Nature
November 2008; Volume 456 (7219) : Pages 239-244
http://dx.doi.org/10.1038/nature07410

The *Phaeodactylum* genome reveals the 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[10,32], Klaus Valentin[7], Fréderic Verret[11], John A. Berges[12], Colin Brownlee[11], Jean-Paul Cadoret[13], Anthony Chiovitti[14], Chang Jae Choi[12], Sacha Coesel[2,32], Alessandra De Martino[1], J. Chris Detter[5], Colleen Durkin[10], Angela Falciatore[2], Jérome Fournet[15], Miyoshi Haruta[16], Marie J. J. Huysman[6,17,] Bethany D. Jenkins[18], Katerina Jiroutova[19,] Richard E. Jorgensen[20], Yolaine Joubert[15], Aaron Kaplan[21,] Nils Kröger[22], Peter G. 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[7,11], Anton Montsant[1,2], Marie-Pierre Oudot–Le Secq[24], Carolyn Napoli[20], Miroslav Obornik[19], Micaela Schnitzler Parker[10], Jean-Louis Petit[9], Betina M. Porcel[9], Nicole Poulsen[25], Matthew Robison[16], Leszek Rychlewski[26], Tatiana A. Rynearson[27], Jeremy Schmutz[4], Harris Shapiro[5], Magali Siaut[2,32], Michele Stanley[28], Michael R. Sussman[16], Alison R. 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]

1. CNRS UMR8186, Department of Biology, Ecole Normale Supérieure, 46 rue d'Ulm, 75005 Paris, France
2. Stazione Zoologica 'Anton Dohrn', Villa Comunale, I-80121 Naples, Italy
3. J. Craig Venter Institute, San Diego, California 92121, USA
4. Joint Genome Institute-Stanford, Stanford Human Genome Center, 975 California Avenue, Palo Alto, California 94304, USA
5. Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, California 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, 27570 Bremerhaven, Germany
8. Fachbereich Biologie, University of Konstanz, 78457 Konstanz, Germany
9. Genoscope, CEA-Institut de Génomique, UMR CNRS no. 8030, 2 rue Gaston Crémieux, 91057 Evry Cedex, France
10. School of Oceanography, University of Washington, Seattle, Washington 98195, USA
11. Marine Biological Association of the UK, The Laboratory, Citadel Hill, Plymouth PL1 2PB, UK
12. Department of Biological Sciences, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin 53201, 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 et Techniques, Université de Nantes, 2 rue de la Houssinière, 44322, BP 92208, 44322 Nantes Cedex 3, France
16. University of Wisconsin Biotechnology Center, 425 Henry Mall, Madison, Wisconsin 53706, USA
17. Laboratory of Protistology and Aquatic Ecology, Ghent University, Krijgslaan 281-S8, B-9000 Ghent, 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. Bio5 Institute and Department of Plant Sciences, University of Arizona, Tucson, Arizona 85719, USA
21. Department of Plant and Environmental Sciences, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel

22. School of Chemistry and Biochemistry, School of Materials Science and Engineering, School of Biology, Georgia Institute of Technology, 901 Atlantic Drive NW, Atlanta, Georgia 30332-0400, USA
23. Leibniz-Institut für Meereswissenschaften, 24105 Kiel, Germany
24. Department of Botany, University of British Columbia, 3529-6270 University Boulevard, Vancouver, British Columbia V6T 1Z4, Canada
25. School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, Georgia 30332-0400, USA
26. BioInfoBank Institute, Limanowskiego 24A/16, 60-744 Poznan, Poland
27. Graduate School of Oceanography, University of Rhode Island, South Ferry Road, Narragansett, Rhode Island 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 Wilmington, 601 South College Road, Wilmington, North Carolina 28403, USA
30. Environmental Biophysics and Molecular Ecology Group, Institute of Marine and Coastal Sciences, Rutgers University, 71 Dudley Road, New Brunswick, New Jersey 08901, USA
31. CNRS UMR7144, Station Biologique de Roscoff, Place George Teissier BP74, 29682 Roscoff Cedex, France
32. Present addresses: University of East Anglia, School of Environmental Sciences, Norwich NR4 7TJ, UK (T.M.); Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103, USA (S.C.); 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 (M.S.).

