The Phaeodactylum genome reveals the dynamic nature and multi-lineage evolutionary history of diatom genomes

Chris Bowler[1,2], Andrew E. Allen[1,3*], Jonathan H. Badger[3*], Jane Grimwood[4*], Kamel Jabbari[1*], Alan Kuo[5*], Uma Maheswari[1*], Cindy Martens[6*], Florian Maumus[1*], Robert P. Otillar[5*], Edda Rayko[1*], Asaf Salamov[5*], Klaas Vandepoele[6*], Bank Beszteri[7], Ansgar Gruber[8], Marc Heijde[1], Michael Katinka[9], Thomas Mock[10], Klaus Valentin[7], Fréderic Vérret[11], John A. Berges[12], Colin Brownlee[11], Jean-Paul Cadoret[13], Anthony Chiovitti[14], Chang Jae Choi[12], Sacha Coesel[2$], Alessandra De Martino[1], J. Chris Detter[5], Colleen Durkin[10], Angela Falciatore[2], Jérome Fournet[15], Miyoshi Haruta[16], Marie Huysman[17], Bethany D. Jenkins[18], Katerina Jiroutova[19], Richard E. Jorgensen[20], Yolaine Joubert[15], Aaron Kaplan[21], Nils Kroeger[22], Peter Kroth[8], Julie La Roche[23], Erica Lindquist[5], Markus Lommer[23],Véronique Martin-Jézéquel[15], Pascal J. Lopez[1], Susan Lucas[5], Manuela Mangogna[2], KarenMcGinnis[20], Linda K. Medlin[7], Anton Montsant[1,2], Marie-Pierre Oudot-Le Secq[24], CarolynNapoli[20], Miroslav Obornik[19], Jean-Louis Petit[9], Betina M. Porcel[9], Nicole Poulsen[25], Matthew Robison[16], Leszek Rychlewski[26], Tatiana A. Rynearson[27], Jeremy Schmutz[4], Micaela Schnitzler Parker[10], Harris Shapiro[5], Magali Siaut[28§], Michele Stanley[28], Michael J., Sussman[16], Alison Taylor[11,29], Assaf Vardi[1,30], Peter von Dassow[31], Wim Vyverman[17], Anusuya Willis[14], Lucjan S. Wyrwicz[26], Daniel S. Rokhsar[5], Jean Weissenbach[9], E. Virginia Armbrust[10], Beverley R. Green[24], Yves Van de Peer[6], and Igor V. Grigoriev[5]

* Equal contribution

1 CNRS UMR8186, Dept of Biology, Ecole Normale Supérieure, 46 rue d'Ulm, Paris, France

2 Stazione Zoologica 'Anton Dohrn,' Villa Comunale, I-80121 Naples, Italy

3 J. Craig Venter Institute, San Diego, CA 92121, USA

[4] Joint Genome Institute-Stanford, Stanford Human Genome Center, 975 California Avenue, Palo Alto, CA 94304, USA

[5] Joint Genome Institute, Genomic Annotation Division, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA

[6] VIB Department of Plant Systems Biology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium

[7] Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12,

Bremerhaven, Germany

[8] Fachbereich Biologie, University of Konstanz, Konstanz, Germany

[9] Genoscope, CEA-Institut de Génomique, UMR CNRS n° 8030, 2 rue Gaston Crémieux,

91057 Evry Cedex, France

[10] School of Oceanography, University of Washington, Seattle, WA 98195, USA

[11] Marine Biological Association of the UK, The Laboratory, Citadel Hill, Plymouth PL1

2PB, UK

[12] Dept. of Biological Sciences, University of Wisconsin, Milwaukee, Wisconsin USA

[13] PBA, IFREMER, BP 21105, 44311 Nantes Cedex 03, France

[14] School of Botany, The University of Melbourne, Victoria, 3010, Australia

[15] EA 2160, Laboratoire Mer, Molécule, Santé, Faculté des Sciences, Université de Nantes,

2 rue de la Houssinière, 44322, BP 92208, Nantes, France

[16] University of Wisconsin Biotechnology Center, 425 Henry Mall, Madison, WI 53706, USA

[17] Lab of Protistology and Aquatic Ecology, Gent University, Krijgslaan 281-S8, B-9000

Gent, Belgium

[18] Department of Cell and Molecular Biology, and Graduate School of Oceanography,

University of Rhode Island, 316 Morrill Hall, 45 Lower College Road, Kingston, Rhode

Island 02881, USA

[19] Biology Centre ASCR, Institute of Parasitology and University of South Bohemia, Faculty

of Science, Branisovska 31, 370 05 Ceske Budejovice, Czech Republic

[20] Institute and Department of Plant Sciences, University of Arizona, Tucson, AZ 85719, USA

[21] Dept of Plant and Environmental Sciences, the Hebrew University of Jerusalem, Israel

[22] School of Chemistry and Biochemistry, School of Materials Science and Engineering,

School of Biology, Georgia Institute of Technology, 901 Atlantic Dr. NW, Atlanta, GA

30332-0400, USA

[23] Leibniz-Institut fuer Meereswissenschaften, 24105 Kiel, Germany

[24] Dept. of Botany, University of British Columbia, 3529-6270 University Boulevard,

Vancouver, B.C., Canada

[25] School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta GA, USA

[26] BioInfoBank Institute, Limanowskiego 24A/16, 60-744 Poznan, Poland

[27] Graduate School of Oceanography, University of Rhode Island, South Ferry Road,

Narragansett, RI 02882-1197, USA

[28] Microbial & Molecular Biology, Scottish Association for Marine Science, Dunstaffnage

Marine Laboratory, Oban, Argyll PA37 1QA, UK

[29] Department of Biology and Marine Biology, The University of North Carolina, 601 South

College Road, Wilmington, NC 28403, USA

[30] Environmental Biophysics and Molecular Ecology Group, Institute of Marine and Coastal

Sciences, Rutgers University, 71 Dudley Road, New Brunswick, NJ 08901 USA

[31] CNRS UMR7144, Station Biologique de Roscoff, Place George Teissier BP74, Roscoff,

France

[$] Current address: Institute for Systems Biology, 1441 N 34th Street, Seattle, WA 98103,

USA

§ Current address: CEA, DSV, IBEB, SBVME, UMR 6191 CNRS/CEA/Université Aix-Marseille, Laboratoire de Bioénergétique et Biotechnologie des Bactéries et Microalgues, Cadarache, Saint-Paul-lez-Durance, F-13108 France

[1]*To whom correspondence may be addressed. E-mail*: cbowler@biologie.ens.fr

May 19, 2011

ACKNOWLEDGMENTS:

Diatom genome sequencing at the JGI (USA) was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory, under contract no. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under contract no. DE-AC52-07NA27344 and Los Alamos National Laboratory under contract no. DE-AC02-06NA25396. P. tricornutum ESTs were generated at Genoscope (France). Funding for this work was also obtained from the EU-funded FP6 Diatomics project (LSHG-CT-2004-512035), the EU-FP6 Marine Genomics Network of Excellence (GOCE-CT-2004-505403), an ATIP 'Blanche' grant from the CNRS (France) and the Agence Nationale de la Recherche (France).

