

# Genome analysis of the smallest free-living eukaryote *Ostreococcus tauri* unveils many unique features

Evelyne Derelle<sup>a,b</sup>, Conchita Ferraz<sup>b,c</sup>, Stephane Rombauts<sup>b,d</sup>, Pierre Rouzé<sup>b,e</sup>, Alexandra Z. Worden<sup>f</sup>, Steven Robbens<sup>d</sup>, Frédéric Partensky<sup>g</sup>, Sven Degroeve<sup>d,h</sup>, Sophie Echeyni<sup>c</sup>, Richard Cooke<sup>i</sup>, Yvan Saeys<sup>d</sup>, Jan Wuyts<sup>d</sup>, Kamel Jabbari<sup>j</sup>, Chris Bowler<sup>k</sup>, Olivier Panaud<sup>l</sup>, Benoît Piégu<sup>i</sup>, Steven G. Ball<sup>k</sup>, Jean-Philippe Ral<sup>k</sup>, François-Yves Bouget<sup>a</sup>, Gwenael Piganeau<sup>a</sup>, Bernard De Baets<sup>h</sup>, André Picard<sup>a,l</sup>, Michel Delseny<sup>l</sup>, Jacques Demaille<sup>c</sup>, Yves Van de Peer<sup>d,m</sup>, and Hervé Moreau<sup>a,m</sup>

<sup>a</sup>Observatoire Océanologique, Laboratoire Arago, Unité Mixte de Recherche 7628, Centre National de la Recherche Scientifique–Université Pierre et Marie Curie–Paris 6, BP44, 66651 Banyuls sur Mer Cedex, France; <sup>c</sup>Institut de Génétique Humaine, Unité Propre de Recherche 1142, Centre National de la Recherche Scientifique, 141 Rue de Cardonille, 34396 Montpellier Cedex 5, France; <sup>d</sup>Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology and <sup>e</sup>Laboratoire Associé de l'Institut National de la Recherche Agronomique (France), Ghent University, Technologiepark 927, 9052 Ghent, Belgium; <sup>f</sup>Rosenstiel School of Marine and Atmospheric Science, University of Miami, 4600 Rickenbacker Causeway, Miami, FL 33149; <sup>g</sup>Station Biologique, Unité Mixte de Recherche 7144, Centre National de la Recherche Scientifique–Université Pierre et Marie Curie–Paris 6, BP74, 29682 Roscoff Cedex, France; <sup>h</sup>Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Coupure Links 653, 9000 Ghent, Belgium; <sup>i</sup>Génome et Développement des Plantes, Unité Mixte de Recherche 5096, Centre National de la Recherche Scientifique–Université de Perpignan, 52, Avenue de Villeneuve, 66860 Perpignan, France; <sup>j</sup>Département de Biologie, Formation de Recherche en Evolution 2910, Centre National de la Recherche Scientifique–Ecole Normale Supérieure, 46 Rue d'Ulm, 75230 Paris Cedex 05, France; and <sup>k</sup>Laboratoire de Chimie Biologique, Unité Mixte de Recherche 8765, Centre National de la Recherche Scientifique–Université Sciences et Technologies de Lille, 59655 Villeneuve d'Ascq, France

Communicated by Marc C. E. Van Montagu, Ghent University, Ghent, Belgium, June 8, 2006 (received for review January 4, 2006)

The green lineage is reportedly 1,500 million years old, evolving shortly after the endosymbiosis event that gave rise to early photosynthetic eukaryotes. In this study, we unveil the complete genome sequence of an ancient member of this lineage, the unicellular green alga *Ostreococcus tauri* (Prasinophyceae). This cosmopolitan marine primary producer is the world's smallest free-living eukaryote known to date. Features likely reflecting optimization of environmentally relevant pathways, including resource acquisition, unusual photosynthesis apparatus, and genes potentially involved in C<sub>4</sub> photosynthesis, were observed, as was downsizing of many gene families. Overall, the 12.56-Mb nuclear genome has an extremely high gene density, in part because of extensive reduction of intergenic regions and other forms of compaction such as gene fusion. However, the genome is structurally complex. It exhibits previously unobserved levels of heterogeneity for a eukaryote. Two chromosomes differ structurally from the other eighteen. Both have a significantly biased G+C content, and, remarkably, they contain the majority of transposable elements. Many chromosome 2 genes also have unique codon usage and splicing, but phylogenetic analysis and composition do not support alien gene origin. In contrast, most chromosome 19 genes show no similarity to green lineage genes and a large number of them are specialized in cell surface processes. Taken together, the complete genome sequence, unusual features, and downsized gene families, make *O. tauri* an ideal model system for research on eukaryotic genome evolution, including chromosome specialization and green lineage ancestry.

genome heterogeneity | genome sequence | green alga | Prasinophyceae | gene prediction

The smallest free-living eukaryote known so far is *Ostreococcus tauri* (1). This tiny unicellular green alga belongs to the Prasinophyceae, one of the most ancient groups (2) within the lineage giving rise to the green plants currently dominating terrestrial photosynthesis (the green lineage) (3, 4). Consequently, since its discovery, there has been great interest in *O. tauri*, which, because of its apparent overall simplicity, a naked, nonflagellated cell possessing a single mitochondrion and chloroplast, in addition to its small size and ease in culturing, renders it an excellent model organism (5). Furthermore, it has been hypothesized, based on its small cellular and genome sizes (2, 6), that it may reveal the “bare limits” of life as a free-living photosynthetic eukaryote, presumably

having disposed of redundancies and presenting a simple organization and very little noncoding sequence.