*: Corresponding author : C. Bowler, email address : cbowler@biologie.ens.fr

---

Abstract:

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, 2]. The genome sequence of the marine centric diatom *Thalassiosira pseudonana* was recently reported, revealing a wealth of information about diatom biology[3, 4, 5]. Here we report the complete genome sequence of the pennate diatom *Phaeodactylum tricornutum* and compare it with that of *T. pseudonana* to clarify evolutionary origins, functional significance and ubiquity of these features throughout diatoms. 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 gene family expansions, differential losses and gains of genes and introns, and differential mobilization of transposable elements. Most significantly, we document the presence of hundreds of genes from bacteria. More than 300 of these gene transfers are found in both diatoms, attesting to their ancient origins, and many are likely to provide novel possibilities for metabolite management and for perception of environmental signals. These findings go a long way towards explaining the incredible diversity and success of the diatoms in contemporary oceans.

Introduction

The sequenced diatoms represent two of the major classes of diatoms – the bi/multipolar centrics (Mediophyceae), to which T. pseudonana belongs, and the pennates (Bacillariophyceae), 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 6,7. Although the youngest, the pennates are by far the most diversified, and they are major components of both pelagic and benthic habitats 7. 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, they are major biofoulers, they include toxic species, and they generally respond most strongly to mesoscale iron fertilization [7,8]. Furthermore, members of the raphid pennate clade can glide actively along surfaces.

The completed *P. tricornutum* genome is approximately 27.4 megabases (Mb), slightly smaller than *T. pseudonana* (32.4 Mb), and *P. tricornutum* is predicted to contain fewer genes (10,402 against 11,776) (Table 1 and Supplementary Fig. 2). 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. In total, 86% of gene predictions had EST support (Supplementary Table 1).

P. tricornutum shares 57% of its genes with *T. pseudonana* (see Supplementary Information for criteria used), of which 1,328 are absent from other sequenced eukaryotes (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 a more distantly related heterokont, the non-photosynthetic oomycete *Phytophthora sojae*. This agrees with the predicted ancient separation (around 700 Ma) of these lineages [9,10]. The divergence between the two diatoms is similar to what is observed between *Saccharomyces cerevisiae* and the related yeast *Kluyveromyces lactis*, and about halfway between *Homo sapiens*/*Takifugu rubripes* (pufferfish) and *H. sapiens*/*Ciona intestinalis* (seasquirt) (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 [11]) is consistent with previous observations [6,7]. As has been found in the two yeasts [12] no major conservation of gene order (synteny) could be detected between the two diatom genomes beyond a few examples of microclusters of up to eight genes (Supplementary Fig. 3). Furthermore, approximately two thirds of intron positions are unique to each species (see Supplementary Information). The widespread intron gain that has been reported in *T. pseudonana* [13] was not found in *P. tricornutum* (Table 1), suggesting that it may be a recent event in the centric 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 [14,15]. The observed high levels of diatom species diversity must therefore have been generated by other mechanisms. While intron gain may be one factor in centric diatoms, the dramatic expansion of diatom-specific copia

retrotransposable elements may have contributed in the *P. tricornutum* genome (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.

