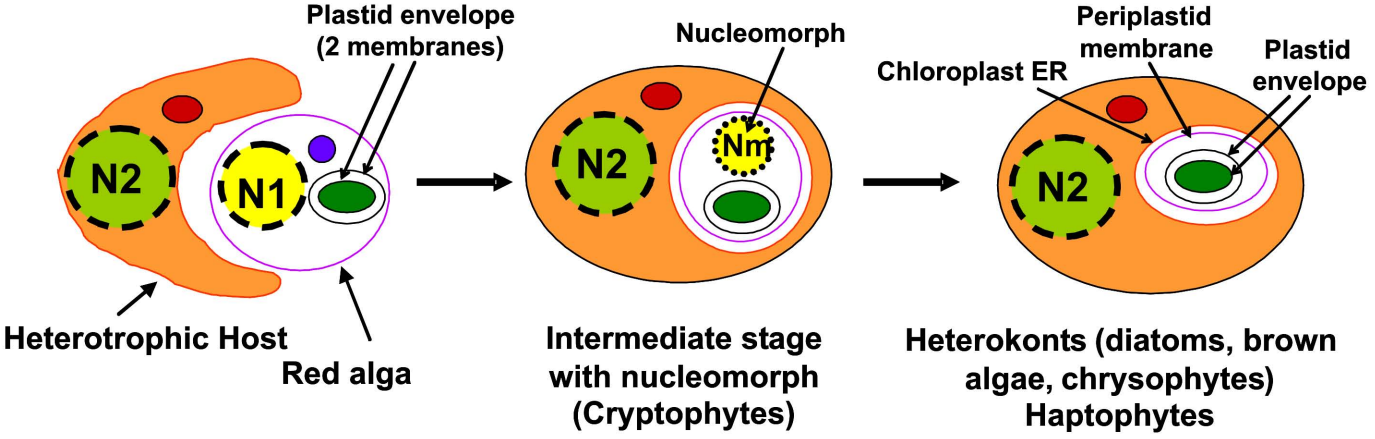
Figure 1. Origin of chloroplasts by secondary endosymbiosis (-es) involving a red algal endosymbiont. The endosymbiont nucleus (N1) disappeared after the transfer of many genes to the host nucleus (N2), except in cryptophyte algae, which have a relict nucleus (nucleomorph, Nm) between the plastid and the two new membranes derived from the red algal plasma membrane (periplastid membrane) and the endomembrane system of the host (chloroplast ER). Not shown are ribosomes attached to the outer surface of the chloroplast ER, or the continuity of chloroplast ER with the rest of the ER system. Nuclear-encoded plastid proteins are translated on these ribosomes and must therefore cross four membranes. The red algal mitochondrion (small blue circle) was also lost.