Since its identification in 1994, *Ostreococcus* has been recognized as a common member of the natural marine phytoplankton assemblage. It is cosmopolitan in distribution, having been found from coastal to oligotrophic waters, including the English Channel, the Mediterranean and Sargasso Seas, and the North Atlantic, Indian, and Pacific Oceans (7–12). Eukaryotes within the picosize fraction (<2- to 3- $\mu$ m diameter) have been shown to contribute significantly to marine primary production (9, 13). *Ostreococcus* itself is notable for its rapid growth rates and potential grazer susceptibility (9, 14). Furthermore, dramatic blooms of this organism have been recorded off the coasts of Long Island (15) and California (11). At the same time, attention has focused on the tremendous diversity of picoeukaryotes (16, 17), which holds true for *Ostreococcus* as well. Recently, *Ostreococcus* strains isolated from surface waters were shown to represent genetically and physiologically distinct ecotypes, with light-regulated growth optima different from those isolated from the deep chlorophyll maximum (18). These findings are similar to the niche adaptations documented in different ecotypes of the abundant marine cyanobacteria *Prochlorococcus* (19, 20).

Overall, marine picophytoplankton play a significant role in primary productivity and food webs, especially in oligotrophic environments where they account for up to 90% of the autotrophic biomass (9, 13, 21, 22). Several recent studies have undertaken a genome sequencing approach to understand the ocean ecology of phytoplankton. To date, these studies have focused on the bacterial component of the plankton, particularly on the picocyanobacteria *Prochlorococcus* (20) and *Synechococcus* (23), for which 9 complete

Conflict of interest statement: No conflicts declared.

Abbreviation: TE, transposable element.

Data deposition: The genome data have been submitted to the European Molecular Biology Laboratory, www.embl.org [accession nos. CR954201 (Chrom 1), CR954202 (Chrom 2), CR954203 (Chrom 3), CR954204 (Chrom 4), CR954205 (Chrom 5), CR954206 (Chrom 6), CR954207 (Chrom 7), CR954208 (Chrom 8), CR954209 (Chrom 9), CR954210 (Chrom 10), CR954211 (Chrom 11), CR954212 (Chrom 12), CR954213 (Chrom 13), CR954214 (Chrom 14), CR954215 (Chrom 15), CR954216 (Chrom 16), CR954217 (Chrom 17), CR954218 (Chrom 18), CR954219 (Chrom 19), and CR954220 (Chrom 20)].

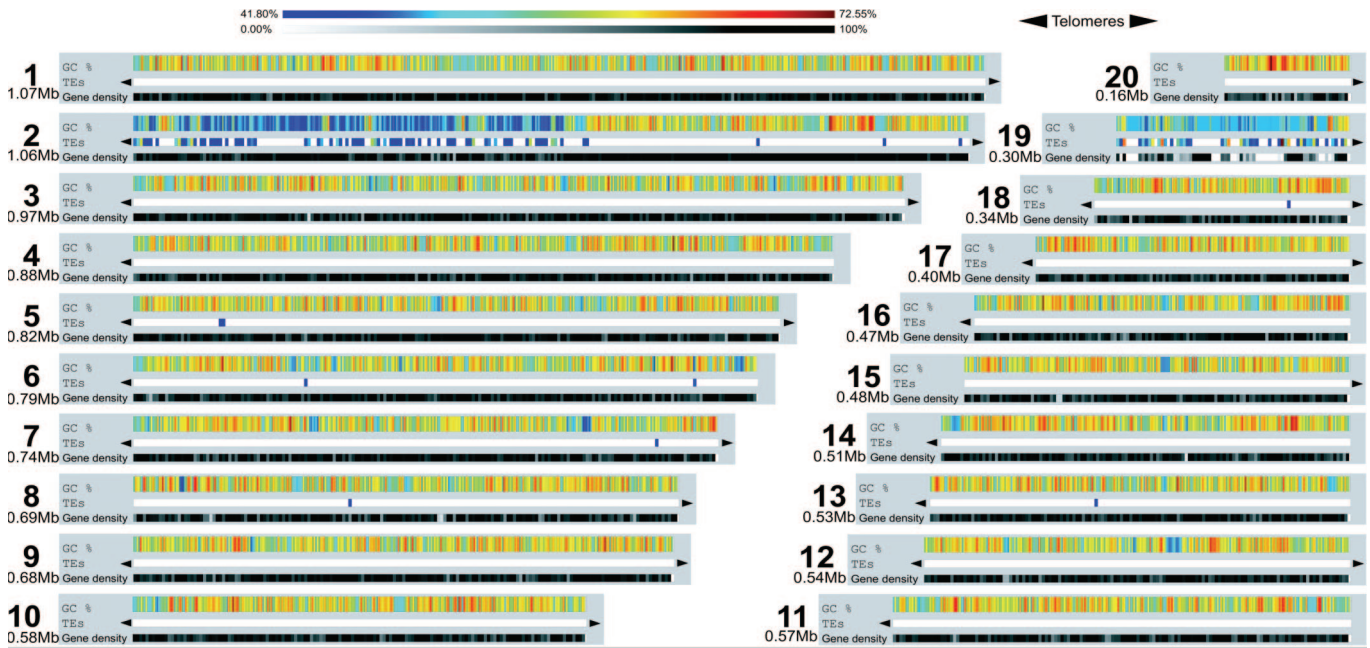
See Commentary on page 11433.

<sup>b</sup>E.D., C.F., S.R., and P.R. contributed equally to this work.

<sup>l</sup>Deceased November 21, 2004.

<sup>m</sup>To whom correspondence may be addressed. E-mail: yves.vandeppeer@psb.ugent.be or h.moreau@obs-banyuls.fr.

© 2006 by The National Academy of Sciences of the USA



**Fig. 1.** General characteristics of the 20 *O. tauri* chromosomes. TEs, transposon frequency. Size is indicated to the left of each chromosome (Mb). Colored bars indicate the percentage G+C content (upper bar) and of transposons (lower bar).

genome sequences are already publicly available and >13 others on the way. Much less is known about eukaryotic phytoplankton, because only one, the diatom *Thalassiosira pseudonana*, has a complete genome sequenced (24). Picoeukaryotes are especially interesting in the context of marine primary production, given the combination of their broad environmental distribution and the fact that their surface area to volume ratio, a critical factor in resource acquisition and success in oligotrophic environments (25), is similar to that of prokaryotic counterparts generally considered superior in uptake and transport of nutrients.

