J. Phycol. **43,** 585–604 (2007)

DOI: 10.1111/j.1529-8817.2007.00342.x

IDENTIFICATION AND COMPARATIVE GENOMIC ANALYSIS OF SIGNALING AND REGULATORY COMPONENTS IN THE DIATOM *THALASSIOSIRA PSEUDONANA*[1]

Anton Montsant

Laboratory of Molecular Plant Biology, CNRS UMR 8186, Department of Biology, Ecole Normale Supérieure, 75230 Paris, France
Laboratory of Cell Signalling, Stazione Zoologica, Villa Comunale, I 80121 Naples, Italy

Andrew E. Allen

Laboratory of Molecular Plant Biology, CNRS UMR 8186, Department of Biology, Ecole Normale Supérieure, 75230 Paris, France
The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

Sacha Coesel

Laboratory of Cell Signalling, Stazione Zoologica, Villa Comunale, I 80121 Naples, Italy
Centre of Marine Sciences, University of Algarve, Campus de Gambelas, 8005-139 Faro, Portugal

Alessandra De Martino

Laboratory of Molecular Plant Biology, CNRS UMR 8186, Department of Biology, Ecole Normale Supérieure, 75230 Paris, France
Laboratory of Cell Signalling, Stazione Zoologica, Villa Comunale, I 80121 Naples, Italy

Angela Falciatore, Manuela Mangogna, Magali Siaut

Laboratory of Cell Signalling, Stazione Zoologica, Villa Comunale, I 80121 Naples, Italy

Marc Heijde, Kamel Jabbari, Uma Maheswari, Edda Rayko, Assaf Vardi

Laboratory of Molecular Plant Biology, CNRS UMR 8186, Department of Biology, Ecole Normale Supérieure, 75230 Paris, France

Kirk E. Apt

Martek Biosciences Corp., 6480 Dobbin Road, MD 21045, USA

John A. Berges

Department of Biological Sciences, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA

Anthony Chiovitti

School of Botany, University of Melbourne, Victoria 3010, Australia

Aubrey K. Davis, Kimberlee Thamatrakoln

Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA 92093, USA

Masood Z. Hadi

Lockheed Martin Corporation, Sandia National Laboratory, PO Box 969, MS-9951, Livermore, CA 94551, USA

Todd W. Lane

Biosystems Research Department, Sandia National Labs, Livermore, CA 94551-0969, USA

J. Casey Lippmeier

Martek Biosciences Corp., 6480 Dobbin Road, MD 21045, USA
Department of Biological Sciences, University of Hull, Hull HU6 7RX, UK

Diego Martinez

DOE Joint Genome Institute, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Micaela S. Parker

School of Oceanography, University of Washington, Seattle, WA 98195, USA

Gregory J. Pazour

Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, MA 01605, USA

Mak A. Saito

Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA

Dan S. Rokhsar

Center for Integrative Genomics, University of California Berkeley, Berkeley, CA, USA
DOE Joint Genome Institute, Walnut Creek, CA 94598, USA

E. Virginia Armbrust[2]

School of Oceanography, University of Washington, Seattle, WA 98195, USA

and

Chris Bowler[3]

Laboratory of Molecular Plant Biology, CNRS UMR 8186, Department of Biology, Ecole Normale Supérieure, 75230 Paris, France
Laboratory of Cell Signalling, Stazione Zoologica, Villa Comunale, I 80121 Naples, Italy

Diatoms are unicellular brown algae that likely arose from the endocytobiosis of a red alga into a single-celled heterotroph and that constitute an algal class of major importance in phytoplankton communities around the globe. The first whole-genome sequence from a diatom species, *Thalassiosira pseudonana* Hasle et Heimdal, was recently reported, and features that are central to diatom physiology and ecology, such as silicon and nitrogen metabolism, iron uptake, and carbon concentration mechanisms, were described. Following this initial study, the basic cellular systems controlling cell signaling, gene expression, cytoskeletal structures, and response to stress have been cataloged in an attempt to obtain a global view of the molecular foundations that sustain such an ecologically successful group of organisms. Comparative analysis with several microbial, plant, and metazoan complete genome sequences allowed the identification of putative membrane receptors, signaling proteins, and other components of central interest to diatom ecophysiology and evolution. *Thalassiosira pseudonana* likely perceives light through a novel phytochrome and several cryptochrome photoreceptors; it may lack the conserved RHO small-GTPase subfamily of cell-polarity regulators, despite undergoing polarized cell-wall synthesis; and it possesses an unusually large number of heat-shock transcription factors, which may indicate the central importance of transcriptional responses to environmental stress. The availability of the complete gene repertoire will permit a detailed biochemical and genetic analysis of how diatoms prosper in aquatic environments and will contribute to the understanding of eukaryotic evolution.

Key index words: **apoptosis; cell cycle; comparative genomics; cryptochrome; cytoskeleton; diatom; flagella; genome; GTP-ase; lateral gene transfer; myosin; oxidative stress; *Phaeodactylum tricornutum*; phytochrome; secondary endosymbiosis; *Thalassiosira pseudonana*; whole-genome analysis; xanthophyll cycle**

Abbreviations: **AP, apurinic/apyrimidinic; BV, biliverdin; CDKs, cyclin-dependent kinases; CSDs, cold-shock domains; ESTs, expressed sequence tags; GAPs, GTPase-activating proteins; GEFs, guanyl-nucleotide exchange factors; GPCR, G-protein-coupled receptors; HATs, histone acetyl transferases; HDACs, histone deacetylases; HK, histidine kinase; Hpt, histidine phosphotransfer; HSFs, heat-shock factors; IFT, intraflagellar transport; LRR, leucine-rich repeats; PCB, phycocyanobilin; PϕB, phytochromobilin; RCC, regulator of chromsome condensation; RIP, repeat induced point; RR, response regulator; SDV, silica deposition vesicle; SODs, superoxide dismutases; TF, transcription factor; VDE, violaxanthin de-epoxidase; ZEP, zeaxanthin epoxidase**

The contemporary oceans cover 70% of the surface of our planet, and even though the photosynthetic organisms living within them constitute only 1% of the earth's photosynthetic biomass, marine ecosystems are responsible for about one-half of global primary productivity (Field et al. 1998, Irigoien et al. 2002). Besides photosynthetic prokaryotes, the most successful organisms are thought to be a class of eukaryotic unicellular algae, known as diatoms.

[1]Received 10 October 2006. Accepted 19 January 2007.
[2]Author for correspondence: e-mail armbrust@ocean.washington.edu.
[3]Author for correspondence: e-mail cbowler@biologie.ens.fr.

On a global scale, diatoms are thought to contribute at least 20% of the annual primary productivity and to play major roles in key biogeochemical cycles (Smetacek 1999).

Two major groups of diatoms can be distinguished according to cell shape. Centric diatoms are radially symmetrical and are thought to have appeared ~200 mya, whereas the bilaterally symmetrical pennate diatoms evolved from the centric diatoms ~70 mya (Kooistra et al. 2003). A hallmark of diatoms is their beautifully ornamented silicified cell wall, known as the frustule. The precision of the nanoscale pattern and architecture of the frustule suggests that understanding diatom cell-wall biosynthesis eventually may be exploitable in nanotechnological applications (Parkinson and Gordon 1999, Lopez et al. 2005). Although in recent years some key components involved in silicon transport and precipitation have been described (Hildebrand et al. 1998, Kroger et al. 1999), and also identified in the *T. pseudonana* genome (Poulsen and Kroger 2004), the mechanisms by which the siliceous material of the frustule is laid down remain largely unknown.

The algal groups of the Heterokontophyta (e.g., diatoms and brown seaweeds), the Haptophyta (e.g., coccolithophorids), and the Cryptophyta (e.g., G*uillardia theta* D. R. A. Hill et Wetherbee) are collectively known as the chromist algae because of the colors conferred by their distinct pigment composition. Together with the Alveolates (dinoflagellate algae and apicomplexan parasites), these diverse eukaryotic groups likely arose from a common secondary endosymbiotic event in which a heterotrophic eukaryote engulfed (or was invaded by) a red eukaryotic alga more than one billion years ago (Patron et al. 2004, Yoon et al. 2004, Li et al. 2006). At the initial stage of this endosymbiosis, the complex cell likely contained at least five different genomes: the nuclear and mitochondrial genomes of the host cell and the nuclear, mitochondrial, and plastid genomes of the red alga. Over evolutionary time most of the cellular structures of the endosymbiont have been lost, and extant diatoms now contain plastids surrounded by four rather than two membranes and a nucleus and mitochondria that are derived from the original heterotrophic host. Diatom cells, therefore, have a number of features that make them highly divergent from the classical cellular structures of higher plants (e.g., a signal peptide-mediated import of proteins into plastids [Apt et al. 2002] or a hypothesized single-cell C4 carbon concentration mechanism [Reinfelder et al. 2000]).

The nuclear, plastid, and mitochondrial genome sequences of the centric diatom *Thalassiosira pseudonana* were recently reported (Armbrust et al. 2004). The nuclear genome is 34 Mb and consists of 24 chromosomes, and the plastid and mitochondrial genomes are 129 kb and 44 kb, respectively. A range of novel metabolic features have been described, such as genes encoding components for formation of silica-based cell walls, biosynthetic enzymes for several types of polyunsaturated fatty acids, and enzymes for utilization of various nitrogenous compounds, including a complete urea cycle (Allen et al. 2006). The *T. pseudonana* genome therefore encodes a unique combination of factors previously thought to be restricted either to photosynthetic or heterotrophic eukaryotes. In this study, exploration of the *T. pseudonana* genome has continued to provide insights into the mechanisms used by diatoms to regulate their cellular homeostasis. Specifically, we have investigated the recognizable gene families controlling cell signaling, transcription, cytoskeletal structures, cell cycle, and stress responses.

MATERIALS AND METHODS

Source of sequence data sets. The complete predicted proteome of *T. pseudonana*, strain CCMP 1335, is available from the Joint Genome Institute (JGI; http://genome.jgi-psf.org/thaps3/thaps3.download.html). In Figure 1 and Tables 1, 3, and S1 (in the supplementary material), the predicted proteomes of several organisms were analyzed along with the diatom predicted proteome for comparison. Their sources were http://merolae.biol.s.u-tokyo.ac.jp/download/ for the unicellular red alga *Cyanidioschyzon merolae* P. DeLuca, R. Taddei et L. Varno (Matsuzaki et al., 2004); http://www.broad.mit.edu/cgibin/annotation/neurospora/downloadlicense.cgi for the fungus *Neurospora crassa* (Galagan et al. 2003); http://genome.jgi-psf.org/ for the parasitic Stramenopiles *Phytophthora ramorum* and *Phytophthora sojae* (Tyler et al. 2006), the green alga *Chlamydomonas reinhardtii* P. A. Dang, the fungus *Phanerochaete chrysosporium*, and the animals *Ciona intestinalis* and *Xenopus laevis*; and http://www.ebi.ac.uk/proteome/index.html for the yeast *Saccharomyces cerevisiae*, the plant *Arabidopsis thaliana*, and *Caenorhabditis elegans*, *Drosophila melanogaster*, and *Mus musculus*.

Protein domain counts. The *T. pseudonana* predicted proteome and the proteome of six reference organisms were compared with the Conserved Domain Database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db = cdd) to identify InterPro domains by means of the RPS-Blast algorithm as described previously (Marchler-Bauer et al. 2003).

Analysis of intragenomic coding sequence similarity. The proteomes of the diatom and six other sequenced eukaryotes, for reference, were compared with themselves using BlastP with a BLOSUM62 matrix (Altschul et al. 1990) and an E-value threshold of E-20. Clusters of similar genes, loosely representing ''gene families,'' were defined as those groups of sequences that presented > 30% identity for at least 50% of their length to a common query. The number of genes within clusters of similar genes was plotted as a function of gene cluster size (Fig. 1a). The query–match pairs employed to build the gene clusters were classified by identity percentage intervals to visualize the level of similarity shared by the components of the gene clusters within each genome (Fig. 1b).

Counts of putative transmembrane kinases and leucine-rich-repeat proteins. Protein predictions that contained protein kinase (IPR000719) or leucine-rich repeat (LRR; IPR001611) domains were identified by means of the InterProScan output keyword search tool (http://www.ebi.ac.uk/InterProScan) implemented in the genome browsers of the organisms in study, and subsequently submitted to the TMHMM algorithm (Krogh et al. 2001) to detect putative transmembrane segments.