DISCLAIMER:

[**LBNL**] This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

The Phaeodactylum genome reveals the dynamic nature and multi-lineage evolutionary history of diatom genomes

Chris Bowler[1,2], Andrew E. Allen[1,3*], Jonathan H. Badger[3*], Jane Grimwood[4*], Kamel Jabbari[1*], Alan Kuo[5*], Uma Maheswari[1*], Cindy Martens[6*], Florian Maumus[1*], Robert P. Otillar[5*], Edda Rayko[1*], Asaf Salamov[5*], Klaas Vandepoele[6*], Bank Beszteri[7], Ansgar Gruber[8], Marc Heijde[1], Michael Katinka[9], Thomas Mock[10], Klaus Valentin[7], Fréderic Vérret[11], John A. Berges[12], Colin Brownlee[11], Jean-Paul Cadoret[13], Anthony Chiovitti[14], Chang Jae Choi[12], Sacha Coesel[2$], Alessandra De Martino[1], J. Chris Detter[5], Colleen Durkin[10], Angela Falciatore[2], Jérome Fournet[15], Miyoshi Haruta[16], Marie Huysman[17], Bethany D. Jenkins[18], Katerina Jiroutova[19], Richard E. Jorgensen[20], Yolaine Joubert[15], Aaron Kaplan[21], Nils Kroeger[22], Peter Kroth[8], Julie La Roche[23], Erica Lindquist[5], Markus Lommer[23], Véronique Martin-Jézéquel[15], Pascal J. Lopez[1], Susan Lucas[5], Manuela Mangogna[2], Karen McGinnis[20], Linda K. Medlin[7], Anton Montsant[1,2], Marie-Pierre Oudot-Le Secq[24], Carolyn Napoli[20], Miroslav Obornik[19], Jean-Louis Petit[9], Betina M. Porcel[9], Nicole Poulsen[25], Matthew Robison[16], Leszek Rychlewski[26], Tatiana A. Rynearson[27], Jeremy Schmutz[4], Micaela Schnitzler Parker[10], Harris Shapiro[5], Magali Siaut[2§], Michele Stanley[28], Michael J. Sussman[16], Alison Taylor[11,29], Assaf Vardi[1,30], Peter von Dassow[31], Wim Vyverman[17], Anusuya Willis[14], Lucjan S. Wyrwicz[26], Daniel S. Rokhsar[5], Jean Weissenbach[9], E. Virginia Armbrust[10], Beverley R. Green[24], Yves Van de Peer[6], Igor V. Grigoriev[5]

[*] Equal contribution
[1] CNRS UMR8186, Dept of Biology, Ecole Normale Supérieure, 46 rue d'Ulm, Paris, France
[2] Stazione Zoologica 'Anton Dohrn,' Villa Comunale, I-80121 Naples, Italy
[3] J. Craig Venter Institute, San Diego, CA 92121, USA
[4] Joint Genome Institute-Stanford, Stanford Human Genome Center, 975 California Avenue, Palo Alto, CA 94304, USA

[5] Joint Genome Institute, Genomic Annotation Division, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA
[6] VIB Department of Plant Systems Biology, Ghent University, Technologiepark 927, B-9052 Ghent, Belgium
[7] Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen 12, Bremerhaven, Germany
[8] Fachbereich Biologie, University of Konstanz, Konstanz, Germany
[9] Genoscope, CEA-Institut de Génomique, UMR CNRS n° 8030, 2 rue Gaston Crémieux, 91057 Evry Cedex, France
[10] School of Oceanography, University of Washington, Seattle, WA 98195, USA
[11] Marine Biological Association of the UK, The Laboratory, Citadel Hill, Plymouth PL1 2PB, UK
[12] Dept. of Biological Sciences, University of Wisconsin, Milwaukee, Wisconsin USA
[13] PBA, IFREMER, BP 21105, 44311 Nantes Cedex 03, France
[14] School of Botany, The University of Melbourne, Victoria, 3010, Australia
[15] EA 2160, Laboratoire Mer, Molécule, Santé, Faculté des Sciences, Université de Nantes, 2 rue de la Houssinière, 44322, BP 92208, Nantes, France
[16] University of Wisconsin Biotechnology Center, 425 Henry Mall, Madison, WI 53706, USA
[17] Lab of Protistology and Aquatic Ecology, Gent University, Krijgslaan 281-S8, B-9000 Gent, Belgium
[18] Department of Cell and Molecular Biology, and Graduate School of Oceanography, University of Rhode Island, 316 Morrill Hall, 45 Lower College Road, Kingston, Rhode Island 02881, USA
[19] Biology Centre ASCR, Institute of Parasitology and University of South Bohemia, Faculty of Science, Branisovska 31, 370 05 Ceske Budejovice, Czech Republic
[20] Institute and Department of Plant Sciences, University of Arizona, Tucson, AZ 85719, USA
[21] Dept of Plant and Environmental Sciences, the Hebrew University of Jerusalem, Israel
[22] School of Chemistry and Biochemistry, School of Materials Science and Engineering, School of Biology, Georgia Institute of Technology, 901 Atlantic Dr. NW, Atlanta, GA 30332-0400, USA
[23] Leibniz-Institut fuer Meereswissenschaften, 24105 Kiel, Germany
[24] Dept. of Botany, University of British Columbia, 3529-6270 University Boulevard, Vancouver, B.C., Canada
[25] School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta GA, USA
[26] BioInfoBank Institute, Limanowskiego 24A/16, 60-744 Poznan, Poland
[27] Graduate School of Oceanography, University of Rhode Island, South Ferry Road, Narragansett, RI 02882-1197, USA
[28] Microbial & Molecular Biology, Scottish Association for Marine Science, Dunstaffnage Marine Laboratory, Oban, Argyll PA37 1QA, UK
[29] Department of Biology and Marine Biology, The University of North Carolina, 601 South College Road, Wilmington, NC 28403, USA
[30] Environmental Biophysics and Molecular Ecology Group, Institute of Marine and Coastal Sciences, Rutgers University, 71 Dudley Road, New Brunswick, NJ 08901 USA
[31] CNRS UMR7144, Station Biologique de Roscoff, Place George Teissier BP74, Roscoff, France
[$] Current address: Institute for Systems Biology, 1441 N 34th Street, Seattle, WA 98103, USA
[§] Current address: CEA, DSV, IBEB, SBVME, UMR 6191 CNRS/CEA/Université Aix-Marseille, Laboratoire de Bioénergétique et Biotechnologie des Bactéries et Microalgues, Cadarache, Saint-Paul-lez-Durance, F-13108 France

Diatoms are photosynthetic secondary endosymbionts found throughout marine and freshwater environments, and are believed to be responsible for around one fifth of the primary productivity on Earth [1-3]. Here we report the complete genome sequence of the marine pennate diatom *Phaeodactylum tricornutum.* By comparison with the sequence of the centric diatom *Thalassiosira pseudonana* [4-6] we have explored the evolutionary origins, functional significance, and ubiquity throughout diatoms of their gene repertoires. In spite of the fact that the pennate and centric lineages have only been diverging for 90 million years, their genome structures are dramatically different and a substantial fraction of genes (~40%) are not shared by these representatives of the two lineages. Analysis of molecular divergence compared with yeasts and metazoans reveals rapid rates of gene diversification in diatoms. Contributing factors include selective expansion of gene families, gains of introns, and differential losses of genes and mobilization of transposable elements. Most significantly, we document the unprecedented presence of hundreds of genes from bacteria. The ancient origins of these lateral gene transfers is testified by the finding that more than 300 are found in both diatoms, and many are likely to provide novel possibilities for metabolite management and for the perception of environmental signals. These findings go a long way toward explaining the incredible diversity and success of the diatoms in contemporary oceans.