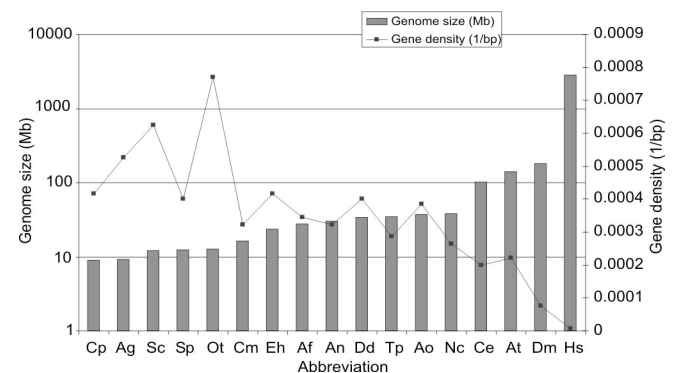
In this article, we describe the complete genome sequence of *O. tauri* OTH95, a strain isolated in the Thau lagoon (France) in which this species makes recurrent, quasimonospecific blooms in summer (1). This genome is particularly significant in that it represents a complete genome sequence of a member of the Prasinophyceae, which diverged at the base of the green lineage (2). It is also the complete genome sequence of a picoeukaryote thought to be of ecological importance to primary production. Analysis of the *O. tauri* genome and comparison with other genomes available to date, including algal, plant, and fungal genomes, allowed delineation of both specific gene features and identification of unique aspects of this genome.

## Results and Discussion

**Global Genome Structure.** Whole genome shotgun sequencing and an oriented walking strategy were used to sequence the genome of *O. tauri* strain OTH95 (Tables 2 and 3, which are published as supporting information on the PNAS web site). A genome size of 12.56 Mb distributed in 20 superscaffolds corresponding to 20 chromosomes was determined by means of sequence assembly (Fig. 1; and Figs. 4 and 5, which are published as supporting information on the PNAS web site), fully consistent with pulsed-field gel electrophoresis results indicating a total size of 12.5 to 13 Mb (Fig. 4 and *Supporting Text*, which are published as supporting information on the PNAS web site). This genome size is similar to that of the yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, despite their larger cell size, but smaller than any other oxyphototrophic eukaryote known so far, including the red alga *Cyanidioschyzon merolae* (26) (Fig. 2 and Table 1). The G+C

content of *O. tauri* is more akin to that of *C. merolae* than to that of plants, fungi, or even *T. pseudonana* (Table 1). As shown in Fig. 2 and Table 1, 8,166 protein-coding genes were predicted in the nuclear genome, making *O. tauri* the most gene dense free-living eukaryote known to date. Only the chromosomes of the nucleomorphs within chlorachniophyte and cryptophyte algae are more gene-dense bodies (27), which are internally contained and not capable of independent propagation. We found that 6,265 genes are supported by homology with known genes in public databases (e-value <10<sup>-5</sup>), of which the majority (46%) were most similar to plant orthologs (Fig. 3). Very few repeated sequences have been found in this genome, except for a long internal duplication of 146,028 kb on chromosome 19. Because the duplicated sequence is >99% identical, it is probably of recent origin.

**Genome Heterogeneity.** In view of what is currently known about eukaryotic nuclear genomes, one of the most striking features of the



**Fig. 2.** Genome size and gene density for various eukaryote genomes. Cp, *Cryptosporidium parvum*; Ag, *Ashbya gossypii*; Sp, *Schizosaccharomyces pombe*; Sc, *Saccharomyces cerevisiae*; Ot, *Ostreococcus tauri*; Cm, *Cyanidioschyzon merolae*; Eh, *Entamoeba histolytica*; Af, *Aspergillus fumigatus*; An, *Aspergillus niger*; Dd, *Dictyostelium discoideum*; Tp, *Thalassiosira pseudonana*; Ao, *Aspergillus oryzae*; Nc, *Neurospora crassa*; Ce, *Caenorhabditis elegans*; At, *Arabidopsis thaliana*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*.

Table 1. General features of the *O. tauri* genome

Feature	<i>O. tauri</i>	<i>T. pseudonana</i>	<i>C. merolae</i>	<i>Arabidopsis thaliana</i>	<i>Ashbya gossypii</i>	<i>S. cerevisiae</i>	<i>S. pombe</i>	<i>Cryptosporidium parvum</i>
Size, Mbp	12.56	34.50	16.52	140.12	9.20	12.07	12.46	9.10
No. of chromosomes	20	24	20	5	7	16	3	8
G+C content, %	58.0 (59.0*)	47.0	55.0	36.0	52.0	38.3	36.0	30.0
Gene number	8,166	11,242	5,331	26,207	4,718	6,563	4,824	3,807
Gene density, kb per gene	1.3	3.5	3.1	4.5	1.9	1.6	2.5	2.4
Mean gene size, bp per gene <sup>†</sup>	1,257	992	1,552	2,232	N.A.	N.A.	1,426	1,795
Mean inter-ORF distance	197	N.A.	1,543	2,213	341	N.A.	952	566
Genes with introns, %	39	N.A.	0.5	79	5	5	43	5
Mean length of introns, bp	103 (187*)	N.A.	248	164	N.A.	N.A.	81	N.A.
Coding sequences, %	81.6	N.A.	44.9	33.0	79.5	N.A.	57.5	75.3
No. of ribosomal RNA units	4	N.A.	3	700–800	50	100–150	200–400	5

Data for the yeast *S. cerevisiae* compiled from refs. 24, 26, 47–49 and from *Saccharomyces* Genome Database at [www.yeastgenome.org](http://www.yeastgenome.org); N.A., not available.

\*Data that exclude chromosomes 2 and 19.

<sup>†</sup>Data that exclude introns.