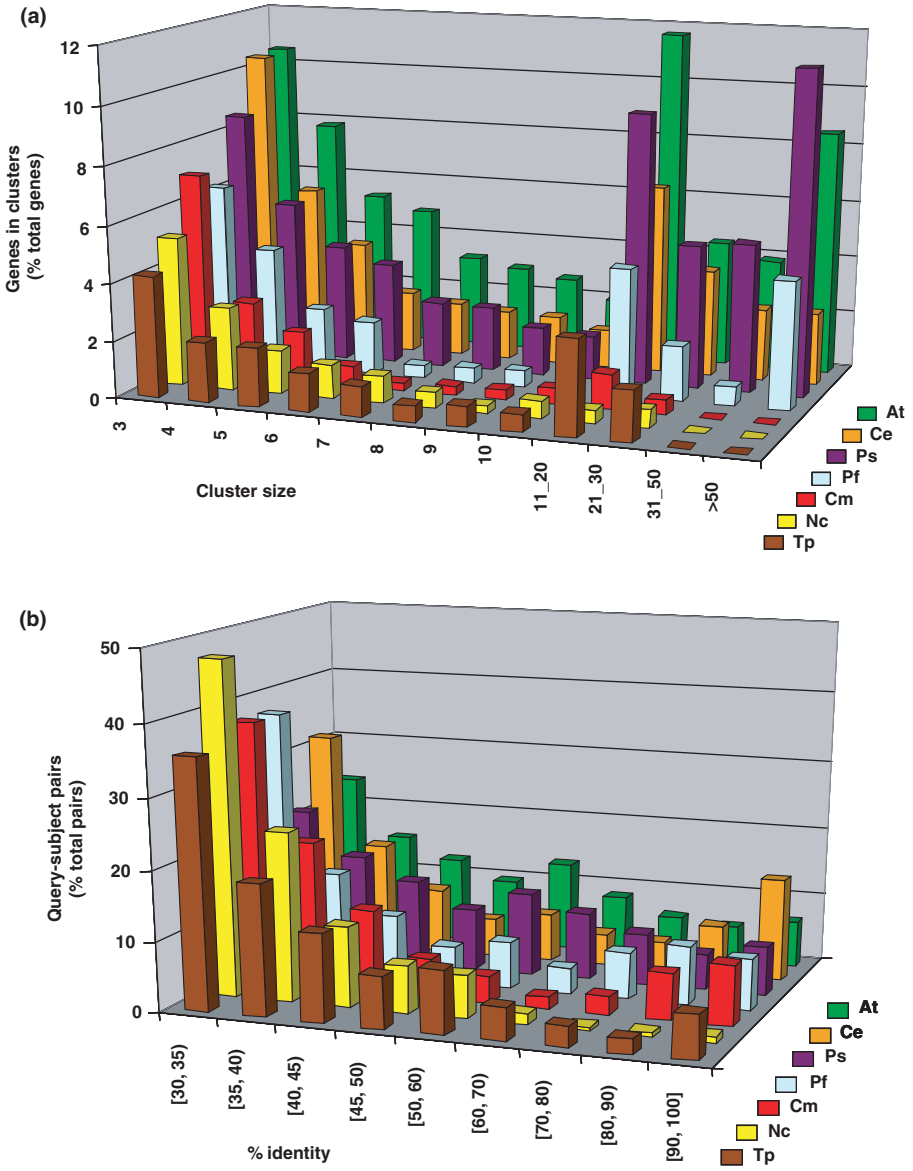

FIG. 1. Clusters of similar proteins in *Thalassiosira pseudonana* and comparison with six unicellular and multicellular model eukaryotes. Clusters of similar proteins within each genome were identified by means of BlastP self-comparison (see Materials and Methods). (a) Percentage of protein predictions within clusters sorted by cluster size; see Table S2 (in the supplementary material) for protein clusters of size two. (b) Identity percentage profile of all the query-subject pairs identified in every genome self-BlastP, with which gene clusters were defined (see Materials and Methods). The fraction of pairs showing 100% identity (and > 50% coverage) in each organism is indicated in Table S2. Tp, *Thalassiosira pseudonana*; Nc, *Neurospora crassa*; Cm, *Cyanidioschyzon merolae*; Pf, *Plasmodium falciparum*; Ps, *Phytophthora sojae*; Ce, *Caenorhabditis elegans*; At, *Arabidopsis thaliana*.

Phylogenetic trees. Phylogenetic trees were employed to gain further understanding of the function and/or evolutionary history of selected genes or gene families. Protein sequences were aligned with BioEdit (Hall 1999), and poorly conserved regions were eliminated. Neighbor-joining (NJ) trees (Tajima and Nei correction) with 1000 bootstrap replicates were subsequently derived with the Treecon package (Vandepeer and Dewachter 1994).

RESULTS AND DISCUSSION

Comparative overview of genetic and functional richness. Similarity to known sequences: More than 11,000 protein-coding genes are predicted in the *T. pseudonana* nuclear genome (Armbrust et al. 2004), and gene density (one gene every 3200 bp) is similar to that found in the comparably sized *Neurospora crassa* genome. Approximately 7200 genes are supported by *T. pseudonana* expressed sequence tags (ESTs), and another 1000 are supported by comparison with 12,000 ESTs from the distantly related pennate diatom *Phaeodactylum tricornutum* Bohlin (Maheswari et al. 2005, Montsant et al. 2005). Of the total predicted nuclear proteins, 5187 display similarity to known proteins from public data bases

TABLE 1. Putative kinase or leucine-rich repeat (LRR) proteins encoded in the genome of *Thalassiosira pseudonana* and in other complete eukaryotic genomes, and subsets that also contain a putative transmembrane (TM) domain.

	Total predicted proteins	Kinase	Kinase + TM	LRR	LRR + TM
Thalassiosira pseudonana	11,397	190	4[a]	92	2[a]
Phytophthora sojae	19,276	424	114	84	23
Phytophthora ramorum	16,066	377	99	77	22
Chlamydomonas reinhardtii	15,256	540	55	93	13
Arabidopsis thaliana	34,536	1189	481	543	297
Saccharomyces cerevisiae	7,098	119	2	11	0
Schizosaccharomyces pombe	5,010	111	1	10	0
Phanerochaete chrysosporium	10,048	172	0	16	0
Ciona intestinalis	14,526	261	37	116	19
Xenopus laevis	20,228	616	80	210	82

[a]See Table S3, annotations 7–12, in the supplementary material.

(score > 200, E-value < E-20) and 6661 have recognizable domains in the InterPro data base (Armbrust et al. 2004, and data not shown). To obtain a view of functional domain diversity within the *T. pseudonana* predicted proteome relative to other organisms, we divided these domains into five major categories according to whether they are equally abundant in other eukaryotes or whether they are over- or underrepresented in the diatom (Table S1 in the supplementary material). Such a comparison revealed a number of characteristics relevant to biological processes involving signaling, cell cycle, cytoskeleton, or transcription, all of which were studied in detail by in silico analysis and manual curation (see the following sections).

Within-genome coding sequence diversity: In studying ''functional richness,'' we further wondered how many different sequences existed within the total pool of predicted genes. The sequence diversity within the *T. pseudonana* predicted proteome was characterized using approaches similar to those previously employed for *N. crassa* (Galagan et al. 2003). The complete predicted proteome of the diatom was compared with itself by means of BlastP, and protein predictions sharing at least 30% amino acid identity over 50% of their lengths were grouped with each other. In doing so, a ''nonredundant'' set of genes was obtained in which most genes constituting a family or derived from recent duplications were grouped together. Clusters of similar proteins were defined in the same fashion within the predicted proteomes of another six unicellular and multicellular eukaryotes (Table S2 in the supplementary material). When gene number was plotted as a function of gene-cluster size, a similar profile was obtained for all the referenced microbes, with multicellular organisms showing a higher proportion of genes in clusters of large size (i.e., multicellular organisms have more multigene families, and these are composed of more members; Fig. 1a). The closest relative of the diatoms among the reference organisms, the oomycete *Phytophthora sojae*, showed a protein family-size profile similar to that of the multicellular organisms (Fig. 1a). However, the diatom and the oomycete have similar numbers of singleton genes (i.e., genes that do not belong to a gene family) and gene clusters (Table S2 in the supplementary material), indicating that the higher gene number of *P. sojae* does not translate into higher within-genome sequence diversity, possibly because gene number increased in the oomycete through multiplication of previously existing families.

In order to estimate the degree of similarity among genes within gene clusters for each of the genomes, the query–subject pairs with which clusters were defined were sorted by % identity, and the corresponding histogram was plotted. As shown previously (Galagan et al. 2003), a repeat-induced-point (RIP) mutation mechanism causes *N. crassa* to have a negligible amount of query–match pairs showing > 70% identity, a feature that was not detected in the diatom or in any other reference organism (Fig. 1b). Besides *Neurospora*, all eukaryotes displayed similar profiles, with *C. elegans* containing a larger fraction of very highly conserved pairs (> 90% identity). Up to 3% of the diatom query–match pairs showed 100% identity (and > 50% coverage), a fraction comparable to that of the flatworm (3.61%) and higher than in the remaining organisms (Table S2 in the supplementary material), some of which may correspond to very recent gene duplication events.

Signal perception. Membrane-bound receptors: Various types of sensing mechanisms have been discovered and extensively studied in plants and animals in the past decades, and a number of transmembrane receptor families can be identified by similarity searches. G-protein-coupled receptors (GPCR) are an extremely important class of receptor proteins in animal cells (FANTOM Consortium & RIKEN group 2002), but *Arabidopsis* is known to contain only one (Arabidopsis Genome Initiative 2000, Colucci et al. 2002). The *T. pseudonana* genome was found to encode two putative GPCRs, and consistently two putative heterotrimeric G proteins, through which GPCRs connect with downstream signal transduction pathways, were also identified (two genes encoding the α and β subunits, none encoding the γ subunit; Table S3, annotations 1–6, in the supplementary material).

On the other hand, InterPro domain searches revealed the presence of ~200 gene models containing a kinase domain and ~100 gene models with LRR (Table S1 in the supplementary material), both of which are common domains in membrane receptors of animals and plants. However, only four kinase-domain- and two LRR-domain-containing protein predictions were found to include putative transmembrane domains (Table S3, annotations

7–12, in the supplementary material). When these counts were obtained for a variety of other genomes for comparison, the oomycetes *P. ramorum* and *P. sojae* (close relatives of the diatoms) showed a similar total number of kinase- or LRR-domain-containing proteins, but the subset of them with a transmembrane region was found to be one to two orders of magnitude higher (Table 1). Complete genome sequences representative of other major phytoplankton taxa, such as coccolithophores, dinoflagellates, and pennate diatoms, will be necessary to conclude whether the utilization of transmembrane kinases and LRR receptors in chromalveolates is, as the parasitic oomycetes would suggest, comparable to that in animals, or whether it is sparse like in the *T. pseudonana* predicted proteome. On the other hand, these types of transmembrane receptors appear to be conspicuously rare in fungi, given that only three putative transmembrane kinases were detected in the three fungal species examined (Table 1).

Understanding the roles of the *T. pseudonana* putative GPCRs and membrane-bound kinases and LRR proteins, and whether these constitute the core of diatom membrane-bound receptors, will require further examination, but it seems likely that other receptor classes await identification. Predicted proteins that may have receptor functions are those of unknown function containing membrane-localizing domains (e.g., domains that define the prenylation amino acid motif CAAX are abundant in the *T. pseudonana* genome).

Photoperception: In addition to being a source of energy, light constitutes an important sensory stimulus. Terrestrial plants utilize three major classes of photoreceptors to sense ambient light in their environment: the red- or far-red-light-absorbing phytochromes and the blue-light-absorbing cryptochromes and phototropins (Falciatore and Bowler 2005). Furthermore, some green algae contain green-light-absorbing rhodopsins (Nagel et al. 2002, Sineshchekov et al. 2002). Green light persists to the greatest depths in coastal waters, but no rhodopsins were identified in the diatom genome. However, the diatom appears likely to use cryptochrome photoreceptors for perception of blue light, and a putative phytochrome was also identified (see below).

Cryptochrome photoreceptors are widespread throughout eukaryotes and are involved principally in regulating circadian rhythms (Cashmore 2003, Lin and Todo 2005). There are three major classes of cryptochromes, all of which are derived from DNA-repair photolyase enzymes. Plant cryptochromes show sequence similarity to the cyclobutane pyrimidine dimer (CPD)–photolyase class (Ahmad and Cashmore 1993), whereas animal cryptochromes are similar to the 6,4-photolyase class. A third class of cryptochrome, termed cry-DASH, was recently identified (Brudler et al. 2003). Neighbor-joining (NJ) segregation of the putative *T. pseudonana* cryptochromes (denoted Cryptochrome Photolyase Family1-4, or CPF1-4) suggested that TpCPF2, TpCPF3, and TpCPF4 belong to the cry-DASH subfamily, whereas TpCPF1 was similar to animal cryptochromes (Fig. 2). By contrast, no orthologs of the typical plant cryptochromes were identified in *T. pseudonana*. Experiments in progress indeed indicate that the diatom CPF1 class encodes a functional cryptochrome photoreceptor (S. Coesel, M. Mangogna, A. Falciatore, and C. Bowler, unpublished data).

The identification of a phytochrome photoreceptor is consistent with an earlier report demonstrating that diatoms can perceive red/far-red light (Leblanc et al. 1999) and contrasts with the absence of such a photoreceptor in the red alga *C. merolae*. Phytochromes were initially thought to be unique to terrestrial plants, but they were subsequently found in algae, cyanobacteria, and even some nonphotosynthetic organisms (Morand et al. 1993, Yeh et al. 1997, Falciatore and Bowler 2005). Phytochrome-like proteins are involved in different functions, such as pigment biosynthesis (Davis et al. 1999, Giraud et al. 2002), control of circadian rhythms (Schmitz et al. 2000), and phototaxis (Yoshihara et al. 2000). Nonetheless, the role of this class of photoreceptor in the marine environment remains unclear, and so the presence of a putative phytochrome in *T. pseudonana* provides a basis for further investigation. In an NJ tree, the putative *T. pseudonana* phytochrome appeared most similar to phytochrome-like sequences from the diatom *P. tricornutum* (M. Siaut, A. Falciatore, and C. Bowler unpublished data) and from two brown algal viruses (Delaroque et al. 2001), revealing a possible ''brown clade'' of phytochromes in which virus-mediated lateral transfer may have contributed to spread phytochrome-like genes among heterokonts (Fig. 3). Although such a hypothesis may sound surprising, it is striking that the only two brown algal viruses that have been sequenced to date both contain genes encoding putative phytochromes, even though viral genomes are typically highly reduced and contain only genes that are essential for viral infection and replication. However, the identification of putative phytochrome genes in brown algal viruses is reminiscent of the photosynthesis genes of cyanobacterial origin recently reported in the genome of marine cyanobacterial phages, proposed to increase the viability of the host during viral invasion (Lindell et al. 2005). Diatoms likely inherited their putative phytochrome-encoding gene from the heterotrophic host rather than from the red algal endosymbiont, as the putative brown clade appeared to be closer to the fungal clade than to the prokaryotic or plant orthologs (Fig. 3).