Diatoms, and heterokonts in general, are believed to be derived from a secondary endosymbiotic process that took place around one billion years ago between a red alga and a heterotrophic eukaryote [16]. 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 [5]. 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 (other heterokonts) (data not shown), which may have originated from the same secondary endosymbiotic event. We used a phylogenomic approach to search for genes of red algal origin in the two diatoms and the two sequenced oomycetes, *P. ramorum* and *P. sojae* [9] using *Cyanidioschyzon merolae* as reference red algal genome [17]. 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 heterokont clade (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 [9]. 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 and manual curation, we could confirm that 587 putative *P. tricornutum* genes clustered with bacteria-only clades or formed a sister group to clades that included only bacterial genes (with or without other heterokonts). This finding indicates that horizontal gene transfer between bacteria and diatoms is pervasive, and is much higher than has been found in other sequenced eukaryotes [18,19]. 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 [20](xylanases and glucanases, prismane, carbon-nitrogen hydrolase, amidohydrolase), functioning of the diatom urea cycle [3] (carbamoyl transferase, carbamate kinase, ornithine cyclodeaminase), and polyamine metabolism related to diatom cell wall silicification [21] (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 (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 [22-24], which may explain the unprecedented levels of HGT events that appear to have occurred. In *P. tricornutum*, bacterial genes are distributed throughout the genome, although several clusters can be observed, as well as regions devoid of bacterial genes (Supplementary Fig. 5). 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), involved in environmental sensing, 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 previously noted in *T. pseudonana* [3,4]. Another domain combination present in both diatoms resembles aureochrome blue-light photoreceptors [25], and *P. tricornutum* contains orthologs of LovK and other light-dependent histidine kinases reported in bacteria [26,27].

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). Diatoms contain many species-specific multicopy gene families, 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 recent 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 [28]. 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 diatom divergence rates [6,7].

Of the gene families found in the diatoms, some contain higher numbers of genes compared with other eukaryotes (Supplementary Table 4, Supplementary Fig. 7), e.g., 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 [21]. 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. 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*). Silaffins and silacidins are proteins/peptides involved in diatom silica formation [29]. *P. tricornutum* contains only one silaffin-like protein, and no homologues to silacidin. Frustulin genes, encoding proteins that form organic constituents of the biosilica cell wall but are not involved in silica formation, are present in large numbers and are highly expressed. Both diatoms contain a similar number of silicon transporters.

Other noteworthy diatom-specific expansions include histidine kinases (see above and Fig. 3), cyclins, and heat shock transcription factors (HSFs). Cyclins are major regulators of the cell cycle in eukaryotes. In addition to members of each of the canonical families of cyclins, we found 10 and 42 diatom-specific cyclin genes in *P. tricornutum* and *T. pseudonana*, respectively. 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 [7]. Conversely, it may be significant that genes encoding RCC1 proteins (Regulators of Chromosome Condensation), also involved in cell cycle control, have been expanded in both diatom genomes (Supplementary Table 4). For the putative HSFs we found 69 copies in *P. tricornutum* and 89 copies in *T. pseudonana* [4]. These numbers represent close to 50% of the total number of

transcription factors in the two sequenced diatoms. The significance of this expansion is unclear, but EST data indicates that the majority are expressed and that some are induced specifically in response to certain growth conditions (Supplementary Fig. 8).

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, from selective gene family expansions and constrictions, and intron gain/loss. It was previously shown that diatoms have retained genes from both partners of the secondary endosymbiosis [3], thus bringing together primary metabolic processes such as photosynthetic carbon fixation and organic nitrogen production via the urea cycle in a single organism [30]. Our studies now suggest that genes acquired after secondary endosymbiosis by gene transfer from bacteria are pervasive in diatoms and represent at least 5% of their gene repertoires. This level of horizontal gene transfer is around one order of magnitude higher than has been found in other eukaryotes, and is similar to the rates found between bacteria [19]. Although our analyses may be biased by the currently poor taxon sampling of whole genome sequences in eukaryotes compared with prokaryotes, they are nonetheless supported by molecular phylogenies. We therefore propose 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 combination of mechanisms reported here may underlie the rapid diversification rates observed in diatoms and may explain why they have come to dominate contemporary marine ecosystems in a relatively short period of time.