*O. tauri* genome is its heterogeneity, a feature which is not only unusual but also perplexing from an evolutionary perspective. Two chromosomes (2 and 19) are different from the other 18, in terms of organization for chromosome 2 and function for chromosome 19 (Fig. 1; and Fig. 6, which is published as supporting information on the PNAS web site). Both of these chromosomes have lower G+C content than the 59% G+C of the other 18 chromosomes (Fig. 1). Chromosome 2 is composed primarily of two blocks, one with a G+C content similar to that of the other chromosomes and the other with a markedly lower G+C content (52%). The average G+C content of the entire chromosome 2 amounts to 55%. Likewise, the G+C content of chromosome 19 (54%) is similar to the atypical region of chromosome 2. Taken together, these two aberrant chromosomes contain 77% of the 417 transposable elements (TEs), or relics thereof, which are identified in the genome (57% in chromosome 2 and 20% in chromosome 19) (Fig. 1 and Table 4, which is published as supporting information on the PNAS web site). Other chromosomes therefore contain very few or no TEs. TEs have a G+C content similar to the rest of the genome and cannot explain the global lower G+C content observed in these two chromosomal regions. Moreover, almost all of the known TE types can be found in the *O. tauri* genome: fifteen class I TE families [i.e., 3 *TY1/Copia*-like LTR-retrotransposons and 12 terminal-repeat

retrotransposons in miniature (TRIMs)], nine transposon families, [4 Mariner-like elements, 2 P instability factors (PIFs), 1 homology and transposition (hAT), 1 foldback, and 1 unclassified (28)], and three miniature inverted repeat transposable element (MITE) families were identified. Only long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and helitrons were not detected. In the case of *O. tauri*, the distribution bias could have two origins: either the species originated from an allopolyploidization event between two donor parents with a genome contrasting for their TE content or there is a strong insertion bias for the TEs on both chromosomes 2 and 19. For most of the TE families, several partial copies or relics can be found throughout the 20 chromosomes (Table 4), indicating their ancient origin in the genome, therefore not supporting the first hypothesis. Nevertheless, further analyses are needed to conclude on this matter.

Chromosome 2 has additional unique features aside from differences in G+C content and the occurrence of many transposons. In particular, codon usage for genes in the low G+C region of this chromosome is different from that of all other chromosomes (Table 5, which is published as supporting information on the PNAS web site). Many of the genes in this low G+C region also contain multiple small introns with specific features (Fig. 7a and b, which is published as supporting information on the PNAS web site).

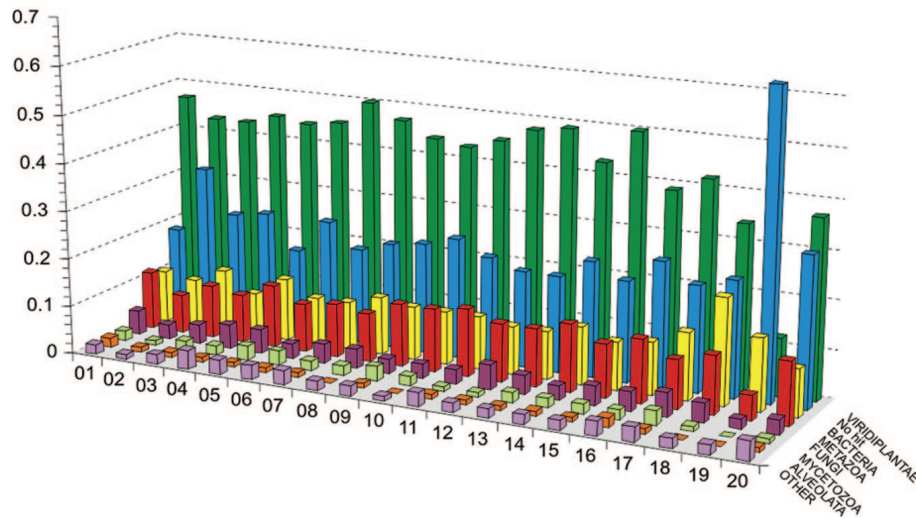


Fig. 3. Taxon distribution of best hits for genes from each of the *O. tauri* chromosomes. Green, viridiplantae; blue, no hit; yellow, bacteria; red, metazoa; pink, fungi; gray, mycetozoa; orange, alveolates; purple, others. Annotation of genes on the low G+C part of chromosome 2 is difficult, and the percentage of genes having no hit on chromosome 2 can be slightly overestimated. See *Genome Heterogeneity* for details.

These two differences make gene modeling more complicated for this region, although at least 61 predicted peptides were supported by ESTs (see Table 6, which is published as supporting information on the PNAS web site). Chromosome 2 small introns differ in many respects from the other introns, such as their size (40–65 bp), composition (they are AT rich and richer by  $\approx 10\%$  than the neighboring exons), and splice sites and branch points that are less conserved than for other introns (Fig. 7b). Interestingly, phylogenetic analysis (see *Materials and Methods*) shows that 43% of the genes on this chromosome, including the small intron-containing genes, have green lineage ancestry (Fig. 3). Of those, 44% cluster specifically (with bootstrap values  $>70\%$ ) with genes of *Chlamydomonas reinhardtii* (data not shown but available on request). Together with the fact that the genes encoded in this region are essential housekeeping genes not duplicated elsewhere in the genome, this observation argues against an alien (horizontal transfer) origin for the low G+C region of chromosome 2. Thus, the origin of the chromosome 2 peculiarities remains elusive. One possibility is that it represents a sexual chromosome. It has been shown before that such chromosomes possess distinctive features for avoiding recombination and are characterized by an unusual richness in transposable elements (29). Meiosis has not been observed in culture, and no equivalent of a mating-type locus has been found akin to that in *C. reinhardtii*. Nevertheless, the presence of most of the core meiotic genes homologous to those identified in other organisms found in *O. tauri* (Table 7, which is published as supporting information on the PNAS web site) is at least a strong indication that *O. tauri* may be a sexual organism (30). Indeed other marine algae known to undergo sexual reproduction commonly suppress this capability in culture (31).

With respect to chromosome 19, phylogenetic analysis shows that only 18% of the peptide-encoding genes are related to the green lineage, a significantly lower percentage than that for the 19 other chromosomes. Others resemble proteins from various origins, mainly bacterial, although generally poorly conserved (Fig. 3; and Table 8, which is published as supporting information on the PNAS web site). Interestingly, most (84%) of the ones having a documented function belong to a few functional categories, primarily encoding surface membrane proteins or proteins involved in the building of glycoconjugates (Table 8). Based on these features, we hypothesize that chromosome 19 is of a different origin than the rest of the genome. This putatively alien material could have yielded some selective advantages in cell surface processes, potentially related, for example, to defense against pathogens or other environmental interactions.