The *T. pseudonana* phytochrome contains an amino-terminal chromophore-binding domain followed by a carboxy-terminal signal-transduction

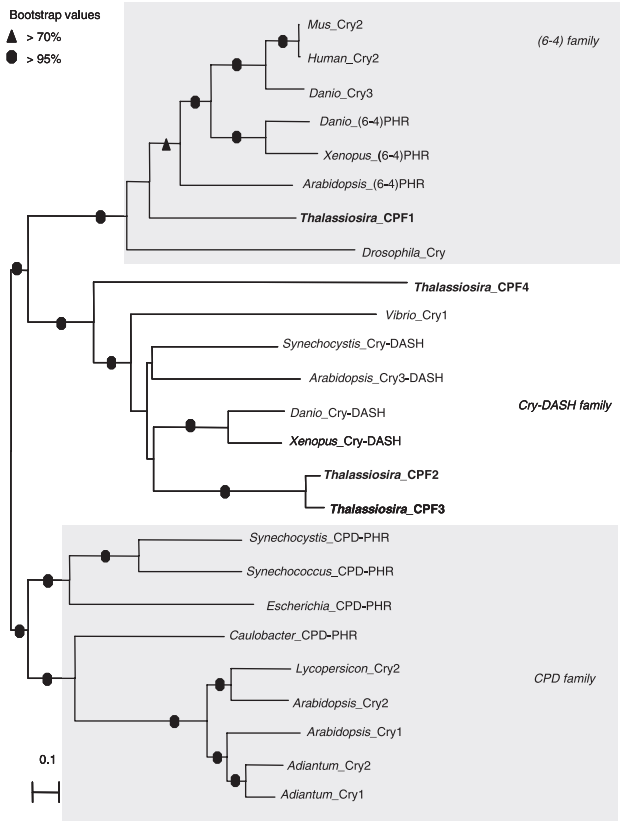

FIG. 2. Phylogenetic analysis of *Thalassiosira pseudonana* cryptochrome family sequences. Members of the cryptochrome/photolyase family from several eukaryotic and prokaryotic lineages were retrieved from GenBank and aligned with the *T. pseudonana* cryptochrome-like predicted proteins, and a highly conserved core of ~435 amino acids (Fig. S1 in the supplementary material) was employed to derive a neighbor-joining tree. The GenBank accession numbers of the proteins employed (or gene model identification for *T. pseudonana* sequences) are given in Table S4 (in the supplementary material).

domain related to the two-component histidine kinase (HK) phosphorelay system (Fig. 4; see below), and it may therefore function as a light-activated kinase. The presence of several genes encoding putative response-regulator (RR) proteins within the *T. pseudonana* genome (see below) indicates that the diatom phytochrome may transduce light signals via typical two-component signal-transduction pathways. The amino-terminal chromophore-binding domain of phytochromes is highly conserved across taxa, with defined histidine and cysteine residues usually being the attachment site of phytochromobilin (PϕB) in plants, phycocyanobilin (PCB) in cyanobacteria, or biliverdin (BV) in heterotrophic bacteria (Wagner et al. 2005). In the *T. pseudonana* putative phytochrome, none of the known chromophore-attachment cysteine residues are conserved, although the histidine residue is present, making the prediction of the chromophore difficult. Searches in the genome for genes encoding enzymes involved in chromophore biosynthesis left either possibility open, because two heme oxygenases, with which *T. pseudonana* can probably produce BV from heme, and a bilin reductase-like gene, which could mediate the production of PϕB or PCB, were identified (Table S3, annotations 13–15, in the supplementary material). Clarification of the chromophore that binds this gene product and its spectral properties appears necessary to understand the function of this gene. Nevertheless, the presence of a phytochrome gene in the genome of *T. pseudonana* is of particular interest, given the strong attenuation of red/far-red light wavelengths in the marine environment within the first few meters of the water column (Kirk 1992).

Intracellular signal transduction. Protein kinases: In eukaryotes examined to date, protein phosphorylation has emerged as a central event in signal transduction. As in *A. thaliana*, protein-kinase-encoding domains are the most abundant domains in the *T. pseudonana* genome (Table S1 in the supplementary material). However, about half of these are tyrosine kinases, which represent only a minor subclass in higher plants. At least two members of the mitogen-activated protein (MAP)–kinase cascade appeared to be conserved in the *T. pseudonana* genome (Table S3, annotations 16–17, in the supplementary material) by searches for conserved domains of the different members of the MAP-kinase family (Jonak et al. 2002). We also detected a putative signaling mechanism reminiscent of bacterial two-component HK systems. The HK-based phosphorelay systems have been described in prokaryotes (Mizuno 1998) and eukaryotes (Oka et al. 2002, Stephenson and Hoch 2002). In prokaryotes this mechanism generally involves only two components—an HK sensor and an RR—whereas in eukaryotes, a third partner—a histidine phosphotransfer (Hpt) protein—transmits the signal from the sensor to the regulator. The diatom genome contains at least eight predicted genes that may encode RRs in various configurations, along with three models containing HK domains, including the putative phytochrome (Fig. 4). However, no predicted Hpt-like proteins were found. Intriguingly, the putative response regulator TpRR7 includes a ''helix-turn-helix'' DNA binding LuxR domain (Fig. 4) that is most similar to that of the PhoB, NarL, and NtrC transcriptional regulators involved in modulating responses to phosphorus and nitrogen availability in bacteria (Rabin and Stewart 1993, Maris et al. 2002), suggesting the presence of a possible prokaryotic nutrient-sensing mechanism in *T. pseudonana*. Diatom genome and EST sequencing projects have recently uncovered several metabolic abilities of bacterial origin related to nutrient status (Allen et al. 2006).

Second messengers: The importance of intracellular calcium for responding to environmental signals was demonstrated *in vivo* in the pennate diatom *P. tricornutum* (Falciatore et al. 2000), and results

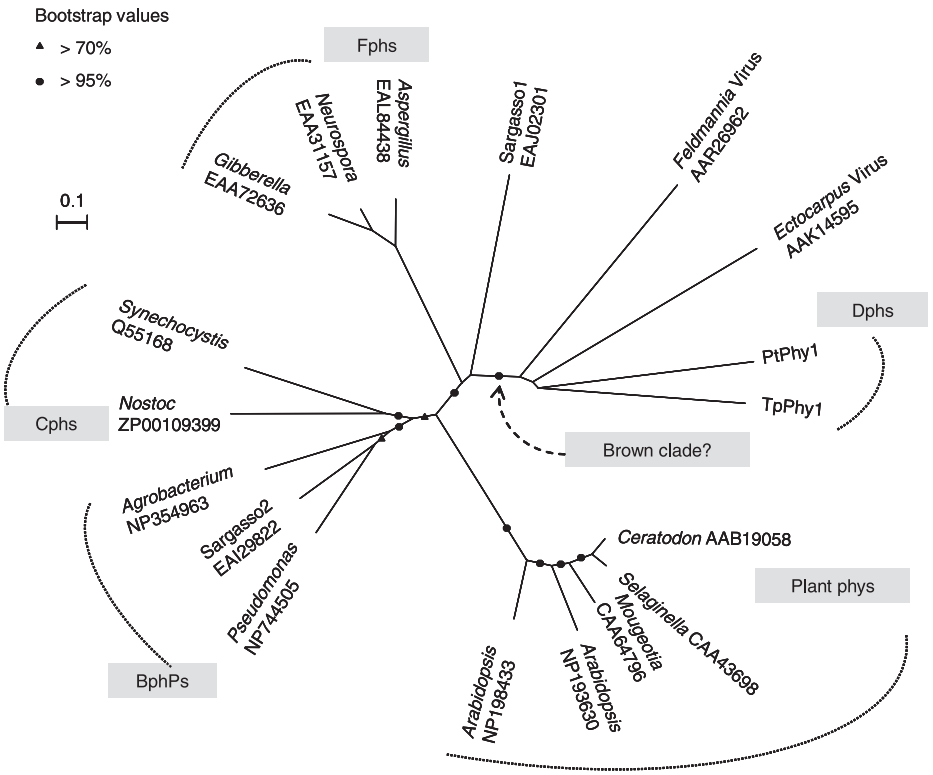

FIG. 3. Phylogenetic analysis of *Thalassiosira pseudonana* phytochrome. Phytochrome-like sequences from multiple species were downloaded from GenBank, and a highly conserved core of ~430 amino acids (Fig. S2 in the supplementary material), encompassing the GAF and histidine kinase domains, was used to build a neighbor-joining tree together with the putative *T. pseudonana* phytochrome (TpPhy) and a *Phaeodactylum tricornutum* phytochrome-like sequence (PtPhy; M. Siaut, A. Falciatore, and C. Bowler, unpublished data). BphPs, bacteriophytochromes; Cphs, cyanobacterial phytochromes; Dphs, diatom phytochromes; Fphs, fungal phytochromes; Plant phys, plant phytochromes. The identification numbers of the amino acid sequences employed are given under the species name. TpPhy1, gene ID 106207 on the JGI Thaps1 genome browser; PtPhy1, GenBank accession number DQ287774.

supporting the utilization of cyclic nucleotides are also available for pennate diatoms (Borowitzka and Volcani 1977, Scala et al. 2002). The utilization of these classical second messengers is likely to be conserved in the centric species *T. pseudonana*. In addition to putative calcium-regulated proteins, the centric diatom genome appeared to encode putative adenylyl cyclase and guanylyl cyclase genes (Table S3, annotations 18–31, in the supplementary material), as well as 39 gene models containing cyclic nucleotide-binding domains (Table S1 in the supplementary material), among which were putative protein kinases and ion channels and exchangers. The utilization of cAMP and cGMP by diatoms would constitute a major difference from land plants (Arabidopsis Genome Initiative 2000). Other metazoan signaling pathways, such as the nuclear steroid receptor, JAK/STAT, Wingless/Wnt, and hedgehog pathways, were not identified in the *T. pseudonana* genome.

Small GTPases: The *T. pseudonana* genome was determined to encode at least 26 classical small GTPases. These molecular switches of ~200 amino acids may have been central to the appearance of eukaryotic cells (Jekely 2003) and are known to be widely conserved across taxa, with five major subfamilies playing regulatory roles in a variety of cellular processes (Takai et al. 2001). Following an approach similar to that employed previously to classify the small GTPases of *Arabidopsis* (Vernoud et al. 2003), an NJ tree was derived to observe the clustering of *T. pseudonana* sequences with well-studied orthologs from yeast and humans. No members of the RHO subfamily (Rho, Rac, Cdc42) could be identified in the diatom by this procedure (Fig. 5), as was recently found to be the case for *Trypanosoma bruceii* (Field 2005). Moreover, no RHO-specific guanyl-nucleotide exchange factors (GEFs) or GTPase-activating proteins (GAPs) were found (e.g., there are no RhoGEF-specific Dbl domains and only one putative RhoGAP domain, both of which are abundant in yeast and animals). The RHO subfamily small GTPases are widespread key regulators of multiple phenomena that require cell polarity (Etienne-Manneville and Hall 2002), and a Rac protein was shown to be a key regulator of cell polarity in the brown macroalga *Fucus distichus* L. (Fowler et al. 2004), suggesting that diatoms may employ

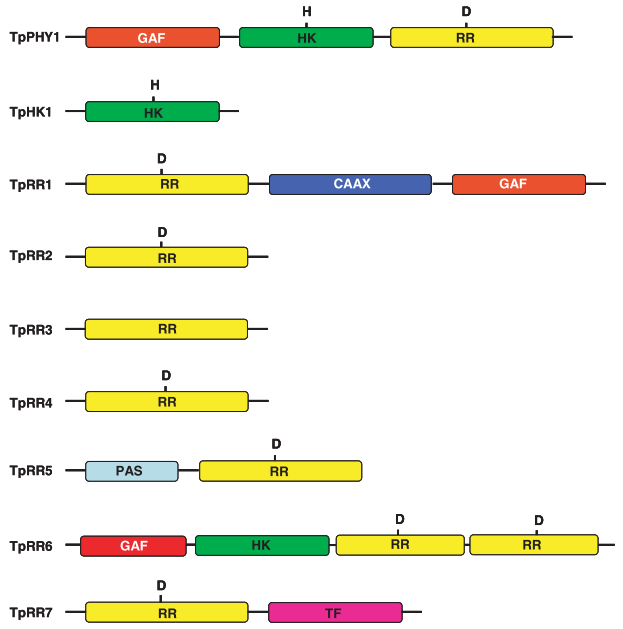

FIG. 4. Schematic of two-component signaling modules within predicted proteins identified in *Thalassiosira pseudonana*. GAF, sensor domain; HK, histidine kinase domain; RR, response-regulator receiver domain; CAAX, prenylation site (for membrane localization); PAS, protein-protein interaction domain; TF, transcription factor DNA-binding motif. Conserved histidine autophosphorylation and aspartate phosphoacceptor sites are indicated, where present. The depicted gene models have the following locus IDs: TpPHY1, TPS_22848; TpHK1, TPS_262298; TpRR1, TPS_20939; TpRR2, TPS_264268; TpRR3, TPS_263389; TpRR4, TPS_264726; TpRR5, TPS_3877; TpRR6, TPS_11819; TpRR7, TPS_33288.

novel mechanisms for controlling polarized processes such as phototactic movement and transport to the silica deposition vesicle (SDV; Zurzolo and Bowler 2001).