The sequenced diatoms represent two of the major classes of diatoms – the Mediophyceae (bi- and multipolar centrics), to which *T. pseudonana* belongs, and the Bacillariophyceae (pennates), to which *P. tricornutum* belongs (Supplementary Fig. 1). The earliest fossil deposit from centrics is at 180 Ma and from pennates is at 90 Ma [7,8]. Although the youngest, the pennate group is by far the most diversified, and they are major components of both pelagic and benthic habitats [8]. They display a range of features, including their bilateral symmetry, that distinguish them from centric species. For example, they have amoeboid isogametes in contrast to motile sperm and oogamy observed in centric species. Members of

the raphid pennate clade also possess a raphe (Supplementary Fig. 1) that permits them to glide along surfaces, they are major biofoulers, they include toxic species, and they generally respond most strongly to mesoscale iron fertilization [8,9].

The completed *P. tricornutum* genome is approximately 27.4 megabases (Mb), slightly smaller than that of *T. pseudonana* (32.4 Mb), and gene density is slightly higher even though the *P. tricornutum* genome is predicted to contain fewer genes (10,402 against 11,776) (Table 1) (www.jgi.doe.gov/phaeodactylum and www.jgi.doe.gov/thalassiosira) (see Supplementary Information). Gene identification and functional analysis was facilitated by the availability of more than 130,000 Expressed Sequence Tags (ESTs) generated from cells grown in 16 different conditions (http://www.biologie.ens.fr/diatomics/EST3). In total, 86% of gene predictions had EST support (Supplementary Table 1). The genome was assembled into 33 large scaffolds ranging from 2.54 Mb to 88 kilobases (Kb), twelve of which contain telomeric repeats (CCCTAA) at both ends (see Supplementary Information and Supplementary Fig. 2).

A combination of *in silico* annotation and manual curation revealed that *P. tricornutum* shares 57% of its genes with *T. pseudonana* (see Supplementary Information for criteria used), of which 1,328 have not been found in other eukaryotes sequenced to date (Table 1). The molecular divergence between the two diatoms was assessed by examining the percent amino acid identity of 4,267 orthologous gene pairs (Fig. 1). We found an average identity of 54.9% between diatom orthologs, compared to approximately 43% between the diatoms and the oomycete *Phytophthora sojae*, in agreement with the predicted ancient separation (around 700 Ma) of these different heterokonts [10-12]. By comparing molecular divergence of orthologous pairs in hemiascomycetes and chordates, it emerges that the divergence between the two diatoms is similar to what is observed between *Saccharomyces cerevisiae* and *Kluyveromyces lactis*, and about halfway between *Homo sapiens*/*Takifugu rubripes* and *H. sapiens*/*Ciona intestinalis* (Fig. 1). The more rapid evolutionary rates of diatoms compared with other organismal groups (e.g., the fish:mammal divergence likely occurred in the Proterozoic era prior to 550 Ma [13]) is consistent with what had been previously observed in rRNA genes [14]. As has been found in the two yeasts [15,16] no major synteny could be detected between the two diatom genomes beyond a few examples of microclusters of up to eight genes (Supplementary Fig. 3). Furthermore, although intron lengths are similar, approximately two thirds of intron positions are unique to each species, with intron positions fully conserved in only 256 orthologs (see Supplementary Information). The widespread intron gain that has been reported in *T. pseudonana* [17] was not found in *P.*

tricornutum (Table 1), suggesting that it is a recent event in the centric diatom (or a secondary intron loss in the pennate diatom).

Large scale within-genome duplication events do not appear to have played a major role in driving the generation of diatom diversity (see Supplementary Information), in contrast to what has been found in yeasts and metazoans [18,19]. The observed high levels of diatom species diversity must therefore have been generated by other mechanisms. While intron gain may have been one factor in centric diatoms, the action of diatom-specific copia retrotransposable elements may also have contributed because we found that they have expanded dramatically in the *P. tricornutum* genome compared to *T. pseudonana* (Table 1; Supplementary Figs. 2 and 4). These elements also appear to have expanded in other pennate diatoms (see Supplementary Information) so they may have been a significant driving force in the generation of pennate diatom diversity through transpositional duplications and subsequent genome fragmentation. Furthermore, the vast majority of transposon insertions are found on only one of the two copies of the diploid genome. The maintenance of heterozygosity suggests that recombination may be suppressed, particularly at loci adjacent to heterozygous transposable elements, which could provide a novel means whereby gene diversity can be generated. In this context it may be significant that the two diatom genomes do not contain genes encoding the three subunits of the INO80 chromatin remodelling complex (INO80, ARP5 and ARP8), recently proposed to be involved in sister chromatid cohesion [20].

A wealth of evidence indicates that diatoms, and heterokonts in general, are derived from a secondary endosymbiotic event that took place around one billion years ago in which a red alga was engulfed (or invaded) by a heterotrophic eukaryote [21,22]. Diatom chloroplast genomes have fewer genes than red algal chloroplast genomes, indicating that a number of chloroplast genes were transferred to the nucleus after secondary endosymbiosis, and a few more genes appear to be in the process of transfer in one diatom species or the other [6]. It is generally thought that the diatom mitochondrion originated from the host, and the mitochondrial gene complement is almost identical to that of haptophytes and cryptophytes (data not shown), which may have originated from the same secondary endosymbiotic event. We used a phylogenomic approach (see Supplementary Information) to search for genes of red algal origin in the two diatoms and the two sequenced oomycetes, *P. ramorum* and *P. sojae* [11] using *Cyanidioschyzon merolae* as reference red algal genome [23]. One hundred and seventy one genes were classified as being of red algal origin based on strong (>85%) bootstrap support for the red alga plus stramenopile clade, and a larger number could be

identified if the level of stringency was reduced (Supplementary Information, Supplementary Table 2). Of the 171 high-scoring genes, 108 were shared between the two diatoms, and 74 (43%) were predicted to be plastid targeted. In addition, 11 of these genes were also present in oomycetes, as expected if the common ancestor of diatoms and oomycetes had a red algal plastid that was subsequently lost in the oomycetes [11]. The results of this survey support a red algal origin for the diatom plastid, and many gene transfers from the red algal nucleus to the host nucleus before the former was lost.