**Genome Compaction.** A second remarkable feature of the *O. tauri* genome is the intense degree of genome compaction, which appears to be the result of several processes. Shortening of intergenic regions is clearly a major factor. The average intergenic size is only 196 bp, which is shorter than that of other eukaryotes having a similar genome size (Table 1). Two other important factors are gene fusion, for which several cases are observed (Fig. 8, which is published as supporting information on the PNAS web site), and reduction of the size of gene families. For example, the gene complement involved in cell division control is one of the most complete across eukaryotes, although there is only one copy of each gene (32). Although this type of reduction is often the case in *O. tauri*, there are some exceptions. For example, the full set of partially redundant enzymes required for polysaccharide metabolism in land plants is present. Here, the maintenance of 27 genes, including multicopy genes, related to synthesis and breakdown of only two types of chemical linkages in the chloroplast, seems excessive for building the semicrystalline starch granule of *O. tauri* (33). Indeed, apicomplexa parasites or even red algae require only 10 genes to build and degrade simple polymers in their cytoplasm (34). *O. tauri* appears to be quite similar to other unicellular organisms in terms of numbers of transcription factors, with no

further reduction than what has commonly been reported. Approximately 2.5–3.8% of predicted proteins of unicellular organisms fall within the category for transcription factors (Table 9, which is published as supporting information on the PNAS web site). This finding is in contrast to multicellular organisms, for which 12–15% of the predicted proteins generally fall within the transcription factor category (see e.g., Table 9).

With respect to pigment biosynthesis and photosynthesis, many genes involved in these pathways are found in multiple copies in other photosynthetic eukaryotes. In *O. tauri*, they also form multigene families, but the copy number is generally lower (e.g., Table 10, which is published as supporting information on the PNAS web site; and see also ref. 35). As expected, *O. tauri* maintains all essential enzymes for carbon fixation (Table 10), and, based on available data for other algae and land plants, homologs are generally present at half the copy number (35). Double sets of several carbon metabolism-related genes, including phosphoglycerate kinase, ribulose-bisphosphate carboxylase, and triosephosphate isomerase, can be found in the *O. tauri* genome. Based on both best hit and subsequent phylogenetic analyses these “doubles” each appear to have different origins (bacterial versus eukaryotic).

***O. tauri* Metabolic Pathways.** *O. tauri* displays some other characteristics unusual for land plants and algae. For instance, the typical genes encoding the major light-harvesting complex proteins associated with photosystem II (LHCII) are lacking. Instead, paralogs encoding prasinophyte-specific chlorophyll-binding proteins are present, making a special antenna as previously observed in *Mantoniella squamata* (35). Interestingly, *O. tauri* also possesses a small set of five *lhca* genes, encoding an LHCI antenna. Combined with the absence of major LHCII protein-encoding genes, this finding supports the hypothesis that the LHCI antenna type is more ancestral than is LHCII (35). Unique features are also seen in the carbon assimilation machinery. Only one carbonic anhydrase (CA), most similar to bacterial  $\beta$ -CA, was identified. No carbon-concentrating mechanism (CCM) genes (36) comparable with those of *C. reinhardtii* or common to organisms that actively or passively enhance inorganic carbon influx were found. However, genes putatively encoding all of the enzymes required for  $C_4$  photosynthesis were identified (see Table 10). Whereas  $C_4$  photosynthesis has yet to be unequivocally shown in unicellular organisms (24, 25, 37, 38),  $C_4$  in the absence of Kranz anatomy is now well documented, especially in *Hydrilla verticillata*, a facultative  $C_4$  aquatic monocot (39). Unlike *T. pseudonana*, which appears to lack plastid-localized NADP-dependent malic enzymes (NADP-ME), *O. tauri* has two NADP-ME orthologs most similar to *H. verticillata* (40) with at least one apparently targeted to the chloroplast based on ChloroP and TargetP predictions. *O. tauri* also has phosphoenolpyruvate (PEP) carboxylase, NADP+ malate dehydrogenase, and pyruvate-orthophosphate dikinase (Table 10), with predicted chloroplast targeting transit peptides in the latter two.  $C_4$  photosynthesis is thought to have evolved multiple times from  $C_3$  ancestors. Although timing is uncertain, it is currently thought to have first evolved 24–35 million years ago in relation to environmental pressures (e.g., declining atmospheric  $CO_2$ ) (36, 38). Interestingly, only one member of the Chlorophyta, the macroalga *Udotea*, has been shown to perform  $C_4$  photosynthesis. *Udotea* utilizes PEP carboxykinase (PEPCK) (NADP-ME being absent) (41), a  $C_4$  photosynthesis form variant to that suggested here, although not yet confirmed experimentally, for *O. tauri*. Despite its energetic cost, if *O. tauri* is capable of  $C_4$  photosynthesis, it could constitute a critical ecological advantage in the  $CO_2$ -limiting conditions of phytoplankton blooms, especially in circumstances where competitors have lower CCM efficiencies (or no CCM at all).