When small GTPase-encoding genes were searched in the genomes of other model eukaryotes and aligned and classified in the same fashion (Figs. S4 and S5 in the supplementary material), the apicomplexan *Plasmodium falciparum* was also found to lack RHO subfamily GTPases (Table 2; Fig. S5a in the supplementary material). However, the oomycete *P. sojae*, more closely related to the diatoms, appeared to have a probable Rac1 ortholog, along with another putative member of the RHO subfamily (Fig. S5b in supplementary material; Table 2). The apicomplexan, in addition, lacks a Ras1-like protein, a feature previously noted in plants (Vernoud et al. 2003), and this was found to be the case also for the green microalga *Ostreococcus tauri* C. Courties et Chret.-Dinet, but not for the red alga *C. merolae* (Fig. S5, c and d, in the supplementary material; Table 2). Relative to the small-GTPase profile of the reference organisms, the regulators of vesicle-trafficking Rab subfamily proteins constitute a high proportion of the total small GTPases of *T. pseudonana* (Table 2). Some unicellular parasites have recently been shown to have much larger Rab subfamilies, possibly reflecting a functional specialization for their specific lifestyles (Lal et al. 2005, Saito-Nakano et al. 2005). Two of the diatom small-GTPases clustered clearly with all Rab proteins, but not within any conserved subclade (Fig. 5). Such diatom-specific Rab variants may play a role in biogenesis or maintenance of membrane structures peculiar to diatoms or closely related organisms (e.g., SDV or plastid outer membranes).

Control of gene expression. Chromatin structure: The nuclear DNA of eukaryotes is typically organized around nucleosomes, which contain two subunits of each of the core histones H2A, H2B, H3, and H4. The *T. pseudonana* genome contains several genes encoding the core histones (as well as the linker histone H1), which are, as frequently observed in other organisms, organized together into gene clusters (Fig. S6 in the supplementary material). A histone H3.3 variant was also not clustered with other histones, in agreement with what is typically found in other eukaryotes (Malik and Henikoff 2003). When components regulating the epigenetic control of gene expression by modification of nucleosome packaging were studied, several putative histone acetyl transferases (HATs) of the GCN5, MYST, and CBP/p300 superfamilies, as well as histone deacetylases (HDACs) belonging to the Rpd3p, HDA1p, and Sir2p families, were identified (Table S3, annotations 32–49, in the supplementary material). Histone methylation is known to be mediated by SET domain proteins (Xiao et al. 2003), of which up to 30 may be encoded in the *T. pseudonana* genome (Table S3, annotations 50–55, in the supplementary material). Methylated and acetylated histone tails are typically recognized by chromodomains and bromodomains, identified, respectively, in 4 and 27 diatom predicted proteins. Four genes encoding putative DNA methyl transferases were also detected (Table S3, annotations 56–59, in the supplementary material), consistent with previous data showing that diatom DNA is methylated (Jarvis et al. 1992).

Transcription factors: Genes encoding the basal machinery for RNA polymerase I-, II-, and III-directed transcription are present within the diatom genome. To examine *T. pseudonana* transcription factor families, we searched for the InterPro (IPR) transcription factor (TF) domains recently summarized in the *A. thaliana* DATF data base (http://datf.cbi.pku.edu.cn; Guo et al. 2005) and compared the abundance of these domains with five species comprising unicellular and multicellular photosynthetic and heterotrophic organisms (Table 3). Twenty-four of the 44 TF domains identified in *A. thaliana* were present in *T. pseudonana*. Most of these are also present in the rest of the referenced eukaryotic organisms, with zinc fingers being of similar relative abundance in all six species. However, some domains widespread in eukaryotes are under- or

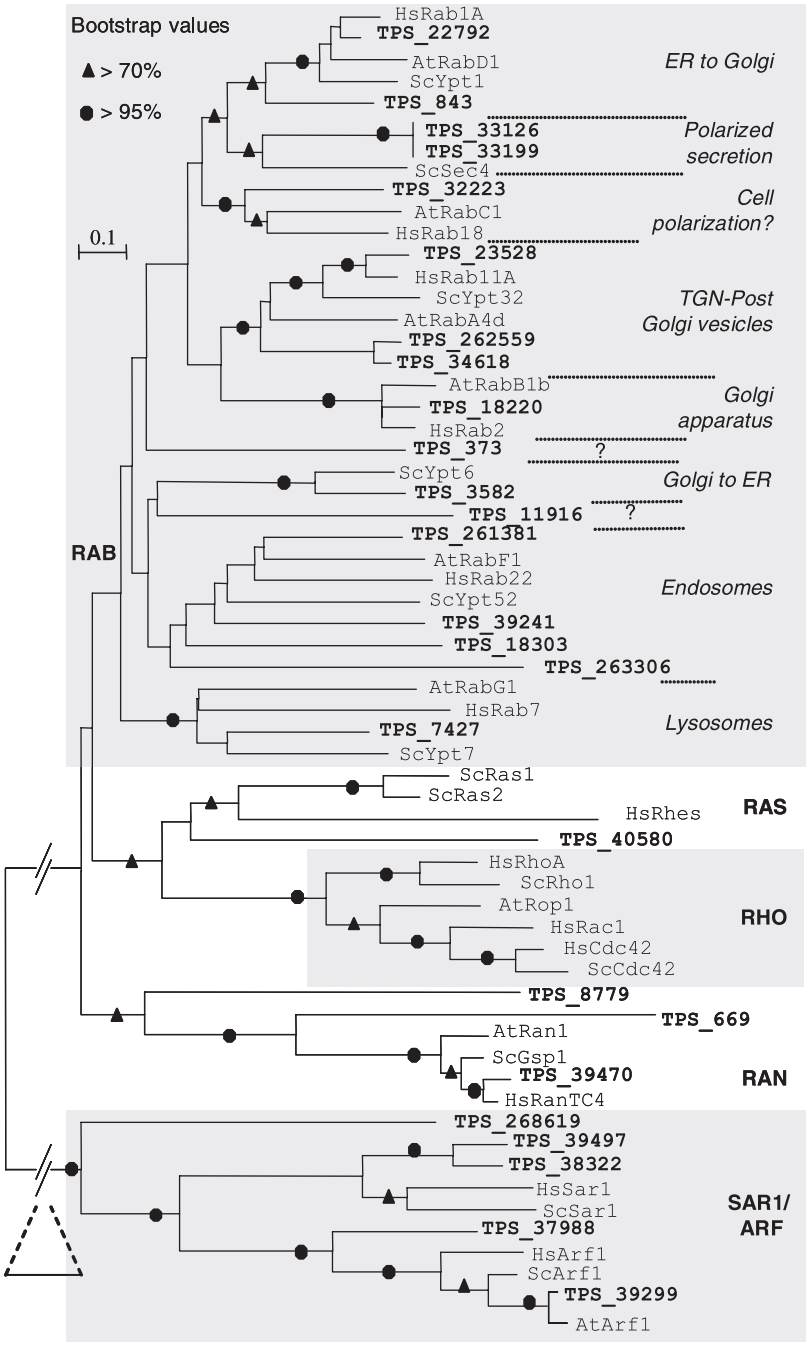

FIG. 5. Classification of the small GTPases of *Thalassiosira pseudonana* into five major subfamilies. The 26 diatom small GTPases were aligned with selected small GTPases from yeast (Sc), human (Hs), or plant (At) genomes, representing the five major subfamilies of small GTPases (Arf, Rab, Rho, Ras, Ran). Upon elimination of unconserved regions, a neighbor-joining tree was derived with sequences of ~150 amino acids (Fig. S3 in the supplementary material). Plant orthologs and site of action of Rab subfamily members are as labeled in Vernoud et al. (2003). The GenBank accession numbers of the sequences employed (or gene locus ID for *T. pseudonana* sequences) are given in Table S5 (in the supplementary material).

overrepresented in the *T. pseudonana* genome. The canonical C2H2 zinc finger and bHLH domains are infrequent in *T. pseudonana* when compared with most eukaryotes, and some domains identifiable in all reference organisms were not detected in the diatom (e.g., the GATA-type zinc finger).

A notable observation is that the *T. pseudonana* genome encodes a much larger number of putative heat-shock factors (HSFs) than any of the remaining reference organisms (Table 3). Although the counts presented here refer to domains rather than complete genes (e.g., IPR signatures may appear

TABLE 2. Small-GTPase subfamily composition in *Thalassiosira pseudonana* and other reference organisms.[a]

Organism	RAS	RHO	RAN	ARF/SAR1-like	RAB	Undefined	Total small GTPases
Thalassiosira pseudonana	1	0	3	5	17	–	26
Phytophthora sojae	3	2	2	12	22	2	43
Plasmodium falciparum	0	0	1	4	11	–	16
Saccharomyces cerevisiae[b]	4	5	2	7	11	–	29
Cyanidioschyzon merolae	1	2	1	5	6	1	16
Ostreococcu tauri	0	1	1	4	7	–	13
Arabidopsis thaliana[b]	0	11	4	21	57	–	93

[a]The alignments and trees derived to perform these counts are shown in Figures S4 and S5 (in the supplementary material).
[b]Previously reported counts for yeast and *A. thaliana* (Vernoud et al. 2003) are provided for comparison.

multiple times in the same gene), a recent report based on complete TF-like sequences also showed that plants and the diatom have a much larger number of HSFs than animals, fungi, or other unicellular organisms (Shiu et al. 2005). These HSFs have been shown to drive the heat-induced transcription of genes regulating homeostatic processes such as cell-wall maintenance, protein turnover, and detoxification (Yamamoto, Mizukami, and Sakurai 2005). Relative to comparably sized genomes, *T. pseudonana* also possesses a large number of cold-shock domains (CSDs; Table 3), initially discovered in a small number of *Escherichia coli* genes that are up-regulated upon sudden shift to lower temperatures. The CSD-containing proteins were then determined to be nucleic-acid-binding peptides conserved in prokaryotes and eukaryotes and were also found to participate in transcriptional and post-transcriptional regulatory functions unrelated to temperarure stress (Graumann and Marahiel 1998, Gualerzi, Giuliodori, and Pon 2003). The multiplicity of DNA binding domains related to temperature stress in the *T. pseudonana* genome opens the possibility of identifying expression networks of importance in diatom stress responses.

On the other hand, two domains, Myb-like SHAQKYF class and RWP-RK, appear in the green and red photosynthetic representatives but not in the heterotrophs. While the former is present in the diatom, the latter is absent. Similarly, the ethylene-responsive TF family and the Auxin/IAA protein family appear to be characteristic of the green lineage, and the diatom genome contains the former but lacks the latter. These observations underscore the chimeric nature of the diatom genome, in which features characteristic of photoautotrophs or of heterotrophs often appear intermingled in similar cellular processes and molecular functions (Armbrust et al. 2004). The probable importance of chromatin-level control of gene expression, discussed above, is reflected in the *T. pseudonana* IPR domain profile, as HMG and SET domains are the third and fifth most-important TF domains in *T. pseudonana* (Table 3), and the regulator of chromosome condensation (RCC) domain is also abundant (Table S1 in the supplementary material). The high abundance of proteins with DEAD box helicase domains (Table S1 in the supplementary material) could suggest that splicing during RNA processing may be an additional regulatory mechanism of importance.

Cytoskeletal structures. Microtubules and microfilaments: The cytoskeleton is known to be a major organizer of cellular activities, such as cell motility, cell division, and organelle transport in eukaryotes, and in diatoms it is an important intermediate of silica deposition (Pickett-Heaps et al. 1990). Cellular cytoskeletons are typically made of tubulin microtubules and actin microfilaments, and in animal cells, a third type, the intermediate filaments, has also been described. No genes encoding intermediate filament components could be found in the *T. pseudonana* genome, but many genes encoding components of tubulin- and actin-based cytoskeletons are present. The *T. pseudonana* genome contains one gene each encoding α- and γ-tubulin, and two encoding β-tubulin, which are expected to assemble into microtubules (Table S3, annotations 60–63, in the supplementary material). Many genes encoding microtubule-binding proteins (e.g., EBP1 and katanin) and microtubule-based motors (e.g., flagellar dynein, cytoplasmic dynein, and 26 predicted proteins with similarity to the motor domain of kinesin) were also detected. On the other hand, five genes encoding actin-like molecules were identified (Table S3, annotations 64–68, in the supplementary material), including conventional actin, expected to assemble into filaments; the centractin subunit of dynactin, involved in coupling cytoplasmic dynein to cargo; and Arp1, another component of the dynactin complex (Schroeder 2004). Putative orthologs of the actin-binding proteins profilin and severin, along with three formin-like peptides, which control actin rearrangements in several polarized processes (Pruyne et al. 2004), were identified. However, genes encoding pleckstrin-like and calponin domains, typical of actin-binding proteins and common in most eukaryotes, are rare in the diatom genome (Table S1 in the supplementary material).