A remarkably high number of *P. tricornutum* predicted genes appear to have been transferred between diatoms and bacteria (784; 7.5% of gene models). Specifically, by searching for orthologous genes in 739 prokaryotic genomes, followed by automated phylogenetic tree construction using Apis (Automated Phylogenetic Inference System; see Supplementary Information) and manual curation, we could confirm that 587 putative *P. tricornutum* genes, outside of other chromalveolates, clustered with bacteria-only clades or formed a sister group to clades that included only bacterial genes. Another 200 sequences failed our alignment criteria for automated tree generation (less than 50% amino acid coverage or e>1-09) but had only bacterial genes in the Blast output (using a cutoff of e<1-05). These findings contrast dramatically with what is found in other chromalveolates [24] and in other eukaryotes in general, and indicate that horizontal gene transfer between bacteria and diatoms is pervasive. Of the 587 identified sequences, 42% are only found in *P. tricornutum* whereas 56% are present in both diatoms (Fig. 2A), testifying to their ancient origin. Only 14 sequences are shared between *P. tricornutum* and *Phytophthora spp.* (Fig. 2A, Supplementary Table 3), suggesting that the vast majority of gene transfers occurred after the divergence of photosynthetic heterokonts and oomycetes.

Many of the genes shared between diatoms and bacteria encode components that are likely to provide novel metabolic capacities, e.g., for organic carbon and nitrogen utilization (xylanases and glucanases, prismane, carbon-nitrogen hydrolase, amidohydrolase), functioning of the diatom urea cycle [4] (carbamoyl transferase, carbamate kinase, ornithine cyclodeaminase), and polyamine metabolism related to diatom cell wall silicification [25] (S-adenosylmethionine (SAM) -dependent decarboxylases and methyl transferases). Others are likely to encode novel cell wall components, and to provide unorthodox mechanisms of DNA replication, repair and recombination for a eukaryotic cell (topoisomerase, DNA primase, DNA ligase and helicase domain proteins) (Supplementary Table 3).

Bacterial genes in diatoms do not appear to be derived from any one specific source but from a range of origins including proteobacteria, cyanobacteria, and archaea (Fig. 2A,B,

Supplementary Table 3). Heterotrophic bacteria and cyanobacteria, especially diazotrophs and planctomycete bacteria, have been found in various intimate associations with diatoms as symbionts, endosymbionts and epibionts [26-28], which may explain the unprecedented levels of horizontal gene transfer that appears to have occurred between diatoms and bacteria. Furthermore, the close relationship between diatom and bacterial genes indicates that many of these events have occurred subsequent to secondary endosymbiosis. In *P. tricornutum*, bacterial genes are distributed throughout the genome, although several clusters can be observed, notably on Scaffolds 6 and 8, as can genomic deserts devoid of bacterial genes (e.g., Scaffold 31) (Supplementary Fig. 5). A further significant observation is that some of these genes in diatoms share bacterial-specific gene fusions that support phylogenetic associations, such as assimilatory nitrite reductase B and D subunits; apparently of planctomycete origin (Fig. 2C).

Bacterial histidine kinase-based phosphorelay two-component systems (TCS) also appear to be highly developed in diatoms. For example, *P. tricornutum* contains a wide range of two-component signalling proteins sometimes organized in novel domain associations (Fig. 3). One of these proteins bears the classical features of bacterial phytochrome photoreceptors, as was previously noted in *T. pseudonana* [4,5]. Another domain combination present in both diatoms strongly resembles aureochrome blue-light photoreceptors [29]. Furthermore, *P. tricornutum* contains orthologs of LovK, a PAS domain-containing histidine kinase that was recently found to regulate light-dependent attachment to substrata in bacteria [30], and other light-dependent histidine kinases that have been reported in bacteria [31]. The fact that *T. pseudonana* does not contain any LovK orthologs is consistent with its pelagic lifestyle.

To identify additional novel features of the diatom gene repertoire we compared the gene family content of the two diatoms with other eukaryotes (Fig. 4, Supplementary Figs. 6 and 7, Supplementary Table 4). Diatoms contain many species-specific multicopy gene families (287 families in *P. tricornutum* and 259 in *T. pseudonana*, consisting of 943 and 716 genes, respectively), as well as large numbers of species-specific single copy genes (denoted orphans in Fig. 4A). The higher number of species-specific gene families in *P. tricornutum* may suggest that the more recently evolved pennate diatoms possess more specialized functions, perhaps related to the heterogeneity of the benthic environments that they commonly inhabit. The centric diatom, by contrast, has retained more features found in other eukaryotes (Fig. 4B, Table 1), such as the flagellar apparatus [32]. We found a similar number of diatom-specific gene families (1,011) and eukaryotic gene families not found in diatoms (1,062), revealing that the rates of gene gain and gene loss are very similar and consistent

with the high diversification rates observed in diatoms. We also found that diatom-specific genes are evolving faster than other genes in diatom genomes (Fig. 4C), providing a further explanation for the rapid evolutionary rates found in diatoms.

Of the gene families found in the diatoms, many of them contain higher numbers of genes compared with other eukaryotes (Supplementary Table 4, Supplementary Fig. 7). Examination of these classes reveals several interesting features (see Supplementary Information), including the over-representation of genes involved in polyamine metabolism. The expansion of polyamine-related components is of interest considering the role of long chain polyamines (LCPA) in silica nanofabrication [25]. Of the eight predicted spermine/spermidine synthase-like genes in *P. tricornutum*, three encode potentially bi-functional enzymes bearing both an aminopropyltransferase domain and a SAM decarboxylase domain. In *T. pseudonana* four of the nine genes are of this type. Although the bi-functional nature of these genes is not unprecedented, it has only been found previously in two bacteria (*Bdellovibrio bacteriovorus* and *Delftia acidovorans*). In addition, a number of these putative enzymes contain a hydrophobic N-terminal domain that is predicted to be either a transmembrane domain or a signal peptide for co-translational import into the endoplasmic reticulum. Other noteworthy diatom-specific expansions include histidine kinases (see above and Fig. 3), heat shock transcription factors (HSFs), and cyclins.

For the putative heat shock transcription factors, we found 69 copies in *P. tricornutum* and 89 copies in *T. pseudonana* [5] (Supplementary Information). These numbers are remarkable considering that they represent close to 50% of the total number of transcription factors in the two sequenced diatoms. The significance of this expansion of HSFs in diatoms is unclear, but because these transcription factors are typically involved in stress responses, our findings may indicate that transcriptional regulation is a major mechanism acting to control responses to stress in these organisms. EST data indicates that the majority of these genes are expressed and that some are induced specifically in response to certain growth conditions (Supplementary Fig. 8).