Resource acquisition is critical to survival in the frequently limiting marine environment, and here *O. tauri* seems to have developed competitive strategies currently thought uncommon amongst eukaryotic algae. Nitrogen is typically a major limiting

nutrient of marine phytoplankton growth. *O. tauri* is known to grow on nitrate, ammonium, and urea (9), and complete sets of genes allowing transport and assimilation of these substrates have been identified (Fig. 9 and Table 11, which are published as supporting information on the PNAS web site). Interestingly, four genes encoding ammonium transporters were identified, two being green lineage-related and the other two prokaryote-like. Eukaryotic algae are generally considered ineffective competitors for ammonium; however, the high number of ammonium transporters in *O. tauri* (unlike e.g., *T. pseudonana*) indicates it may be a strong competitor for this resource. All other genes related to nitrogen acquisition and assimilation are found in a single copy, including those for nitrate, again in contrast to *T. pseudonana*. It is notable that eight of the genes involved in nitrate uptake and assimilation are found next to each other on chromosome 10 (Fig. 9A), as well as four genes for urea assimilation genes on chromosome 15 (Fig. 9B). A comparable clustering of nitrate assimilation genes was also observed in *C. reinhardtii* (41) but grouping fewer genes. This organization is reminiscent of prokaryotes, especially cyanobacteria (20), and indicates a possible selective pressure for optimization of nitrate and urea uptake and assimilation, although experimental evidence for the regulation of expression of these genes is currently lacking. The nitrite reductase (NIR) apoenzyme has a unique structure, with two additional redox domains at the C terminus of canonical ferredoxin-NIR, rubredoxin and cytochrome b5 reductase (Fig. 8). This structure should allow this enzyme to use NAD(P)H directly as reducing substrate, which may also contribute to optimization of the pathway. Within this cluster, Snt encodes a protein with weak similarity to sulfate transporters. Nonetheless, its specific position in the cluster suggests that Snt probably encodes a molybdate transporter, a gene predicted to exist but so far unidentified in any species. Taken together with the possibility that *O. tauri* may be capable of C<sub>4</sub> photosynthesis and the relatively high surface area to volume ratio of this tiny phytoplankton, these various ways to optimize nitrogen assimilation could yield a major competitive advantage over other unicellular phytoplankton. This adaptation would be particularly important to its relative success under environmental scenarios, such as intense bloom conditions, where limitation of multiple resources can be encountered.

Finally, *O. tauri* displays a few traits seemingly more characteristic of land plants than green algae. These traits include the absence of genes encoding the three subunits of the light-independent protochlorophyllide reductase. Thus, like angiosperms, chlorophyll can only be synthesized during the day, owing to the light-dependent protochlorophyllide oxido-reductase gene, present in two copies in the genome. In contrast, the large number of kinase-encoding and calcium-binding domains (Table 12, which is published as supporting information on the PNAS web site) suggests that, as in *Arabidopsis* and *Chlamydomonas*, phosphorelay-based calcium-dependent signal transduction systems are commonly used. However, tyrosine kinases appear to be more highly represented in *O. tauri* than in plants, as is also the case in *Chlamydomonas*.

In conclusion, the genome structure of *O. tauri* generally follows predictions of compaction and streamlining that might be driven by its specific lifestyle and ecology. However, the heterogeneity we reveal here concerning two chromosomes raises the challenge of elucidating its origin, which could either be a reminiscence of this alga's ancient nature or on the contrary more recent adaptations to its environmental niche. It also raises the question of whether this type of heterogeneity is in fact not unique to *O. tauri*, but rather a common feature of some eukaryotes, given that current understanding of eukaryotic genomes relies on a genome database so far dominated by "higher organisms". Understanding features specific to success in the marine environment as well as of evolutionary processes within the green lineage relies on new hypotheses and further experimentation for which this complete genome sequence provides a powerful resource. The exceptional features unveiled in the genome of this ubiquitous, ancient, autonomous unicell high-

light the fundamental level at which we might reconsider current paradigms.

## Materials and Methods

**BAC Library.** Genomic DNA was prepared by embedding *O. tauri* cells in agarose strings, subsequently lysed with proteinase K and partially digested by HindIII. DNA fragments were separated according to size by using pulsed-field gel electrophoresis and electroeluted from the gel. DNA fragments were then ligated to pINDIGO BAC5-HindIII cloning ready (Epicentre Technologies) at a molar ratio insert/vector of 10/1. The ligation product was mixed with EC100 electrocompetent cells (Epicentre Technologies) and electroporated. After 20 h at 37°C on LB chloramphenicol (12.5 µg/ml) plates, recombinant colonies were picked into 384-well microtitre plates containing 60 µl of 2YT medium plus 5% glycerol and 12.5 µg/ml chloramphenicol, grown for 18 h at 37°C, duplicated and stored at -80°C. Two BAC libraries having inserts of ≈50 kb and 130 kb, were prepared, representing a 7-fold coverage of the genome. Clones of both libraries were spotted on high-density filters for further hybridizations, and their ends were sequenced.

**Shotgun Libraries.** Purified DNA was broken by sonication, and, after filling ends, DNA fragments ranging from 1 to 5 kb were separated in an agarose gel. Blunt-end fragments were inserted into pBluescript II KS (Stratagene) digested with EcoRV and dephosphorylated. About 60,000 clones were isolated from four independent *O. tauri* shotgun libraries. Plasmid DNA from recombinant *Escherichia coli* strains was extracted according to the TempliPhi method (GE Healthcare), and inserts were sequenced on both strands by using universal forward and reverse M13 primers and the ET DYEnamic terminator kit (GE Healthcare). Sequences were obtained with MegaBace 1000 automated sequencers (GE Healthcare). Data were analyzed, and contigs were assembled by using Phred-Phrap (42) and Consed software packages. Gaps were filled through primer-directed sequencing by using custom made primers.

**cDNA Library.** Two cDNA libraries were generated from cultures grown under different conditions to improve the representation of the expressed sequences. Exponentially growing cells sampled at various stages of the cell cycle of cultures synchronized by light/dark cycles were mixed with a stationary stage culture. Poly(A) mRNAs from the different cultures were isolated and then mixed together. One cDNA library was created in the λ ZAP vector (Stratagene) and the second in the Gateway system according to the manufacturer's instructions (Invitrogen). The average insert size analyzed on agarose gels was ≈1.5 kb for both libraries. Sequences were obtained by using the forward primer, and single reads were assembled in contigs by using Phred-Phrap (42).