Methods Summary

High molecular weight DNA was extracted from axenic cultures of *P. tricornutum* accession Pt1 8.6 (deposited as CCMP2561 in the Provasoli-Guillard National Center for Culture of Marine Phytoplankton) and used to construct replicate libraries containing inserts of 2-3 Kb, 6-8 Kb, and 35-40 Kb. Using the JGI JAZZ assembler, approximately 556,000 reads involving 564 Mb of sequence were trimmed, filtered for short reads, and assembled. All low quality areas and gaps were identified and converted into targets for manual finishing. The draft genome sequence of *T. pseudonana* [3] was finished in a similar way. Both diatom genomes were annotated using the JGI annotation pipeline, which combines several gene

prediction, annotation and analysis tools. Assemblies and annotations of each genome are available through the JGI Genome Portal at www.jgi.doe.gov/phaeodactylum and www.jgi.doe.gov/thalassiosira. cDNA libraries were constructed from mRNA extracted from *P. tricornutum* cultures grown in sixteen different conditions. More than 130,000 ESTs were generated and their expression across the different libraries can be visualized at http://www.biologie.ens.fr/diatomics/EST3. Full information about all methods used for the analyses reported here is available in Supplementary Information.

References

1. **Falkowski, P. G., Barber, R. T. & Smetacek, V. Biogeochemical controls and feedbacks on ocean primary production. *Science* 281, 200-206 (1998).**
2. **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).**
3. **Armbrust, E. V. et al. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* 306, 79-86 (2004).**
4. **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).**
5. **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).**
6. **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).**
7. **Kooistra, W. H. C. F., Gersonde, R., Medlin, L. K. & Mann, D. G. in *Evolution of Primary Producers in the Sea* (eds. Falkowski, P. G. & Knoll, A. H.) 207-249 (Academic Press, Inc., 2007).**
8. **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).**
9. **Tyler, B. M. et al. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313, 1261-6 (2006).**
10. **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).**
11. **Kumar, S. & Hedges, S. B. A molecular timescale for vertebrate evolution. *Nature* 392, 917-920 (1998).**
12. **Dujon, B. Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet* 22, 375-87 (2006).**
13. **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).**
14. **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).**

15. 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).
16. 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).
17. Matsuzaki, M. et al. Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. *Nature* 428, 653-7 (2004).
18. 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).
19. Keeling, P. J. & Palmer, J. D. Horizontal gene transfer in eukaryotic evolution. *Nature Rev. Genet.* 9, 605-618 (2008).
20. Kroth, P. G. et al. A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from whole genome analysis and comparative genomic analyses with Thalassiosira pseudonana and other photoautotrophs. *PLoS One* 3, e1426 (2008).
21. 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).
22. Carpenter, E. J. & Janson, S. Intracellular cyanobacterial symbionts in the marine diatom *Climacodium frauenfeldianum* (Bacillariophyceae). *J. Phycol.* 36, 540-544 (2000).
23. Schmid, A.-M. M. Endobacteria in the diatom *Pinnularia* (Bacillariophyceae). I. Scattered ct-nucleoids explained: DAPI-DNA complexes stem from exoplastidial bacteria boring into the chloroplasts. *J. Phycol.* 39, 122-138 (2003).
24. 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).
25. Takahashi, F. et al. AUREOCHROME, a photoreceptor required for photomorphogenesis in stramenopiles. *Proc Natl Acad Sci U S A.* 104, 19625-19630 (2007).
26. 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).
27. Swartz, T. E. et al. Blue-light-activated histidine kinases: two-component sensors in bacteria. *Science* 317, 1090-1093 (2007).
28. Merchant, S. S. et al. The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. *Science* 318, 245-250 (2007).
29. Sumper, M. & Brunner, E. Silica biomineralization in diatoms: the model organism *Thalassiosira pseudonana*. *Chembiochem.* 9, 1187-1194 (2006).
30. 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).

Supplementary information accompanies the paper on **www.nature.com/nature.**

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.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.B. (cbowler@biologie.ens.fr). Genome assemblies together with predicted gene models and annotations have been deposited at DDBJ/EMBL/GenBank under the project accessions ABQD00000000 and AAFD00000000, respectively, and the versions described in this paper represent the first version, ABQD01000000, for *P. tricornutum*, and the second version, AAFD02000000, for *T. pseudonana*. *P. tricornutum* EST expression profiles can be explored at http://www.biologie.ens.fr/diatomics/EST3, which also provides links to gene models on the JGI genome browser. ESTs have been deposited at NCBI dbEST with GenBank accession numbers CD374840-CD384835, BI306757-BI307753, CT868744-CT950687, and CU695349-CU740080.