Another diatom-specific gene family expansion encodes cyclins, major regulators of the cell cycle in other eukaryotes [33]. In this case, 10 and 42 diatom-specific cyclin genes have been found in the *P. tricornutum* and *T. pseudonana* genomes, respectively, in addition to members of each of the canonical families of cyclins. The function of this new class of cyclins must await experimental investigation, although we have already found that in *P. tricornutum* the majority are expressed at specific stages of the cell cycle (data not shown). The dramatic expansion of this gene family may reflect the unusual characteristics of diatom life cycles due

to the rigid nature of their cell wall, such as the control of cell size reduction, the activation of sexual reproduction at a critical size threshold, and life in rapidly changing and unpredictable environments. Conversely, it may be significant that genes encoding RCC1 proteins (Regulators of Chromosome Condensation), also involved in cell cycle control [34], have been expanded in both diatom genomes (Supplementary Table 4).

In conclusion, through our comparative analyses we have revealed diverse origins of diatom genes. Diatom-specific genes may have arisen by genome rearrangements and subsequent domain recombinations due to the action of diatom-specific transposable elements and from selective gene family expansions and constrictions. The maintenance of the diploid chromosomes in a heterozygous condition provides an additional means to promote gene diversification, and may imply that recombination mechanisms are somewhat relaxed in diatoms. It was previously shown that diatoms have retained genes from both partners of the secondary endosymbiosis [4], thus bringing together primary metabolic processes such as photosynthetic carbon fixation and organic nitrogen production via the urea cycle in a single organism [35]. Our studies now reveal that genes acquired after secondary endosymbiosis by gene transfer from bacteria are pervasive in diatoms and represent at least 5% of their gene repertoires. Although horizontal gene transfer between bacteria is now established as a common event [36], it is much rarer in eukaryotes and has only been found in specialized instances such as in obligate pathogens [37-39] and as a result of transfer from intracellular bacteria in Drosophila [40], and at much lower levels than reported here. Our data suggest that gene transfer from bacteria to diatoms and perhaps vice versa has been a common event in marine environments and has been a major driving force during diatom evolution. It has also brought together highly unorthodox combinations of genes permitting non-canonical management of carbon and nitrogen in primary metabolism and the sensing of external stimuli adapted to aquatic environments. The presence of nitrite reductase and carbamate kinase, which bring novel capabilities in nitrogen metabolism, together with the unusual configurations of two-component signalling components, are examples of bacterial genes that are likely to perform useful functions in a eukaryotic context. We propose that this combination of mechanisms may underlie the rapid evolution and diversification rates observed in diatoms and may explain why they have come to dominate contemporary marine ecosystems.

Acknowledgements

Diatom genome sequencing at the Joint Genome Institute (Walnut Creek, CA, USA) was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396. *P. tricornutum* ESTs were generated at Genoscope (Evry, Paris). Funding for this work was also obtained from the EU-funded FP6 Diatomics project (LSHG-CT-2004-512035), the EU-FP6 Marine Genomics Network of Excellence (GOCE-CT-2004-505403), an ATIP "Blanche" grant from CNRS, and the Agence Nationale de la Recherche (France). We are grateful to Mathieu Muffato and Hugues-Roest Crollius for the analysis reported in Supplementary Fig. 3A.

References

1. **Smetacek, V. Diatoms and the ocean carbon cycle. *Protist* 150, 25-32 (1999).**
2. **Falkowski, P. G., Barber, R. T. & Smetacek, V. Biogeochemical controls and feedbacks on ocean primary production. *Science* 281, 200-206 (1998).**
3. **Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* 281, 237-40 (1998).**
4. **Armbrust, E. V. et al. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306, 79-86 (2004).**
5. **Montsant, A. et al. Identification and comparative genomic analysis of signaling and regulatory components in the diatom Thalassiosira pseudonana. *J. Phycol.* 43, 585-603 (2007).**
6. **Oudot-Le Secq, M.-P. et al. Chloroplast genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*: Comparison with other plastid genomes of the red lineage. *Mol. Gen. Genom.* 277, 427-439 (2007).**
7. **Sims, P. A., Mann, D. G. & Medlin, L. K. Evolution of the diatoms: Insights from fossil, biological and molecular data. *Phycologia* 45, 361-402 (2006).**
8. **Kooistra, W. H. C. F., Gersonde, R., Medlin, L. K. & Mann, D. G. in *The origin and evolution of the diatoms: their adaptation to a planktonic existence* 207-249 (2007).**
9. **de Baar, H. J. W. et al. Synthesis of iron fertilization experiments: From the iron age in the age of enlightenment. *Journal of Geophysical Research-Oceans* 110 (2005).**
10. **Baldauf, S. L. The deep roots of eukaryotes. *Science* 300, 1703-1706 (2003).**
11. **Tyler, B. M. et al. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313, 1261-6 (2006).**
12. **Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G. & Bhattacharya, D. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol.* 21, 809-818 (2004).**

13. **Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* 392, 917-920 (1998).**
14. **Kooistra, W. H. C. F. & Medlin, L. K. Evolution of the diatoms (Bacillariophyta): IV. A reconstruction of their age from small subunit rRNA coding regions and the fossil record. *Molecular Phylogenetics and Evolution* 6, 391-407 (1996).**
15. **Dujon, B. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet* 22, 375-87 (2006).**
16. **Dujon, B. et al. Genome evolution in yeasts. *Nature* 430, 35-44 (2004).**
17. **Roy, S. W. & Penny, D. A very high fraction of unique intron positions in the intron-rich diatom *Thalassiosira pseudonana* indicates widespread intron gain. *Mol Biol Evol* 24, 1447-57 (2007).**
18. **Scannell, D. R., Byrne, K. P., Gordon, J. L., Wong, S. & Wolfe, K. H. Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature* 440, 341-345 (2006).**
19. **Semon, M. & Wolfe, K. H. Reciprocal gene loss between Tetraodon and zebrafish after whole genome duplication in their ancestor. *Trends Genet.* 23, 108-112 (2007).**
20. **Ogiwara, H., Enomoto, T. & Seki, M. The INO80 chromatin remodeling complex functions in sister chromatid cohesion. *Cell Cycle* 6, 1090-1095 (2007).**
21. **Bhattacharya, D., Archibald, J. M., Weber, A. P. & Reyes-Prieto, A. How do endosymbionts become organelles? Understanding early events in plastid evolution. *Bioessays* 29, 1239-1246 (2007).**
22. **Keeling, P. A brief history of plastids and their hosts. *Protist* 155, 3-7 (2004).**
23. **Matsuzaki, M. et al. Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature* 428, 653-7 (2004).**
24. **Martens, C., Vandepoele, K. & Van de Peer, Y. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc Natl Acad Sci U S A.* 105, 3427-3432 (2008).**
25. **Kroger, N., Deutzmann, R., Bergsdorf, C. & Sumper, M. Species-specific polyamines from diatoms control silica morphology. *Proc. Natl. Acad. Sci. U S A* 97, 14133-14138 (2000).**
26. **Carpenter, E. J. & Janson, S. Intracellular cyanobacterial symbionts in the marine diatom *Climacodium frauenfeldianum* (Bacillariophyceae). *J. Phycol.* 36, 540-544 (2000).**
27. **Schmid, A.-M. M. Endobacteria in the diatom *Pinnularia* (Bacillariophyceae). I. Scattereed ct-nucleoids explained: DAPI-DNA complexes stem from exoplastidial bacteria boring into the chloroplasts. *J. Phycol.* 39, 122-138 (2003).**
28. **Zehr, J. P., Carpenter, E. J. & Villareal, T. A. New perspectives on nitrogen-fixing microorganisms in tropical and subtropical oceans. *Trends Microbiol.* 8, 68-73 (2000).**
29. **Takahashi, F. et al. AUREOCHROME, a photoreceptor required for photomorphogenesis in stramenopiles. *Proc Natl Acad Sci U S A.* 104, 19625-19630 (2007).**
30. **Purcell, E. B., Siegal-Gaskins, D., Rawling, D. C., Fiebig, A. & Crosson, S. A photosensory two-component system regulates bacterial cell attachment. *Proc Natl Acad Sci U S A.* 104, 18241-18246 (2007).**
31. **Swartz, T. E. et al. Blue-light-activated histidine kinases: two-component sensors in bacteria. *Science* 317, 1090-1093 (2007).**
32. **Merchant, S. S. et al. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 318, 245-250 (2007).**