TABLE 3. Plant transcription factor families of *Thalassiosira pseudonana*.[a]

		Tp		At		Cr		Cm		Sc		Mm	
IPR ID	**Description**	no.	*Rank*	no.	*Rank*	no.	*Rank*	no.	*Rank*	no.	*Rank*	no.	*Rank*
IPR000232	Heat-shock factor	65	1	32	26	2	23	3	13	6	12	10	23
IPR001005	Myb-related domain/Trihelix	39	2	659	2	42	2	48	1	23	3	76	12
IPR000910	High mobility group	31	3	22	31	9	14	–		10	10	112	7
IPR001965	Cys-rich zinc finger/Plant homeodomain finger	29	4	222	5	41	4	16	3	17	5	143	6
IPR001214	SET domain	28	5	72	15	42	1	7	7	2	20	76	13
IPR001356	Homeobox domain	17	6	156	8	5	17	7	6	12	7	400	3
IPR000571	RING finger protein/Zinc finger domains	15	7	95	13	19	6	6	9	7	11	84	11
IPR001092	Basic/helix-loop-helix dimerization region	13	8	264	3	41	3	14	4	–		198	5
IPR007087	Zinc-finger C2H2 domain	13	8	248	4	28	5	17	2	57	1	1138	1
IPR000047	Helix-turn-helix motif	11	10	33	24	–		–		1	21	93	8
IPR004827	Basic-leucine zipper (bZIP) motif	10	11	128	11	16	7	5	10	20	4	91	9
IPR001789	HisAsp phosphorelay signaling, similar to Myb	10	11	67	16	13	11	–		5	13	–	
IPR003347	Jumonji, JmjC	10	11	27	27	8	15	4	12	5	14	49	14
IPR001471	DNA-binding /ethylene-responsive element binding proteins	9	14	202	6	15	8	–		–		–	
IPR005172	Tesmin/TS01-like, CXC domain	8	15	16	33	3	21	2	17	–		4	24
IPR000253	Forkhead-associated domain	7	16	23	28	9	13	3	15	16	6	34	16
IPR006447	Myb like SHAQKYF class	5	17	129	10	6	16	9	5	–		–	
IPR002059	Cold-shock protein, DNA binding	5	17	5	>40	3	19	–		–		14	21
IPR003958	CCAAT-binding factor	4	19	38	20	4	18	2	18	4	16	10	23
IPR000007	Tubby family proteins	3	20	22	29	3	19	–		11	8	205	4
IPR004022	DNA-binding homeobox and Different Transcription factors	1	21	10	39	–		1	20	2	18	4	24
IPR003316	Transcription factor E2 F/ Dimerisation Partner (TDP)	1	21	22	30	2	24	3	14	–		30	17
IPR003349	Jumonji, JmjN	1	21	13	34	3	20	1	21	3	17	13	22
IPR002100	MADS-box	1	21	183	7	1	27	1	22	4	15	16	20
IPR009057	Homeodomain-like	1	21	764	1	1	29	1	24	28	2	500	2
IPR000315	Zn-finger, B-box	–		39	19	2	26	2	16	–		86	10
IPR000679	Zn-finger, GATA type	–		41	18	13	10	6	8	10	9	22	19
IPR001606	AT-rich interaction domain	–		12	36	2	25	4	11	2	19	27	18
IPR002910	Floricaula/leafy	–		2	42	–		–		–		–	
IPR003035	Plant regulator RWP-RK	–		18	32	13	9	1	23	–		–	
IPR003311	AUX/IAA protein family	–		37	21	2	22	–		–		–	
IPR003340	Transcription factor B3	–		123	12	–		–		–		–	
IPR003441	No apical meristem (NAM) protein	–		148	9	–		–		–		–	
IPR003657	DNA-binding WRKY	–		75	14	1	28	–		–		–	
IPR003851	Zn-finger, Dof type	–		37	22	–		–		–		–	
IPR004333	SBP domain	–		36	23	12	12	–		–		–	
IPR005202	GRAS transcription factor	–		64	17	–		–		–		–	
IPR005333	TCP transcription factor	–		32	25	–		–		–		–	
IPR006510	Putative zinc finger domain, LRP1	–		11	37	–		–		–		–	
IPR006511	LRP1, C-terminal	–		10	38	–		–		–		–	
IPR006779	DNA-binding protein S1FA	–		4	41	–		–		–		–	
IPR006780	YABBY protein	–		12	35	–		–		–		–	
IPR006957	Ethylene insensitive 3	–		6	40	–		–		–		–	

[a]The InterPro (IPR) domains characteristic of transcription factors identified in *Arabidopsis thaliana* (Guo et al. 2005) were searched in the *T. pseudonana* predicted proteome; sequences are ranked according to their abundance in the predicted proteomes of the respective reference organism (1 is the most abundant). Tp, *Thalassiosira pseudonana*, At, *Arabidopsis thaliana*; Cr, *Chlamydomonas reinhardtii*; Cm, *Cyanidioschyzon merolae*; Sc, *Saccharomyces cerevisiae*, Mm, *Mus musculus*.

Myosin motors: The *T. pseudonana* genome encodes at least 12 putative myosin-like motor proteins, the major drivers of cargo transport along actin filaments. The vast myosin superfamily described so far in eukaryotic cells was classified into 17 types based on their conserved motor head domain sequences (Sellers 2000), and recently as many as 37 classes were discriminated by also considering domain composition of the myosin tails, a diversity that was proposed to support a major unikont/bikont split early in eukaryote evolution (Richards and Cavalier-Smith 2005). The *T. pseudonana* myosins appeared to cluster in three major groups, termed myosin types A–C, in an NJ tree of bikont head domains (Fig. 6a). Diatom and oomycete type-A myosins are most similar to plant orthologs when similarity searches in public data bases are performed. Interestingly, four of the diatom plant-like myosins appeared to lack a dimerization domain, a feature in common with a monomeric

class XIII myosin of the green alga *Acetabularia cliftonii*, proposed to be involved in organelle maintenance (Vugrek et al. 2003). The chromalveolate type-B and type-C myosins defined independent clades (Fig. 6) and appeared most similar to orthologs from opisthokonts and other heterotrophic eukaryotes in multiple similarity searches. It is therefore possible that the chromalveolate lineages inherited type-A myosins from the red algal endosymbiont and types B and C from the host nucleus. The fact that 9 out of the 12 diatom myosins are type A (Fig. 6b), compared with only 3 out of 15 in *Phytophthora* (not shown), a close relative that lost the plastid, further supports such a view.

Flagella: Flagella are produced only during sperm formation in centric diatoms and are unusual in remaining assembled and functional during chromosome separation and subsequent division (Manton et al. 1970). As expected from their 9 + 0 axonemal structure, no central pair or radial spoke proteins are present. Most eukaryotic cilia and flagella assemble via translocation of large protein complexes along microtubules in a process known as intraflagellar transport (IFT; Rosenbaum and Witman 2002). Genes encoding IFT complex B subunits were found, along with genes for the kinesin II motor that transports IFT cargo to the ciliary tip (Table S3, annotations 69–80, in the

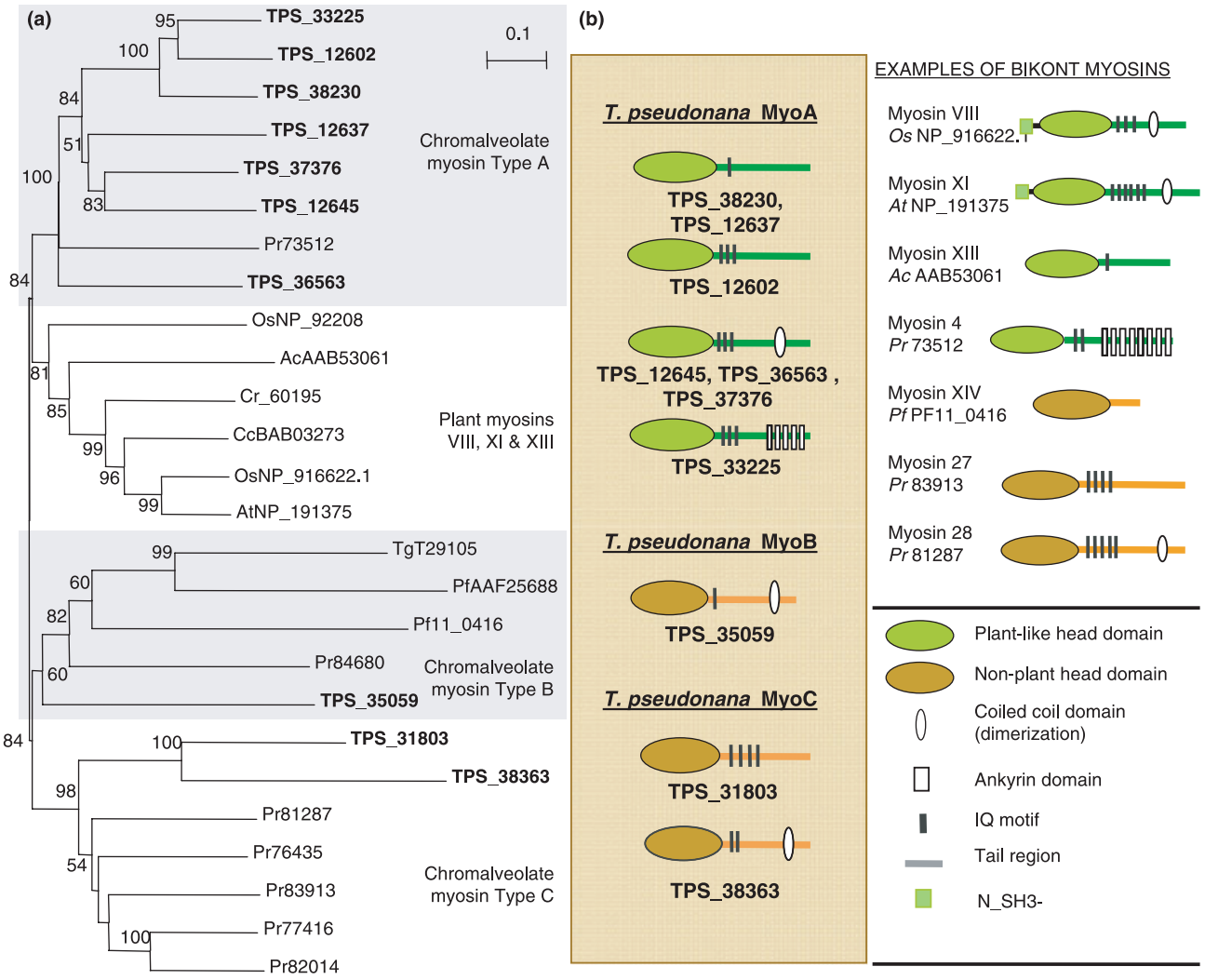

FIG. 6. The myosin family of *Thalassiosira pseudonana*. (a) A neighbor-joining tree of the *T. pseudonana* myosins. Myosin-like proteins were identified in the *T. pseudonana* genome browser, and their conserved head domains (390 amino acids, Fig. S7 in the supplementary material) were aligned with similar sequences from species of the green lineage, the apicomplexans, and the oomycetes. Bootstrap values above 50% (of 1000) are shown. (b) Schematic of the myosin-like proteins of *T. pseudonana*. The domain content of the myosin-like protein predictions of the diatom was studied by means of the InterProScan tool. Myosin types are expressed as roman numerals after Sellers (2000) and as arabic numbers after Richards and Cavalier-Smith (2005). The genome locus IDs or GenBank accession numbers are indicated following the two-letter species code. Ac, *Acetabularia cliftonii*; At, *Arabidopsis thaliana*; Cc, *Chara corallina*; Cr, *Chlamydomonas reinhardtii*; Os, *Oryza sativa*; Pf, *Plasmodium falciparum*; Pr, *Phytophthora ramorum*; Tg, *Toxoplasma gondii*; Tp, *T. pseudonana*.

supplementary material). However, no components of a cell-body-directed IFT motor (heavy and light-intermediate chain of cytoplasmic dynein 2/1b) or IFT complex A subunits were found, suggesting that *T. pseudonana* either uses a different motor for retrograde transport or that its IFT does not run in both directions. The lack of IFT complex A proteins in a ciliated species is surprising and supports the idea that IFT complex A and complex B have fundamentally different functions (Piperno et al. 1998).

Tracking absolute and subjective time. Circadian clock: The circadian clock allows organisms to adapt to daily changing light conditions by anticipating the night–day transitions and is therefore of particular importance for photosynthetic organisms. Although the clocks of different eukaryotic organisms consist of different individual components, the mechanism is generally based on a core oscillator that receives inputs from sensors of extracellular stimuli (e.g., photoreceptors) and signals downstream to a number of output pathways that control cell cycle and production of photosynthesis-related peptides (Salome and McClung 2004, Unsal-Kacmaz et al. 2005). Although no complete conserved circadian oscillator was distinguished from the *T. pseudonana* genome, domains common in plant core oscillators (e.g., MYB, PAS, CCT motif, B-Box or F-Box; Somers 2001) appear in multiple predicted proteins, and a putative ortholog of the Circadian Phase Modifier CpmA, a member of a cyanobacterial output pathway thought to regulate the photosystem proteins psbAI, psbAII, and KaiA (Katayama et al. 1999), was identified (Table S3, annotation 81, in the supplementary material). Together with the aforementioned photoreceptors, likely major input sources for the core oscillator, the fragmentary picture obtained from the genome sequence may provide a starting point for experimental work.