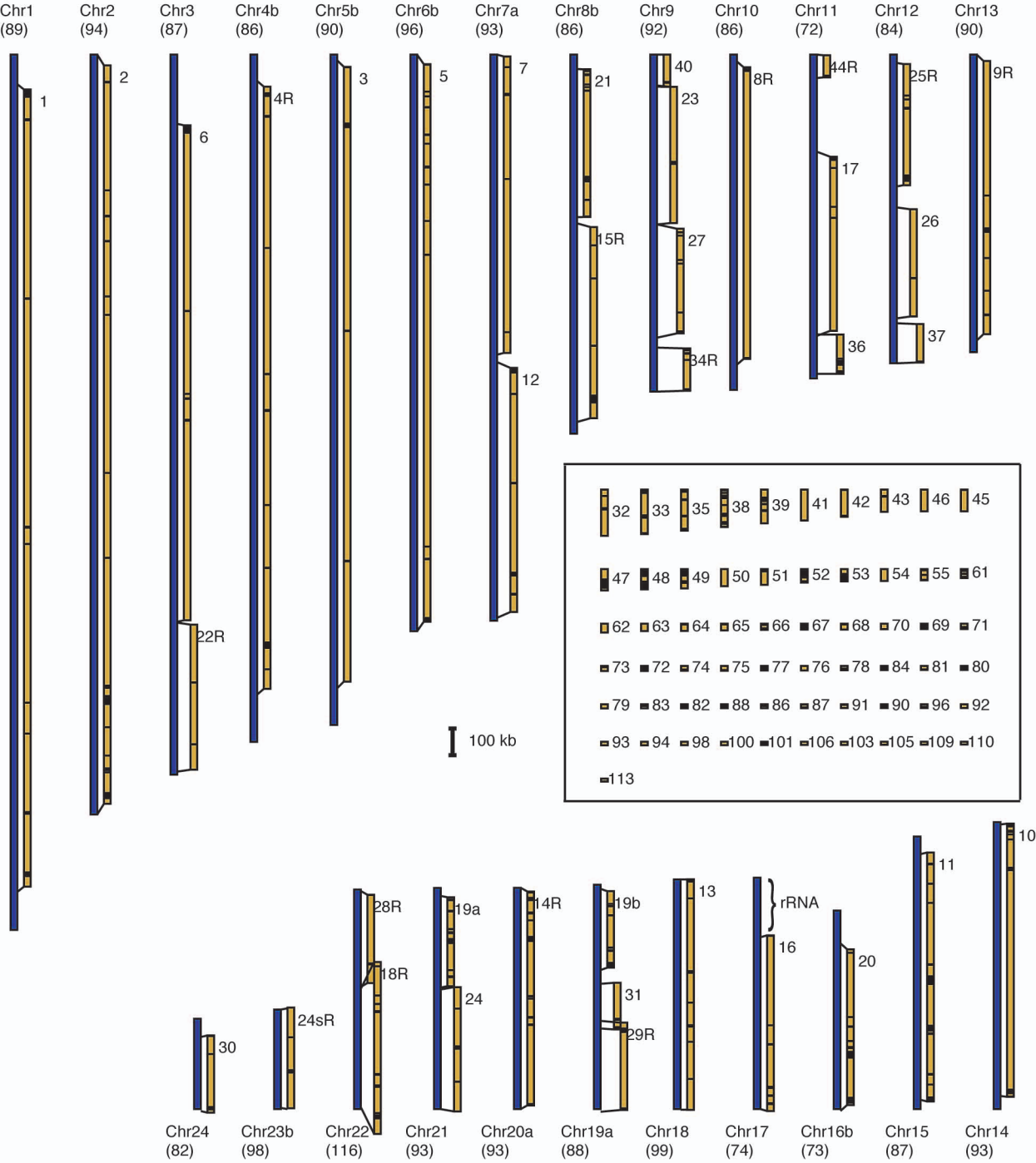
Figure 2. Schematic representation of 24 chromosomes in *T. pseudonana*, including 9 chromosomes with 2 distinguishable haplotypes (indicated with an ‘a’ or ‘b’ after chromosome number). Blue lines represent optical maps of chromosomes, gold-colored lines represent whole genome scaffold sequences with gaps shown as horizontal black lines. All sequences are shown to scale except that small gaps are shown with a minimum line width so all gaps are visible. Scaffolds are labeled by their number and are represented 5' to 3' down the page. Numbers in parentheses indicate the percentage of each chromosome's length spanned by mapped scaffolds. A letter 'R' after a scaffold indicates reverse complementation. Scaffolds assigned to the map based on comparison between the map and *in silico* scaffold digests are positioned adjacent to the matching segment of the map. Lines connect the ends of each placed scaffold with the end points