Individual contributions C.B. coordinated *Phaeodactylum* genome annotation and manuscript preparation. E.V.A. coordinated *Thalassiosira* genome annotation. D.S.R. and I.V.G. coordinated diatom genome sequencing and analysis at JGI. J.W. coordinated EST sequencing at Genoscope. A.E.A., J.H.B., J.G., K.J., A.K., U.M., C.M., F.M., R.P.O., E.R., A.S. and K.V. made equivalent and substantial contributions to the data presented. B.B., A.G.,

M.H., M.K., T.M., K.V. and F.V. also made significant contributions. E.V.A., B.R.G., Y.V.d.P. and I.V.G assisted in data interpretation and manuscript preparation. Other authors contributed as members of the *Phaeodactylum* genome sequencing consortium.

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 [10]. 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) [11]. The centric:pennate divergence, on the other hand, has been dated to at least 90 Ma [7]. 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 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 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.

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 LovHK) are indicated above (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.

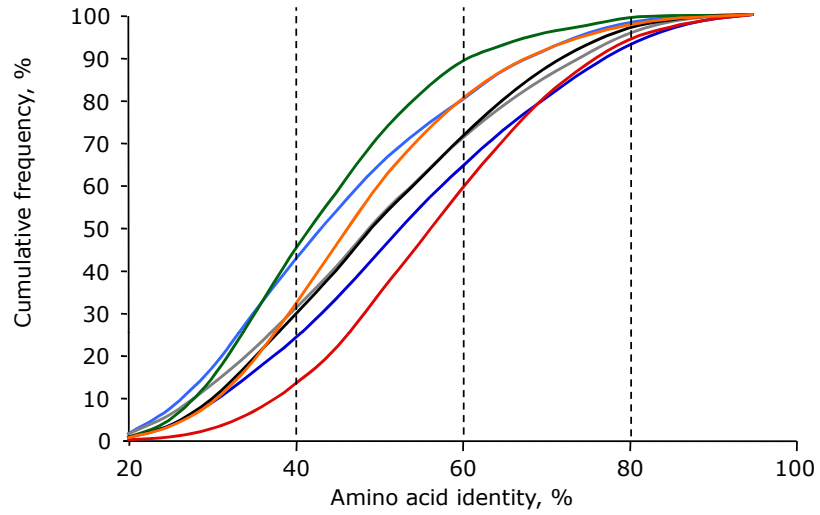
	P. tricornutum	*T. pseudonana*
Genome size	27.4 Mb	32.4 Mb
Predicted genes	10,402	11,776
Core genes*	3,523	4,332
Diatom-specific genes*	1,328	1,407
Unique genes*	4,366	3,912
Introns	8,169	17,880
Introns/gene	0.79	1.52
LTR retrotransposon content	5.8%	1.1%

* 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.