**Genome Annotation.** The genomic sequence of *O. tauri* was annotated by using the EuGène (43) gene finding system with Splice-Machine (44) signal sensor components trained specifically on *O. tauri* datasets. A set of 152 GT donor and 152 AG acceptor sites was constructed to optimize the SpliceMachine context representations and to train the splice site sensors that were used to recognize *O. tauri* splice sites. We found GT donor sites to be highly conserved, which resulted in a highly accurate donor site signal sensor. For acceptor sites, the AG consensus pattern was less conserved, whereas the branch point motif was again highly conserved. Splice-Machine was able to extract this branch point pattern and to use it to recognize AG acceptor sets, again resulting in a highly accurate acceptor site sensor. The content sensor used by EuGène to recognize coding sequences is an interpolated Markov model that was computed from 145 *O. tauri* ORFs and 167 intron sequences (used as background). Training EuGène requires the estimation of scaling parameters from known *O. tauri* genes within their genomic

context. As such, 17 genomic *O. tauri* sequences that each contained abutting genes were constructed and used to train EuGène.

Peptides for two deviant chromosomes, numbers 2 and 19, were modeled by using EuGène and SpliceMachine trained specifically on low GC chromosome 2 special genes. A set of 253 GT donor and 253 AG acceptor sites was constructed to optimize the SpliceMachine context representations and to train the sensors used to recognize the splice sites on these two deviant chromosomes. In contrast to the splice sites of the normal *O. tauri* genes, these GT-AG splice sites were less conserved, resulting in less accurate splice site sensors. However, splice site recognition accuracy was boosted by incorporating intron length constraints (introns in these genes are shorter than in so-called normal genes, with lengths typically between 40 and 60 bp, compared with 170–190 bp for the 18 other chromosomes) at the level of gene recognition. The interpolated Markov model used by EuGène to recognize the special coding sequences was computed from 43 *O. tauri* ORFs and 209 intron sequences (used as background). Ten genes within their genomic context were used to optimize the scaling parameters within EuGène.

The data sources used to complement the *ab initio* part of EuGène were composed of *O. tauri* expressed sequence tags (ESTs), proteins, and genomic sequences. ESTs sequenced over the course of the project were aligned on the genome and used as the most reliable source of extrinsic information. For BlastX, the Swissprot protein dataset (v. 42), *C. merolae* proteins (26), publicly available *C. reinhardtii* proteins, and predicted proteins from Sargasso Sea environmental sequences (45) were used in a decreasing order of priority to avoid error propagation, because the latter dataset is the least reliable.

The functional annotation resulted from the synthesis of InterPro and Gene Ontology (GO) assignments based on domain occurrences in the predicted proteins by using the InterPro scripts, BlastP against the clusters of eukaryotic orthologous groups (KOG) database, and a top four of BlastP hits (e-value  $<10^{-5}$ ) against the nonredundant UniProt database. Throughout this process, genes and pathways of particular importance were curated manually by

specialists and integrated into the genome annotation. The resulting database is publicly available at <http://bioinformatics.psb.ugent.be/genomes/ostreococcus-auri/in> a format that includes browse and query options.

**Phylogenetic Analyses.** Homologous genes of *O. tauri* were searched for in public databases by using BlastP. All top hits were retrieved (up to a significant rise in e-value), and the amino acid sequences were aligned by using ClustalW. Alignment columns containing gaps were removed when a gap was present in  $>10\%$  of the sequences. To reduce the chance of including misaligned amino acids, all positions in the alignment left or right from the gap were also removed until a column in the sequence alignment was found where the residues were conserved in all genes included in our analyses. Column conservation was determined as follows: For every pair of residues in the column, the BLOSUM62 value was retrieved. If at least half of the pairs had a BLOSUM62 value = 0, the column was considered as conserved.

Neighbor-joining trees were constructed by using TreeCon (46), based on Poisson- and Kimura-corrected distances. Bootstrap analyses with 500 replicates were performed to test the significance of the nodes. Genes were only ascribed to a certain taxon if supported at a bootstrap level  $>70\%$ .

This paper is dedicated to André Picard, who passed away in November 2004. He made a major contribution to the field of cell biology applied to marine models. We thank B. Khadaroo and C. Schwartz for technical help in preparation of cDNA libraries, X. Sabau for macroarrays, C. Courties and P. Lagoda for discussions, and F. Dierick and E. Bonnet for bioinformatics help. We also thank I. Grigoriev, J. Grimwood, and B. Palenik for sharing information that helped confirm our assembly. This work was supported by the Génomole Languedoc-Roussillon and the French research ministry and by Région Bretagne (Phostreo) Grant 1043-266-2003 (to F.P. and A.Z.W.). A.Z.W. acknowledges support from a Gordon and Betty Moore Foundation investigator grant. S. Robbens thanks the Institute for the Promotion of Innovation by Science and Technology in Flanders. The work presented here was conducted within the framework of the “Marine Genomics Europe” European Network of Excellence (2004–2008) (GOCE-CT-2004-505403).