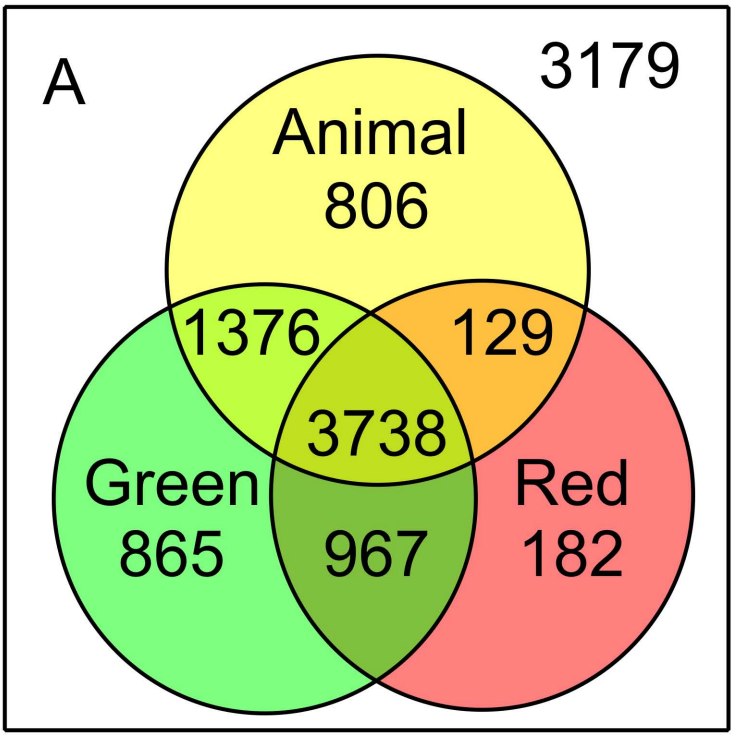
of the aligning segment of the map. Scaffolds that cannot be unambiguously placed on the map are shown in box.

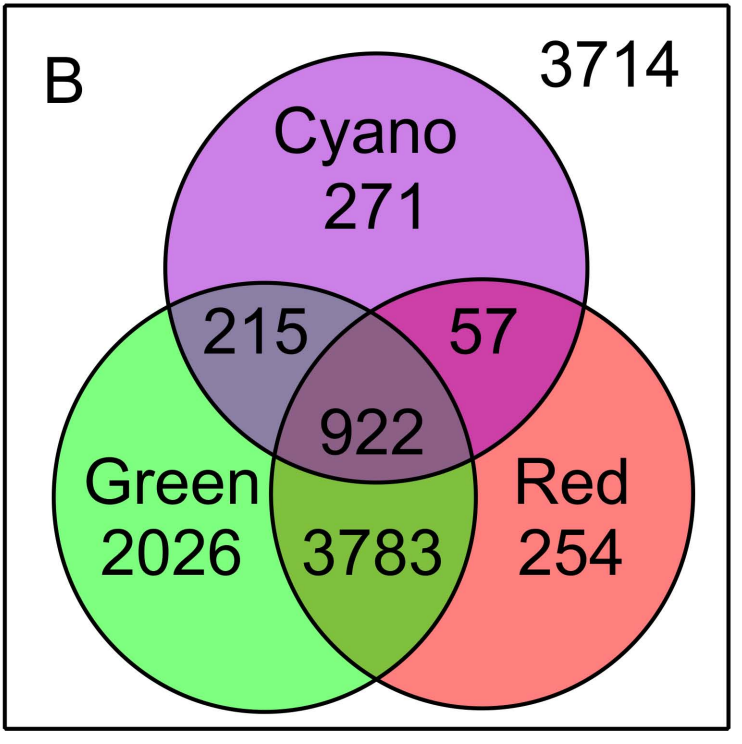
Figure 3. Diatom proteins with homologs (BLAST scores ≥ 100) in other organisms. A) Venn diagram of the distribution of *T. pseudonana* proteins with homology to proteins from *A. thaliana* (green), *C. merolae* (red) or *M. musculus* (animal). B) Same as A, but with *Nostoc sp*. PCC 7120 instead of *M. musculus.* Number outside the circles indicates the number of *T. pseudonana* proteins with no homology to the examined proteomes.