33. Bloom, J. & Cross, F. R. Multiple levels of cyclin specificity in cell-cycle control. *Nature Reviews Molecular Cell Biology* 8, 149-160 (2007).
34. Moore, J. D. The Ran-GTPase and cell-cycle control. *Bioessays* 23, 77-85 (2001).
35. Allen, A. E., Vardi, A. & Bowler, C. An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Curr Opin Plant Biol* 9, 264-73 (2006).
36. Snel, B., Bork, P. & Huynen, M. A. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* 12, 17-25 (2002).
37. Opperdoes, F. R. & Michels, P. A. Horizontal gene transfer in trypanosomatids. *Trends Parasitol.* 23, 470-476 (2007).
38. Loftus, B. et al. The genome of the protist parasite Entamoeba histolytica. *Nature* 433, 865-868 (2005).
39. Carlton, J. M. et al. Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science* 315, 207-212 (2007).
40. Hotopp, J. C. et al. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. *Science* 317, 1753-1756 (2007).
41. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52, 696-704 (2003).
42. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21, 2104-2105 (2005).

Figure Legends

Figure 1 Molecular divergence between *P. tricornutum* and *T. pseudonana*.

A. Summary of numbers of orthologous pairs (reciprocal best hits at $e<10^{-10}$) for each organism comparison and their mean percentage identities.

B. Analysis of molecular divergence between the diatoms and other heterokonts, and comparison with selected hemiascomycetes and chordates. The diatom:oomycete pair displays the lowest amino acid identity (43.3%), in agreement with their proposed ancient separation, around 700 Ma [12]. The divergence between the pennate and centric diatom is very similar to the fish:mammal divergence, which likely occurred in the Proterozoic era (550 Ma) [13]. The centric:pennate divergence, on the other hand, has been dated to at least 90 Ma [8]. In the figure, we represent the cumulative frequencies of amino acid identity across each set of potential orthologous pairs.

Figure 2 Bacterial genes in diatoms.

A. Venn diagrams showing how many of the bacterial genes identified in *P. tricornutum* are also found in other heterokonts (left), and which bacterial classes are most related phylogenetically (right). In each case, the venn diagrams indicate the number of trees in which the designated taxa occur within the same clade or in a sister clade of *P. tricornutum*.

B. Breakdown of different bacterial groups that occur in the same clade or in a sister clade of *P. tricornutum*. Unique denotes a gene found only in a particular bacterial class, Shared denotes a gene that is most similar to a gene of that specific bacterial class but that is also present in other bacterial groups.

C. PhyML [41] maximum likelihood tree (-loglk=22358.321320) as inferred from the amino acid sequences of the large subunit of NAD(P)H assimilatory nitrite reductase (nirB). The choice of model was WAG with gamma-distributed rates (alpha=0.80), as suggested by a ProtTest [42] analysis of the alignment. Numbers above selected branches indicate ML bootstrap support (100 replicates). In most cases, the large (nirB) and small (nirD) subunits of NAD(P)H assimilatory nitrite reductase are encoded by distinct ORFs, but in diatoms and planctomycetes the nirD and nirB ORFs have been fused to encode a single gene product. A total of 587 trees show evidence for prokaryotic origins of diatom genes and are available in Supplementary Information as a supplementary file.

Figure 3 Domain structures of two-component systems (TCS) found in *P. tricornutum*. Domains are illustrated schematically and *P. tricornutum* Protein IDs are indicated on the left. Proteins corresponding to putative photoreceptors (aureochrome, phytochrome and LovK) are indicated first (in grey, above the horizontal line). Different domains likely to be involved in signalling are indicated schematically. For further information about TCS see Supplementary Information. Domain abbreviations are PAS: Per/Arnt/Sim, B-ZIP: Basic region Leucine Zipper, GAF: cGMP phosphodiesterase/Adenylyl cyclase/FhlA, PHY: Phytochrome, HK: Histidine Kinase, RR: Response Regulator, LRR: Leucine-Rich Repeat, LUX R: LuxR transcriptional activator, CHASE: Cyclases/Histidine kinases Associated Sensory Extracellular.

Figure 4 Shared and unique gene families.

A. Venn diagram representation of shared/unique gene families in *P. tricornutum*, *T. pseudonana*, Viridiplantae (i.e., plants and green algae) & red algae, and other eukaryotes (i.e., other chromalveolates and Opisthokonta (i.e., fungi and metazoa)). In addition to the total number of gene families specific to *P. tricornutum* and *T. pseudonana*, the number of families consisting of a single gene (denoted 'orphans') is also indicated. For example, of the 3,710 gene families that are only found in *P. tricornutum*, 3,423 consist of single copy genes whereas 287 gene families have at least two members.

B. Venn diagram of the distribution of *P. tricornutum* (left) and *T. pseudonana* (right) gene families with homology to proteins from the Viridiplantae & red algae, Opisthokonta and other chromalveolates (including the other diatom). The numbers outside the circles indicate the number of *P. tricornutum* (left) or *T. pseudonana* (right) gene families with no homology to the examined proteomes.

C. Percent amino acid identity plot of orthologs (based on reciprocal best hits) of different classes of diatom genes identified in A. Numbers in parentheses indicate the number of orthologs per class. 'Diatom' corresponds to genes only found in *P. tricornutum* and *T. pseudonana* (members of the 1,011 gene families shown in A); 'core' corresponds to genes present in all eukaryotic groups (members of the 1,666 gene families shown in A), and 'all' corresponds to all orthologous gene pairs in *P. tricornutum* and *T. pseudonana*.

Table 1 Major features of the *P. tricornutum* and *T. pseudonana* genomes.

	P. tricornutum	*T. pseudonana*
Genome size	27.4 Mb	32.4 Mb
Predicted genes	10,402	11,776
Core genes*	3,523	4,332
Diatom-specific genes*	1,328	1,407
Unique genes*	4,366	3,912
Introns	8,169	17,880
Introns/gene	0.79	1.52
LTR retrotransposon content	5.8%	1.1%

* Different classes of genes were assigned by comparing the *P. tricornutum* and *T. pseudonana* predicted proteomes with those from two plants, three green algae, one red alga, three metazoans, two fungi, and ten other chromalveolates (see Supplementary Information) by all-against-all BLASTP using an E-value cutoff of E-5. Core genes are defined as being present in representatives from all these eukaryotic groups, diatom-specific genes are only present in both of the diatoms but not elsewhere, and unique genes are only found in one of the two diatoms. The different numbers of diatom-specific genes in the two diatoms is a consequence of species-specific gene duplication events.