- Courties, C., Vaquer, A., Troussellier, M., Lautier, J., Chrétiennot-Dinet, M. J., Neveux, J., Machado, C. & Claustre, H. (1994) *Nature* **370**, 255.
- Courties, C., Perasso, R., Chrétiennot-Dinet, M.-J., Gouy, M., Guillou, L. & Troussellier, M. (1998) *J. Phycol.* **34**, 844–849.
- Baldauf, S. L. (2003) *Science* **300**, 1703–1706.
- Yoon, H. S., Hackett, J. D., Ciniglia, C., Pinto, G. & Bhattacharya, D. (2004) *Mol. Biol. Evol.* **21**, 809–818.
- Chrétiennot-Dinet, M.-J., Courties, C., Vaquer, A., Neveux, J., Claustre, H., Lautier, J. & Machado, M. C. (1995) *Phycologia* **34**, 285–292.
- Derelle, E., Ferraz, C., Lagoda, P., Eychenié, S., Cooke, R., Regad, F., Sabau, X., Courties, C., Delseny, M., Demaille, J., et al. (2002) *J. Phycol.* **38**, 1150–1156.
- Diez, B., Pedrós-Alió, C. & Massana, R. (2001) *Appl. Environ. Microbiol.* **67**, 2932–2941.
- Guillou, L., Eikrem, W., Chrétiennot-Dinet, M.-J., Le Gall, F., Massana, R., Romari, K., Pedrós-Alió, C. & Vault, D. (2004) *Protist* **155**, 193–214.
- Worden, A. Z., Nolan, J. K. & Palenik, B. (2004) *Limnol. Oceanogr.* **49**, 168–179.
- Zhu, F., Massana, R., Not, F., Marie, D. & Vault, D. (2005) *FEMS Microbiol. Ecol.* **52**, 79–92.
- Countway, P. D. & Caron, D. A. (2006) *Appl. Environ. Microbiol.* **72**, 2496–2506.
- Worden, A. Z. (2006) *Aquat. Microb. Ecol.* **43**, 165–175.
- Li, W. K. W. (1994) *Limnol. Oceanogr.* **39**, 169–175.
- Fouilland, E., Descolas-Gros, C., Courties, C., Collos, Y., Vaquer, A. & Gasc, A. (2004) *Microb. Ecol.* **48**, 103–110.
- O’Kelly, C. J., Sieracki, M. E., Their, E. C. & Hobson, I. C. (2003) *J. Phycol.* **39**, 850–854.
- López-García, P., Rodríguez-Valera, F., Pedrós-Alió, C. & Moreira, D. (2001) *Nature* **409**, 603–607.
- Moon-van der Staay, S. Y., De Watcher, R. & Vault, D. (2001) *Nature* **409**, 607–610.
- Rodríguez, F., Derelle, E., Guillou, L., Le Gall, F., Vault, D. & Moreau, H. (2005) *Environ. Microbiol.* **7**, 853–859.
- Moore, L. R., Rocap, G. & Chisholm, S. W. (1998) *Nature* **393**, 464–467.
- Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., et al. (2003) *Nature* **424**, 1042–1047.
- Campbell, L., Holla, H. A. & Vault, D. (1994) *Limnol. Oceanogr.* **39**, 954–961.
- Rocap, G., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. (2002) *Appl. Environ. Microbiol.* **68**, 1180–1191.
- Palenik, B., Brahmashya, B., Larimer, F. W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E. E., McCarren, J., et al. (2003) *Nature* **424**, 1037–1042.
- Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., et al. (2004) *Science* **306**, 79–86.
- Raven, J. A. & Kübler, J. E. (2002) *J. Phycol.* **38**, 11–16.
- Matsuzaki, M., Misumi, O., Shin-i, T., Maruyama, S., Takahara, M., Miyagishima, S.-y., Mori, T., Nishida, K., Yagisawa, F., Nishida, K., et al. (2004) *Nature* **428**, 653–657.
- Gilson, P. R. (2001) *Genome Biol.* **2**, 1022.1–1022.5.
- Feschotte, C. & Wessler, S. R. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 280–285.
- Fraser, J. A. & Heitman, J. (2004) *Mol. Microbiol.* **51**, 299–306.
- Ramesh, M. A., Malik, S.-B. & Logsdon, J. M., Jr. (2005) *Curr. Biol.* **15**, 185–191.
- Chepurnov, V. A., Mann, D. G., Sabbe, K. & Vyverman, W. (2004) *Int. Rev. Cytol.* **237**, 91–154.
- Robbens, S., Khadaroo, B., Camasses, A., Derelle, E., Ferraz, C., Inzé, D., Van de Peer, Y. & Moreau, H. (2005) *Mol. Biol. Evol.* **22**, 589–597.
- Ral, J.-P., Derelle, E., Ferraz, C., Wattedled, F., Farinas, B., Corellou, F., Buléon, A., Slomianny, M.-C., Delvalle, D., d’Hulst, C., et al. (2004) *Plant Physiol.* **136**, 3333–3340.
- Coppin, A., Varré, J.-S., Lienard, L., Dauvillée, D., Guérardel, Y., Soyer-Gobillard, M.-O., Buléon, A., Ball, S. & Tomavo, S. (2005) *J. Mol. Evol.* **60**, 257–267.
- Six, C., Worden, A. Z., Rodríguez, F., Moreau, H. & Partensky, F. (2005) *Mol. Biol. Evol.* **22**, 2217–2230.
- Giordano, M., Beardall, J. & Raven, J. A. (2005) *Annu. Rev. Plant Biol.* **56**, 99–131.
- Reinfelder, J. R., Milligan, A. J. & Morel, F. M. M. (2004) *Plant Physiol.* **135**, 2106–2111.
- Sage, R. F. (2004) *New Phytol.* **161**, 341–370.
- Rao, S. K., Magnin, N. C., Reiskind, J. B. & Bowes, G. (2002) *Plant Physiol.* **130**, 876–886.
- Bowes, G., Rao, S. K., Estavillo, G. M. & Reiskind, J. B. (2002) *Funct. Plant Biol.* **29**, 379–392.
- Quesada, A., Galván, A., Schnell, R. A., Lefebvre, P. A. & Fernández, E. (1993) *Mol. Gen. Genet.* **240**, 387–394.
- Ewing, B., Hillier, L., Wendl, M. C. & Green, P. (1998) *Genome Res.* **8**, 175–185.
- Schiex, T., Moisan, A. & Rouzé, P. (2001) *Lect. Notes Comput. Sci.* **2066**, 111–125.
- Degroeve, S., Saeyes, Y., De Baets, B., Rouzé, P. & Van de Peer, Y. (2005) *Bioinformatics* **21**, 1332–1338.
- Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., et al. (2004) *Science* **304**, 66–74.
- Van de Peer, Y. & De Wachter, R. (1997) *Comput. Appl. Biosci.* **13**, 227–230.
- Abrahamson, M. S., Templeton, T. J., Enomoto, S., Abrahamant, J. E., Zhu, G., Lancto, C. A., Deng, M., Liu, C., Widmer, G., Zipori, S., et al. (2004) *Science* **304**, 441–445.
- Haas, B. J., Wortman, J. R., Ronning, C. M., Hannick, L. I., Smith, R. K., Jr., Maiti, R., Chan, A. P., Yu, C., Farzad, M., Wu, D., et al. (2005) *BMC Biol.* **3**, 7.
- Dietrich, F. S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pöhlmann, R., Luedi, P., Choi, S., et al. (2004) *Science* **304**, 304–307.