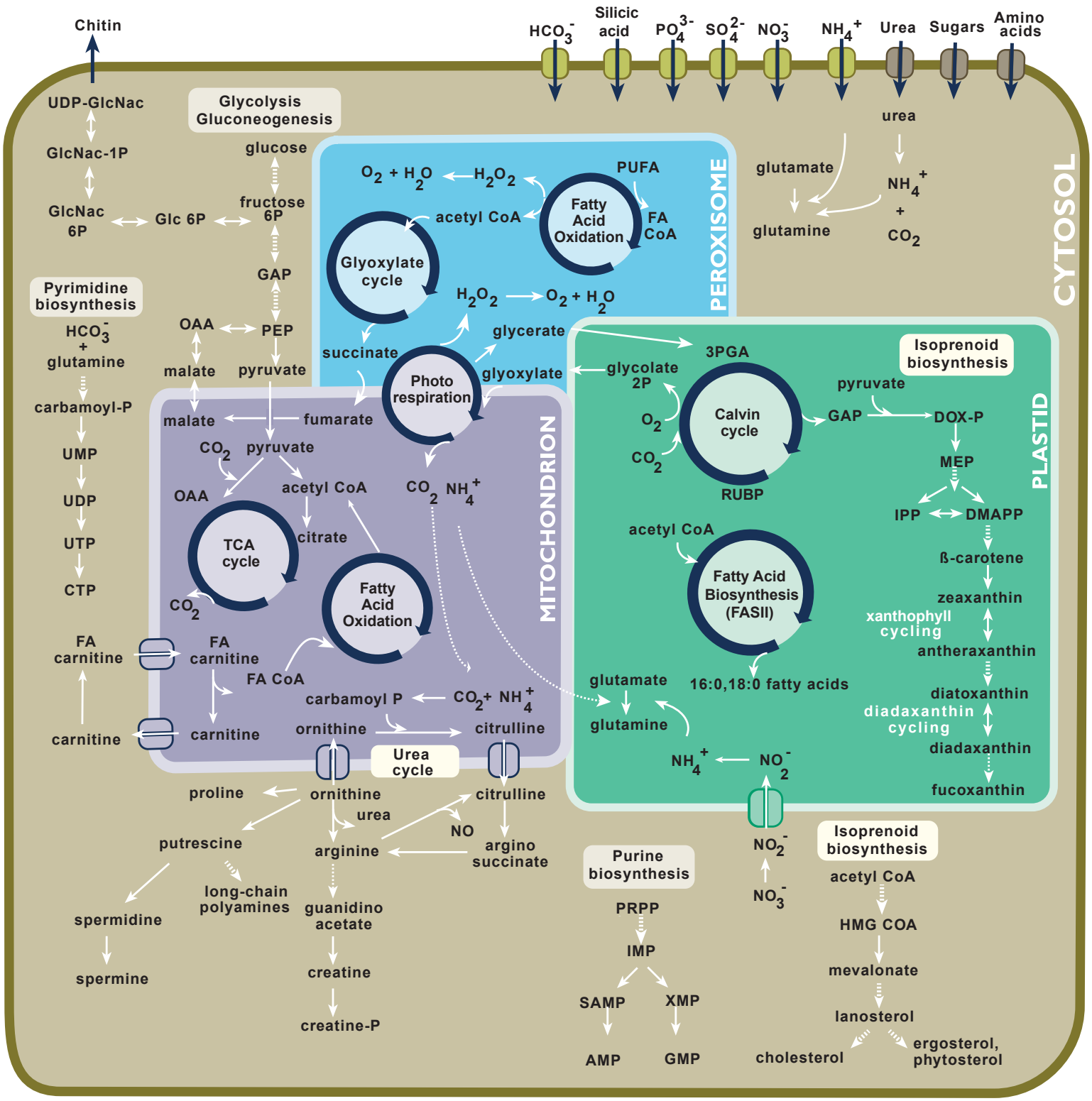
Figure 4. Novel combinations of metabolic pathways and key components of nutrient transport in *T. pseudonana*. Metabolic steps are represented by arrows: solid arrows indicate direct steps in a pathway, dashed arrows indicate that known multiple steps in a pathway are not shown, and dotted arrows represent hypothesized steps. Transporters or pathways localized to mitochondrion and plastid are supported by identification of targeting presequences. Pathways localized to peroxisome are based primarily on similar localization in other organisms (see text for details). Transporters of unknown localization are shown in the outer membrane.


