2008

The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes

Chris Bowler

Andrew E. Allen

*See next page for additional authors*

Creative Commons License

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.

Follow this and additional works at: https://digitalcommons.uri.edu/gsofacpubs

Citation/Publisher Attribution


Available at: https://doi.org/10.1038/nature07410

This Article is brought to you for free and open access by the Graduate School of Oceanography at DigitalCommons@URI. It has been accepted for inclusion in Graduate School of Oceanography Faculty Publications by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons@etal.uri.edu.
The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes


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 Earth1–4. The genome sequence of the marine centric diatom *Thalassiosira pseudonana* was recently reported, revealing a wealth of information about diatom biology5–8. 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.

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 180 million years (Myr) old and that from pennates is 90 Myr old27. Although being 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, by contrast with the 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\(^8\). Furthermore, members of the raphid pennate clade can glide actively along surfaces.

The completed \textit{P. tricornutum} genome is approximately 27.4 megabases (Mb) in size, which is slightly smaller than \textit{T. pseudonana} (32.4 Mb), and \textit{P. tricornutum} is predicted to contain fewer genes (10,402 as opposed to 11,776; Table 1, 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 under 16 different conditions. In total, 86\% of gene predictions had EST support (Supplementary Table 1).

\textit{P. tricornutum} shares 57\% of its genes with \textit{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 percentage amino acid identity of 4,267 orthologous gene pairs (Table 2, Fig. 1). We found an average identity of 54.9\% between diatom orthologues, in comparison with approximately 43\% (Table 2, Fig. 1). The diatom–oomycete pair displays the lowest amino acid identity (43.3\%), in agreement with their proposed ancient separation, around 700 Myr ago\(^9\). The divergence between the pennate and centric diatom is similar to the fish–mammal divergence, which probably occurred in the Proterozoic era (550 Myr ago)\(^10\). The centric–pennate divergence, on the other hand, has been dated to at least 90 Myr ago. In the figure, we represent the cumulative frequencies of amino acid identity across each set of potential orthologous pairs shown in Table 2.

Large-scale within-genome duplication events do not appear to have played a major role in driving the generation of diatom diversity (Supplementary Information), by contrast with what has been found in yeasts and metazoans\(^11\). The observed high levels of diatom species diversity must therefore have been generated by other mechanisms. Whereas intron gain may be one factor in centric diatoms, the dramatic expansion of diatom-specific copia-retrotransposable elements may have contributed to the \textit{P. tricornutum} genome (Table 1, Supplementary Figs 2, 4). These elements also appear to have expanded in other pennate diatoms (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\(^12\). 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\(^1\). It is generally thought that the diatom mitochondrion originated in the host, and the mitochondrial gene complement is almost identical to that of haptophytes and cryptophytes, which are other algal phyla that may have originated from the same secondary endosymbiotic event. We used a phylogenetic approach to search for genes of red algal origin in the two diatoms and the two sequenced oomycetes, \textit{Phytophthora ramorum} and \textit{Phytophthora sojae}, using \textit{Cyamidioschyzon merolae} as reference red algal genome\(^13\). We classified 171 genes as being of red algal origin, on

<table>
<thead>
<tr>
<th>Pairwise comparison</th>
<th>Mean identity (%)</th>
<th>Number of compared pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{Phaeodactylum tricornutum}/Thalassiosira pseudonana</td>
<td>54.9</td>
<td>4,267</td>
</tr>
<tr>
<td>\textit{Phaeodactylum tricornutum}/\textit{Phytophthora sojae}</td>
<td>43.3</td>
<td>2,952</td>
</tr>
<tr>
<td>\textit{Saccharomycyes cerevisiae}/\textit{Debaryomyces hansenii}</td>
<td>50.1</td>
<td>2,694</td>
</tr>
<tr>
<td>\textit{Saccharomycyes cerevisiae}/\textit{Kluyveromyces lactis}</td>
<td>54.8</td>
<td>4,246</td>
</tr>
<tr>
<td>\textit{Saccharomycyes cerevisiae}/\textit{Candida glabrata}</td>
<td>58.2</td>
<td>4,884</td>
</tr>
<tr>
<td>\textit{Homo sapiens}/\textit{Ciona intestinalis}</td>
<td>52.6</td>
<td>5,208</td>
</tr>
<tr>
<td>\textit{Homo sapiens}/\textit{Takifugu rubripes}</td>
<td>61.4</td>
<td>10,225</td>
</tr>
</tbody>
</table>

Summary of numbers of orthologous pairs (reciprocal best hits with an expected cut-off value of 10\(^{-10}\)) for each organism comparison and their mean percentage identities.
the basis of 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. The results of this survey support there being 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 confirmed 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. Of the 587 identified sequences, 42% are only found in P. tricornutum whereas 56% are present in both diatoms (Fig. 2a), attesting to their ancient origin. Only 73 sequences are shared between P. tricornutum and Phytophthora spp. (Fig. 2a, Supplementary Table 3), 59 of which are also present in T. pseudonana, 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, for example for organic carbon and nitrogen utilization (xylanases and glucanases, prasmine, carbon-nitrogen hydrodase, amidohydrolase), functioning of the diatom urea cycle (carbamoyl transferase, carbonic anhydrase, ornithine cycloclase) and polyamine metabolism relevant to diatom cell wall silicification (S-adenosylmethionine (SAM)-dependent decarboxylases and methyltransferases). Others are likely to encode novel cell wall components, and to provide unconventional 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 close associations with diatoms, which may explain the unprecedented levels of horizontal gene transfer events that appear to have occurred. In P. tricornutum, bacterial genes are distributed throughout the genome, although several clusters, as well as regions devoid of bacterial genes, can be observed (Supplementary Fig. 3). Some of these genes in diatoms share bacterial-specific gene fusions that support phylogenetic associations, such as assimilatory nitrite reductase B and D subunits; these are apparently of planctomycete origin (Fig. 2c).