A

Pairwise comparison			Mean identity %	Number of compared pairs
Phaeodactylum tricornutum	/	*Thalassiosira pseudonana*	54.9	4,267
Phaeodactylum tricornutum	/	*Phytophthora sojae*	43.3	2,952
Saccharomyces cerevisiae	/	*Debaryomyces hansenii*	50.1	2,694
Saccharomyces cerevisiae	/	*Kluyveromyces lactis*	54.8	4,246
Saccharomyces cerevisiae	/	*Candida glabrata*	58.2	4,484
Homo sapiens	/	*Ciona intestinalis*	52.6	5,208
Homo sapiens	/	*Takifugu rubripes*	61.4	10,225

B

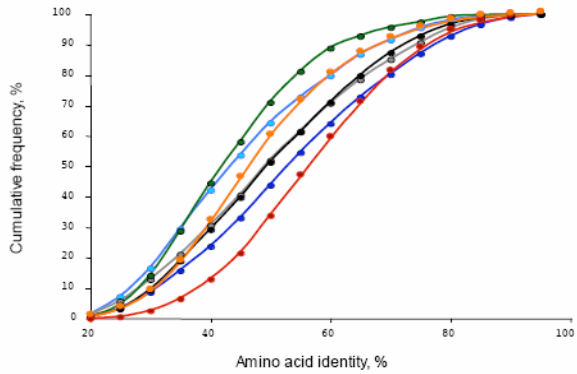


Figure 1

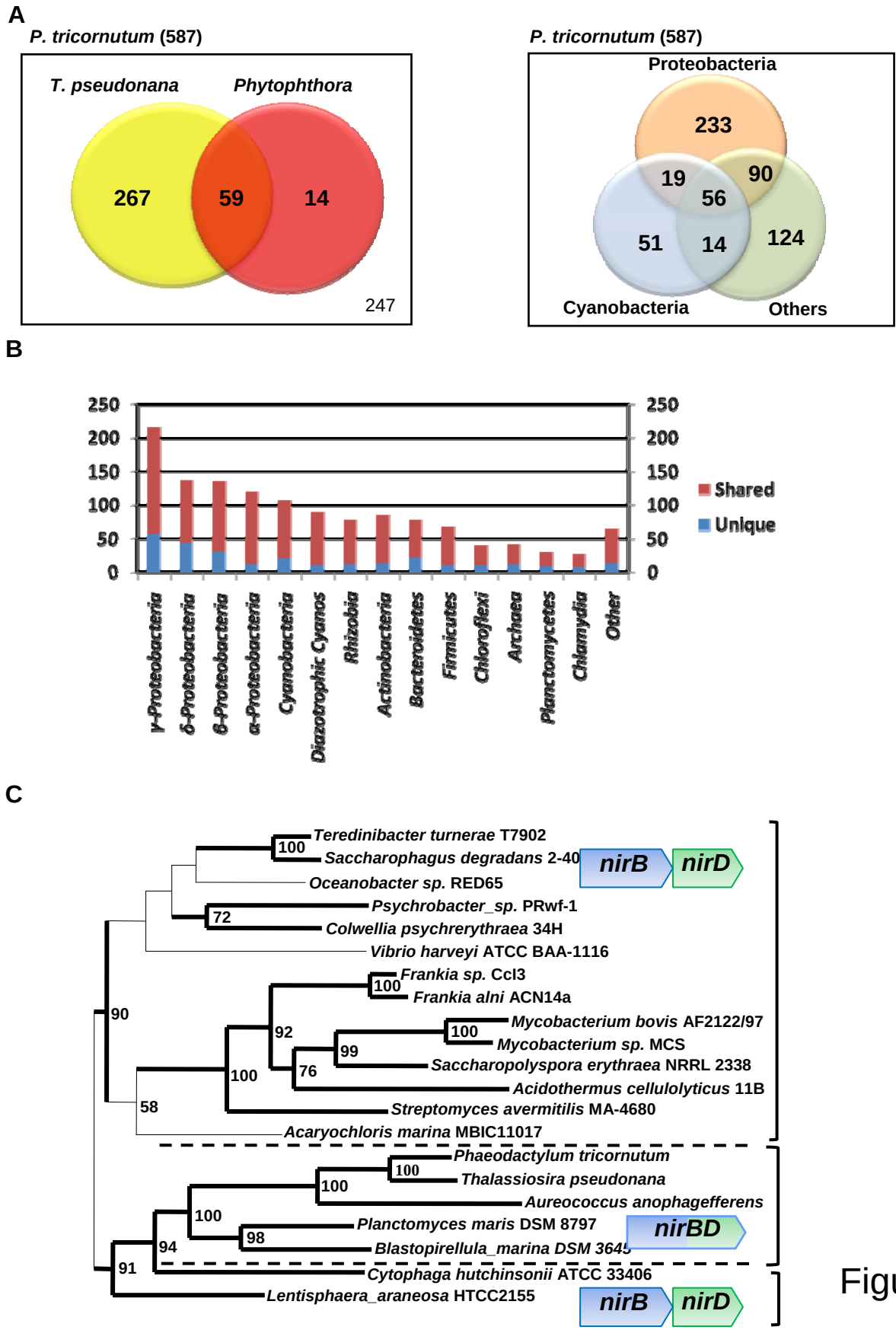

A
P. tricornutum (587)
T. pseudonana
Phytophthora
267
59
14
247
P. tricornutum (587)
Proteobacteria
233
19
90
56
51
14
124
Cyanobacteria
Others
B
250
200
150
100
50
0
Shared
Unique
γ-Proteobacteria
δ-Proteobacteria
β-Proteobacteria
α-Proteobacteria
Cyanobacteria
Diazotrophic Cyanos
Rhizobia
Actinobacteria
Bacteroidetes
Firmicutes
Chloroflexi
Archaea
Planctomycetes
Chlamydia
Other
C
Teredinibacter turnerae T7902
Saccharophagus degradans 2-40
Oceanobacter sp. RED65
Psychrobacter_sp. PRwf-1
Colwellia psychrerythraea 34H
Vibrio harveyi ATCC BAA-1116
Frankia sp. CcI3
Frankia alni ACN14a
Mycobacterium bovis AF2122/97
Mycobacterium sp. MCS
Saccharopolyspora erythraea NRRL 2338
Acidothermus cellulolyticus 11B
Streptomyces avermitilis MA-4680
Acaryochloris marina MBIC11017
Phaeodactylum tricornutum
Thalassiosira pseudonana
Aureococcus anophagefferens
Planctomyces maris DSM 8797
Blastopirellula_marina DSM 3645
Cytophaga hutchinsonii ATCC 33406
Lentisphaera_araneosa HTCC2155
100
72
100
100
92
99
76
100
90
58
100
100
100
98
94
91
nirB
nirD
nirBD
nirB
nirD