Plastid envelope
(2 membranes)
N2
N1
Heterotrophic Host
Red alga
Nucleomorph
N2
Nm
Intermediate stage
with nucleomorph
(Cryptophytes)
Periplastid
membrane
Chloroplast ER
Plastid
envelope
N2
Heterokonts (diatoms, brown
algae, chrysophytes)
Haptophytes


Chr1 (89)
Chr2 (94)
Chr3 (87)
Chr4b (86)
Chr5b (90)
Chr6b (96)
Chr7a (93)
Chr8b (86)
Chr9 (92)
Chr10 (86)
Chr11 (72)
Chr12 (84)
Chr13 (90)
1
2
6
22R
4R
3
5
7
12
21
15R
40
23
27
34R
8R
44R
17
36
25R
26
37
9R
32 33 35 38 39 41 42 43 46 45
47 48 49 50 51 52 53 54 55 61
62 63 64 65 66 67 68 70 69 71
73 72 74 75 77 76 78 84 81 80
79 83 82 88 86 87 91 90 96 92
93 94 98 100 101 106 103 105 109 110
113
100 kb
10
11
rRNA
16
13
20
28R
18R
19a
24
14R
19b
31
29R
30
24sR
Chr24 (82)
Chr23b (98)
Chr22 (116)
Chr21 (93)
Chr20a (93)
Chr19a (88)
Chr18 (99)
Chr17 (74)
Chr16b (73)
Chr15 (87)
Chr14 (93)


A
3179
Animal
806
1376
129
3738
Green
865
967
Red
182


B
3714
Cyano
271
215
57
922
Green
2026
3783
Red
254


Chitin
HCO_3^-
Silicic acid
PO_4^{3-}
SO_4^{2-}
NO_3^-
NH_4^+
Urea
Sugars
Amino acids
CYTOSOL
UDP-GlcNac
GlcNac-1P
GlcNac 6P
Glc 6P
Glycolysis Gluconeogenesis
glucose
fructose 6P
GAP
PEP
pyruvate
OAA
malate
Pyrimidine biosynthesis
HCO_3^- + glutamine
carbamoyl-P
UMP
UDP
UTP
CTP
PEROXISOME
$O_2 + H_2O$
H_2O_2
PUFA
Fatty Acid Oxidation
FA CoA
acetyl CoA
Glyoxylate cycle
succinate
glycerate
glyoxylate
Photo respiration
MITOCHONDRION
fumarate
CO_2
NH_4^+
citrate
TCA cycle
Fatty Acid Oxidation
FA carnitine
carnitine
FA CoA
carbamoyl P
ornithine
citrulline
Urea cycle
proline
urea
NO
argino succinate
arginine
putrescine
spermidine
spermine
long-chain polyamines
guanidino acetate
creatine
creatine-P
glutamate
glutamine
urea
NH_4^+
CO_2
PLASTID
3PGA
glycolate 2P
Calvin cycle
RUBP
GAP
pyruvate
DOX-P
MEP
IPP
DMAPP
Isoprenoid biosynthesis
ß-carotene
zeaxanthin
xanthophyll cycling
antheraxanthin
diatoxanthin
diadaxanthin cycling
diadaxanthin
fucoxanthin
Fatty Acid Biosynthesis (FASII)
16:0,18:0 fatty acids
NO_2^-
NO_3^-
Purine biosynthesis
PRPP
IMP
SAMP
XMP
AMP
GMP
Isoprenoid biosynthesis
acetyl CoA
HMG COA
mevalonate
lanosterol
cholesterol
ergosterol, phytosterol