Life cycle: The basic components that control progression through the cell cycle, such as cyclins and cyclin-dependent kinases (CDKs), are very highly conserved in all eukaryotes, and we found this to be the case also in the diatom. The CDKs can be discriminated mainly on the basis of their cyclin-binding domain motif. According to this criterion, we identified seven different CDKs in *T. pseudonana*, three of which were observed to be highly related to known CDKs from alveolates and opisthokonts (Fig. 7). In addition, up to 48 cyclin-like proteins were identified, along with several members of the retinoblastoma pathway and a probable POLO-kinase, but not a Cdc25 gene. Interestingly, the *T. pseudonana* genome encodes an ortholog of the *Schizosaccharomyces pombe* Bub1 protein, which controls the metaphase–anaphase checkpoint (Bernard et al. 1998) and is absent in the *A. thaliana* genome.

Although a sexual cycle has not been described for *T. pseudonana*, other species within the same genus (e.g., *T. weissflogii*) are known to go through a meiotic cycle to produce male and female gametes (Round et al. 1990). The *T. pseudonana* genome encodes some meiosis-specific components, such as the spindle checkpoint regulator Mad2, the primase for replication initiation, the RP-A single-strand-binding protein, and the meiotic recombination components spo11, mre11, rad50, and rad51 (Table S3, annotations 82–87, in the supplementary material). Although no matches were detected for Rec8 or Spo13, known to be involved in double-strand break repair in organisms that undergo recombination (Lee et al. 2002, Kitajima et al. 2003), the presence of meiotic components in combination with flagellar genes indicates that *T. pseudonana* has been a sexual species in its recent past.

Stress responses. DNA repair: Genomic integrity is subject to constant threats by mutagenizing chemicals, photons, and heat, which can lead to a variety of perturbations in normal DNA structure (e.g., modified bases, abasic sites, intra- and interstrand cross-links, single- and double-strand breaks). Given the variety of threatening agents and potential lesions, a corresponding diversity of DNA-repair mechanisms exists in eukaryotes. Members of major DNA-repair systems were identified in the diatom genome, including putative class I and class II photolyases, nucleotide excision pathway components, and the apurinic/apyrimidinic (AP) endonuclease EXOII subfamily (Table S3, annotations 88–98, in the supplementary material). None of the latter, however, contained the Ref1 domain known to be involved in the DNA binding activity of several oncoproteins (Xanthoudakis and Curran 1992, Xanthoudakis et al. 1994). In addition, the *T. pseudonana* genome encodes putative components of the Ku70–Ku80 complex (Table S3, annotations 99–100, in the supplementary material), a strong discriminator of the paired and mispaired residues at the 3′ position of DNA nicks, but no other conserved members of the nonhomologous end-joining pathway were detected.

Oxidative stress: Like all eukaryotes, diatoms utilize a number of means to avoid damage from reactive oxygen species generated during photosynthesis and aerobic metabolism. One of the most important mechanisms to rapidly dissipate excessive excitation energy in higher plants and green algae involves the reversible conversion of violaxanthin and zeaxanthin, known as the xanthophyll cycle. In addition to this cycle, diatoms possess the diadimexanthin (DD) cycle, a second xanthophyll cycle that comprises the interconversion of diadinoxanthin and diatoxanthin and is thought to be responsible for the rapid and efficient photoprotection measured in planktonic diatoms (Lavaud et al. 2002). While green photosynthetic organisms such as *A. thaliana* or *C. reinhardtii* have one gene each encoding violaxanthin de-epoxidase (VDE) and zeaxanthin epoxidase (ZEP; Lange and Ghassemian 2003, Lohr et al. 2005), in the

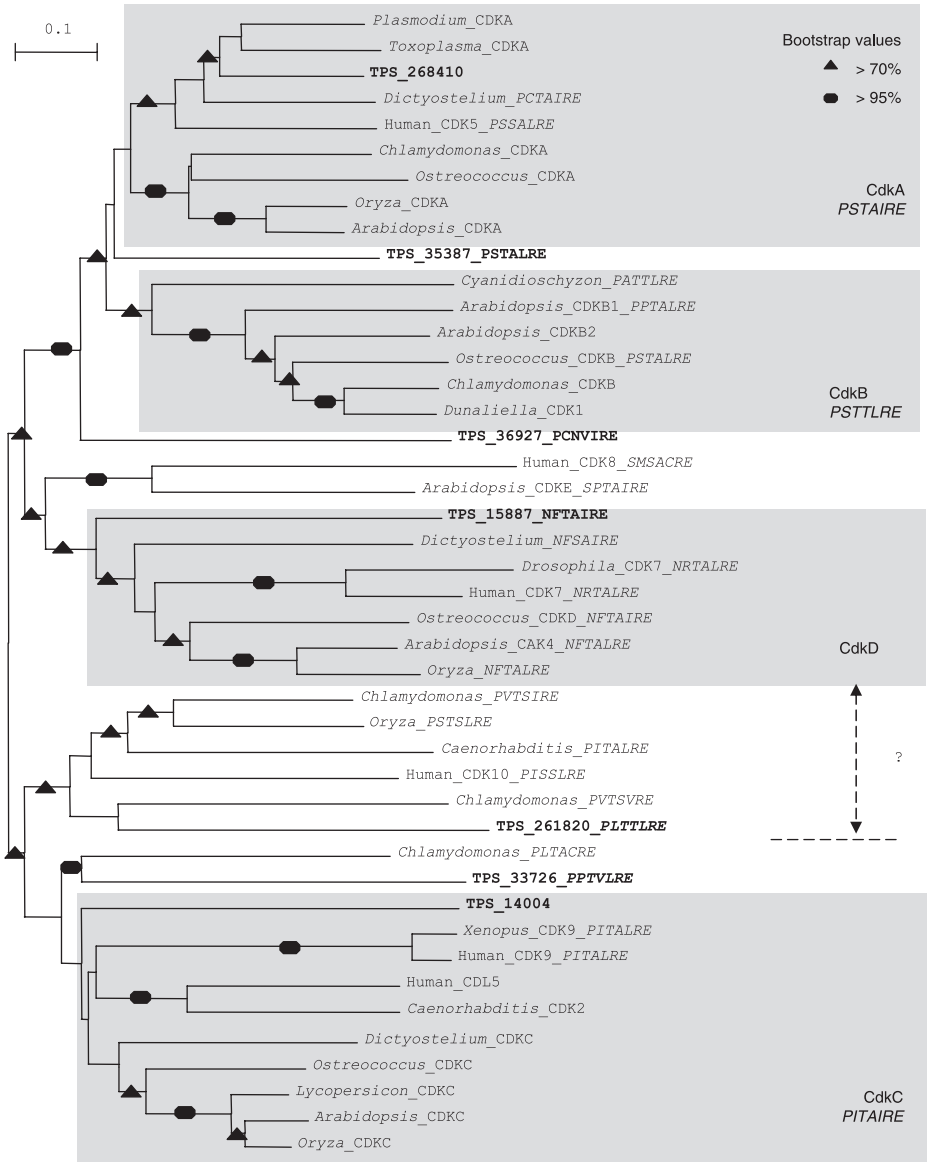

FIG. 7. Phylogenetic analysis of the cyclin-dependent kinase (CDK)–like proteins encoded in the *Thalassiosira pseudonana* genome. The seven putative CDKs identified in the diatom genome were trimmed to a conserved core (~280 amino acids; Fig. S8 in the supplementary material), along with CDKs from animal, protozoan, and plant species, and used to build a neighbor-joining tree. Conserved motifs are indicated under the CDK subfamily name. In orthologs with variations of the canonical motif, the motif sequence is indicated after the sequence name. GenBank accession numbers or protein prediction identification numbers of all sequences employed are indicated in Table S6 (in the supplementary material).

T. pseudonana genome we found two copies of VDE and two copies of ZEP (Table S3, annotations 101–104, in the supplementary material), possibly reflecting the presence of this second diatom-specific xanthophyll cycle. We also recognized components of pathways associated with the utilization of antioxidants and antioxidative enzymes such as ascorbate, α-tocopherol, and glutathione peroxidases and reductases (Table S3, annotations 105–116, in the supplementary material), in agreement with the enzyme activities measured in *T. pseudonana* as a response to UV radiation (Rijstenbil 2001). A putative prokaryotic-type catalase/peroxidase was identified that may break down the H_2O_2 generated during the β-oxidation of fatty acids and the glyoxylate cycle, which likely takes place within the peroxisome (Armbrust et al. 2004). The *T. pseudonana* genome encodes two Fe-type superoxide dismutases (SODs) and two Mn-type SODs (at least one likely to be localized to mitochondria) that act to convert oxygen radicals to hydrogen peroxide (Table S3, annotations 117–120, in the supplementary material). Similar to *P. falciparum*, no obvious match was found for a Cu/Zn-type SOD or for a Ni-containing

SOD recently discovered in marine cyanobacteria (Eitinger 2004).

Biotic stress: In their aquatic environments, diatoms are exposed to a variety of predators and pathogenic organisms (Holfeld 2000, Nagasaki et al. 2004, Nagasaki et al. 2005). Diatom cells damaged by copepod feeding release reactive unsaturated aldehydes shown to have an antiproliferative effect on copepod egg hatching and larval development, although the ecological implications of this remain controversial (Miralto et al. 1999, Irigoien et al. 2002, Ianora et al. 2004). Such a defense mechanism has not been reported in *T. pseudonana*, but its genome encodes several putative components of oxylipin production pathways (e.g., phospholipases and cytochrome p450–like predicted proteins; Table S1, in the supplementary material). However, no lipoxygenase/hydroperoxide lyases, hypothesized to be required for this pathway (Pohnert 2002), could be detected. The presence of genes with similarity to components of the prostaglandin biosynthesis pathway (Table S3, annotations 122–123, in the supplementary material), associated with the inflammatory response of the mammalian immune system, may suggest other metabolic pathways related to defense-signaling cascades. As pathogens threaten to become established, higher plants commonly activate a hypersensitive response that derives from an oxidative burst and limits pathogen spread by killing infected cells and activating transcription of cell-defense proteins. The *T. pseudonana* genome encodes multiple proteins with LRR motifs (Table 1), but none contain the nucleotide-binding or TOLL/TIR domains that additionally define disease-defense R-proteins. However, we identified more than 10 plantlike pathogen-related or hypersensitive response-induced proteins; two putative subunits of the NADPH respiratory burst oxidase, recently found to be up-regulated during pathogen attack in a red alga (Herve et al. 2005); and two putative nitric oxide synthases, which may play a central role in detection of biochemical signals from wounded neighboring cells in natural bloom conditions (Vardi et al. 2006). At least five copies of genes that encode proteins belonging to the Multi Antimicrobial and Toxin Extrusion (MATE) family, recently shown to be part of a disease-resistance signal transduction pathway in plants (Nawrath et al. 2002), were also found (Table S3, annotations 141–145, in the supplementary material).

Unicellular apoptosis: When oxidative stress exceeds antioxidant capacity, apoptotic pathways can be induced. Internally triggered cell death has been identified in unicellular organisms (Ameisen 2002), and it may play a major role in phytoplankton bloom succession and collapse (Vardi et al. 1999, Bidle and Falkowski 2004, Vardi et al. 2006). No clear homologs of important elements of metazoan apoptotic pathways, such as p53, caspases, or the BCL2 family of apoptosis regulators, were observed in *T. pseudonana*. However, we could identify at least five apoptosis-related metacaspases (Table S3, annotations 146–150, in the supplementary material), present in *A. thaliana*, fungi, and several eukaryotic microbes (Koonin and Aravind 2002, Madeo et al. 2002). Furthermore, we identified apoptosis-associated nuclear factors E2 F and DP1, found in metazoans and plants, but not previously reported in unicellular organisms, as well as three putative serpins (serine proteinase inhibitors) that have been implicated in host defense in animals (Irving et al. 2000; Table S3, annotations 151–165, in the supplementary material). Interestingly, a serpin for which there is abundant EST support and a plantlike pathogenesis-related protein were recently detected in a proteomic study of a *T. pseudonana* cell-wall protein-enriched fraction, and their pattern of expression through the cell cycle correlated with that of cell-wall-synthesis marker genes (Frigeri et al. 2006). Although the apoptotic elements present in *T. pseudonana* appear largely similar to those of other eukaryotic microbes and plants, no homologs were found for TIR adaptor proteins or AP-ATPases, both of which are abundant in *A. thaliana*. Further examination of these features in other chromalveolates should help elucidate the origin of cell-death pathways and their ecological role in phytoplankton bloom termination.

CONCLUSION

This study complements the overview initiated by Armbrust et al. (2004) of major features of diatom biology that can be investigated from the availability of a first whole-genome sequence for this type of organism. In both reports, novel assortments of metabolic and regulatory components have been inferred from the genome sequence. Whereas Armbrust et al. (2004) focused on the general features of diatom primary metabolism (e.g., carbon, nitrogen, and silica metabolism), the current manuscript describes mechanisms of cell regulation and maintenance of cell homeostasis. More specifically, we have highlighted the basic features of signal transduction, including receptors and canonical signaling cascades, transcription factors and chromatin-level control, cytoskeleton, cell cycle, and stress-response pathways. The analysis reveals an unusual assortment of pathways and components that have never previously been found together in the same organism, reflecting the unusual phylogenetic history of diatoms.