Figure 2

Figure 3

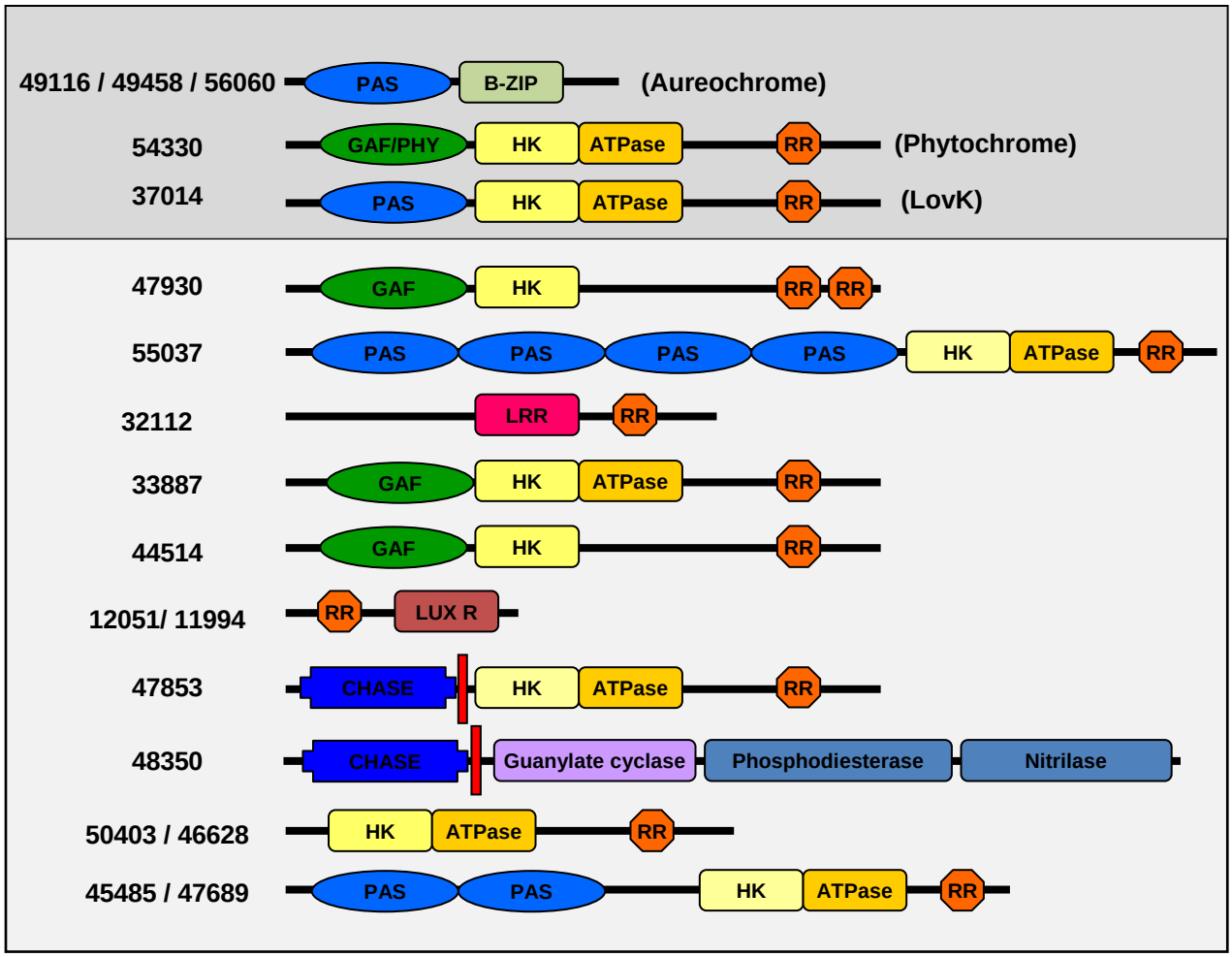
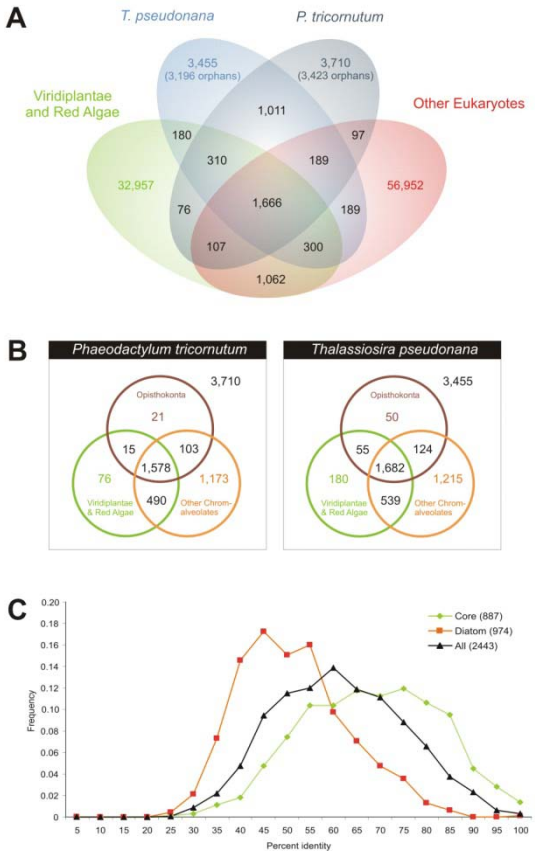

49116 / 49458 / 56060
PAS
B-ZIP
(Aureochrome)
54330
GAF/PHY
HK
ATPase
RR
(Phytochrome)
37014
PAS
HK
ATPase
RR
(LovK)
47930
GAF
HK
RR
RR
55037
PAS
PAS
PAS
PAS
HK
ATPase
RR
32112
LRR
RR
33887
GAF
HK
ATPase
RR
44514
GAF
HK
RR
12051/ 11994
RR
LUX R
47853
CHASE
HK
ATPase
RR
48350
CHASE
Guanylate cyclase
Phosphodiesterase
Nitrilase
50403 / 46628
HK
ATPase
RR
45485 / 47689
PAS
PAS
HK
ATPase
RR

Figure 4


A
T. pseudonana
P. tricornutum
3,455
(3,196 orphans)
3,710
(3,423 orphans)
Viridiplantae
and Red Algae
Other Eukaryotes
1,011
180
97
310
189
32,957
56,952
76
1,666
189
107
300
1,062
B
Phaeodactylum tricornutum
Thalassiosira pseudonana
3,710
Opisthokonta
21
15
103
1,578
76
1,173
490
Viridiplantae & Red Algae
Other Chrom-alveolates
3,455
Opisthokonta
50
55
124
1,682
180
1,215
539
Viridiplantae & Red Algae
Other Chrom-alveolates
C
Core (887)
Diatom (974)
All (2443)
Frequency
Percent identity