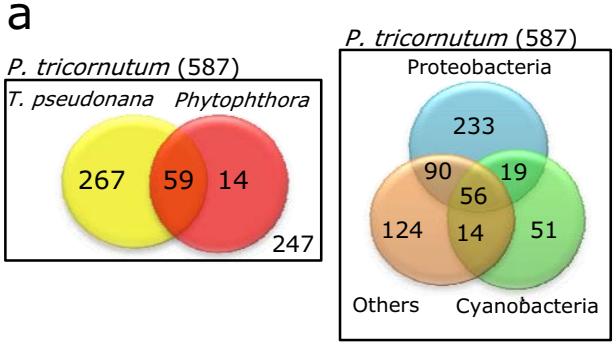
a

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

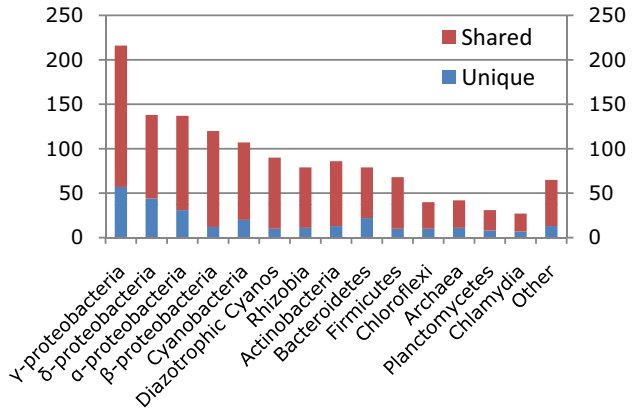
b



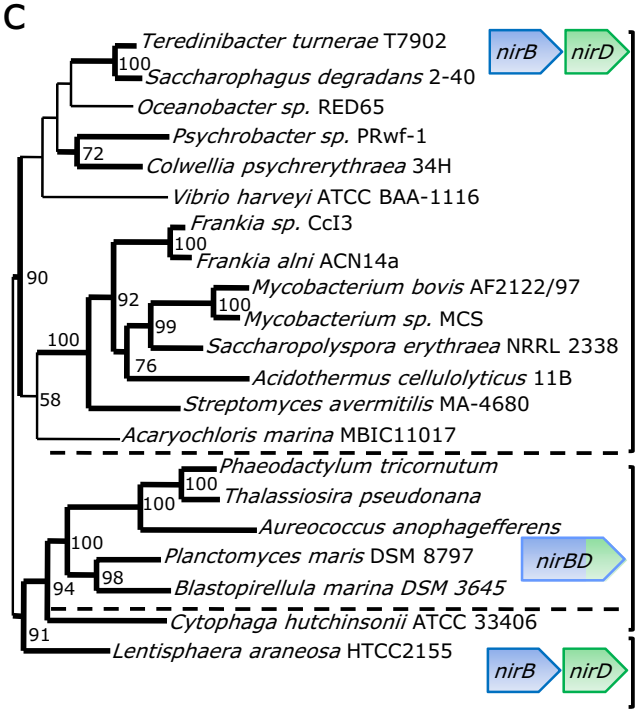