Of potential major significance is that diatoms may possess a novel class of phytochrome that represents a ''brown'' clade, and that the *T. pseudonana* genome appears to lack conventional Rho-type small GTPases, generally considered to be essential regulators of polarized processes such as localized cell-wall synthesis and tactic movement in eukaryotes. The

extraordinarily high number of heat-shock TFs may be of significant interest for future work on diatom stress responses. Finally, the putative membrane-bound environmental sensing mechanisms described here are also very promising experimental targets, in view of recent studies that have revealed sophisticated chemical-based systems for the transduction of external signals in marine diatoms (Falciatore et al. 2000, Vardi et al. 2006). However, the relative paucity of environmental sensing mechanisms will be a major challenge to address experimentally. Another highly successful group of photosynthetic organisms in the marine environment, the prokaryotic prochlorophytes, were also determined to have a poor content of classical sensing mechanisms for the perception of external signals (as opposed to their freshwater counterparts), which may suggest that the marine environment is sufficiently uniform and unchanging that a large number of receptor-based signaling systems is not required (Mary and Vaulot 2003).

In summary, the results of the in silico–based analyses described here and in Armbrust et al. provide a rather complete blueprint of *T. pseudonana* biology that should serve as cornerstones for future laboratory-based and in situ–based experimental studies aimed at understanding diatom biology and the position of these organisms in the eukaryotic evolutionary tree.

This work was performed under the auspices of the U.S. Department of Energy's (DOE) Office of Science, Biological and Environmental Research Program and 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. Additional funding was provided by the European Union (contracts QLRT-2001-01226, LSHG-CT-2004-512035, and GOCE-CT-2004-505403 to C. B.), the CNRS ATIP ''Blanche'' programme (2JE144 to C. B.), and the U.S. DOE (DE-FG03-02ER63471 to E. V. A.).

Ahmad, M. & Cashmore, A. R. 1993. *HY4* gene of *A. thaliana* encodes a protein with characteristics of a blue-light photoreceptor. *Nature* 366:162–6.

Allen, A. E., Vardi, A. & Bowler, C. 2006. An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Curr. Opin. Plant Biol.* 9:264–73.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. 1990. Basic local alignment search tool. *J. Mol. Evol.* 215: 403–10.

Ameisen, J. C. 2002. On the origin, evolution, and nature of programmed cell death: a timeline of four billion years. *Cell Death Differ.* 9:367–93.

Apt, K. E., Zaslavkaia, L., Lippmeier, J. C., Lang, M., Kilian, O., Wetherbee, R., Grossman, A. R. & Kroth, P. G. 2002. In vivo characterization of diatom multipartite plastid targeting signals. *J. Cell Sci.* 115:4061–9.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815.

Armbrust, E. V., Berges, J. B., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306:79–86.

Bernard, P., Hardwick, K. & Javerzat, J. P. 1998. Fission yeast bub1 is a mitotic centromere protein essential for the spindle checkpoint and the preservation of correct ploidy through mitosis. *J. Cell Biol.* 143:1775–87.

Bidle, K. D. & Falkowski, P. G. 2004. Cell death in planktonic, photosynthetic microorganisms. *Nat. Rev. Microbiol.* 2: 643–55.

Borowitzka, L. J. & Volcani, B. E. 1977. Role of silicon in diatom metabolism. VIII. Cyclic AMP and cyclic GMP in synchronized cultures of *Cylindrotheca fusiformis*. *Arch. Microbiol.* 112:147–52.

Brudler, R., Hitomi, K., Daiyasu, H., Toh, H., Kucho, K., Ishiura, M., Kanehisa, M., Roberts, V. A., Todo, T., Tainer, J. A. & Getzoff, E. D. 2003. Identification of a new cryptochrome class. Structure, function, and evolution. *Mol. Cell* 11:59–67.

Cashmore, A. R. 2003. Cryptochromes: enabling plants and animals to determine circadian time. *Cell. Mol. Life Sci.* 114:537–43.

Colucci, G., Apone, F., Alyeshmerni, N., Chalmers, D. & Chrispeels, M. J. 2002. *GCR1*, the putative *Arabidopsis* G protein-coupled receptor gene is cell cycle-regulated, and its overexpression abolishes seed dormancy and shortens time to flowering. *Proc. Natl. Acad. Sci. U. S. A.* 99:4736–41.

Davis, S. J., Vener, A. V. & Vierstra, R. D. 1999. Bacteriophytochromes: phytochrome-like photoreceptors from nonphotosynthetic eubacteria. *Science* 286:2517–20.

Delaroque, N., Mueller, D. G., Bothe, G., Pohl, T., Knippers, R. & Boland, W. 2001. The complete DNA sequence of the *Ectocarpus siliculosus* Virus EsV-1 genome. *Virology* 287:112–32.

Eitinger, T. 2004. In vivo production of active nickel superoxide dismutase from *Prochlorococcus marinus* MIT9313 is dependent on its cognate peptidase. *J. Bacteriol.* 186:7821–5.

Etienne-Manneville, S. & Hall, A. 2002. Rho GTPases in cell biology. *Nature* 420:629–35.

Falciatore, A. & Bowler, C. 2005. The evolution and function of blue and red light photoreceptors. *Curr. Top. Dev. Biol.* 68: 317–50.

Falciatore, A., d'Alcala, M. R., Croot, P. & Bowler, C. 2000. Perception of environmental signals by a marine diatom. *Science* 288:2363–6.

FANTOM Consortium & the RIKEN Genome Exploration Research Group Phase I and II Team. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420: 563–73.

Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281:237–40.

Field, M. C. 2005. Signalling the genome: the Ras-like small GTPase family of trypanosomatids. *Trends Parasitol.* 21:447–50.

Fowler, J. E., Vejlupkova, Z., Goodner, B. W., Lu, G. & Quatrano, R. S. 2004. Localization to the rhizoid tip implicates a *Fucus distichus* Rho family GTPase in a conserved cell polarity pathway. *Planta* 219:566–866.

Frigeri, L. G., Radabaugh, T. R., Haynes, P. A. & Hildebrand, M. 2006. Identification of proteins from a cell wall fraction of the diatom *Thalassiosira pseudonana*: insights into silica structure formation. *Mol. Cell. Proteomics* 5:182–93.

Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., FitzHugh, W., et al. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* 422:859–68.

Giraud, E., Fardoux, J., Fourrier, N., Hannibal, L., Genty, B., Bouyer, P., Dreyfus, B. & Vermeglio, A. 2002. Bacteriophytochrome controls photosystem synthesis in anoxygenic bacteria. *Nature* 417:202–5.

Graumann, P. L. & Marahiel, M. A. 1998. A superfamily of proteins that contain the cold-shock domain. *Trends Biochem. Sci.* 23: 286–90.

Gualerzi, C. O., Giuliodori, A. M. & Pon, C. L. 2003. Transcriptional and post-transcriptional control of cold-shock genes. *J. Mol. Biol.* 331:527–39.

Guo, A., He, K., Liu, D., Bai, S., Gu, X., Wei, L. & Luo, J. 2005. DATF: a database of *Arabidopsis* transcription factors. *Bioinformatics* 21:2568–9.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* 41: 95–8.

Herve, C., Tonon, T., Collen, J., Corre, E. & Boyen, C. 2005. NADPH oxidases in eukaryotes: red algae provide new hints! *Curr. Genet.* 49:190–204.

Hildebrand, M., Dahlin, K. & Volcani, B. E. 1998. Characterization of a silicon transporter gene family in *Cylindrotheca fusiformis*: sequences, expression analysis, and identification of homologs in other diatoms. *Mol. Gen. Genet.* 260:480–6.

Holfeld, H. 2000. Infection of the single-celled diatom *Stephanodiscus alpinus* by the chytrid *Zygorhizidium*: parasite distribution within host population, changes in host cell size, and host-parasite size relationship. *Limnol. Oceanogr.* 45:1440–4.

Ianora, A., Miralto, A., Poulet, S. A., Carotenuto, Y., Buttino, I., Romano, G., Casotti, R., Pohnert, G., Wichard, T., Colucci-D'Amato, L., Terrazzano, G. & Smetacek, V. 2004. Aldehyde suppression of copepod recruitment in blooms of a ubiquitous planktonic diatom. *Nature* 429:403–7.

Irigoien, X., Harris, R. P., Verheye, H. M., Joly, P., Runge, J., Starr, M., Pond, D., Campbell, R., Shreeve, R., Ward, P., Smith, A. N., Dam, H. G., Peterson, W., Tirelli, V., Koski, M., Smith, T., Harbour, D. & Davidson, R. 2002. Copepod hatching success in marine ecosystems with high diatom concentrations. *Nature* 419:387–9.

Irving, J. A., Pike, R. N., Lesk, A. M. & Whisstock, J. C. 2000. Phylogeny of the serpin superfamily: implications of patterns of amino acid conservation for structure and function. *Genome Res.* 10:1845–64.

Jarvis, E. E., Dunahay, T. G. & Brown, L. M. 1992. DNA nucleoside and methylation in several species of microalgae. *J. Phycol.* 28:356–62.

Jekely, G. 2003. Small GTPases and the evolution of the eukaryotic cell. *Bioessays* 25:1129–38.

Jonak, C., Okresz, L., Bogre, L. & Hirt, H. 2002. Complexity, cross talk and integration of plant MAP kinase signalling. *Curr. Opin. Plant Biol.* 5:415–24.

Katayama, M., Tsinoremas, N. F., Kondo, T. & Golden, S. S. 1999. *CpmA*, a gene involved in an output pathway of the cyanobacterial circadian system. *J. Bacteriol.* 181:3516–24.

Kirk, J. 1992. The nature and measurement of the light environment in the ocean. *In* Falkowski, P. & Woodhead, A. [Eds.] *Primary Productivity and Biogeochemical Cycles in the Sea.* Plenum Press, New York, pp. 9–29.

Kitajima, T. S., Yokobayashi, S., Yamamoto, M. & Watanabe, Y. 2003. Distinct cohesin complexes organize meiotic chromosome domains. *Science* 300:1152–5.

Kooistra, W. H. C. F., DeStefano, M., Mann, D. G. & Medlin, L. K. 2003. The phylogeny of the diatoms. *Prog. Mol. Subcell. Biol.* 33:59–97.

Koonin, E. V. & Aravind, L. 2002. Origin and evolution of eukaryotic apoptosis: the bacterial connection. *Cell Death Differ.* 9:394–404.

Kroger, N., Deutzmann, R. & Sumper, M. 1999. Polycationic peptides from diatom biosilica that direct silica nanosphere formation. *Science* 286:1129–32.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305: 567–80.

Lal, K., Field, M. C., Carlton, J. M., Warwicker, J. & Hirt, R. P. 2005. Identification of a very large Rab GTPase family in the parasitic protozoan *Trichomonas vaginalis*. *Mol. Biochem. Parasitol.* 143:226–35.

Lange, B. M. & Ghassemian, M. 2003. Genome organization in *Arabidopsis thaliana*: a survey for genes involved in isoprenoid and chlorophyll metabolism. *Plant Mol. Biol.* 51:925–48.

Lavaud, J., Rousseau, B., van Gorkom, H. J. & Etienne, A. L. 2002. Influence of the diadinoxanthin pool size on photoprotection in the marine planktonic diatom *Phaeodactylum tricornutum*. *Plant Physiol.* 129:1398–406.

Leblanc, C., Falciatore, A., Watanabe, M. & Bowler, C. 1999. Semi-quantitative RT-PCR analysis of photoregulated gene expression in marine diatoms. *Plant Mol. Biol.* 40:1031–44.

Lee, B. H., Amon, A. & Prinz, S. 2002. Spo13 regulates cohesin cleavage. *Genes Dev.* 16:1672–81.

Li, S., Nosenko, T., Hackett, J. D. & Bhattacharya, D. 2006. Phylogenomic analysis identifies red algal genes of endosymbiotic origin in the chromalveolates. *Mol. Biol. Evol.* 23:663–74.

Lin, C. & Todo, T. 2005. The cryptochromes. *Genome Biol.* 6:220–8.

Lindell, D., Jaffe, J. D., Johnson, Z. I., Church, G. M. & Chisholm, S. W. 2005. Photosynthesis genes in marine viruses yield proteins during host infection. *Nature* 438:86–9.

Lohr, M., Im, C. S. & Grossman, A. R. 2005. Genome-based examination of chlorophyll and carotenoid biosynthesis in *Chlamydomonas reinhardtii*. *Plant Physiol.* 138:490–515.

Lopez, P. J., Descles, J., Allen, A. E. & Bowler, C. 2005. Prospects in diatom research. *Curr. Opin. Biotechnol.* 16:180–6.

Madeo, F., Herker, E., Maldener, C., Wissing, S., Lachelt, S., Herlan, M., Fehr, M., Lauber, K., Sigrist, S. J., Wesselborg, S. & Frohlich, K. U. 2002. A caspase-related protease regulates apoptosis in yeast. *Mol. Cell* 9:911–7.

Maheswari, U., Montsant, A., Goll, J., Krishnaswamy, S., Rajyashri, K. R., Patell, V. M. & Bowler, C. 2005. The diatom EST database. *Nucleic Acids Res.* 33:D344–7.

Malik, H. S. & Henikoff, S. 2003. Phylogenomics of the nucleosome. *Nat. Struct. Biol.* 10:882–91.

Manton, I., Kowallik, K. & von Stosch, H. A. 1970. Observations on the fine structure and development of the spindle at mitosis and meiosis in a marine centric diatom (*Lithodesmium undulatum*). IV. The second meiotic division and conclusion. *J. Cell Sci.* 7:407–43.

Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S. Q., Hurwitz, D. I., et al. 2003. CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* 31:383–7.