Bacterial histidine-kinase-based phosphorelay two-component systems, which are 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 photo receptors.

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 panel), and which bacterial classes are most related phylogenetically (right panel). 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 (log likelihood ratio, −22,358.321320) as inferred from the amino acid sequences of the large subunit of NAD(P)H assimilatory nitrite reductase (NirB). The choice of protein evolution model was WAG with gamma-distributed rates (x = 0.80), as suggested by a ProtTest analysis of the alignment (see Supplementary Information for methods). Numbers above selected branches indicate maximum likelihood bootstrap support (100 replicates). Gene fusions and distinct open reading frames are indicated adjacent to the appropriate clades. In most cases, the large (NirB) and small (NirD) subunits of NAD(P)H assimilatory nitrite reductase are encoded by distinct open reading frames, but in diatoms and planctomycetes the nirD and nirB open reading frames 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.
Shared and unique gene families. a P. tricornutum contains orthologues of LovHK and T. pseudonana Guanylate cyclase. Nitrilase P. tricornutum Other eukaryotes role of long-chain polyamines in silica nanofabrication in polyamine metabolism are over-represented. The expansion of silicon formation in P. tricornutum contains only one silafillin-like protein, and no homologues of silacidin. Frustulin genes, encoding proteins that form organic constituents of the biosilica cell wall but as previously noted in T. pseudonana". Another domain combination present in both diatoms resembles aureochrome blue-light photoreceptors and P. tricornutum contains orthologues of LovHK and other light-dependent histidine kinases reported in bacteria. 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, 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. 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 proteins are evolving faster than other genes in diatom genomes (Fig. 4c), providing a further explanation for the rapid diatom divergence rates.

Of the gene families found in the diatoms, some contain higher numbers of genes in comparison with other eukaryotes (Supplementary Table 4, Supplementary Fig. 7); for example, genes involved in polyamine metabolism are over-represented. The expansion of polyamine-related components is of interest in consideration of the role of long-chain polyamines in silica nanofabrication. Of the eight predicted spermine/spermidine synthase-like genes in P. tricornutum, three encode potentially bifunctional enzymes bearing both an aminopropyltransferase domain and a SAM decarboxylase domain. Interestingly, the only other organisms containing such bifunctional proteins are T. pseudonana (four copies) and the bacteria Bdellovibrio bacteriovorus and Delftia acidovorans. Silafillins and silacidins are proteins/peptides believed to be involved in diatom silica formation. P. tricornutum contains only one silafillin-like protein, and no homologues of silacidin. Frustulin genes, encoding proteins that form organic constituents of the biosilica cell wall but as previously noted in T. pseudonana. Another domain combination present in both diatoms resembles aureochrome blue-light photoreceptors, and P. tricornutum contains orthologues of LovHK and other light-dependent histidine kinases reported in bacteria. 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, 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. 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 proteins are evolving faster than other genes in diatom genomes (Fig. 4c), providing a further explanation for the rapid diatom divergence rates.

Of the gene families found in the diatoms, some contain higher numbers of genes in comparison with other eukaryotes (Supplementary Table 4, Supplementary Fig. 7); for example, genes involved in polyamine metabolism are over-represented. The expansion of polyamine-related components is of interest in consideration of the role of long-chain polyamines in silica nanofabrication. Of the eight predicted spermine/spermidine synthase-like genes in P. tricornutum, three encode potentially bifunctional enzymes bearing both an aminopropyltransferase domain and a SAM decarboxylase domain. Interestingly, the only other organisms containing such
METHODS SUMMARY

High-molecular-weight DNA was extracted from axenic cultures of Phaeodactylum tricornutum accession P1.8.6 (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 Joint Genome Institute (IGI) 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 was finished in a similar way. Both diatom genomes were annotated using the IGI annotation pipeline, which combines several gene prediction, annotation and analysis tools. Complementary DNA libraries were constructed from messenger RNA extracted from P. tricornutum cultures grown under 16 different conditions. More than 130,000 ESTs were generated. Full information about all methods used for the analyses reported here is available in Supplementary Information.

Received 20 June; accepted 8 September 2008. Published online 15 October 2008.

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements 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 Diatoms 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). We are grateful to M. Muffato and H.-R. Crollius for the analysis reported in Supplementary Fig. 3a.


Author Information Assemblies and annotations of the *P. tricornutum* and *T. pseudonana* genomes are available through the JGI Genome Portal at http://www.jgi.doe.gov/phaeodactylum and http://www.jgi.doe.gov/thalassiosira. Genome assemblies together with predicted gene models and annotations have been deposited at DDBJ, EMBL and GenBank under the project accessions ABQD000000000 and AAFD000000000, respectively. The versions described in this paper are the first version, ABQD010000000, for *P. tricornutum*, which includes complete chromosomes 3 and 11 (CP001142 and CP001141), and the second version, AAFD020000000, for *T. pseudonana*, also including complete chromosomes 7 and 18 (CP001160 and CP001159). *P. tricornutum* EST expression profiles can be found 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, B306757-B307753, CT868744-CT950687 and CU695384-CU740080. Reprints and permissions information is available at www.nature.com/reprints. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. Correspondence and requests for materials should be addressed to C.B. (cbowler@biologie.ens.fr).