a

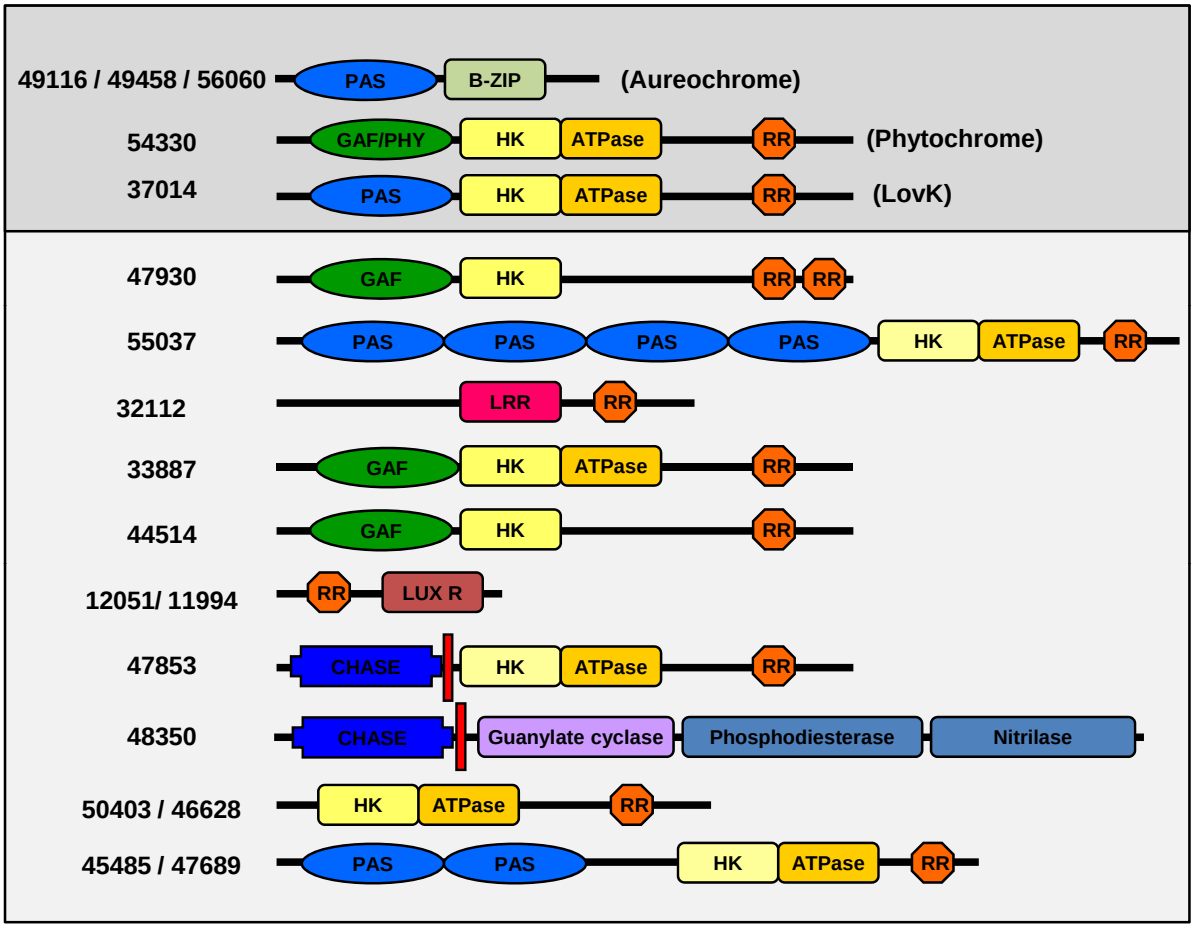

P. tricornutum (587)
T. pseudonana
Phytophthora
267
59
14
247
P. tricornutum (587)
Proteobacteria
233
90
19
56
124
14
51
Others
Cyanobacteria