Maris, A. E., Sawaya, M. R., Kaczor-grzeskowiak, M., Jarvis, M. R., Bearson, S. M., Kopka, M. L., Schroder, I., Gunsalus, R. P. & Dickerson, R. E. 2002. Dimerization allows DNA target site recognition by the NarL response regulator. *Nat. Struct. Biol.* 9:771–8.

Mary, I. & Vaulot, D. 2003. Two-component systems in *Prochlorococcus* MED4: genomic analysis and differential expression under stress. *FEMS Microbiol. Lett.* 226:135–44.

Matsuzaki, M., Misumi, O., Shin-I, T., Maruyama, S., Takahara, M., Miyagishima, S. Y., Mori, T., et al. 2004. Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature* 428:653–57.

Miralto, A., Barone, G., Romano, G., Poulet, S. A., Ianora, A., Russo, G. L., Buttino, I., Mazzarella, G., Laabir, M., Cabrini, M. & Giacobbe, M. G. 1999. The insidious effect of diatoms on copepod reproduction. *Nature* 402:173–6.

Mizuno, T. 1998. His-Asp phosphotransfer signal transduction. *J. Biochem.* 123:555–63.

Montsant, A., Jabbari, K., Maheswari, U. & Bowler, C. 2005. Comparative genomics of the pennate diatom *Phaeodactylum tricornutum*. *Plant Physiol.* 137:500–13.

Morand, L. Z., Kidd, D. G. & Lagarias, J. C. 1993. Phytochrome levels in the green alga *Mesotaenium caldariorum* are light regulated. *Plant Physiol.* 101:97–104.

Nagasaki, K., Tomaru, Y., Katanozaka, N., Shirai, Y., Nishida, K., Itakura, S. & Yamaguchi, M. 2004. Isolation and characterization of a novel single-stranded RNA virus infecting the bloom-forming diatom *Rhizosolenia setigera*. *Appl. Environ. Microbiol.* 70:704–11.

Nagasaki, K., Tomaru, Y., Takao, Y., Nishida, K., Shirai, Y., Suzuki, H. & Nagumo, T. 2005. Previously unknown virus infects marine diatom. *Appl. Environ. Microbiol.* 71:3528–35.

Nagel, G., Ollig, D., Fuhrmann, M., Kateriya, S., Musti, A. M., Bamberg, E. & Hegemann, P. 2002. Channelrhodopsin-1: a light-gated proton channel in green algae. *Science* 296:2395–8.

Nawrath, C., Heck, S., Parinthawong, N. & Metraux, J. P. 2002. EDS5, an essential component of salicylic acid-dependent

signaling for disease resistance in *Arabidopsis*, is a member of the MATE transporter family. *Plant Cell* 14:275–86.
Oka, A., Sakai, H. & Iwakoshi, S. 2002. His-Asp phosphorelay signal transduction in higher plants: receptors and response regulators for cytokinin signaling in *Arabidopsis thaliana*. *Genes Genet. Syst.* 77:383–91.
Parkinson, J. & Gordon, R. 1999. Beyond micromachining: the potential of diatoms. *Trends Biotechnol.* 17:190–6.
Patron, N. J., Rogers, M. B. & Keeling, P. J. 2004. Gene replacement of fructose-1,6-bisphosphate aldolase supports the hypothesis of a single photosynthetic ancestor of chromalveolates. *Eukaryot. Cell* 3:1169–75.
Pickett-Heaps, J., Schmid, A. -M. M. & Edgar, L. A. 1990. The cell biology of diatom valve formation. In Round, F. E. & Chapman, D. J. [Eds.] *Progress in Phycological Research*. Biopress Ltd., Bristol, UK, pp. 1–168.
Piperno, G., Siuda, E., Henderson, S., Segil, M., Vaananen, H. & Sassaroli, M. 1998. Distinct mutants of retrograde intraflagellar transport (IFT) share similar morphological and molecular defects. *J. Cell Biol.* 143:1591–601.
Pohnert, G. 2002. Phospholipase A2 activity triggers the wound-activated chemical defense in the diatom *Thalassiosira rotula*. *Plant Physiol.* 129:103–11.
Poulsen, N. & Kroger, N. 2004. Silica morphogenesis by alternative processing of silaffins in the diatom *Thalassiosira pseudonana*. *J. Biol. Chem.* 279:42993–9.
Pruyne, D., Legesse-Miller, A., Gao, L., Dong, Y. & Bretscher, A. 2004. Mechanisms of polarized growth and organelle segregation in yeast. *Annu. Rev. Cell Dev. Biol.* 20:559–91.
Rabin, R. S. & Stewart, V. 1993. Dual response regulators (NarL and NarP) interact with dual sensors (NarX and NarQ) to control nitrate- and nitrite-regulated gene expression in *Escherichia coli* K-12. *J. Bacteriol.* 175:3259–68.
Reinfelder, J. R., Kraepiel, A. M. L. & Morel, F. M. M. 2000. Unicellular C-4 photosynthesis in a marine diatom. *Nature* 407:996–9.
Richards, T. A. & Cavalier-Smith, T. 2005. Myosin domain evolution and the primary divergence of eukaryotes. *Nature* 436:1113–8.
Rijstenbil, J. W. 2001. Effects of periodic, low UVA radiation on cell characteristics and oxidative stress in the marine planktonic diatom *Ditylum brightwellii*. *Eur. J. Phycol.* 36:1–8.
Rosenbaum, J. L. & Witman, G. B. 2002. Intraflagellar transport. *Nat. Rev. Mol. Cell Biol.* 3:813–25.
Round, F. E., Crawford, R. M. & Mann, D. G. 1990. *The Diatoms: Biology and Morphology of the Genera*. Cambridge University Press, London, 747 pp.
Saito-Nakano, Y., Loftus, B. J., Hall, N. & Nozaki, T. 2005. The diversity of Rab GTPases in *Entamoeba histolytica*. *Exp. Parasitol.* 110:244–52.
Salome, P. A. & McClung, C. R. 2004. The *Arabidopsis thaliana* clock. *J. Biol. Rhythms* 19:425–35.
Scala, S., Carels, N., Falciatore, A., Chiusano, M. L. & Bowler, C. 2002. Genome properties of the diatom *Phaeodactylum tricornutum*. *Plant Physiol.* 129:993–1002.
Schmitz, O., Katayama, M., Williams, S. B., Kondo, T. & Golden, S. S. 2000. CikA, a bacteriophytochrome that resets the cyanobacterial circadian clock. *Science* 289:765–8.
Schroeder, T. A. 2004. Dynactin. *Annu. Rev. Cell Dev. Biol.* 20:759–79.
Sellers, J. R. 2000. Myosins: a diverse superfamily. *Biochim. Biophys. Acta* 1496:3–22.
Shiu, S. H., Shih, M. C. & Li, W. H. 2005. Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiol.* 139:18–26.
Sineshchekov, O. A., Jung, K. H. & Spudich, J. L. 2002. Two rhodopsins mediate phototaxis to low- and high-intensity light in *Chlamydomonas reinhardtii*. *Proc. Natl. Acad. Sci. U. S. A.* 99:8689–94.
Smetacek, V. 1999. Diatoms and the ocean carbon cycle. *Protist* 150:25–32.
Somers, D. E. 2001. Clock-associated genes in *Arabidopsis*: a family affair. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 356:1745–53.
Stephenson, K. & Hoch, J. A. 2002. Evolution of signalling in the sporulation phosphorelay. *Mol. Microbiol.* 46:297–304.
Takai, Y., Sasaki, T. & Matozaki, T. 2001. Small GTP-binding proteins. *Physiol. Rev.* 81:153–208.
Tyler, B. M., Tripathy, S., Zhang, X., Dehal, P., Jiang, R. H., Aerts, A., Arredondo, F. D., et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313:1261–6.
Unsal-Kacmaz, K., Mullen, T. E., Kaufmann, W. K. & Sancar, A. 2005. Coupling of human circadian and cell cycles by the timeless protein. *Mol. Cell. Biol.* 25:3109–16.
Vandepeer, Y. & Dewachter, R. 1994. Treecon for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. *Comput. Appl. Biosci.* 10:569–70.
Vardi, A., Berman-Frank, I., Rozenberg, T., Hadas, O., Kaplan, A. & Levine, A. 1999. Programmed cell death of the dinoflagellate *Peridinium gatunense* is mediated by CO_2 limitation and oxidative stress. *Curr. Biol.* 9:1061–4.
Vardi, A., Formiggini, F., Casotti, R., De Martino, A., Ribalet, F., Miralto, A. & Bowler, C. 2006. A stress surveillance system based on calcium and nitric oxide in marine diatoms. *PLoS Biol.* 4:e60 (e-pub ahead of print).
Vernoud, V., Horton, A. C., Yang, Z. & Nielsen, E. 2003. Analysis of the small GTPase gene superfamily of *Arabidopsis*. *Plant Physiol.* 131:1191–208.
Vugrek, O., Sawitzky, H. & Menzel, D. 2003. Class XIII myosins from the green alga *Acetabularia*: driving force in organelle transport and tip growth? *J. Muscle Res. Cell Motil.* 24:87–97.
Wagner, J. R., Brunzelle, J. S., Forest, K. T. & Vierstra, R. D. 2005. A light-sensing knot revealed by the structure of the chromophore-binding domain of phytochrome. *Nature* 438:325–31.
Xanthoudakis, S. & Curran, T. 1992. Identification and characterization of Ref-1, a nuclear protein that facilitates AP-1 DNA-binding activity. *EMBO J.* 11:653–65.
Xanthoudakis, S., Miao, G. G. & Curran, T. 1994. The redox and DNA-repair activities of Ref-1 are encoded by nonoverlapping domains. *Proc. Natl. Acad. Sci. U. S. A.* 4:23–7.
Xiao, B., Wilson, J. R. & Gamblin, S. J. 2003. SET domains and histone methylation. *Curr. Opin. Struct. Biol.* 13:699–705.
Yamamoto, A., Mizukami, Y. & Sakurai, H. 2005. Identification of a novel class of target genes and a novel type of binding sequence of heat shock transcription factor in *Saccharomyces cerevisiae*. *J. Biol. Chem.* 280: 11911–9.
Yeh, K. C., Wu, S. H., Murphy, J. T. & Lagarias, J. C. 1997. A cyanobacterial phytochrome two-component light sensory system. *Science* 277:1505–8.
Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G. & Bhattacharya, D. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol. Biol. Evol.* 21:809–18.
Yoshihara, S., Suzuki, F., Fujita, H., Geng, X. X. & Ikeuchi, M. 2000. Novel putative photoreceptor and regulatory genes required for the positive phototactic movement of the unicellular motile cyanobacterium *Synechocystis sp.* PCC 6803. *Plant Cell Physiol.* 41:1299–304.
Zurzolo, C. & Bowler, C. 2001. Exploring bioinorganic pattern formation in diatoms. A story of polarized trafficking. *Plant Physiol.* 127:1339–45.

Supplementary Material

Tables of InterPro domain counts (Table S1), data complementing the analysis of similar-gene clusters (Table S2), annotated gene models (Table S3), and accession numbers of sequences employed in Figures 2, 5, and 7 (Tables S4–S6) are available online. Alignments used to derive neighbor-joining trees (Figs. S1–S4, S7, and S8), the small-GTPase trees for reference microbes on which Table 2 counts are based (Fig. S5), and a scheme of histone clusters (Fig. S6) are also provided as supplementary figures.

The following supplementary materials are available as part of the online article from http://www.blackwell-synergy.com:

Table S1. InterPro conserved domains identified in *Thalassiosira pseudonana* and comparison with six reference eukaryotes.

Table S2. Clusters of similar coding sequences in *Thalassiosira pseudonana* and in other model eukaryotes.

Table S3. A selection of annotated gene predictions relevant to signaling, gene expression, cytoskeletal regulation, and stress response.

Table S4. Identification numbers of the protein sequences used to derive the cryptochrome tree in Figure 2.

Table S5. GenBank accession numbers of the protein sequences used to derive Figures 5 and S4 (in the supplementary material).

Table S6. Identification numbers of the protein sequences used to derive Figure 7.

Figure S1. Alignment of the cryptochrome sequences used to derive the tree shown in Figure 2 of main text.

Figure S2. Alignment of the phytochrome sequences used to derive the tree shown in Figure 3 of main text.

Figure S3. Alignment of the small GTPase sequences used to derive the tree shown in Figure 5 of main text.

Figure S4. A. Alignment of the sequences used to derive a small-GTPase tree for *Plasmodium falciparum*. B. Alignment of the sequences used to derive a small-GTPase tree for *Phytophthora sojae*. C. Alignment of the sequences used to derive a small-GTPase tree for *Ostreococcus tauri*. D. Alignment of the sequences used to derive a small-GTPase tree for *Cyanidioschyzon merolae*.

Figure S5. Neighbor-joining trees of the small GTPases of *Plasmodium falciparum* (a), *Phytophthora sojae* (b), *Ostreococcus tauri* (c), and *Cyanidioschyzon merolae* (d) along with well-characterized small GTPases from yeast and humans.

Figure S6. Schematic of histone clusters identified in the *Thalassiosira pseudonana* genome.

Figure S7. Alignment of the myosin head domain sequences used to derive the tree shown in Figure 6 of main text.

Figure S8. Alignment of the CDK sequences used to derive the tree shown in Figure 7 of main text.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1529-8817.2007.00342.x (This link will take you to the article abstract).

Please note: Blackwell Publishing is not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.