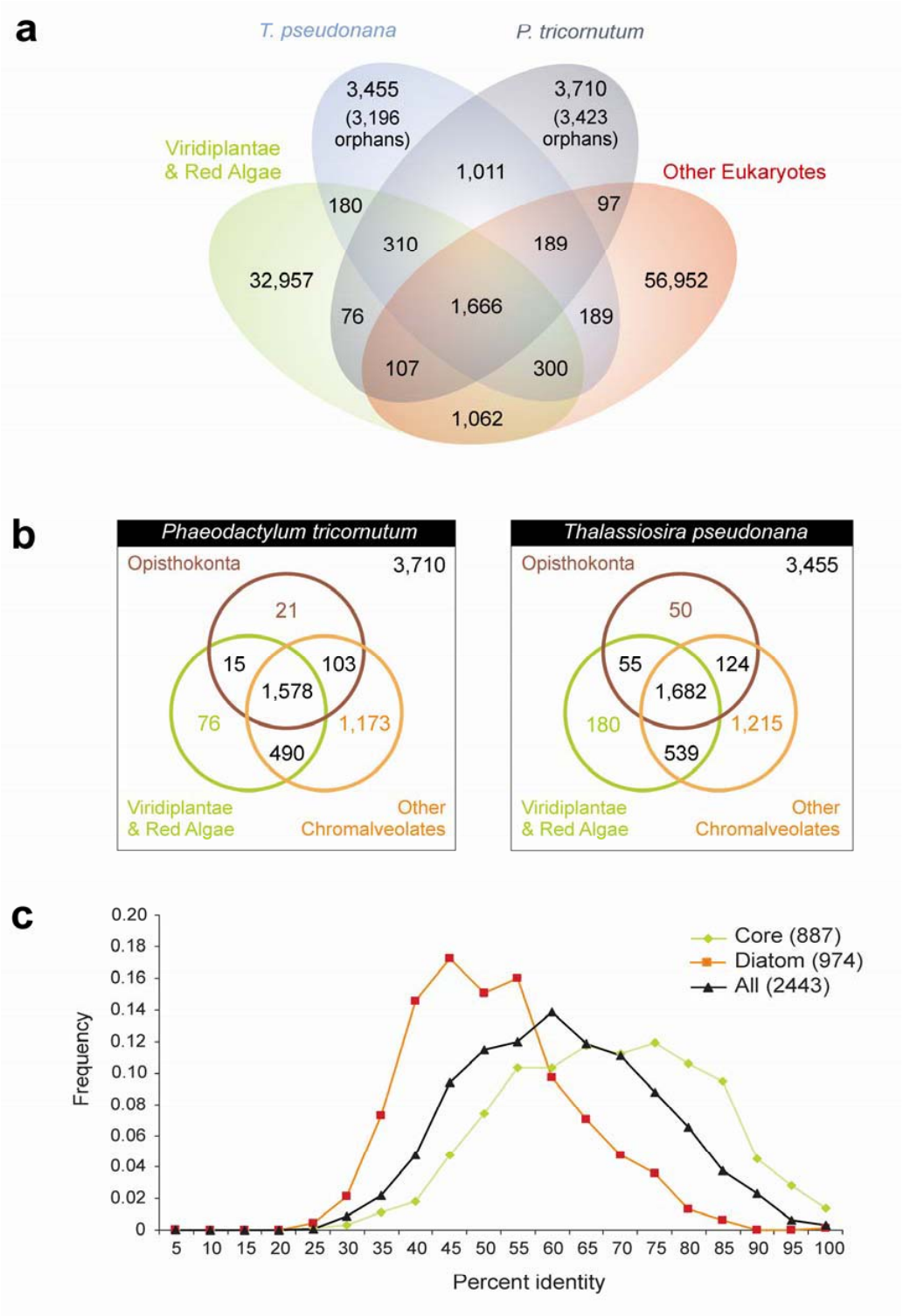
b


250
200
150
100
50
0
Shared
Unique
γ-proteobacteria
δ-proteobacteria
α-proteobacteria
β-proteobacteria
Cyanobacteria
Diazotrophic Cyanos
Rhizobia
Actinobacteria
Bacteroidetes
Firmicutes
Chloroflexi
Archaea
Planctomycetes
Chlamydia
Other

c


Teredinibacter turnerae T7902
nirB
nirD
100
Saccharophagus degradans 2-40
Oceanobacter sp. RED65
Psychrobacter sp. PRwf-1
72
Colwellia psychrerythraea 34H
Vibrio harveyi ATCC BAA-1116
Frankia sp. CcI3
100
Frankia alni ACN14a
90
92
Mycobacterium bovis AF2122/97
100
Mycobacterium sp. MCS
99
100
Saccharopolyspora erythraea NRRL 2338
76
Acidothermus cellulolyticus 11B
58
Streptomyces avermitilis MA-4680
Acaryochloris marina MBIC11017
Phaeodactylum tricornutum
100
Thalassiosira pseudonana
100
Aureococcus anophagefferens
100
Planctomyces maris DSM 8797
nirBD
98
94
Blastopirellula marina DSM 3645
Cytophaga hutchinsonii ATCC 33406
91
Lentisphaera araneosa HTCC2155
nirB
nirD


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


a
T. pseudonana
P. tricornutum
3,455
(3,196
orphans)
3,710
(3,423
orphans)
Viridiplantae
& Red Algae
1,011
Other Eukaryotes
180
97
310
189
32,957
56,952
76
1,666
189
107
300
1,062
b
Phaeodactylum tricornutum
Opisthokonta
3,710
21
15
103
1,578
76
1,173
490
Viridiplantae
& Red Algae
Other
Chromalveolates
Thalassiosira pseudonana
Opisthokonta
3,455
50
55
124
1,682
180
1,215
539
Viridiplantae
& Red Algae
Other
Chromalveolates
c
Core (887)
Diatom (974)
All (2443)
Frequency
0.20
0.18
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0
5 10 15 20 25 30 35 40 45 50 55 60 65 70 75 80 85 90 95 100
Percent identity